Methods in Molecular Biology 2222

Springer Protocols

Pascale Besse *Editor*

Molecular Plant Taxonomy

Methods and Protocols

Second Edition

MOREMEDIA

Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, UK

For further volumes:
http://www.springer.com/series/7651

For over 35 years, biological scientists have come to rely on the research protocols and methodologies in the critically acclaimed *Methods in Molecular Biology* series. The series was the first to introduce the step-by-step protocols approach that has become the standard in all biomedical protocol publishing. Each protocol is provided in readily-reproducible step-by-step fashion, opening with an introductory overview, a list of the materials and reagents needed to complete the experiment, and followed by a detailed procedure that is supported with a helpful notes section offering tips and tricks of the trade as well as troubleshooting advice. These hallmark features were introduced by series editor Dr. John Walker and constitute the key ingredient in each and every volume of the *Methods in Molecular Biology* series. Tested and trusted, comprehensive and reliable, all protocols from the series are indexed in PubMed.

Molecular Plant Taxonomy

Methods and Protocols

Second Edition

Edited by

Pascale Besse

UMR PVBMT, Universite de la Reunion, St Pierre, Réunion, France

Humana Press

Editor
Pascale Besse
UMR PVBMT
Universite de la Reunion
St Pierre, Réunion, France

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-0716-0996-5 ISBN 978-1-0716-0997-2 (eBook)
https://doi.org/10.1007/978-1-0716-0997-2

This Humana imprint is published by the registered company Springer Science+Business Media, LLC, part of Springer Nature.
The registered company address is: 1 New York Plaza, New York, NY 10004, U.S.A.

Preface

Plant taxonomy is an ancient discipline facing new challenges with the availability of a vast array of modern molecular technologies. The literature reviews and protocols that appear as chapters in this book were selected to provide conceptual as well as technical guidelines to plant taxonomists and geneticists. This second edition of *Molecular Plant Taxonomy* appeared necessary to take into account the increasing use of next-generation sequencing (NGS) technologies for many applications in plant taxonomy.

The introductive Chapter 1 allows the reader to travel through the historical aspects of plant taxonomy with a focus on the strengths, limitations, and the future of molecular techniques. Chapter 2 then proposes guidelines to choose the best sequence and molecular technique to be used according to the taxonomic question addressed. A temporal landscape of the most commonly used techniques is also provided. Both chapters are prerequisite readings to understand the concepts underlying the "plant taxonomy" discipline and to fully appreciate the strengths and limits of each molecular technique presented in this book.

DNA extraction protocols specifically focused on recalcitrant plant species (Chapter 3) and herbarium specimens (Chapter 4) are proposed. The latter will ensure the development of an "integrative" taxonomic approach by allowing the use of ancient DNA references from herbarium specimens together with present-date accessions in DNA analyses.

Next-generation sequencing technologies have opened a new era for molecular plant taxonomy. This revised edition provides literature review and wet-lab protocols and/or decision flowcharts covering whole chloroplast (Chapter 5) and mitochondrial (Chapter 6) genome sequencing, now more and more replacing the Sanger sequencing of specific regions described in the earlier version of this book. We also chose to present an updated protocol for microsatellite markers isolation based on Illumina sequencing (Chapter 11) to complement classical enriched library construction described in the first version. This NGS-based method is powerful enough to reveal numerous microsatellite loci, which are markers of choice for molecular plant taxonomy. New methods to discover single nucleotide polymorphism (SNP) markers from sequenced pangenomes (Chapter 9) are also described, together with the simple and powerful genotyping-by-sequencing (GBS) method to develop SNP markers without any need for whole genome sequencing and assembly, perfectly suited for many plant species (Chapter 10).

This book still provides detailed literature reviews and detailed wet-lab protocols for many multilocus PCR-based profiling methods that have been shown to be very efficient in resolving many molecular plant taxonomy issues: amplified fragment length polymorphism (AFLP, Chapter 12), random amplified polymorphic DNA (RAPD, Chapter 13) and their multiple derived techniques, inter-simple sequence repeats (ISSR, Chapter 14), and the use of a range of methods tagging retrotransposable elements (Chapter 15). It also provides a protocol for Sanger sequencing and data analysis of the widely used internal transcribed spacer (ITS) nuclear region in plants (Chapter 7), and the usefulness and power of this ITS region together with that of various chloroplast regions as a "DNA barcoding" tool is reviewed and assessed (Chapter 8): it is now clear that using these simple "barcode tools" as defined by the CBOL (consortium for the barcoding of life) for resolving plant taxonomy will not be sufficient, particularly in some plant groups. We rather highly recommend that molecular approaches are used within an "integrative taxonomy" framework, combining a

range of nucleic acid and cytogenetic data together with other crucial information (taxonomy, morphology, anatomy, ecology, reproductive biology, biogeography, paleobotany, etc.), which will help not only to best circumvent species delimitation but also to resolve the evolutionary processes in play. In this respect, Chapters 17, 18, and 19, covering cytogenetic techniques such as flow cytometry, chromosome banding, fluorescent in situ hybridization (FISH), and genomic in situ hybridization (GISH), are essential to provide tools allowing the assessment of plant genome size, ploïdy, aneuploidy, reproductive mode, species relationships, and interspecific hybrids. Moreover, the generation of large sets of SNP markers through NGS technologies now allows detailed population genomics studies (Chapter 16) that can help to resolve the evolutionary processes in play in natural populations through the analysis of population structure, the inference of population splits and exchanges, and the detection of footprints of natural or artificial selection. Although the primary focus of plant taxonomy is on the delimitation of species, molecular approaches now provide a better understanding of evolutionary processes, at species and population level, a particularly important issue for some taxonomic complex groups and a prerequisite to resolve speciation processes. This is essential when one wants to apply plant taxonomy to conservation issues.

St Pierre, Réunion, France *Pascale Besse*

Contents

Contributors

Kantipudi Nirmal Babu • *Indian Institute of Spices Research, Kozhikode, Kerala, India*

Philipp E. Bayer • *School of Biological Sciences, University of Western Australia, Perth, Australia*

Guillaume Besnard • *CNRS-UPS-IRD, UMR5174, EDB, Université Paul Sabatier, Toulouse, France*

Pascale Besse • *UMR PVBMT, Universite de la Reunion, St Pierre, Réunion, France*

Svetlana Boronnikova • *Department of Botany and Genetics of Plants, Faculty of Biology, Perm State University, Perm, Russia*

Carine Charron • *CIRAD, UMR PVBMT, St Pierre, La Réunion, France*

Monica F. Danilevicz • *School of Biological Sciences, University of Western Australia, Perth, Australia*

Denis Da Silva • *Université de La Réunion, UMR PVBMT, St Pierre, La Réunion, France*

Jérôme Duminil • *DIADE, University of Montpellier, IRD, Montpellier, France*

David Edwards • *School of Biological Sciences, University of Western Australia, Perth, Australia*

Félicien Favre • *Université de La Réunion, UMR PVBMT, St Pierre, La Réunion, France*

Cassandria G. Tay Fernandez • *School of Biological Sciences, University of Western Australia, Perth, Australia*

Myriam Gaudeul • *Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, Sorbonne Université, Ecole Pratique des Hautes Etudes, Université des Antilles, CNRS, Paris, France*

Chrissen E. C. Gemmill • *School of Science, University of Waikato, Hamilton, New Zealand*

Ananduchandra Giridhari • *Indian Institute of Spices Research, Kozhikode, Kerala, India*

Ella R. P. Grierson • *Plant & Food Research, Palmerston North, New Zealand*

Michel Grisoni • *CIRAD, UMR PVBMT, St Pierre, La Réunion, France*

Berthold Heinze • *Department of Genetics, Austrian Federal Research Centre for Forests (BFW), Vienna, Austria*

Cyril Jourda • *CIRAD, UMR PVBMT, St Pierre, La Réunion, France*

Ruslan Kalendar • *Department of Agricultural Sciences, Viikki Plant Science Centre and Helsinki Sustainability Centre, University of Helsinki, Helsinki, Finland; National Laboratory Astana, Nazarbayev University, Nur-Sultan, Kazakhstan*

Illathidath Payatatti Vijesh Kumar • *Indian Institute of Spices Research, Kozhikode, Kerala, India*

Ilia J. Leitch • *Department of Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, Richmond, Surrey, UK*

Thibault Leroy • *Montpellier Institute of Evolutionary Sciences (ISEM), Université de Montpellier, Montpellier, France; Department of Botany and Biodiversity Research, University of Vienna, Vienna, Austria*

Jacob I. Marsh • *School of Biological Sciences, University of Western Australia, Perth, Australia*

Divakaran Minoo • *Providence Women's College, Kozhikode, Kerala, India*

Alexander Muterko • *The Federal Research Center Institute of Cytology and Genetics, Novosibirsk, Russian Federation*

Jaume Pellicer • *Department of Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, Richmond, Surrey, UK; Department of Biodiversity, Institut Botànic de Barcelona (IBB, CSIC-Ajuntament de Barcelona), Barcelona, Spain*
Nathalie Piperidis • *SRA, Sugar Research Australia, Te Kowai, QLD, Australia*
Robyn F. Powell • *Department of Comparative Plant and Fungal Biology, Royal Botanic Gardens, Kew, Richmond, Surrey, UK*
Fatima Pustahija • *Faculty of Forestry, University of Sarajevo, Sarajevo, Bosnia and Herzegovina*
Muliyar Krishna Rajesh • *Central Plantation Crops Research Institute, Kasaragod, Kerala, India*
Ronan Rivallan • *CIRAD, UMR AGAP, Montpellier, France; AGAP, University of Montpellier, CIRAD, INRAe, Montpellier SupAgro, Montpellier, France*
Odile Robin • *University Paris-Saclay, CNRS, AgroParisTech, Ecologie Systématique Evolution, Orsay, France*
Quentin Rougemont • *Département de Biologie, Institut de Biologie Intégrative et des Systèmes (IBIS), Université Laval, Quebec, QC, Canada*
Germinal Rouhan • *Institut de Systématique, Evolution, Biodiversité (ISYEB), Muséum national d'Histoire naturelle, Sorbonne Université, Ecole Pratique des Hautes Etudes, Université des Antilles, CNRS, Paris, France*
Kukkamgai Samsudeen • *Central Plantation Crops Research Institute, Kasaragod, Kerala, India*
Mervi Seppänen • *Department of Agricultural Sciences, Viikki Plant Science Centre and Helsinki Sustainability Centre, University of Helsinki, Helsinki, Finland*
Thotten Elampilay Sheeja • *Indian Institute of Spices Research, Kozhikode, Kerala, India; Division of Crop Improvement and Biotechnology, ICAR-Indian Institute of Spices Research, Kozhikode, Kerala, India*
Sonja Siljak-Yakovlev • *University Paris-Saclay, CNRS, AgroParisTech, Ecologie Systématique Evolution, Orsay, France*
Erinjery Jose Suraby • *Indian Institute of Spices Research, Kozhikode, Kerala, India*
Vedrana Vičić-Bočkor • *Faculty of Science, Department of Molecular Biology, University of Zagreb, Zagreb, Croatia*
Hélène Vignes • *CIRAD, UMR AGAP, Montpellier, France; AGAP, University of Montpellier, CIRAD, INRAe, Montpellier SupAgro, Montpellier, France*
Lenka Záveská Drábková • *Laboratory of Pollen Biology, Institute of Experimental Botany of the Czech Academy of Sciences, Prague, Czech Republic*

List of Abbreviations

ADBC	Advancing digitization of biological collections
AFLP	Amplified fragment length polymorphism
AFLP-RGA	Resistance gene analog anchored AFLP
AIMS	Amplification of insertion mutagenized sites
AMP-PCR	Anchored microsatellite-primed PCR
APG	Angiosperm phylogeny group
AP-PCR	Arbitrary primed polymerase chain reaction
CAPS	Cleaved amplified polymorphic sequences
CBC	Compensatory base change
CBD	Convention on biological diversity
CBOL	Consortium for barcoding of life
cDNA-AFLP	Complementary DNA AFLP
CETAF	Consortium of European taxonomic facilities
CID	Cultivar identification diagram
CNV	Copy number variation
cpDNA	Chloroplast DNA
CTAB	Cetyltrimethyl ammonium bromide
DAF	DNA amplification fingerprinting
DArT	Diversity array technology
DD	Differential display
DNA	Deoxyribonucleic acid
DOI	Digital object identifiers
EDGF	Enzymatic digestion and glass-fiber filtration
EDIT	European Distributed Institute of Taxonomy
ERV	Endogenous retrovirus
FCM	Flow cytometry
FCSS	Flow cytometric seed screening
FISH	Fluorescence in situ hybridization
GBS	Genotyping by sequencing
GEA	Genotype-environment associations
GISH	Genomic in situ hybridization
GPA	Genotype-phenotype associations
GSPC	Global strategy for plant conservation
GTI	Global taxonomic initiative
HTS	High-throughput sequencing technologies
IAM	Infinite allele model
ICN	International Code of Nomenclature for algae, fungi, and plants
iDigBio	Integrated Digitized Biocollections
Indel	Insertion/deletion
iPBS	Inter-PBS amplification
IPNI	International Plant Names Index
IRAP	Inter-retrotransposons amplified polymorphism
ISSR	Inter-simple sequence repeats
ISTR	Inverse sequence tagged repeat
ITS	Internal transcribed spacer

LARDs	Large retrotransposon derivatives
LCNG	Low-copy nuclear genes
LD	Linkage disequilibrium
LGT	Lateral gene transfer
LINE	Long interspersed repetitive element
LTR	Long terminal repeat
M-AFLP	Microsatellite-amplified fragment length polymorphism
MGE	Mobile genetic element
MITE	Miniature inverted-repeat transposable element
MOTU	Molecular operational taxonomic units
MSAP	Methylation-sensitive amplified polymorphism
mtDNA	Mitochondrial DNA
mtpt	Plastid-derived region
nDNA	Nuclear DNA
NGS	Next-generation sequencing
NOR	Nucleolar organizing region
nrDNA	Nuclear ribosomal RNA genes
NSF	National Science Foundation
ORF	Open reading frame
PAV	Presence and absence variation
PBS	Primer binding site
PCA	Principal component analysis
PCR	Polymerase chain reaction
PEET	Partnerships for Enhancing Expertise in Taxonomy
pGWAS	Population genome-wide association study
PLOP	Pre-labeled oligonucleotide probes
QTL	Quantitative trait loci
RAD	Restriction-site reduced complexity approach
RAHM	Random amplified hybridization microsatellites
RAMPO	Random amplified microsatellite polymorphism
RAPD	Randomly amplified polymorphic DNA
rDNA	Ribosomal DNA
RE	Restriction Enzyme
REMAP	Retrotransposon-microsatellite amplification polymorphism
RFLP	Restriction fragments length polymorphism
RL	Restriction ligation
RNA	Ribonucleic acid
RTE	Retrotransposable elements
SAMPL	Selective amplification of microsatellite polymorphic loci
SCAR	Sequence characterized amplified region
SDAFLP	Secondary digest AFLP
SINE	Short interspersed nuclear element
SMM	Stepwise mutation model
SNP	Single nucleotide polymorphism
SRAP	Sequence-related amplified polymorphism
SSAP	Sequence-specific amplified polymorphism
SSCP	Single strand conformation polymorphism
SSR	Simple sequence repeats
STR	Simple tandem repeats

TCG	Taxonomic complex group
TDF	Transcript-derived fragment
TE	Transposable element
TE-AFLP	Three endonucleases AFLP
TRIM	Terminal-repeat retrotransposons in miniature
TSDs	Target site duplications
VNTR	Variable number of tandem repeats

Chapter 1

Plant Taxonomy: A Historical Perspective, Current Challenges, and Perspectives

Germinal Rouhan and Myriam Gaudeul

Abstract

Taxonomy is the science that explores, describes, names, and classifies all organisms. In this introductory chapter, we highlight the major historical steps in the elaboration of this science, which provides baseline data for all fields of biology and plays a vital role for society but is also an independent, complex, and sound hypothesis-driven scientific discipline.

In a first part, we underline that plant taxonomy is one of the earliest scientific disciplines that emerged thousands of years ago, even before the important contributions of the Greeks and Romans (e.g., Theophrastus, Pliny the Elder, and Dioscorides). In the fifteenth–sixteenth centuries, plant taxonomy benefited from the Great Navigations, the invention of the printing press, the creation of botanic gardens, and the use of the drying technique to preserve plant specimens. In parallel with the growing body of morpho-anatomical data, subsequent major steps in the history of plant taxonomy include the emergence of the concept of natural classification, the adoption of the binomial naming system (with the major role of Linnaeus) and other universal rules for the naming of plants, the formulation of the principle of subordination of characters, and the advent of the evolutionary thought. More recently, the cladistic theory (initiated by Hennig) and the rapid advances in DNA technologies allowed to infer phylogenies and to propose true natural, genealogy-based classifications.

In a second part, we put the emphasis on the challenges that plant taxonomy faces nowadays. The still very incomplete taxonomic knowledge of the worldwide flora (the so-called taxonomic impediment) is seriously hampering conservation efforts that are especially crucial as biodiversity has entered its sixth extinction crisis. It appears mainly due to insufficient funding, lack of taxonomic expertise, and lack of communication and coordination. We then review recent initiatives to overcome these limitations and to anticipate how taxonomy should and could evolve. In particular, the use of molecular data has been era-splitting for taxonomy and may allow an accelerated pace of species discovery. We examine both strengths and limitations of such techniques in comparison to morphology-based investigations, we give broad recommendations on the use of molecular tools for plant taxonomy, and we highlight the need for an integrative taxonomy based on evidence from multiple sources.

Key words Classification, Floras, DNA, History, Molecular taxonomy, Molecular techniques, Morpho-anatomical investigations, Plant taxonomy, Species, Taxonomic impediment

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222,
https://doi.org/10.1007/978-1-0716-0997-2_1,

1 Introduction

Adapting the famous aphorism of Theodosius Dobzhansky [1], could we dare to say that nothing in biology makes sense except in the light of taxonomy? Maybe yes, considering that most of biology relies on identified—and so described—species that are end products of taxonomy. Taxonomic information is obviously crucial for studies that analyze the distribution of organisms on Earth, since they need taxonomic names for inventories and surveys. But names are also needed to report empirical results from any other biological study dealing with, e.g., biochemistry, cytology, ecology, genetics, or physiology: even if working an entire life on a single species, e.g., *Arabidopsis thaliana* (L.) Heynh., a molecular biologist will focus all his/her research on numerous plants that all represent this species as delimited by taxonomy. Thus, taxonomy provides names, but it is not only a 'biodiversity-naming' service: it is also a scientific discipline requiring theoretical, empirical, and epistemological rigor [2]. Names represent scientific hypotheses on species boundaries, and to put forward such hypotheses involves gathering information from characters of the organisms and adopting a species concept (*see* **Note 1** for an overview of the main species concepts). Morphology, anatomy, and genetics are the main sources of characters used in today's plant taxonomy. Not without noting that these types of characters all bring potentially valuable evidence, the focus of this book is on the use of nucleic acids—and genome size and chromosomes—for a reliable and efficient taxonomy.

Before discussing how to choose genomic regions to be studied in order to best deal with particular taxonomic issues (Chapter 2), this chapter aims to summarize the history of taxonomy and to highlight that plant molecular taxonomy emerged from an ancient discipline that has been, and is still, central to other scientific disciplines and plays a vital role for society. We will also give a brief overview of the general background into which plant taxonomy is performed today and propose some general considerations about molecular taxonomy.

2 Taxonomy and Taxon: Terminology and Fluctuating Meanings

It is not before 1813 that the Swiss botanist Augustin Pyramus De Candolle (1778–1841) invented the neologism 'taxonomy' from the Greek *ταξις* (order) and *νόμος* (law, rule) and published it for the first time in his book *Théorie élémentaire de la Botanique* ('Elementary Theory of Botany,' [3]). He defined this scientific discipline as the 'theory of the classifications applied to the vegetal kingdom,' which he considered as one of the three components

of botany along with glossology—'the knowledge of the terms used to name plant organs,' and phytography—'the description of plants in the most useful way for the progress of science.'

Much later, the Global Biodiversity Assessment of the United Nations Environment Programme (UNEP; [4]) defined taxonomy as 'the theory and practice of classifying organisms,' including the classification itself but also the delimitation and description of taxa, their naming, and the rules that govern the scientific nomenclature. Today, depending on the authors, taxonomy is viewed either as a synonym for the 'systematics' science—also called biosystematics [5, 6]—including the task of classifying species, or only as a component of systematics restricted to the delimitation, description, and identification of species. This latter meaning of taxonomy emerged lately, with the advent of phylogenetics as another component of systematics that allows classifications based on the evolutionary relationships among taxa [7].

Thus, it is ironical that taxonomy and systematics, which deal in particular with classifications and relationships between organisms, often themselves require clarifications on their relative circumscriptions and meanings before being used [8]. This book will consider plant taxonomy in the broadest sense, from, e.g., species delimitation based on different molecular techniques—to focuses on population genomics methods, or studies resolving interspecific and intergeneric hybrids.

Incidentally, it is interesting to note that the word 'taxon'—plural: taxa—was invented much later (Lam, in [9]) than 'taxonomy': a taxon is a theoretical entity intended to replace terms such as 'taxonomic group' or 'biodiversity unit' [10], and 'taxon' refers to a group of any rank in the hierarchical classification, e.g., species, genus, or family.

3 A Historical Perspective to Plant Taxonomy

3.1 One of the Earliest Scientific Disciplines

Delimiting, describing, naming, and classifying organisms are activities whose origins are obviously much older than the word 'taxonomy'—which dates back to the nineteenth century; see above. The use of oral classification systems likely even predated the invention of the written language ca. 5600 years ago. Then, as for all vernacular classifications, the precision of the words used to name plants was notably higher for plants that were used by humans. There was no try to link names and organisms in hierarchical classifications since the known plants were all named following their use: some were for food, others for medicines, poisons, or materials. As early as that time, several hundreds of plant organisms of various kinds were identified, while relatively few animals were known and named—basically those that were hunted or feared [11].

These early classifications, that were exclusively utilitarian, persisted until the fifteenth–sixteenth centuries although some major advances were achieved, mainly by ancient Greeks and Romans. It was perceptible that the Greeks early considered plants not just as useful, but also as beautiful, taking a look at paintings in Knossos (1900 BC) that indeed show useful plants like barley, fig, and olive, but also narcissus, roses, and lilies. The Greek Theophrastus (372–287 BC), famous as the successor of Aristotle at the head of the Lyceum, is especially well known as the first botanist and the author of the first written works on plants. Interested in naming plants and finding an order in the diversity of plants, he could have been inspired by Aristotle who started his *Metaphysics* book with the sentence: 'All men by nature desire to know.' Theophrastus is indeed the first one to provide us with a philosophical overview of plants, pointing out important fundamental questions for the development of what will be later called taxonomy, such as 'what have we got?' or 'how do we differentiate between these things?' He was moreover the first one to discuss relationships among plants, and to suggest ways to group them not just based on their usefulness or uses. Thus, in his book *Enquiry into Plants*, he described ca. 500 plants—probably representing all known plants at that time—that he classified as trees, shrubs, undershrubs, and herbs. He also established a distinction between flowering and nonflowering plants, between deciduous and evergreen trees, and between plants that grew in water and those that did not. Even if 80% of the plants included in his works were cultivated, he had realized that 'most of the wild kinds have no names, and few know about them,' highlighting the need to recognize, describe, and name plants growing in the wild [12]. Observing and describing the known plants, he identified many characters that were valuable for later classifications. For instance, based on his observations of plants sharing similar inflorescences—later named 'umbels'—he understood that, generally, floral morphology could help to cluster plants into natural groups and, several centuries later, most of these plants showing umbels were indeed grouped in the family Umbelliferae—nowadays Apiaceae.

Theophrastus was way ahead of his time, to such a point that his botanical ideas and concepts became lost during many centuries in Europe. But his works survived in Persia and Arabia, before being translated back into Greek and Latin and rediscovered in Europe in the fifteenth century. During this long Dark Age for botany—like for all other natural sciences—in Europe, the Roman Pliny the Elder (23–79 AD) and the Greek Dioscorides (~40–90 AD), in first century AD, have however been two important figures. Although they did not improve the existing knowledge and methods about the description, naming, or classifications of plants, they compiled the available knowledge and their written works were renowned and widely used. The *Naturalis Historia* of Pliny

(77 AD) was indeed a rich encyclopedia of the natural world, gathering 20,000 facts and observations reported by other authors, mostly from Greeks like Theophrastus. At the same time in Greece, plants were almost only considered and classified in terms of their medical properties. The major work of Dioscorides *De Materia Medica* (ca. 77 AD) was long the sole source of botanical information (but at that time, botany was only considered in terms of pharmacology) and was repeatedly copied until the fifteenth century in Europe. Juliana's book—*Juliana Anicia Codex*, sixth century; Fig. 1—is the most famous of these copies, well known because it innovated by adding beautiful and colorful plants illustrations to the written work of Dioscorides. If some paintings could be seen as good visual aids to identification—which should be considered as an advance for taxonomy—others, however, were fanciful [12]. All those plant books, called 'herbals' and used by herbalists—who had some knowledge about remedies extracted from plants—throughout the Middle Ages, did not bring any other substantial progress.

3.2 Toward a Scientific Classification of Plants

With the Renaissance, the fifteenth and sixteenth centuries saw the beginning of the Great Navigations—e.g., C. Columbus discovered the New World from 1492; Vasco da Gama sailed all around Africa to India from 1497; F. Magellan completed the first circumnavigation of Earth in 1522—allowing to start intensive and large-scale naturalist explorations around the world: most of the major territories, except Australia and New Zealand, were discovered as soon as the middle of the sixteenth century, greatly increasing the number of plants that were brought back in Europe either by sailors themselves or naturalists on board. At that time, herbalists still played a major role in naming and describing plants, in association with illustrators who were producing realistic illustrations. But naming and classifying so numerous exotic and unknown plants from the entire world would not have been possible without three major inventions. Firstly, the invention of the Gutenberg's printing press with moveable type system (1450–1455) made written works on plants largely available in Europe—the first Latin translation of Theophrastus' books came out in 1483. Secondly, the first botanic gardens were created in Italy in the 1540s, showing the increasing interest of the population for plants and allowing teaching botany. Thirdly, in the botanic garden of Pisa, the Italian Luca Ghini (1490–1556) invented a revolutionary method for preserving—and so studying—plants, consisting in drying and pressing plants to permanently store them in books as 'hortus siccus' (dried garden), today known as 'herbaria'—or 'herbarium specimens.' These perennial collections of dried plants were—and are still—a keystone element for plant taxonomy and its development: from that time, any observation and experimental result could be linked to specific plant specimens available for further identification, study of

Fig. 1 Painting of a *Cyclamen* plant, taken from the Juliana's book, showing the flowering stems rising from the upper surface of the rounded corm. According to Dioscorides, those plants were used as purgatives, antitoxins, skin cleansers, labor inducers, and aphrodisiacs

morphology, geographic distribution, ecology, or any other features. In short, Ghini provided with herbaria the basis of reproducibility that is an essential part of the scientific method [13].

A student of Ghini, Andrea Cesalpino (1519–1603), was the first one since the Ancient Greeks to take over the work of Theophrastus, and to discuss it. He highlighted that plants should be

classified in a more natural and rational way than the solely utilitarian thinking. Convinced that all plants have to reproduce, he provided a new classification system primarily based on seeds and fruits: in *De Plantis libri XVI* (1583), he described 1500 plants that he organized into 32 groups such as the Umbelliferae and Compositae—currently Apiaceae and Asteraceae, respectively. Cesalpino also made a contribution to the naming of plant names, sometimes adding adjectives to nouns designing a plant, e.g., he distinguished *Edera spinosa* (spiny ivy) from *Edera terrestris* (creeping ivy). This could be seen as a prefiguration of the binomial naming system that was established in the eighteenth century and is still used in taxonomy. But the science of scientific naming was only starting and plants—like other living beings—were usually characterized by several words forming polynomial Latin names: for instance, tomato was designed as *Solanum caule inermi herbaceo, foliis pinnatis incisis*, which means 'Solanum with a smooth herbaceous stem and incised pinnate leaves' [14] (Fig. 2).

Cesalpino contributed to the emergence of the concept of natural classification, i.e., a classification reflecting the 'order of Nature.' This latter expression involved different interpretations and classifications through the history of taxonomy, but a natural classification was always intended to reflect the relationships among plants. Because the Evolutionary thought was not developed yet, it basically resulted in clustering plants with similar morphological features. So, it must be noted that the distinction between artificial and natural classifications—respectively named 'systems' and 'methods' at the end of the eighteenth century—is a modern interpretation of the past classifications. Taking advantage of both technical progresses like microscopy—in the seventeenth century—and scientific methods inspired by Descartes (1596–1650), several attempts were made to reach such a natural classification. For example, Bachmann—also known as Rivin or Rivinus (1652–1723)—based his classification on the corolla shape in *Introductio* ad rem *herbariam* in 1690. Altogether, the major interest of these classifications is that they triggered investigations on many morpho-anatomical characters that could be used by later taxonomists to describe and circumscribe plant species. The British John Ray (1627–1705) innovated by not relying anymore on a single characteristic to constitute groups of plants: he suggested natural groupings 'from the likeliness and agreement of the principal parts' of the plants, based on many characters—mostly relative to leaves, flowers, and fruits. He documented more than 17,000 worldwide species in *Historia Plantarum* (1686–1704) and distinguished flowering vs. nonflowering plants, and plants with one cotyledon, which he named 'monocotyledons,' vs. plants with two cotyledons, 'dicotyledons.' Ray also played a major role in the development of plant taxonomy—and more generally of plant science—by creating the first text-based dichotomous keys that he used as a means to classify plants [15].

Fig. 2 Herbarium specimen from the Tournefort's Herbarium (housed at the Paris national Herbarium, Muséum national d'Histoire naturelle, MNHN) displaying a label with the hand-written polynomial name '*Aconitum caeruleum, glabrum, floribus consolid(ae) regalis*'

In contrast to Ray and his method intended to be natural, his French contemporary Joseph Pitton de Tournefort (1656–1708) explored, in his *Elements de Botanique* (1694), the possibility of classifying plants based on only a few characteristics related to the corolla of flowers, creating an artificial system. The success of Tournefort's system resulted from the ease to identify groups of plants based on the number and relative symmetry of the petals of a

flower. Within his system, Tournefort precisely defined 698 entities—'*Institutiones rei herbariae*,' 1700—each being called a genus, plural: genera. The genus concept was new and contributed to a better structuration of the classification.

3.3 Naming Plant Names: Major Advances by Linnaeus

In spite of the numerous new ideas and systems produced from the 16th to the middle of the eighteenth century, names of plants still consisted in polynomial Latin names, i.e., a succession of descriptors following the generic name. This led to a rather long, complicated, and inoperative means to designate plants and became problematic in the context of the Great Explorations, which allowed the discovery of more and more plants from all over the world (major explorations with naturalists on board included, e.g., the circumnavigation of La Boudeuse under Bougainville from 1766 to 1769, and the travels to the Pacific of J. Cook between 1768 and 1779). To overcome this impediment involving the naming of plants, the Swedish Carolus Linnaeus (1707–1778) took a critical step forward for the development of taxonomy.

He suggested dissociating the descriptors of the plant from the name itself, because according to him, the name should only serve to designate the plant. Therefore, he assigned a 'trivial name' to each plant (more than 6000 plants in *Species Plantarum*, 1753) [16] and this name was binomial, only consisting of two words: the 'genus' followed by the 'species,' e.g., *Adiantum capillus-veneris* is a binomen created by Linnaeus that is still known and used as such to designate the Venus-hair fern. Although there had been some attempts of binomials as early as Theophrastus (followed by Cesalpino and a few others), Linnaeus succeeded in popularizing his system as new, universal—applied for all plants and, later on, even for animals in *Systema Naturae* [17], and long-lasting. Truly, *Species Plantarum* [16] has been a starting point for setting rules in plant taxonomy. Used since Linnaeus until today, the binomial system along with other principles for the naming of plants were developed, standardized, synthesized, and formally accepted by taxonomists into a code of nomenclature—initially called 'Laws of botanical nomenclature' [18] and nowadays called the International Code of Nomenclature for algae, fungi, and plants (ICN). The current code is slightly evolving every 6 years, after revisions are adopted at an international botanical congress.

Linnaeus also proposed his own artificial classification. With the goal to describe and classify all plants—and other living beings—that were 'put on Earth by the Creator,' he grouped them based on the number and arrangement of stamens and pistils within flowers—contrary to Tournefort, who only focused on petals. He called this classification a 'sexual system,' referring to the fundamental role of flowers in sexual reproduction (Fig. 3). This system included five hierarchical categories: varieties, species, genera, orders—equivalent to current families, and classes.

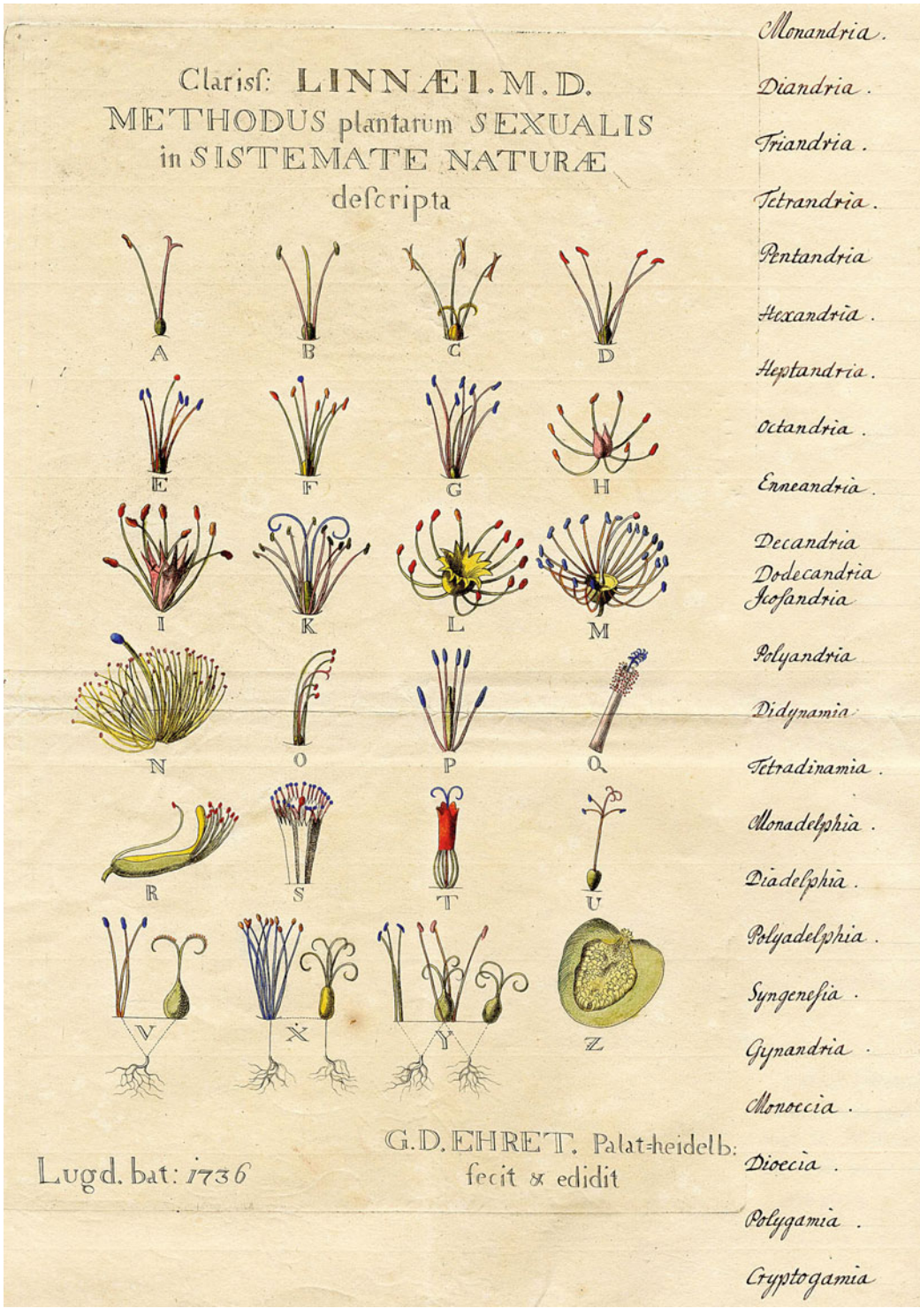


Fig. 3 Linnaeus's sexual system as drawn by G. D. Ehret for the Hortus Cliffortianus (1735–1748); this illustration shows the 24 classes of plants that were defined by Linnaeus according to the number and arrangements of stamens

3.4 The Advent of the Theory of Evolution and Its Decisive Impact on Taxonomy

The end of the eighteenth century was conducive to revolutionary ideas in France, including new principles to reach the natural classification. Studying how to arrange plants in space for creating the new royal garden of the Trianon in the Palace of Versailles, Bernard de Jussieu (1699–1777) applied the key principle of subordination of characters, which will be published in 1789 by his nephew Antoine Laurent de Jussieu (1748–1836) in *Genera Plantarum* [19]. Bernard and A. L. de Jussieu stated that a species, genus, or any other taxon of the hierarchical classification should group plants showing character constancy within the given taxon, as opposed to the character variability observed among taxa. Since not all characters are useful at the same level of the classification, the principle of subordination led to a character hierarchy: characters displaying higher variability should be given less weight than more conserved ones in plant classifications. As a result, B. and A. L. de Jussieu subordinated the characters of flowers—judged more variable and therefore less suitable at higher levels—to the more conserved characters of seeds and embryos. It was the first application of this principle in taxonomy, and it could be interpreted today as a way to limit homoplasy, though the concept of homoplasy had not been elaborated yet [20].

Whereas botanical taxonomy had long been preponderant and faster in its development than its zoological counterpart, the trend was reversed at the beginning of the nineteenth century, especially with the application of the principle of subordination of characters to animals by the French biologists Jean-Baptiste de Lamarck (1744–1829) and Georges Cuvier (1769–1832). New questions then arose in the mind of taxonomists, who were not only interested in naming, describing, and classifying organisms anymore, but also in elucidating how the observed diversity had been generated. Early explanatory theories included the theory of the transmutation of species, proposed by Jean-Baptiste de Lamarck in 1809 in his *Philosophie zoologique* [21]. This was the first theory to suggest the evolution of species, although it involved several misleading assumptions such as the notion of spontaneous generations. Charles Darwin (1809–1882) published his famous theory of evolution in *On the Origin of Species* (1859) [22], and introduced the central concept of descent with modification that later received extensive support and is still accepted today. This implied that useful characters in taxonomy, the so-called homologous characters, are those inherited from a common ancestor. Darwin indeed predicted that 'our classifications will come to be, as far as they can be so made, genealogies' (Darwin 1859, p. 486) [22]. In other words, since the history of life is unique, only one natural classification is possible that reflects the phylogeny. This latter word was however not coined by Darwin himself, but in 1866 in his *Generelle Morphologie der Organismen* [23] by Ernst Haeckel (1834–1919), who is commonly known for the first illustration of a phylogeny,

although Dayrat [24] evidenced that all Haeckel's illustrations should not be interpreted as real evolutionary-based phylogenetic trees [23] (Fig. 4). However, Darwin did not provide any new techniques or approaches to reconstruct the phylogeny or assist practicing taxonomists in their work [25] and, in spite of his major contributions, plant taxonomists therefore kept applying the method of classification described by B. and A. L. de Jussieu even after the onset of the evolutionary thought.

3.5 New Methods and New Sources of Characters for a Modern Taxonomy

In the 1960s, facing the subjectivity of the existing methods to reconstruct phylogenies, the new concept of numerical taxonomy proposed an entirely new way of examining relationships among taxa. Robert Sokal (1926–2012) and Peter Sneath (1923–2011) started developing this concept in 1963 [26], and elaborated it as an objective method of classification. The method consisted in a quantitative analysis of overall similarities between taxa, based on a characters-by-taxa data matrix—with characters divided into character states—and resulting in pairwise distances among taxa. But this method was not based on any evolutionary theory and the resulting diagrams could therefore not be reasonably interpreted in an evolutionary context, or as an evolutionary classification. Nevertheless, this theory flourished for a while, greatly benefiting from rapid advances in informatics.

A crucial change in the way botanists practice taxonomy occurred with the development of the cladistic theory and reconstruction of phylogenies—using diagrams called cladograms—to infer the evolutionary history of taxa. Willi Hennig (1913–1976) initiated this revolution with his book *Grundzüge einer Theorie der Phylogenetischen Systematik*, published in 1950 [27], but his ideas were much more widely diffused in 1966 with the English translation entitled *Phylogenetic Systematics* [28]. The primary principle of cladism, or cladistics, is not to use the overall similarity among taxa to reconstruct the phylogeny, since similarity does not necessarily reflect an actual close evolutionary relationship. Instead, Hennig only based the phylogenetic classification on derived characters, i.e., the characters that are only inherited from the last common ancestor to two taxa—as opposed to the primitive characters. Every taxonomic decision, from a species definition to a system of higher classification, was to be treated as a provisional hypothesis, potentially falsifiable by new data [29]. This new method benefited from an increasing diversity of sources of characters to be considered, thanks to the important technological advances accomplished in the 1940s and 1950s in cytology, ecology, and especially in genetics.

The discovery of the double helical structure of the DNA molecule in 1953, by James Watson and Francis Crick, followed by the possibility to target specific fragments of the genome for selectively amplifying DNA—the Polymerase Chain Reaction

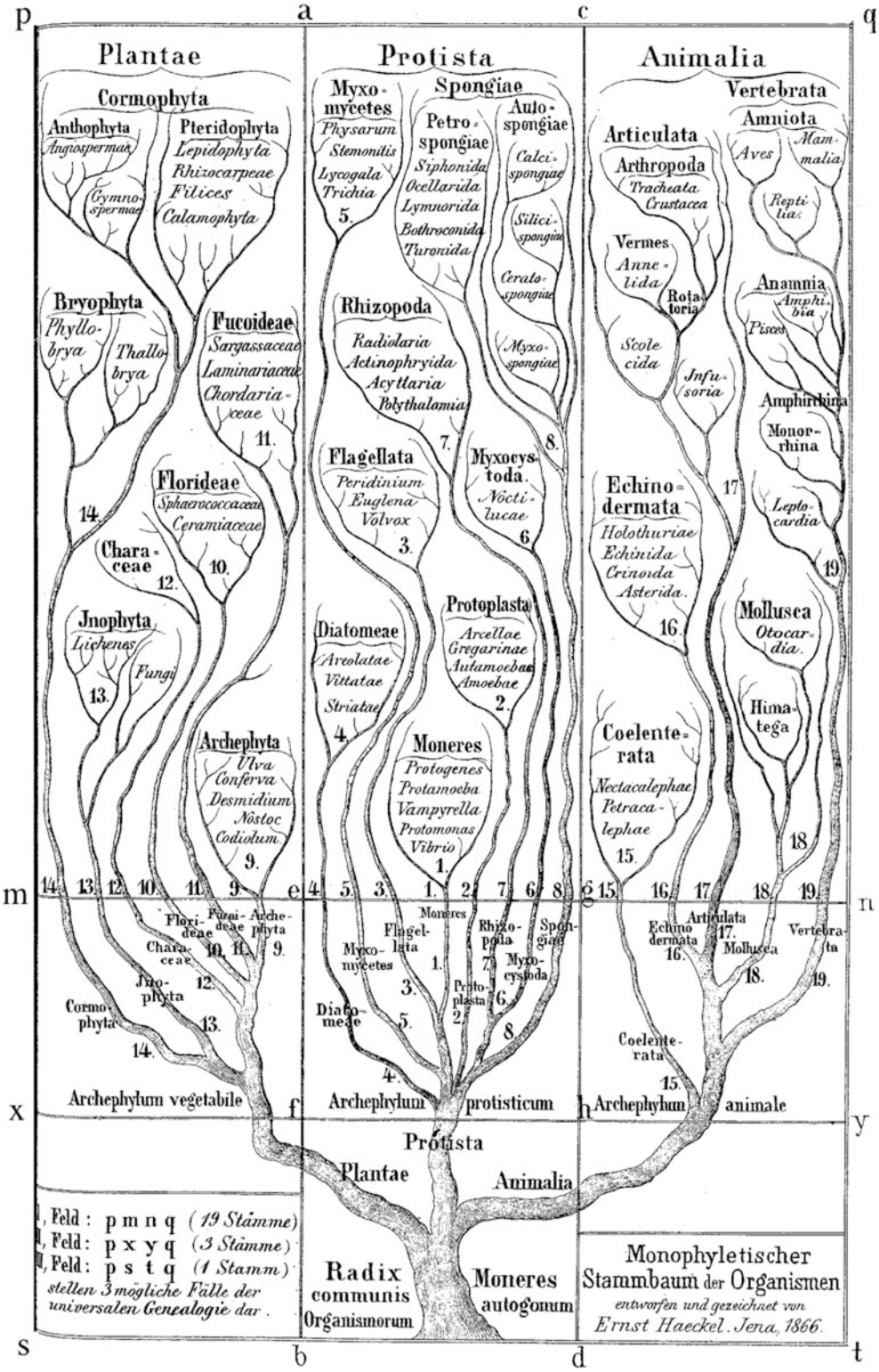


Fig. 4 Illustration from 'Monophyletischer Stammbaum der Organismen' (Haeckel 1866): plants form one of the three main branches of the monophyletic genealogical tree of organisms

(PCR) was invented by Karry Mullis in 1986 [30]—have dramatically changed biology. In particular, the introduction of DNA sequence data has been era-splitting for plant taxonomy, offering access to numerous characters and statistical approaches. Thus, at the turn of the twenty-first century, the use of molecular data and new tree-building algorithms—with probabilistic approaches—led the Angiosperm Phylogeny Group (APG) to better circumscribe all orders and families of flowering plants [31–34]. Similarly, the Pteridophyte Phylogeny Group reached a consensual classification for free-sporing vascular plants (ferns and lycophytes) to the genus level [35]. Such collaborative initiatives have improved to a great extent our understanding of the plant classification based on evolutionary relationships. Many long-standing views of deep-level relationships were drastically modified at the ordinal level, and to a lesser extent at the familial level in flowering plants. One of the most striking changes is the abandonment of the long-recognized monocot-dicot split, since monocots—class Liliopsida—were found to be derived from within a basal grade of families that were traditionally considered as dicots—class Magnoliopsida. Another outstanding finding resulting from analyses of molecular data has been that horsetails and ferns together are the closest relatives to seed plants, necessitating the abandonment of the prevailing view that ferns and horsetails represent paraphyletic successive grades of increasing complexity in early vascular plant evolution, which eventually led to the more complex seed plants, and ultimately to angiosperms [36]. Thus, the more or less intuitive classifications proposed since the beginning of the twentieth century [37–41] have progressively been less used, as a consequence of the modifications brought to the classification by molecular results [42].

Taxonomy took advantage of molecular data not only for improving plant classification or species delineation, but also for species-level identification with the development of the DNA barcoding initiative since the early 2000s. DNA barcoding is based on the premise that a short standardized DNA sequence can allow distinguishing individuals of different species because genetic variation between species is expected to exceed that within species. It was first promoted by Paul Hebert for animals [43] and later supported by international alliances of research organizations like the Consortium for Barcoding of Life (CBOL; http://barcoding.si.edu), which includes a Plant working group, or the China Plant Barcoding of Life Group.

The long history of plant taxonomic research and its numerous contributors, both for theoretical concepts and the practical accumulation of knowledge, allowed the development of an independent, complex, and sound hypothesis-driven scientific discipline that explores, describes, documents the distribution of, and classifies taxa. It is clearly not restricted to, e.g., identifying specimens

and establishing species lists, but it nevertheless also provides basic knowledge that is required to address a wide range of research questions and serve stakeholders in government agencies and international biodiversity organizations (for management of agriculture pests, development of new pharmaceutical compounds, control of trade in endangered species, management of natural resources, etc.; [29, 44–48]). However, taxonomy is faced with the enormous existing plant diversity, and one still unanswered question resides in the extent of plant diversity: how many species are there on Earth?

4 Plant Taxonomy Today: Current Challenges, Methods, and Perspectives

4.1 How Many Plant Species Are There?

Linnaeus' *Species Plantarum*, published in 1753, was one of the first key attempts to document the diversity of plants on a global scale [16]. In this work, Linnaeus recognized more than 6000 species but erroneously concluded that 'the number of plants in the whole world is much less than commonly believed, I ascertained by fairly safe calculation [...] it hardly reaches 10,000' [16]. Later on, in 1824, the Swiss A.P. de Candolle, in his *Prodromus Systematis Naturalis Regni Vegetabilis* [49], aimed to produce a flora of the world: he included 58,000 species in seven volumes. Today, we know that the magnitude of plant diversity is much larger, although we are uncertain of the exact number of plant species.

There are two questions in estimating the total number of plant species: the first one is *how many species have already been described*; the second one is *how many more species are presently unknown to science*.

Our uncertainty about the number of described species is mostly due to the fact that taxonomists sometimes gave different names to the same species inadvertently, especially in the past due to poor communication means between distant scientists. This led to the existence of multiple names for a single biological entity, a phenomenon called synonymy. As a consequence, we know that more than 1,064,908 vascular plant names were published, as evidenced by the International Plant Names Index (IPNI) [50, 51], but they would actually represent only 223,000 to 422,000 accepted species—depending on the method of calculation ([46, 52] and references therein, [53, 54]), with the most recent estimates of 383,671 [51] and 351,176 according to The Leipzig Catalogue of Vascular Plants (LCVP) v.1.0.2 by Freiberg et al. (unpublished). In addition, the disagreement on a single species concept (*see* **Note 1**) among plant taxonomists means that species counts can easily differ by an order of magnitude or more when the same data are examined by different botanists [55]. This leads to a taxonomic inflation, i.e., an increased number of species in a given group that is not due to an actual discovery of

new species [56–58]. In practice, this can occur when, e.g., different botanists do not recognize the same number of species in a given taxonomic group—the 'splitters' vs. the 'lumpers'—or when one botanist describes subspecies while another one elevates them to the rank of species.

The estimation of the total number of plant species on Earth is also obviously hampered by our uncertainty about the extent of the unknown plant diversity: how many more species there are to discover? The exploration of plant diversity allows the description of ca. 2000 new plant species every year [46, 47, 59] although part of which may turn out to be synonyms based on future thorough monographic revisions. Based on a model of the rates of plant species description, Joppa et al. [60] estimated that there should be an increase of 10–20% in the current number of flowering plant species. This means that, based on the estimation of 352,000 flowering plant species [46], they predicted the actual diversity between 390,000 and 420,000 species for this group. Meanwhile, Mora et al. [61] used higher taxonomy data, i.e., they extrapolated the global number of plant species based on the strong negative correlation between the taxonomic rank and the number of higher taxa—which is better known than the total number of species. As a result, focusing on land plants, they suggested an expected increase of 38% in the number of species, from 215,000 in Catalogue of Life [62] to 298,000 predicted species.

These numbers make clear that our knowledge of plant diversity is still very incomplete and that even estimates of its magnitude remain highly controversial and speculative, highlighting the need for more taxonomic studies.

4.2 Current Threats on Plant Diversity, the Taxonomic Impediment, and Some Initiatives to Overcome It

At the Sixth Conference of the Parties to the Convention on Biological Diversity (CBD) held in 2002, more than 180 countries adopted the Global Strategy for Plant Conservation (GSPC). It included 16 specific targets that were to be achieved by 2010, with the goal to halt the loss of plant diversity [46]. The Strategy was updated in 2010 (at the Tenth meeting of the Conference of the Parties) and it is now implemented within the broader framework of the Strategic Plan for Biodiversity 2011–2020. The first and most fundamental target of the Strategy was initially to complete a 'widely accessible working list of known plant species, as a step towards a complete world flora' [46, 59]. After the completion of this list in 2010 [53], Target 1 was slightly modified to developing 'an online flora of all know plants' (http://www.cbd.int/gspc/targets.shtml; http://www.worldfloraonline.org/). This target aims to provide baseline taxonomic information, i.e., a list of the accepted names for all known plant species, linked to their synonyms but also to biological information such as geographic distribution and basic identification tools. Since species are basic units of analysis in several areas of biogeography, ecology, and

macroevolution and also the currency for global biodiversity assessments [63], the lack of such taxonomic information is a critical bottleneck for research, conservation, and sustainable use of plant diversity [46], and was called the 'taxonomic impediment' at the Second Conference of the Parties to the CBD (decision II/8). This is especially critical at a time when biodiversity faces its sixth extinction crisis: most newly described species occur in hotspots of diversity, often in tropical dense forests, where protected areas are scarce, the level of habitat destruction (due to anthropic activities) is high, and the impact of climate change is strong. Newly described species are also likely to be characterized by locally low abundance and small geographic ranges, enhancing their risk of extinction [64]. Therefore, botanists must engage into a race to describe and name species before they go extinct. This is especially true since plants still lag far behind many animal groups in contributing to global conservation planning, despite their essential role in structuring most ecosystems [65]. In addition to the major conservation concern, there are a multitude of possible concrete examples of beneficial application of taxonomic discovery such as the identification of new wild species adaptable for agriculture, timber or fibers; new genes for enhancement of crop productivity; and new classes of pharmaceuticals. Also, basic taxonomic knowledge is a prerequisite to monitor and anticipate the spread of invasive plants, and to better understand ecosystem services [45, 47].

Several factors limit the efficiency of botanists in documenting plant diversity. However, recent improvements and future optimistic perspectives must also be underlined, and numerous contributions have been made to imagine and propose what the 'Taxonomy for the twenty-first century' should and could be (*see*, e.g., [29, 66], and the whole Theme Issue of the *Philosophical Transactions of the Royal Society of London, Series B* that they coordinated; *see also* [67–69]), and what roles, challenges, and opportunities should be for the 'Botanists of the twenty-first century' (https://unesdoc.unesco.org/ark:/48223/pf0000243791.locale=fr).

First, one limiting factor is the general lack of funding and, in particular, the lack of resources devoted to the basic field activity of collecting new material [55, 66, 67, 70]. Field explorations are also made difficult by practical limitations such as ease of access to remote areas or safety concerns in some parts of the world that may be politically unstable [44, 45, 59]. However, we currently know a renewed age of exploration and discovery, supported by several national or international initiatives. This is particularly true in the United States, where the 'Planetary Biodiversity Inventories: Mission to an (Almost) Unknown Planet' program, was launched in 2003, aiming to complete the world species inventory for some selected taxa, with individual project awards of ca. US$3 million over a 5-year duration (http://nsf.gov/pubs/2006/nsf06500/nsf06500.htm) [55, 71]. Other leading initiatives like 'Our Planet

Reviewed' (http://www.laplaneterevisitee.org/en) much contribute to inventories in hotspots. In Europe, several major museums and botanical gardens established the Consortium of European Taxonomic Facilities (CETAF) in 1996, which in turn created the European Distributed Institute of Taxonomy (EDIT) in 2006, for a 5-year period, under the European Union sixth Framework Programme. This worldwide network of excellence brought together 29 leading European, but also North American and Russian, institutions with the goal to increase both the scientific basis and capacity for biodiversity conservation. Developing countries also participated to this international effort by developing similar national or multinational programs, e.g., in Brazil, Mexico, and Africa [72]. On a global scale, the Global Taxonomic Initiative (GTI) was launched in 1998 by the Conference of the Parties to the CBD, and was later related to the GSPC, in order to remove or reduce the 'taxonomic impediment.' In addition to institutional breakthroughs, modern means of travel have facilitated access to remote places where many species occur. As a result, although today botanical expeditions could probably not be as prolific as those reported during the great naturalists explorations of the eighteenth and nineteenth centuries in terms of new species descriptions (e.g., in 1770, Sir Joseph Banks collected specimens representing as many as 110 new genera and 1300 new species in Australia; White 1772 in [59]), important discoveries occurred in the recent past and provide evidence for the vitality of contemporary botanists: for instance, the Malagasy endemic *Takhtajania perrieri* (Capuron) Baranova & J.-F.Leroy (Winteraceae) was first collected in 1909 and thought to have gone extinct, but was rediscovered in 1994 [73], i.e., almost 90 years after its first collection. Other examples suggest that some showy, sometimes abundant plants still remain to be described, even in geographical areas that are supposed to be well prospected: a new genus and species of conifer, *Wollemia nobilis* W.G.Jones, K.D.Hill & J.M.Allen, was observed in the 1990s only ca. 150 km from Sydney (Australia), and was shown to belong to a well-known family of charismatic trees (Araucariaceae), including only two other genera [74]. In 2007, Thulin and collaborators reported the discovery of a conspicuous and dominating tree in the Somali National Regional State (Ogaden) in eastern Ethiopia [75, 76]. This tree, *Acacia fumosa* Thulin (Fabaceae), covers an area as large as Crete but was hitherto unknown to science. The location of this species in an African war zone and the inaccessibility of the area probably explain that it had never been collected and remained undescribed so far. To cite a last example of recent striking botanical discovery, we can mention the description of a new palm genus and species from Madagascar, *Tahina spectabilis* J.Dransf. & Rakotoarin. [77]. The trees grow to 18 m high and leaves reach 5 m in diameter, making them the most massive palms ever found in Madagascar. However, the small census size

(less than a hundred individuals), limited habitat, and rare reproduction events lead to serious conservation concerns for the species.

A second crucial issue for enhancing our knowledge of plant diversity is the lack of taxonomic expertise. This is at least partly due to the lack of credit given to works of descriptive taxonomy (e.g., species lists, floras, or monographs) compared to peer-reviewed publications in high-impact journals [46, 70, 78, 79]. The global number of species described over time has increased over the past 250 years [60, 80], but this remains clearly not sufficient to counteract the increasing rate of species extinctions, and many species are at risk of disappearing before being described. Although taxonomists have most likely increased the efficiency of their efforts since the mid-1700s, the involvement of more numerous people into the tasks of exploring and describing the biodiversity is needed: in the United States, the NSF's Partnerships for Enhancing Expertise in Taxonomy (PEET) program allowed the training of new generations of taxonomists since 1995 [78, 81], and enjoyed much success. In addition, in some regions (e.g., in Costa Rica or Papua New Guinea), local people called 'parataxonomists' contribute to specimens collection and species recognition based on rough morphological criteria, in collaboration with taxonomic experts [45]. This is also in line with the growing body of 'citizen scientists,' who are often amateurs and offer their help to accumulate data, e.g., on the presence/absence of a given species in a given region, or the distribution of a morphological character across space. Because they are usually organized as large networks, they represent an immense and increasingly important workforce and make possible some tasks that would otherwise not have been possible because of, e.g., limited time and funding [80]. However, sound knowledge and experience of professional taxonomists remain critical [46, 66, 67, 70, 82] and capacity building in tropical countries—where the greatest diversity of life is concentrated—should therefore be a priority [59].

A third identified impediment to our taxonomic knowledge was—and still is, to a certain extent—the problem of communication and coordination, of tracing the accumulated publication records, of deciphering the complex synonymy, and of chasing the scattered (and sometimes in poor condition) material, especially type specimens that are housed in herbaria around the world [59, 66, 67]. Worldwide natural history collections contain 390 million plant specimens [83] and their importance has recently been made even more prominent by the finding that they house many new species that remain to be described [84]: researchers analyzing the time lapse between flowering plant sample collection and new species recognition estimated that only 16% were described within 5 years of being collected for the first time, and that nearly 25% of new species descriptions involved specimens of

more than 50 years old. The median time lag between the earliest specimen collection and the publication of the new plant species description was ca. 30 [84] or 32 years [85]. Such a lag time (also called 'shelf life') is longer for herbarium specimens than for all other taxonomic groups [85]. Natural History Collections thus act as a reservoir of potential new species. Therefore, although one limiting step of species discovery may be the capacity to undertake field work (as suggested above), access and examination of existing herbarium collections by experts are another bottleneck. This is however now partly overcome by programs such as the European SYNTHESIS (from 2004 to 2017 superseded by SYNTHESIS+ in 2019; https://www.synthesys.info/about-synthesys.html) that provide funded researcher visits to specimens housed by diverse institutions, or by increased international collaborations and a better access to information and specimens, thanks to modern data-sharing technologies [46–48, 61, 69]. As an example, a major step was accomplished thanks to funding from the Andrew W. Mellon Foundation and subsequent institutional commitments to database and image name-bearing type specimens—on which the species original descriptions are based—and deposit these data in the central repository JSTOR Plant Science [82]. At an even larger scale, several major herbaria—including the Paris Herbarium, which is one of the biggest/richest in the world with ca. eight million specimens [86]—achieved large-scale digitization of all their vascular plant specimens, in order to make them freely available as high-quality photographs on the web—through both the herbarium database https://science.mnhn.fr/ and the platform e-ReColNat https://www.recolnat.org/fr/—that gather images for all natural history collections from France (and *see* **Note 2**). In the United States, the National Science Foundation (NSF), through its Advancing Digitization of Biological Collections (ADBC) program, developed a strategic plan for a 10-year coordinated effort to digitize and mobilize images and data associated with all biological research collections of the country in a freely available online platform. This will ensure increased accessibility of all valuable information and is being made possible by the establishment of a central National Resource for Digitization of Biological Collections (called iDigBio for 'Integrated Digitized Biocollections'; https://www.idigbio.org/).

For a better diffusion of taxonomic revisions, Godfray [66] claimed the need for a 'unitary web-based and modernized taxonomy' (*see also* [87]). Without opting for such a drastic evolution, a revision of the International Code of Nomenclature (ICN) has nevertheless encouraged a change dynamics toward electronic publications: at the International Botanical Congress held in Melbourne in July 2011, purely electronic descriptions were judged valid for the publication of new species (Art. 29), as opposed to the previous requirement to publish in traditional, printed publication [88]. But based on the following 8 years, it must be concluded that

the new applicable rule did not accelerate the rate of plant species description or participation in biodiversity discovery as was hoped [89]. Also, whereas the current taxonomic knowledge is mostly made available in paper format as monographs, floras, and field guides, many internet taxonomy initiatives exist and catalogue species names, lists of museums specimens, and identification keys and/or other biological information. These websites include, e.g., IPNI (www.ipni.org), The Plant List (www.theplantlist.org), GBIF (www.gbif.org), Species 2000/ITIS Catalogue of Life (www.catalogueoflife.org), Tree of Life (www.tolweb.org), and Encyclopedia of Life (www.eol.org), to cite only a few (*see* [55, 66]).

4.3 Molecular Taxonomy and the Need for an Accelerated Pace of Species Discovery

In addition to increased efforts towards exploration in the field, various initiatives to promote and develop taxonomic expertise, generalization of collaborative work, and improved access to natural history collections and literature, major advances in technology also provide new opportunities to facilitate and accelerate the rate of species discovery at a time of increasing need to monitor and manage biodiversity. The goal of accelerating the pace of species discovery was made especially clear by the promoters of the DNA barcode initiative [90, 91], but more generally, the use of molecular tools for taxonomic purpose emerged in the 1990s—or even in the 1970s if considering allozyme markers—and has quickly become an area of intense activity.

Today, most recognized species have been delineated and described based on morphological evidence: in general, they have been delimited based on one or more qualitative or quantitative morphological characters that show no—or very little—overlap with other species [92]. The initial enthusiasm for molecular taxonomy most probably came from the additional and complementary information that it provided. Also, molecular taxonomy requires an expertise that is nowadays more broadly distributed than that for thorough morphological investigations, it makes use of tools that are not specific to a particular group of plants, and it may appear more prone to scientific publications in peer-reviewed journals than more traditional, taxonomic studies. We synthesize, here, several other characteristics—both strengths and limitations—of molecular taxonomy that one should keep in mind when initiating taxonomic studies using molecular tools.

4.3.1 Strengths and Limitations of Molecular Taxonomy

First, it must be noted that the resemblance criterion within a species, on which is based the morphological approach to delimit species, suffers exceptions and can lead to erroneous conclusions. Before the various reproductive systems of plants were well understood, male and female individuals from a single—e.g., dioecious—species were sometimes described as two distinct species based on morphological investigations. For example, in the orchid genus *Catasetum* Rich. ex Kunth plants are functionally dioecious (i.e.,

with female and male flowers situated on distinct individuals) and can morphologically differ so much from each other that taxonomists of the nineteenth century assigned individuals of the same species to different genera (*Monachanthus* Lindl. and *Myanthus* Lindl.) [93]. Other species descriptions incorporated characters that were in fact due to anther-smut disease caused by the fungus *Microbotryum violaceum* (Pers.) G. Deml & Oberw.: anthers of infected plants are filled with dark-violet fungal spores instead of yellow pollen [94]. As a result, *Silene cardiopetala* Franch., for example, was distinguished from *Silene tatarinowii* Regel by its dark anthers but should likely be treated as the same species. More generally, because the phenotype of a plant is influenced both by its genotype but also by its environment—and the interaction between the genotype and the environment, called phenotypic plasticity—the observations of herbarium specimens collected in the field may be somewhat misleading. Molecular taxonomy should avoid this possible bias since it is based on neutral markers that are in principle independent of environmental conditions. However, the influence of the environment is mostly true for vegetative characters and usually less problematic for reproductive characters. In addition, the use of several morphological characters should limit the problem since all traits are unlikely to be affected in the same way [95].

Second, several studies showed that, in comparison to the traditional morphological criterion for delimiting species, molecular tools sometimes allow the detection of additional, so-called 'cryptic' species that could not be distinguished on morphological grounds only. This may happen when species emerged in the recent past, due to morphological stasis, or to morphological convergences [96]. The existence of such cryptic species was reported, e.g., on temperate or tropical plants ([97, 98] and references therein; for an animal example, *see* [99]).

Third, in addition to the primary goal of species delimitation, the use of genetic tools may allow to better understand the evolutionary process at work within taxonomically complex groups, where taxa are sometimes difficult—or even impossible—to delineate. These groups are often characterized by uniparental reproduction, e.g., self-fertilization or apomixis, and reticulate evolution, due to, e.g., hybridization and introgression, which preclude the delineation of discrete and unambiguous taxonomic entities. In such cases—e.g., in the genera *Sorbus*, *Epipactis*, and *Taraxacum* ([100–102] respectively, cited in [103])—principles of conservation biology suggest that the evolutionary processes that generate and maintain diversity should themselves be preserved because they are even more important than the presently observed taxa [103, 104]. In this perspective, molecular tools can yield very useful information, usually based on a population sampling.

Fourth, from a practical point of view, a key strength of molecular taxonomy is that it can be performed on any life stage—even some that bear no or only few morphological characters such as seeds, seedlings, or fern gametophytes [105, 106]—and almost any type of material, e.g., leaves, cambium [107–110], bark [111], dry wood [112], and roots [113, 114]. Therefore, the use of molecular characters for taxonomic purposes appears especially suitable for organisms that require years before flowering and/or fully developing, or when access to some other key—e.g., reproductive—characters is difficult.

The ubiquitous character of the DNA molecule in living beings can also become a problem, and care should be taken to only isolate DNA from the target material and exclude DNAs of any other animal, vegetal, or fungal organisms living around or in the plant under study—e.g., parasitic insects, epiphyllous mosses, and endophytic fungi.

Another practical limitation of molecular taxonomy is the cost, as molecular lab facilities and often rather expensive consumables are needed. This cost may be especially limiting in developing countries [115, 116], although it is ever decreasing thanks to the spread of molecular analyses, which are more and more commonly employed, and to technological advances that allow cheaper and less time-consuming analyses, see below.

An important parameter that is shared by 'traditional' and molecular taxonomy studies is sampling strategy and sampling effort. Taxonomy is based on a comparative approach that requires the investigation of as many specimens/samples as possible in order to catch all the extent of natural variation. Therefore, the quality of taxonomic studies partly relies on a thorough sampling of specimens/samples to be surveyed, and a biased sampling may cause erroneous conclusions. As an example, *Marsilea azorica* Launert & Paiva, which was thought to be a local endemic and critically endangered species of the Azores archipelago, was recently shown to be conspecific to an Australian native species that is widely cultivated and invasive in Florida, *Marsilea hirsuta* R.Br [117]: because the spread of *M. hirsuta* out of Australia was not documented when the *Marsilea* specimens from the Azores were examined by Launert and Paiva in 1983 [118], the botanists did not include the Australian taxa into their survey, and erroneously described the species as new to science.

On more theoretical and conceptual grounds, some claim that, in comparison with 'traditional'—typically morphology-based—taxonomy, the use of molecular tools may avoid bias due to the subjectivity of a given taxonomist, who could have a priori ideas on species delimitation. However, the acquisition of a molecular dataset also implies some more or less subjective choices, e.g., on the distinction of orthologs vs. paralogs, on defining character

homology when sequences of different lengths must be aligned to form a square matrix, or on the statistical analysis to carry out after the data are produced (*see* e.g. [95, 119–121]); the latter choice, on data analysis, is closely related to the adoption of a given species concept (*see* **Note 1**). Also, because our current technological capacities do not allow the routine inclusion of the whole genome in taxonomic analyses, choices must be made on the genomic compartment(s) to survey—nuclear, mitochondrial, or chloroplastic, the molecular technique(s) to use, and the precise, individual marker(s) to consider (Chapter 2). The choice of a limited number of markers is required, in practice, although multiple independent loci might often be necessary to solve the possible disagreement between gene trees and species trees, and to uncover the common reticulate evolution—due to horizontal DNA transfer, hybridization, and polyploidization events—and incomplete lineage sorting in plants [55, 95, 121–123]. The extent of genome coverage by molecular markers is partly dependent on the molecular technique that is used, and there is often a trade-off between the possibility—due to time and cost limitations—of surveying numerous markers and the information content provided by each marker. For example, it is usually achievable to include a large number of anonymous markers based on length polymorphism—such as RAPD or AFLP markers—but the number of DNA regions that could be sequenced, representing highly informative data, is much more limited with the traditional Sanger method. However, rapid advances in Next-Generation Sequencing (NGS) technologies have resulted in huge cost reduction and offer incredible new opportunities for producing billions of base pairs of accurate DNA sequence data in a few hours [124–126]. Studying whole chloroplast genomes or multiple nuclear loci might therefore become routine even in non-model species, and begin, obviously, to revolutionize plant molecular taxonomy [127]. Then, the main bottleneck is probably cleaning up and assembling the sequence reads to generate useable data, and major improvements in bioinformatics would be needed to deal with such huge amounts of data [124, 125].

Another limitation of molecular taxonomy is the possible lack of genetic divergence when sister-species have very recent origins because they will share alleles due to recent ancestry and, if reproductive isolation is not complete, to ongoing gene flow, i.e., hybridization. This lack of genetic variation can nevertheless be accompanied by some level of morphological differentiation, leading to the exact symmetrical situation to cryptic taxa—where one could observe genetic but no morphological distinction; see above. The absence or extremely weak genetic divergence was observed, e.g., in the young and species-rich neotropical genus *Inga* Mill. (Fabaceae) [128], and striking examples of such morphological

diversification but weak genetic variation are also provided by cases of adaptive radiations, where species rapidly adapt to different environments—e.g., in the Hawaiian silverswords (Asteraceae) [129], the Asian genus *Rheum* L. [130], or the widespread columbine genus *Aquilegia* L. [131]; for more examples, *see* [132]. In such cases of recent diversification, the delimitation of species will usually be based on allele frequency changes rather than diagnostic changes [133] and can benefit from recently developed coalescent-based methods (e.g., [134, 135]). The time required for genetic divergence to build up after speciation will depend on the mutation and fixation rates—and the fixation rate depends on the number of reproductively effective individuals. Because of different fixation rates between (diploid) nuclear and (haploid) organelle genomes, studies based on nuclear vs. organelle DNA markers may yield contrasted results on species limits. Such contrasted results are also made likely by the horizontal organelle DNA transfers that occasionally occur, especially among closely related species.

Molecular markers can also suffer from homoplasy, i.e., markers can show similar character states that, however, do not derive from a common ancestor. In this case, they do not inform on the genealogy of taxa and, because they do not reflect a shared evolutionary history, they may be misleading on evolutionary and, as a consequence, on taxonomic relationships. This is especially problematic for highly variable markers, e.g., microsatellites, and for DNA sequences that are only composed of four types of monomer (A, C, G, and T): as a result, a substitution at any one position has a high probability of being a reversal or a convergence, i.e., of being homoplasic [95, 121]. It is therefore critical to take this caveat into account when analyzing and interpreting molecular data.

Another drawback of molecular taxonomy is that name-bearing type specimens often do not permit DNA analyses because of nonoptimal drying and storage conditions, resulting in DNA deterioration ([136]; the same limit also obviously applies to most plant fossils, which do not contain DNA). Consequently, a comparison of the supposedly new species with known species may not be possible on a molecular basis and prevent a rigorous taxonomic, comparative approach. As part of their 'plea for DNA taxonomy,' Tautz et al. [133] proposed to identify neotypes for all known species in cases of unavailable genetic information from the original types, so that these neotypes could constitute new reference records for further studies. However, this proposal received very limited support (*see*, e.g., [119, 137]). Besides, recent progresses have been made in the extraction of DNA from herbarium specimens [138, 139] and the genetic analysis of such material will very likely benefit from NGS technologies [126]. But so far, given the usually low-quantity and often degraded DNA that is extracted from

herbarium specimens, the most commonly employed molecular techniques are microsatellite markers—because their short length makes amplification more likely than that of longer DNA stretches (*see* e.g. [140]), and organelle DNA sequencing, because their multiple copies per cell represent more abundant template DNA for PCR than nuclear loci. Most of the published studies report the successful exploitation of specimens up to ca. 100–150 years old, with DNA sequences produced usually ca. 500 pb long [141–145]. But the kind of material—e.g., presence of PCR-inhibiting substances [146]—and the speed and method of drying appear more important than the actual age of the sample [141–143, 147], and some botanists managed to obtain DNA sequences from even older specimens and/or longer DNA regions, e.g., Ames & Spooner sequenced ca. 440-bp DNA fragments from potato material from the early eighteenth century [148], and Andreasen et al. [147] sequenced 800-bp DNA fragments from a specimen collected in the late eighteenth century. The successful use of aged seeds has also been reported [138, 149].

4.3.2 The Definitive Need for an Integrative Taxonomy

The use of molecular data in plant taxonomy has been era-splitting and highly successful in many instances, but we also highlighted some limits and cautions to consider when adopting this approach. Most importantly, a species description solely based on molecular evidence would obviously seem critically disconnected from the natural history of the species, i.e., its life-history traits, ecological requirements, co-occurring species, biotic interactions, etc. As such, molecular tools may indeed accelerate the rate of species discovery but would actually be a poor contribution to our knowledge and understanding of plant diversity and evolution. Such a use of molecular taxonomy could even end up with the exact opposite of the expected outcome if funders only aim to basically delineate and count species with no other ambition; indeed, gathering further biological information is an essential prerequisite to make a general use of the taxonomic knowledge, efficiently preserve the existing diversity, and allow its continued evolution. Botanists have long realized this and promoted the use of multiple independent sources of data, and/or the use of several analytical methods on the same dataset to corroborate the delimitation and provide a thorough and detailed description of species. As early as 1961, Simpson (p. 71) wrote 'It is an axiom of modem taxonomy that the variety of data should be pushed as far as possible to the limits of practicability' [6]. In agreement, Alves and Machado [150] wrote that 'Taxonomy should be based on all available evidence.' This awareness gave rise to the advent of what is now called 'integrative taxonomy,' where taxonomic hypotheses are cross-validated by several lines of evidence ([29, 121, 150–155], and many others). As sources of relevant characters, many fields of biology might contribute to

taxonomic studies: they include morpho-anatomy which takes advantage of new techniques such as Scanning Electron Microscopy (SEM), remotely operable digital microscopy, computer-assisted tomography, confocal laser microscopy, and automatic image processing for morphometry [155, 156], cytometry and cytogenetics (*see* Chapters 17–19) but also palynology, physiology, chemistry (production of secondary compounds), breeding relationships, and ecological niche modelling—we are not aware of currently available examples in plant taxonomy but for animals, *see* [157–159]. Other sources of information will also most probably be more widely used in the future, such as transcriptomics [160, 161], metabolomics [162], proteomics [95], and even phenomics—Munck et al. [163] showed, in barley, that the fingerprint of a near-infrared spectrum from an individual represents a coarse-grained overview of the whole physiochemical composition of its phenome, with the phenomic profile resulting from the combined effects of the entire genome, proteome, and metabolome [164]. The diversity of approaches involved in modern plant taxonomy is consistent with the observations by Joppa et al. [80] that (a) today's biologists who describe species are not only contributing to the field of taxonomy, but also active in other fields/disciplines, and (b) most new species are nowadays described by several authors whereas descriptions by a single author were common around 1900.

It is also clear that end users of taxonomy such as conservation planners need an operational, character-based, and cheap way to discriminate species [91, 115, 150]. This could tend to diminish the perceived potential of molecular taxonomy, but in this perspective and in spite of the shortcomings that we have just underlined, molecular taxonomy obviously has a great role to play. DNA can aid to delimit taxa, and to group specimens among which to find morphological—or other types of—affinities in further investigations (*see*, e.g., [165, 166]). Such clusters of individuals, characterized by close genetic relationships, are sometimes referred to as 'molecular operational taxonomic units' (MOTU) [167], before their genuine taxonomic statuses are evaluated by gathering additional data. Markmann and Tautz [168] called this approach, based on an initial molecular assessment, the 'reverse taxonomy' (*see also* [151, 152]). The fruitful link between 'traditional' and molecular taxonomy should be accompanied by an analogous link between herbarium vouchers, plant samples for DNA extraction, and DNA extracts [169, 170]. The curation of such collections and the maintenance of a dynamic link between them will provide a long-lasting and reliable framework for taxonomic investigations, and will permit the critical re-evaluation of taxa delimitations at any time, based on both herbarium and DNA material.

Duminil and Di Michele [95] reviewed studies comparing species delimitations based on morphological traits and molecular markers. They found both cases of congruence and incongruence between the two types of data. As suggested above, cases of incongruence were either due to stronger molecular discrimination between species—suggesting the existence of cryptic species—or, on the contrary, to stronger morphological differentiation, due to processes like local adaptation, phenotypic plasticity, or neutral morphological polymorphism (e.g., [171]). Conflicting results often trigger more in-depth studies using as many loci as possible and, if possible, loci that originate from different genomes, with the goal to better understand the patterns and processes of plant evolution and diversification (e.g. [172–178]).

Taxonomic circumscriptions are scientific hypotheses, which are ideally validated by evidence from multiple sources, and molecular methods offer the opportunity to yield high-potential information. However, there is not a single, best method to be used in all plant groups and the molecular taxonomist will have to face multiple questions: before anything, it is necessary to identify the optimal sampling strategy, the most suited genomic compartment(s) to examine, the right technique(s) to use, and the adequate method (s) of statistical analysis to extract the relevant information about species limits and relationships [120, 121]. In addition to the complementarity of 'traditional' and genetic approaches, molecular taxonomy itself will often require to gather and compare patterns based on several types of data—e.g., nuclear vs. cytoplasmic markers or markers with different rates of evolution. The goal of this book is to present the possible alternatives of molecular taxonomy, their practical implications in the lab, current analytical tools that are available, and theoretical consequences for data interpretation. The empirical and analytical approaches used for a molecular taxonomic study, together with the conclusions drawn from the data, will also obviously depend on the species concept that is adopted and on the choice of operational criteria to delimit species—*see* **Note 1**).

5 Notes

1. Species concepts and contemporary criteria for species delimitation.

   Species delimitation obviously depends on what a species is and, although the species is often seen as the fundamental unit of evolution, its definition has long remained highly debated.

   The existence of species itself is somewhat controversial, especially in plants where asexuality, hybridization, and polyploidy may render the definition and delimitation of species

complex and fuzzy. Some argue that species are 'arbitrary constructs of the human mind' while others claim that they are objective, discrete entities. Reviewing the available data (both in plants and animals), Rieseberg et al. [179] showed that discrete phenotypic clusters exist in most genera (>80%), although the correspondence of taxonomic species to these clusters is poor (<60% and not different between plants and animals). In addition, crossability experiments indicate that as much as 70% of plant taxonomic species and 75% of plant phenotypic clusters correspond to reproductively independent lineages.

The proliferation of alternative species concepts really started in the 1970s. It gave rise to several decades of debate and taxonomic instability because many concepts were incompatible in that they lead to the recognition of different species boundaries and different number of species. This was called the 'species problem.'

Morphological approaches have dominated species delimitation for centuries, starting with the purely typological (i.e., essentialist) pre-Darwinian view. But most contemporary biologists are familiar with the idea that species are groups of actually or potentially interbreeding natural populations, which are reproductively isolated from other such groups (the 'Biological Species Concept'), whether or not they differ in phenotypic characters that are readily apparent.

However, another, unified species concept has now emerged. It originated as early as the beginning of the twentieth century (with, e.g., E. B. Poulton), became well established during the period of the Modern Evolutionary Synthesis (with the great leaders T. Dobzhansky, E. Mayr, G. G. Simpson, and S. Wright), and was recently largely promoted by de Queiroz [180, 181]. This unified concept reconciles previous, at least partially incompatible species concepts. It considers species as separately evolving metapopulation lineages and is called the 'General (metapopulation) Lineage Concept.' Other properties of species, which used to be treated as necessary (and sufficient) properties to recognize a species as such (e.g., reproductive isolation, monophyly; *see* Table 1), are now only seen as different lines of evidence, or 'operational criteria,' relevant to assessing lineage separation. The unified species concept is actually not a new concept, but simply the clear separation of the theoretical concept from the operational criteria that are used for the empirical application of this concept.

Operational criteria can be either tree-based or non-tree based (e.g., direct tests of crossability, indirect estimates of gene flow, statistical clustering algorithms) [198], and new methods are still being developed (e.g., analyzing multilocus genetic data in a coalescent framework). Criteria differ in their

Table 1
Some alternative contemporary species concepts/criteria

Name of the species concept/ criterion	Definition of the species	Major contributor(s)	Ref.
Interbreeding species concept [forms the basis for the general (metapopulation) lineage concept]	A group of potentially interbreeding populations	Wright 1940, Mayr 1942, Dobzhansky 1950	[182–184]
[a]Isolation species concept [often called the biological species concept]	A group of potentially interbreeding populations that is reproductively isolated from other such groups	Poulton 1904, Mayr 1942, Dobzhansky 1970	[184–186]
Phenetic species concept	A group that forms a phenetic cluster (quantitative difference)	Sokal and Crovello 1970	[187]
Ecological species concept	A group that shares the same niche or adaptive zone	Van Vaalen 1976	[188]
[a]Evolutionary species concept [corresponds closely to the general (metapopulation) lineage concept]	A lineage (i.e., an ancestral-descendant sequence of populations) evolving separately from others and with its own evolutionary role and tendencies	Simpson 1951, Wiley 1978	[189, 190]
Phylogenetic species concept—character diagnosability version	An irreducible (basal) cluster of organisms, diagnosably distinct from other such clusters, and within which there is a parental pattern of ancestry and descent (fixed qualitative character) The diagnostic character can be from any trait (e.g., morphological or molecular) and of any significance (e.g., a single base pair)	Cracraft 1989	[191]
Phylogenetic species concept—reciprocal monophyly version	A group that shows monophyly (consisting of an ancestor and all of its descendants, and commonly inferred from the possession of shared derived character states)	Rosen 1979, Donoghue 1985, Mishler 1985	[192–194]
Genealogical species concept	A group that shows monophyly for all (or at a consensus of) gene genealogies in the genome	Baum and Shaw 1995	[195]
Genotypic species concept	A group recognizable on the basis of multiple, unlinked, inherited genetic markers A pair of such genotypic clusters is recognizable if the frequency distribution of genotypes is bimodal or multimodal, and	Mallet 1995	[196]

(continued)

Table 1
(continued)

Name of the species concept/ criterion	Definition of the species	Major contributor(s)	Ref.
	strong heterozygote deficits and linkage disequilibria are evident between the clusters		
[a]Cohesion species concept	A group that is characterized by cohesion mechanisms, including reproductive isolation, recognition mechanisms, and ecological selection, as well as by genealogical distinctiveness	Templeton 1998	[197]

[a]Combined species concepts, i.e., concepts using a combination of morphological, ecological, phylogenetic, and reproductive criteria

suitability to some particular species (e.g., sexual vs. asexual), their requirements in terms of type of data and sampling, and their strengths and limitations. It must also be noted that most of them will require researchers to make some qualitative judgments at some point.

The commonly observed incompatibility between various criteria stems from the fact that various properties actually arise at different stages in the process of speciation: as lineages diverge, they become distinguishable in terms of quantitative traits, diagnosable in terms of fixed character states, reproductively incompatible, they evolve distinct ecologies, they pass through polyphyletic, paraphyletic, and monophyletic stages, etc. These changes commonly do not occur at the same time, and they are not even necessarily expected to occur in a specific order. De Queiroz [180] qualifies this transition period, from one ancestral species to two divergent species, a 'grey zone,' where alternative species definitions can come into conflict. But as lineages diverge, the number of species criteria satisfied will increase and allow a highly corroborated hypothesis of lineage separation and species delimitation.

2. A new world-class research Infrastructure is now being built in Europe, the Distributed System of Scientific Collections (DiSSCo), that will work for the digital unification of all natural science assets under common curation, access, policies and practices, and that aims to ensure that the data is easily Findable, Accessible, Interoperable and Reusable.

References

1. Dobzhansky T (1973) Nothing in biology makes sense except in the light of evolution. Am Biol Teach 35:125–129
2. de Carvalho MR, Bockmann FA, Amorim DS, de Vivo M, de Toledo-Piza M, Menezes NA, de Figueiredo JL, Castro RMC, Gill AC, McEachran JD, Compagno LJV, Schelly RC, Britz R, Lundberg JG, Vari RP, Nelson G (2005) Revisiting the taxonomic impediment. Science 307:353–353
3. Candolle AP (1813) Théorie Élémentaire de la botanique, ou Exposition des principes de la classification naturelle et de l'art de décrire et d'étudier les végétaux
4. Heywood VH, Watson RT (1995) Global biodiversity assessment. Cambridge University Press, Cambridge
5. Mayr E (1969) Principles of systematic zoology. Mcgraw-Hill, New. York
6. Simpson GG (1961) Principles of animal taxonomy. Columbia University Press, New York
7. Tillier S (2000) Systématique - Ordonner la diversité du Vivant. Rapport sur la science et la technologie de l'Académie des sciences n°11. Éditions Tec & Doc
8. Small E (1989) Systematics of biological systematics (or taxonomy of taxonomy). Taxon 38:335–356
9. Sprague JL, Lanjouw J, Andreas CH (1948) Minutes of the Utrecht conference. Chronica Botanica 12(1/2):12
10. Morton CV (1957) The misuse of the term taxon. Taxon 6(5):155
11. Raven PH (2004) Taxonomy: where are we now? Philos Trans R Soc Lond B Biol Sci 359:729–730
12. Pavord A (2005) The naming of names: the search for order in the world of plants. Bloomsbury, New York
13. Funk VA, Hoch PC, Prather LA, Wagner WL (2005) The importance of vouchers. Taxon 54:127–129
14. Knapp S (2012) What's in a name? A history of taxonomy. http://www.nhm.ac.uk/nature-online/science-of-natural-history/taxonomy-systematics/history-taxonomy. Accessed Jan 2012
15. Griffing LR (2011) Who invented the dichotomous key? Richard Waller's watercolors of the herbs of Britain. Am J Bot 98:1911–1923
16. Linnaeus C (1753) Species Plantarum. Stockholm
17. Linnaeus C (1758) Systema naturae, 10th edn. Stockholm
18. Candolle AP (1867) Lois de la nomenclature botanique adoptées par le Congrès international de botanique: tenu à Paris en août 1867. H. Georg, Geneva
19. Jussieu AL (1789) Genera plantarum. Herissant, Paris
20. Philippe H, Lecointre G, VanLe HL, LeGuyader H (1996) A critical study of homoplasy in molecular data with the use of a morphologically based cladogram, and its consequences for character weighting. Mol Biol Evol 13:1174–1186
21. Lamarck J-BPAM (1809) Philosophie zoologique. Dentu, Paris
22. Darwin C (1859) On the origin of species by means of natural selection, or the preservation of favoured races in the struggle for life. John Murray, London
23. Haeckel E (1866) Generelle Morphologie der Organismen. Reimer, Berlin
24. Dayrat B (2003) The roots of phylogeny: how did Haeckel build his trees? Syst Biol 52:515–527
25. Davis PH, Heywood PH (1963) Principles of angiosperm taxonomy. Oliver and Boyd, Edinburgh/London
26. Sneath PHA, Sokal RR (1963) Principles of numerical taxonomy, 7th edn. W. H. Freeman, San Francisco
27. Hennig W (1950) Grundzüge einer Theorie der phylogenetischen Systematik. Deutscher Zentralverlag, Berlin
28. Hennig W (1966) Phylogenetic systematics (tr. D. Davis and R. Zangerl), University of Illinois Press, Urbana
29. Godfray HCJ, Knapp S (2004) Taxonomy for the twenty-first century - introduction. Philos Trans R Soc Lond B Biol Sci 359:559–569
30. Mullis K, Faloona F, Scharf S, Saiki R, Horn G, Erlich H (1986) Specific enzymatic amplification of DNA Invitro - the polymerase chain-reaction. Cold Spring Harb Symp Quant Biol 51:263–273
31. Bremer K, Chase MW, Stevens PF, Anderberg AA, Backlund A, Bremer B, Briggs BG, Endress PK, Fay MF, Goldblatt P, Gustafsson MHG, Hoot SB, Judd WS, Kallersjo M, Kellogg EA, Kron KA, Les DH, Morton CM, Nickrent DL, Olmstead RG, Price RA, Quinn CJ, Rodman JE, Rudall PJ, Savolainen V, Soltis DE, Soltis PS, Sytsma KJ, Thulin M, Grp AP (1998) An ordinal classification for the families of flowering plants. Ann Mo Bot Gard 85:531–553

32. Bremer B, Bremer K, Chase MW, Reveal JL, Soltis DE, Soltis PS, Stevens PF, Anderberg AA, Fay MF, Goldblatt P, Judd WS, Kallersjo M, Karehed J, Kron KA, Lundberg J, Nickrent DL, Olmstead RG, Oxelman B, Pires JC, Rodman JE, Rudall PJ, Savolainen V, Sytsma KJ, van der Bank M, Wurdack K, Xiang JQY, Zmarzty S, Grp AP (2003) An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG II. Bot J Linn Soc 141:399–436
33. Bremer B, Bremer K, Chase MW, Fay MF, Reveal JL, Soltis DE, Soltis PS, Stevens PF, Anderberg AA, Moore MJ, Olmstead RG, Rudall PJ, Sytsma KJ, Tank DC, Wurdack K, Xiang JQY, Zmarzty S, Grp AP (2009) An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG III. Bot J Linn Soc 161:105–121
34. The Angiosperm phylogeny group (2016) An update of the angiosperm phylogeny group classification for the orders and families of flowering plants: APG IV. Bot J Linn Soc 181:1–20. https://doi.org/10.1111/boj.12385
35. PPGI (2016) A community-derived classification for extant lycophytes and ferns. J Syst Evol 54:563–603. https://doi.org/10.1111/jse.12229
36. Pryer KM, Schneider H, Smith AR, Cranfill R, Wolf PG, Hunt JS (2001) Horsetails and ferns are a monophyletic group and the closest living relatives to seed plants. Nature 409:618–621. https://doi.org/10.1038/35054555
37. Bessey CE (1915) The phylogenetic taxonomy of flowering plants. Ann Mo Bot Gard 2:109–164
38. Cronquist A (1981) An integrated system of classification of flowering plants. Columbia University Press, New York
39. Stebbins GL (1974) Flowering plants: evolution above the species level. Belknap press, Cambridge
40. Takhtajan A (1997) Diversity and classification of flowering plants. Columbia University Press, New York
41. Thorne RF (1976) A phylogenetic classification of the Angiospermae. Evol Biol 9:35–106
42. Soltis DE, Soltis PS, Endress PK, Chase MW (2005) Phylogeny and evolution of angiosperms. Sinauer associates, Sunderland
43. Hebert PDN, Cywinska A, Ball SL, DeWaard JR (2003) Biological identifications through DNA barcodes. Proc R Soc Lond B Biol Sci 270:313–321
44. May RM (2004) Tomorrow's taxonomy: collecting new species in the field will remain the rate-limiting step. Philos Trans R Soc Lond B Biol Sci 359:733–734
45. May RM (2011) Why worry about how many species and their loss? PLoS Biol 9:e1001130
46. Paton AJ, Brummitt N, Govaerts R, Harman K, Hinchcliffe S, Allkin B, Lughadha EN (2008) Towards target 1 of the global strategy for plant conservation: a working list of all known plant species - progress and prospects. Taxon 57:602–611
47. Wilson EO (2003) The encyclopedia of life. Trends Ecol Evol 18:77–80
48. Wilson EO (2004) Taxonomy as a fundamental discipline. Philos Trans R Soc Lond B Biol Sci 359:739–739
49. Candolle AP (1824-1873) Prodromus systematis naturalis regni vegetabilis. Sumptibus Sociorum Treuttel et Würtz, Parisii
50. International Plant Names Index (2012) Published on the Internet http://www.ipni.org. Accessed July 2019
51. Lughadha EN, Govaerts R, Belyaeva I, Black N, Lindon H, Allkin R, Magill RE, Nicolson N (2016) Counting counts: revised estimates of numbers of accepted 407 species of flowering plants, seed plants, vascular plants and land plants with a review 408 of other recent estimates. Phytotaxa 272:82–88
52. Scotland RW, Wortley AH (2003) How many species of seed plants are there? Taxon 52:101–104
53. The Plant List (2019) Version 1. Published on the Internet http://www.theplantlist.org/. Accessed July 2019
54. Wortley AH, Scotland RW (2004) Synonymy, sampling and seed plant numbers. Taxon 53:478–480
55. Mallet J, Willmott K (2003) Taxonomy: renaissance or tower of babel? Trends Ecol Evol 18:57–59
56. Isaac NJB, Mallet J, Mace GM (2004) Taxonomic inflation: its influence on macroecology and conservation. Trends Ecol Evol 19:464–469
57. Meiri S, Mace GM (2007) New taxonomy and the origin of species. PLoS Biol 5:1385–1386
58. Pillon Y, Chase MW (2007) Taxonomic exaggeration and its effects on orchid conservation. Conserv Biol 21:263–265
59. Crane PR (2004) Documenting plant diversity: unfinished business. Philos Trans R Soc Lond B Biol Sci 359:735–737

60. Joppa LN, Roberts DL, Pimm SL (2011) How many species of flowering plants are there? Proc R Soc B Biol Sci 278:554–559
61. Mora C, Tittensor DP, Adl S, Simpson AGB, Worm B (2011) How many species are there on earth and in the ocean? PLoS Biol 9: e1001127
62. Bisby FA, Roskov YR, Orrell TM, Nicolson D, Paglinawan LE et al (2010) Species 2000 & ITIS Catalogue of Life: 2010 Annual Checklist. Digital resource at http://www.catalogueoflife.org/annual-checklist/2010. Species 2000, Reading, UK
63. Caldecott JO, Jenkins MD, Johnson TH, Groombridge B (1996) Priorities for conserving global species richness and endemism. Biodivers Conserv 5:699–727
64. Joppa LN, Roberts DL, Myers N, Pimm SL (2011) Biodiversity hotspots house most undiscovered plant species. Proc Natl Acad Sci U S A 108:13171–13176
65. Callmander MW, Schatz GE, Lowry PP (2005) IUCN red list assessment and the global strategy for plant conservation: taxonomists must act now. Taxon 54:1047–1050
66. Godfray HCJ (2002) Challenges for taxonomy - the discipline will have to reinvent itself if it is to survive and flourish. Nature 417:17–19
67. Funk VA (2006) Floras: a model for biodiversity studies or a thing of the past? Taxon 55:581–588
68. Wheeler QD, Raven PH, Wilson EO (2004) Taxonomy: impediment or expedient? Science 303:285–285
69. Wheeler QD, Knapp S et al (2012) Mapping the biosphere: exploring species to understand the origin, organization and sustainability of biodiversity. Syst Biodivers 10(1):1–20
70. Ebach MC, Valdecasas AG, Wheeler QD (2011) Impediments to taxonomy and users of taxonomy: accessibility and impact evaluation. Cladistics 27:550–557
71. Ronquist F, Gardenfors U (2003) Taxonomy and biodiversity inventories: time to deliver. Trends Ecol Evol 18:269–270
72. Joly CA (2006) Taxonomy: programmes developing in the south too. Nature 440:24–24
73. Schatz GE, Lowry PP, Ramisamihantanirina A (1998) Takhtajania perrieri rediscovered. Nature 391:133–134
74. Jones WG, Hill KD, Allen JM (1995) *Wollemia nobilis*, a new living Australian genus and species in the Araucariaceae. Telopea 6:173–176
75. Mabberley DJ (2009) Exploring Terra Incognita. Science 324:472–472
76. Thulin M (2007) Acacia fumosa sp nov (Fabaceae) from eastern Ethiopia. Nord J Bot 25:272–274
77. Dransfield J, Rakotoarinivo M, Baker WJ, Bayton RP, Fisher JB, Horn JW, Leroy B, Metz X (2008) A new coryphoid palm genus from Madagascar. Bot J Linn Soc 156:79–91
78. Agnarsson I, Kuntner M (2007) Taxonomy in a changing world: seeking solutions for a science in crisis. Syst Biol 56:531–539
79. Crisci JV (2006) One-dimensional systematist: perils in a time of steady progress. Syst Bot 31:217–221
80. Joppa LN, Roberts DL, Pimm SL (2011) The population ecology and social behaviour of taxonomists. Trends Ecol Evol 26:551–553
81. Rodman JE, Cody JH (2003) The taxonomic impediment overcome: NSF's partnerships for enhancing expertise in taxonomy (PEET) as a model. Syst Biol 52:428–435
82. Bebber DP et al (2012) Big hitting collectors make massive and disproportionate contribution to the discovery of plant species. Proc R Soc Lond B Biol Sci 279(1736):2269–2274
83. Thiers B (2019) Index herbarium: a global directory of public herbaria and associated staff. New York Botanical Garden's Virtual Herbarium. http://sweetgum.nybg.org/ih/
84. Bebber DP, Carine MA, Wood JRI, Wortley AH, Harris DJ, Prance GT, Davidse G, Paige J, Pennington TD, Robson NKB, Scotland RW (2010) Herbaria are a major frontier for species discovery. Proc Natl Acad Sci U S A 107:22169–22171
85. Fontaine B, Perrard A, Bouchet P (2012) 21 years of shelf life between discovery and description of new species. Curr Biol 22(22): R943–R944. https://doi.org/10.1016/j.cub.2012.10.029
86. Le Bras G, Pignal M, Jeanson ML, Muller S, Aupic C, Carré B, Flament G, Gaudeul M, Gonçalves C, Invernón VR, Jabbour F, Lerat E, Lowry PP, Offroy B, Pérez Pimparé E, Poncy O, Rouhan G, Haevermans T (2017) The French Muséum national d'histoire naturelle vascular plant herbarium collection dataset. Scientific Data 4:170016
87. Godfray HCJ, Clark BR, Kitching IJ, Mayo SJ, Scoble MJ (2007) The web and the structure of taxonomy. Syst Biol 56:943–955
88. Knapp S, McNeill J, Turland NJ (2011) Changes to publication requirements made at the XVIII International Botanical Congress in Melbourne - what does e-publication mean for you? PhytoKeys (6):5–11
89. Nicolson N, Challis K, Tucker A, Knapp S (2017) Impact of e-publication changes in the International Code of Nomenclature for

algae, fungi and plants (Melbourne Code, 2012) - did we need to "run for our lives"? BMC Evol Biol 17:116. https://doi.org/10.1186/s12862-017-0961

90. Hebert PDN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. Syst Biol 54:852–859
91. Savolainen V, Cowan RS, Vogler AP, Roderick GK, Lane R (2005) Towards writing the encyclopaedia of life: an introduction to DNA barcoding. Philos Trans R Soc B Biol Sci 360:1805–1811
92. Wiens JJ (2007) Species delimitation: new approaches for discovering diversity. Syst Biol 56:875–878
93. Pannell JR (2009) Mating-system evolution: succeeding by celibacy. Curr Biol 19: R983–R985
94. Hood ME, Antonovics J (2003) Plant species descriptions show signs of disease. Proc R Soc Lond B Biol Sci 270:S156–S158
95. Duminil J, Di Michele M (2009) Plant species delimitation: a comparison of morphological and molecular markers. Plant Biosystems 143:528–542
96. Bickford D, Lohman DJ, Sodhi NS, Ng PKL, Meier R, Winker K, Ingram KK, Das I (2007) Cryptic species as a window on diversity and conservation. Trends Ecol Evol 22:148–155
97. Grundt HH, Kjolner S, Borgen L, Rieseberg LH, Brochmann C (2006) High biological species diversity in the arctic flora. Proc Natl Acad Sci U S A 103:972–975
98. Pillon Y, Hopkins HCF, Munzinger J, Amir H, Chase MW (2009) Cryptic species, gene recombination and hybridization in the genus Spiraeanthemum (Cunoniaceae) from New Caledonia. Bot J Linn Soc 161:137–152
99. Dulvy NK, Reynolds JD (2009) BIODIVERSITY skates on thin ice. Nature 462:417–417
100. Robertson A, Newton AC, Ennos RA (2004) Multiple hybrid origins, genetic diversity and population genetic structure of two endemic Sorbus taxa on the Isle of Arran, Scotland. Mol Ecol 13:123–134
101. Squirrell J, Hollingsworth PM, Bateman RM, Tebbitt MC, Hollingsworth ML (2002) Taxonomic complexity and breeding system transitions: conservation genetics of the Epipactis leptochila complex (Orchidaceae). Mol Ecol 11:1957–1964
102. van Dijk PJ (2003) Ecological and evolutionary opportunities of apomixis: insights from Taraxacum and Chondrilla. Philos Trans R Soc Lond Ser B Biol Sci 358:1113–1121
103. Ennos RA, French GC, Hollingsworth PM (2005) Conserving taxonomic complexity. Trends Ecol Evol 20:164–168
104. Ennos RA, Whitlock R, Fay MF, Jones B, Neaves LE, Payne R, Taylor I, De Vere N, Hollingsworth PM (2012) Process-based species action plans: an approach to conserve contemporary evolutionary processes that sustain diversity in taxonomically complex groups. Bot J Linn Soc 168:194–203
105. Li FW, Tan BC, Buchbender V, Moran RC, Rouhan G, Wang CN, Quandt D (2009) Identifying a mysterious aquatic fern gametophyte. Plant Syst Evol 281:77–86
106. Van Deynze A, Stoffel K (2006) High-throughput DNA extraction from seeds. Seed Sci Technol 34:741–745
107. Asif MJ, Cannon CH (2005) DNA extraction from processed wood: a case study for the identification of an endangered timber species (Gonystylus bancanus). Plant Mol Biol Report 23:185–192
108. Colpaert N, Cavers S, Bandou E, Caron H, Gheysen G, Lowe AJ (2005) Sampling tissue for DNA analysis of trees: trunk cambium as an alternative to canopy leaves. Silvae Genetica 54:265–269
109. Rachmayanti Y, Leinemann L, Gailing O, Finkeldey R (2006) Extraction, amplification and characterization of wood DNA from Dipterocarpaceae. Plant Mol Biol Report 24:45–55
110. Tibbits JFG, McManus LJ, Spokevicius AV, Bossinger G (2006) A rapid method for tissue collection and high-throughput isolation of genomic DNA from mature trees. Plant Mol Biol Report 24:81–91
111. Novaes RML, Rodrigues JG, Lovato MB (2009) An efficient protocol for tissue sampling and DNA isolation from the stem bark of Leguminosae trees. Genet Mol Res 8:86–96
112. Deguilloux MF, Pemonge MH, Petit RJ (2002) Novel perspectives in wood certification and forensics: dry wood as a source of DNA. Proc R Soc B Biol Sci 269:1039–1046
113. Hiiesalu I, Opik M, Metsis M, Lilje L, Davison J, Vasar M, Moora M, Zobel M, Wilson SD, Partel M (2012) Plant species richness belowground: higher richness and new patterns revealed by next-generation sequencing. Mol Ecol 21:2004–2016
114. Kesanakurti PR, Fazekas AJ, Burgess KS, Percy DM, Newmaster SG, Graham SW, Barrett SCH, Hajibabaei M, Husband BC (2011) Spatial patterns of plant diversity belowground as revealed by DNA barcoding. Mol Ecol 20:1289–1302
115. Dunn CP (2003) Keeping taxonomy based in morphology. Trends Ecol Evol 18:270–271
116. Santos LM, Faria LRR (2011) The taxonomy's new clothes: a little more about the DNA-based taxonomy. Zootaxa 3025:66–68

117. Schaefer H, Carine MA, Rumsey FJ (2011) From European priority species to invasive weed: Marsilea azorica (Marsileaceae) is a misidentified alien. Syst Bot 36:845–853
118. Launert GOE, Paiva JAR (1983) Iconographia selecta florae Azoricae. Coimbra 2:159
119. Lipscomb D, Platnick N, Wheeler Q (2003) The intellectual content of taxonomy: a comment on DNA taxonomy. Trends Ecol Evol 18:65–66
120. Sites JW, Marshall JC (2004) Operational criteria for delimiting species. Annu Rev Ecol Evol Syst 35:199–227
121. Stace CA (2005) Plant taxonomy and biosystematics - does DNA provide all the answers? Taxon 54:999–1007
122. Linder CR, Rieseberg LH (2004) Reconstructing patterns of reticulate evolution UN plants. Am J Bot 91:1700–1708
123. Vriesendorp B, Bakker FT (2005) Reconstructing patterns of reticulate evolution in angiosperms: what can we do? Taxon 54:593–604
124. Egan AN, Schlueter J, Spooner DM (2012) Applications of next-generation sequencing in plant biology. Am J Bot 99:175–185
125. Harrison N, Kidner CA (2011) Next-generation sequencing and systematics: what can a billion base pairs of DNA sequence data do for you? Taxon 60:1552–1566
126. Straub SCK, Parks M, Weitemier K, Fishbein M, Cronn RC, Liston A (2012) Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. Am J Bot 99:349–364
127. Bakker FT (2015) DNA sequences from plant herbarium tissue (Chapter 8). In: Hörandl E, Appelhans MS (eds) Next-generation sequencing in plant systematics. International Association for Plant Taxonomy (IAPT). https://doi.org/10.14630/000009
128. Richardson JE, Pennington RT, Pennington TD, Hollingsworth PM (2001) Rapid diversification of a species-rich genus of neotropical rain forest trees. Science 293:2242–2245
129. Baldwin BG, Sanderson MJ (1998) Age and rate of diversification of the Hawaiian silversword alliance (Compositae). Proc Natl Acad Sci U S A 95:9402–9406
130. Wang AL, Yang MH, Liu JQ (2005) Molecular phylogeny, recent radiation and evolution of gross morphology of the rhubarb genus Rheum (Polygonaceae) inferred from chloroplast DNA trnL-F sequences. Ann Bot 96:489–498
131. Hodges SA, Arnold ML (1994) Columbines - a geographically widespread species flock. Proc Natl Acad Sci U S A 91:5129–5132
132. Linder HP (2008) Plant species radiations: where, when, why? Philos Trans R Soc B Biol Sci 363:3097–3105
133. Tautz D, Arctander P, Minelli A, Thomas RH, Vogler AP (2003) A plea for DNA taxonomy. Trends Ecol Evol 18:70–74
134. Hey J, Pinho C (2012) Population genetics and objectivity in species diagnosis. Evolution 66:1413–1429
135. Knowles L, Carstens B (2007) Delimiting species without monophyletic gene trees. Syst Biol 56:887–895
136. Staats M, Cuenca A, Richardson JE, Vrielink-van Ginkel R, Petersen G, Seberg O, Bakker FT (2011) DNA damage in plant herbarium tissue. PLoS One 6:e28448
137. Seberg O, Humphries CJ, Knapp S, Stevenson DW, Petersen G, Scharff N, Andersen NM (2003) Shortcuts in systematics? A commentary on DNA-based taxonomy. Trends Ecol Evol 18:63–65
138. Lister DL, Bower MA, Howe CJ, Jones MK (2008) Extraction and amplification of nuclear DNA from herbarium specimens of emmer wheat: a method for assessing DNA preservation by maximum amplicon length recovery. Taxon 57:254–258
139. Wandeler P, Hoeck PEA, Keller LF (2007) Back to the future: museum specimens in population genetics. Trends Ecol Evol 22:634–642
140. Cozzolino S, Cafasso D, Pellegrino G, Musacchio A, Widmer A (2007) Genetic variation in time and space: the use of herbarium specimens to reconstruct patterns of genetic variation in the endangered orchid Anacamptis palustris. Conserv Genet 8:629–639
141. Erkens RHJ, Cross H, Maas JW, Hoenselaar K, Chatrou LW (2008) Assessment of age and greenness of herbarium specimens as predictors for successful extraction and amplification of DNA. Blumea 53:407–428
142. Drabkova L, Kirschner J, Vlcek C (2002) Comparison of seven DNA extraction and amplification protocols in historical herbarium specimens of Juncaceae. Plant Mol Biol Report 20:161–175
143. Jankowiak K, Buczkowska K, Szweykowska-Kulinska Z (2005) Successful extraction of DNA from 100-year-old herbarium specimens of the liverwort Bazzania trilobata. Taxon 54:335–336
144. Korpelainen H, Pietilainen M (2008) Effort to reconstruct past population history in the fern Blechnum spicant. J Plant Res 121:293–298
145. Savolainen V, Cuenoud P, Spichiger R, Martinez MDP, Crevecoeur M, Manen JF (1995)

The use of herbarium specimens in DNA Phylogenetics - evaluation and improvement. Plant Syst Evol 197:87–98

146. Ribeiro RA, Lovato MB (2007) Comparative analysis of different DNA extraction protocols in fresh and herbarium specimens of the genus Dalbergia. Genet Mol Res 6:173–187
147. Andreasen K, Manktelow M, Razafimandimbison SG (2009) Successful DNA amplification of a more than 200-year-old herbarium specimen: recovering genetic material from the Linnaean era. Taxon 58:959–962
148. Ames M, Spooner DM (2008) DNA from herbarium specimens settles a controversy about origins of the European potato. Am J Bot 95:252–257
149. Walters C, Reilley AA, Reeves PA, Baszczak J, Richards CM (2006) The utility of aged seeds in DNA banks. Seed Sci Res 16:169–178
150. Alves RJV, Machado MD (2007) Is classical taxonomy obsolete? Taxon 56:287–288
151. DeSalle R, Egan MG, Siddall M (2005) The unholy trinity: taxonomy, species delimitation and DNA barcoding. Philos Trans R Soc B Biol Sci 360:1905–1916
152. DeSalle R (2006) Species discovery versus species identification in DNA barcoding efforts: response to Rubinoff. Conserv Biol 20:1545–1547
153. Schlick-Steiner BC, Steiner FM, Seifert B, Stauffer C, Christian E, Crozier RH (2010) Integrative taxonomy: a multisource approach to exploring biodiversity. Annu Rev Entomol 55:421–438
154. Dayrat B (2005) Towards integrative taxonomy. Biol J Linn Soc 85:407–415
155. Wheeler QD (2005) Losing the plot: DNA "barcodes" and taxonomy. Cladistics 21:405–407
156. Corney DPA, Clark JY, Tang HT, Wilkin P (2012) Automatic extraction of leaf characters from herbarium specimens. Taxon 61 (1):231–244
157. Raxworthy CJ, Ingram CM, Rabibisoa N, Pearson RG (2007) Applications of ecological niche modeling for species delimitation: a review and empirical evaluation using day geckos (Phelsuma) from Madagascar. Syst Biol 56:907–923
158. Rissler LJ, Apodaca JJ (2007) Adding more ecology into species delimitation: ecological niche models and phylogeography help define cryptic species in the black salamander (Aneides flavipunctatus). Syst Biol 56:924–942
159. Wiens JJ, Graham CH (2005) Niche conservatism: integrating evolution, ecology, and conservation biology. Annu Rev Ecol Evol Syst 36:519–539
160. Brautigam A, Gowik U (2010) What can next generation sequencing do for you? Next generation sequencing as a valuable tool in plant research. Plant Biol 12:831–841
161. Cronn R, Knaus BJ, Liston A, Maughan PJ, Parks M, Syring JV, Udall J (2012) Targeted enrichment strategies for next-generation plant biology. Am J Bot 99:291–311
162. Rodriguez-Fernandez JI, De Carvalho CJB, Pasquini C, De Lima KMG, Moura MO, Arizaga GGC (2011) Barcoding without DNA? Species identification using near infrared spectroscopy. Zootaxa 46–54
163. Munck L, Jespersen BM, Rinnan A, Seefeldt HF, Engelsen MM, Norgaard L, Engelsen SB (2010) A physiochemical theory on the applicability of soft mathematical models-experimentally interpreted. J Chemom 24:481–495
164. Cruickshank RH, Munck L (2011) It's barcoding Jim, but not as we know it. Zootaxa 2933:55–56
165. Andres-Sanchez S, Rico E, Herrero A, Santos-Vicente M, Martinez-Ortega MM (2009) Combining traditional morphometrics and molecular markers in cryptic taxa: towards an updated integrative taxonomic treatment for Veronica subgenus Pentasepalae (Plantaginaceae sensu APG II) in the western Mediterranean. Bot J Linn Soc 159:68–87
166. Schlick-Steiner BC, Seifert B, Stauffer C, Christian E, Crozier RH, Steiner FM (2007) Without morphology, cryptic species stay in taxonomic crypsis following discovery. Trends Ecol Evol 22:391–392
167. Blaxter M, Mann J, Chapman T, Thomas F, Whitton C, Floyd R, Abebe E (2005) Defining operational taxonomic units using DNA barcode data. Philos Trans R Soc B Biol Sci 360:1935–1943
168. Markmann M, Tautz D (2005) Reverse taxonomy: an approach towards determining the diversity of meiobenthic organisms based on ribosomal RNA signature sequences. Philos Trans R Soc B Biol Sci 360:1917–1924
169. Pleijel F, Jondelius U, Norlinder E, Nygren A, Oxelman B, Schander C, Sundberg P, Thollesson M (2008) Phylogenies without roots? A plea for the use of vouchers in molecular phylogenetic studies. Mol Phylogenet Evol 48:369–371
170. Puillandre N, Bouchet P, Boisselier-Dubayle MC, Brisset J, Buge B, Castelin M, Chagnoux S, Christophe T, Corbari L,

Lambourdiere J, Lozouet P, Marani G, Rivasseau A, Silva N, Terryn Y, Tillier S, Utge J, Samadi S (2012) New taxonomy and old collections: integrating DNA barcoding into the collection curation process. Mol Ecol Resour 12:396–402

171. Gemeinholzer B, Bachmann K (2005) Examining morphological and molecular diagnostic character states of Cichorium intybus L. (Asteraceae) and C-spinosum L. Plant Syst Evol 253:105–123
172. Bacon CD, McKenna MJ, Simmons MP, Wagner WL (2012) Evaluating multiple criteria for species delimitation: an empirical example using Hawaiian palms (Arecaceae: Pritchardia). BMC Evol Biol 12:23
173. Barrett CF, Freudenstein JV (2011) An integrative approach to delimiting species in a rare but widespread mycoheterotrophic orchid. Mol Ecol 20:2771–2786
174. Koffi KG, Heuertz M, Doumenge C, Onana JM, Gavory F, Hardy OJ (2010) A combined analysis of morphological traits, chloroplast and nuclear DNA sequences within Santiria trimera (Burseraceae) suggests several species following the biological species concept. Plant Ecol Evol 143:160–169
175. Ley AC, Hardy OJ (2010) Species delimitation in the central African herbs Haumania (Marantaceae) using georeferenced nuclear and chloroplastic DNA sequences. Mol Phylogenet Evol 57:859–867
176. Meudt HM, Lockhart PJ, Bryant D (2009) Species delimitation and phylogeny of a New Zealand plant species radiation. BMC Evol Biol 9:111
177. Schmidt-Lebuhn AN (2007) Using amplified fragment length polymorphism (AFLP) to unravel species relationships and delimitations in Minthostachys (Labiatae). Bot J Linn Soc 153:9–19
178. Zeng YF, Liao WJ, Petit RJ, Zhang DY (2010) Exploring species limits in two closely related Chinese oaks. PLoS One 5:e15529
179. Rieseberg LH, Troy TE, Baack EJ (2006) The nature of plant species. Nature 440:524–527
180. de Queiroz K (2005) Ernst Mayr and the modern concept of species. Proc Natl Acad Sci U S A 102:6600–6607
181. de Queiroz K (2007) Species concepts and species delimitation. Syst Biol 56(6):879–886
182. Wright S (1940) The statistical consequences of Mendelian heredity in relation to speciation. In: Huxley J (ed) The new systematics. Oxford university press, London, pp 161–183
183. Mayr E (1942) Systematics and the origin of species. Columbia university press, New York
184. Dobzhansky T (1950) Mendelian populations and their evolution. Am Nat 84:401–418
185. Poulton EB (1904) What is a species? Proc Entomol Soc Lond 1903:lxxviicxvi
186. Dobzhansky T (1970) Genetics of the evolutionary process. Columbia University Press, New York
187. Sokal RR, Crovello TJ (1970) The biological species concept: a critical evaluation. Am Nat 104:107–123
188. Van Valen L (1976) Ecological species, multispecies, and oaks. Taxon 25:233–239
189. Simpson GG (1951) The species concept. Evolution 5:285–298
190. Wiley EO (1978) The evolutionary species concept reconsidered. Syst Zool 21:17–26
191. Cracraft J (1989) Speciation and its ontology: the empirical consequences of alternative species concepts for understanding patterns and processes of differentiation. In: Otte D, Endler JA (eds) Speciation and its consequences. Sinauer Associates, Sunderland, pp 28–59
192. Rosen DE (1979) Fishes from the uplands and intermontane basins of Guatemala: revisionary studies and comparative geography. Bull Am Mus Nat Hist 162:267–376
193. Donoghue MJ (1985) A critique of the biological species concept and recommendations for a phylogenetic alternative. Bryologist 88:172–181
194. Mishler BD (1985) The morphological, developmental, and phylogenetic basis of species concepts in bryophytes. Bryologist 88:207–214
195. Baum DA, Shaw KL (1995) Genealogical perspectives on the species problem. In: Hoch PC, Stephenson AG (eds) Experimental and molecular approaches to plant biosystematics. Missouri Botanical Garden, St. Louis, pp 289–303
196. Mallet J (1995) A species definition for the modern synthesis. Trends Ecol Evol 10:294–299
197. Templeton AR (1998) Species and speciation: geography, population structure, ecology, and gene trees. In: Howard DJ, Berlocher SH (eds) Endless forms: species and speciation. Oxford university press, New York, pp 32–43
198. Sites JW, Marshall JC (2003) Delimiting species: a renaissance issue in systematic biology. Trends Ecol Evol 18(9):462–470

Chapter 2

Guidelines for the Choice of Sequences for Molecular Plant Taxonomy

Pascale Besse

Abstract

This chapter presents an overview of the major plant DNA sequences and molecular methods available for plant taxonomy. Guidelines are provided for the choice of sequences and methods to be used, based on the DNA compartment (nuclear, chloroplastic, mitochondrial), evolutionary mechanisms, and the level of taxonomic differentiation of the plants under survey.

Key words Nuclear DNA, Chloroplast DNA, Mitochondrial DNA, Repeated DNA, Low copy DNA, Evolution, Molecular plant taxonomy

1 The Plant Genome and Regions Targeted for Molecular Plant Taxonomy

The nuclear genome in plants is very complex as in many eukaryotes, as illustrated by the "C-value enigma" [1, 2]: although the overall haploid DNA content (C-value) increases with apparent biological complexity, some species have more DNA in their haploid genome than some more complex organisms. Also, for a similar level of biological complexity, some species, such as plants, exhibit a surprisingly wide range of C-values (Fig. 1). This apparent discrepancy can be in part explained by the occurrence of variable amounts of repetitive DNA in the genomes (Fig. 1), most of which is constituted by noncoding sequences [4].

1.1 Repeated Nuclear DNA Sequences

Most nuclear sequences targeted in molecular taxonomy experiments belong to the category of highly repetitive DNA. Nuclear ribosomal RNA genes (nrDNA) are tandemly (side by side) repeated and located at a few loci in plant genomes [5–7] (Fig. 2). These, and particularly the ITS (internal transcribed spacers) [8, 9], have long been widely used for resolving plant taxonomic issues, initially using restriction analysis and then sequencing (Chapter 7). Microsatellite markers, also called STR

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_2,

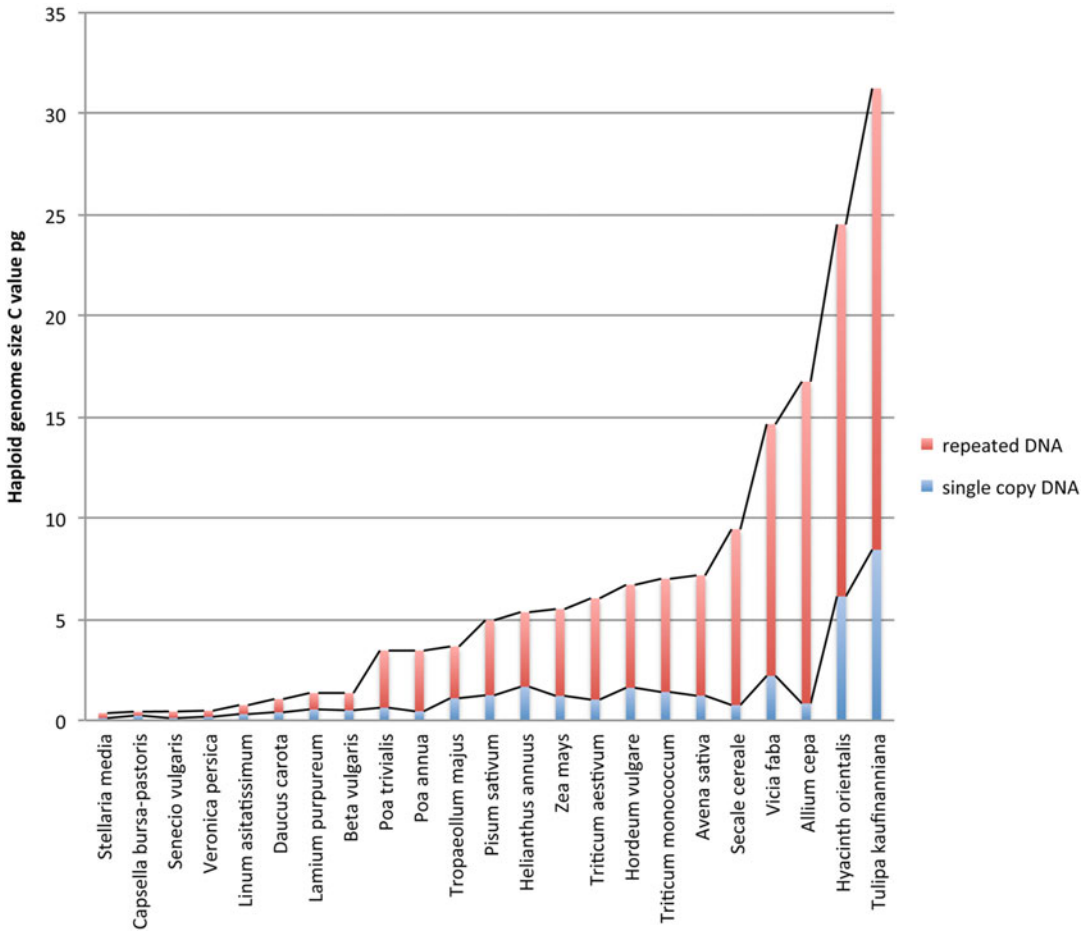


Fig. 1 Haploid genome size and composition for different plant species (graph built from data taken in [3])

(simple tandem repeats) or SSR (simple sequence repeats), are tandem repeats of small stretches of noncoding DNA sequences, discovered in 1989 and named after the discovery of minisatellite and satellite DNA, which exhibited a similar tandem structure [10] (Fig. 2). Microsatellites are also widely used for diversity studies in plants, either as powerful single locus markers easily amplified by PCR (Chapter 11) or in multi-locus profiling methods revealing regions between adjacent SSRs (inter-simple sequence repeats, ISSR) by PCR amplification (Fig. 3)(Chapter 14).

Transposable elements (TEs) represent another class of repeated DNA, but the elements are dispersed across the genome instead of being tandemly repeated and these also can represent an important part of the plant nuclear genome. Two main classes of TEs exist in plants: class I retrotransposons (which transpose through an RNA copy which is then reverse transcribed into DNA and inserted at a new site—in a "copy/paste" mode) and class II transposons (which are excised and transposed directly as

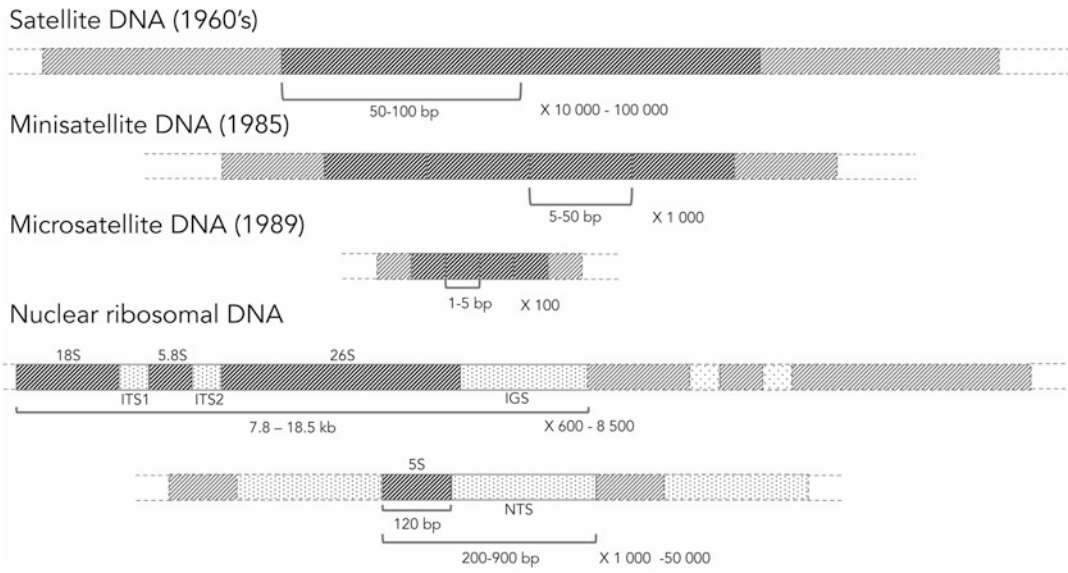


Fig. 2 Tandem repeat sequences used for molecular plant taxonomy: structure and number of tandem repeats

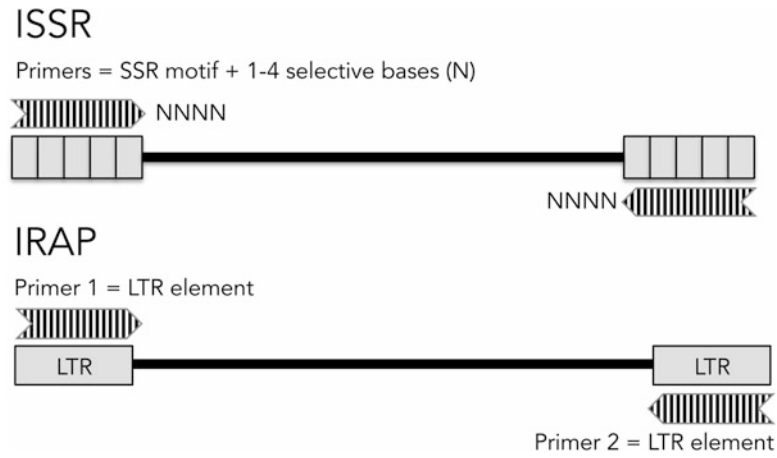


Fig. 3 : Multi-locus profiling methods using either SSR or retrotransposons as anchors

DNA—in a "cut/paste" mode) (Fig. 4). Class I retrotransposons are more numerous in genomes than class II as the original copy of the transposon is retained after transposition. In maize for example, LTR (Long Terminal Repeats) retrotransposons represent up to 70% of the nuclear genome [11]. Class I transposons (either LTR-retrotransposons or non LTR-SINEs, Short Interspersed Nuclear Elements) are now commonly used for phylogenetic and taxonomic studies. Many studies use multi-locus PCR-based profiling methods such as Inter-Retrotransposons Amplified Polymorphism (IRAP) (Fig. 3), which amplifies regions between

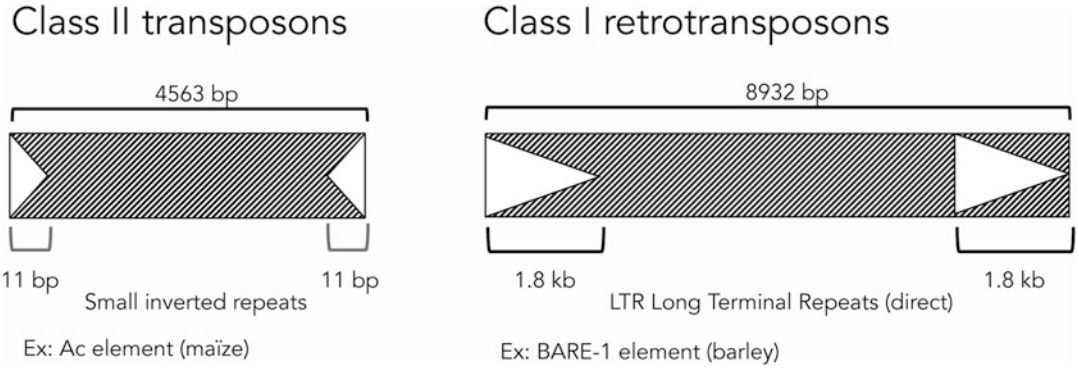


Fig. 4 Transposable elements in plants

adjacent LTR repeats of LTR-retrotransposons (Chapter 15). As described for eukaryotes [12], SINE elements are considered as perfect makers and are also being sequenced to build robust plant phylogenies although these studies are restricted to a limited number of plant species (mainly cultivated species) for which SINEs have been described and isolated [13, 14].

1.2 Low Copy Nuclear Genes and SNPs

Contrary to ribosomal DNA nuclear genes, low-copy nuclear genes (LCNG) do not suffer the possible disadvantages of concerted evolution, paralogy and homoplasy [8, 9, 15, 16] that can be particularly limiting for taxonomic studies in recent hybrids or polyploids (Chapter 7). However, care must be taken if using low-copy genes belonging to multigenic families for which paralogy and concerted evolution issues might still be problematic [17].

Despite their advantages, single-copy nuclear genes have not, for a long time, so much been used for plant taxonomy as they are much more difficult to isolate and characterize, contrary to chloroplast DNA or ribosomal nuclear DNA which have been extensively used because they are easily amplified using universal primers (as described in the earlier version of this book [18]) (and *see* Chapter 5). This situation has however changed rapidly [16, 17, 19]. With the availability and affordability of new sequencing technologies [20, 21] [22], it is now becoming feasible to assess variations at a wide range of single or low copy genes in nuclear genomes giving access to powerful phylogenomic analyses [23, 24]. The availability of an increasing number of complete plant genome sequences [25] now allows single nucleotide polymorphisms (SNPs) to be searched for and analyzed (Fig. 5) (Chapter 9). Even without a complete genome sequence, various sequence-based SNP assays can be designed [26]. Simple methods such as GBS (Genotyping By Sequencing) now allow to reveal numerous SNPs markers without sequencing whole genomes [27] (Chapter 10). GBS technique is now often preferred to the high-throughput DNA-array technology DArT (Diversity Array Technology) [28, 29].

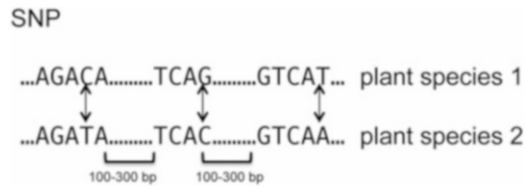


Fig. 5 SNPs in plants

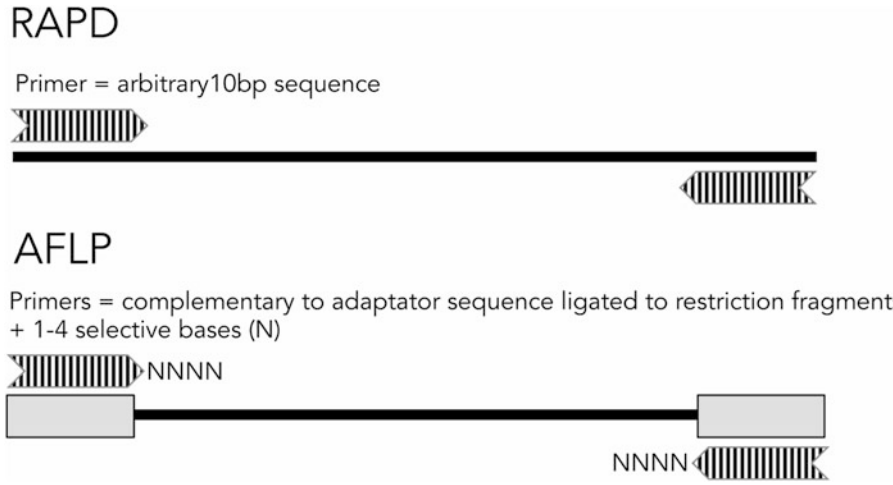


Fig. 6 Markers revealed by RAPD and AFLP

1.3 Anonymous Sequences

Many molecular technologies also rely on revealing variations at randomly picked anonymous sequences in genomes. In such techniques, the importance is not the nature of the target sequence itself, but rather the high throughput of the technology, which allows revealing numerous markers (loci) covering the genome. The aim is to give an as-accurate-as-possible view of the genome diversity. This is the case for Amplified Fragment Length Polymorphism (AFLP) (Chapter 12), Randomly Amplified Polymorphic DNA (RAPD) (Chapter 13), and associated techniques (Fig. 6) which use primers with arbitrary sequence to amplify genomic regions. Some multi-locus profiling techniques use a combination of AFLP associated with the revelation of either SSR loci (Selective Amplification of Microsatellite Polymorphic Loci, SAMPL) [30] (Chapter 12) or LTR-retrotransposons (Sequence-Specific Amplified Polymorphism, SSAP) (Chapter 15), others combine anchor primers in both SSR- and LTR-retrotransposons conserved regions (Retrotransposon-Microsatellite Amplification Polymorphism, REMAP) (Chapter 15) (Fig. 7).

1.4 Organellar DNA

In plants, genetic information is also carried in the mitochondrial as well as chloroplast genomes (organellar DNA). Although mitochondrial genome (mtDNA) has received little attention in plant

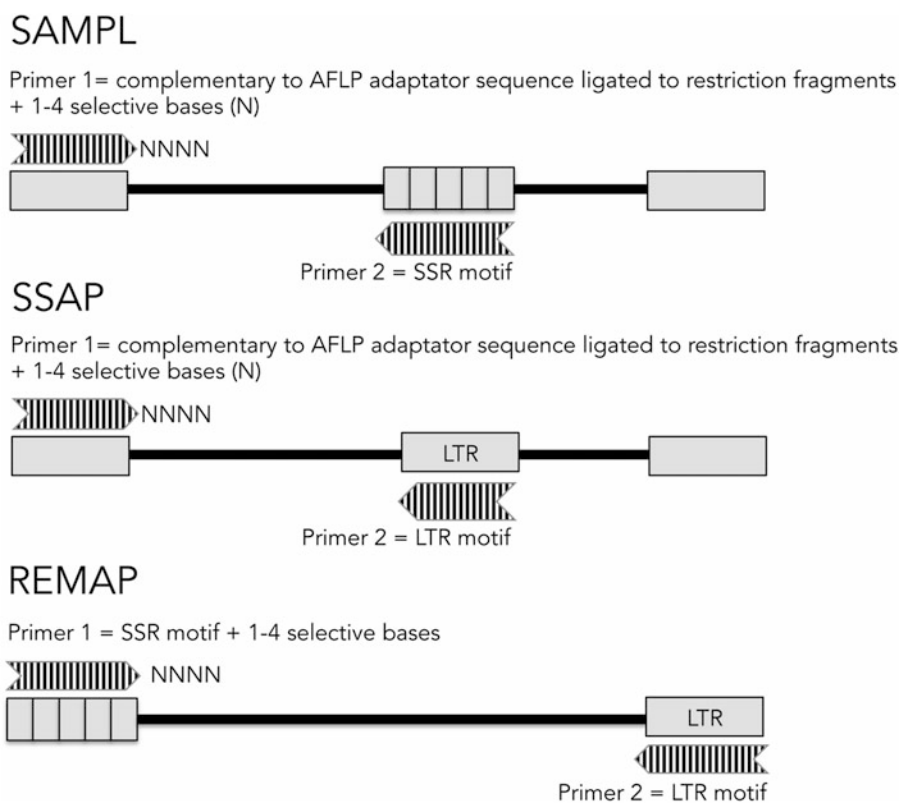


Fig. 7 Multi-locus profiling methods using a combination of anchors based on AFLP, SSR, or LTR

taxonomic studies (but see Chapter 6) because of numerous rearrangements and low levels of sequence variation, chloroplast DNA (cpDNA) has been widely used in molecular plant phylogeny (Chapter 5). When in earlier years the use of these sequences simply relied on PCR Sanger sequencing of target mitochondrial [31] or chloroplast genes or intergenic regions [18], as described in the earlier version of this book, nowadays with the affordability of NGS (next-generation sequencing) techniques, researchers are more and more engaged in whole chloroplast and mitochondrial genome sequencing (Chapters 5 and 6).

2 Evolutionary Considerations

The molecular clock hypothesis suggests that nucleotide substitutions occur at a roughly constant rate between and within evolutionary lineages across time [32] and has given rise to different models to estimate this evolutionary rate and its constancy [33]. According to the neutral theory of evolution, the speed of this rate (the amount of molecular variation accumulated over

time) depends on the structural and functional constraints of the molecule [34]. This can be illustrated by noncoding DNA molecules (such as introns or intergenic sequences) evolving much faster than coding DNA as they accumulate more variations over time. Also it is now well admitted that third position bases in codons evolve much faster than other positions due to the redundancy of the genetic code [34] (less functional constraint on the third position allows for more variations to accumulate over time). Most markers generated using RAPD or AFLP technologies have been shown by genome mapping experiments to cluster around the centromeres of chromosomes [35–38], a heterochromatin region with mainly noncoding sequences. Consequently, these markers often reveal an important amount of variation.

The evolutionary rate of a molecule is also driven by its evolutionary mechanisms. Microsatellite markers are the most variable molecules known to date. They are mostly noncoding molecules and vary in length (due to the variation in the number of tandem repetitions or VNTR) due to replication slippage (stepwise mutation model SMM [39]), which occurs at a high frequency (10^{-6} to 10^{-2}) in plants [40]. Microsatellites with shorter motifs and greater number of repeats are more prone to replication slippage and are thus the most variable [41]. ISSR, SAMPL, and REMAP markers, which use a microsatellite locus as an anchor, also beneficiate to a certain extent from the microsatellite length hypervariability. Minisatellite sequences that tend to evolve through unequal crossing-over (infinite allele model IAM [39]), which is a phenomenon with greater frequency than simple base mutations, also vary in length (i.e., number of tandem repeats) with great frequency. For this reason both types of sequences have been used for generating powerful DNA fingerprints in human [42, 43] and subsequently in numerous species, including plants.

Most tandemly repeated sequences in the genome evolve through what is known as "concerted evolution" or molecular drive [44, 45], which involves mechanisms such as unequal crossing over or biased gene conversion. Over time, the sequences that compose a family of tandem repeats within an individual genome are maintained similar thanks to this concerted evolution [7, 46–48]. Such sequences also tend to be maintained identical through close lineages within a species and will therefore display a slower evolutionary rate than molecules without concerted evolution.

In the cpDNA, like in the nDNA, intergenic noncoding sequences evolve faster than coding sequences. For example, by testing 7 different sequences on a range of land plants, [49] classified these sequences by order of variation as follows: *psb*K-*psb*I > *trn*H-*psb*A > *atp*F-*atp*H > *mat*K > *rpo*B > *rpo*C1 > *rbc*L, illustrating that cpDNA intergenic regions are more variable than coding regions. Globally, in plants, organellar sequences evolve more slowly than nuclear sequences: mtDNA evolves 3 times

slower than cpDNA, which in turn evolves 2 times slower than nDNA (average synonymous substitution rates per site per year for mtDNA and cpDNA are $0.2–1.0 \times 10^{-9}$ and $1.0–3.0 \times 10^{-9}$, respectively [50]) (Chapter 6). Even the most variable of intergenic regions in cpDNA is less variable than nuclear ITS: ITS reveals 2.81% sequence divergence in a range of plant families compared to 1.24% divergence for *trn*H-*psb*A, one of the most variable intergenic cpDNA regions [51].

Finally, Class I TEs are good classification criteria to evaluate species phylogenetic relationships, their mode of transposition ("copy-paste" mode) makes them numerous and implies no ambiguity in the ancestral state definition, which is, for a given locus, the absence of TE [12, 13]. Class II TEs are less appropriate for phylogenetic issues mainly because of their direct mode of transposition ("cut-paste" mode). However, it is of note that TEs show possibilities of horizontal transfer [52], which can lead to erroneous classifications (TE phylogenetic trees not concordant with species phylogenetic history) [53, 54].

3 Choice of Sequences for Molecular Taxonomy

These evolutionary considerations are of primary importance when one wants to use a DNA sequence to infer phylogenetic relationships between a set of accessions. Two questions have to be considered when starting a molecular taxonomy project:

1. What is the degree of time divergence between the accessions under study? Do we want to address variations at the intraspecific level (population level) or are we comparing species from the same genus or different genera from the same family or above?
2. What is the evolutionary rate of the molecule that will be used to infer relationships between accessions?

The rule to keep in mind is that the further we need to go in evolutionary times, the slower the molecule must evolve. Going too far with too much diverging sequences will lead to homoplasy (characters identical by state, not by descent) through convergence or reversion. On the contrary, slow-evolving sequences will not be enough discriminating for groups that have evolved recently (Fig. 8). Figure 9 illustrates this rule: if a very slow-evolving sequence is used, it might be unable to differentiate the two hypothetical species under study (Fig. 9A). A sequence with an intermediate rate of evolution and concerted evolution would allow the identification of each species, but would be unable to reveal any intraspecific variability (Fig. 9B). To reach such level of informativeness, one would need to use a single copy gene (Fig. 9C) or a

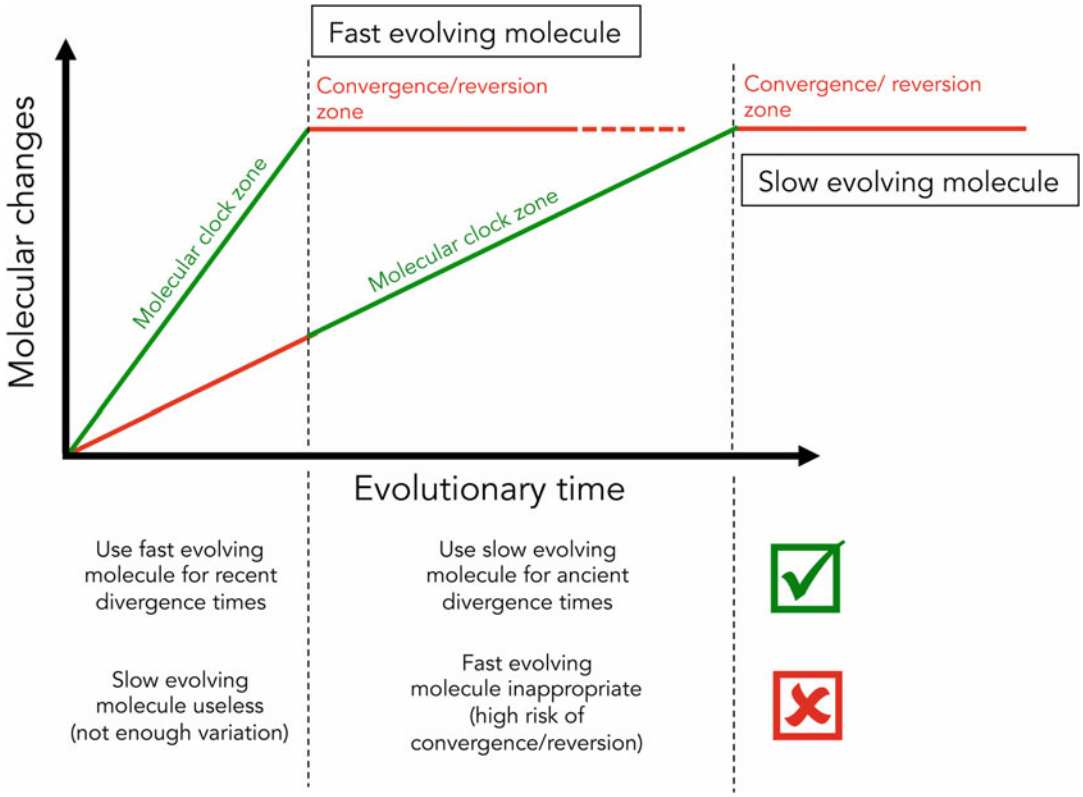


Fig. 8 Illustration of the usefulness of rapidly evolving versus slow-evolving sequences in molecular taxonomy assessment of recently or anciently diverged groups. The curvilinear relationship between molecular changes and time is represented theoretically, starting with a constant accumulation rate (molecular clock hypothesis), which plateaus as a consequence of the saturation of the sequence over time. The faster the sequence evolves, the faster the plateau is reached

microsatellite marker (Fig. 9D), but the latter, due to high evolutionary rate, may generate homoplasy (*), which could lead to erroneous interpretations if comparing species A and B, as individual 4B would appear more related to species A than to individuals from species B. Such rapidly evolving sequences are therefore not appropriate for studying relationships at taxonomic levels too high.

Guidelines for the choice of sequences to be used depending on the level of taxonomic divergence are illustrated (Fig. 10). It must be kept in mind that the level of taxonomic differentiation can vary considerably depending on the species group, therefore one always needs to perform preliminary tests of various sequences on a representative subset of accessions to assess their power in differentiating our own individuals, species, or genera of interest.

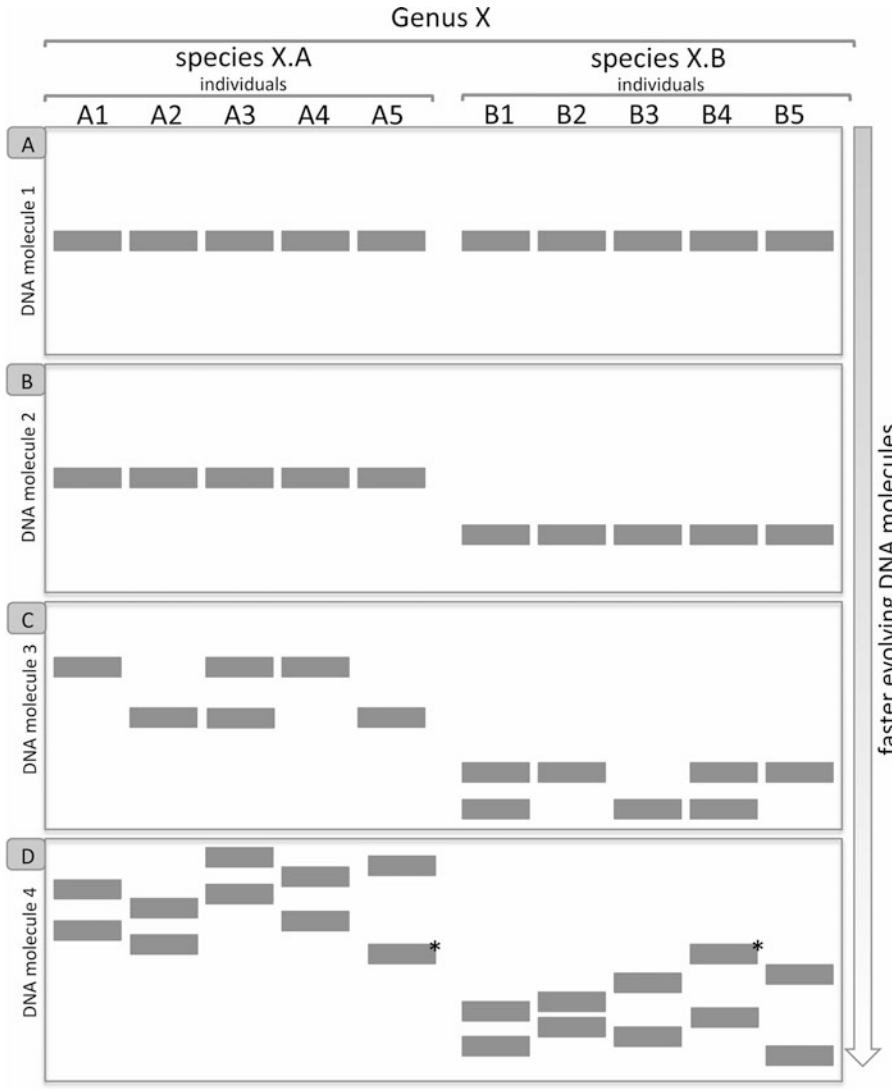


Fig. 9 Illustration of the differentiation power of DNA molecules depending on their evolutionary rates. Panels A to D represent hypothetical electrophoresis results using four different DNA molecules with increasing evolutionary rate. * indicates fortuitous co-migrating bands i.e. homoplasy

4 Genetic Considerations

Knowledge of the mode of inheritance of the molecules under study is also of great importance. Nuclear sequences are inherited in a Mendelian fashion, with contribution from both parents. Organellar (chloroplastic and mitochondrial) sequences are almost always uniparentally inherited (generally maternally, but see [55]). This can have important consequences when building a molecular phylogeny, as individuals or species of interspecific origin will

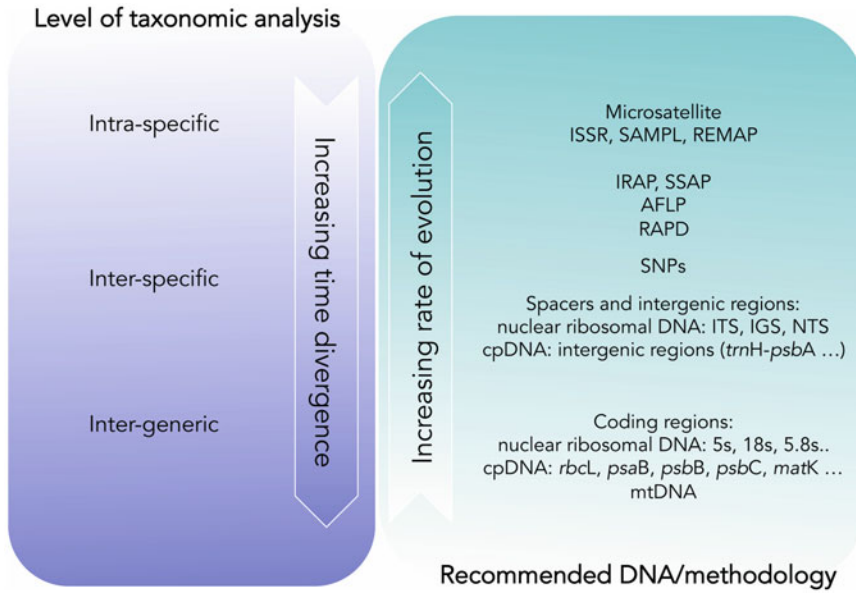


Fig. 10 General guidelines for the choice of markers to be used for plant taxonomy

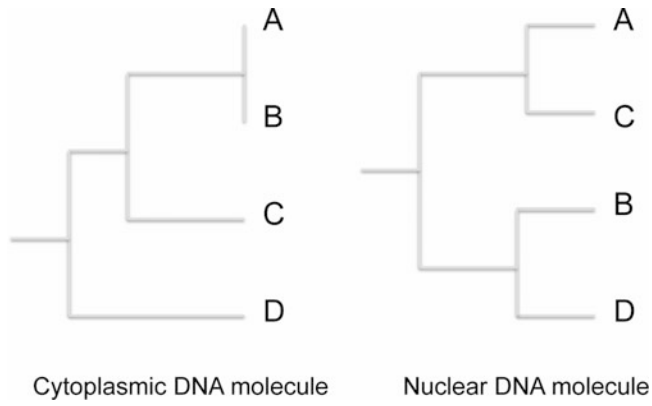


Fig. 11 A hypothetic phylogeny involving a hybrid species B whose maternal parent is species A

appear inconsistently on the trees generated with each type of marker (Fig. 11): a species B of hybrid origin will be grouped with its mother species A using cytoplasmic sequences, although it will appear different from it on the nuclear tree.

AFLP, RAPD, ISSR, and other multilocus profiling methods generate >90% dominant markers [56]. The polymorphism revealed is mainly due to mutations in the hybridization region of one of the primers, leading to either amplification of the locus (presence) or null allele (absence of amplification), i.e., a dominant system (Fig. 12). Consequently, such methods provide only biallelic markers.

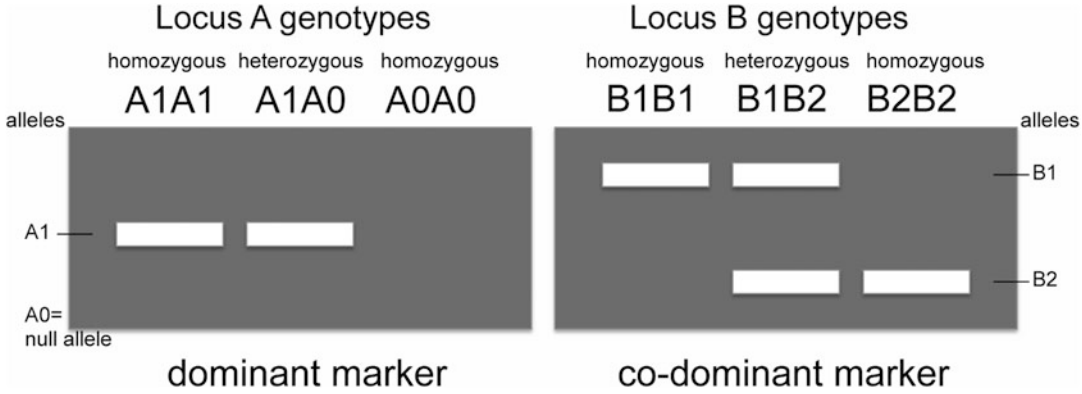


Fig. 12 Different genetic profiles: dominant versus codominant markers

On the other hand, microsatellite are very powerful monolocus markers as they are multiallelic and codominant (Fig. 12). They are indeed widely used in molecular ecology and population genetics studies as heterozygous loci can be clearly identified and allelic frequencies can be calculated to test for deviations from Hardy Weinberg equilibrium. One microsatellite multi-allelic marker provides as much genetic information as four to ten biallelic AFLP markers [57].

SNPs markers are monolocus, codominant, but are biallelic. Indeed, they evolve through the infinite sites (IAM) model: given the low rate of substitutions in genomes (the average synonymous substitution rate in plant nuclear genome is about $5.0–30.0 \times 10^{-9}$ per site per year [50]) the probability of more than one mutation at a given site is negligible, therefore each SNP is almost exclusively found only with two different states among the 4 possible (A, G, C, or T). For population genetic studies, it will be necessary to compensate the low allelic diversity of SNP markers by increasing the number of studied loci (2 to 6 times more SNP locus are needed as compared to microsatellites [58] to reach the same level of informativeness) particularly for populations with low levels of differentiation [59].

5 Analyzing Results

Most techniques presented (microsatellites, RAPD, AFLP, ISSR, IRAP, REMAP...) will generate fragment length data (different band sizes visualized and coded after electrophoretic separation). They are coded as 0/1 (absence/presence) or allelic size (when not dominant). Most of these sequences can be used to construct dendrograms using distance-based methods (generally using UPGMA or Neighbor Joining simple clustering analyses). Such procedures as applied for RAPD markers are detailed in Chapter 13.

On the other hand, sequence data can be analyzed either using distance-based methods or more powerfully using character-based cladistics methods (e.g., using maximum parsimony or maximum likelihood), allowing true phylogenetic trees to be constructed rather than phenetic trees. A detailed step-by-step phylogenetic data analysis protocol is available in the earlier version of this book [60]. Always remember that the tree built is a sequence tree, not a species tree. For all the reasons discussed above, using different sequences can lead to different trees reflecting the different evolutionary patterns of the sequences under study (Fig. 11).

In search for an integrative approach to draw the best taxonomic identification of species, it seems necessary to account not only for the level of differentiation using neutral molecular markers, but also to assess information related to nonneutral markers, which may represent markers under selection and provide information regarding fitness or adaptation of species [61–63]. Such approaches are conducted using population level studies, and can be performed using SNPs markers, using a new era of data analyses called population genomics (Chapter 16).

6 A Temporal Landscape of the most Commonly Used Methods for Molecular Plant Taxonomy

Interrogating NCBI database (as per 17/09/2019 for vascular plants) gives a good representation of the evolving landscape of techniques used over time for molecular plant taxonomy. As a whole it is clear that molecular techniques have increasingly been used since years 2000 for plant taxonomy. Searching for bibliographic references in *PubMed* for plant taxonomy papers using the various techniques described in the present book (Fig. 13) shows that microsatellite markers remain still largely and stably used in plant taxonomy and population genetic studies (with a mean contribution of 33% since 1999). On the other hand, random priming methods such as AFLP or RAPDs, are decreasing in use: they still account for 16% of the references found in 2018 while they reached 56% in 2010 and 94% in 1996. Retrotransposon-tagging techniques such as IRAP and REMAP have always remained more anecdotic with a maximum of 3% of the publications. In parallel, the increasing use of next-generation sequencing (NGS) techniques (SNP, GBS) is striking: they represented 12% of the papers published in 2010 and reached 56% in 2018.

Regarding sequences that are used to build molecular phylogenies in plants (Fig. 14), a search in NCBI-nucleotide database reveals that chloroplast DNA sequences and ITS nuclear sequences remain the most published, mitochondrial DNA sequences representing only 4% of the sequences produced in the 2015–2019 period.

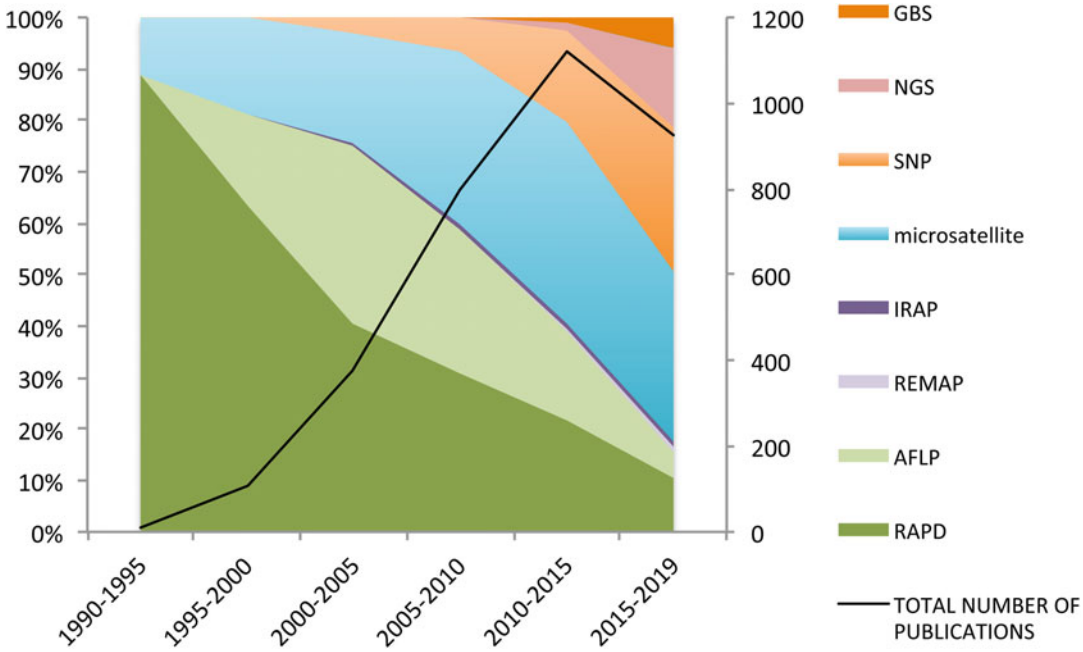


Fig. 13 Total number of publications on molecular plant taxonomy for the mentioned techniques per 5-year period (right axis) and percentage of each technique used (left axis), as referenced in NCBI-PubMed

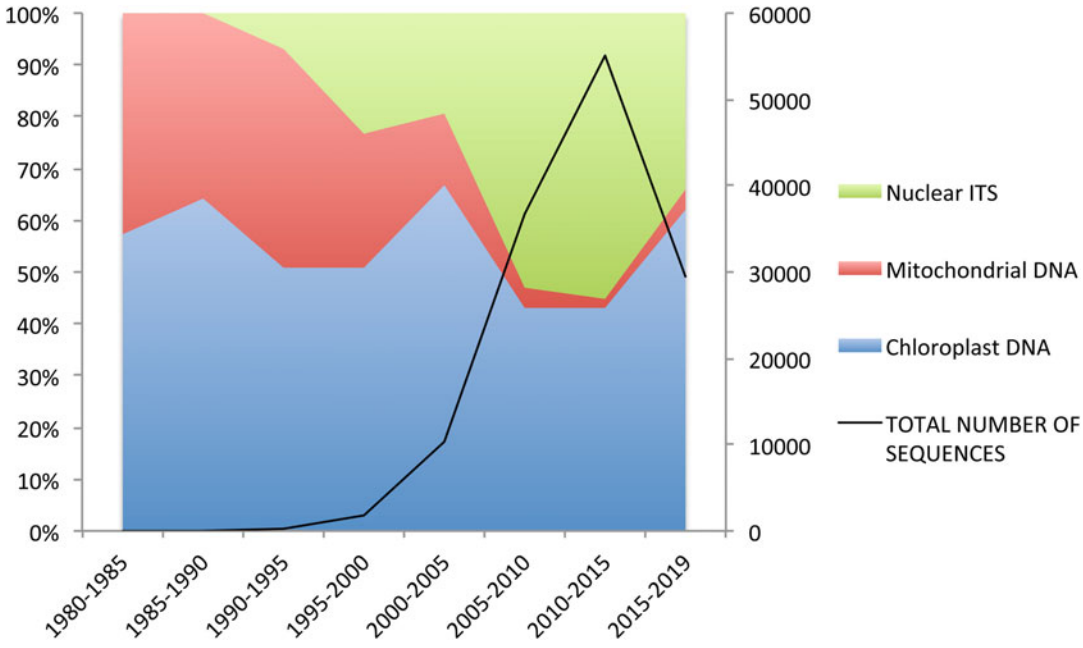


Fig. 14 Total number of sequences published per 5-year period (right axis) and percentage (left axis) of nuclear ITS, mitochondrial or chloroplast vascular plant DNA sequences deposited in NCBI-nucleotide database (excluding whole genome sequences)

7 Further Exploration: Chromosomal Organization

In plants, genome organization is very complex and polyploidy can be an important speciation mode [64]. It will be almost impossible to differentiate, for example, a diploid species from a related autopolyploid species using simply molecular markers. Molecular taxonomy can be greatly enhanced in some taxonomic complex plant groups by not only assessing phylogenetic relationships, but also genome organization to determine introgression, hybridization, or polyploidization (either by analyzing chromosomes or simply genome size) (Chapters 17–19).

References

1. Gregory TR (2001) Coincidence, coevolution, or causation? DNA content, cell size, and the C-value enigma. Biol Rev 76:65–101
2. Thomas CA (1971) The genetic organization of chromosomes. Annu Rev Genet 5:237–256
3. Flavell R, Bennett M, Smith J, Smith D (1974) Genome size and the proportion of repeated nucleotide sequence DNA in plants. Biochem Genet 12:257–269
4. Schmidt T, Heslop-Harrison JS (1998) Genomes, genes and junk: the large-scale organization of plant chromosomes. Trends Plant Sci 3:195–199
5. Hamby RK, Zimmer EA (1992) Ribosomal RNA as a phylogenetic tool in plant systematics. In: Soltis PS, Soltis DE, Doyle JJ (eds) Molecular systematics of plants. Chapman & Hall, New York
6. Schaal BA, Learn GH (1988) Ribosomal DNA variations between and among plant populations. Ann Mo Bot Gard 75:1207–1216
7. Hillis DM, Dixon MT (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. Q Rev Biol 66:411–453
8. Alvarez IA, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. Mol Phylogenet Evol 29:417–434
9. Poczai P, Hyvönen J (2010) Nuclear ribosomal spacer regions in plant phylogenetics: problems and prospects. Mol Biol Rep 37:1897–1912
10. Ellegren H (2004) Microsatellites: simple sequences with complex evolution. Nat Rev Genet 5:435–445
11. SanMiguel P, Bennetzen JL (1998) Evidence that a recent increase in maize genome size was caused by the massive amplification of intergene retrotransposons. Ann Bot 82:37–44
12. Ray DA (2007) SINEs of progress: Mobile element applications to molecular ecology. Mol Ecol 16:19–33
13. Deragon JM, Zhang X (2006) Short interspersed elements (SINEs) in plants: origin, classification, and use as phylogenetic markers. Syst Biol 55:949–956
14. Schmidt T (1999) LINEs, SINEs and repetitive DNA: non-LTR retrotransposons in plant genomes. Plant Mol Biol 40:903–910
15. Feliner GN, Rosselló JA (2007) Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. Mol Phylogenet Evol 44:911–919
16. Zimmer EA, Wen J (2013) Using nuclear gene data for plant phylogenetics: progress and prospects. Phylogenet Evol 66:539–550
17. Small RL, Cronn RC, Wendel JF (2004) Use of nuclear genes for phylogeny reconstruction in plants. Aust Syst Bot 17:145–170
18. Heinze B, Koziel-Monte A, Jahn D (2014) Analysis of variation in chloroplast DNA sequences. In: Molecular plant taxonomy. Springer, pp 85–120
19. Schlötterer C (2004) The evolution of molecular markers — just a matter of fashion? Nat Rev Genet 5:63–69
20. Thudi M, Li Y, Jackson SA, May GD, Varshney RK (2012) Current state-of-art of sequencing technologies for plant genomics research. Briefings Funct Genomics 11:3–11
21. Hudson ME (2008) Sequencing breakthroughs for genomic ecology and evolutionary biology. Mol Ecol Resour 8:3–17
22. Borevitz JO, Ecker JR (2004) Plant genomics: the third wave. Annu Rev Genomics Hum Genet 5:443–477

23. Grover CE, Salmon A, Wendel JF (2012) Targeted sequence capture as a powerful tool for evolutionary analysis1. Am J Bot 99:312–319
24. Timme RE, Bachvaroff TR, Delwiche CF (2012) Broad Phylogenomic sampling and the sister lineage of land plants. PLoS One 7:1–8
25. Exposito-Alonso M, Drost H, Burbano HA, Weigel D (2019) The earth BioGenome project: opportunities and challenges for plant genomics and conservation. Plant J 102
26. Syvänen A-C (2001) Accessing genetic variation: genotyping single nucleotide polymorphisms. Nat Rev Genet 2:930–942
27. Poland JA, Rife TW (2012) Genotyping-by-sequencing for plant breeding and genetics. Plant Genome 5:92–102
28. Jaccoud D, Peng K, Feinstein D, Killian A (2001) Diversity arrays: a solid state technology for sequence information independent genotyping. Nucleic Acids Res 29:1–7
29. James KE, Schneider H, Ansell SW, Evers M, Robba L, Uszynski G, Pedersen N, Newton AE, Russell SJ, Vogel JC, Kilian A (2008) Diversity arrays technology (DArT) for pan-genomic evolutionary studies of non-model organisms. PLoS One 3:1–11
30. Goulao LF, Oliveira CM (2014) Multilocus profiling with AFLP, ISSR, and SAMPL. In: Molecular plant taxonomy. Springer, pp 211–231
31. Duminil J (2014) Mitochondrial genome and plant taxonomy. In: Molecular plant taxonomy. Springer, pp 121–140
32. Zuckerkandl E, Pauling L (1965) Evolutionary divergence and convergence in proteins. In: Bryson V, Vogel HJ (eds) Evolving genes and proteins. Academic Press, New York, pp 97–166
33. Arbogast BS, Edwards SV, Wakeley J, Beerli P, Slowinski JB (2002) Estimating divergence times from molecular data on phylogenetic and population genetic timescales. Annu Rev Ecol Syst 33:707–740
34. Kimura M, Ohta T (1974) On some principles governing molecular evolution*(population genetics/mutational pressure/negative selection/random drift). Proc Natl Acad Sci U S A 71:2848–2852
35. Saliba-Colombani V, Causse M, Gervais L, Philouze J (2000) Efficiency of RFLP, RAPD, and AFLP markers for the construction of an intraspecific map of the tomato genome. Genome 43:29–40
36. Qi X, Stam P, Lindhout P (1998) Use of locus-specific AFLP markers to construct a high-density molecular map in barley. Theor Appl Genet 96:376–384
37. Saal B, Wricke G (2002) Clustering of amplified fragment length polymorphism markers in a linkage map of rye. Plant Breed 121:117–123
38. Young WP, Schuppet JM, Keim P (1999) DNA methylation and AFLP marker distribution in the soybean genome. Theor Appl Genet 99:785–792
39. Shriver MO, Jin L, Chakraborty R, Boerwinkle E (1993) VNTR allele frequency distributions under the stepwise mutation model: a computer simulation approach. Genetics 134:983–993
40. Bhargava A, Fuentes FF (2010) Mutational dynamics of microsatellites. Mol Biotechnol 44:250–266
41. Buschiazzo E, Gemmell NJ (2006) The rise, fall and renaissance of microsatellites in eukaryotic genomes. BioEssays 28:1040–1050
42. Jeffreys AJ, Wilson V, Thein SL (1985) Individual-specific fingerprints of human DNA. Nature 316:76–79
43. Jobling MA, Gill P (2004) Encoded evidence: DNA in forensic analysis. Nat Rev Genet 5:739–752
44. Dover G (1982) Molecular drive: a cohesive mode of species evolution. Nature 299:111–117
45. Dover G (1994) Concerted evolution, molecular drive and natural selection. Curr Biol 4:1165–1166
46. Plohl M, Luchetti A, Meštrović N, Mantovani B (2008) Satellite DNAs between selfishness and functionality: structure, genomics and evolution of tandem repeats in centromeric (hetero)chromatin. Gene 409:72–82
47. Nei M, Rooney AP (2005) Concerted and birth-and-death evolution of multigene families*. Annu Rev Genet 39:121–152
48. Ganley ARD, Kobayashi T (2007) Highly efficient concerted evolution in the ribosomal DNA repeats: Total rDNA repeat variation revealed by whole-genome shotgun sequence data. Genome Res 17:184–191
49. Hollingsworth ML, Andra Clark A, Forrest LL, Richardson J, Pennington RT, Long DG, Cowan R, Chase MW, Gaudeul M, Hollingsworth PM (2009) Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Mol Ecol Resour 9:439–457
50. Wolfe KH, Li W-H, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci 84:9054–9058

51. Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. Proc Natl Acad Sci 102:8369–8374
52. Fortune PM, Roulin A, Panaud O (2008) Horizontal transfer of transposable elements in plants. Commun Integr Biol 1:74–77
53. Capy P, Anxolabéhère D, Langin T (1994) The strange phylogenies of transposable elements: are horizontal transfers the only explanation? Trends Genet 10:7–12
54. Syvanen M (1994) Horizontal gene transfer: evidence and possible consequences. Annu Rev Genet 28:237–261
55. McCauley DE, Sundby AK, Bailey MF, Welch ME (2007) Inheritance of chloroplast DNA is not strictly maternal in Silene vulgaris (Caryophyllaceae): evidence from experimental crosses and natural populations. Am J Bot 94:1333–1337
56. Bensch S, Åkesson M (2005) Ten years of AFLP in ecology and evolution: why so few animals? Mol Ecol 14:2899–2914
57. Mariette S, Le Corre V, Austerlitz F, Kremer A (2002) Sampling within the genome for measuring within-population diversity: trade-offs between markers. Mol Ecol 11:1145–1156
58. Morin PA, Luikart G, Wayne RK (2004) SNPs in ecology, evolution and conservation. Trends Ecol Evol 19:208–216
59. Morin PA, Martien KK, Taylor BL (2009) Assessing statistical power of SNPs for population structure and conservation studies. Mol Ecol Resour 9:66–73
60. De Bruyn A, Martin DP, Lefeuvre P (2014) Phylogenetic reconstruction methods: an overview. In: Molecular plant taxonomy. Springer, pp 257–277
61. Allendorf FW, Hohenlohe PA, Luikart G (2010) Genomics and the future of conservation genetics. Nat Rev Genet 11:697–709
62. Ouborg NJ, Pertoldi C, Loeschcke V, Bijlsma RK, Hedrick PW (2010) Conservation genetics in transition to conservation genomics. Trends Genet 26:177–187
63. Ouborg N, Vergeer P, Mix C (2006) The rough edges of the conservation genetics paradigm for plants. J Ecol 94:1233–1248
64. Adams KL, Wendel JF (2005) Polyploidy and genome evolution in plants. Curr Opin Plant Biol 8:135–141

Chapter 3

Isolation and Purification of DNA from Complicated Biological Samples

Ruslan Kalendar, Svetlana Boronnikova, and Mervi Seppänen

Abstract

The isolation of nucleic acids from a biological sample is an important step for many molecular biology applications and medical diagnostic assays. This chapter describes an efficient protocol using established acidic CTAB (with a pH value of 5.0 to 6.8) based extraction method for isolation and/or purification of high molecular weight genomic DNA from a range of fresh and difficult sources from plant, animal, fungi, and soil material. This protocol is suitable for many sequencing and genotyping applications, including large-scale sample screening.

Key words High-quality DNA, DNA extraction, Plant tissues, Woody plants, Animal tissues, Herbarium specimens, Food, Soil, CTAB

1 Introduction

Nucleic acid sequences have a variety of applications in the field of molecular biology. They are a valuable tool in many analytical and application techniques used in the field of molecular biology, health, medicine (gene therapy, diagnostics, and recombinant protein expression), forensics, and food science. Some examples of these techniques include next-generation sequencing applications, genotyping with DNA fingerprinting, detection of pathogens, and forensic identification of biological samples and environmental samples contaminated with different biological entities [1–8].

To be used as a diagnostic tool, the target nucleic acid sequence should be free of contaminants that inhibit PCR and other downstream applications. Such contaminants chemically or mechanically block or inhibit chemical and enzymatic reactions, including denaturation and hybridization of nucleic acids, and other applications used in molecular biology methods. Contaminants can also degrade or modify the nucleic acid. These include high-molecular substances, such as polysaccharides and polyphenols, as well as substances of lower molecular weight, such as pigments, secondary

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_3,

metabolites, lipids, humic substances, low-molecular enzyme inhibitors, and oligonucleotides. Therefore, in order to be able to use nucleic acids from biological materials for further analysis, it is important that these substances are eliminated entirely from the sample.

Isolating DNA or RNA that is sufficiently purified from contaminants is complicated by the diversity and complex composition of biological material from which DNA and RNA are isolated. Biological material consists of cells and tissues. Cells in liquid media, such as blood, lymph, milk, urine, and feces, and cells in culture, on an agarose or polyacrylamide gel, in soil, or in solution, usually include significant amounts of contaminants that must be removed from the DNA or RNA before molecular biology experiments. The presence of chemical or mechanical crosslinks between DNA chains and with contaminants interweaving with DNA leads to partial or complete inhibition of DNA denaturation and the appearance of artifacts. The quality of nucleic acids directly influences problems and artifacts produced by molecular biology procedures downstream. Thus, for efficient DNA amplification, for example, using the PCR method or isothermal DNA amplification, complete separation of nucleic acid strands at all lengths is required.

A variety of DNA extraction and purification methods have been developed [9–23], and are known for different characteristics. Ionic ion exchange resins were used to purify a nucleic acid already in 1953 [24]. Nucleic acids, proteins, and other contaminants are bound on a solid support by anion exchange. Nucleic acids are then eluted in a high salt concentration (7 M urea or 1.2 M NaCl) [18] and further purified by ethanol precipitation. Ultracentrifugation in a gradient of sucrose or cesium salts has also been used to purify DNA. Nucleic acids are separated from other macromolecules in accordance with their sedimentation coefficient, before extraction with phenol or phenol/chloroform and precipitation with ethanol or isopropanol. Conventional protocols for the extraction of DNA or RNA from cells are well known in the field, and described in *Molecular Cloning*, Sambrook et al. [25]. For DNA, these protocols typically include a cell lysis step, solubilization of DNA, enzymatic or chemical extraction, and separation of DNA from impurities such as proteins, RNA, and other substances [26].

A wide spectrum of methods has been developed for the purification of nucleic acids by filtration on a microporous carrier [27]. The microporous membrane as a matrix for binding and support for DNA purification has many advantages, such as compactness and ease of development. It allows differential control of the elution of desired molecules and the removal of undesirable components in the liquid phase, in parallel for a larger number of samples in a shorter period of time compared to other approaches. There are several methods based on the binding of nucleic acids on a sorbent and then washing unwanted impurities, followed by

elution of nucleic acids. Silicon dioxide particles (SiO_2, silica, sand with particle size of 10–50 μm), fiberglass (glass microfiber filters Grade GF/A), microballs, hydroxyapatite, anion exchange resins, and diatomite are used as sorbents. Nucleic acids bind reversibly on particles of synthetic silica gel in buffers containing high concentrations of chaotropic salts (sodium iodide, sodium perchlorate, or guanidine thiocyanate) or urea. Unbound cell components are then washed out, after which the pure nucleic acid is eluted from the sorbent with an aqueous solution with low ionic strength [20, 28–30]. Several companies offer DNA and RNA purification kits based on this approach. The kits contain columns with membranes of sorbents based on silicon dioxide and microporous glass. Centrifugation or vacuum filtration is used to bind nucleic acids with the sorbent, followed by washing and elution. The use of glass microfiber filters as a sorbent for purifying nucleic acids does not always result in sufficiently pure DNA for subsequent use in molecular biology protocols [12, 13, 20, 23, 24, 31, 32].

Probably the most promising approach for the isolation and purification of nucleic acids is the use of electroelution techniques [33–36]. The electroelution procedure allows the purification of very clean DNA for use in all molecular biology applications. It effectively separates DNA from compounds, including high-molecular substances such as polysaccharides and polyphenols, as well as from pigments and humic substances that interfere with subsequent DNA quantification and amplification. For example, the SageELF electrophoresis system (Sage Science, Inc. USA) is commercially used to separate DNA or protein samples by size and then fractionate the whole sample. The system is equipped with pulsed-field electrophoresis for resolving large DNA.

We have developed a protocol for the isolation of DNA from biological samples using a lysis buffer (pH < 7, but preferably less than 6) containing acidic organic and inorganic salts of sodium or potassium (acetate, propionate, formate, citrate) or weak acids of zwitterionic buffering agents: HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid), MOPS (3-(N-morpholino)propanesulfonic acid), or MES (2-morpholin-4-ylethanesulfonic acid) [37].

The combination of acid lysis buffer and subsequent extraction with chloroform allows highly selective separation of polysaccharides, pigments, proteins, and other cellular components in the organic phase, which possesses the original color of the pigment, while in the aqueous phase the remaining pure DNA and RNA acquires a completely transparent color (*see* **Note 1**).

During DNA or RNA extraction using a lysing buffer with a weakly acidic pH (range from 5.0 to 6.8), oxidation processes and enzymatic and chemical reactions are almost completely blocked. As a result, covalent bonds are not formed between DNA and phenolics or polysaccharide components. During subsequent extraction with chloroform, contaminants that inhibit PCR

(polysaccharides, polyphenols, peptides, lipids, and pigments) are selectively removed to the interphase and an organic phase. The aqueous phase containing the DNA is collected and mixed with an equal volume of simple alcohol to precipitate the DNA. Finally, the DNA is purified by precipitation or filtration through a column with a glass/cellulose microfiber filter (*see* **Note 2**).

This protocol for DNA isolation is universal for most biological specimens. The method will effectively isolate DNA from whole blood, bones, plant samples, soil, herbarium, mycelium of fungi, and tissues rich in secondary metabolites, polysaccharides, and pigments. DNA samples obtained using the proposed method can be used in studies where the presence of contaminants in nucleic acids is undesirable; for example, during cloning, sequencing, and genotyping (*see* **Note 3**).

2 Materials

Prepare all solutions using ultrapure Milli-Q water and analytical grade reagents. Prepare and store all reagents at room temperature unless otherwise specified and away from direct sunlight. Diligently follow all waste disposal regulations when disposing waste materials.

2.1 DNA Extraction

1. Glass beads 6 mm or tungsten carbide beads 3 mm.
2. TissueLyser II bead mill or similar Mixer Mill system, the adapter set 2 × 24 or set 2 × 96 (QIAGEN).
3. NanoDrop™ 2000/2000c Spectrophotometers or similar equipment for RNA (or DNA) concentration measures.
4. Chloroform:isoamyl alcohol mix (24:1).
5. 100% Isopropanol (2-propanol).
6. 70% Ethanol.
7. 10 mM Tris–HCL pH 8.0.
8. 0.5 M Na_3EDTA.
9. Ribonuclease A solution: 10 mg/mL in 50% glycerol, 10 mM Tris–HCL pH 8.0.
10. TE buffer: 1 mM Na_3EDTA, 10 mM Tris–HCl adjusted to pH 8.0.
11. CTAB DNA extraction buffer: 2% cetyltrimethylammonium bromide (CTAB), 1.5 M NaCl, 10 mM Na_3EDTA, 100 mM 4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES-acid) or 3-(N-morpholino)propanesulfonic acid (MOPS-acid). Combine 20 g CTAB, 25 g HEPES-acid (or 21 g MOPS-acid) and 20 ml of 0.5 M Na_3EDTA dissolved in 500 ml of Milli-Q water; then add 300 ml 5 M NaCl and bring final volume to 1 L.

2.2 Gel Electrophoresis

1. Electrophoresis Tris-Acetate-EDTA buffer (1× TAE): 40 mM Tris-base, 20 mM Acetic Acid, 1 mM EDTA (pH 8.0).
2. Gel loading buffer (10×): 20% (w/w) Polysucrose 400, 100 mM Tris–HCl pH 8.0, 10 mM EDTA, ~0.01% bromophenol blue. Dissolve 20 g Polysucrose 400 (Ficoll 400) in 80 mL 10× TE buffer. Add bromophenol blue according to the desired color intensity. Store at +4 °C.
3. DNA Ladder for electrophoresis: 100–10,000 base range. The DNA Ladder is diluted with 1× gel loading buffer to final concentration 25 ng/μL.
4. Agarose Basic for DNA Electrophoresis.

2.3 Equipment

1. Power supply (minimum 300 V, 400 mA) for electrophoresis.
2. Horizontal electrophoresis apparatus without special cooling. Most commercially available medium- or large-scale horizontal DNA gel electrophoresis systems are suitable. We routinely employ an apparatus with a run length of 10 cm.
3. UV transilluminator, for visualization of Ethidium Bromide-stained or SYBR Green-stained nucleic acids, with a viewing area of 20 × 20 cm.
4. Imaging system.
5. Spectrophotometer.

3 Methods

All lab procedures are performed at room temperature. This protocol was tested with different samples from herbarium specimens, seeds including plant samples containing significant amounts of contaminants and polysaccharides (*Medicago sativa, Vicia faba, Lupinus angustitolius, Colocasia esculenta*), as well as with woody plants, soil samples, and animal tissue, like blood.

3.1 DNA Extraction Protocol

3.1.1 Tissue Grinding

1. This step can be performed using either a TissueLyser II or a mortar and pestle. The TissueLyser II option is preferred because less time is required, more samples can be extracted, and cross-contamination is minimized.
2. Collect the tissue sample (the sample mass should not exceed 50–100 mg) in 2 mL Eppendorf Safe-Lock microcentrifuge tube containing a glass ball. Place the samples (plant leaves) and the TissueLyser II adapters in an ultralow freezer and store the frozen tissue at −80 °C. Cooling is not required for dry samples (herbarium specimen).
3. Powder the tissue by shaking in the presence of the steel (glass) balls at 30 Hz for 2–10 min. Proper grinding of plant samples

with a TissueLyser II is a crucial step and the plant tissue should be ground to a fine powder after the disruption. However, for some plants one disruption step may not be enough. In these cases, repeat the disruption for 5 min at 30 Hz until the sample is thoroughly and equally homogenized.

3.1.2 Extraction of DNA from Ground Tissue

1. Add 1 mL of preheated CTAB DNA extraction buffer with 1 μL Ribonuclease A solution to the tissue powder and mix in the TissueLyser II for 1 min at 30 Hz.
2. Incubate at 65 °C for 1 h (long incubation increases DNA yield).
3. Centrifuge at maximum speed in a microcentrifuge for 2 min to remove nonsoluble debris.
4. Transfer the entire clarified supernatant to a new 2 mL microcentrifuge tube containing an equal volume of chloroform.
5. Mix well for 5 min in the TissueLyser II at 30 Hz.
6. Centrifuge at maximum speed in a microcentrifuge for 2 min.
7. Transfer the entire clarified upper aqueous layer to a new 2 mL microcentrifuge tube which contains an equal or half the volume of 2-propanol, and vortex thoroughly.
8. Centrifuge at maximum speed in a microcentrifuge for 2–5 min. A whitish DNA pellet should be visible.
9. Discard supernatant and wash the pellet by adding 1.8 mL 70% ethanol; vortex thoroughly. At this stage, DNA samples can be stored at room temperature or refrigerated.
10. Centrifuge at maximum speed for 2–5 min and carefully discard the supernatant by decanting or with a micropipette. A whitish DNA pellet should be visible during discarding of supernatant.
11. Ensure the DNA pellet does not dry and dissolve immediately in 300 μL TE buffer, pH 8.0 at 55 °C for 10–20 min.

3.2 DNA Analysis

The spectrophotometric method of DNA quantitation is commonly used to determine both concentration and relative purity of nucleic acids in a solution. A spectrophotometer is used to measure the absorbance and purity of DNA samples. Pure DNA exhibited an A_{260}/A_{230} ratio in the range of 1.8–2.0 and is acceptable down to ratios of about 1.5. Smaller values around or even below 1.0 indicate significant amounts of impurities, contamination with polysaccharides.

The rapid agarose gel electrophoresis method provides a much more accurate quantitation of the genomics DNA. The integrity of the genomic DNA samples extracted is analyzed by electrophoresis on an agarose gel. Atleast 4 samples loaded onto a gel, at least one lane should contain a series of DNA fragments of known sizes so

that a standard curve can be constructed to allow the calculation of the size of unknown DNA fragments. The most commonly used molecular weight markers are calf thymus DNA or DNA Ladder. DNA Ladder usually cover a wide range of DNA sizes.

1. Mix 10 μL of the DNA solution prepared in the previous section with 2 μL of gel loading buffer (10×) in tube or plate, a quick spin with centrifugation at 14,000 RPM (16,873 × *g*) for a few seconds.
2. Prepare 1% agarose gel in 1× TAE electrophoresis buffer containing ethidium bromide. The agarose gel must be completely melted in the microwave and then allowed to slowly cool until its temperature drops to about 50–60 °C. At that point, if desired, add the ethidium bromide solution at a rate of 20 μL per 100 mL, to bring the final concentration to 0.5 μg per mL.
3. Load the sample into one of the wells. In the adjacent wells, load equal volumes of a series of DNA concentration standards (e.g., calf thymus in the range of 25–500 ng/μL) or DNA Ladder.
4. Run the gel at 50 V when the bromophenol blue tracking dye has migrated at least 2 cm from the wells, the run can be stopped.
5. Examine the gel on an ultraviolet light transilluminator. Intact DNA will be visible as a band near wells. A smear extending from the well to the dye front indicates that the DNA has been fragmented. The images can be saved in a Gel Documentation System (*see* **Note 4**).
6. From the gel photo, estimate the quantity of DNA in the test samples by comparison to the DNA concentration standards. The yield should be in the range 5–15 μg of DNA per 300 μL, with an average size of above 50 kb.

4 Notes

1. The original techniques using CTAB for DNA isolation was first developed by Murray and Thompson in 1980 [13]. The original protocol described by the authors contains an alkaline pH on Tris buffer. However, under alkaline conditions DNA extraction takes place with oxidative processes, causing a change in the color of the lysis solution from green to brown (for plants). Therefore, water-soluble polymer polyvinylpyrrolidone (PVP) and reducing agent β-mercaptoethanol were added to the lysis buffer. The use of CTAB, a cationic detergent, facilitates the separation of polysaccharides during purification, while additives such as PVP can help remove polyphenols. Buffers based on CTAB are also used to purify

DNA from plant tissues and their metabolites. Polyphenols are compounds that contain more than one phenolic ring (e.g., tannin), a structure that binds very effectively to DNA. They occur naturally in plants, but they also form when tissue is damaged (roasting). When plant tissues are homogenized, polyphenols are synthesized by the polyphenol oxidase released. The addition of PVP prevents polyphenols from binding to DNA and phenolic rings. The presence of chemical crosslinks between the chains and impurities from the tissue, or mechanical spatial entanglements of DNA in the presence of polysaccharides, leads to partial or complete inhibition of DNA denaturation and the appearance of artifacts. When DNA is isolated, certain groups of polysaccharides form a viscous, jelly-like, uniform mass with the DNA. Serious and damaging effects are exerted by oxidants of different biochemical nature, including phenolic compounds.

2. The implementation of our protocol of isolation and purification of total DNA from a biological sample is achieved as follows. Preliminary steps are homogenization of tissue sample to complete destruction within a few seconds to minutes. Both dry and liquid samples can be used. For blood samples, start with a red blood cell lysis step and precipitation of leukocytes. Lysis of the samples is carried out using weakly acidic (pH 5.0–6.8) DNA extraction buffers containing acidic zwitterionic agents (MOPS or HEPES) during incubation at 55–65 °C. Organic extraction with chloroform results in contaminants selectively separated into interphase and an organic phase. The aqueous phase containing the DNA is collected and mixed with an equal volume of simple alcohol to precipitate the DNA. As a result, an aqueous solution containing DNA becomes completely transparent, while the organic phase possesses the original color of the pigment (or brown for hemoglobin). In some cases, organic extraction with chloroform is not possible and DNA must be precipitated immediately or purified on a column.
3. The composition of the lysis solution contains inorganic salts (sodium chloride), within the effective concentration in the range of 1–4 M. The optimal concentration of the detergent is 1.5% CTAB. To increase the efficiency of DNA extraction, proteinase K can be added to the acidic lysing solution, which retains proteolytic activity at high ionic strength and low pH values, even in the presence of strong detergents and chaotropic agents. The subsequent extraction with chloroform increases the purity of the isolated DNA, especially from complex samples (thermally treated raw materials, blood, herbarium specimen, and soils). The effective concentration of chloroform is 1–2 volumes of the total lysate. Further, the

DNA is precipitated from the aqueous phase with a water-soluble organic solvent, such as a simple alcohol (it is preferable to use isopropanol). Depending on the biological material, DNA can be precipitated by filtration through a column with a glass microfiber filter, for example, glass microfiber filters (Grade GF/A) or through cellulose paper [26]. Finally, DNA is washed by precipitation or filtration in a solution of 80% ethanol and dissolved in low ionic buffered water.

4. The most frequent cause of bad DNA resolution is improper choice of agarose concentration.

   Low percentage agarose gels should be used to resolve high-molecular-weight DNA fragments and high percentage gels for low-molecular-weight DNAs. Trailing and smearing of DNA bands are most frequently observed with high-molecular-weight DNA fragments. This is often caused by overloading the DNA sample or running gels at high voltages.

Acknowledgments

The work was partially supported by the Government of Perm Krai, research project No C-26/174.3 on 31.01.2019.

References

1. Welsh J, McClelland M (1990) Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res 18(24):7213–7218. https://doi.org/10.1093/nar/18.24.7213
2. Kalendar R, Schulman AH (2014) Transposon-based tagging: IRAP, REMAP, and iPBS. Methods Mol Biol 1115:233–255. https://doi.org/10.1007/978-1-62703-767-9_12
3. Kalendar R, Amenov A, Daniyarov A (2019) Use of retrotransposon-derived genetic markers to analyse genomic variability in plants. Funct Plant Biol 46(1):15–29. https://doi.org/10.1071/fp18098
4. Arcot SS, Wang Z, Weber JL, Deininger PL, Batzer MA (1995) Alu repeats: a source for the genesis of primate microsatellites. Genomics 29(1):136–144. https://doi.org/10.1006/geno.1995.1224
5. Drabek J, Smolikova M, Kalendar R, Pinto FAL, Pavlousek P, KleparnIk K, Frebort I (2016) Design and validation of an STR hexaplex assay for DNA profiling of grapevine cultivars. Electrophoresis 37(23–24):3059–3067. https://doi.org/10.1002/elps.201600068
6. Kalendar R, Antonius K, Smykal P, Schulman AH (2010) iPBS: a universal method for DNA fingerprinting and retrotransposon isolation. Theor Appl Genet 121(8):1419–1430. https://doi.org/10.1007/s00122-010-1398-2
7. Kalendar R, Shustov AV, Seppänen MM, Schulman AH, Stoddard FL (2019) Palindromic sequence-targeted (PST) PCR: a rapid and efficient method for high-throughput gene characterization and genome walking. Sci Rep 9(1):17707. https://doi.org/10.1038/s41598-019-54168-0
8. Ghonaim M, Kalendar R, Barakat H, Elsherif N, Ashry N, Schulman AH (2020) High-throughput retrotransposon-based genetic diversity of maize germplasm assessment and analysis. Mol Biol Rep 47(3):1589–1603. https://doi.org/10.1007/s11033-020-05246-4
9. Green MR, Sambrook J (2012) Molecular cloning: a laboratory manual, vol 1. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY. https://books.google.com/books?id=DgqZtgAACAAJ
10. Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K (2003) Short protocols in molecular biology. Wiley, New York
11. Ausubel FM, Brent R, Kingston RE, Moore DD, Seidman JG, Smith JA, Struhl K (eds)

(1988) Current protocols in molecular biology. Wiley, New York. https://www.wiley.com/en-us/Current+Protocols+in+Molecular+Biology+-p-9780471503385

12. Allen GC, Flores-Vergara MA, Krasynanski S, Kumar S, Thompson WF (2006) A modified protocol for rapid DNA isolation from plant tissues using cetyltrimethylammonium bromide. Nat Protoc 1(5):2320–2325. https://doi.org/10.1038/nprot.2006.384
13. Murray MG, Thompson WF (1980) Rapid isolation of high molecular weight plant DNA. Nucleic Acids Res 8(19):4321–4325. https://doi.org/10.1093/nar/8.19.4321
14. Pereira JC, Chaves R, Bastos E, Leitao A, Guedes-Pinto H (2011) An efficient method for genomic DNA extraction from different molluscs species. Int J Mol Sci 12(11): 8086–8095. https://doi.org/10.3390/ijms12118086
15. Lade BD, Patil AS, Paikrao HM (2014) Efficient genomic DNA extraction protocol from medicinal rich Passiflora foetida containing high level of polysaccharide and polyphenol. Springerplus 3:457. https://doi.org/10.1186/2193-1801-3-457
16. Inglis PW, Pappas MCR, Resende LV, Grattapaglia D (2018) Fast and inexpensive protocols for consistent extraction of high quality DNA and RNA from challenging plant and fungal samples for high-throughput SNP genotyping and sequencing applications. PLoS One 13(10):e0206085. https://doi.org/10.1371/journal.pone.0206085
17. Anderson CB, Franzmayr BK, Hong SW, Larking AC, van Stijn TC, Tan R, Moraga R, Faville MJ, Griffiths AG (2018) Protocol: a versatile, inexpensive, high-throughput plant genomic DNA extraction method suitable for genotyping-by-sequencing. Plant Methods 14:75. https://doi.org/10.1186/s13007-018-0336-1
18. Blin N, Stafford DW (1976) A general method for isolation of high molecular weight DNA from eukaryotes. Nucleic Acids Res 3(9):2303–2308. https://doi.org/10.1093/nar/3.9.2303
19. Doyle J, Doyle J (1990) Isolation of plant DNA from fresh tissue. Focus 12:13–15. http://ci.nii.ac.jp/naid/10003365693/en/
20. Ndunguru J, Taylor NJ, Yadav J, Aly H, Legg JP, Aveling T, Thompson G, Fauquet CM (2005) Application of FTA technology for sampling, recovery and molecular characterization of viral pathogens and virus-derived transgenes from plant tissues. Virol J 2:45. https://doi.org/10.1186/1743-422X-2-45
21. Couch JA, Fritz PJ (1990) Isolation of DNA from plants high in polyphenolics. Plant Mol Biol Report 8(1):8–12. https://doi.org/10.1007/bf02668875
22. Springer NM (2010) Isolation of plant DNA for PCR and genotyping using organic extraction and CTAB. Cold Spring Harb Protoc 2010(11):pdb prot5515. https://doi.org/10.1101/pdb.prot5515
23. Kato H, Caceres AG, Mimori T, Ishimaru Y, Sayed AS, Fujita M, Iwata H, Uezato H, Velez LN, Gomez EA, Hashiguchi Y (2010) Use of FTA cards for direct sampling of patients' lesions in the ecological study of cutaneous leishmaniasis. J Clin Microbiol 48 (10):3661–3665. https://doi.org/10.1128/JCM.00498-10
24. Jones A (1953) The isolation of bacterial nucleic acids using cetyltrimethylammonium bromide (cetavlon). Biochim Biophys Acta 10(4):607–612. https://doi.org/10.1016/0006-3002(53)90304-7
25. Sambrook J, Fritsch EF, Maniatis T (1989) Molecular cloning: a laboratory manual, 2nd edn. Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY
26. Zou Y, Mason MG, Wang Y, Wee E, Turni C, Blackall PJ, Trau M, Botella JR (2017) Nucleic acid purification from plants, animals and microbes in under 30 seconds. PLoS Biol 15(11):e2003916. https://doi.org/10.1371/journal.pbio.2003916
27. Tolosa JM, Schjenken JE, Civiti TD, Clifton VL, Smith R (2007) Column-based method to simultaneously extract DNA, RNA, and proteins from the same sample. BioTechniques 43(6):799–804. https://doi.org/10.2144/000112594
28. Vogelstein B, Gillespie D (1979) Preparative and analytical purification of DNA from agarose. Proc Natl Acad Sci 76(2):615–619. https://doi.org/10.1073/pnas.76.2.615
29. Thompson JD, Cuddy KK, Haines DS, Gillepsie D (1990) Extraction of cellular DNA from crude cell lysate with glass. Nucleic Acids Res 18(4):1074. https://doi.org/10.1093/nar/18.4.1074
30. Hoss M, Paabo S (1993) DNA extraction from Pleistocene bones by a silica-based purification method. Nucleic Acids Res 21(16): 3913–3914. https://doi.org/10.1093/nar/21.16.3913
31. Drabkova LZ (2014) DNA extraction from herbarium specimens. Methods Mol Biol 1115:69–84. https://doi.org/10.1007/978-1-62703-767-9_4
32. Bost DA, Greenfield L (2014) Compositions and methods of selective nucleic acid isolation. USA patent US7537898B2

33. Zarzosa-Alvarez AL, Sandoval-Cabrera A, Torres-Huerta AL, Bermudez-Cruz RM (2010) Electroeluting DNA fragments. J Vis Exp (43):2136. https://doi.org/10.3791/2136
34. Stellwagen NC (2009) Electrophoresis of DNA in agarose gels, polyacrylamide gels and in free solution. Electrophoresis 30(Suppl 1): S188–S195. https://doi.org/10.1002/elps.200900052
35. Kallmeyer J, Smith DC (2009) An improved electroelution method for separation of DNA from humic substances in marine sediment DNA extracts. FEMS Microbiol Ecol 69(1): 125–131. https://doi.org/10.1111/j.1574-6941.2009.00684.x
36. Strychalski EA, Konek C, Butts EL, Vallone PM, Henry AC, Ross D (2013) DNA purification from crude samples for human identification using gradient elution isotachophoresis. Electrophoresis 34(17):2522–2530. https://doi.org/10.1002/elps.201300133
37. Kalendar R (2019) Universal DNA isolation protocol. protocolsio, Berkeley, CA. https://doi.org/10.17504/protocols.io.z2jf8cn

Chapter 4

Herbarium Specimens: A Treasure for DNA Extraction, an Update

Lenka Záveská Drábková

Abstract

With the expansion of molecular techniques, the historical collections have become widely used. The last boom started with using next- and second-generation sequencing in which massive parallel sequencing replaced targeted sequencing and third-generation technology involves single molecule technology. Studying plant DNA using these modern molecular techniques plays an important role in understanding evolutionary relationships, identification through DNA barcoding, conservation status, and many other aspects of plant biology. Enormous herbarium collections are an important source of material especially for taxonomic long-standing issues, specimens from areas difficult to access or from taxa that are now extinct. The ability to utilize these specimens greatly enhances the research. However, the process of extracting DNA from herbarium specimens is often fraught with difficulty related to such variables as plant chemistry, drying method of the specimen, and chemical treatment of the specimen. The result of these applications is often fragmented DNA. The reason new sequencing approaches have been so successful is that the template DNA needs to be fragmented for proper library building, and herbarium DNA is exactly that. Although many methods have been developed for extraction of DNA from herbarium specimens, the most frequently used are modified CTAB and DNeasy Plant Mini Kit protocols. Nine selected protocols in this chapter have been successfully used for high-quality DNA extraction from different kinds of plant herbarium tissues. These methods differ primarily with respect to their requirements for input material (from algae to vascular plants), type of the plant tissue (leaves with incrustations, sclerenchyma strands, mucilaginous tissues, needles, seeds), and further possible applications (PCR-based methods, microsatellites, AFLP or next-generation sequencing).

Key words AFLP, DNA extraction, Difficult plant tissues, Herbarium specimens, Microsatellites, Next-generation sequencing, PCR

1 Introduction

Hundreds of protocols for DNA preparation from various types of tissues have been published over the last few decades. Plant and mainly herbarium plant samples DNA extraction frequently present a challenge in the first stage of each study, because of the extraction of any given taxon may require time-consuming optimization of the extraction protocols. The problem of DNA extraction is crucial

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_4,

for further analyses of herbarium samples. The satisfactory quality of DNA is essential for the success of the whole molecular study. Most future molecular taxonomic studies will probably be partly or entirely based on DNA extracts from herbarium specimens because of the easy accessibility and richness of herbarium collections. Nowadays, the term "museomics" firstly used by zoologists [1] is frequently used for studies using large-scale analyses of DNA from herbarium samples. The last boom started with using next- and second-generation sequencing in which massive parallel sequencing replaced targeted sequencing and third-generation technology involves single molecule technology. This chapter, will use the term "next-generation" sequencing for all these kinds of sequencing technologies.

However, DNA isolation from dried specimens usually requires some modifications to frequently used protocols [2] because of the small amount of dry herbarium tissue available. The herbarium material is dried and stored on herbarium sheets in packages. If the specimens are air-dried at up to 42 °C [3], they contain a useful amount of high-molecular-weight DNA. Air-drying is considered to be better than the preservation of tissues in silica gel or anhydrous $CaSO_4$ [4]. In general, old air-dried material that has not been treated with chemical preservatives, high temperatures, or microwaves has the best chance of yielding useful DNA [3]. To preserve DNA well, it is necessary to dry plants as fast as possible. Extraction results depend on how the plant material is prepared, how many times the collection is disinfected, and the type of chemicals or procedures used. For instance, DNA was seriously degraded in leaves that were microwaved [4–6], boiled in water, or immersed in chemical solutions. Another important factor is the regular herbarium treatment used to keep specimens free of pests. Fumigation methods have been changed from time to time [7], making it difficult to be sure about the DNA quality. These kinds of post-mortem DNA damage can come in different ways as described [8]: (1) double-stranded damage, usually resulting from loss of A and G bases (depurination), and (2) breaks in the sugar–phosphate backbone of the DNA molecule, causing reduction of PCR-amplifiable template DNA. Herbarium DNA is usually degraded into small fragments with low molecular weight probably as a direct result of heat-treatment of the specimens [9–11]. As summarized [12], single-stranded damage on can lead to the generation of erroneous sequence information or so-called miscoding lesions. Damaged nucleotides in herbarium DNA, caused by oxidative stress and (or) heating, may include a-puric sites (loss of A and G bases), de-aminated cytosine residues resulting in uracil, or oxidized guanine residues, as found in studies in vivo and on ancient DNA [13, 14]. PCR amplification of these sites may result in damage-specific nucleotide mis-incorporations [15–17]. This type of damage is therefore in principle polymerase-bypassable, which

led to incorrect bases in the inferred sequence [12]. The extent of DNA degradation in herbarium specimens appears to be related to the condition of the fresh leaf rather than the year in which it was dried [18]. The DNA from herbarium specimens have been satisfactorily obtained from vascular plants about 200 years old and 100 years old from lichens.

Obtaining high-quality DNA depends on the extraction technique used. Traditional plant DNA extraction protocol by Doyle and Doyle [19] was used 18,118 times to date (e.g., cited in Google Scholar, 11.8.2020). This method is widely used for herbarium material, but sometimes with modifications [20, 21]. Several DNA isolation techniques that are useful for dry plant tissue from herbarium specimens have been described [2, 3, 22, 23]. Herbarium specimens have been frequently used during the last decade [20, 24–32]. The most convenient organ to sample is the leaf. However, also the seeds and pollen are efficient and inexpensive sources for DNA [33].

There are many different protocols more or less satisfactorily used for DNA extraction from herbarium samples of different group of plants with specific types of tissues. It is not possible to present all of them in this book. I do not give detailed literature evidence for many different protocols used but a simple comparison of articles published during the last 10 years clearly shows the most frequently used methods were the modified CTAB method [19] and DNeasy Plant Mini Kit (Qiagen). I selected a few of the most used protocols during the last 10 years. This chapter should serve as a tool for projects involving DNA extraction from herbarium specimens of different plants.

1.1 Main Isolation Difficulties

One problem with extraction from herbarium specimens is a very low yield of plant material. Many people work with plant taxa that are rare or grown in inaccessible locations, making it difficult to obtain fresh plant material. The use of dried plants from historical collections becomes essential for representative taxonomic sampling. Another problem is the quantity of the suitable tissue available. Many plants have a very limited leaf tissue volume, and the sampling for a nonproblematic extraction (yielding a sufficient amount of DNA) would cause serious damage to the herbarium specimen. Nowadays, this problem is overcome by third-generation sequencing, which involves single molecule technology (Oxford Nanopore technology).

Undoubtedly, especially good homogenization is essential. Another crucial point is a longer and repeated precipitation. Many protocols for DNA extraction use liquid nitrogen for grinding the plant material. The homogenization of plant material is easier and faster, but the simultaneous processing of multiple samples in mortars in one laboratory table leads to the loss of DNA and

contamination of the samples. When PCR products are analyzed by sequencing, the contamination is revealed. Other techniques, such as RFLP and RAPD, do not detect this type of mistake.

A good alternative is the use of bead-mills. These cylinders disrupt the plant tissues in microcentrifuge tubes in the mixer mill (or Tissue Lyser) without risk of contamination. Insufficient disruption of starting material leads to low yield and comprised purity. Pulverizing plant material with a mixer mill is easier and produces DNA of more reliable quality than grinding with liquid nitrogen in a mortar [34].

2 Materials

Plants have cell walls mostly comprised of cellulose and some other complex polysaccharide or another chemical compounds or have tissues with mucilaginous substances. All these compounds may influence the quality and yield of extracted DNA even in fresh samples and herbarium samples as well. Furthermore, extraction of DNA from herbarium specimens has always been difficult due to the preservation conditions or liquids in which specimens are preserved.

2.1 Key to Choice of Protocols

1. According to plant group

Vascular plants or conifers	*Method 1*
Mosses	*Method 1*
Lichens	*Method 6*
Algae	*Method 7*
Mushrooms/fungi	*Method 8*

2. According to type of plant tissue

Plant material with leaf incrustation or containing sclerenchyma strands	*Method 1*
Plants with needles	*Method 1*
Seeds	*Method 3*
Plant material containing polysaccharides or phenolic compounds	*Method 4*
Plant material containing mucilaginous tissues	*Method 5*

3. According to type of next procedures

DNA extraction for PCR based methods	*Method 1*
DNA extraction for AFLP	*Method 2*
DNA extraction for microsatellites	*Method 9*
Extraction of ultrashort DNA molecules for "next-generation" sequencing	*Method 10*

2.2 General Equipment for all Protocols

1. Manual pipettes.
2. Centrifuge for microcentrifuge tubes.
3. Vortex.
4. Thermal heating-block or water bath for incubation and pre-heating of Buffers (up to 65 °C).
5. Equipment for sample disruption and homogenization (TissueLyser or Mixer Mill) including TissueLyser Adapter Set and cylinders (Tungsten carbide beads or ceramic cylinders).
6. 1.5–2 mL microcentrifuge tubes
7. Disposable tips.
8. Ice.
9. Personal protection equipment (lab coat, gloves).

2.3 Method 1: DNeasy Plant Mini Kit (QIAgen) for Plants with Leaves Containing Sclerenchyma Strands

1. Buffers from DNeasy Plant Mini Kit (AP1, AP2, AP3/E, AW, AE).
2. RNase A (100 mg/mL).
3. 100% ethanol
4. Tungsten carbide beads.
5. QIAshredder Mini spin column.
6. DNeasy Mini spin column.

2.4 Method 2: DNeasy Plant Mini Kit (QIAgen) Modified for AFLP

The following are in addition to the items needed for Method 1

1. Proteinase K (19.45 mg/mL).
2. Mortar and pestle.
3. Liquid nitrogen.
4. Quartz sand.

2.5 Method 3: DNA Extraction from Seeds

The following are in addition to the items needed for Method 1:

1. Tween 20.
2. Liquid nitrogen.
3. 10% bleach solution (sodium hypochlorite).

2.6 Method 4: The STE/CTAB Method for Micro-Scale DNA Extraction from Polysaccharide-Rich Plants

1. STE (Sucrose-Tris-EDTA): 0.25 M sucrose, 0.03 M Tris, 0.05 M EDTA.
2. 2× CTAB (cetyltrimethylammonium bromide) extraction Buffer: 100 mM Tris–HCl (pH = 8.0), 1.4 M NaCl, 20 mM EDTA (pH = 8.0), 2% (w/v) PVPP (polyvinyl polypyrrolidine), 0.1% (v/v) β-mercaptoethanol (include to the solution immediately prior to use)
3. Chloroform.
4. Isopropanol.
5. 80% ethanol
6. TE Buffer solution: 10 mM Tris–HCl (pH = 8), 1 mM EDTA.
7. Liquid nitrogen.

2.7 Method 5: Modified CTAB Adapted Method for Mucilaginous Tissues

1. 2× CTAB (cetyltrimethylammonium bromide) extraction Buffer: 100 mM Tris-HCl (pH = 8.0), 1.4 M NaCl, 20 mM EDTA (pH = 8.0), 2% (w/v) PVPP (polyvinyl polypyrrolidine), 0.1% (v/v) β-mercaptoethanol (include to the solution immediately prior to use)
2. β-mercaptoethanol,
3. SEVAG: chloroform/isoamyl alcohol 24:1.
4. Ice-cold isopropanol.
5. Isopropanol.
6. TE Buffer: 10 mM Tris–HCl (pH = 8), 1 mM EDTA.
7. RNase (10 mg/mL).
8. 2.5 M Sodium Acetate (NaOAc)
9. 95% ethanol
10. 70% ethanol
11. Vacuum desiccator.

2.8 Method 6: Modified CTAB Method for Fungi and Lichen Forming Fungi

1. Extraction Buffer: 1% (w/v) CTAB, 1 M NaC1, 100 mM Tris, 20 mM EDTA (pH = 8.0), 1% (w/v) PVPP (polyvinyl polypyrrolidine) (include to the solution immediately prior to use).
2. Precipitation Buffer: 1% (w/v) CTAB, 50 mM Tris–HC1, 10 mM EDTA, 40 mM NaCl.
3. 1.2 M NaC1
4. SEVAG: chloroform/isoamyl alcohol 24:1.
5. RNase A (10 mg/mL).
6. Isopropanol.
7. 70% ethanol
8. PCR grade water (nuclease-free water).
9. TE Buffer solution: 10 mM Tris–HCl (pH = 8), 1 mM EDTA.
10. Liquid nitrogen.

2.9 Method 7: CTAB/HNO_3 Method for Algae

1. 2% CTAB extraction Buffer: 100 mM Tris–HCl (pH = 8.0), 1.4 M NaCl, 20 mM EDTA (pH = 8.0), 0.1% (w/v) PVPP (polyvinyl polypyrrolidine), 0.2% (v/v) β-mercaptoethanol (added freshly)
2. Binding Buffer: 6 M NaI, 0.1 M Na_2SO_3.
3. HNO_3 (1 mL, 5 M).
4. Washing Buffer: 20 mM Tris–HCl (pH 8), 1 mM EDTA, 0.1 mM NaCl solution, 18 mL 100% ethanol.
5. TE Buffer solution: 10 mM Tris–HCl (pH = 8), 1 mM EDTA.
6. 0.45-μm membrane filter (Whatman)
7. MilliQ filtered de-ionized water.
8. Silica gel.

2.10 Method 8: DNA Extraction from Dried Mushrooms Using Enzymatic Digestion and Glass-Fiber Filtration (EDGF)

1. Proteinase K (20 mg/mL).
2. Lysis Buffer (LB): 100 mM NaCl, 50 mM Tris–HCl (pH 8.0), 10 mM EDTA (pH 8.0), 0.5% (w/v) SDS.
3. Binding Buffer (BB): 6 M GuSCN, 20 mM EDTA (pH 8.0), 10 mM Tris–HCl (pH 6.4), 4% (v/v) Triton X-100.
4. Binding mix (BM): 50 mL of ethanol (96%) thoroughly mixed with 50 mL of BB.
5. Protein wash Buffer (PWB): 70 mL of ethanol (96%), 26 mL of BB.
6. Wash Buffer (WB): ethanol (60%), 50 mM NaCl, 10 mM Tris–HCl (pH 7.4), 0.5 mM EDTA (pH 8.0).
7. TE Buffer solution: 10 mM Tris–HCl (pH = 8), 1 mM EDTA.
8. PCR plate (e.g., Sorenson 96-well UltraAmp).
9. PALL collar.
10. Glass fiber filtration (GF) membrane (Whatman).
11. Aluminum cover.

2.11 Method 9: NucleoSpin Plant II Kit (Macherey-Nagel) Used for Microsatellites

1. Buffers from NucleoSpin Plant II kit (PL1, PL2, PC, PW1, PW2, PE).
2. RNase A (100 mg/mL).
3. 96–100% ethanol
4. NucleoSpin Plant II Column**.**

2.12 Method 10: Extraction of Ultrashort DNA Molecules for "Next-Generation" Sequencing

1. Buffers from DNeasy Plant Mini Kit (AP2, AP3/E, AW, AE).
2. PTB extraction Buffer: 1% SDS, 10 mM Tris, pH 8.0, 5 mM NaCl. 50 mM DTT, 0.4 mg/mL proteinase K, 10 mM EDTA, 2.5 mM N -phenacylthiazolium bromide (PTB).
3. Dithiothreitol (DTT).

4. 100% ethanol
5. Tungsten carbide beads.
6. MinElute Purification columns (Qiagen).

3 Methods

3.1 DNeasy Plant Mini Kit (QIAgen)

Several commercially available DNA extraction kits are very popular for high-quality extracted DNA and ease of use. For extraction of herbarium specimens it is usually necessary to modify the manufacturer protocol. The most valuable part of the kits is silica-gel-membrane spin columns for convenient extraction of high-quality DNA, especially for PCR.

All procedures should be carried out at room temperature unless different conditions are specified (e.g., sample incubation on ice).

3.1.1 Method 1: DNeasy Plant Mini Kit (QIAgen) for Plants with Leaves Containing Sclerenchyma Strands

This extraction protocol was modified for monocots [24] and is useful for all PCR-based applications. As PCR requires only minute amounts of DNA, it suggests that herbarium collections will become more valuable as sources of material for molecular studies and analyses based on PCR technique [25]. However, herbarium samples do require special extraction and reaction conditions (the most crucial points are emphasized in Notes).

Mechanical disruption of plant material proved to be a limiting step when handling multiple samples in parallel [34]. Therefore, the tissue should be ground in the mixer mill (Tissue Lyser) with tungsten carbide beads or ceramic cylinders. This procedure is optimal for sufficient homogenization of hard leaf structure of, e.g., *Juncaceae*, *Cyperaceae*, and *Pinaceae*. This extraction is presented according to the QIAgen protocol with a modification for dried samples. These modifications were introduced mainly in the laboratory of the Institute of Botany, Copenhagen University (G. Petersen, personal communication).

1. Place 0.5–1 g of dried leaf tissue together with 3 mm tungsten carbide beads (2 or 3 pieces) into a 1.5 mL microcentrifuge tube. Place the tubes into the TissueLyser Adapter Set, and fix into the clamps of the TissueLyser. Grind the samples for 1–3 min at 30 Hz (*see* **Note 1**).
2. Add 450 μL Buffer AP1 and 4 μL RNase A to a maximum of 20 mg dried disrupted plant tissue and vortex powerfully (*see* **Note 2**).
3. Incubate the mixture for 30 min at 65 °C for cell lysis. Mix 2 or 3 times during incubation by inverting tube.
4. Add 130 μL Buffer AP2 to the lysate, mix, and incubate for 5 min on ice.

5. Pipet the lysate into the QIAshredder Mini spin column placed in a 1.5 µL collection tube, and centrifuge for 2 min at 20,000 × *g* (*see* **Note 3**).
6. Transfer the flow-through fraction from the previous step into a new tube without disturbing the cell-debris pellet. Usually 450 µL of lysate is recovered (*see* **Note 4**).
7. To 450 µL lysate add 675 µL Buffer AP3/E. Reduce the amount of Buffer AP3/E to 1.5 volumes if different volume of lysate than 450 µL is obtained. Mix it by pipetting. It is important to pipet Buffer AP3/E directly onto the cleared lysate and to mix immediately.
8. Pipet 650 µL of the mixture from the previous step, including any precipitate that may have formed, into the DNeasy Mini spin column placed in a 2 mL collection tube. Centrifuge for 1 min at 6000 × *g*, and discard the flow-through. Reuse the collection tube in the next step.
9. Repeat previous step with remaining sample. Discard flow-through and collection tube. Place the DNeasy Mini spin column into a new 2 mL collection tube, add 500 µL Buffer AW, and centrifuge for 1 min at 6000 × *g*. Discard the flow-through and reuse the collection tube in the next step.
10. Add 500 µL Buffer AW to the DNeasy Mini spin column, and centrifuge for 2 min at 20,000 × *g* to dry the membrane (*see* **Note 5**).
11. Transfer the DNeasy Mini spin column to a 1.5 mL or 2 mL microcentrifuge tube, and pipet 50 µL Buffer AE directly onto the DNeasy membrane. Incubate for 10 min at room temperature (15–25 °C), and then centrifuge for 1 min at 6000 × *g* to elute.
12. Repeat the elution step.
13. Store at −20 °C or − 80 °C (*see* **Notes 6** and **7**).

3.1.2 Method 2: DNeasy Plant Mini Kit (QIAgen) Modified for AFLP

This extraction protocol was modified for dried vascular plants [35] and successfully used for AFLP.

1. Grind the plant tissue in a mortar with quartz sand and about 3 mL liquid nitrogen into a very fine powder.
2. Preheat a total of 500 µL AP1 Buffer (60 °C), add to the sample and grind until the mixture is completely homogeneous.
3. After grinding add 4 µL RNase and 4 µL Proteinase K.
4. Transfer the mixture to an Eppendorf tube and incubate at 60 °C for 1 h.

5. Add 150 µL AP2 Buffer and follow the Qiagen extraction protocol to the final step, in which elute the DNA with 50 µL preheated (60 °C) AE Buffer.
6. Store at −20 °C or −80 °C (*see* **Note 8**).

3.1.3 Method 3: DNA Extraction from Seeds

This extraction protocol was modified [22] for seeds of vascular plants.

1. Remove seed coats from seeds after a preliminary 10-min soak in 10% bleach solution containing a drop of Tween 20.
2. Ground whole embryos, or separated embryonic axes and cotyledons into a powder in the presence of liquid nitrogen.
3. Extract and purify DNA using the DNeasy Mini Kit according to Subheading 3.1.1. or the manufacturer's instructions.

3.2 CTAB Modified Methods

The CTAB extraction methods are based on the well-established CTAB extraction procedure [19]. However, there are some modifications for different types of plant tissues. Two protocols modified for mucilaginous tissues and fungi and lichen forming fungi follow.

3.2.1 Method 4: The STE/CTAB Method for Micro-Scale DNA Extraction from Polysaccharide-Rich Plants

This extraction protocol was modified [36] for polysaccharide-rich plant tissues.

1. Place 0.5–1 g of dried leaf tissue in a microcentrifuge tube with a sterile grinder. Snap freeze by suspending the tube in liquid nitrogen and grind to a fine powder (*see* **Note 9**).
2. Add 1 mL of freshly made STE to the ground plant tissue. Vortex, then centrifuge at 2000 × *g* for 10 min. Discard supernatant and repeat STE wash.
3. Add 600 µL of CTAB solution and incubate at 60 °C for 40 min with occasional shaking.
4. Add 600 µL chloroform and shake vigorously to homogenize. Pulse centrifuge to 7000 × *g*.
5. Remove upper aqueous layer with a wide-bore pipette tip into a new microcentrifuge tube. Add 600 µL of room-temperature isopropanol and invert gently.
6. Leave at room temperature for 1–5 min and transfer DNA pellet using a wide-bore pipette tip into a microcentrifuge tube containing 800 µL of 80% ethanol. Wash pellet by gently inverting several times. Remove DNA pellet to a new microcentrifuge tube and repeat ethanol wash.
7. Dry the pellet at room temperature and suspend in 30–60 µL of TE.

3.2.2 Method 5: Modified CTAB Adapted Method for Mucilaginous Tissues

This extraction protocol was modified [37] for plant mucilaginous tissues. The DNA obtained by this extraction can be used not only for PCR-based techniques, but also for AFLP.

1. Add 750 μL of 2× CTAB Buffer and 3.0 μL of β-mercaptoethanol to Eppendorf tubes.
2. Grind 0.5–1.0 g of tissue with liquid nitrogen and sterilized sand until finely powdered.
3. Add a spatula-tip of powdered tissue to each tube and mix well.
4. Incubate in a water bath at 55–60 °C for 1–5 h, mixing every 15 min.
5. Add 700 μL of SEVAG to each tube and mix thoroughly. Centrifuge at 9240 × *g* for 10–15 min. Transfer the aqueous phase to a new Eppendorf tube.
6. Add 0.33 vol of ice-cold isopropanol and store at −30 °C for at least 1 h.
7. Spin at 9240–13,305 × *g* for 10 min at room temperature. Discard supernatant without disturbing the pellet. Vacuum dry. Repeat **steps 6** and **7** two to four times if the aqueous phase is viscous.
8. Resuspend pellet in 100–200 μL of TE. Add 1–2 μL of RNase. Mix well and incubate for 30 min at 37 °C.
9. Add 20 μL (0.1 vol) of NaOAc and 500 μL (2–2.5 vol) ice-cold 95% ethanol and store at −20 °C for ≥30 min. Spin at 9240–13,305 × *g* for 5 min. Discard supernatant.
10. Wash pellet with 1 mL of 70% ethanol. Do not disturb the pellet. Spin at 9204 × *g* for 4 min and pour off ethanol. Vacuum-dry pellet. Do not overdry (*see* **Note 10**).
11. Resuspend pellet in 100–200 μL of TE. Store at −20 °C.

3.2.3 Method 6: Modified CTAB Method for Fungi and Lichen Forming Fungi

This extraction protocol was modified [18] and adapted for fungi and lichen-forming fungi [38].

The best results were obtained from liquid nitrogen frozen samples [18]. Samples can be disrupted without liquid nitrogen by grinding the material in powdered glass. DNA extracted in this way gave good amplifications, although the total DNA yield was reduced compared to liquid nitrogen preparations. A mortar and pestle can also be used without additional abrasives although this did not prove practical for either large numbers or small amounts of material.

1. Put 3–100 mg of material into 1.5 mL tubes and place in a container with liquid nitrogen for 5–10 min. Then remove from the container and place in an insulated rack. Add liquid nitrogen to the tube and grind the material with a sterile

precooled sharp glass bar. Sterilize glass bars in flame immediately prior to use.

2. Add 0.5 mL of pre-warmed extraction Buffer to the ground material. Add PVPP to the Buffer immediately prior to use. Mix the tubes by inverting several times and then heat in a water bath for 30 min at 70 °C.
3. Add one volume of SEVAG. Mix by inverting the tube and centrifuge for 5 min at 10,000 × *g* at room temperature.
4. Collect the upper aqueous phase in a new tube and discard the slurry and lower layers (*see* **Note 11**).
5. Add two volumes of precipitation Buffer to the supernatant and mix well by inversion for 2 min.
6. Centrifuge the mixture for 15 min at 13,000 × *g* at room temperature and collect the pellet.
7. Resuspend the pellet in 350 μL of 1.2 M NaCl and add one volume of SEVAG. Mix vigorously and centrifuge for 5 min at 10,000 × *g* at room temperature.
8. For RNA-free DNA add 2 μL of RNase A to the sample and incubate at 37 °C for 30 min.
9. Remove the upper phase to a new tube and add 0.6 volume of isopropanol. Mix by inversion and place the tube at −20 °C for 15 min.
10. Collect the final pellet by centrifugation for 20 min at 13,000 × *g* at 4 °C. Wash the final pellet with 1 mL of 70% ethanol and recollect by centrifugation for 3 min at 13,000 × *g* at 4 °C. Then drain the pellet and dry at 50 °C prior to resuspension in either PCR-grade water or TE Buffer.

3.2.4 Method 7: CTAB/HNO_3 Method for Algae

This extraction protocol was modified [39] for brown macroalgae.

1. Grind the plant tissue and add 2% CTAB Buffer.
2. Clarify the binding Buffer by filtration through a 0.45-μm membrane filter.
3. Prepare silica fines by placing 20–30 g of silica gel into *c.* 500 mL of milliQ filtered de-ionized water and stirring for *c.* 1 h. After stirring, allow the silica to settle for *c.* 15 min.
4. Transfer the supernatant to 50 mL plastic tubes and centrifuge for 5 min at 1250 × *g*.
5. Remove most of the supernatant from each 50 mL tube; leave only a small amount for resuspension of the pelleted particles and subsequent consolidation into one tube.
6. Transfer aliquots (c. 1 mL) of the consolidated particles to 2 mL plastic tubes.

7. Add HNO_3 to each 2 mL tube prior to heating at 95 °C to 100 °C for 30 min in a vented hood.
8. After cooling, centrifuge the tubes at 13,000 × *g* for 1 min and discard the supernatant.
9. Wash the silica pellet by resuspending in *c.* 2 mL milliQ-filtered de-ionized water and centrifuge for 1 min at 13,000 × *g*. Discard the supernatant.
10. Repeat the washing step five times prior to a final resuspension with an equal volume of milliQ-filtered de-ionized water.
11. Add 6.8 mL of the washing Buffer.
12. Elute in the TE Buffer.

3.3 Method 8: DNA Extraction from Dried Mushrooms Using Enzymatic Digestion and Glass-Fiber Filtration (EDGF)

This extraction protocol was described [40] for animal tissues and modified [41] for dried mushrooms.

1. Add a small amount of sample (1–2 mm^3) to each well of a 96-well PCR plate. Instruments should be flame sterilized between samples to avoid cross contamination. Last well can be left blank and used as a negative control.
2. Mix 5 mL of LB and 0.5 mL of Proteinase K (20 mg/mL) in a sterile container and dispense 100 μL to each well. Cover each row with caps and incubate at 56 °C overnight (8–16 h) to allow digestion.
3. Centrifuge at 1000 × *g* for 1 min.
4. Add 100 μL of BM to each sample. Mix by pipetting up and down a few times.
5. Remove cap strips/cover and transfer the lysate (about 150 μL) from the wells of microplate into the wells of the PALL glass fiber filtration (GF) plate placed on top of a square-well block. Seal the plate with adhesive cover.
6. Centrifuge at 1500 × *g* for 10 min to bind DNA to the GF membrane.
7. Add 250 μL of PWB to each well of the GF plate. Seal with a new adhesive cover and centrifuge at 1500 × *g* for 5 min. Discard the flowthrough.
8. Add 300 μL of WB to each well of the GF plate. Seal with a new cover and centrifuge at 1500 × *g* for 10 min.
9. To avoid incomplete WB removal, open the cover to relieve the vacuum that may have formed in the wells, seal the plate again and centrifuge the plates again at 1500 × *g* for 5 min. Discard the flow-through.
10. Repeat **steps 8** and **9**.

11. Remove the cover. Place the GF plate on a clean square-well block and incubate at 56 °C for 30 min to evaporate residual ethanol.
12. Position a PALL collar on a collection plate and place plate and collar on top of a clean square-well block. Place GF PALL plate with DNA bound to the membrane on top of a PCR plate. Dispense 50 μL of 0.1× TE Buffer or water, pre-warmed at 56 °C, directly onto the membrane of each well of GF plate and incubate at room temperature for a few minutes and then seal plate.
13. Centrifuge at 1500 × *g* for 10 min to collect the eluted DNA. Remove the GF plate and discard it.
14. Cover DNA plate with an aluminum cover. Keep at 4 °C for temporary storage or at −20 °C for long-term storage.

3.4 Method 9: NucleoSpin Plant II Kit (Macherey-Nagel) Used for Microsatellites

This extraction protocol was used [42] prior to microsatellite data analysis (*see* **Note 12**).

1. Homogenize up to 20 mg dry weight plant.
2. Transfer the resulting powder to a new tube and add 400 μL Buffer PL1. Vortex the mixture thoroughly (*see* **Note 13**). Alternatively, transfer the resulting powder to a new tube and add 300 μL Buffer PL2. Vortex the mixture thoroughly. If the sample cannot be resuspended easily, additional Buffer PL2 can be added.
3. Add 10 μL RNase A solution and mix sample thoroughly. Incubate the suspension for 10 min at 65 °C. Alternatively, add 75 μL Buffer PL3, mix thoroughly and incubate for 5 min on ice to precipitate SDS completely (*see* **Note 14**).
4. Place a NucleoSpin Filter into a new collection tube (2 mL) and load the lysate onto the column. Centrifuge for 2–5 min at 11,000 × *g*, collect the clear flow-through and discard the NucleoSpin Filter. If all liquid has not passed the filter, repeat the centrifugation step. If a pellet is visible in the flow-through, transfer the clear supernatant to a new 1.5 mL microcentrifuge tube.
5. Add 450 μL Buffer PC and mix thoroughly by pipetting up and down (5 times) or by vortexing.
6. Place a NucleoSpin Plant II Column into a new collection tube (2 mL) and load a maximum of 700 μL of the sample (*see* **Note 15**). Centrifuge for 1 min at 11,000 × *g* and discard the flowthrough.
7. Preheat Buffer PE to 65 °C.

8. Add 400 µL Buffer PW1 to the NucleoSpin Plant II Column. Centrifuge for 1 min at 11,000 × *g* and discard the flowthrough.
9. Add 700 µL Buffer PW2 to the NucleoSpin Plant II Column. Centrifuge for 1 min at 11,000 × *g* and discard the flowthrough.
10. Add another 200 µL Buffer PW2 to the NucleoSpin Plant II Column. Centrifuge for 2 min at 11,000 × *g* in order to remove wash Buffer and dry the silica membrane completely.
11. Place the NucleoSpin Plant II Column into a new 1.5 mL microcentrifuge tube. Pipette 50 µL Buffer PE (65 °C) onto the membrane. Incubate the NucleoSpin Plant II Column for 5 min at 65 °C. Centrifuge for 1 min at 11,000 × *g* to elute the DNA. Repeat this step with another 50 µL Buffer PE (65 °C) and elute into the same tube.

3.5 Method 10: Extraction of Ultrashort DNA Molecules for "Next-Generation" Sequencing: Modification of DNeasy Plant Mini Kit (QIAgen)

Recently, many commercially available DNA extraction kits can be used for extraction of plant herbarium DNA, but sometimes modifications to standard protocols are often necessary to improve the DNA yield. This method is chosen from [43–45], who optimized it and successfully tested it in many different herbarium samples for "next-generation" sequencing. A modified extraction method combines a N-phenacylthiazolium bromide lysis Buffer, DNeasy Plant Mini Kit (Qiagen), and MinElute Purification columns (Qiagen). Old specimens contain much shorter DNA fragments than fresh material, therefore to efficiently retrieve short molecules, this modified extraction method, which combines PTB lysis Buffer with MinElute Purification columns (Qiagen), is utilized. PTB cleaves glucose-derived protein crosslinks [46], and releases DNA trapped within sugar-derived condensation products [47]. To compare the choice of extraction Buffer that has a great impact on the length distribution of molecules recovered from herbarium specimens, *see* [43].

1. Prepare 1.2 mL PTB extraction Buffer per sample.
2. Place 0.5–1 g of dried leaf tissue together with 3 mm tungsten carbide beads (2 or 3 pieces) into a 1.5 mL microcentrifuge tube. Place the tubes into the TissueLyser Adapter Set, and fix into the clamps of the TissueLyser. Grind the samples for 1–3 min at 30 Hz (*see* **Note 1**).
3. Add the sample to 1.2 mL of PTB extraction Buffer in a 2-mL or larger tube, and vortex to homogenize thoroughly. The mixture should be somewhat fluid, not a dry cake in the tube [44]. Add more PTB extraction Buffer, if necessary, to achieve the desired consistency (*see* **Note 16**).

4. Incubate the mixture at 37 °C with constant agitation for 18–24 h.
5. Centrifuge the mixture at 9000 × *g* for 5 min. The samples should separate into a dense mass of tissue and about 500–700 μ L of supernatant. If the tissue is not suitably compacted (i.e., if more than a very small amount of visible debris is suspended in the supernatant), centrifuge for an additional 2 min at up to 16,000 × *g*.
6. Transfer the supernatant from each tube to a new 1.5- or 2-mL tube, and estimate the recovered volume for the next step.
7. Add 0.325 volumes of Qiagen Buffer AP2, mix, and incubate on ice 5 min.
8. Pipet the lysate into the QIAshredder Mini spin column placed in a 1.5 μL collection tube, and centrifuge for 2 min at 20,000 × *g* (*see* **Note 3**).
9. Transfer the flowthrough fraction from the previous step into a new tube without disturbing the cell-debris pellet. Usually 450 μL of lysate is recovered (*see* **Note 4**).
10. To 450 μL lysate add 675 μL Buffer AP3/E. Reduce the amount of Buffer AP3/E to 1.5 volumes if different volume of lysate than 450 μL is obtained. Mix it by pipetting. It is important to pipet Buffer AP3/E directly onto the cleared lysate and to mix immediately.
11. Pipet 650 μL of the mixture from the previous step, including any precipitate that may have formed, into the MinElute Purification column (Qiagen) placed in a 2 mL collection tube. Centrifuge for 1 min at 6000 × *g*, and discard the flowthrough. Reuse the collection tube in the next step.
12. Repeat previous step with remaining sample. Discard flowthrough and collection tube. Place the MinElute Purification column into a new 2 mL collection tube, add 750 μL PE Buffer, and centrifuge for 1 min at 6000 × *g*. Discard the flowthrough and reuse the collection tube in the next step.
13. Add 750 μL PE Buffer to the MinElute Purification column, and centrifuge for 2 min at 20,000 × *g* to dry the membrane (*see* **Note 17**).
14. Transfer MinElute Purification column to a 1.5 mL or 2 mL microcentrifuge tube, and pipet 50 μL Buffer AE directly onto the column membrane. Incubate for 1 min at room temperature (15–25 °C), and then centrifuge for 1 min at 6000 × *g* to elute.
15. Repeat the elution step.
16. Store at −20 °C or − 80 °C (*see* **Notes 6** and **7**).

4 Notes

1. Proper grinding of plant samples with a TissueLyser or Mixer Mill is the crucial step. The plant tissue should be ground to a fine powder after the disruption. However, for some plants one disruption step may not be sufficient. In that case repeat the disruption for 1 min at 30 Hz until the sample is thoroughly and equally homogenized.
2. It is necessary to remove tissue clumps, because tissue clumps will not lyse properly and therefore decrease yield of DNA. If the small amount of sample is expected, use longer precipitation or repeat it.
3. It may be necessary to cut the end off the pipet tip to apply the lysate to the QIAshredder Mini spin column. The QIAshredder Mini spin column removes most precipitates and cell debris, but a small amount will pass through and form a pellet in the collection tube.
4. It is crucial not to disturb the pellet. In case you do that, repeat **step 5**. For herbarium specimens usually less lysate is recovered. In this case, determine the volume for the next step.
5. It is important to dry the membrane of the DNeasy Mini spin column since residual ethanol may interfere with subsequent reactions. Discard flow-through and collection tube.
6. Preferably short-term storage in TE (or AE Buffer) at −25 °C, for long-term storage use −80 °C.
7. The exclusion of samples based on visualization of total DNA on agarose gel alone is gratuitous. This statement is also valid for other techniques as AFLP (see below). For PCR the best results require short length of products (optimum of 300–350 to 500 bp). Higher number of PCR cycles are recommended.
8. Use short AFLP fragments, up to 300 bp (depending on the quality/quantity of DNA and chromatograms). To compensate for using only part of the chromatogram, it may be necessary to increase the number of primer combinations in order to obtain a sufficient number of polymorphic fragments. Even samples for which DNA appearance on the agarose gel showed small amount and/or low quality may in some cases work well for AFLP.
9. To obtain a fine powder is the most crucial step.
10. If the pellet is disturbed, centrifuge again.
11. Do not disturb the lower layers and the pellet. If you do so, centrifuge again.
12. Proceed with cell lysis using Buffer PL1 or alternatively with Buffer PL2 or Buffer PL3. Test the different Buffers and

choose the Buffer most appropriate to the plant tissue or the plant species used: Buffer PL1 is based on the established CTAB procedure. Additionally, the SDS-based Lysis Buffer PL2 is provided by the manufacturer, which requires subsequent protein precipitation by potassium acetate (Precipitation Buffer PL3).

13. If the sample cannot be resuspended easily, additional Buffer PL1 can be added.
14. The maximum loading capacity of the NucleoSpin Plant II Column is 700 μL. For higher sample volumes repeat the loading step.
15. Extraction with N-phenacylthiazolium bromide (PTB) Buffer decreased median fragment length by 35% when compared with cetyl-trimethyl ammonium bromide (CTAB) [43].
16. Residual ethanol from PE Buffer will not be completely removed unless the flowthrough is discarded before this additional centrifugation for 1 min at maximum speed.

Acknowledgments

Upgrade of the study was supported by GAČR 19-02699S.

References

1. Guschanski K, Krause J, Sawyer S, Valente LM, Bailey S, Finstermeier K, Sabin R, Gilissen E, Sonet G, Nagy ZT et al (2013) Next-generation museomics disentangles one of the largest primate radiations. Syst Biol 62:539–554
2. Rogers SO (1994) Phylogenetic and taxonomic information from herbarium and mummified DNA. In: Adams RP et al (eds) Conservation of plant genes II: utilization of ancient and modern DNA, Monographs in systematic botany from the Missouri Botanical Garden, vol 48. Missouri Botanical Garden
3. Tailor JW, Swann EC (1994) Dried samples: soft tissues, DNA from herbarium specimens. In: Herrmann B, Hummel S (eds) Ancient DNA. Springer Verlag, Heidelberg
4. Hall DW (1981) Microwave: a method to control herbarium insects. Taxon 30:818–819
5. Hill SR (1983) Microwave and the herbarium specimen: potential dangers. Taxon 32:614–615
6. Bacci M, Checcuccii A, Checcuccii G, Palandek MR (1983) Microwave drying of herbarium specimens. Taxon 34:649–653
7. Metsger DA, Byers SC (1999) Managing the modern herbarium, an interdisciplinary approach. Society for the preservation of natural history collections, Washington DC, p 384
8. Lindahl T, Andersson A (1972) Rate of chain breakage at apurinic sites in double-stranded deoxyribonucleic acid. Biochemistry 11:3618–3623
9. Doyle JJ, Dickson EE (1987) Preservation of plant species for DNA restriction endonuclease analysis. Taxon 36:715–722
10. Pyle MM, Adams RP (1989) In situ preservation of DNA in plant specimens. Taxon 38:576–581
11. Harris SA (1993) DNA analysis of tropical plant species: an assessment of different drying methods. Plant Syst Evol 188:57–64
12. Bakker FT (2017) Herbarium genomics: skimming and plastomics from archival specimens. Webbia 72:35–45
13. Lindahl T (1993) Instability and decay of the primary structure of DNA. Nature 362:709–715
14. Pääbo S, Poinar H, Serre D, Jaenicke-Despres-V, Hebler J, Rohland N, Kuch M, Krause J,

Vigilant L, Hofreiter M (2004) Genetic analyses from ancient DNA. Ann Rev Genet 38:645–679

15. Hofreiter M, Jaenicke V, Serre D, von Haeseler A, Pääbo S (2001) DNA sequences from multiple amplifications reveal artifacts induced by cytosine deamination in ancient DNA. Nucleic Acids Res 29:4793–4799
16. Gilbert MTP, Hansen AJ, Willerslev E, Rudbeck L, Barnes I, Lynnerup N, Cooper A (2003) Distribution patterns of postmortem damage in human mitochondrial DNA. Am J Hum Genet 72:48–61
17. Stiller M, Green RE, Ronan M, Simons JF, Du L, He W, Egholm M, Rothberg JM, Keates SG, Ovodov ND et al (2006) Patterns of nucleotide misincorporations during enzymatic amplification and direct large-scale sequencing of ancient DNA. Proc Natl Acad Sci U S A 103:13578–13584
18. Rogers SO, Bendich AJ (1994) Extraction of total cellular DNA from plants, algae and fungi. In: Gelvin SB, Schilperoort RA (eds) Plant molecular biology manual. Kluwer Academic Publishers, Dordrecht, pp 1–8
19. Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11–15
20. Ribeiro RA, Lovato MB (2007) Comparative analysis of different DNA extraction protocols in fresh and herbarium specimens of the genus *Dalbergia*. Genet Mol Res 6:173–187
21. Agostini G, Lüdtke R, Echeverrigaray S, de Souz-Chies TT (2011) Genomic DNA extraction from herbarium samples of *Cunila* D. Royen ex L. (Lamiaceae) and *Polygala* L. (Polygalaceae). Conservation Genet Resour 3:37–39
22. Wittzell H (1999) Chloroplast DNA variation and reticulate evolution in sexual and apomictic sections of dandelions. Mol Ecol 8:2023–2035
23. Ristaino JB, Groves CT, Parra GR (2001) PCR amplification of the Irish potato famine pathogen from historic specimens. Nature 411 (6838):695–697
24. Drábková L, Kirschner J, Vlček Č (2002) Historical herbarium specimens in molecular taxonomy of the Juncaceae: a comparison of DNA extraction and amplification protocols. Plant Mol Biol Rep 20(2):161–175
25. De Castro O, Menale B (2004) PCR amplification of Michele Tenore's historical specimens and facility to utilize an alternative approach to resolve taxonomic problems. Taxon 53:147–151
26. Jankowiak K, Buczkowska K, Szweykowska-Kulinska Z (2005) Successful extraction of DNA from 100-year-old herbarium specimens of the liverwort *Bazzania trilobata*. Taxon 54:335–336
27. Asif MJ, Cannon CH (2005) DNA extraction from processed wood: a case study for identification of an endangered timber species (*Gonystylus bancanus*). Plant Mol Biol Rep 23 (2):185–192
28. Erkens RHJ, Cross H, Maas JW, Hoenselaar K, Chatrou LW (2008) Assessment of age and greenness of herbarium specimens as predictors for successful extraction and amplification of DNA. Blumea 53:407–428
29. Lister DL, Bower MA, Howe CJ, Jones MK (2008) Extraction and amplification of nuclear DNA from herbarium specimens of emmer wheat: a method for assessing DNA preservation by maximum amplicon length recovery. Taxon 57:254–258
30. Andreasen K, Manktelow M, Razafimandimbison SG (2009) Successful DNA amplification of a more than 200-year-old herbarium specimen: recovering genetic material from the Linnaean era. Taxon 58:959–962
31. Poczai P, Taller J, Szabo I (2009) Molecular genetic study of a historical *Solanum* (Solanaceae) herbarium specimen collected by Paulus Kitaibel in the 18th century. Acta Bot Hung 51:337–346
32. Sohrabi M, Myllis L, Soili S (2010) Successful DNA sequencing of a 75 year-old herbarium specimen of *Aspicilia aschabadensis* (J. Steiner) Mereschk. Lichenologist 42:626–628
33. Walters C, Reilley AA, Reeves PA, Baszczak J, Richards CM (2006) The utility of aged seeds in DNA banks. Seed Sci Res 16:169–178
34. Csaikl UM, Bastion H, Brettschneider R, Gauch S, Metr A, Schauerte M, Schulz F, Sperisen C, Vornam B, Ziegenhagen B (1998) Comparative analysis of different DNA extraction protocols: a fast, universal maxi-preparation of high quality plant DNA for genetic evaluation and phylogenetic studies. Plant Mol Biol Rep 16:69–86
35. Lambertini C, Frydenberg J, Gustafsson MHG, Brix H (2008) Herbarium specimens as a source of DNA for AFLP fingerprinting of *Phragmites* (Poaceae): possibilities and limitations. Plant Syst Evol 272:224–231
36. Shepherd LD, McLay TGB (2011) Two micro-scale protocols for the isolation of DNA from polysaccharide-rich plant tissue. J Plant Res 124:311–314
37. Cota-Sánchez JH, Remarchuk K, Ubayasena K (2006) Ready-to-use DNA extracted with a CTAB method adapted for herbarium specimens and mucilaginous plant tissue. Plant Mol Biol Rep 24:161–167

38. Cubero OF, Crespo A, Fatehi J, Bridge PD (1999) DNA extraction and PCR amplification method suitable for fresh, herbarium- stored, lichenized, and other fungi. Plant Syst Evol 216:243–249
39. Ivanova NV, Dewaard JR, Hebert PDN (2006) An inexpensive, automation-friendly protocol for recovering high-quality DNA. Mol Ecol Notes 6:998–1002
40. Dentinger BTM, Margaritescu S, Moncalvo J-M (2010) Rapid and reliable high-throughput methods of DNA extraction for use in barcoding and molecular systematics of mushrooms. Mol Ecol 10:628–633
41. Hoarau G, Coyer JA, Stam TW, Olsen JL (2007) A fast and inexpensive DNA extraction/purification protocol for brown macroalgae. Mol Ecol Notes 7:191–193
42. Malenica N, Šimon S, Besendorfer V, Malecić E, Kontić JK, Pejić I (2011) Whole genome amplification and microsatellite genotyping of herbarium DNA revealed the identity of an ancient grapevine cultivar. Naturwissenschaften 98:763–772
43. Gutaker RM, Reiter E, Furtwangler A, Schuenemann VJ, Burbano HA (2017) Extraction of ultrashort DNA molecules from herbarium specimens. BioTechniques 62:76–79
44. Kistler L (2012) Ancient DNA extraction from plants. Methods Mol Biol 840:71–79
45. Dabney J, Knapp M, Glocke I, Gansauge MT, Weihmann A, Nickel B, Valdiosera C, Garcia N et al (2013) Complete mitochondrial genome sequence of a middle Pleistocene cave bear reconstructed from ultrashort DNA fragments. Proc Natl Acad Sci U S A 110:15758–15763
46. Vasan S, Zhang X, Zhang X, Kapurniotu A, Bernhagen J, Teichberg S, Basgen J, Wagle D et al (1996) An agent cleaving glucose-derived protein crosslinks in vitro and in vivo. Nature 382:275–278
47. Poinar HN, Hofreiter M, Spaulding WG, Martin PS, Stankiewicz BA, Bland H, Evershed RP, Possnert G, Paabo S (1998) Molecular coproscopy: dung and diet of the extinct ground sloth *Nothrotheriops shastensis*. Science 281:402–406

Chapter 5

Sequencing of Complete Chloroplast Genomes

Berthold Heinze

Abstract

In this chapter, frequently used methods for elucidating sequence and structure of chloroplast genomes are reviewed, as a current best practice guide. This concerns methods for DNA extraction, sequencing library preparation, and bioinformatics (assembly, verification, annotation, and sequence comparisons). Recommendations for standard data reporting practices are given—chloroplast genome sequencing reports can be highly formalized, and publication in the form of standard data reports is the best option for comparison and meta-analysis purposes.

Key words Chloroplast genome, High-throughput sequencing, Next-generation sequencing, Bioinformatics, Data reporting standards

1 Introduction

The number of completely sequenced chloroplast genomes has exploded in the last few years. While the numbers were in the tens in the early years of the millennium when it became possible to sequence entire nuclear genomes (e.g., *Populus trichocarpa* as the first tree species and third plant species overall, [1]), they were in the hundreds when we published the chloroplast genomes of the date palm [2] and *Syzygium cumini* [3], and a previous version of this book chapter [4]. Our recent publication of a conifer genome, *Abies alba* [5], also contained its completely sequenced chloroplast genome, and by now the entries of completely sequenced chloroplast genomes in NCBI's GenBank are approaching 4000 (accessed on January 27, 2020). This surge in sequenced chloroplast genomes came about through technical advances in sequencing ("next generation" or "high-throughput" methods), but also in

Electronic Supplementary Material The online version of this chapter (https://doi.org/10.1007/978-1-0716-0997-2_5) contains supplementary material, which is available to authorized users.

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_5,

bioinformatics, with several key programs or web services that are highly used for these purposes.

For this review, a small database of publications reporting complete sequences of chloroplast genomes was collated. It contains mainly articles published in the recent few years (approximately 2017–2019), and the main sources were the journal PLOS ONE (where I handled many manuscript submissions as an editor), and a collection of articles published in an e-book [6]; this selection is completely arbitrary, but it gives a good overview of currently preferred methods. By browsing the database and further reports of chloroplast genome sequencing, it becomes clear that methods and reports have become highly standardized, and that there is not a great variety, neither in wet lab techniques nor in bioinformatics. Therefore, the chapter will summarize these workflows and will give hints at alternatives to commonly used methods. The report of the sequencing of the deadly nightshade *Solanum dulcamara* [7] will serve as a guiding example, because it is comprehensive and the authors went back to their references (sequenced genomes in databases) and corrected these. This is important, as omissions, errors, or inconsistencies that once get a hold in databases will spread by taking over annotations and other features for the newly sequenced chloroplasts; thus perpetuating common shortcomings and errors in reference sequences.

Wet lab techniques being done in-house are now often reduced to DNA extraction (and sometimes quality checks), while library preparation and sequencing itself is outsourced to specialized genome centers. Both are highly reliant on commercial kits, for which exhaustive methods descriptions are available. For this reason, the chapter will not list individual step-by-step instructions, but rather it will review the most important questions that researchers who want to sequence chloroplast genomes are faced with. Similarly for bioinformatics, the possibilities will be listed and commented, but not be detailed in a step-by-step mode. The chapter will close with recommendations for reporting, which would benefit from applying a simple scheme in the form of a "data report."

2 Review of Methods

2.1 Database of Research Articles

The database contains approximately 50 entries of methods from published articles that describe chloroplast DNA sequencing (Data S1) in the form of an Excel table, with text and information from the articles, edited into a common format. It lists citations, journal source, digital object identifiers (DOI), plant taxonomic groups, purpose of study, material used, and DNA extraction and sequencing methods (in categories), as well as information on bioinformatic procedures for assembly, annotation, software tools used, and

any other useful hints. The following sections try to extract the essentials from the comparison of the methods present in the database entries.

2.2 Plant Material and DNA Extraction

Most studies in the database use fresh leaves (or other types of fresh plant materials). If not used immediately, tissue samples are placed on ice, or immediately frozen and stored at deep-freeze temperatures (e.g., −80 °C). Liquid nitrogen is sometimes used for shock frosting. It can also be used for breaking cell walls with a mortar and pestle, or with the help of a shaking mill. The resulting powder is then used for DNA extraction.

Alternatively, silica gel for gentle, but thorough drying of the material may be a convenient way to transport and store material, especially if collected in the field. Tee filter bags are convenient for placing individual samples separately into the silica gel, and they can be labelled easily. A few studies list cambium as a source of plant material. Obtaining cambium, the growth layer of tree trunks, is often more convenient than getting twigs or leaves from high tree crowns. A leather punch with approximately 1–2 cm diameter can serve for this purpose. A piece of cambium (with thin layers of bark and wood attached at opposite sides) can likewise be plunged into silica gel for drying and transport. In the laboratory, these layers are removed (e.g., with a scalpel, sharp knife, or razor blade), and the cambium layer is further processed.

There is an example of flower buds (which are not necessarily a photosynthetic tissue, thus may not be the best option for obtaining chloroplast DNA) and another describing the use of young leaf buds (which, on the contrary, consist of tightly packed cells, thus lots of DNA). Needles (in the case of conifers) can be used much in the same way, but may be harder to homogenize. A single article in the collection describes isolation of chloroplasts prior to DNA extraction. This may avoid an issue in bioinformatics—how to extract sequencing reads that belong to chloroplast DNA from the rest of the (nuclear and mitochondrial) DNA? There are ways of dealing with this in silico (see sections below), but enriching for chloroplasts in the first place may avoid this issue. Storing fresh leaves in a refrigerator (at 4 °C) at least overnight may reduce levels of chlorophyll, which can cause oxidation and problems associated with it, like browning and inhibition of enzymes (polymerase, ligase) in later steps.

The amount of plant material for DNA extraction varies widely, but this has to do with whether fresh or dry material is used. The smallest amounts listed are approximately 100 mg (presumably dry), whereas up to 20 g may be used as in the study of Shi et al. (2012 [8]), who worked on improvements in chloroplast isolation.

A high proportion of the reviewed articles used the well-known CTAB (cetyl trimethyl ammonium bromide) method (originally by [9]) or slight modifications of it for extracting the total genomic

DNA from plant tissues (18 articles). This method can easily be scaled to various amounts of starting material. Modifications include the addition of antioxidant chemicals or other buffer components. DNA is precipitated and re-dissolved in this method. Alternatively, commercial kits are popular for DNA extraction, including the DNeasy Plant Mini Kit (Qiagen, Hilden, Germany; listed 10 times) or the Plant Genomic DNA Kit of TIANGEN (Beijing, China; 4 times). These kits usually apply the principle of purifying DNA by binding it to silica membranes in the presence of high concentrations of chaotropic salts, and releasing it under low salt conditions. Others listed include the MagicMag Genomic DNA Micro Kit (Sangon Biotech, Shanghai, China), the Genome Wizard Kit (Promega, Madison, WI, USA), GenElute Plant Genomic DNA Miniprep kit (Sigma-Aldrich, St. Louis, MO, USA), HP Plant DNA Kit D2485-01 (Omega Bio-Tek, Santa Clara, CA, USA), and the Gentra Puregene Tissue Kit (QIAGEN, Hilden, Germany). The Invitrogen DNA Plantzol Reagent was used in only one method among the approximately 50. Two studies employed specific plant chloroplast isolation kits (Genmed Scientific Inc., Arlington, USA, and BTN120308, Beijing, China; the latter in combination with their Column Plant DNA Extraction Kit). The rest of the studies applied specific protocols, often optimized for the taxonomic group investigated. For example, one of these was based on the Dellaporta protocol [10], which uses sodium dodecyl sulfate (SDS) instead of CTAB, and high salt conditions, and another one involved DEDTCA (diethyl dithiocarbamate) as the surfactant. The quality and concentration of DNA are often checked by agarose gel electrophoresis (this is still highly recommended) and spectrometric/fluorometric methods, respectively. Among the latter, the Qubit and Nanodrop instruments are most popular, but their results can somewhat differ from each other, as Qubit is less sensitive to low-molecular DNA (which may therefore go undetected), and the Nanodrop instruments are sensitive to SDS carry-over on the lens. Capillary electrophoresis can replace both quality and quantity checks (e.g., the Bioanalyzer 2100 of Aligent, Santa Clara, CA, USA).

2.3 Construction of Sequencing Libraries

Few studies list the specific methods for fragmenting DNA. The sizes of DNA fragments for high-throughput sequencing are much smaller than the ones obtained with the various DNA extraction methods. A narrow size spectrum is important, as polymerase chain reaction (PCR) protocols are employed in nearly all workflows (as smaller insert sizes are preferentially amplified), and bioinformatic assembly methods incorporate the approximate insert size in their algorithms. Only the ends of the inserts are sequenced in most methods, thus their physical distance is an important piece of information for assembly. The Covaris shearing instruments (Woburn, MA, USA) is an industry standard. Several commercial

kits employ a different principle, in that a DNA-fragmenting enzyme is incubated with the DNA for various amounts of time in order to obtain a defined fragment size range. None of the methods screened in the database mention steps for size selection of DNA fragments. This could be easily done, e.g., by preparative agarose electrophoresis.

Specific size ranges listed in the various studies range from 150 basepairs (bp) to 20 kilobasepairs (kb), but the latter only applies for specific long-range techniques/instruments (PacBio SMRT and Nanopore sequencing). The most common ranges are 300 bp (5 times), 350 bp (5 times), and 500 bp (12 times). Just one article mentions a larger size range (600–800 bp), while another one used 800 bp inserts for the (long-range sequencing) PacBio SMRT technique. A single example of a mate-pair library of 5 kb insert size is present. Large inserts (mate-pair libraries) offer the advantage of better dealing with the large inverted repeats present in most chloroplast genomes, as the fragments will often span the junctions of these with the single-copy regions.

Frequently, libraries are prepared in the genome centers or by commercial service providers; in such cases, details of the library preparation methods are often lacking in the reports. Where they are listed, Illumina kits and protocols are the favorites and include the Nextera, TruSeq, and other variants. New England Biolab's NEBNext kits (in various versions) are second in line. It would be desirable to have this information in all chloroplast genome sequencing reports. Specific sequencing instruments (e.g., the Ion-Torrent and PacBio instruments) require specific library methods and kits.

2.4 Sequencing Platforms and Modes

The overwhelming sequencing mode is "paired-end" (30 articles). There is one mate-pair, and one single-end sequencing strategy among the articles in the database. Again, the sequencing mode is not reported in every case, although this would be highly desirable. Similarly, most studies sequenced for 2 × 150 bp (14), fewer with 2 × 100/101 bp or 2 × 125 bp (four to five). Longer reads can be done, in the case of Illumina sequencing, e.g., on the MiSeq machines (2 × 250 bp and 2 × 300 bp, total of six studies). Unfortunately, the latter ones did not report insert sizes in all cases, and not even read lengths are reported in each case (they can be partially deduced from the type of instrument employed, see below).

Illumina sequencing instruments have developed into a sort of industry standard in the most recent years. They were used in 40 of the 50 cases. The various instruments produce different read lengths, e.g., 2 × 100 bp on HiSeq 2000, 2× 125 or 150 bp on HiSeq 2500, and 2 × 150 bp also on HiSeq 4000 and HiSeq X. As mentioned above, MiSeq can cope with 2 × 250 or 300 bp reads. A single report in the collection used Roche's 454 GS FLX Titanium

platform, another one the IonTorrent, and two examples have made use of the substantially longer read sequencing capacities of the PacBio system (the RS II platform in both cases). One report mentions the NanoPore platform as an alternative; it is also used for much longer reads. Just one example of a PCR-based strategy found its way into the collection; these authors designed primers for long-range amplicons and sequenced those with Sanger technology. This strategy is now almost obsolete, but it has advantages in that DNA quality is less of an issue; it can be paralleled for a higher number of samples; and there is no need to employ bioinformatic routines to separate reads of chloroplast origin from others. Nevertheless, the need for (species-)specific primer design, and the need to sequence longer fragments in sections are work-intensive drawbacks. As only up to 1000 bp can be sequenced in one Sanger run, it would be necessary to sequence such long fragments of, e.g., 10,000 bp in ten or more (overlapping) sections, with specific primers designed for each. The chloroplast primer database [11] offers a collection of "universal" primers anchored in genes; these primers will be very helpful in a PCR-based sequencing strategy.

Data reporting standards for the amount of sequence obtained are poor in this collection of articles (at least in the methods sections). Only eleven articles mention the amount of raw or clean data in Gigabasepairs (Gb: range of 1.4 to 20.86) or in numbers of reads (seven articles; range, approximately 600,000 to 152 million reads; typical median numbers are from 3 to 60 million reads). The numbers of reads and amount of sequence data obtained should always be reported.

2.5 Raw Data (Read) Processing

There is a need to filter out low-quality reads. Next-generation/high-throughput methods yield sequences along with a quality assessment figure for each base. By comparing these quality levels among all reads, or by using a pre-set cutoff, reads with ambiguous sequences are removed. This is essentially also done in traditional Sanger sequencing (but not always as an obvious step)—sometimes by computer programs that assess the chromatograms (electropherograms), sometimes by visual inspection by the user. For next-generation/high-throughput sequencing, a number of commercial or free software solutions exist. CLC (Aarhus, Denmark) is a commercial solution; its Genomic Workbench runs on ordinary PCs and its graphic interface offers a convenient way to visualize data for a quick overview (there are also versions that run on servers). CLC programs have been used six times by methods in the database for this purpose. GENEIOUS (Auckland, New Zealand) is another commercial solution and quite popular among chloroplast sequencers. It can be employed for various tasks in chloroplast sequencing by NGS methods. Other, stand-alone solutions from the scientific literature/internet are FastQC (https://www.bioinformatics.

babraham.ac.uk/projects/fastqc/; mentioned by six articles), NGSQC Toolkit ([12]; nine times), Trimmomatic ([13]; ten times), PRINSEQ lite (http://prinseq.sourceforge.net/index.html; twice), CUTADAPT [14] and SICKLE (https://github.com/najoshi/sickle; once each), and SMRT Analysis (also mentioned only once; specific for PacBio). Another mentioned possibility for PacBio experiments is their long read correction tool LSC, which corrects the more error-prone PacBio reads with Illumina paired-end reads. Also GENEIOUS can perform trimming and filtering. However, a relatively high number of articles do not mention specific trimming and filtering procedures. Only a few among those specify the parameters explicitly; this is certainly the best practice.

2.6 Selection of Chloroplast Reads and Assembly

Most methods work with sequence reads from the overall genome and thus contain a mixture of nuclear, mitochondrial, and chloroplast sequences. There is a need to somehow select or filter for only chloroplast reads. These are often more numerous (e.g., [1]), but that criterion alone will not suffice, as there are also numerous reads from mitochondria, and from nuclear repeats. Furthermore, the sequence of chloroplast origin can, over evolutionary times, "travel" to the mitochondrial and nuclear genomes and get incorporated there [1]. Because these sequences and genes will most often lose their function, their mutation rates are high. Reads of such origins must thus be excluded from assembly, as they may introduce polymorphisms caused by mutations of the mitochondrial or nuclear copies. This is not a trivial task.

The easiest way to cope with this issue is to use a reference chloroplast genome against which the reads are aligned; this is what most methods in the database did. This can also be tricky, as it will only identify homologous sequences that are present in the reference genome. For most cases, where the reference is closely related, this will not be a big issue, as the chloroplast genomes are often very similar. There are exceptions, however, for example in parasitic plants (which do not depend on a fully functioning photosynthetic apparatus, with the corresponding chloroplast genes becoming dispensable).

The showcase example in this respect is [15] who used not just one reference genome, but a collection of 1688 complete plastome sequences from GenBank, against which the reads were filtered. This requires higher computer power of course, but should still be possible on desktop machines. A similar approach is to use BLAST against only chloroplast entries in GenBank.

As for software for assembly, there are a handful of preferred choices in the methods reviewed: CLC Genomic Workbench or similar products of the same company (mentioned seven times in various versions; a general multipurpose assembler that is capable of assembling chloroplast genomes on desktop PCs); GENEIOUS

(various versions; four times, also for multiple purposes); SOAPdenovo [16] (various versions, seven times), usually employed for large genome sequencing projects; SPADES (mostly versions 3.*, seven times); NOVOPLASTY [17] (various versions; six times), an interesting approach that starts from confirmed chloroplast sequence and tries to extend the assembly by adding reads at both ends; and a few others used in two to three methods each: the GetOrganelle python pipeline [18], MITObim [19] (both are also biting and iteration approach), ABYSS [20] (an assembler for entire nuclear genomes); and VELVET [21].

2.7 Improving the Assembly

There are interesting approaches of how to refine the initial assembly. Such an assembly may suffer from various problems. Many such problems can be due to the presence of repeats. The large inverted repeats often only become evident in reference-guided assembly approaches. "Ordinary" assemblers do not expect a genome to close in a circle, but that is what a chloroplast genome is usually assumed to do (even if the actual conformation may be more complicated [22]). Smaller repeats, often present in the spacers between the genes, are another issue. Introns, especially those in the transfer RNA genes (tRNAs), can be similar to each other and thus may confound assemblers.

Again, [7] provide good guidance, by combining and comparing a reference mapping with two de novo methods and inspection to resolve ambiguities. Sanger-based sequencing of PCR products (predicted on the basis of the initial assembly) is a simple and effective way to confirm the assembly in places where there is doubt. Especially, the junctions between the inverted repeats (IRs) and the single copy regions require such attention, as they often shift a little bit even among closely related species. IRScope is a program designed for supporting this purpose [23].

Gaps in the initial assembly are another issue that is often encountered. There is a special function to deal with gaps called GapCloser in the SOAP package that does not require additional labwork (like in the case of PCR/Sanger sequencing). Such gaps can also be tackled by selecting high-quality flanking reads as seeds for further local assembly. Another approach that is generally recommended is to map back clean reads onto the initial assembly; in this way, inconsistencies will be identified and can be corrected.

2.8 Description of the Assembly

Few of the reviewed methods explicitly describe the coverage of the final genome—i.e., the average number of reads supporting each position (nucleotide). It can be directly obtained from more versatile assemblers like CLC Genomic Workbench or SOAP.

The usual next step is to define the genes and other features along the genome. This process is called annotation. There are three main routes that authors have taken in this respect. The first, and the longest established, is the DOGMA web service for

chloroplast and mitochondria annotation [24]. It works on the basis of GenBank entries, to which it compares the submitted sequence, and returns annotations for common chloroplast genes. While it is still online, the server has "come in the years." A recent message (15 May 2019) on the website says that it is no longer accepting new users, and will be disabled completely soon. DOGMA required users to fine-tune start and stop codons of genes manually. Luckily, there are alternatives; one (the most popular) of them is cpGAVAS [25, 26], which returns a similar full annotation. The third option that is often cited is GENEIOUS, which has a function to transfer annotations from aligned genomes (i.e., from reference genomes to new assemblies). Two other software packages are mentioned, but only twice each: GeSeq [27] and Verdant [28]. However, the GeSeq website lists several other, alternative tools. Many other authors have relied on ("manual") BLAST searches and corrections. In the face of this, the approach by [7] and also [29], to compare the results of several annotation tools, is the gold standard. However, even that requires enough insight so that the correct starts and stops (as well as intron borders) can be selected from those suggested. Overwhelmingly, authors have made use of the tRNAscan-SE software [30] (often in addition to automatic tools) to correctly annotate tRNA genes. An alternative is ARAGORN [31] (two mentions), or again, BLAST searches followed by inspection and correction. It goes without saying that the final annotated sequences must be submitted to one of the nucleotide databases (GenBank, EBI, DDBJ).

Most reports contain a graphical representation of the resulting chloroplast "ring" chromosome; the well-established OGDRAW software/server [32] is by far the most often used for this purpose. Alternatives are GenomeVx [33] and CGView [34]; CIRCOS [35] is a full-blown option for comparing entire nuclear genomes, but apparently also works for the much smaller chloroplast genomes.

Repeats make it more difficult to assemble chloroplast genomes (see above), but they are an essential feature and should be described. The Tandem Repeats Finder [36] identifies direct (tandem) repeats according to various parameters to be set by the user. Ten database entries followed this route. However, as Amiryusefi et al. (2108) [7] have pointed out, redundant repeats which are placed entirely within other repeats and/or duplicated (parts of) tRNA genes should not be counted in the analysis. REPuter [37], of similar age, searches for further repeat types. Most authors who have used Tandem Repeat Finder have also employed REPuter (but not vice versa: 15 methods only employed the latter, which may be sufficient for this purpose). The shortest repeats, those of simple sequences often called "microsatellites," are often discovered with the MISA tool (a Perl script [38]; 23 cases). Once again, overall more authors used MISA than both MISA and REPuter, but a few employed Phobos [39] instead of MISA. The third alternative is MSATCOMMANDER [40].

2.9 Phylogenetic and General Comparison of Chloroplast Genomes

I will only give an overview on methods for phylogenetic analyses of chloroplasts, as these may go beyond the more technical scope of this chapter. mVISTA [41] and MAUVE [42] perform whole-genome alignment visualization. Especially, mVISTA is almost universally used by the authors in the database, most often in the so-called Shuffle-Lagan mode. MAFFT [43], which is constantly being updated [44], performs the alignment of protein coding sequences (amino acids). For calculating key parameters for DNA sequence comparison and evolution, many authors turn to DnaSP [45], which has a similar continuous history of development [46]. MEGA (currently at version 10, [47]) performs similar tasks, and can be used across different computing platforms. Further investigations mentioned more than once are RNA editing (programs PREPACT [48] and PREP-cp [49]). However, the gold standard (again presented by Amiryousefi et al. [7]) is to actually sequence RNA from active chloroplasts and compare these to the genomic sequence. Defective RNA editing may lead to cytonuclear incompatibilities [22], which again can be environment dependent. The same authors [22] also point out that if entire chloroplasts, or entire gene sets, are used in phylogenies, genes or sites under positive selection may blur the picture (and the resulting tree). Codon usage is most often analyzed with the CodonW software [50], and the actual model of nucleotide substitution (the "ease" or probability of the different types of mutation) is often analyzed with jModelTest [51]. Most studies have compared the 70–80 common protein genes for phylogenies. However, even intergenic sequence, if properly aligned and analyzed for the types of mutation present [52], can provide a lot of insight into phylogenetic evolution. RAxML [53] is the clear favorite for maximum likelihood estimations of phylogenic relationships (a new alternative is EasyCodeML [54]); followed by MrBAYES [55] for Bayesian inference. PAUP* [56], which is based on maximum parsimony, is still popular. MEGA can also be used to build trees and assess their significance. SNP detection can be done by SNiPlay [57].

3 Recommendations for Reporting Chloroplast Sequences

The analysis of the database shows that the description of chloroplast DNA sequences is highly formalized—there is often a set of a few choices only for each step, and many authors follow this very similar sequence of steps. These steps can be standardized to a high degree. This will be exemplified here by [7] in a table (Table 1).

General recommendations include describing the exact sources of the material analyzed (species and lower taxonomic ranks), and mentioning the mating system of the species. Highly outbred species (where each individual represents a completely different genotype) are different in this respect from inbred species (where

Table 1
Example of data reporting standards derived from Amiryousefi et al. 2008 [7]

Authors, year	Amiryousefi et al. 2018
Source journal	PLOS ONE
Species/taxa	*Solanum dulcamara*
Species/taxa additional info	
Citation	PLoS ONE 13 (4): e0196069.
DOI	https://doi.org/10.1371/journal.pone.0196069
Purpose	Transcription, correcting annotations
Material	Fresh leaves (plants)
Treatment/storage	
Amount of sample	
DNA extraction method	Modified high-salt protocol of Shi et al. (2012): Shi C, Hu N, Huang H, Gao J, Zhao Y-J, Gao L-Z. An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. PLoS ONE 2012; 7:e31468. https://doi.org/10.1371/journal.pone.0031468 PMID: 22384027; multiply-primed rolling circle amplification (RCA): Atherton RA, McComish BJ, Shepherd LD, Berry LA, Albert NW, Lockhart PJ. Whole genome sequencing of enriched chloroplast DNA using the Illumina GAII platform. Plant Meth. 2010; 6:22; REPLI-g Mini Kit (Qiagen)
Quality checks of DNA	Agarose gel electrophoresis, Qubit; Agilent Technologies 2100 Bioanalyzer using a DNA 1000 chip
Library insert size	300 bp
Library kit/method	IlluminaTruSeq DNA sample prep kit
Sequencing mode	Paired-end
Sequencing length	2 × 150 bp
Sequencing instrument	Illumina MiSeq
Raw/clean data (Gb)	
Raw/clean data (reads)	
Further info on data amounts	
Trimming/filtering procedures	Trimmomatic
Extraction of chloroplast reads (software)	Use of a reference genome/de novo
Assembly software	GENEIOUS v9.1.7, VELVET v1.2.10 (de novo)
Additional steps for assembly	Reference mapping and two de novo methods compared and inspected; sanger-based gap closure and IR junction verification
Coverage	

(continued)

Table 1
(continued)

Annotation	DOGMA, cpGAVAS, VERDANT, GeSeq; inspected and curated all annotation manually; local BLAST searches to confirm the position of CDS; confirmed start and stop codons manually and by comparison to RNAseq; reconfirmed any internal stop codons. Reannotation followed a two-step protocol—Software tools DOGMA to GeSeq
Annotation (additional)	tRNAscan-SE
Further processing	
Drawing	OGDraw v1.2
Repeat analysis (1)	Redundant repeats found entirely within other repeats as well as duplicated parts of tRNAs pruned
Repeat analysis (2)	REPuter; manually inspected output file and located repeats in GENEIOUS (because REPuter overestimates number of repeats)
Repeat analysis (3)	MISA
(Whole-genome) alignment	Sequences aligned, compared, and manually curated (compared to new reference); mVISTA; MAFFT
Analysis of nucleotide variability	
Further examinations	IRscope (expansion and contraction of the inverted repeat IR regions at junction sites examined and plotted)
Codon usage/gene selective pressure analysis	Codon frequency and relative synonymous codon usage (RSCU) calculated on the basis of protein-coding genes using an in-house script; MEGA v7.0.21 (computed overall mean of pairwise distances of 80 protein-coding genes based on Kimura 2-parameter model)
Phylogeny: substitution model testing	jModelTest2
Phylogeny: regions used	35 complete chloroplast genomes
Phylogeny: tree building	RAxML-NG (maximum likelihood ML under three different strategies)
Phylogeny: further details	(1) One of the IR regions removed from all plastid genomes to reduce overrepresentation of duplicated sequences; (2) same data matrix partitioned by gene, exon, intron, and intergenic spacer regions ($n = 258$) and allowed separate base frequencies, α-shape parameters, and evolutionary rates to be estimated for each; (3) inferred best-fitting partitioning strategy with PartitionFinder2 for alignment ($n = 24$)

each individual is essentially genetically identical). The purpose of the study should be made clear—in most cases it will be the description of the new sequence, along with a comparison to related species. Therefore, the most appropriate form of such a report in the future would be a Data Report with its own digital object identifier (DOI). These can be kept simple, provide valuable

data for others to use in their research, and are citable as a digital resource. Many journals now encourage authors to use this format for large datasets. In the past, GenBank entries of chloroplast sequence without corresponding journal articles have made it difficult to assess the significance of, e.g., sequence polymorphisms as compared to new sequence (e.g., [2]).

The nature of the plant material used for the purpose of chloroplast sequencing (e.g., leaves) should be mentioned, as well as the amount and the state/treatment of the material. Next is a summary of the DNA extraction method. This can be done by citation in most cases. Quality checks on the DNA obtained should be done and mentioned. Most methods of sequencing now work with libraries; the sizes of the DNA inserts (and how they were fragmented) should be mentioned, as well as the library preparation protocol. This is often done by service provider laboratories; nevertheless, they should report these methods. The same applies to the sequencing mode (e.g., single-end or paired-end), the actual sequencing machine (instrument), the amount of raw sequence data, and the numbers of reads (as well as their average lengths and length range). After trimming for quality (mentioning the methods to do so), the statistics of cleaned sequence used further downstream should be given.

The strategy to extract chloroplast reads from the total DNA should be given in sufficient detail (e.g., use of a reference genome), and the software used to assemble these reads. The most important settings of the software should be mentioned. If there are additional steps (or several alternative ways) of assembly, they should be described in the same detail. As a result, the general coverage of reads per nucleotide should be given.

Steps for the annotation can also be highly formalized by mentioning the software and its settings, and whether corrections were done manually. It is very important to mention the reference genomes (and time points of access to these) against which the corrections are made, as errors and inconsistencies in database-stored sequences will be propagated easily. Best practice would be to have RNA sequencing data available for independent confirmation of gene translation starts and stops. For annotation of tRNAs, tRNA-SE is still a standard, along with the slightly younger ARAGORN, but there may be more up-to-date tools coming up. This is a general issue – newer software tools may continue to appear, but given that the current ones cope with chloroplast genomes in an efficient and satisfactory way, the "peak" of development in this area is probably past.

Visualization tools for chloroplast genomes all converge on a standard form of representation of single genomes as a ring chromosome. It would be desirable to have more advanced tools that would allow for better graphical comparison of multiple chloroplasts, e.g., in concentric rings. While experienced graphic

designers would have no problems with creating such images from the coordinate data of features in the sequence (see, e.g., the example of the *Populus trichocarpa* chloroplast in the supplementary material to [1] which we did in this way), new stand-alone graphic design software would certainly be widely used.

Repeat analysis is included as a feature in most reports. The well-established software packages are all suitable; what is necessary to be reported are details of the settings used. Repeat structures can be complex and interwoven, especially in the case of introns [58]. Many authors design primers for highly variable microsatellites/simple sequence repeats (SSRs). These are of limited use beyond the species they were designed for. Quite often, the microsatellites will be found in areas for which previously published and tested primers already exist [11].

Phylogenetic studies involving whole chloroplasts vary substantially because of different purposes and data availability. Data reporting standards should include the sequence base (which genes, whether introns or intergenic sequences, are included, etc.) for each step/analysis, the software and settings for alignments, steps in calculation parameters of sequence variability, and whether codon usage and selective pressure was analyzed (and by which strategy, software, and settings). The substitution model should be selected based on a test for genes; there is not yet a good consensus for the choice of an appropriate model for introns and intergenic spacers, however. Tree building is dominated by a few software packages, each specializing on a single approach (maximum likelihood/Bayesian/maximum parsimony).

4 Conclusions

Sequencing of chloroplast genomes is now a quite straightforward exercise; along with its execution, data reporting can be formalized to a high degree. Future investigators are encouraged to report along the recommendations given in this chapter (e.g., as a Data Report or similar, with its own DOI), as this will make the reports highly comparable. This should lead to better oversight of the progress in this field, and to enhanced possibilities for advanced studies with these sequences. It will also help to improve methods for even higher throughput and still better standardization.

References

1. Tuskan GA, DiFazio S, Jansson S, Bohlmann J, Grigoriev I, Hellsten U, Putnam N, Ralph S, Rombauts S, Salamov A, Schein J, Sterck L, Aerts A, Bhalerao RR, Bhalerao RP, Blaudez D, Boerjan W, Brun A, Brunner A, Busov V, Campbell M, Carlson J, Chalot M, Chapman J, Chen G-L, Cooper D, Coutinho PM, Couturier J, Covert S, Cronk Q, Cunningham R, Davis J, Degroeve S, Dejardin A, dePamphilis C, Detter J, Dirks B, Dubchak I, Duplessis S, Ehlting J, Ellis B, Gendler K, Goodstein D, Gribskov M,

Grimwood J, Groover A, Gunter L, Hamberger B, Heinze B, Helariutta Y, Henrissat B, Holligan D, Holt R, Huang W, Islam-Faridi N, Jones S, Jones-Rhoades M, Jorgensen R, Joshi C, Kangasjarvi J, Karlsson J, Kelleher C, Kirkpatrick R, Kirst M, Kohler A, Kalluri U, Larimer F, Leebens-Mack J, Leple J-C, Locascio P, Lou Y, Lucas S, Martin F, Montanini B, Napoli C, Nelson DR, Nelson C, Nieminen K, Nilsson O, Pereda V, Peter G, Philippe R, Pilate G, Poliakov A, Razumovskaya J, Richardson P, Rinaldi C, Ritland K, Rouze P, Ryaboy D, Schmutz J, Schrader J, Segerman B, Shin H, Siddiqui A, Sterky F, Terry A, Tsai C-J, Uberbacher E, Unneberg P, Vahala J, Wall K, Wessler S, Yang G, Yin T, Douglas C, Marra M, Sandberg G, Van de Peer Y, Rokhsar D (2006) The genome of black cottonwood, *Populus trichocarpa* (Torr. & Gray). Science 313(5793):1596–1604

2. Khan A, Khan I, Heinze B, Azim M (2012) The chloroplast genome sequence of date palm (*Phoenix dactylifera* L. cv. 'Aseel'). Plant Mol Biol Report 30(3):666–678
3. Asif H, Khan A, Iqbal A, Khan IA, Heinze B, Azim MK (2013) The chloroplast genome sequence of *Syzygium cumini* (L.) and its relationship with other angiosperms. Tree Genet Genomes 9(3):867–877. https://doi.org/10.1007/s11295-013-0604-1
4. Heinze B, Koziel-Monte A, Jahn D (2014) Analysis of variation in chloroplast DNA sequences. In: Besse P (ed) Molecular plant taxonomy, Methods in molecular biology, vol 1115. Humana Press, pp 85–120
5. Mosca E, Cruz F, Gómez-Garrido J, Bianco L, Rellstab C, Brodbeck S, Csilléry K, Fady B, Fladung M, Fussi B, Gömöry D, González-Martínez SC, Grivet D, Gut M, Hansen OK, Heer K, Kaya Z, Krutovsky KV, Kersten B, Liepelt S, Opgenoorth L, Sperisen C, Ullrich KK, Vendramin GG, Westergren M, Ziegenhagen B, Alioto T, Gugerli F, Heinze B, Höhn M, Troggio M, Neale DB (2019) A reference genome sequence for the European silver fir (*Abies alba* Mill.): a community-generated genomic resource. G3 9(7):2039–2049. https://doi.org/10.1534/g3.119.400083
6. Sabater B (2018) Evolution and function of the chloroplast. Current investigations and perspectives. Int J Mol Sci 19(10):3095
7. Amiryousefi A, Hyvönen J, Poczai P (2018) The chloroplast genome sequence of bittersweet (*Solanum dulcamara*): plastid genome structure evolution in Solanaceae. PLoS One 13(4):e0196069. https://doi.org/10.1371/journal.pone.0196069
8. Shi C, Hu N, Huang H, Gao J, Zhao Y-J, Gao L-Z (2012) An improved chloroplast DNA extraction procedure for whole plastid genome sequencing. PLoS One 7(2):e31468. https://doi.org/10.1371/journal.pone.0031468
9. Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11–15
10. Dellaporta SL, Wood J, Hicks JB (1983) A plant DNA minipreparation: version II. Plant Mol Biol Reporter 1:19–21
11. Heinze B (2007) A database of PCR primers for the chloroplast genomes of higher plants. Plant Methods 3:4. https://doi.org/10.1186/1746-4811-3-4
12. Patel RK, Jain M (2012) NGS QC toolkit: a toolkit for quality control of next generation sequencing data. PLoS One 7(2):e30619. https://doi.org/10.1371/journal.pone.0030619
13. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics (Oxford, England) 30(15):2114–2120. https://doi.org/10.1093/bioinformatics/btu170
14. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17(1):3. https://doi.org/10.14806/ej.17.1.200
15. Jiang M, Chen H, He S, Wang L, Chen AJ, Liu C (2018) Sequencing, characterization, and comparative analyses of the plastome of *Caragana rosea* var. *rosea*. Int J Mol Sci 19(5):1419
16. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y, Tang J, Wu G, Zhang H, Shi Y, Liu Y, Yu C, Wang B, Lu Y, Han C, Cheung DW, Yiu S-M, Peng S, Xiaoqian Z, Liu G, Liao X, Li Y, Yang H, Wang J, Lam T-W, Wang J (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1(1):18–18. https://doi.org/10.1186/2047-217X-1-18
17. Dierckxsens N, Mardulyn P, Smits G (2016) NOVOPlasty: de novo assembly of organelle genomes from whole genome data. Nucleic Acids Res 45(4):e18
18. Jin J-J, Yu W-B, Yang J-B, Song Y, dePamphilis CW, Yi T-S, Li D-Z (2019) GetOrganelle: a fast and versatile toolkit for accurate *de novo* assembly of organelle genomes. bioRxiv:256479. https://doi.org/10.1101/256479

19. Hahn C, Bachmann L, Chevreux B (2013) Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads—a baiting and iterative mapping approach. Nucleic Acids Res 41(13): e129–e129. https://doi.org/10.1093/nar/gkt371
20. Jackman SD, Vandervalk BP, Mohamadi H, Chu J, Yeo S, Hammond SA, Jahesh G, Khan H, Coombe L, Warren RL, Birol I (2017) ABySS 2.0: resource-efficient assembly of large genomes using a bloom filter. Genome Res 27(5):768–777. https://doi.org/10.1101/gr.214346.116
21. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18(5): 821–829. https://doi.org/10.1101/gr.074492.107
22. Bock DG, Andrew RL, Rieseberg LH (2014) On the adaptive value of cytoplasmic genomes in plants. Mol Ecol 23(20):4899–4911. https://doi.org/10.1111/mec.12920
23. Amiryousefi A, Hyvönen J, Poczai P (2018) IRscope: an online program to visualize the junction sites of chloroplast genomes. Bioinformatics 34(17):3030–3031. https://doi.org/10.1093/bioinformatics/bty220
24. Wyman SK, Jansen RK, Boore JL (2004) Automatic annotation of organellar genomes with DOGMA. Bioinformatics 20(17):3252–3255
25. Liu C, Shi L, Zhu Y, Chen H, Zhang J, Lin X, Guan X (2012) CpGAVAS, an integrated web server for the annotation, visualization, analysis, and GenBank submission of completely sequenced chloroplast genome sequences. BMC Genomics 13(1):715
26. Shi L, Chen H, Jiang M, Wang L, Wu X, Huang L, Liu C (2019) CPGAVAS2, an integrated plastome sequence annotator and analyzer. Nucleic Acids Res 47(W1): W65–W73. https://doi.org/10.1093/nar/gkz345
27. Tillich M, Lehwark P, Pellizzer T, Ulbricht-Jones ES, Fischer A, Bock R, Greiner S (2017) GeSeq – versatile and accurate annotation of organelle genomes. Nucleic Acids Res 45(W1):W6–W11. https://doi.org/10.1093/nar/gkx391
28. McKain MR, Hartsock RH, Wohl MM, Kellogg EA (2016) Verdant: automated annotation, alignment and phylogenetic analysis of whole chloroplast genomes. Bioinformatics 33(1):130–132. https://doi.org/10.1093/bioinformatics/btw583
29. Kahraman K, Lucas SJ (2019) Comparison of different annotation tools for characterization of the complete chloroplast genome of *Corylus avellana* cv Tombul. BMC Genomics 20(1): 874. https://doi.org/10.1186/s12864-019-6253-5
30. Lowe TM, Eddy SR (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res 25(5):955–964. https://doi.org/10.1093/nar/25.5.955
31. Laslett D, Canback B (2004) ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. Nucleic Acids Res 32(1):11–16. https://doi.org/10.1093/nar/gkh152
32. Lohse M, Drechsel O, Bock R (2007) OrganellarGenomeDRAW (OGDRAW): a tool for the easy generation of high-quality custom graphical maps of plastid and mitochondrial genomes. Curr Genet 52(5):267–274
33. Conant GC, Wolfe KH (2008) GenomeVx: simple web-based creation of editable circular chromosome maps. Bioinformatics 24(6): 861–862
34. Stothard P, Wishart DS (2005) Circular genome visualization and exploration using CGView. Bioinformatics 21(4):537–539
35. Krzywinski MI, Schein JE, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA (2009) Circos: an information aesthetic for comparative genomics. Genome Res 19(9): 1639–1645. https://doi.org/10.1101/gr.092759.109
36. Benson G (1999) Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res 27(2):573–580. https://doi.org/10.1093/nar/27.2.573
37. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R (2001) REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res 29(22):4633–4642. https://doi.org/10.1093/nar/29.22.4633
38. Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics 33(16): 2583–2585. https://doi.org/10.1093/bioinformatics/btx198
39. Leese F, Mayer C, Held C (2008) Isolation of microsatellites from unknown genomes using known genomes as enrichment templates. Limnol Oceanogr Methods 6(9):412–426. https://doi.org/10.4319/lom.2008.6.412
40. Faircloth BC (2008) Msatcommander: detection of microsatellite repeat arrays and automated, locus-specific primer design. Mol Ecol Resour 8(1):92–94. https://doi.org/10.1111/j.1471-8286.2007.01884.x

41. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I (2004) VISTA: computational tools for comparative genomics. Nucleic Acids Res 32(suppl 2):W273–W279
42. Darling ACE, Mau B, Blattner FR, Perna NT (2004) Mauve: multiple alignment of conserved genomic sequence with rearrangements. Genome Res 14(7):1394–1403
43. Katoh K, Misawa K, Ki K, Miyata T (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res 30(14): 3059–3066. https://doi.org/10.1093/nar/gkf436
44. Rozewicki J, Li S, Amada KM, Standley DM, Katoh K (2019) MAFFT-DASH: integrated protein sequence and structural alignment. Nucleic Acids Res 47(W1):W5–W10. https://doi.org/10.1093/nar/gkz342
45. Rozas J, Rozas R (1995) DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. Bioinformatics 11(6):621–625. https://doi.org/10.1093/bioinformatics/11.6.621
46. Rozas J, Ferrer-Mata A, Sánchez-DelBarrio JC, Guirao-Rico S, Librado P, Ramos-Onsins SE, Sánchez-Gracia A (2017) DnaSP 6: DNA sequence polymorphism analysis of large data sets. Mol Biol Evol 34(12):3299–3302. https://doi.org/10.1093/molbev/msx248
47. Kumar S, Stecher G, Li M, Knyaz C, Tamura K (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. Mol Biol Evol 35(6):1547–1549. https://doi.org/10.1093/molbev/msy096
48. Lenz H, Rüdinger M, Volkmar U, Fischer S, Herres S, Grewe F, Knoop V (2010) Introducing the plant RNA editing prediction and analysis computer tool PREPACT and an update on RNA editing site nomenclature. Curr Genet 56(2):189–201. https://doi.org/10.1007/s00294-009-0283-5
49. Mower JP (2009) The PREP suite: predictive RNA editors for plant mitochondrial genes, chloroplast genes and user-defined alignments. Nucleic Acids Res 37(Web Server issue): W253–W259. https://doi.org/10.1093/nar/gkp337
50. Sharp PM, Bailes E, Grocock RJ, Peden JF, Sockett RE (2005) Variation in the strength of selected codon usage bias among bacteria. Nucleic Acids Res 33(4):1141–1153. https://doi.org/10.1093/nar/gki242
51. Posada D (2008) jModelTest: phylogenetic model averaging. Mol Biol Evol 25(7): 1253–1256. https://doi.org/10.1093/molbev/msn083
52. Lockwood JD, Aleksic JM, Zou J, Wang J, Liu J, Renner SS (2013) A new phylogeny for the genus *Picea* from plastid, mitochondrial, and nuclear sequences. Mol Phylogenet Evol 69(3):717–727
53. Stamatakis A (2014) RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics 30(9): 1312–1313. https://doi.org/10.1093/bioinformatics/btu033
54. Gao F, Chen C, Arab DA, Du Z, He Y, Ho SYW (2019) EasyCodeML: a visual tool for analysis of selection using CodeML. Ecol Evol 9(7):3891–3898. https://doi.org/10.1002/ece3.5015
55. Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17(8):754–755. https://doi.org/10.1093/bioinformatics/17.8.754
56. Swofford D (2002) PAUP*. Phylogenetic analysis using parsimony (*and other methods). Version 4, vol Version 4.0. Sinauer Associates, Sunderland. https://doi.org/10.1111/j.0014-3820.2002.tb00191.x
57. Dereeper A, Nicolas S, Le Cunff L, Bacilieri R, Doligez A, Peros J-P, Ruiz M, This P (2011) SNiPlay: a web-based tool for detection, management and analysis of SNPs. Application to grapevine diversity projects. BMC Bioinformatics 12(1):134
58. Fussi B (2010) Phylogeography, flowering phenology and cytonuclear interactions of Populus alba and P. tremula. Dissertation thesis, University of Vienna, Faculty of Life Sciences, Vienna

Chapter 6

Utility of the Mitochondrial Genome in Plant Taxonomic Studies

Jérôme Duminil and Guillaume Besnard

Abstract

Size, structure, and sequence content lability of plant mitochondrial genome (mtDNA) across species has sharply limited its use in taxonomic studies. Historically, mtDNA variation has been first investigated with RFLPs, while the development of universal primers then allowed studying sequence polymorphisms within short genomic regions (<3 kb). The recent advent of NGS technologies now offers new opportunities by greatly facilitating the assembly of longer mtDNA regions, and even full mitogenomes. Phylogenetic works aiming at comparing signals from different genomic compartments (i.e., nucleus, chloroplast, and mitochondria) have been developed on a few plant lineages, and have been shown especially relevant in groups with contrasted inheritance of organelle genomes. This chapter first reviews the main characteristics of mtDNA and the application offered in taxonomic studies. It then presents tips for best sequencing protocol based on NGS data to be routinely used in mtDNA-based phylogenetic studies.

Key words DNA polymorphism, Genome assembly, Heteroplasmy, Lateral gene transfer (LGT), Mitogenome, Next Generation Sequencing (NGS), Organelle inheritance, Organellar genome, Phylogeny, Phylogeography, Plastid-derived region (*mtpt*)

1 Mitochondrial Genomes

1.1 Origin of Mitochondrial Genomes

The mitochondrial genome originated from a eubacterial ancestor. More specifically, it is now widely accepted that the mitochondria originated from a single endosymbiotic event which involved a α-proteobacteria-like organism and a common cellular ancestor of eukaryotes [1]. This symbiotic relationship between a primitive eukaryote nucleus and an aerobic bacteria—the future mitochondria—has enabled the eukaryote to evolve an aerobic lifestyle. In relation with this new endosymbiotic habit, the "resident" mitochondrial genome has undergone a reductive evolution, characterized, for example, by a loss of coding capacities [2]. The gene content reduction of mitochondrial genomes has been primarily attributable to either gene loss or mitochondria-to-nucleus gene transfers [3]. This process has been interpreted as a consequence of

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_6,

deleterious accumulation in organelle genomes [2], and as a necessity for multicellular organism to keep the function originally coded by organelle genomes. Gene transfer from the mitochondria to the nucleus has been demonstrated to be an ongoing process in plants [4], which explains that the mitochondrial gene content varies across distantly related plant lineages [5, 6].

1.2 Mitochondrial Structure and Genome Size

Land plant mitochondrial genomes (mtDNA or mitogenomes) are usually represented as circular maps (e.g., [7, 8]), yet mtDNA structure is highly variable and should be seen as a complex, dynamic mixture of forms [9–12]. Indeed, plant mtDNA is composed of multiple alternative subgenomic forms (isoforms) that can recombine due to the presence of large repeats. This population of isoforms is thus composed of highly complex structures, linear molecules, open circles of variable size, and supercoiled molecules. Such a structural lability leads to some difficulties for the definition of universal primers and for the full assembly of mitogenome (but see below).

In sharp contrast to the relative small and homogenous mtDNA size in animals (usually between 16 and 20 kb; [13]) and fungi (between 19 and 100 kb; [14]), land plant mitogenome is large and variable in size (between 104 kb in the moss *Anomodon rugelii* and 11.3 Mb in the angiosperm *Silene conica*; [15]). This important size variation can be observed between closely related species [16]. Thus, a comparative study demonstrated that mtDNA size variation within the *Silene* genus might be related to variable mutation rates, with an accumulation of noncoding sequences in mitogenomes presenting higher mutation rates [15]. Angiosperm mtDNA size variation among species is mainly related to differences in the size of noncoding regions, especially large repeats, and alien sequences acquired from intercellular gene transfer and/or interspecific horizontal gene transfer [17, 18]. Plastid-derived (the so-called *mtpt* regions) and nuclear-derived nucleotide sequences represent, respectively, from 1% to 12% and from 0.1% to 13.4% of the mitogenome [17, 19, 20]. Lateral gene transfers (LGT) resulting from mitogenome fusion between distantly related species have been documented, especially in epiphytic and holoparasitic plants [21–23].

1.3 Gene Arrangement and the Importance of Homologous Recombination

Due to the presence of numerous repeated regions and to the putative co-existence of more than one type of mitochondrial genome in a cell (heteroplasmy; [24]), recombinations are frequent within the mtDNA, and gene arrangement (synteny) in higher plants vary enormously [25]. Besides the large size, recombination activity is the most distinctive feature of these genomes [26]. Gene arrangement of mtDNA in higher plants varies enormously due to the presence of repeated regions, source of recombination within and between mtDNA genomes [25]. Cole et al. [27] have

demonstrated that rates of mtDNA rearrangements can be very variable between species from the same genus. Importantly, rearrangements lead to the possibility to generate chimerical genes, potentially involved in some traits of interest, such as the cytoplasmic male sterility [28]. Fortunately, mtDNA coding sequences are highly conserved, facilitating the identification of conserved regions within which universal primers can be defined [29, 30] and that can be easily assembled using next-generation sequencing data.

1.4 Molecular Evolutionary Rates of the mtDNA

In opposition to animals, plant mitochondrial genes evolve very slowly. Comparing silent (synonymous) substitution rates among coding sequences from three genomic compartments in plants [i.e., nuclear DNA (nDNA), chloroplast DNA (cpDNA), and mtDNA], Wolfe et al. [31] have demonstrated that mitochondrial genes evolve three times slower ($0.2–1.1 \times 10^{-9}$ substitutions per synonymous site per year) than chloroplast genes ($1.1–2.9 \times 10^{-9}$ substitutions per synonymous site per year), which in turn evolve two times slower than the nuclear genes (up to 31.5×10^{-9} substitutions per synonymous site per year). These results were further confirmed by Gaut et al. [32] on the comparison of genes from all three genomes between maize and rice. Interestingly, as outlined by Muse [33], the similarity obtained between Wolfe and Gaut studies, albeit different levels of evolutionary divergence were addressed, might indicate that plant nucleotide substitution features have been constant over higher plant evolution. This is somewhat nuanced by Drouin et al. [34], who, based on the comparison of 12 genes in 27 seed plant species, demonstrated that the overall relative rate of synonymous substitutions of mitochondrial, chloroplast, and nuclear genes is 1:3:10 if averaged across studied seed plants, 1:2:4 in gymnosperms, 1:3:16 in angiosperms, and that they go up to 1:3:20 in basal angiosperms. Though this low molecular evolutionary rate of mitochondrial genes appeared to concern most of plant species, some exceptions were demonstrated (e.g., within *Pelargonium*, *Plantago*, *Silene*; [15, 35, 36]). The generality of slow synonymous sequence evolution in mitogenomes has been investigated across a large and taxonomically widely distributed set of seed plants [37]. According to this study, earlier findings were confirmed for roughly 80–90% of the studied species, indicating that a surprising number of taxa depart from this common pattern by presenting either an accelerated or a slower synonymous substitution rate. Moreover, Mower et al. [37] demonstrated that both patterns of faster and slower evolutionary rates can be found within the same species at different genes supporting the idea that all genes evolve independently from one another. Albeit this observation might be related to different artifacts (see the discussion in [37]), independent evidences for mutation rate variation among genes were acquired [27, 38]. Therefore, the general idea remains

that mitochondrial genes evolve at a slow rate, and that mtDNA polymorphism is very low within one species and even between closely related species. This explains the limited use of mtDNA in phylogeography and phylogeny, though the demonstration of molecular rate heterogeneity within some plant lineages [37, 38] might support the idea that it is worth investigating if this pattern holds true for a given species.

1.5 Mode of Inheritance of the Mitochondrial Genome

Mitochondrial genomes are generally uniparentally inherited (usually maternally) in seed plants, though some species have been shown to present a paternal (some coniferous species) or a biparental inheritance [39]. Uniparental inheritance of organelle genomes is more and more seen as an evolutionarily unstable trait [40]. The uniparental inheritance of organellar genomes, together with slow molecular evolutionary rates, explain their success as molecular markers in phylogeography studies (reviewed in [41]). Mode of inheritance has been shown theoretically and experimentally to have a major effect on the estimation of the among-population genetic differentiation: maternally inherited genomes generally experience more subdivision than paternally or biparentally inherited ones [42]. Thus, in conifers, G_{ST} is almost always larger at mtDNA markers than at cpDNA markers, while it is nearly similar at both markers in angiosperms, where both are generally maternally inherited [42].

2 Mitochondrial Molecular Markers in Phylogenetics and Taxonomy

2.1 Genomic Resources: Complete Mitochondrial Genome

The first land plant complete mitogenome was obtained for the liverwort *Marchantia polymorpha* [43]. The number of plant species whose complete mtDNA sequence is available is now 221 (Fig. 1). In comparison, 4020 plant species were completely sequenced for their chloroplast genome (data compiled in October 2019 according to https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles/).

2.2 Use of mtDNA in Phylogeography and Phylogenetics

Due to the supposed absence of recombination within the cpDNA molecule and its slightly faster evolutionary rate, cpDNA-derived molecular markers were more popular in phylogeography and phylogenetic studies than mtDNA-derived ones. However, the acquisition of mtDNA data can also be very interesting to characterize species evolutionary history, notably in addition to and comparison with cpDNA data [44–50].

Historically, mtDNA variation was first evaluated with the Restriction Fragments Length Polymorphisms technique (RFLPs; [51, 52]). This approach allowed revealing large rearrangements that were particularly useful to investigate linkage disequilibrium between chloroplast and mitochondrial polymorphisms within

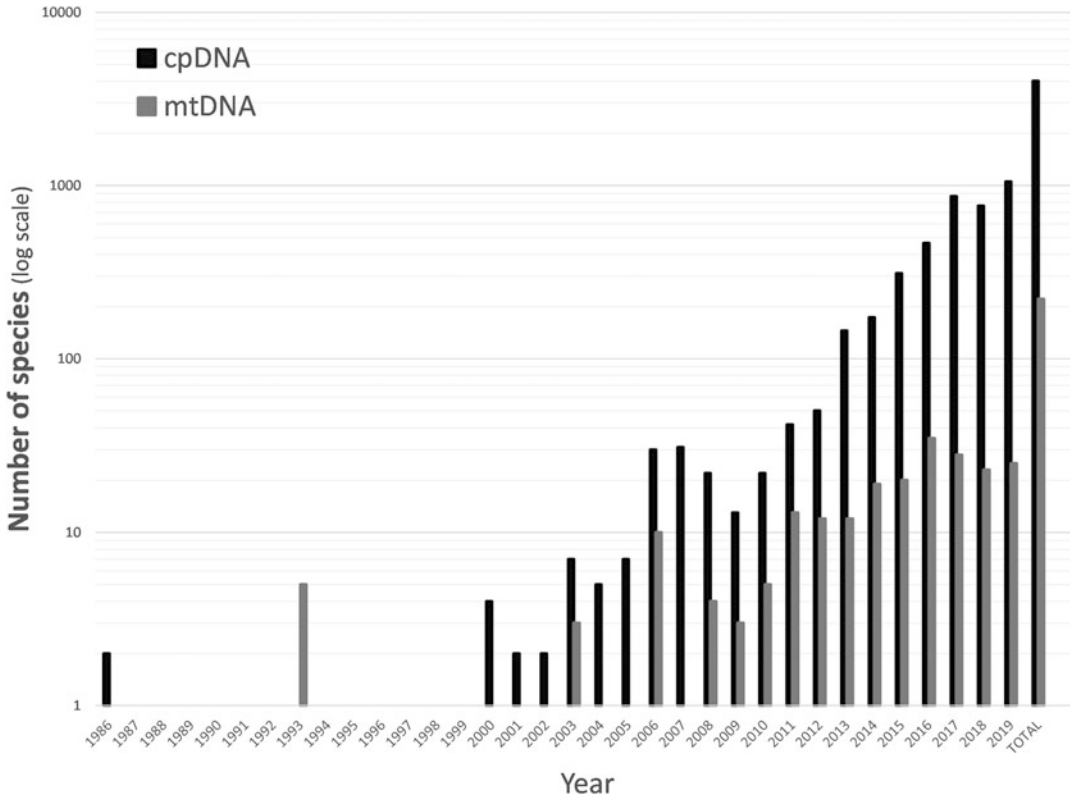


Fig. 1 Number of full mtDNA and cpDNA sequences published per year for plant species

some species due to the common maternal inheritance of cytoplasmic organelles [53, 54]. With the advent of PCR methods, the amplification of mtDNA loci with universal primers became more popular [29, 30, 55–58]. Universal primers were defined in conserved regions (exons) and were used to amplify mtDNA regions with conserved micro-synteny. Polymorphism in amplified fragment has been revealed with various methods: PCR-RFLP [59–63], RFLP-SSR [57, 60, 64], mtDNA-SSR [46], the variable number of tandem repeats (VNTR) in minisatellite regions [65–67], and finally in DNA sequences [68–72]. The choice of the candidate loci was limited and depended on the taxonomic level addressed by the phylogenetic study. At the lowest taxonomic level (intraspecific or among closely related species), intergenic or intronic sequences were particularly interesting. Instead, at higher taxonomic levels, polymorphism from coding sequences was generally used. Based on such approaches, only a few studies have combined cpDNA and mtDNA polymorphisms to reconstruct the phylogeography of species (whereas cpDNA has remained the most frequently used marker). Yet, some contrasted phylogeographic patterns have been revealed in some species. This was particularly true in conifers, in which cpDNA and mtDNA can be transmitted by different

parents, providing complementary information on species pollen- and seed-mediated gene flow [47, 48].

We have now entered the high-throughput sequencing area. This offers new opportunities for the use of mtDNA in phylogeography and phylogenetics. Given the low polymorphism nature of mtDNA, acquiring long mtDNA fragments, or even full mtDNA genomes allows capturing useful genomic variations. However, these new technologies also bring along new challenges, notably in terms of mtDNA assembly and comparison between species. As mentioned above, the mtDNA assembly is complicated by the presence of numerous short and long repeated fragments (some reaching more than 10 kb), as well as exogenous fragments (intercellular gene transfers and/or LGTs; [17]). Assembly of nonrepeated sequences is feasible on relatively long contigs (>10 kb; [73, 74]), but the integration of all fragments in a master chromosome can be challenging [75]. The combined use of long reads (Oxford Nanopore Technologies) with short reads (Illumina technologies) can help mtDNA assembly [12, 76], with the possibility to observe recombination in long repeated regions (alternative conformation of mitogenomes). The parallel reconstruction of plastid and mitochondrial genomes is also necessary to resolve the assembly of *mtpt* regions; by applying a step-by-step approach, it was possible to reconstruct the master chromosomes in Oleaceae even on very fragmented DNA from old herbarium specimens [50, 75]. All parts of mitogenome are, however, not informative for phylogeographic or taxonomical studies, since some regions are not shared between species, even at the genus level, whereas some homologous regions are not necessarily orthologous (because they could be recurrently transferred, in particular from the plastome). As a consequence, mtDNA phylogenies should be reconstructed with the pan-mitogenome or the core fragments (shared by all mitogenomes; e.g., Wang et al., [76]), and thus focus on regions with functional genes (i.e., exons and introns). Using this approach, the comparison of phylogenetic topologies obtained with cpDNA and mtDNA have shown subtle differences in several groups [27, 50, 77, 78], but it can also demonstrate strong informative incongruences [45, 49]. At the species level, the use of complete mitogenomes could be possible, but beforehand, orthology of *mtpt* regions has to be verified (by testing, for each *mtpt*, phylogenetic clustering of accessions of the same species compared to other genera). Such a strategy has been applied on the olive tree, and allowed resolving the phylogeny of maternally inherited genome, which was not possible with the plastome only [50]. Overall, at this taxonomical level, more information was recovered from the whole mtDNA (>0.6 Mb) than from the plastome (ca. 0.15 Mb).

3 Tips for Sequencing Protocol Based on NGS

This chapter finally aims at providing tips for a good sequencing protocol based on NGS data to assemble mtDNA sequences for phylogenetic studies (see [79] for methods based on a PCR approach). The protocol is defined to sequence conserved mtDNA regions (i.e., parts of the pan-mitogenome) among distantly related species. The approach is relatively simple and is based on shot-gun sequencing of total genomic DNA (the so-called "genome skimming" approach; [80]). Considering the high number of cytoplasmic organites in a cell, organelle DNA is expected to be highly represented in such data (ca. 5–10% of total genomic reads). These data can thus be used for the assembly of different genomic regions from both the nuclear genome (especially the ribosomal DNA cluster that is highly repeated) and the organellar genomes (e.g., [50, 74, 77]).

3.1 DNA Purification

For studied samples, total genomic DNA has to be extracted with an appropriate protocol, that allows recovering a relatively clean extract with enough double-stranded DNA (at least 50 ng). For instance, the BioSprint 15 DNA Plant Kit (Qiagen Inc.) has been successfully used for distinct plant groups, including relatively old museum specimens (e.g., [81]). With this method, each leaf sample needs to be ground in a 2-mL tube containing three tungsten beads with a TissueLyser (Qiagen Inc., Texas) before starting the DNA purification procedure. Double-stranded DNA concentration of final extracts is then quantified.

DNA quantification can be done on an agarose gel (if DNA is not degraded) or by absorbance measurement. When using quantification on agarose gel, PCR products can be quantified, respectively, to a standard DNA ladder. In this aim, PCR products as well as the DNA ladder have to be analyzed on an agarose gel electrophoresis system. Most ladders have a standard band that corresponds to a standard amount of DNA per μL. The PCR product concentration is roughly quantified according to the intensity of the standard. If using quantification by absorbance measurement, the PCR product concentration can be accurately quantified using a microvolume spectrophotometer (e.g., PicoGreen or Qubit, ThermoFisher). Spectrophotometer-based quantification is more accurate than gel-based quantification.

3.2 Construction of Libraries and Sequencing

Between 50 and 500 ng of double-stranded DNA are usually used to construct sequencing libraries with a kit (e.g., TruSeq DNA Sample, Illumina) or following a home-made, well-established procedure such as the one described in the supplemental online material of Mariac et al. [82]. For herbarium specimens, DNA libraries can be generated without prior DNA sonication because the DNA

is supposedly moderately to highly degraded (e.g., [81]). A pool of 24 to 96 libraries can be bulked in equimolar concentrations before sequencing.

Each sample is then paired-end sequenced (usually reads of 150 bp) on a sequencer lane (e.g., HiSeq or NovaSeq, Illumina). Bridge amplification is performed to generate clusters, and paired-end reads are collected on the sequencer.

4 Recommendations for Bioinformatics Analyses

Before starting the assembly, duplicated reads have to be removed and overlapping paired-end reads can be merged. Because no automated approach of full plant mitogenome assembly based on short-read data is currently available, we recommend to focus on the conserved mtDNA regions and map reads on a reference that has been previously defined on complete mitogenome. For instance, in the Oleaceae family, 36 protein-coding genes and 16 introns, for a total of ca. 55 kb) have been targeted [50]. It is better using as reference genes from a complete mitogenome that is closely phylogenetically related to your model species (the list can be found here: https://www.ncbi.nlm.nih.gov/genome/browse#!/organelles/). The merging of overlapping paired-end reads can be done using BBMERGE [BBTOOLS] as implemented in GENEIOUS v. 9.0.5 [83]. Overlapping paired-end reads ("merged reads") and non-overlapping paired-reads ("unmerged reads") are then used for the mitogenome assembly as described in Van de Paer et al. [75]. We recommend to check the quality of the mapping and the homogeneity of the sequencing depth to detect any chimeric genes or duplication/deletion among the targeted regions.

References

1. Gray MW, Burger G, Franz Lang B (2001) The origin and early evolution of mitochondria. Genome Biol 2:reviews1018. https://doi.org/10.1186/gb-2001-2-6-reviews1018
2. Andersson SGE, Kurland CG (1998) Reductive evolution of resident genomes. Trends Microbiol 6:263–268. https://doi.org/10.1016/s0966-842x(98)01312-2
3. Palmer JD, Adams KL, Cho Y et al (2000) Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proc Natl Acad Sci U S A 97:6960–6966. https://doi.org/10.1073/pnas.97.13.6960
4. Adams KL, Daley DO, Qiu YL et al (2000) Repeated, recent and diverse transfers of a mitochondrial gene to the nucleus in flowering plants. Nature 408:354–357
5. Adams KL, Qiu YL, Stoutemyer M et al (2002) Punctuated evolution of mitochondrial gene content: high and variable rates of mitochondrial gene loss and transfer to the nucleus during angiosperm evolution. Proc Natl Acad Sci U S A 99:9905–9912. https://doi.org/10.1073/pnas.042694899
6. Adams K, Palmer JD (2003) Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol Phylogenet Evol 29:380–395. https://doi.org/10.1016/S1055-7903(03)00194-5
7. Lonsdale DM, Hodge TP, Fauron CMR (1984) The physical map and organisation of

the mitochondrial genome from the fertile cytoplasm of maize. Nucleic Acids Res 12:9249–9261. https://doi.org/10.1093/nar/12.24.9249

8. Palmer JD, Shields CR (1984) Tripartite structure of the *Brassica campestris* mitochondrial genome. Nature 3074:437–440. https://doi.org/10.1038/307437a0
9. Palmer JD, Herbon LA (1988) Plant mitochondrial DNA evolves rapidly in structure, but slowly in sequence. J Mol Evol 28:87–97. https://doi.org/10.1007/bf02143500
10. Backert S, Nielsen BL, Börner T (1997) The mystery of the rings: structure and replication of mitochondrial genomes from higher plants. Trends Plant Sci 2:477–483. https://doi.org/10.1016/S1360-1385(97)01148-5
11. Morley SA, Nielsen BL (2017) Plant mitochondrial DNA. Front Biosci 22:1023–1032. https://doi.org/10.2741/4531
12. Kozik A, Rowan BA, Lavelle D et al (2019) The alternative reality of plant mitochondrial DNA: one ring does not rule them all. PLoS Genet 15:e1008373. https://doi.org/10.1371/journal.pgen.1008373
13. Boore JL (1999) Animal mitochondrial genomes. Nucleic Acids Res 27:1767–1780. https://doi.org/10.1093/nar/27.8.1767
14. Bullerwell CE, Gray MW (2004) Evolution of the mitochondrial genome: protist connections to animals, fungi and plants. Curr Opin Microbiol 7:528–534. https://doi.org/10.1016/j.mib.2004.08.008
15. Sloan DB, Alverson AJ, Chuckalovcak JP et al (2012) Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. PLoS Biol 10:e1001241. https://doi.org/10.1371/journal.pbio.1001241
16. Alverson AJ, Wei X, Rice DW et al (2010) Insights into the evolution of mitochondrial genome size from complete sequences of *Citrullus lanatus* and *Cucurbita pepo* (Cucurbitaceae). Mol Biol Evol 27:1436–1448. https://doi.org/10.1093/molbev/msq029
17. Marienfeld J, Unseld M, Brennicke A (1999) The mitochondrial genome of *Arabidopsis* is composed of both native and immigrant information. Trends Plant Sci 4:495–502. https://doi.org/10.1016/S1360-1385(99)01502-2
18. Choi IS, Schwarz EN, Ruhlman TA et al (2019) Fluctuations in Fabaceae mitochondrial genome size and content are both ancient and recent. BMC Plant Biol 19:448. https://doi.org/10.1186/s12870-019-2064-8
19. Kubo T, Newton KJ (2008) Angiosperm mitochondrial genomes and mutations. Mitochondrion 8:5–14. https://doi.org/10.1016/j.mito.2007.10.006
20. Wang D, Wu YW, Shih ACC et al (2007) Transfer of chloroplast genomic DNA to mitochondrial genome occurred at least 300 Mya. Mol Biol Evol 24:2040–2048. https://doi.org/10.1093/molbev/msm133
21. Rice DW, Alverson AJ, Richarson AO et al (2013) Horizontal transfer of entire genomes via mitochondrial fusion in the angiosperm *Amborella*. Science 342:1468–1473. https://doi.org/10.1126/science.1246275
22. Gandini CL, Sanchez-Puerta MV (2017) Foreign plastid sequences in plant mitochondria are frequently acquired via mitochondrion-to-mitochondrion horizontal transfer. Sci Rep 7:43402. https://doi.org/10.1038/srep43402
23. Sanchez-Puerta M, García LE, Wohlfeiler J et al (2017) Unparalleled replacement of native mitochondrial genes by foreign homologs in a holoparasitic plant. New Phytol 214:376–387. https://doi.org/10.1111/nph.14361
24. Kmiec B, Woloszynska M, Janska H (2006) Heteroplasmy as a common state of mitochondrial genetic information in plants and animals. Curr Genet 50:149–159. https://doi.org/10.1007/s00294-006-0082-1
25. Schuster W, Brennicke A (1994) The plant mitochondrial genome: physical structure, information content, RNA editing, and gene migration to the nucleus. Annu Rev Plant Physiol Plant Mol Biol 45:61–78. https://doi.org/10.1146/annurev.pp.45.060194.000425
26. Woloszynska M (2010) Heteroplasmy and stoichiometric complexity of plant mitochondrial genomes-though this be madness, yet there's method in't. J Exp Bot 61:657–671. https://doi.org/10.1093/jxb/erp361
27. Cole LW, Guo W, Mower JP et al (2018) High and variable rates of repeat-mediated mitochondrial genome rearrangement in a genus of plants. Mol Biol Evol 35:2773–2785. https://doi.org/10.1093/molbev/msy176
28. Schnable PS, Wise RP (1998) The molecular basis of cytoplasmic male sterility and fertility restoration. Trends Plant Sci 3:175–180. https://doi.org/10.1016/S1360-1385(98)01235-7
29. Demesure B, Sodzi N, Petit RJ (1995) A set of universal primers for amplification of polymorphic non-coding regions of mitochondrial and chloroplast DNA in plants. Mol Ecol 4:129–131. https://doi.org/10.1111/j.1365-294x.1995.tb00201.x
30. Duminil J, Pemonge MH, Petit RJ (2002) A set of 35 consensus primer pairs amplifying genes and introns of plant mitochondrial

DNA. Mol Ecol Notes 2:428–430. https://doi.org/10.1046/j.1471-8286.2002.00263.x

31. Wolfe KH, Li WH, Sharp PM (1987) Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A 84:9054–9058. https://doi.org/10.1073/pnas.84.24.9054
32. Gaut BS, Morton BR, McCaig BC et al (1996) Substitution rate comparisons between grasses and palms: synonymous rate differences at the nuclear gene *Adh* parallel rate differences at the plastid gene *rbcL*. Proc Natl Acad Sci U S A 93:10274–10279. https://doi.org/10.1073/pnas.93.19.10274
33. Muse SV (2000) Examining rates and patterns of nucleotide substitution in plants. Plant Mol Biol 42:25–43. https://doi.org/10.1023/A:1006319803002
34. Drouin G, Daoud H, Xia J (2008) Relative rates of synonymous substitutions in the mitochondrial, chloroplast and nuclear genomes of seed plants. Mol Phylogenet Evol 49:827–831. https://doi.org/10.1016/J.YMPEV.2008.09.009
35. Cho Y, Mower JP, Qiu YL et al (2004) Mitochondrial substitution rates are extraordinarily elevated and variable in a genus of flowering plants. Proc Natl Acad Sci U S A 101:17741–17746. https://doi.org/10.1073/pnas.0408302101
36. Parkinson CL, Mower JP, Qiu YL et al (2005) Multiple major increases and decreases in mitochondrial substitution rates in the plant family Geraniaceae. BMC Evol Biol 5:73. https://doi.org/10.1186/1471-2148-5-73
37. Mower JP, Touzet P, Gummow JS et al (2007) Extensive variation in synonymous substitution rates in mitochondrial genes of seed plants. BMC Evol Biol 7:135. https://doi.org/10.1186/1471-2148-7-135
38. Barr CM, Keller SR, Ingvarsson PK et al (2007) Variation in mutation rate and polymorphism among mitochondrial genes of *Silene vulgaris*. Mol Biol Evol 24:1783–1791. https://doi.org/10.1093/molbev/msm106
39. Birky CW Jr (2001) The inheritance of genes in mitochondria and chloroplasts: laws, mechanisms, and models. Annu Rev Genet 35:125–148. https://doi.org/10.1146/annurev.genet.35.102401.090231
40. Greiner S, Sobanski J, Bock R (2015) Why are most organelle genomes transmitted maternally? BioEssays 37:80–94. https://doi.org/10.1002/bies.201400110
41. Petit RJ, Vendramin GG (2007) Plant phylogeography based on organelle genes: an introduction. In: Weiss S, Ferrand N (eds) Phylogeography of Southern Europe Refugia. Springer, Dordrecht, pp 23–101. https://doi.org/10.1007/1-4020-4904-8_2
42. Petit RJ, Duminil J, Fineschi S et al (2005) Comparative organization of chloroplast, mitochondrial and nuclear diversity in plant populations. Mol Ecol 14:689–701. https://doi.org/10.1111/j.1365-294X.2004.02410.x
43. Oda K, Yamato K, Ohta E et al (1992) Gene organization deduced from the complete sequence of liverwort *Marchantia polymorpha* mitochondrial DNA. A primitive form of plant mitochondrial genome. J Mol Biol 223:1–7. https://doi.org/10.1016/0022-2836(92)90708-r
44. Dumolin-Lapegue S, Kremer A, Petit RJ (1999) Are chloroplast and mitochondrial DNA variation species independent in oaks? Evolution 53:1406–1413. https://doi.org/10.1111/j.1558-5646.1999.tb05405.x
45. Govindarajulu R, Parks M, Tennessen JA et al (2015) Comparison of nuclear, plastid, and mitochondrial phylogenies and the origin of wild octoploid strawberry species. Am J Bot 102:544–554. https://doi.org/10.3732/ajb.1500026
46. Hosaka K, Sanetomo R (2009) Comparative differentiation in mitochondrial and chloroplast DNA among cultivated potatoes and closely related wild species. Genes Genet Syst 84:371–378. https://doi.org/10.1266/ggs.84.371
47. Jaramillo-Correa JP, Beaulieu J, Ledig FT et al (2006) Decoupled mitochondrial and chloroplast DNA population structure reveals Holocene collapse and population isolation in a threatened Mexican-endemic conifer. Mol Ecol 15:2787–2800. https://doi.org/10.1111/j.1365-294X.2006.02974.x
48. Liepelt S, Bialozyt R, Ziegenhagen B (2002) Wind-dispersed pollen mediates postglacial gene flow among refugia. Proc Natl Acad Sci U S A 99:14590–14594. https://doi.org/10.1073/pnas.212285399
49. Rydin C, Wikström N, Bremer B (2017) Conflicting results from mitochondrial genomic data challenge current views of Rubiaceae phylogeny. Am J Bot 104:1522–1532. https://doi.org/10.3732/ajb.1700255
50. Van de Paer C, Bouchez O, Besnard G (2018) Prospects on the evolutionary mitogenomics of plants: a case study on the olive family

(Oleaceae). Mol Ecol Resour 18:407–423. https://doi.org/10.1111/1755-0998.12742

51. Levings CS III, Pring DR (1976) Restriction endonuclease analysis of mitochondrial DNA from normal and Texas cytoplasmic male-sterile maize. Science 193:158–160. https://doi.org/10.1126/science.193.4248.158
52. Quetier F, Vedel F (1977) Heterogenous population of mitochondrial DNA molecules in higher plants. Nature 268:365–368. https://doi.org/10.1038/268365a0
53. Desplanque B, Viard F, Bernard J et al (2000) The linkage disequilibrium between chloroplast DNA and mitochondrial DNA haplotypes in *Beta vulgaris* ssp. *maritima* (L.): the usefulness of both genomes for population genetic studies. Mol Ecol 9:141–154. https://doi.org/10.1046/j.1365-294x.2000.00843.x
54. Besnard G, Khadari B, Baradat P et al (2002) Combination of chloroplast and mitochondrial DNA polymorphisms to study cytoplasm genetic differentiation in the olive complex (*Olea europaea* L.). Theor Appl Genet 105:139–144. https://doi.org/10.1007/s00122-002-0868-6
55. Dumolin-Lapegue S, Pemonge MH, Petit RJ (1997) An enlarged set of consensus primers for the study of organelle DNA in plants. Mol Ecol 6:393–397. https://doi.org/10.1046/j.1365-294x.1997.00193.x
56. Froelicher Y, Mouhaya W, Bassene JB et al (2011) New universal mitochondrial PCR markers reveal new information on maternal citrus phylogeny. Tree Genet Genomes 7:49–61. https://doi.org/10.1007/s11295-010-0314-x
57. Jaramillo-Correa JP, Bousquet J, Beaulieu J et al (2003) Cross-species amplification of mitochondrial DNA sequence-tagged-site markers in conifers: the nature of polymorphism and variation within and among species in *Picea*. Theor Appl Genet 106:1353–1367. https://doi.org/10.1007/s00122-002-1174-z
58. Jeandroz S, Bastien D, Chandelier A et al (2002) A set of primers for amplification of mitochondrial DNA in *Picea abies* and other conifer species. Mol Ecol Notes 2:389–392. https://doi.org/10.1046/j.1471-8286.2002.00271.x
59. Boonruangrod R, Desai D, Fluch S et al (2008) Identification of cytoplasmic ancestor genepools of *Musa acuminata* Colla and *Musa balbisiana* Colla and their hybrids by chloroplast and mitochondrial haplotyping. Theor Appl Genet 118:43–55. https://doi.org/10.1007/s00122-008-0875-3
60. Godbout J, Jaramillo-Correa JP, Beaulieu J et al (2005) A mitochondrial DNA minisatellite reveals the postglacial history of jack pine (*Pinus banksiana*), a broad-range North American conifer. Mol Ecol 14:3497–3512. https://doi.org/10.1111/j.1365-294X.2005.02674.x
61. San Jose-Maldia L, Uchida K, Tomaru N (2009) Mitochondrial DNA variation in natural populations of Japanese larch (*Larix kaempferi*). Silvae Genet 58:234–241. https://doi.org/10.1515/sg-2009-0030
62. Moriguchi Y, Kang KS, Lee KY et al (2009) Genetic variation of *Picea jezoensis* populations in South Korea revealed by chloroplast, mitochondrial and nuclear DNA markers. J Plant Res 122:153–160. https://doi.org/10.1007/s10265-008-0210-8
63. Naydenov K, Senneville S, Beaulieu J et al (2007) Glacial vicariance in Eurasia: mitochondrial DNA evidence from Scots pine for a complex heritage involving genetically distinct refugia at mid-northern latitudes and in Asia Minor. BMC Evol Biol 7:233. https://doi.org/10.1186/1471-2148-7-233
64. Burban C, Petit RJ (2003) Phylogeography of maritime pine inferred with organelle markers having contrasted inheritance. Mol Ecol 12:1487–1495. https://doi.org/10.1046/j.1365-294x.2003.01817.x
65. Bastien D, Favre JM, Collignon AM et al (2003) Characterization of a mosaic minisatellite locus in the mitochondrial DNA of Norway spruce [*Picea abies* (L.) Karst.]. Theor Appl Genet 107:574–580. https://doi.org/10.1007/s00122-003-1284-2
66. Honma Y, Yoshida Y, Terachi T et al (2011) Polymorphic minisatellites in the mitochondrial DNAs of *Oryza* and *Brassica*. Curr Genet 57:261–270. https://doi.org/10.1007/s00294-011-0345-3
67. Yoshida Y, Matsunaga M, Cheng D et al (2012) Mitochondrial minisatellite polymorphisms in fodder and sugar beets reveal genetic bottlenecks associated with domestication. Biol Plant 56:369. https://doi.org/10.1007/s10535-012-0101-7
68. Avtzis DN, Aravanopoulos FA (2011) Host tree and insect genetic diversity on the borderline of natural distribution: a case study of *Picea abies* and *Pityogenes chalcographus* (Coleoptera, Scolytinae) in Greece. Silva Fenn 45:157–164. https://doi.org/10.14214/sf.37
69. Eckert AJ, Tearse BR, Hall BD (2008) A phylogeographical analysis of the range disjunction for foxtail pine (*Pinus balfouriana*, Pinaceae): the role of Pleistocene glaciation. Mol Ecol 17:1983–1997. https://doi.org/10.1111/j.1365-294X.2008.03722.x
70. Edwards EJ, Nyffeler R, Donoghue MJ (2005) Basal cactus phylogeny: implications of *Pereskia* (Cactaceae) paraphyly for the transition to the

cactus life form. Am J Bot 92:1177–1188. https://doi.org/10.3732/ajb.92.7.1177

71. Goodall-Copestake WP, Pérez-Espona S, Harris DJ et al (2010) The early evolution of the mega-diverse genus *Begonia* (Begoniaceae) inferred from organelle DNA phylogenies. Biol J Linn Soc 101:243–250. https://doi.org/10.1111/j.1095-8312.2010.01489.x
72. Gugger PF, Gonzalez-Rodriguez A, Rodriguez-Correa H et al (2011) Southward Pleistocene migration of Douglas-fir into Mexico: phylogeography, ecological niche modeling, and conservation of "rear edge" populations. New Phytol 189:1185–1199. https://doi.org/10.1111/j.1469-8137.2010.03559.x
73. Donnelly K, Cottrell J, Ennos RA et al (2017) Reconstructing the plant mitochondrial genome for marker discovery: a case study using Pinus. Mol Ecol Resour 17:943–954. https://doi.org/10.1111/1755-0998.12646
74. Malé PJG, Bardon L, Besnard G et al (2014) Genome skimming by shotgun sequencing helps resolve the phylogeny of a pantropical tree family. Mol Ecol Resour 14:966–975. https://doi.org/10.1111/1755-0998.12246
75. Van de Paer C, Hong-Wa C, Jeziorski C et al (2016) Mitogenomics of *Hesperelaea*, an extinct genus of Oleaceae. Gene 594:197–202. https://doi.org/10.1016/J.GENE.2016.09.007
76. Wang S, Song Q, Li S et al (2018) Assembly of a complete mitogenome of *Chrysanthemum nankingense* using Oxford Nanopore long reads and the diversity and evolution of Asteraceae mitogenomes. Genes 9:547. https://doi.org/10.3390/genes9110547
77. Fonseca LHM, Lohmann LG (2020) Exploring the potential of nuclear and mitochondrial sequencing data generated through genome-skimming for plant phylogenetics: a case study from a clade of neotropical lianas. J Syst Evol 58:18–32. https://doi.org/10.1111/jse.12533
78. Wang X, Cheng F, Rohlsen D et al (2018) Organellar genome assembly methods and comparative analysis of horticultural plants. Hortic Res 5:3. https://doi.org/10.1038/s41438-017-0002-1
79. Duminil J (2014) Mitochondrial genome and plant taxonomy. In: Besse P (ed) Molecular plant taxonomy. Methods in molecular biology (Methods and protocols), vol 1115. Humana Press, Totowa, NJ, pp 121–140. https://doi.org/10.1007/978-1-62703-767-9_6
80. Straub SCK, Parks M, Weitemier K et al (2012) Navigating the tip of the genomic iceberg: next-generation sequencing for plant systematics. Am J Bot 99:349–364. https://doi.org/10.3732/ajb.1100335
81. Zedane L, Hong-Wa C, Murienne J et al (2016) Museomics illuminate the history of an extinct, paleoendemic plant lineage (*Hesperelaea*, Oleaceae) known from an 1875 collection from Guadalupe Island, Mexico. Biol J Linn Soc 117:44–57. https://doi.org/10.1111/bij.12509
82. Mariac C, Scarcelli N, Pouzadou J et al (2014) Cost-effective enrichment hybridization capture of chloroplast genomes at deep multiplexing levels for population genetics and phylogeography studies. Mol Ecol Resour 14:1103–1113. https://doi.org/10.1111/1755-0998.12258
83. Kearse M, Moir R, Wilson A et al (2012) Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics 28:1647–1649. https://doi.org/10.1093/bioinformatics/bts199

Chapter 7

Nuclear Ribosomal RNA Genes: ITS Region

Pascale Besse

Abstract

Despite possible drawbacks (intraspecific polymorphisms and possible fungal contamination), sequencing of the ITS region of the ribosomal RNA genes remains one of the most popular nuclear sequences used for plant taxonomy and phylogeny. A protocol for PCR amplification and sequencing of this region using universal plant primers is provided.

Key words Ribosomal DNA, ITS, Sanger sequencing, PCR

1 Introduction

Since early reviews [1] and the general agreement around the necessity to use biparentally inherited nuclear markers together with monoparentally inherited ones such as chloroplast or mitochondrial DNA (Chapter 2), nuclear ribosomal RNA genes (nrDNA) have received increasing attention in plant taxonomy and phylogeny. One of the reasons is that such genes provide significant information in phylogenetic research because they are composed of different regions (both coding and noncoding) that are conserved differently and thus provide information at different taxonomic levels [2] (*see* Chapter 2). In particular, spacer regions of nrDNA are useful for plant systematics from species to generic levels [3]. Another related reason for such popularity is the easy PCR amplification of this region, provided by PCR primers designed in conserved coding regions surrounding a more variable spacer region. Ribosomal genes are arranged in tandem repeats and are subjected to concerted evolution, which results in the homogenization of the sequences at the tandem array, individual, population, and species levels mainly through genomic mechanisms like unequal crossing over [4, 5]. Homogeneous nrDNA sequences are therefore generally found within one genome [2]. This implies reduced levels of intraspecific variation (as compared to interspecific) therefore allowing a reduced intraspecific sampling effort. It

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_7,

also provides ease of analysis (because nrDNA is abundant and uniform) [2].

In plants, nrDNA are generally arranged in two distinct sets of tandem repeats. The first one is composed of 5 s nrDNA and the ond of 18 s + 5.8 s + 26 s nrDNA (Chapter 2). The latter is the most frequently used for plant phylogeny and taxonomy. It is present at one or a few loci (with hundred to thousands of tandem copies) [6], and when transcriptionally active, these regions are referred to as NORs (Nucleolar Organizer Regions). It comprises different spacer regions. The intergenic spacer IGS, which separates adjacent 18 s + 5.8 s + 26 s nrDNA units, contains many reiterated sub-repeats within its sequence and is very variable both in sequence and in length [3]. This leads to difficulties for correctly aligning IGS sequences. It therefore has not received as much attention as the internal transcribed spacers (ITS), which are flanking the 5.8S RNA gene region (between the 18 s and the 26 s RNA genes)(Fig. 1). This entire ITS region (ITS1 + 5.8 s + ITS2) can be easily amplified using universal primers in the conserved coding regions [3] (Fig. 1), as the total size is up to 700 bp in angiosperms [1], although in some other seed plants such as gymnosperms, it can be much longer, up to 1500–3700 bp [9]. The ITS region has become highly popular in plant phylogeny [9], as witnessed by the constant increase of Embryophyta ITS sequences available in the NCBI database since 2003 (381,076 ITS sequences as per 30th of March 2020) (Fig. 2). As a comparison, much fewer hits (59,519) are obtained for 5 s nrDNA. This region provides different levels of informativeness: the central 5.8 s RNA gene is highly conserved due to evolutionary constrains, whereas the surrounding ITS spacers are highly variable (particularly the ITS2 spacer) and more informative. This can be illustrated by an alignment of these regions made from Poaceae sequences (Fig. 3).

Despite early warnings [13], attention has been focused only recently on the possible drawbacks in using nrDNA (and therefore ITS) for phylogenetic studies [10]. Concerted evolution does not always act immediately after organismal processes (such as hybridization or polyploidization) or after genomic changes (duplication, recombination) [3, 10]. Concerted evolution efficiency may also

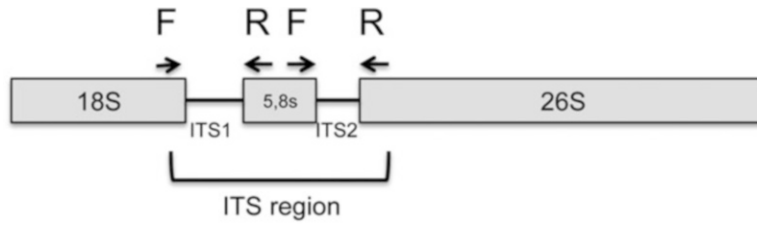


Fig. 1 Structure of the ITS region of the nuclear ribosomal RNA genes and schematic location of primers from Table 1

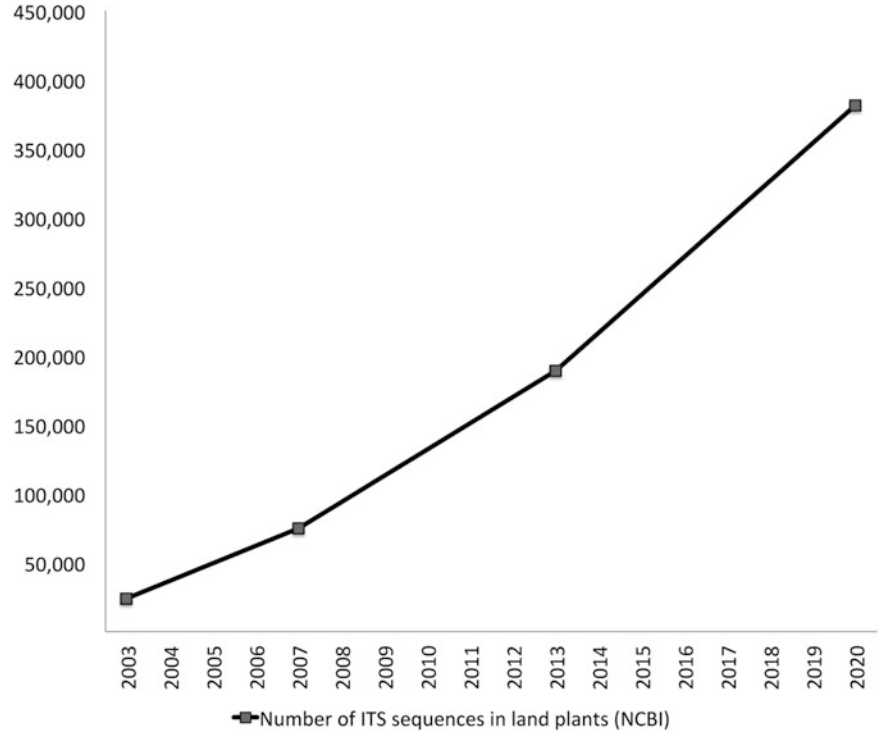


Fig. 2 Number of nucleotide ITS sequences for land plants in NCBI: 2003 and 2007 values are from [10], 2013 value was as searched on the 17th of June 2013 (189,026 hits) [11], 2020 value, see text

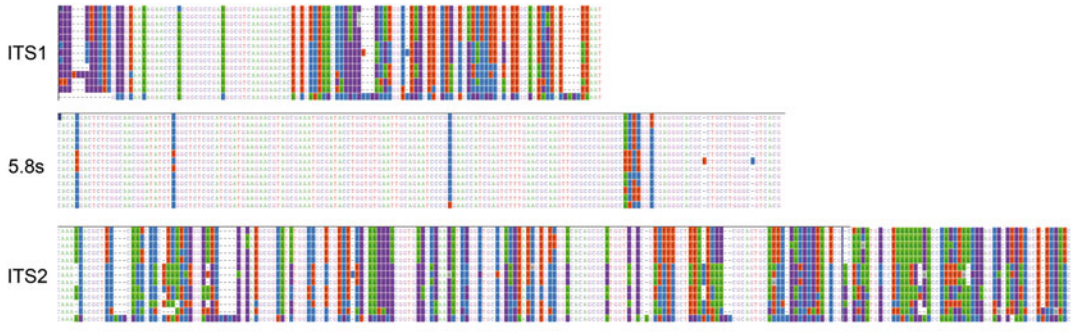


Fig. 3 Variable sites (highlighted) in the ITS1, 5.8 s rRNA gene and ITS2 regions of various Poaceae species following alignment with Mega 5 [12]: *Bromus carinatus* (AY367948), *Bromus gunckelii* (AY367947), *Bromus berteroanus* (AY367946), *Bromus striatus* (AY367945), *Bromus cebadilla* (AY367944), *Festuca matthewsii* (AY524836), *Festuca madida* (AY524833), *Festuca novae-zelandiae* (AY524832), *Agropyron cristatum* (L36480), *Thinopyrum bessarabicum* (L36506), *Lolium perenne* (L36517), *Poa alpina* (AY327793), and *Oryza sativa* (DQ996015)

vary across loci and taxons [10]. This may induce intraindividual nrDNA polymorphism [13]. If hybrids and allopolyploids are recent, they might retain paralogous copies of their nrDNA genes. On the other hand, in some hybrid species or polyploids, one of the parental nrDNA can be more or less rapidly selectively eliminated [9, 14, 15]. Intermediate cases of partial additivity are also found [9]. The PCR product obtained may therefore represent a mixture of sequences (in various concentrations) sharing the same priming sites, but located at one or more locus on one or more chromosomes, and representing either paralogous or orthologous sequences [9, 13]. It is important to be aware of the possibility of such intraspecific nrDNA heterogeneity, which may thus result not only from the presence of homeologous loci due to recent hybridization (with or without polyploidization), but as well from [9, 15, 16]:

- Low concerted evolution rates: different sequences will coexist within a single locus.
- Duplication.
- Allelic variants (heterozygosity).
- Amplification of nonfunctional copies (pseudogenes) with different evolutionary constraints [17].
- Possibilities of contamination by fungal DNA (as the same primers are used for plants) [9].

Paralogous copies resulting from hybridization or allopolyploidization processes can efficiently be utilized to study these processes. Many examples are reviewed in [3, 10]. The problem is more when paralogous sequences are mistaken for orthologous sequences, which will lead to wrong inference in species relationships [13, 18]. The occurrence of nrDNA intraspecific heterogeneity (not due to hybrids or allopolyploids) has been documented in a range of taxa as reviewed by [18]. Only a detailed study involving cloning and cytogenetics (e.g., FISH to reveal array number and chromosomal distribution) (*see* Chapter 18) may help resolve the origin of this heterogeneity. Very thorough flowchart diagrams were designed [16] to help unravel part of these problems. The least that should be done is:

(1) To check, after PCR amplification (under stringent conditions), that only one clear band is obtained. Otherwise subsequent cloning and analysis of each PCR product will be required. (2) After sequencing, always do a BLAST search to check for possible contamination (particularly from fungus). (3) It is also very important to verify the congruence of the obtained ITS tree with other marker trees such as chloroplast gene trees: if differences are detected, they could be due to the hybrid status of some species, but isolated polyphyly could indicate paralogous sequence. A more detailed

procedure depicted in very thorough flowchart diagrams is available [16] if necessary.

A strategy to distinguish between paralogous (pseudogene) and orthologous copies by using nucleotide diversification patterns to determine if sequences are functionally constrained using tree-based approaches was also proposed [18]: A comparison of 5.8 s and ITS trees is performed. Functional copies should have a slower rate of evolution of 5.8 s compared to ITS, whereas pseudogenes should show equal evolutionary rates in 5.8 s and ITS.

The ITS region was early proposed [19] as a powerful tool for plant DNA barcoding, following testing on a large plant sampling (99 species covering 80 genera from 53 different families) which showed more divergence (2.81%) than the most variable intergenic chloroplastic region *trn*H-*psb*A (1.24%). Nevertheless, two chloroplastic genes (*mat*K and *rbc*L) have been selected by CBOL in 2009 as the universal plant barcode system [20] mainly because of the previously mentioned possible drawbacks in ITS analysis (*see also* Chapter 8). Recently, the high success rate of the ITS2 region to identify species in dicotyledons (76.1%), monocotyledons (74.2%), gymnosperms (67.1%), ferns (88.1%), and mosses (77.4%) was further demonstrated [21]. Moreover, ribosomal RNAs can form secondary structures that offer new prospects for phylogenies [3]. The secondary structure of ITS2 is evolutionary constrained (it is composed of four helixes throughout eukaryotes (Fig. 4)) because it is important for ribosome synthesis; on the other hand its sequence is highly variable because ITS2 is not present in the mature ribosome [22]. Both levels of information

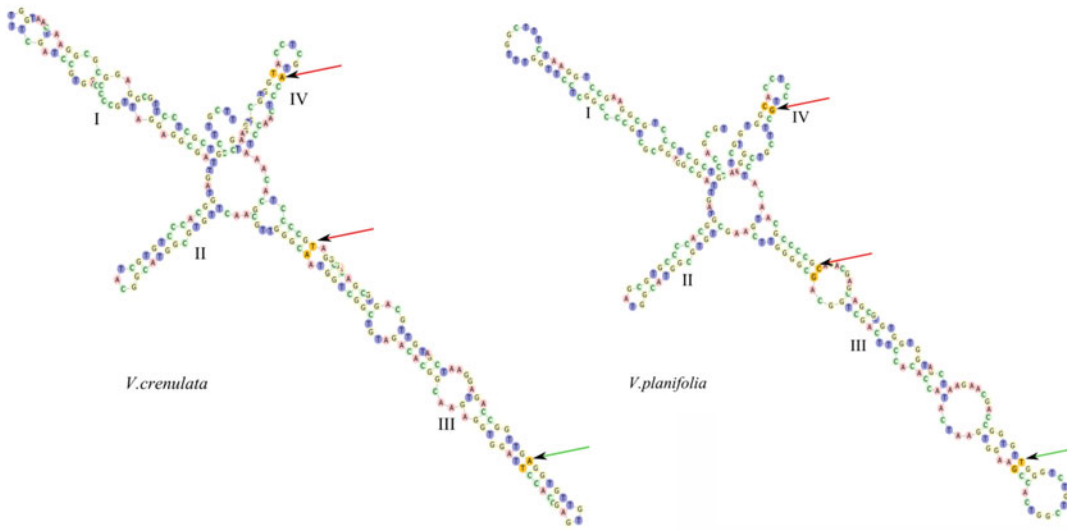


Fig. 4 Example of ITS2 secondary structures of two different orchid species, *Vanilla planifolia* and *V. crenulata*, showing the occurrence of the typical four helixes. CBCs between the two species are indicated by arrows (the green arrows show CBCs located within the critical conserved 30pb zone)

are precious and allow the construction of robust phylogenies at various taxonomic levels with increased resolution as compared to simple sequence analyses [22–26]. Moreover, the existence of CBCs (Compensatory Base Changes) between individuals was shown to be a powerful way to identify species [27]. Particularly, CBCs located within a 30 pb highly conserved region in the 5′ part of helix III were shown to be indicative of sexual incompatibility [28](Fig. 4). An ITS2 database was constructed in 2006 [29], now version IV (http://its2.bioapps.biozentrum.uni-wuerzburg.de/) [30–32], allowing ITS2 secondary structure predictions and phylogenetic analyses. The secondary structure of ITS1 can be predicted as well but is much more variable and therefore more difficult to analyze and use [3].

The Chinese Barcoding of Life group [33] conducted a very thorough (6286 samples from 1757 angiosperm and gymnosperm species) comparative (with chloroplastic genes *rbc*L and *mat*K and intergene *trn*H–*psb*A) research on ITS efficiency/universality. Themselves and others [34] advocate for the incorporation of the ITS region as a supplementary barcode for land plants. Adding ITS to the official plant barcode system (*rbc*L + *mat*K) indeed brings discrimination success from 49.7% to 77.4% (*and see* Chapter 8). Furthermore, [33] showed that contrary to what was feared [9], very low problems with fungal contamination (only in 2% of the samples studied) and a very low occurrence of intraindividual multiple copies of nrDNA (only 7.4% of the individuals) were detected.

ITS is therefore a highly suitable and powerful region for resolving plant taxonomic and phylogenetic issues in most plant lineages, as long as one is aware of its possible (but hopefully rare) limits.

2 Materials

All solutions must be made up using sterile deionized water (Milli-Q water), and all chemicals must be analytical reagent grade. As in all molecular biology procedures, work surfaces should be cleaned and gloves should be worn for all procedures.

2.1 PCR

1. PCR machine (thermocycler).
2. PCR plates or PCR tubes.
3. Taq polymerase: GoTaq® DNA Polymerase (Promega) is well suited, 5 U/μL.
4. Appropriate Taq polymerase buffer (e.g., Green Flexi Buffer for GoTaq®DNA Polymerase) (5×).
5. $MgCl_2$ 25 mM (if not present in buffer).

Table 1
Universal plant primers for the ITS region (always use a combination of a forward and a reverse primer; *see* Fig. 1)

Primer name	Sequence (5′- > 3′)	Reference
18 s forward primers		
ITS1	TCCGTAGGTGAACCTGCGG	[7]
ITS5	GGAAGTAAAAGTCGTAACAAGG	[7]
17SE	ACGAATTCATGGTCCGGTGAAGTGTTC	[8]
26 s reverse primers		
ITS4	TCCTCCGCTTATTGATATGC	[7]
26SE	TAGAATTCCCCGGTTCGCTCGCCGTTAC	[8]
5.8 s reverse primer		
ITS2	GCTGCGTTCTTCATCGATGC	[7]
5.8 s forward primer		
ITS3	GCATCGATGAAGAACGCAGC	[7]

6. dNTP mix (10 mM each) (add 10 µL of each dNTP solution at 100 mM to 60 µL Milli-Q water).
7. Universal plant ITS primers (Table 1) (5 µM).
8. Plant DNA (10 ng/µL).
9. Sterile Milli-Q water.

2.2 Electrophoresis

1. Electrophoresis apparatus (gel tray, combs, power supply).
2. Standard transilluminator (302 nm with 6 × 15 W tubes).
3. High-resolution agarose and standard agarose (molecular biology grade) (*see* **Note 1**).
4. 10× TRIS (tris(hydroxymethyl)aminomethane)-borate (TBE) buffer: 10.8 g TRIS base, 5.5 g boric acid, 0.7 g ethylenediaminetetraacetic acid. (EDTA)-Na_2 in100 mL H_2O. This TBE buffer is diluted to 1× in Milli-Q water for use.
5. Fluorescent nucleic acid gel stain: GelRed™ 1000× in water (*see* **Note 2**) (ethidium bromide can also be used if preferred).

3 Methods

3.1 PCR Reaction

For each PCR reaction (25 µL) (*see* **Note 3**):

1. Deposit 2.5 µL template DNA in tube or well of the plate.
2. Add 1.5 µL $MgCl_2$ 25 mM.

3. Add 0.5 µL dNTPmix 10 mM.
4. Add 1.5 µL of each primer 5 µM (forward and reverse).
5. Add 5 µL of PCR buffer 5×.
6. Add 12.3 µL Milli-Q sterile water.
7. Add 0.2 µL (1 U) DNA polymerase.

3.2 PCR Program

1. Pre-denaturation at 95 °C for 3 min.
2. 35 cycles with 95 °C for 45 s, 60 °C for 45 s, and 72 °C for 1 min 30.
3. Final elongation step at 72 °C for 7 min.
4. Maintain at 4 °C.

3.3 PCR Quality Verification on Agarose Gel

1. Prepare a 2% mixture of agarose in TBE 1× (2 g agarose for 100 mL).
2. Bring to boil in a microwave oven.
3. Add 5 µL RedGel™ for 50 µL 2% agarose/TBE.
4. Cool down, and then pour the gel. Let the gel cool down and prepare for migration.
5. Add 10 µL of PCR solution and loading dye.
6. Deposit in the well.
7. Run migration for appropriate time and observe gel over trans-illuminator (*see* **Note 4**).

3.4 Sequencing

A large number of private companies perform sequencing reaction directly from PCR products that can be sent by express mail (either sealed or vacuum dried in a SpeedVac). Generally the primers used have to be provided (*see* **Note 5**).

4 Notes

1. We use high-resolution agarose gels rather than standard agarose gels to check for the purity of the amplified fragment and insure that only one band is amplified. Further routine checks can be made on standard agarose gels, which can be re-thawed and reused six times.
2. We prefer using GelRed™ than ethidium bromide (EB) as with standard Ames test, as measured in two bacterial strains, GelRed™ has been confirmed to be substantially safer than EB. GelRed™ is not mutagenic at all dosages in the absence of the S9 fraction. With S9 metabolic activation, GelRed™ showed weak mutagenicity only at the highest dosage (50 µg/plate or 18.5 µg/mL), well above the normal concentration used for gel staining.

We however use EB safety rules when handling GelRed™: solution pipetting is made under a fume hood, and wear gloves and a lab coat. Whether GelRedTM waste solution can be directly poured into the drain may depend on local regulations despite its nonmutagenicity and noncytotoxicity. Alternatively, GelRedTM solution may be disposed by adding 25–50 mL bleach (regular household bleach) to each gallon (~4 L) of the waste staining solution, and let the mixture react for at least 8 h before pouring the solution to a sink (practically, you may simply accumulate your GelRedTM waste solution in a jar containing appropriate amount of bleach). For precast gels, you can simply let the gels dry out first and then let the dried waste go in regular trash bag (together with gloves and other wastes which are autoclaved prior to disposal).

3. Generally a PCR mix (**steps 2–7**) is prepared for the desired number of reactions (allow for 10% variation) and then aliquoted in the wells of the PCR plates or in the PCR tubes containing the DNA samples. Work on ice.
4. At this stage the amplification should give a unique clear band. If more than one band is obtained, try to use more stringent PCR conditions (increase annealing temperature, lower $MgCl_2$ concentration, use species-specific primers rather than universal ones). If the problem still appears, the different bands will have to be cloned and sequenced.
5. Always ask for a double sequencing reaction, e.g., forward and reverse, which helps to check the quality of the sequence (consistently poor reactions despite specific primers and stringent PCR conditions might be due to heterozygous state or heterogeneity in the sequences—paralogous copies).

Acknowledgments

Denis Da Silva (Université de La Réunion, UMR PVBMT) and Michel Grisoni (CIRAD, UMR PVBMT) are acknowledged for their help in the ITS amplification protocol design. Adeline Trebel and Nicolas Cazanove are thanked for running ITS2 secondary structure analyses as part of their master 1 thesis work.

References

1. Baldwin BG, Sanderson MJ, Porter JM, Wojciechowski MF, Campbell CS, Donoghue MJ (1995) The ITS region of nuclear ribosomal DNA: a valuable source of evidence on angiosperm phylogeny. Ann Mo Bot Gard 82:247–277
2. Hillis DM, Dixon MT (1991) Ribosomal DNA: molecular evolution and phylogenetic inference. Q Rev Biol 66:411–453
3. Poczai P, Hyvönen J (2010) Nuclear ribosomal spacer regions in plant phylogenetics: problems and prospects. Mol Biol Rep 37:1897–1912

4. Arnheim N (1983) Concerted evolution of multigene families. In: Nei M, Koehn RK (eds) Evolution of genes and proteins. Sinauer, Sunderland, MA
5. Dover G (1994) Concerted evolution, molecular drive and natural selection. Curr Biol 4:1165–1166
6. Rogers SO, Bendich AJ (1987) Ribosomal RNA genes in plants: variability in copy number and in the intergenic spacer. Plant Mol Biol 9:509–520
7. White TJ, Bruns T, Lee S, Taylor J (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ, White TJ (eds) PCR protocols. A guide to methods and applications, vol 18. Academic Press, San Diego, pp 315–322
8. Sun Y, Skinner D, Liang G, Hulbert S (1994) Phylogenetic analysis of sorghum and related taxa using internal transcribed spacers of nuclear ribosomal DNA. Theor Appl Genet 89:26–32
9. Álvarez I, Wendel JF (2003) Ribosomal ITS sequences and plant phylogenetic inference. Mol Phylogenet Evol 29:417–434
10. Calonje M, Martín-Bravo S, Dobeš C, Gong W, Jordon-Thaden I, Kiefer C, Kiefer M, Paule J, Schmickl R, Koch MA (2009) Non-coding nuclear DNA markers in phylogenetic reconstruction. Plant Syst Evol 282:257–280
11. Besse P (2014) Nuclear ribosomal RNA genes: ITS region. In: Molecular plant taxonomy. Springer, pp 141–149
12. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739
13. Buckler ES, Ippolito A, Holtsford TP (1997) The evolution of ribosomal DNA divergent paralogues and phylogenetic implications. Genetics 145:821–832
14. Bachmann K (1994) Molecular markers in plant ecology. New Phytol 126:403–418
15. Zimmer EA, Wen J (2013) Using nuclear gene data for plant phylogenetics: progress and prospects. Mol Phylogenet Evol 66:539–550
16. Feliner GN, Rosselló JA (2007) Better the devil you know? Guidelines for insightful utilization of nrDNA ITS in species-level evolutionary studies in plants. Mol Phylogenet Evol 44:911–919
17. Besnard G, Rubio de Casas R, Vargas P (2007) Plastid and nuclear DNA polymorphism reveals historical processes of isolation and reticulation in the olive tree complex (Olea europaea). J Biogeogr 34:736–752
18. Bailey CD, Carr TG, Harris SA, Hughes CE (2003) Characterization of angiosperm nrDNA polymorphism, paralogy, and pseudogenes. Mol Phylogenet Evol 29:435–455
19. Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. Proc Natl Acad Sci 102:8369–8374
20. CBOL Plant Working Group, Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL (2009) A DNA barcode for land plants. Proc Natl Acad Sci 106:12794–12797
21. Yao H, Song J, Liu C, Luo K, Han J, Li Y, Pang X, Xu H, Zhu Y, Xiao P, Chen S (2010) Use of ITS2 region as the universal DNA barcode for plants and animals. PLoS One 5:e13
22. Keller A, Förster F, Müller T, Dandekar T, Schultz J, Wolf M (2010) Including RNA secondary structures improves accuracy and robustness in reconstruction of phylogenetic trees. Biol Direct 5:4
23. Coleman AW (2003) ITS2 is a double-edged tool for eukaryote evolutionary comparisons. Trends Genet 19:370–375
24. Schultz J, Wolf M (2009) ITS2 sequence–structure analysis in phylogenetics: a how-to manual for molecular systematics. Mol Phylogenet Evol 52:520–523
25. Buchheim MA, Sutherland DM, Schleicher T, Förster F, Wolf M (2012) Phylogeny of Oedogoniales, Chaetophorales and Chaetopeltidales (Chlorophyceae): inferences from sequence-structure analysis of ITS2. Ann Bot 109:109–116
26. Wolf M, Koetschan C, Mueller T (2014) ITS2, 18S, 16S or any other RNA—simply aligning sequences and their individual secondary structures simultaneously by an automatic approach. Gene 546:145–149
27. Müller T, Philippi N, Dandekar T, Schultz J, Wolf M (2007) Distinguishing species. RNA 13:1469–1472
28. Coleman AW (2009) Is there a molecular key to the level of "biological species" in eukaryotes? A DNA guide. Mol Phylogenet Evol 50:197–203
29. Schultz J, Müller T, Achtziger M, Seibel PN, Dandekar T, Wolf M (2006) The internal transcribed spacer 2 database—a web server for (not only) low level phylogenetic analyses. Nucleic Acids Res 34:W704–W707

30. Ankenbrand MJ, Keller A, Wolf M, Schultz J, Förster F (2015) ITS2 database V: twice as much. Mol Biol Evol 32:3030–3032
31. Merget B, Koetschan C, Hackl T, Förster F, Dandekar T, Müller T, Schultz J, Wolf M (2012) The ITS2 database. J Vis Exp (61): e3806
32. Koetschan C, Hackl T, Müller T, Wolf M, Förster F, Schultz J (2012) ITS2 database IV: interactive taxon sampling for internal transcribed spacer 2 based phylogenies. Mol Phylogenet Evol 63:585–588
33. Li D-Z, Gao L-M, Li H-T, Wang H, Ge X-J, Liu J-Q, Chen Z-D, Zhou S-L, Chen S-L, Yang J-B, Fu C-X, Zeng C-X, Yan H-F, Zhu Y-J, Sun Y-S, Chen S-Y, Zhao L, Wang K, Yang T, Duan G-W (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. Proc Natl Acad Sci U S A 108:19641–19646
34. Hollingsworth PM (2011) Refining the DNA barcode for land plants. Proc Natl Acad Sci U S A 108:19451–19452

Chapter 8

Plant DNA Barcoding Principles and Limits: A Case Study in the Genus *Vanilla*

Pascale Besse, Denis Da Silva, and Michel Grisoni

Abstract

Powerful DNA barcodes have been much more difficult to define in plants than in animals. In 2009, the international Consortium for the Barcoding Of Life (CBOL) chose the combination of the chloroplast genes (rbcL + matK) as the proposed official barcode for plants. However, this system has got important limits. First, any barcode system will only be useful if there is a clear barcode gap and if species are monophyletic. Second, chloroplast and mitochondrial (COI gene used for animals) barcodes will not be usable for discriminating hybrid species. Moreover, it was also shown that, using chloroplast regions, maximum species discrimination would be around 70% and very variable among plant groups. This is why many authors have more recently advocated for the addition of the nuclear ITS region to this barcode because it reveals more variations and allows the resolution of hybrid or closely related species. We tested different chloroplast genes (rbcL, matK, psaB, psbC) and the nuclear ITS region in the genus *Vanilla*, a taxonomically complex group and therefore a good model to test for the efficiency of different barcode systems. We found that the CBOL official barcode system performed relatively poorly in *Vanilla* (76% species discrimination), and we demonstrate that adding ITS to this barcode system allows to increase resolution (for closely related species and to the subspecies level) and to identify hybrid species. The best species discrimination attained was 96.2% because of one paraphyletic species that could not be resolved.

Key words DNA barcoding, ITS, rbcL, matK, Barcode gap, Species discrimination

1 Introduction

1.1 Barcoding History in Plants

Stoeckle et al. [1], in the "Taxonomy, DNA, and the Barcode of Life" meeting held in New York in 2003, proposed the use of a single universal short DNA sequence for molecular identification of all living species, so-called DNA barcoding, later officialized through an international consortium named CBOL (Consortium for the Barcoding Of Life). *Imagine a world in which any person, anywhere, at any time can identify any species at little or no cost. That world is technologically upon us.* It was in these exact terms that the global aim of this ambitious (other would say unrealistic) project was defined. Nevertheless, the idea of a universal barcoding system was appealing.

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_8,

In animals, the 5′ end of the mitochondrial COI (cytochrome *c* oxidase subunit 1) gene (≈ 650 bp length) was rapidly identified as the best DNA barcode [2]. This was because (a) the mitochondrial genome is present in numerous copies in a cell, allowing easy sequencing; (b) in animal mitochondrial genome, the order of the genes is highly conserved; (c) the gene shows a high substitution rate [3] and therefore a high polymorphism between species; and (d) there is a low intraspecific polymorphism thanks to maternal transmission [4].

However, COI was rapidly demonstrated to be not suitable for plant barcoding. Indeed the mitochondrial genome in plants shows low variation rates (low polymorphism) and numerous rearrangements [5], advocating for the use of a different barcode system for plant species. From 2005, many research groups tested a large variety of sequences for finding the best barcode system for plants [6–15] as termed the Holy Grail quest by some authors [16]. An illustrative "enthusiasm heatmap" summary of all the sequences tested can be found in [17]. The chloroplast (cp) genome was chosen as the genome of choice for barcoding in plants because of its maternal inheritance and therefore low intraspecific variation like the mitochondrial genome. However, cpDNA is less variable in plants than is mitochondrial genome in animals (Chapter 2). Some authors recommended the use of noncoding intergenic regions such as trnH-psbA [8], because these regions show a much higher rate of variation and better species discrimination than coding cpDNA regions (Chapter 2). Nevertheless, most authors favored the use of cpDNA genes, such as rbcL or matK [12, 13] as being a more appropriate DNA barcode sequence (more similar to the animal barcode choice). However, these cpDNA coding regions show low variation. It became rapidly evident that DNA barcoding in plants would necessitate the use of a multilocus system [5]. The initial suggestion of a universal and unique gene for DNA barcode was long gone. As it appeared as the best system recognized through publications testing all possible previously suggested barcodes [9, 10], the CBOL proposed in 2009 [18] to use the combination (rbcL + matK) as the official barcode for plants. It is the system currently recognized in the CBOL barcoding database BOLD (www.boldsystems.org/) [19, 20].

1.2 Limits of DNA Barcoding

DNA barcoding suffers a range of important limitations. First, it can only be applied to identify monophyletic species and will obviously fail in case of polyphyly or paraphyly (Fig. 1). An ideal DNA barcode should generate non-overlapping distributions between intra- and interspecific distances, so-called barcoding gap (Fig. 1). The absence of a barcoding gap will make impossible the definition of a threshold value to identify species, generating either false negatives (species missed) or false positives (false species) (Fig. 1). Because of hybridization and greater levels of gene-tree paraphyly

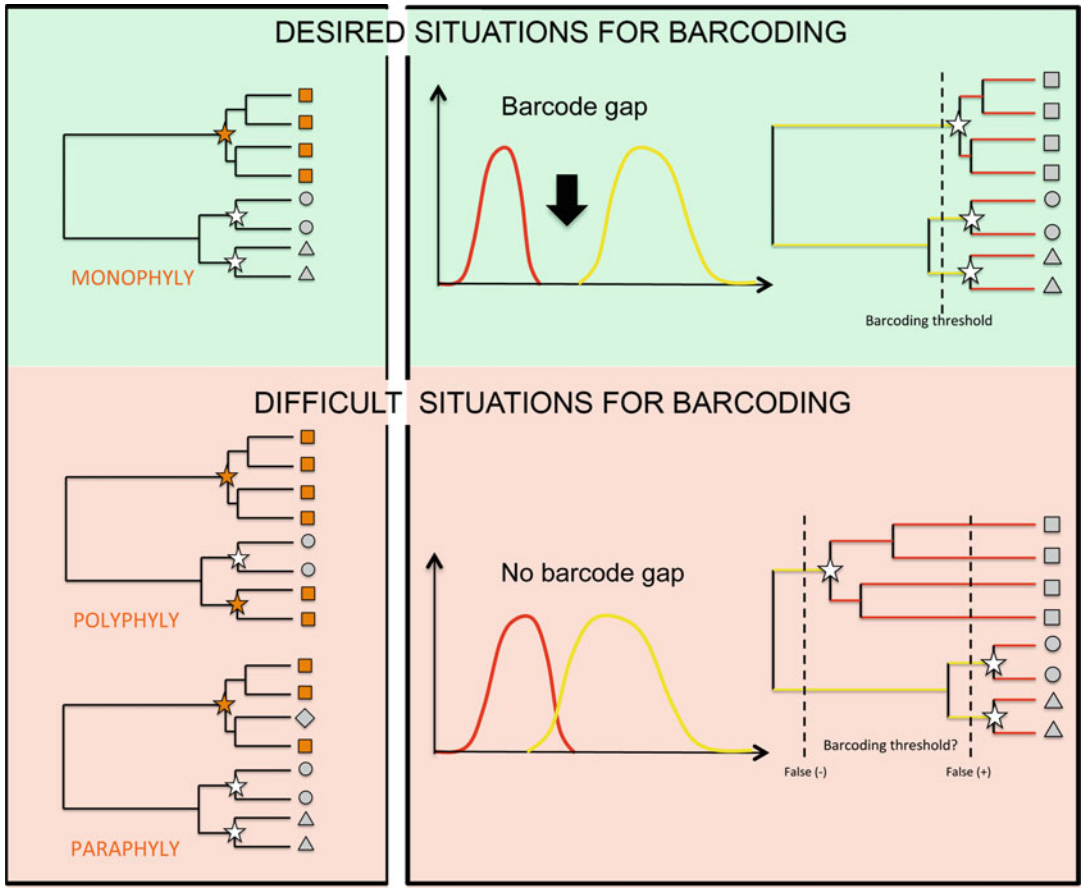


Fig. 1 Schematic representation of favorable and unfavorable situations for DNA barcoding. Only monophyletic species are appropriate (polyphyletic and paraphyletic species are not), and they need to display a clear barcode gap (i.e., a gap between frequency distributions of intraspecific (in red) versus interspecific (in yellow) distances. If monophyletic species show no barcode gap, it will be impossible to define a barcoding threshold

in plants, barcode gaps are less important than in animals, making plants much more difficult to barcode accurately [21].

Pursuing the search for a suitable barcode system in plants, it became rapidly quite clear that, in plants, even with the best barcode or combination of barcodes, the maximum discrimination power would reach only 70% of plant species (even with the best combination of more than two chloroplast markers) and would largely vary across plant families [9] [10]. It is predicted that plant groups that are long-lived, with polyploidy, hybridization, closely related autogamous lineages, recent speciation, narrow species limits, or poor seed dispersal, will show lower discrimination success with DNA barcodes [17]. In particular, it is clear that because of maternal transmission of the chloroplast genome (like the mitochondrial genome), a species of hybrid origin will not be differentiable from the maternal parental species (Fig. 11 in Chapter 2).

Many publications were written, particularly by the CBOL Chinese group [22], to plead for the inclusion of the ITS gene in the CBOL barcode system. ITS barcode limitations (Chapter 7) (lower universality, fungal contamination, paralogous gene copies) were carefully addressed and shown to be acceptable. Indeed, the ITS region shows much more variations than any of the chloroplast barcodes (see also Table S1 in [17] for a detailed review of papers). This strongly stresses the need to refine the current CBOL barcode system, as acknowledged by [23] [17]. Currently, efforts are made to develop standardized protocols, and researchers are discussing the usefulness of the ITS2 spacer [24–26] against that of the ITS1 spacer [27] to provide a simple and short barcode. Some research groups have already set up barcode databases for ITS2 in plants [25, 28] and eukaryotes [29, 30] allowing the use of both sequence and secondary structure variations as powerful barcodes (Chapter 7).

1.3 A Case Study in the Genus Vanilla

Vanilla Plum. ex Miller is an ancient genus in the Orchidaceae family, Vanilloideae subfamily, Vanilleae tribe, and Vanillinae subtribe [31]. *Vanilla* species are distributed throughout the tropics between the 27th north and south parallels, in Africa, America, and Asia. Over 100 species have been described in the genus [32]. Taxonomic classification is complex in this genus as it is based on morphological variations in vegetative traits (which show important intraspecific variations) and on floral traits (but flowers are ephemeral and rarely available in herbarium specimens) [33]. A DNA barcode system for this genus is therefore highly desirable.

Vanilla can be considered as a TCG, a "Taxonomic Complex Group" [34], because it shows both a sexual and a uniparental reproduction mode (vegetative reproduction) [33, 35], interspecific hybridization [36, 37] [38], and polyploidy [33, 39].

This genus is therefore a good model to assess the discrimination power of various DNA barcodes. From our previous published [40] and unpublished [41] [42] work on molecular taxonomy in the genus *Vanilla*, we obtained sequence data for the chloroplast gene regions rbcL, matK, psaB, and psbC and the nuclear ribosomal DNA ITS region. The objective was to test which of these loci were more appropriate to be used for specimen identification in the genus *Vanilla*.

2 Testing Barcoding Sequences for Vanilla

A total of 52 accessions maintained in the *Vanilla* collection of the Biological Resources Center (BRC) Vatel in Reunion Island were studied (Table 1). This dataset represents a total of 26 *Vanilla* species, representative of the genus diversity. All specimens used here were previously identified based on an integrative combination

Table 1
***Vanilla* species studied and their classification in subgenera, sections, and morphological groups [32]**

Subgenera and section	Morphological group	Species	Subspecies	*n*
Subgen. *Xanata*	*V. planifolia*	*V. planifolia*		5
Sect. *Xanata*		*V. bahiana*		4
		V. ×tahitensis		3
		V. sotoarenasii		1
		V. phaeantha		2
		V. ensifolia		1
		V. insignis		1
		V. odorata		1
	V. hostmanii	*V. cribbiana*		1
	V. pompona	*V. chamissonis*		1
		V. pompona	*pittieri*	1
		V. pompona	*pompona*	2
		V. pompona	*grandiflora* α	4
		V. pompona	*grandiflora* γ	1
	V. palmarum	*V. palmarum*		1
		V. lindmaniana		2
Subgen. *Xanata*	*V. africana*	*V. africana*		1
Sect. *Tethya*	*V. albida*	*V. crenulata*		2
		V. albida		1
	V. aphylla	*V. aphylla*		1
	V. barbellata	*V. dilloniana*		2
	V. francoisii	*V. francoisii*		2
	V. imperialis	*V. imperialis*		2
		V. polylepis		1
	V.phalaenopsis	*V. phalaenopsis*		1
		V. roscheri		1
		V. perrieri		3
		V. humblotii		2
Subgen. *Vanilla*	*V. mexicana*	*V. mexicana*		2

Subspecies are according to [48]. The number of studied accessions from the BRC Vatel collection is given (*n*)

of morphological observations (they are maintained as living plants in shade houses) and various unpublished or published molecular phylogenetic studies [40, 41]. It includes species from the *Vanilla* subgenus *Vanilla* (ancestral membranous American species), from the *Vanilla* subgenus *Xanata* section *Xanata* (American leafy species), and from the *Vanilla* subgenus *Xanata* section *Tethya* (African and Asian leafy and leafless species and American leafless species) [32] [35, 40]. For 13 of the species, two to eight different individuals were assessed to gain information on the levels of intra-specific diversity. We used sequences from the chloroplast genes rbcL, matK, psaB, and psbC and from the nuclear ribosomal RNA spacer ITS region (ITS1 + 5.8 s + ITS2) obtained during our various unpublished and published studies [40, 41]. The sequences of the primers that were used for PCR amplification are indicated in Table 2.

Table 2
Primer sequences used (5′-3′)

Sequence	Primer name	Primer sequence 5′-3′
rbcL_part1	RcbL33L	CTCCTGACTACGAAACCAAAGA
	RcbL730R	TCTCTGGCAAATACCGCTCT
rbcL_part2	RcbL453L	TCGTCCCCTATTGGGATGTA
	RcbL1231R	CCTCATTACGAGCTTGCACA
matK	matK743F	CTTCTGGAGTCTTTCTTGAGC
	matK1520R	CGGATAATGTCCAAATACCAAATA
psaB_part1	PsaB49L	CCGTCGCAAGGAAAACTATAA
	PsaB848R	TTCGGGATTGGTCACAGTAT
psaB_part2	PsaB766L	AGACCCTTATGYCCACGYC
	PsaB1526R	GCTTGGCAAGGAAATTTTGA
psbC_part1	PsbC25L	GGTCTGGCTCTGAACCTACG
	PsbC786R	GGGCTAAGGGTCAARTTGGT
psbC_part2	PsbC596L	TCCTTTCCATTCTTCGGTTATG
	PsbC1379R	AAGAACCTAAAGGAGCATGAGTC
ITS	AB101F	ACGAATTCATGGTCCGGTGAAGTGTTC
	AB102R	TAGAATTCCCCGGTTCGCTCGCCGTTAC

Table 3
Percentage of variation revealed for each studied sequence in the genus *Vanilla*

Sequence	Size (bp)	Variable sites	% variation (%)
ITS	774	305	39.40
matK	695	85	12.20
rbcL	1115	53	4.80
psaB	1327	53	4.00
psbC	1223	44	3.60

3 Which Locus Shows the Greatest Sequence Diversity?

Sequence alignment was performed with MUSCLE using the Mega5 software [43]. Genetic distances were then computed in Mega5 as the number of bp differences revealed between two sequences. The data is used to assess the level of diversity for each sequence as the number of variable sites compared to the total length of the sequence (Table 3). The nuclear ITS spacer region is the most variable of the sequences with 39.4% of variable sites. Among the chloroplast genes, matK is the most variable (12.2% of variable sites). The rbcL, psaB, and psbC genes are the longest sequences but the least variable (3.6–4.8%).

In order to determine which locus revealed the most variable genetic distances (measured as the number of bp differences revealed) in the studied dataset, we produced graphs to compare inter-accession divergences for each pair of loci (Fig. 2a). Wilcoxon signed-rank tests (BiostaTGV, biostatgv.sentiweb.fr) were used to test the significance of the differences and show that the range of distances revealed for the different loci are all significantly different from one another (Fig. 2b). This allows the following classification by order of decreasing diversity: ITS >> matK > psaB > rbcL > psbC.

4 Is there a Barcoding Gap in Vanilla?

The frequency distribution of intra- and interspecific genetic distances (expressed as bp differences) was assessed first for each studied sequence at a global scale [44] (Fig. 3). The range of bp differences between individuals is variable between sequences; it ranges from 0–33 bp for the four chloroplast sequences to 0–170 bp for the ITS sequence. There is an overlap in the distribution of the intraspecific and interspecific distances for the five studied sequences, showing the absence of a barcoding gap in *Vanilla* at global scale (Fig. 3). This precludes any possible use of

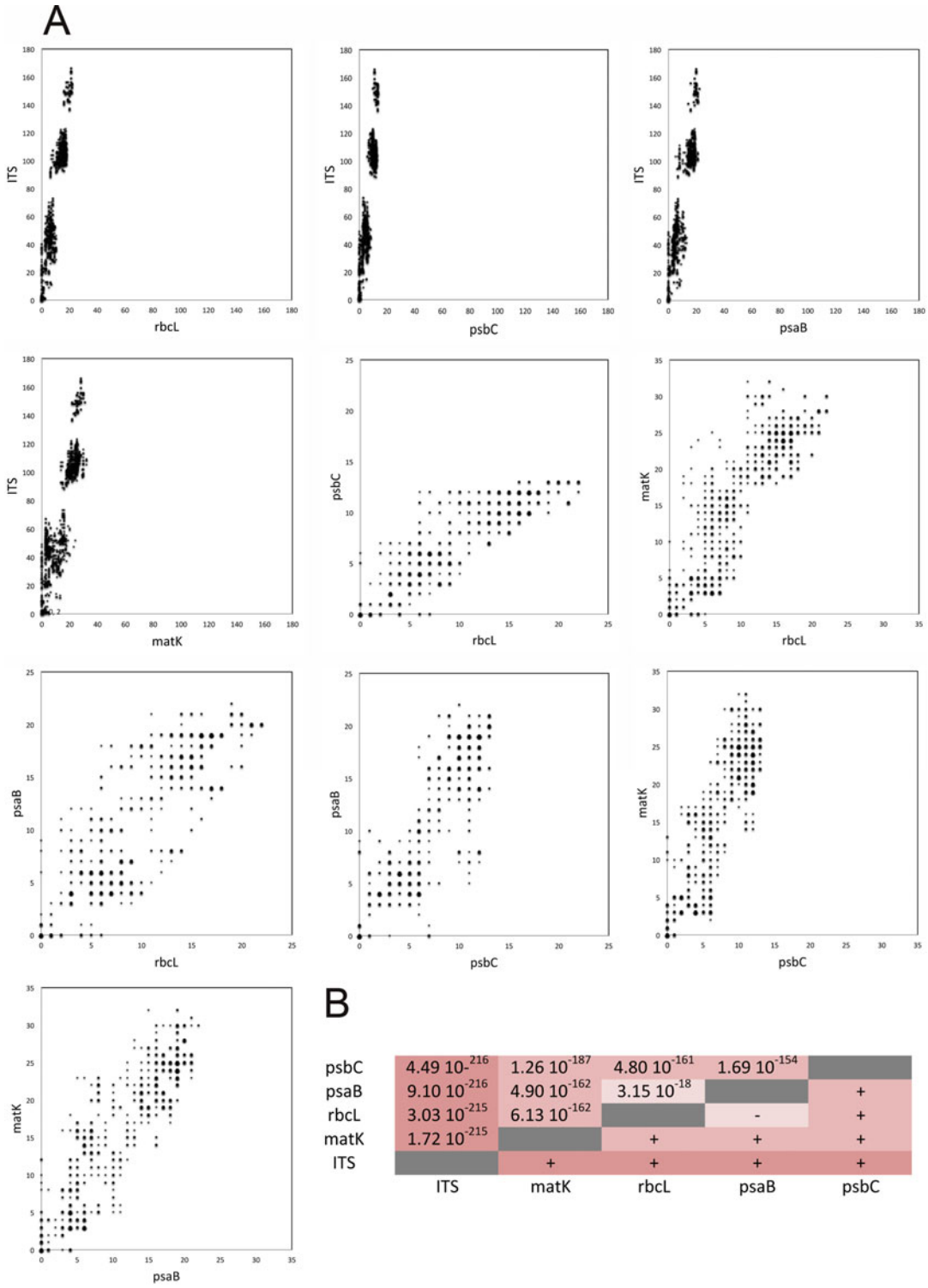


Fig. 2 (**a**) Pairwise genetic distances (number of bp differences) for all locus combinations using the five sequences tested. (**b**) Results of the Wilcoxon signed-rank tests for each combination of locus tested: all p values are highly significant and indicated in the upper panel, in the lower panel + indicates that locus on vertical line is significantly more variable, – indicates that locus on the vertical line is significantly less variable, than the locus on the horizontal line

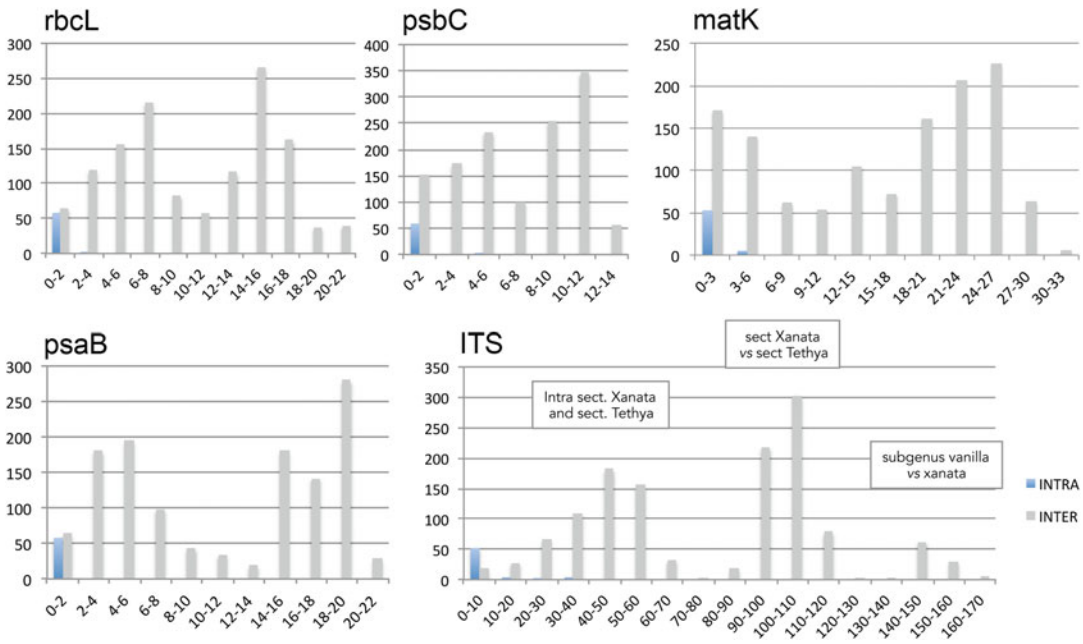
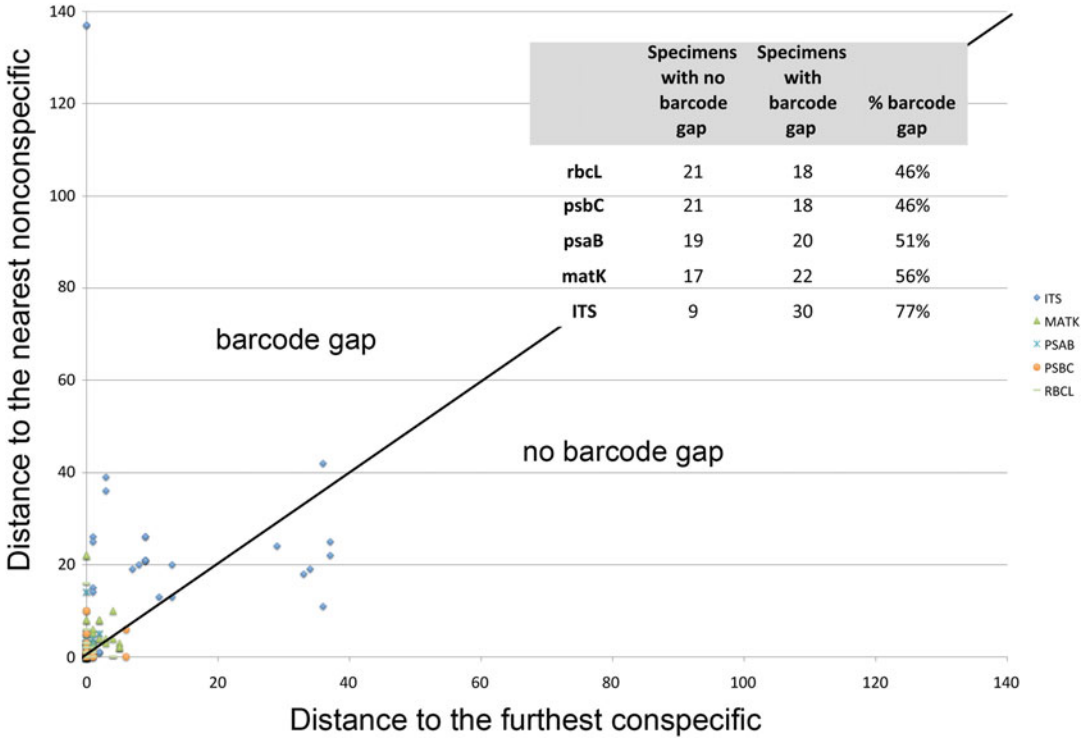


Fig. 3 Barcoding gap study at global scale. Frequency (*y* axis) distribution of the genetic distances (number of bp differences) (*x* axis) for the five loci tested on 52 individuals. Intraspecific distances are indicated in blue, interspecific distances in gray

a general threshold for species delimitation (Fig. 1) [45], which could allow the identification of specimens based on distance data. Interestingly, the distributions of interspecific distances are bimodal for rbcL, psaB, matK, and psbC and clearly trimodal with clear gaps for ITS. These gaps correspond to gaps between sections (*Xanata* vs *Tethya*) or subgenera (*Vanilla* vs *Xanata*). A distance threshold method might be straightforward only for section subdivision (sect. *Xanata* vs sect. *Tethya*) or subgenus subdivision using ITS as a triage tool for preliminary sorting of unknown specimens. The absence of global barcoding gap has been described as a rule in plants rather than an exception, with important coalescence depth variation observed between species [44]. The genus *Vanilla* is no exception, especially as a TCG with recent speciation events and hybridizations.

Identification success of a specimen might however be efficient, even if intraspecific distances for this species exceed interspecific distances for other species [44, 46]. To further assess barcoding gap at this local scale [44] and see more precisely the possible limits for specimen identification, we also performed dot plots for each individual in the dataset for which at least one conspecific was available: the distance to the furthest conspecific was plotted against the distance to the nearest nonconspecific (Fig. 4). Only those individuals falling above the 1:1 slope (largest intraspecific distance inferior to the smallest interspecific distance) would be identifiable



	Specimens with no barcode gap	Specimens with barcode gap	% barcode gap
rbcL	21	18	46%
psbC	21	18	46%
psaB	19	20	51%
matK	17	22	56%
ITS	9	30	77%

Fig. 4 Barcoding gap study at local scale. Dot plot distribution of the distance (number of bp differences) to the furthest conspecific (*x* axis) against the distance to the nearest nonconspecific (*y* axis), for each individual (39 total) and each of the five tested loci. The 1:1 line represents the limit where the difference between x and y is zero (i.e.,the limit of the local barcoding gap). The table describes, for each locus, the number of individuals with (above the 1:1 line) or without (onto or below the 1:1 line) a local barcoding gap

(presence of a barcode gap). The results show that at local scale, only 46% (with rbcL and psbC) to a maximum 77% (with ITS) specimens display a clear barcode gap (Fig. 4).

The absence of clear barcoding gaps both at global and local scale might hamper barcoding success in the genus *Vanilla*. It is of note that, given our sampling, yet important, but not exhaustive, barcoding gap was estimated here by default at the intraspecific level (limited number of accessions per species) and sometimes by excess at the interspecific level (not all species of the genus have been surveyed, including very related ones). Nevertheless, results show that ITS should be the most appropriate marker for the genus *Vanilla*.

5 Which Locus Shows the Greatest Level of Species Discrimination?

Three methods were used for assessing species (specimen) discrimination: distance, blast, and tree-building. Barcode sequences were considered identical when the genetic distance between them was

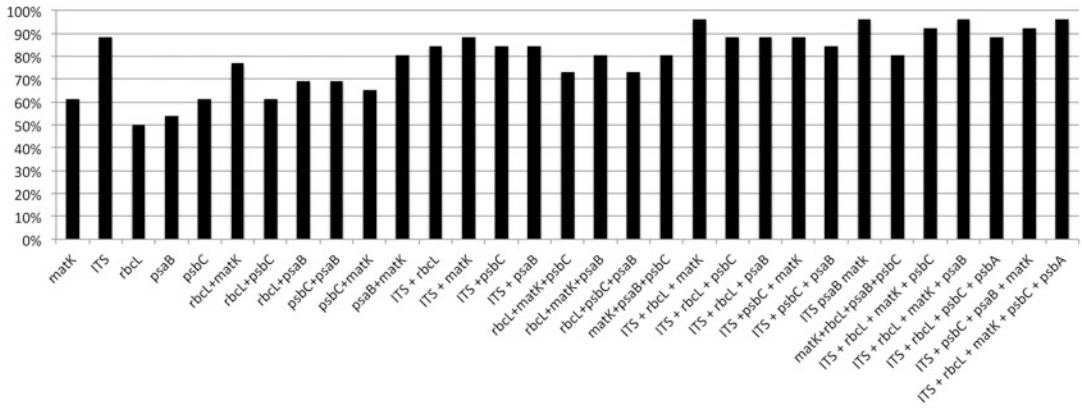


Fig. 5 Percentage of the 26 *Vanilla* species discriminated using each sequence and all possible combinations (one to five) of sequences

null (no bp difference). In the genetic distance matrix generated in Mega5, if any accession from one species had the same barcode sequence as another accession from another species, the two species were considered as not distinguishable. Species represented by a unique accession possessing a unique barcode were considered identified. For species represented by more than one accession, we verified that the species was represented by a monophyletic group on the neighbor joining tree (constructed using Mega5) for it to be identified. If not, we used a Blastn test online (blast.ncbi.nlm.nih.gov), with the align function against our own database built from the current dataset. When each accession of a species was correctly identified to the right species as first hit, the species was considered identified.

The ITS sequence allows to identify 88.5% of the assessed species (23 of 26) (Fig. 5). The best chloroplast sequences are matK and psbC, which allow to discriminate 61.5% of the species (16 of 26 species), and the least efficient for species discrimination is rbcL with 50% (13 of 26 species) (Fig. 5). The order of efficiency of the sequences for species discrimination is therefore ITS > matK > psaB = psbC > rbcL.

6 What Is the Discrimination Gain Obtained by Combining Multiple Loci?

We also tested all possible multiple combination sets of loci following the concatenation of the individual sequences in Mega5. Globally and as expected, there is an important increase in the ability to discriminate species (63.1–96.2% in mean values for all combinations) when combining one to five sequences (Fig. 6).

The best two-marker combination in our dataset is (ITS + matK) with 88.5% discrimination (23 species of 26) (Fig. 5).

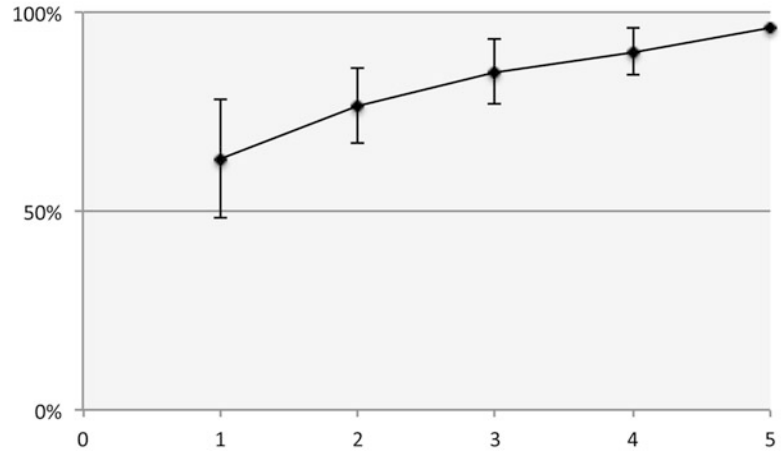


Fig. 6 Percentage of species discrimination using one single or two to five combined sequences (mean values and standard deviation are represented)

Adding one locus (either psaB or rbcL) to this combination increases the resolution to 96.2% (25 species of 26), which appears as the maximum % discrimination possible in our dataset even with the best four (ITS + rbcL + matK + psaB) or even five loci (Fig. 5). ITS brings higher resolution than chloroplast genes: the best chloroplast combinations (either with two, three, or four sequences) only reach 80.8% discrimination (Fig. 5), which is lower than the discrimination achieved by ITS on its own.

7 Conclusion: Which DNA Barcode for Vanilla and What Limits?

The CBOL selected (rbcL + matK) DNA barcode for land plants only allows a poor discrimination of 76.9% (20 species out of 26) for *Vanilla* (Fig. 5). Our results are however concordant with results found in other species, and we confirm that matK is the most variable of the chloroplast markers. This CBOL system is not fully adapted for the *Vanilla* genus. Adding ITS to the international barcode system, on the other hand, allows the maximum (96.2%) discrimination possible in our dataset. Practically, we therefore recommend to use a sequential strategy by first testing for ITS (as the most variable barcode) and then add matK and rbcL to increase resolution for some groups if needed. A neighbor joining tree representing species grouping using this (rbcL + matK + ITS) barcode is proposed for illustration (Fig. 7) (note that this tree has no evolutionary sense as it mixes sequences with different rates of evolution and from different genomes, so it is only given as a visual guide for discussion).

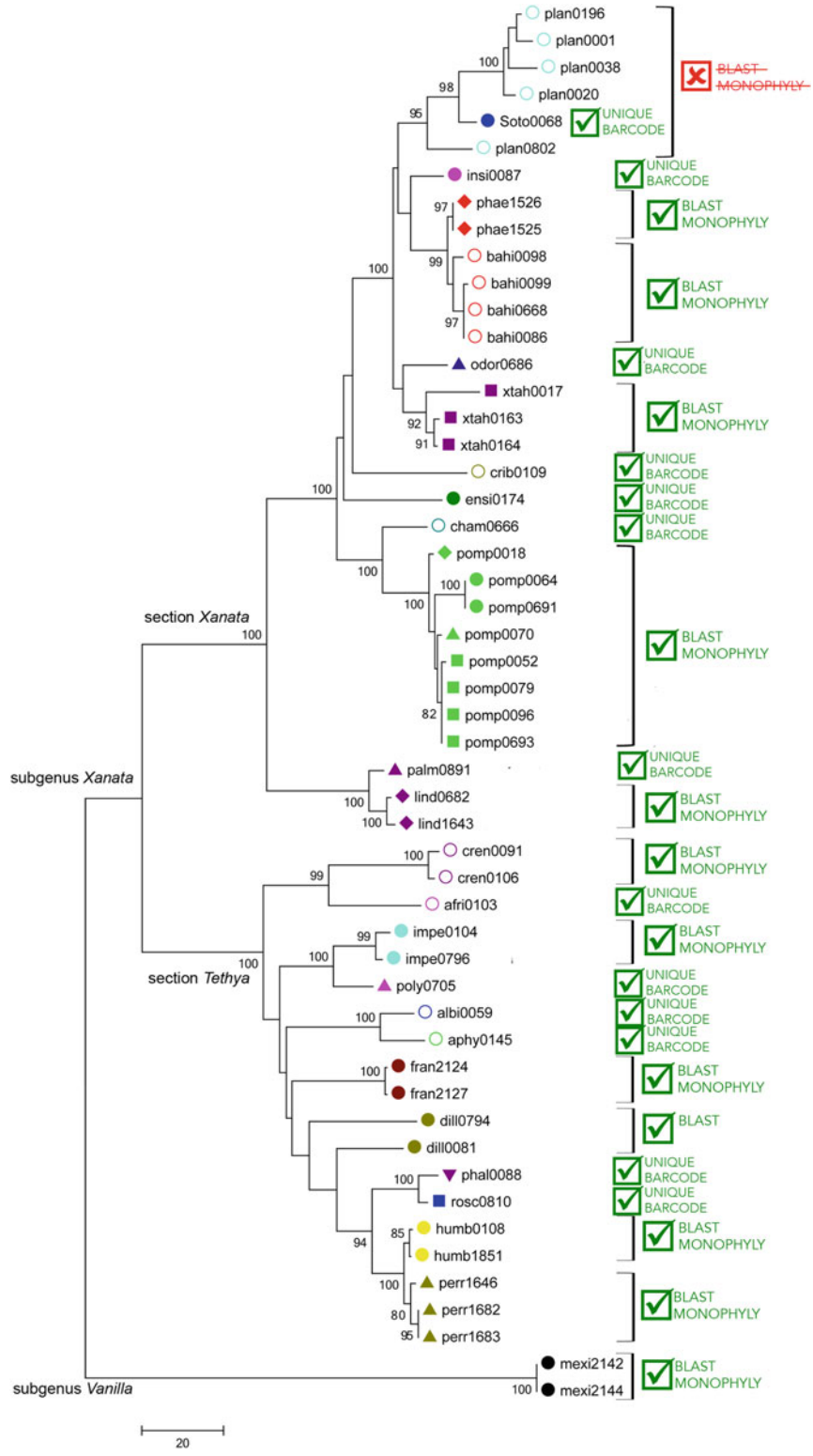


Fig. 7 (rbcL + matK + ITS) neighbor joining tree (500 bootstraps) showing the discrimination of the 26 studied *Vanilla* species. Accession codes are given as species name abbreviation followed by official CR code number from the BRC Vatel collection (e.g., plan0196 is *V. planifolia* accession CR0196). Discrimination success is shown, as detected by Blast and/or by monophyly (for species with at least two accessions) or by a unique barcode (for species represented by a single accession)

7.1 Closely Related Species

ITS allows to discriminate very closely related species from the *V. planifolia* group such as *V. bahiana*, *V. phaeantha*, and *V. insignis* (Fig. 7) which cannot be resolved using any of the tested chloroplast barcodes alone (data not shown). Similarly, some of the species studied in the present dataset, such as those from the *V. phalaenopsis* group, have evolved recently (4.4 Mya) [40]. The combination of (rbcL + matK + ITS) (Fig. 7) gives a discriminant barcode for the four tested species from this group, solely due to the use of ITS, because matK and rbcL on their own are unable of such a resolution (data not shown).

7.2 Non-monophyletic Species

Only one species could not be discriminated with the proposed system: *V. planifolia* stays unresolved because of the recently described *V. sotoarenasii* species which appears among the *V. planifolia* accessions making *V. planifolia* paraphyletic and therefore failing the Blastn test if *V. sotoarenasii* is included in the test database (Fig. 7). *V. sotoarenasii* is different morphologically but closely related to *V. planifolia*; it formed a clear-cut nested clade within *V. planifolia* using ITS but was not differentiated from *V. planifolia* using matK [47]. This pattern of a recently evolved nested new species is one obvious limit of the barcoding approach.

The (rbcL + matK + ITS) barcode system is useful to identify specimens from the three subspecies present within the *V. pompona* complex (Figs. 7 and 8). These subspecies are species that were recently reclassified to the subspecies rank [48]. When using only (rbcL + matK), *V. pompona* subsp. *grandiflora* is monophyletic (Fig. 8). ITS is more powerful to resolve the two subgroups α and γ within this subspecies, but the subsp. *grandiflora* then appears as paraphyletic, as noted previously [48] (Fig. 8). The combined barcode with the three markers here appears powerful and complementary to resolve both the species and its two subgroups.

Another interesting, but different case is the one of *V. dilloniana*. This species can only be resolved based on Blast analysis, but it is polyphyletic with our barcode system (Fig. 7). This is due to high intraspecific variation in the ITS region. Indeed, rbcL or matK on their own would resolve it as a monophyletic group (data not shown). Therefore in that case, the CBOL barcode would be more appropriate.

7.3 Hybrid Species

We previously mentioned the limit of the CBOL barcode to discriminate hybrid species. A very good example in our dataset is the one of *V.* ×*tahitensis*, which has been shown to be of hybrid origin between *V. planifolia* x *V. odorata* [38]. When using chloroplast sequences, this species is nested within the *V. planifolia* accessions (data not shown but see the phylogeny in [40] for an illustration). Only using a nuclear sequence such as ITS allows to discriminate this hybrid species from the maternal *V. planifolia* donor species

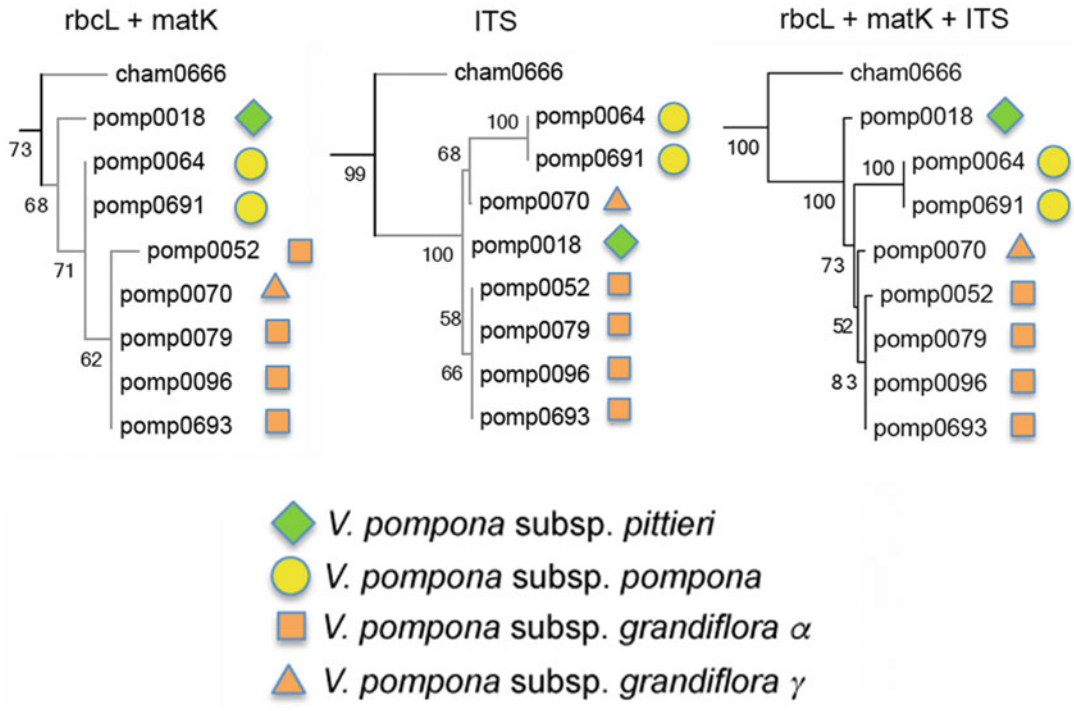


Fig. 8 The resolution power of the different barcodes selected to resolve subspecies and subgroups within the *V. pompona* species complex

(Fig. 6). It is however important to point out that our sampling of *V.* ×*tahitensis* is limited and accessions might fall in two different clusters, either related to the *V. odorata* parent (like here) or more related to the *V. planifolia* parent (as shown in [38]), making the use of a barcoding approach very limited for such hybrid species.

7.4 Conclusion

Vanilla is yet another example of seed plant for which the addition of ITS to the recommended (rbcL + matK) system is essential. However given the complexity of the genus, the absence of a clear barcoding gap, and the existence of closely related, hybrid as well as non-monophyletic species in the genus, a simple barcoding tool cannot resolve all taxonomic issues. Indeed, in some situations, a DNA barcode can be of great help, but it is unfortunately also just an additional species concept. We deeply agree, particularly for *Vanilla,* that barcoding should only be used as an aid to specimen identification (against a known and well-characterized DNA database with verified vouchers), but species delimitation and new species discovery as applied to conservation issues should use barcoding only as a triage tool in preliminary assessments, species delimitation being only possible if adding more DNA loci and doing so within an integrative taxonomy framework [44].

Acknowledgments

We thank M. Duvigneau for generating many of the sequences used and A. Vancassel for preliminary analyses of barcode data during their MSc thesis. This work was partly funded by the Vanitax (ANR Bibliothèque du Vivant) and ANR (ANR11-EBIM-005-01) and Reunion Regional Council (DGADD/PE/20120590 and 20120589) through the VaBiome project (ERA-NET-Net-Biome funded project).

References

1. Stoeckle M (2003) Taxonomy, DNA, and the bar code of life. Bioscience 53:796–797. https://doi.org/10.1641/0006-3568(2003)053[0796:TDATBC]2.0.CO;2
2. Hebert PDN, Ratnasingham S, deWaard JR (2003) Barcoding animal life: cytochrome c oxidase subunit 1 divergences among closely related species. Proc Biol Sci 270(Suppl 1): S96–S99. https://doi.org/10.1098/rsbl.2003.0025
3. Brown WM, George M, Wilson AC (1979) Rapid evolution of animal mitochondrial DNA. Proc Natl Acad Sci 76:1967–1971. https://doi.org/10.1073/pnas.76.4.1967
4. Avise JC (2000) Phylogeography: the history and formation of species. Harvard university press, Cambridge
5. Kress WJ, Wurdack KJ, Zimmer EA, Weigt LA, Janzen DH (2005) Use of DNA barcodes to identify flowering plants. Proc Natl Acad Sci 102:8369–8374. https://doi.org/10.1073/pnas.0503123102
6. Chase MW, Salamin N, Wilkinson M, Dunwell JM, Kesanakurthi RP, Haidar N, Savolainen V (2005) Land plants and DNA barcodes: short-term and long-term goals. Philos Trans R Soc B Biol Sci 360:1889–1895. https://doi.org/10.1098/rstb.2005.1720
7. Chase MW, Cowan RS, Hollingsworth PM, van den Berg C, Madriñán S, Petersen G, Seberg O, Jørgsensen T, Cameron KM, Carine M (2007) A proposal for a standardised protocol to barcode all land plants. Taxon 56:295–299. https://doi.org/10.1002/tax.562004
8. Cowan RS, Chase MW, Kress WJ, Savolainen V (2006) 300,000 species to identify: problems, progress, and prospects in DNA barcoding of land plants. Taxon 55:611–616. https://doi.org/10.2307/25065638
9. Fazekas AJ, Burgess KS, Kesanakurti PR, Graham SW, Newmaster SG, Husband BC, Percy DM, Hajibabaei M, Barrett SCH (2008) Multiple multilocus DNA barcodes from the plastid genome discriminate plant species equally well. PLoS One 3:e2802. https://doi.org/10.1371/journal.pone.0002802
10. Hollingsworth ML, Andra Clark A, Forrest LL, Richardson J, Pennington RT, Long DG, Cowan R, Chase MW, Gaudeul M, Hollingsworth PM (2009) Selecting barcoding loci for plants: evaluation of seven candidate loci with species-level sampling in three divergent groups of land plants. Mol Ecol Resour 9:439–457. https://doi.org/10.1111/j.1755-0998.2008.02439.x
11. Kress WJ, Erickson DL (2007) A two-locus global DNA barcode for land plants: the coding rbcL gene complements the non-coding trnH-psbA spacer region. PLoS One 2:e508. https://doi.org/10.1371/journal.pone.0000508
12. Lahaye R, van der Bank M, Bogarin D, Warner J, Pupulin F, Gigot G, Maurin O, Duthoit S, Barraclough TG, Savolainen V (2008) DNA barcoding the floras of biodiversity hotspots. Proc Natl Acad Sci U S A 105:2923–2928. https://doi.org/10.1073/pnas.0709936105
13. Newmaster S, Fazekas A, Ragupathy S (2006) DNA barcoding in land plants: evaluation of rbcL in a multigene tiered approach. Botany 84:335–341. https://doi.org/10.1139/b06-047
14. Taberlet P, Coissac E, Pompanon F, Gielly L, Miquel C, Valentini A, Vermat T, Corthier G, Brochmann C, Willerslev E (2007) Power and limitations of the chloroplast trn L (UAA) intron for plant DNA barcoding. Nucleic Acids Res 35:e14–e14. https://doi.org/10.1093/nar/gkl938
15. Presting GG (2006) Identification of conserved regions in the plastid genome: implications for DNA barcoding and biological function. Botany 84:1434–1443. https://doi.org/10.1139/b06-117
16. Rubinoff D, Cameron S, Will K (2006) Are plant DNA barcodes a search for the holy

grail? Trends Ecol Evol 21:1–2. https://doi.org/10.1016/j.tree.2005.10.019

17. Hollingsworth PM, Graham SW, Little DP (2011) Choosing and using a plant DNA barcode. PLoS One 6:e19254. https://doi.org/10.1371/journal.pone.0019254
18. CBOL Plant Working Group, Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M, Chase MW, Cowan RS, Erickson DL (2009) A DNA barcode for land plants. Proc Natl Acad Sci 106:12794–12797. https://doi.org/10.1073/pnas.0905845106
19. Ratnasingham S, Hebert PD (2007) BOLD: the barcode of life data system (http://www.barcodinglife.org). Mol Ecol Notes 7:355–364. https://doi.org/10.1111/j.1471-8286.2007.01678.x
20. Liu D, Liu L, Guo G, Wang W, Sun Q, Parani M, Ma J (2013) BOLDMirror: a global mirror system of DNA barcode data. Mol Ecol Resour 13:991–995. https://doi.org/10.1111/1755-0998.12048
21. Fazekas AJ, Kesanakurti PR, Burgess KS, Percy DM, Graham SW, Barrett SCH, Newmaster SG, Hajibabaei M, Husband BC (2009) Are plant species inherently harder to discriminate than animal species using DNA barcoding markers? Mol Ecol Resour 9(Suppl s1):130–139. https://doi.org/10.1111/j.1755-0998.2009.02652.x
22. Li D-Z, Gao L-M, Li H-T, Wang H, Ge X-J, Liu J-Q, Chen Z-D, Zhou S-L, Chen S-L, Yang J-B, Fu C-X, Zeng C-X, Yan H-F, Zhu Y-J, Sun Y-S, Chen S-Y, Zhao L, Wang K, Yang T, Duan G-W (2011) Comparative analysis of a large dataset indicates that internal transcribed spacer (ITS) should be incorporated into the core barcode for seed plants. Proc Natl Acad Sci U S A 108:19641–19646. https://doi.org/10.1073/pnas.1104551108
23. Hollingsworth PM (2011) Refining the DNA barcode for land plants. Proc Natl Acad Sci U S A 108:19451–19452. https://doi.org/10.1073/pnas.1116812108
24. Chen S, Yao H, Han J, Liu C, Song J, Shi L, Zhu Y, Ma X, Gao T, Pang X, Luo K, Li Y, Li X, Jia X, Lin Y, Leon C (2010) Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. PLoS One 5:e8613. https://doi.org/10.1371/journal.pone.0008613
25. Yao H, Song J, Liu C, Luo K, Han J, Li Y, Pang X, Xu H, Zhu Y, Xiao P, Chen S (2010) Use of ITS2 region as the universal DNA barcode for plants and animals. PLoS One 5:e13102. https://doi.org/10.1371/journal.pone.0013102
26. Coleman AW (2003) ITS2 is a double-edged tool for eukaryote evolutionary comparisons. Trends Genet 19:370–375. https://doi.org/10.1016/S0168-9525(03)00118-5
27. Wang X-C, Liu C, Huang L, Bengtsson-Palme-J, Chen H, Zhang J-H, Cai D, Li J-Q (2015) ITS1: a DNA barcode better than ITS2 in eukaryotes? Mol Ecol Resour 15:573–586. https://doi.org/10.1111/1755-0998.12325
28. Li H, Bai H, Yu S, Han M, Ning K (2018) Holmes-ITS2: consolidated ITS2 resources and search engines for plant DNA-based marker analyses. bioRxiv 263541. https://doi.org/10.1101/263541
29. Merget B, Koetschan C, Hackl T, Förster F, Dandekar T, Müller T, Schultz J, Wolf M (2012) The ITS2 database. J Vis Exp (61): e3806. https://doi.org/10.3791/3806
30. Ankenbrand MJ, Keller A, Wolf M, Schultz J, Förster F (2015) ITS2 database V: twice as much. Mol Biol Evol 32:3030–3032. https://doi.org/10.1093/molbev/msv174
31. Arenas MS, Cameron K (2003) Vanilla. In: Genera Orchidacearum: Orchidoideae. Oxford University Press, New York, pp 321–334
32. Arenas MAS, Cribb P (2009) A new infrageneric classification and synopsis of the genus *Vanilla* Plum. ex Mill.(Orchidaceae: Vanillinae). Lankesteriana 9:355–398. https://doi.org/10.15517/lank.v0i0.12071
33. Bory S, Brown S, Duval M-F, Besse P (2010) Evolutionary processes and diversification in the genus Vanilla. In: Vanilla. Crc Press, Florida, pp 15–29
34. Ennos RA, French GC, Hollingsworth PM (2005) Conserving taxonomic complexity. Trends Ecol Evol 20:164–168. https://doi.org/10.1016/j.tree.2005.01.012
35. Rodolphe G, Séverine B, Michel G, Pascale B (2011) Biodiversity and evolution in the Vanilla genus. In: The dynamical processes of biodiversity –case studies of evolution and spatial distribution, pp 1–27
36. Nielsen LR, Siegismund HR (1999) Interspecific differentiation and hybridization in *Vanilla* species (Orchidaceae). Heredity 83:560–567. https://doi.org/10.1046/j.1365-2540.1999.00588.x
37. Nielsen LR (2000) Natural hybridization between *Vanilla claviculata* (W. Wright) Sw. and *V. barbellata* Rchb. f.(Orchidaceae): genetic, morphological, and pollination experimental data. Bot J Linn Soc 133:285–302. https://doi.org/10.1111/j.1095-8339.2000.tb01547.x
38. Lubinsky P, Cameron KM, Molina MC, Wong M, Lepers-Andrzejewski S, Gómez-Pompa A, Kim S (2008) Neotropical roots of

a Polynesian spice: the hybrid origin of Tahitian vanilla, Vanilla tahitensis (Orchidaceae). Am J Bot 95:1040–1047. https://doi.org/10.3732/ajb.0800067

39. Bory S, Catrice O, Brown S, Leitch IJ, Gigant R, Chiroleu F, Grisoni M, Duval M-F, Besse P (2008) Natural polyploidy in Vanilla planifolia (Orchidaceae). Genome 51:816–826. https://doi.org/10.1139/G08-068
40. Bouetard A, Lefeuvre P, Gigant R, Bory S, Pignal M, Besse P, Grisoni M (2010) Evidence of transoceanic dispersion of the genus *Vanilla* based on plastid DNA phylogenetic analysis. Mol Phylogenet Evol 55:621–630. https://doi.org/10.1016/j.ympev.2010.01.021
41. Duvignau M (2012) Utilisation de marqueurs moléculaires pour une avancée dans la résolution de la phylogénie du genre# Vanilla. Master 2 thesis, Montpellier 2
42. Gigant R (2008) L'aphyllie dans le genre Vanilla: étude phylogénétique des espèces de l'Océan Indien. Master 2 thesis Biodiversité et écosystèmes tropicaux, Université de la Réunion
43. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, Kumar S (2011) MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28:2731–2739. https://doi.org/10.1093/molbev/msr121
44. Collins RA, Cruickshank RH (2013) The seven deadly sins of DNA barcoding. Mol Ecol Resour 13:969–975. https://doi.org/10.1111/1755-0998.12046
45. Meyer CP, Paulay G (2005) DNA barcoding: error rates based on comprehensive sampling. Plos Biol 3:e422. https://doi.org/10.1371/journal.pbio.0030422
46. Blagoev G, Hebert P, Adamowicz S, Robinson E (2009) Prospects for using DNA barcoding to identify spiders in species-rich genera. ZooKeys 16:27. https://doi.org/10.3897/zookeys.16.239
47. Azofeifa-Bolaños JB, Gigant LR, Nicolás-García M, Pignal M, Tavares-González FB, Hágsater E, Salazar-Chávez GA, Reyes-López D, Archila-Morales FL, García-García JA (2017) A new vanilla species from Costa Rica closely related to *V. planifolia* (Orchidaceae). Eur J Taxon 284:1–26. https://doi.org/10.5852/ejt.2017.284
48. Soto-Arenas MA, Dressler RL (2010) A revision of the Mexican and Central American species of *Vanilla* Plumier ex Miller with a characterization of their ITS region of the nuclear ribosomal DNA. Lankesteriana 9:285–354. https://doi.org/10.15517/lank.v0i0.12065

Chapter 9

High-Throughput Genotyping Technologies in Plant Taxonomy

Monica F. Danilevicz, Cassandria G. Tay Fernandez, Jacob I. Marsh, Philipp E. Bayer, and David Edwards

Abstract

Molecular markers provide researchers with a powerful tool for variation analysis between plant genomes. They are heritable and widely distributed across the genome and for this reason have many applications in plant taxonomy and genotyping. Over the last decade, molecular marker technology has developed rapidly and is now a crucial component for genetic linkage analysis, trait mapping, diversity analysis, and association studies. This chapter focuses on molecular marker discovery, its application, and future perspectives for plant genotyping through pangenome assemblies. Included are descriptions of automated methods for genome and sequence distance estimation, genome contaminant analysis in sequence reads, genome structural variation, and SNP discovery methods.

Key words Single nucleotide polymorphism, SNP, Presence and absence variation, PAV, Pangenome, Mash, Phylogenetic

1 Introduction

Next-generation sequencing technology has provided cost-effective approaches to large-scale resequencing of plant genotypes, enabling DNA barcoding and revolutionizing ecological and taxonomic plant studies [1, 2]. DNA barcoding is a technique that characterizes species using short conserved DNA sequences. This sequencing technology has also enabled ancient DNA analysis, providing information about the former plant communities and the environmental conditions present at that period [3, 4]. Molecular markers are DNA tools complementary to phenotypic analyses; they allow characterization of the underlying genetic variation between different individuals. The majority of the genetic variations are not visible at the phenotypic level, though they can assist the assessment of plant communities [5, 6].

Single nucleotide polymorphisms (SNPs) have emerged as the most widely used genotyping markers for plants [7]. SNP markers

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_9,

are abundant, heritable, and unaffected by the environment. Genome structure variations such as copy number variations (CNVs) and and insertion/deletions (indels) are also used as molecular markers. Comparing structural variation enables the estimation of linkage disequilibrium, syntenic analyses, as well as the study of genome rearrangement across taxa [8]. It is possible to infer statistical associations between genetic markers and environmental variables within a species through the use of molecular markers [9–12]. Molecular markers have numerous applications in plant taxonomy, as genome variation analysis can help unravel complex genetic structures and map plant populations. Molecular markers also have the potential to broaden our understanding of demographic histories and genetic evolution. Novel markers can expand our ability to reliably identify and characterize individuals from genomic samples, which may be particularly useful for investigating remote populations.

1.1 What Are SNPs?

Single nucleotide polymorphisms (SNPs) are the most frequent forms of genetic variation. They are single nucleotide differences in the DNA strand at specific loci [13]. SNPs are preferred for plant genetic and genomic analyses because they are widely distributed throughout the genome and thus capable of providing a high density of markers near a loci of interest [14]. Previous studies have estimated that SNPs appear every 100–300 bp in most crops [13, 15–17]. They present codominant inheritance and chromosome-specific location and are highly reproducible due to their low mutation rate [18]. The aforementioned characteristics have led to the widespread use of SNPs in studies, including demographic inferences about species and genotype distribution in a given location. The distribution characterization in turn can shed light on the roles of multiple climatic and geological events shaping the population structure [19].

SNPs are excellent markers for studying complex genetic traits, providing genetic diagnostics, and germplasm identification. SNPs have been used extensively in crop studies, and several databases are available to investigate these variations such as cropSNPdb for *Brassica* and bread wheat [20]; autoSNPdb for SNP identification in barley, rice, and *Brassica* [14]; Panzea for maize genotypes [21]; and CerealsDB 3.0 for data on cereal crops [22].

Recently, a method for genotyping ancient DNA based on SNP identification has been developed using genotyping by sequencing [23]. SNPs can assist in differentiating related sequences, both within an individual and between individuals within a population. SNPs are direct markers for tracing the nature of allelic variants, which can be useful in comparison studies between wild and related domesticated species, potentially revealing novel allelic variants suitable for introgression into improved crop varieties [9]. A study assessed the genomic variation of 3010 rice genomes,

identifying 29 million SNPs and other genetic variations. Through the analysis of SNPs and other molecular markers, they were able to suggest several previously unreported rice subpopulations that correlate to their geographic location [24]. In wheat, the analysis of SNPs from 4506 genomes enabled the reconstruction of wheat domestication history and phylogeography, using landrace and cultivar varieties from 105 countries [25].

1.2 Gene Presence and Absence Variation

The plant genome contains significant structural variation between different genotypes. These structural variations may occur in the form of copy number variations (CNVs), inversions, translocations, and presence/absence variations (PAVs). PAV is considered an extreme form of copy number variation, in which a segment of the genome is entirely absent in some individuals. Structural variations contribute substantially to the genetic diversity of major species, as these variable regions are often associated with phenotypic traits [26–28]. It is estimated that 20% of the soybean and maize whole genomes are variable [29, 30], whereas in *Brassica oleracea*, PAV affects 20% of the genes [26] and in wheat approximately 40% of the genome is variable [27]. The soybean genome contains roughly 5000 large PAVs (>500 bp) across wild and domesticated populations [31], with 133 genes present in PAV regions [32]. Researchers identified a cluster of ten genes affected by PAVs in *Papaver somniferum* (opium poppy); this gene cluster controls the production of noscapine, an antitumor alkaloid. It was observed that structural variation is able to decrease or even halt the production of noscapine, demonstrating the economic potential of PAV characterization [33].

PAV analysis can be used to trace the recombination history within a distinct species or population. The comparison of geographically diverse mitochondrial genomes from *Silene noctiflora*, *S. turkestanica*, and *S. undulata* showed the extensive fragmentation events in the mtDNA and its assortment profile. Through analyzing the structural variation of mtDNA, it was possible to reconstruct a sexual-like recombination history between the species [34]. In recent studies, PAV analysis enabled the ancestry estimation of protein domain families in the *Solanaceae* family [35] and uncovered a rare allele present in a PAV region capable of regulating fruit flavor [36]. The characterization of PAV regions may assist in understanding the complex relationships between variation in gene family size and its rate of evolution. Genes identified in PAV regions can be directly linked to phenotype and fitness and can be used to trace a gene's history to a significant ecological event [37]. Novel PAV discovery is especially useful for breeding crops, which as a result of artificial selection often have lower genetic diversity compared to their wild relatives.

1.3 Molecular Marker Applications in Plant Taxonomy

Molecular markers are based on alterations of the DNA sequence that may be associated with the plant's phenotype, such as SNPs and structural variations. The increasing availability of sequence data, generated from genome sequencing projects, provides a valuable resource for the discovery of novel molecular markers associated with phenotypes of interest. The growing availability of DNA data improves the coverage of intraspecific genome variability and informs the geographical distribution of varieties. Molecular markers can be integrated into tools for fast and accurate identification of plant taxonomy and even in situ identification of plant varieties. For instance, a taxonomic method for the classification of the *Aurantioideae* subfamily has been developed using SNPs from chloroplast genes [38]. This method enables a cost-effective cladistic analysis in large collections at a subfamily level, which was not feasible with previous methods, and can easily be expanded to classify other plant species [38]. Another application of molecular markers is the characterization of ancient DNA. Ancient DNA samples are often highly degraded; therefore the use of SNPs can be advantageous as small fragments can be amplified and compared [39]. In the Tehuacán valley (Mexico), specimens of maize dating at a similar age of 5300–4970 y B.P. were genetically analyzed uncovering that the earliest maize from San Marcos was already inbred, possibly from an isolated founder population [40].

The identification of new molecular markers largely depends on the use of appropriate bioinformatic tools and well-curated marker databases, some of which have been reviewed by Scheben et al. [41] and Singh, Singh [42].

1.4 Pangenomes as the Future of Molecular Markers and Gene Variance Identification

With increased access to genomic data, it became clear that a substantial portion of genome varies between individuals within the same species, suggesting that a single reference genome is not sufficient to represent its genetic diversity [26, 43, 44]. Recently, there has been an increase in the number of pangenomes assembled, which requires sequencing, assembly, and comparison of several lineages to characterize genetic variation. Currently, the majority of the software packages have been designed for the analysis of microbial pangenomes, reviewed in Xiao et al. [45], although some can be adapted for the analysis of plants [26].

A pangenome represents a collection of individuals, in which the genomic region present in all individuals is labelled "core," whereas the genomic portion exclusive to some of the individuals is "dispensable" [46]. The *Brassica* and the soybean pangenomes have been shown to exhibit a higher density of SNPs in dispensable compared to core regions, and this trend seems to be followed in other pangenomes [26, 29, 47].

The majority of research has focused on SNP identification by comparing a single or multiple individuals to a reference genome. However, this approach fails to identify SNPs in the dispensable

region, when they are absent in the reference genome [43]. The use of pangenome references allows for the recognition of SNPs occurring in both the core and dispensable regions. The discovery of SNPs in dispensable regions can facilitate the identification of novel alleles and the characterization of novel metabolic pathways [47]. Markers from the dispensable region may also aid the discovery of molecular fingerprinting targets for population genetic studies and the reconstruction of phylogenetic histories [48]. The pangenome is an ideal model for the characterization of CNVs and PAVs, as it defines the core genome and the variations occurring at the dispensable region [49]. The increased level of information contained within pangenomes, particularly regarding gene structure, SNPs, and variable regions, expands our understanding of how dispensable regions evolve, which has potential for improving the resolution of plant phylogenies [48, 50].

1.5 Tools for Plant Genotyping and Taxonomy

1.5.1 Identification of SNPs In Silico

The major challenge of SNP discovery is not their identification but the differentiation of true SNPs from the often more abundant sequence errors. It is estimated that $6.4 \pm 1.24\%$ of sequences are mutated during Illumina sequencing and $0.24 \pm 0.06\%$ errors per base occur, most of which are single nucleotide substitutions [51, 52]. There are several potential sources of error; therefore it is essential to perform a stringent quality assessment during read processing to differentiate between sequence errors and true polymorphisms [52]. SNP genotyping methods usually begin with read quality trimming, followed by mapping, processing of the mapped reads, variant calling, and finally variant filtering [53]. The recurrence of a polymorphism at a particular loci increases the confidence of the SNP being a true polymorphism; however several other filtering steps can be implemented to increase confidence in assaying a true variance [54].

The tools for SNP calling are mostly heuristic-based or probability-based algorithms, both relying heavily on the abundance and quality of data [43]. The two primary types of SNP calling tools are haplotype-based callers and single site-based callers such as Samtools/BCFtools [55]. The choice of software and quality restrictions can greatly impact the proportion of SNPs encountered, as most of the variant callers substantially disagree on the SNPs and other structural variations found [54, 56, 57]. Variant calling tools used for calling SNPs and other types of sequence variation have enabled the discovery of population-specific polymorphisms used for a wide range of applications including phylogenetic studies [58]. The continued advancement of genome sequencing technology and pangenome assembly ensures that tools for variant calling will continue to be in high demand and will be expanded to meet future requirements.

1.5.2 Distance Estimation Methods in Genotyping

The discovery of molecular markers is highly dependent on which genotype is used as reference, which makes the choice of reference genome very important [43]. Sequence distance estimation can be applied to choose the most appropriate reference genome; it can also assist in the clustering of similar genotypes. Distance estimation analysis can be applied to whole-genome phylogenies, classification of protein families, identification of horizontally transferred genes, and detection of recombined sequences [59]. There are two main types of tools to cluster and perform distance estimation on large genomic data: alignment-dependent and alignment-free tools. The alignment-dependent tools are based on strict evolutionary assumptions that may not be reflected by the reality of living organisms [59]. These assumptions can hinder proper clustering of similar sequences. For instance, protein superfamily sequences sometimes fail to cluster due to their highly variable primary sequences, and even though their tridimensional structure is conserved, it is hard to evaluate through assembled sequences [60]. Alignment-dependent tools are subject to a substantial decrease in accuracy if gaps are allowed, particularly impacting nucleotide comparisons [61]. They are computationally demanding and time-consuming, which can limit the amount of data used, inhibiting multi-genome scale data analysis [59]. In contrast, alignment-free tools do not rely on dynamic programming, which makes them less computationally demanding than alignment-dependent tools [62]. This makes alignment-free tools particularly useful for large sequencing data estimations, and they will likely become the preferred tools for future genomic data management. In addition, alignment-free tools do not depend on evolutionary sequence assumptions, which can hamper clustering of similar sequences presenting structural variation [59]. Alfree is one of many freely available web tools that can be used to run small alignment-free analysis [59]. This web application has 27 tools available for testing and has the ability to process a maximum of 50 sequences of 200,000 nucleotides/amino acids.

For larger data analyses, there are other more powerful alignment-free tools that may assist in deciding upon the most appropriate reference genome for SNP discovery, to perform a rapid triage, cluster data, and assign species labels to mixed samples and to identify mis-tracked or low-quality samples [63–67]. Alignment-free tools are able to rapidly assess clusters of thousands of genomes at a time, enabling the identification of outlier varieties [68]. Alignment-free tools have great potential to be applied to large-scale genomic management and emerging long-read, single-molecule sequencing technologies.

2 Materials

2.1 Bioinformatics Requirements for Mash Analyses

The analysis must be run in the command line from a Linux or OS X machine. Download and install the following binary files:

1. Sample files in FASTQ and MSH format are available as Data_R1.fastq, Data_R2.fastq, Data 2_R1_val_1.fq.msh, and Data 3_R1_val_1.fq.msh at our group website, to be used in this analysis (http://appliedbioinformatics.com.au/index.php/Sample_data).
2. Trim Galore version 0.5.0 (https://github.com/FelixKrueger/TrimGalore) is a flexible pipeline which includes Cutadapt [69] and FastQC [70] for trimming the adaptors from raw sequencing reads and assessing the quality of the remaining data [71].
3. Mash version 2.1.1 (available at https://github.com/marbl/Mash) is a tool kit capable of generating sketch files, estimating distance using MinHash, and genome contamination screening [66].
4. A reference sketch file created from microorganism species, publicly available at https://gembox.cbcb.umd.edu/Mash/refseq.genomes.k21s1000.msh courtesy of Ondov et al. [66].

2.2 Bioinformatics Requirements for SNP Identification Using BCFtools

The variant calling analysis requires (a) sequence reads from any sequencer machine in FASTQ format and (b) a current assembled genome of the reference species in FASTA format. The analysis must be run in the command line on a Linux or OS X machine; download the files and install the following tools:

1. Sample sequence files in FASTQ format are available as Seq_R1.fastq and Seq_R2.fastq, at our group website for download and usage in this analysis. The sample files are from *Brassica oleracea*; thus the *B. oleracea* reference genome must be downloaded to perform the read assembly. The sample files are available at http://appliedbioinformatics.com.au/index.php/Sample_data, the *B. oleracea* genome can be downloaded from http://plants.ensembl.org/Brassica_oleracea/Info/Index, and alternatively you can use the *B. oleracea* pangenome as reference, available at http://www.brassicagenome.net/databases.php.
2. HISAT2 version 2.1.0 (https://ccb.jhu.edu/software/hisat2/manual.shtml#obtaining-hisat2) [72] is an alignment program for mapping next-generation sequencing reads from any sequencer machine. HISAT2 requires a reference genome.
3. Samtools version 1.2 (available at http://samtools.sourceforge.net/; [73]) contains a suite of utilities designed for

manipulating alignments in the SAM/BAM format, including sorting, merging, indexing, and compressing.

4. BCFtools version 1.4.1 (http://www.htslib.org/doc/bcftools.html) is a flexible data management program to manage SNPs and Indels in VCF or BCF format.
5. VCFlib version 1.0.0 (https://github.com/vcflib/vcflib.git) provides a set of tools to manipulate VCF format files and perform VCF comparison, format conversion, filtering and subsetting, annotation, and ordering.

3 Methods

3.1 Mash Analyses

3.1.1 Mash for Distance Estimation

Mash is an alignment-free tool that estimates the distance between sample sequences against one or more reference sequences using MinHash. The MinHash probabilistic approach described in Broder [74] enables the comparison of large datasets such as genomic data, by dividing the sequences in small segments entitled "hash." The hash comparison allows for rapidly cluster analysis. Mash has extremely low memory and CPU requirements, making distance estimation of several genomic datasets feasible. It can use assembled or unassembled sequences as the input, which will be reduced to compressed sketch representations used for the distance calculation. The distance estimation returns the Jaccard index (i.e., the fraction of shared k-mers), *p*-value, and Mash distance, which estimates the rate of sequence mutation under a simple evolutionary model [66, 75].

i. `trim_galore -q 20 --phred33 --fastqc --illumina --length 99 --trim-n --paired Data_R1.fastq Data_R2.fastq`

Trim Galore is used to prepare raw reads for the following analysis (*see* **Note 1**). We have two input paired-end sequence files, Data_R1.fastq and Data_R2.fastq. The option "--phred33" instructs the use of ASCII+33 quality scores as Phred for quality trimming, "-q 20" defines the minimum quality score, and "--illumina" indicates the type of sequence adaptors to trim. The "--fastqc" and "--paired" indicate the type of input file. The "--length 99" establishes the minimum length of the sequence read, and "--trim-n" allows to remove the "N" bases from either side of the sequence reads. Unidentified nucleotides must be trimmed to allow a better sequence comparison. By default, the output of trim galore ends in "_val_1.fq" and _val_2.fq." In this case, it will generate two files, "Data_R1_val_1.fq" and "Data_R2_val_2.fq."

ii. `mash sketch –r –m 3 Data_R1_val_1.fq Data_R1_val_2.fq`

The Mash run is performed in two steps using "Mash sketch" followed by "Mash dist." The "Mash sketch" step must be performed individually for each sample you want to perform the distance estimation. It creates a reduced representation of the sequence as simplified k-mers "words" that are used for the distance estimation. It is important to perform this step with all the query and reference sequences, and each must be done separately. To create this sketch, you must first indicate the type of file followed by the input file name (*see* **Note 2**). The flag "-r" must be used for genomic read input, and "-m 3" is used to discard all k-mers that appear less than three times, as these are likely to be sequencing errors. Another potentially important flag is the "-s" flag, which defines the sketch size per sample. By default this is set to 1000, so 1000 min-hashes will be stored per sample. For many highly similar samples, it is recommended to increase the sketch size to 10,000 or even more in order to find rare differences. The output file will be used as input for the second Mash step. The output files end in ".msh," one per individual. In this example, Mash will generate a file named "Data_R1_val_1.fq.msh."

iii. `mash dist reference_file.msh Data_R1_val_1.fq.msh Data1_R1_val_1.fq.msh`

The "Mash dist" step estimates the global and pairwise mutation distance of each query sequence against the reference, using previously generated ".msh" files for the samples. At the distance estimation step, it is important that both reference and query files have matching k-mer sizes. In this step, it is possible to compare multiple sketched sequences against a sketched reference sequence file. More options on how to adapt the "dist" tool to your specific data type are described in **Note 3**. This command will print a table of distances; an example is given in Table 1.

3.1.2 Investigate Sample Contamination with Mash

In Mash, the "screen" can be used to quickly check for contamination in sequencing read samples. For this type of analysis, reference genomes with approximately 10x coverage should suffice. The reference genome file must contain the genome sketches of all the

Table 1
Example Mash dist output table for one reference and three samples to compare, with an added header

Reference filename	Query filename	Distance	*p*-value	Number of shared hashes
referencefile.msh	Data_R1_val_1.fq.msh	0.2630222	7.27863e-06	2/1000
referencefile.msh	Data1_R1_val_1.fq.msh	0.2630222	5.18662e-06	2/1000
referencefile.msh	Data2_R1_val_1.fq.msh	0.0759181	0	113/1000

potential contaminants for your sample (as provided in Subheading 2.1), and the file must be in MSH format following the instructions given at Subheading 3.1. The query samples can be in FASTA, FASTQ, or MSH format.

i. `mash screen -w –i 0 reseq.genomes.k21s1000.msh Data_R1_val_1.fq`

The "Mash screen" tool was designed to search for genome containment within the sample analyzed. The reference and query inputs can be single or multiple files, and the files must be separated by space for the latter. "Mash screen" uses a hash count system, in which each aligned hash counts points used to determine the genomes identified in the sample. It is possible for the same hash to align to multiple species, thus scoring points for more than one species. The "-w" option is used to reduce the redundancy of hash alignment using a winner-takes-all approach. In other words, a hash that aligns to multiple references will only add points to the best matched reference, leading to less output redundancy. The "-i" sets the minimum identity level show in the output, and "-i 0" will output identities with at least one shared hash. The output file reports the species genomes identified alongside contamination warnings where required. The results will be displayed in six columns: identity of the given reads with the reference in percentage, ranging from 0 to 1, number of shared hashes between reads and reference, median multiplicity (the median of how often each shared hash appears in the entire pool), *p*-value, query ID, and query comment (usually the number of sequences with hits and the first few hit IDs). An example output table is given in Table 2.

Table 2
Example output of the mash pool command, with an added header

Identity	Number of shared hashes	Median multiplicity	*p*-value	Query ID
0.719686	1/1000	1	0.0746507	GCF_000760155.1_ASM76015v1_genomic.fna.gz
0.719686	1/1000	1	0.0746507	GCF_000760175.1_ASM76017v1_genomic.fna.gz
0.719686	1/1000	1	0.0746507	GCF_000760235.1_ASM76023v1_genomic.fna.gz
0.719686	1/1000	2	0.0746507	GCF_000760555.1_ASM76055v1_genomic.fna.gz
0.719686	1/1000	1	0.0746507	GCF_000760675.1_R_fas_A3b_genomic.fna.gz
0.719686	1/1000	2	0.0746507	GCF_000760775.1_R_fas_A78_genomic.fna.gz
0.743837	2/1000	24	0.0028556	GCF_000760795.1_R_fas_GIC26_genomic.fna.gz
0.719686	1/1000	1	0.0746507	GCF_000760155.1_ASM76015v1_genomic.fna.gz

3.2 SNP Identification Using BCFtools

The workflow described below applies some of the latest tools for SNP discovery from sequence reads data to produce a variation file. This variant file is in a tab-delimited format that concisely describes reference-indexed variations between individuals or populations. The output file describes polymorphic loci, which can be applied for taxonomic analysis.

i. `trim_galore -q 20 --phred33 --fastqc --illumina --length 99 --trim-n --paired Seq_R1.fastq Seq_R2.fastq`

It is important to check the quality and remove adaptors from the sequence reads used in this analysis. Trim Galore can be used to prepare the sequence data for the analysis; in this step we are using two input files, Seq_R1.fastq and Seq_R2.fastq, which are paired sequence reads from Illumina sequencer (*see* **Note 1**). The option "--phred33" instructs the use of ASCII+33 quality scores as Phred for quality trimming, "-q 20" defines the minimum quality score, and "--illumina" indicates the type of sequence adaptors to trim. The "--fastqc" and "--paired" indicate the type of input file. The "--length 99" establishes the minimum length of the sequence read, and "--trim-n" allows to remove the "N" bases from either side of the sequence reads. Unidentified nucleotides must be trimmed to allow a better sequence comparison. By default, the output of Trim Galore ends in "_val_1.fq" and "_val_2.fq." In this case, it will generate two files, "Seq_R1_val_1.fq" and "Seq_R2_val_2.fq."

ii. `hisat2-build -p number_of_threads -f reference_genome.fasta output_indexed_genome ;`

HISAT2 is an alignment program for mapping next-generation sequencing reads to an indexed reference genome (*see* **Note 4**). In addition to using one global Graph Full-text index in Minute space (GFM) that represents the general population, HISAT2 uses a large set of small GFM indexes that collectively cover the whole genome [72]. The "hisat2-build" command above is used to build the index file from the reference genome.

iii. `hisat2 -p number_of_threads --no-softclip -x indexed_genome -1 Seq_R1_val_1.fq -2 Seq_R2_val_2.fq -S Seq_R1R2_val.sam`

HISAT2 is now used to map the sequence reads used for variant discovery, based on the previous indexed genome. The "--no-softclip" flag prevents soft clipping to increase mapping confidence; "-S" ensures the mapping will be stored in SAM format (*see* **Note 5**). HISAT2 outputs a report on the screen regarding the percentage of reads aligned to the reference genome, as shown in Fig. 1.

```
10533338 reads; of these:
  10533338 (100.00%) were paired; of these:
    2903059 (27.56%) aligned concordantly 0 times
    6286374 (59.68%) aligned concordantly exactly 1 time
    1343905 (12.76%) aligned concordantly >1 times
    ----
    2903059 pairs aligned concordantly 0 times; of these:
      46048 (1.59%) aligned discordantly 1 time
    ----
    2857011 pairs aligned 0 times concordantly or discordantly; of these:
      5714022 mates make up the pairs; of these:
        4159299 (72.79%) aligned 0 times
        1378922 (24.13%) aligned exactly 1 time
        175801 (3.08%) aligned >1 times
80.26% overall alignment rate
```

Fig. 1 Alignment output and mapping percentage of the reads from sample file provided by HISAT2

iv. `samtools view -b Seq_R1R2_val.sam > Seq_R1R2_val.bam`

The assembled sequences in SAM must be converted to BAM format. It is possible to quickly convert to BAM files using Samtools, one file at time. The "-b" option sets the output to be created in BAM format.

v. `samtools sort Seq_R1R2_val.bam Seq_R1R2_val.sorted.bam`

Each BAM file must be sorted individually and converted to BCF prior to variant discovery. The alignments in the BAM file will be sorted by leftmost coordinates using the command above.

vi. `samtools faidx reference_genome.fasta`

Create an index file of the genome assembly used; the index file enables efficient access to arbitrary regions within the genome reference. The "samtools faidx" tool accepts other file formats; please *see* **Note 6** for more information.

vii. `samtools mpileup -f reference_genome.fasta -q 30 -Q 20 --per-sample-mF -g -b list.txt > raw_output.bcf`

Convert the sorted BAM file to binary call format (BCF) using "samtools mpileup" (*see* **Note 7**). It is possible to use multiple BAM files (each BAM file is considered a sample), and they must be in a space delimited list. In this command the BCF file is generated. The list of input BAM files (using their full names, one file per line) is specified with the option "-b. The"-f" indicates the reference genome is in FASTA format, "-Q" sets the minimum base

quality, and "-q" sets the minimum mapping quality for an alignment. The "--per-sample-mF" flag sets (a) the minimum number of gapped reads for indel candidates and (b) the minimum fraction of gapped reads per sample, to increase sensitivity of calling. These thresholds can be set manually using "-m" and "-F" flags." The option "-g" computes genotype likelihoods and outputs them in BCF.

viii. `bcftools call --multiallelic-caller --variants-only -g -Ov -S raw_output.bcf -o variant_output.vcf`

The "bcftools call" replaces the former "bcftools view" in samtools versions >0.1.19 and will be used to create a VCF file consisting only of the variants encountered. Bcftools call uses an alternative model for multiallelic and rare-variant calling, which is recommended for most tasks (--multiallelic-caller). The option "-g" adds the VCF blocks of homozygous reference calls to the output file, and "--variants-only" outputs only variant sites. The "-Ov" sets the output to be produced in VCF format, and "-o" indicates the name of the output file.

ix. `vcffilter -f "DP > 10 & QUAL > 30" variant_output.vcf`

VCFtool performs further filtering of the samples (*see* **Note 8**). The "vcffilter" tool estimates the likelihood of a true SNP by counting the number of reads supporting the polymorphism. The minimum read number is established using "DP <10." The minimum SNPs' quality threshold is set using "QUAL >30." For the command described below, SNPs with quality scores and read alignment depth below 30 and 10, respectively, are not considered reliable [76]. This will provide you with a VCF file containing all the SNPs in the group of samples provided and can be used as input for GWAS and MAS analysis.

4 Notes

1. When using Trim Galore, check the documentation available at https://github.com/FelixKrueger/TrimGalore/blob/master/Docs/Trim_Galore_User_Guide.md for options that best suit your sequencing data, as it may have optimized settings available. The adaptor used for sequencing can be specified using -a/--adapter option; otherwise it will auto-detect whether the Illumina universal, Nextera transposase, or Illumina small RNA adapter sequence was used.

2. Mash sketch can receive one or multiple files as inputs; either case must be indicated by the option: "-i" for one fasta input, "-r" for fastq file inputs, or "-l" for a list of file inputs, which presents the specific paths to each sequence file separated one per line. If using sequence reads, the following options can be added: "-b" for Bloom filter of a defined size, "-m" to define minimum copies of each k-mer required to pass the noise filter, "-c" to display the target coverage, and "-g" to indicate the genome size for *p*-value calculation.
3. The "Mash dist" output can be modified to be displayed in table format using the "-t" option. The options "-v" and "-d" can be used to set the maximum *p*-value and the maximum distance to report, respectively.
4. Alternative software for de novo assembly is available if the species you are analyzing does not have a reference genome; in that case you can use Velvet [77], SOAPdenovo2 [78], MaSuRCA [79], ABySS [79], or others.
5. HISAT2 is the reference-based alignment tool used for this pipeline; it requires a modest amount of RAM to perform the mapping steps. HISAT2 soft clips reads by default, which can lead to false-positive alignments; therefore this option was disabled.
6. Samtools faidx tool can perform the indexing of the whole FASTA file or extract subsequence from previously indexed reference sequences which is useful to retrieve specific sequences into a separate file. The input file for indexing can be compressed in the BGZF, FASTA, or FASTQ format (if using FASTQ format, it must be indicated by using –fastq).
7. In the mpileup format, each line represents a genomic position consisting of the chromosome name, 1-based coordinate, reference base, number of reads covering the site, read bases, base qualities, and alignment mapping qualities. Information on matches, mismatches, indels, strands, mapping quality, and the start and end of reads are encoded in the read base column. More information on the mpileup options for further manipulating the data can be found at http://www.htslib.org/doc/samtools-1.2.html.
8. Quality and SNP frequency filtering help remove very low abundance variants in the sequence samples, which are often caused by sequencing/mapping errors. The parameters for SNP quality and read depth chosen here are relatively arbitrary, but often used for variant calling in plant sequencing data. The parameters should be altered based on factors including but not limited to the depth of sequencing and the quality of the sequencing reads.

References

1. Hebert PDN, Gregory TR (2005) The promise of DNA barcoding for taxonomy. Syst Biol 54(5):852–859. https://doi.org/10.1080/10635150500354886
2. Dick CW, Webb CO (2012) Plant DNA barcodes, taxonomic management, and species discovery in tropical forests. In: DNA barcodes. Springer, pp 379–393
3. Parducci L, Bennett KD, Ficetola GF, Alsos IG, Suyama Y, Wood JR, Pedersen MW (2017) Ancient plant DNA in lake sediments. New Phytol 214(3):924–942
4. Sønstebø JH, Gielly L, Brysting AK, Elven R, Edwards M, Haile J, Willerslev E, Coissac E, Rioux D, Sannier J (2010) Using next-generation sequencing for molecular reconstruction of past Arctic vegetation and climate. Mol Ecol Resour 10(6):1009–1018
5. Pornon A, Escaravage N, Burrus M, Holota H, Khimoun A, Mariette J, Pellizzari C, Iribar A, Etienne R, Taberlet P (2016) Using metabarcoding to reveal and quantify plant-pollinator interactions. Sci Rep 6:27282
6. Yesson C, Jackson A, Russell S, Williamson CJ, Brodie J (2018) SNPs reveal geographical population structure of Corallina officinalis (Corallinaceae, Rhodophyta). Eur J Phycol 53(2):180–188. https://doi.org/10.1080/09670262.2017.1402373
7. Batley J, Edwards D (2007) SNP applications in plants. In: Association mapping in plants. Springer, pp 95–102
8. Tang H, Bowers JE, Wang X, Ming R, Alam M, Paterson AH (2008) Synteny and collinearity in plant genomes. Science 320(5875):486–488
9. Andrello M, Henry K, Devaux P, Verdelet D, Desprez B, Manel S (2017) Insights into the genetic relationships among plants of Beta section Beta using SNP markers. Theor Appl Genet 130(9):1857–1866
10. Brito PH, Edwards SV (2009) Multilocus phylogeography and phylogenetics using sequence-based markers. Genetica 135(3):439–455. https://doi.org/10.1007/s10709-008-9293-3
11. Koch MA, Kiefer C (2006) Molecules and migration: biogeographical studies in cruciferous plants. Plant Syst Evol 259(2):121–142. https://doi.org/10.1007/s00606-006-0416-y
12. Atwell S, Huang YS, Vilhjálmsson BJ, Willems G, Horton M, Li Y, Meng D, Platt A, Tarone AM, Hu TT (2010) Genome-wide association study of 107 phenotypes in Arabidopsis thaliana inbred lines. Nature 465(7298):627
13. Edwards D, Forster JW, Chagné D, Batley J (2007) What are SNPs? In: Association mapping in plants. Springer, pp 41–52
14. Duran C, Appleby N, Clark T, Wood D, Imelfort M, Batley J, Edwards D (2009) AutoSNPdb: an annotated single nucleotide polymorphism database for crop plants. Nucleic Acids Res 37(suppl_1):D951–D953
15. Bhattramakki D, Dolan M, Hanafey M, Wineland R, Vaske D, Register JC, Tingey SV, Rafalski A (2002) Insertion-deletion polymorphisms in 3′ regions of maize genes occur frequently and can be used as highly informative genetic markers. Plant Mol Biol 48(5):539–547. https://doi.org/10.1023/a:1014841612043
16. Bundock PC, Eliott FG, Ablett G, Benson AD, Casu RE, Aitken KS, Henry RJ (2009) Targeted single nucleotide polymorphism (SNP) discovery in a highly polyploid plant species using 454 sequencing. Plant Biotechnol J 7(4):347–354. https://doi.org/10.1111/j.1467-7652.2009.00401.x
17. Hayashi K, Hashimoto N, Daigen M, Ashikawa I (2004) Development of PCR-based SNP markers for rice blast resistance genes at the Piz locus. Theor Appl Genet 108(7):1212–1220
18. Scheben A, Batley J, Edwards D (2017) Genotyping-by-sequencing approaches to characterize crop genomes: choosing the right tool for the right application. Plant Biotechnol J 15(2):149–161
19. Reyes-Velasco J, Manthey JD, Bourgeois Y, Freilich X, Boissinot S (2018) Revisiting the phylogeography, demography and taxonomy of the frog genus Ptychadena in the Ethiopian highlands with the use of genome-wide SNP data. PLoS One 13(2):e0190440
20. Scheben A, Verpaalen B, Lawley CT, Chan CKK, Bayer PE, Batley J, Edwards D (2018) CropSNPdb: a database of SNP array data for Brassica crops and hexaploid bread wheat. Plant J 98(1):142–152
21. Zhao W, Canaran P, Jurkuta R, Fulton T, Glaubitz J, Buckler E, Doebley J, Gaut B, Goodman M, Holland J (2006) Panzea: a database and resource for molecular and functional diversity in the maize genome. Nucleic Acids Res 34(suppl_1):D752–D757
22. Wilkinson PA, Winfield MO, Barker GLA, Tyrrell S, Bian X, Allen AM, Burridge A, Coghill JA, Waterfall C, Caccamo M (2016)

CerealsDB 3.0: expansion of resources and data integration. BMC Bioinformatics 17(1):256

23. Suyama Y, Matsuki Y (2015) MIG-seq: an effective PCR-based method for genome-wide single-nucleotide polymorphism genotyping using the next-generation sequencing platform. Sci Rep 5:16963
24. Wang W, Mauleon R, Hu Z, Chebotarov D, Tai S, Wu Z, Li M, Zheng T, Fuentes RR, Zhang F, Mansueto L, Copetti D, Sanciangco M, Palis KC, Xu J, Sun C, Fu B, Zhang H, Gao Y, Zhao X, Shen F, Cui X, Yu H, Li Z, Chen M, Detras J, Zhou Y, Zhang X, Zhao Y, Kudrna D, Wang C, Li R, Jia B, Lu J, He X, Dong Z, Xu J, Li Y, Wang M, Shi J, Li J, Zhang D, Lee S, Hu W, Poliakov A, Dubchak I, Ulat VJ, Borja FN, Mendoza JR, Ali J, Li J, Gao Q, Niu Y, Yue Z, Naredo MEB, Talag J, Wang X, Li J, Fang X, Yin Y, Glaszmann J-C, Zhang J, Li J, Hamilton RS, Wing RA, Ruan J, Zhang G, Wei C, Alexandrov N, McNally KL, Li Z, Leung H (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557(7703):43–49. https://doi.org/10.1038/s41586-018-0063-9
25. Balfourier F, Bouchet S, Robert S, De Oliveira R, Rimbert H, Kitt J, Choulet F, Paux E (2019) Worldwide phylogeography and history of wheat genetic diversity. Sci Adv 5(5):eaav0536. https://doi.org/10.1126/sciadv.aav0536
26. Golicz AA, Bayer PE, Barker GC, Edger PP, Kim H, Martinez PA, Chan CKK, Severn-Ellis-A, McCombie WR, Parkin IAP, Paterson AH, Pires JC, Sharpe AG, Tang H, Teakle GR, Town CD, Batley J, Edwards D (2016) The pangenome of an agronomically important crop plant Brassica oleracea. Nat Commun 7:13390
27. Montenegro JD, Golicz AA, Bayer PE, Hurgobin B, Lee H, Chan C-KK, Visendi P, Lai K, Doležel J, Batley J, Edwards D (2017) The pangenome of hexaploid bread wheat. Plant J 90(5):1007–1013. https://doi.org/10.1111/tpj.13515
28. Zhang Y, Xia R, Kuang H, Meyers BC (2016) The diversification of plant NBS-LRR defense genes directs the evolution of MicroRNAs that target them. Mol Biol Evol 33(10):2692–2705. https://doi.org/10.1093/molbev/msw154
29. Li Y-H, Zhou G, Ma J, Jiang W, Jin L-G, Zhang Z, Guo Y, Zhang J, Sui Y, Zheng L, Zhang S-S, Zuo Q, Shi X-H, Li Y-F, Zhang W-K, Hu Y, Kong G, Hong H-L, Tan B, Song J, Liu Z-X, Wang Y, Ruan H, CKL Y, Liu J, Wang H, Zhang L-J, Guan R-X, Wang K-J, Li W-B, Chen S-Y, Chang R-Z, Jiang Z, Jackson SA, Li R, Qiu L-J (2014) De novo assembly of soybean wild relatives for pan-genome analysis of diversity and agronomic traits. Nat Biotechnol 32(10):1045–1052. https://doi.org/10.1038/nbt.2979
30. Morgante M, De Paoli E, Radovic S (2007) Transposable elements and the plant pan-genomes. Curr Opin Plant Biol 10(2):149–155
31. Lam H-M, Xu X, Liu X, Chen W, Yang G, Wong F-L, Li M-W, He W, Qin N, Wang B (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection. Nat Genet 42(12):1053
32. McHale LK, Haun WJ, Xu WW, Bhaskar PB, Anderson JE, Hyten DL, Gerhardt DJ, Jeddeloh JA, Stupar RM (2012) Structural variants in the soybean genome localize to clusters of biotic stress-response genes. Plant Physiol 159(4):1295–1308
33. Winzer T, Gazda V, He Z, Kaminski F, Kern M, Larson TR, Li Y, Meade F, Teodor R, Vaistij FE (2012) A Papaver somniferum 10-gene cluster for synthesis of the anticancer alkaloid noscapine. Science 336(6089):1704–1708
34. Wu Z, Sloan DB (2019) Recombination and intraspecific polymorphism for the presence and absence of entire chromosomes in mitochondrial genomes. Heredity 122(5):647
35. Wang P, Moore BM, Panchy NL, Meng F, Lehti-Shiu MD, Shiu S-H (2018) Factors influencing gene family size variation among related species in a plant family, Solanaceae. Genome Biol Evol 10(10):2596–2613. https://doi.org/10.1093/gbe/evy193
36. Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM, Burzynski-Chang EA, Fish TL, Stromberg KA, Sacks GL, Thannhauser TW, Foolad MR, Diez MJ, Blanca J, Canizares J, Xu Y, van der Knaap E, Huang S, Klee HJ, Giovannoni JJ, Fei Z (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet 51(6):1044–1051. https://doi.org/10.1038/s41588-019-0410-2
37. Dlugosch KM, Parker IM (2008) Founding events in species invasions: genetic variation, adaptive evolution, and the role of multiple introductions. Mol Ecol 17(1):431–449. https://doi.org/10.1111/j.1365-294X.2007.03538.x
38. Oueslati A, Ollitrault F, Baraket G, Salhi-Hannachi A, Navarro L, Ollitrault P (2016) Towards a molecular taxonomic key of the Aurantioideae subfamily using chloroplastic SNP diagnostic markers of the main clades

genotyped by competitive allele-specific PCR. BMC Genet 17(1):118. https://doi.org/10.1186/s12863-016-0426-x

39. Wutke S, Ludwig A (2019) Targeted PCR amplification and multiplex sequencing of ancient DNA for SNP analysis. In: Ancient DNA: methods and protocols. Springer New York, New York, NY, pp 141–147. https://doi.org/10.1007/978-1-4939-9176-1_15
40. Vallebueno-Estrada M, Rodríguez-Arévalo I, Rougon-Cardoso A, Martínez González J, García Cook A, Montiel R, Vielle-Calzada J-P (2016) The earliest maize from San Marcos Tehuacán is a partial domesticate with genomic evidence of inbreeding. Proc Natl Acad Sci 113(49):14151. https://doi.org/10.1073/pnas.1609701113
41. Scheben A, Yuan Y, Edwards D (2016) Advances in genomics for adapting crops to climate change. Current Plant Biology 6:2–10
42. Singh BD, Singh AK (2015) Marker-assisted plant breeding: principles and practices. Springer, New Delhi
43. Hurgobin B, Edwards D (2017) SNP discovery using a Pangenome: has the single reference approach become obsolete? Biology 6(1):21. https://doi.org/10.3390/biology6010021
44. Veeckman E, Ruttink T, Vandepoele K (2016) Are we there yet? Reliably estimating the completeness of plant genome sequences. Plant Cell 28(8):1759–1768
45. Xiao J, Zhang Z, Wu J, Yu J (2015) A brief review of software tools for pangenomics. Genomics Proteomics Bioinformatics 13(1):73–76
46. Marroni F, Pinosio S, Morgante M (2014) Structural variation and genome complexity: is dispensable really dispensable? Curr Opin Plant Biol 18:31–36
47. Yao W, Li G, Zhao H, Wang G, Lian X, Xie W (2015) Exploring the rice dispensable genome using a metagenome-like assembly strategy. Genome Biol 16(1):187. https://doi.org/10.1186/s13059-015-0757-3
48. Vernikos G, Medini D, Riley DR, Tettelin H (2015) Ten years of pan-genome analyses. Curr Opin Microbiol 23:148–154
49. Lu F, Romay MC, Glaubitz JC, Bradbury PJ, Elshire RJ, Wang T, Li Y, Li Y, Semagn K, Zhang X, Hernandez AG, Mikel MA, Soifer I, Barad O, Buckler ES (2015) High-resolution genetic mapping of maize pan-genome sequence anchors. Nat Commun 6:6914
50. Bosi E, Fondi M, Orlandini V, Perrin E, Maida I, de Pascale D, Tutino ML, Parrilli E, Giudice AL, Filloux A (2017) The pangenome of (Antarctic) Pseudoalteromonas bacteria: evolutionary and functional insights. BMC Genomics 18(1):93
51. Fox EJ, Reid-Bayliss KS, Emond MJ, Loeb LA (2014) Accuracy of next generation sequencing platforms. Next Gener Seq Appl 1:1000106
52. Pfeiffer F, Gröber C, Blank M, Händler K, Beyer M, Schultze JL, Mayer G (2018) Systematic evaluation of error rates and causes in short samples in next-generation sequencing. Sci Rep 8(1):10950. https://doi.org/10.1038/s41598-018-29325-6
53. Nielsen R, Paul JS, Albrechtsen A, Song YS (2011) Genotype and SNP calling from next-generation sequencing data. Nat Rev Genet 12(6):443
54. Yu X, Sun S (2013) Comparing a few SNP calling algorithms using low-coverage sequencing data. BMC Bioinformatics 14(1):274. https://doi.org/10.1186/1471-2105-14-274
55. Li H (2011) A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. Bioinformatics 27(21):2987–2993
56. O'Rawe JA, Ferson S, Lyon GJ (2015) Accounting for uncertainty in DNA sequencing data. Trends Genet 31(2):61–66
57. Mielczarek M, Szyda J (2016) Review of alignment and SNP calling algorithms for next-generation sequencing data. J Appl Genet 57(1):71–79. https://doi.org/10.1007/s13353-015-0292-7
58. Lee T-H, Guo H, Wang X, Kim C, Paterson AH (2014) SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. BMC Genomics 15(1):162
59. Zielezinski A, Vinga S, Almeida J, Karlowski WM (2017) Alignment-free sequence comparison: benefits, applications, and tools. Genome Biol 18(1):186. https://doi.org/10.1186/s13059-017-1319-7
60. Chattopadhyay AK, Nasiev D, Flower DR (2015) A statistical physics perspective on alignment-independent protein sequence comparison. Bioinformatics 31(15):2469–2474
61. Gardner PP, Wilm A, Washietl S (2005) A benchmark of multiple sequence alignment programs upon structural RNAs. Nucleic Acids Res 33(8):2433–2439. https://doi.org/10.1093/nar/gki541
62. Bonham-Carter O, Steele J, Bastola D (2013) Alignment-free genetic sequence comparisons: a review of recent approaches by word analysis. Brief Bioinform 15(6):890–905
63. Bromberg R, Grishin NV, Otwinowski Z (2016) Phylogeny reconstruction with

alignment-free method that corrects for horizontal gene transfer. PLoS Comput Biol 12(6): e1004985

64. Didier G, Debomy L, Pupin M, Zhang M, Grossmann A, Devauchelle C, Laprevotte I (2007) Comparing sequences without using alignments: application to HIV/SIV subtyping. BMC Bioinformatics 8(1):1
65. Ondov BD, Starrett GJ, Sappington A, Kostic A, Koren S, Buck CB, Phillippy AM (2019) Mash screen: high-throughput sequence containment estimation for genome discovery. bioRxiv:557314. https://doi.org/10.1101/557314
66. Ondov BD, Treangen TJ, Melsted P, Mallonee AB, Bergman NH, Koren S, Phillippy AM (2016) Mash: fast genome and metagenome distance estimation using MinHash. Genome Biol 17(1):132. https://doi.org/10.1186/s13059-016-0997-x
67. Saw AK, Raj G, Das M, Talukdar NC, Tripathy BC, Nandi S (2019) Alignment-free method for DNA sequence clustering using fuzzy integral similarity. Sci Rep 9(1):3753. https://doi.org/10.1038/s41598-019-40452-6
68. Li Y, He L, Lucy He R, Yau SST (2017) A novel fast vector method for genetic sequence comparison. Sci Rep 7(1):12226. https://doi.org/10.1038/s41598-017-12493-2
69. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17(1):10–12
70. Andrews S (2010) FastQC: a quality control tool for high throughput sequence data
71. Krueger F (2015) Trim galore. A wrapper tool around Cutadapt and FastQC to consistently apply quality and adapter trimming to FastQ files 516:517
72. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12(4):357
73. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, Subgroup GPDP (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079. https://doi.org/10.1093/bioinformatics/btp352
74. Broder AZ (1997) On the resemblance and containment of documents. In: Proceedings. compression and complexity of SEQUENCES 1997 (Cat. No. 97TB100171). IEEE, pp 21–29
75. Fan H, Ives AR, Surget-Groba Y, Cannon CH (2015) An assembly and alignment-free method of phylogeny reconstruction from next-generation sequencing data. BMC Genomics 16(1):522
76. Brockman W, Alvarez P, Young S, Garber M, Giannoukos G, Lee WL, Russ C, Lander ES, Nusbaum C, Jaffe DB (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. Genome Res 18(5):763–770
77. Zerbino DR, Birney E (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res 18(5):821–829
78. Luo R, Liu B, Xie Y, Li Z, Huang W, Yuan J, He G, Chen Y, Pan Q, Liu Y (2012) SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. Gigascience 1(1):18
79. Zimin AV, Marçais G, Puiu D, Roberts M, Salzberg SL, Yorke JA (2013) The MaSuRCA genome assembler. Bioinformatics 29(21): 2669–2677

Chapter 10

Genotyping-by-Sequencing Technology in Plant Taxonomy and Phylogeny

Félicien Favre, Cyril Jourda, Pascale Besse, and Carine Charron

Abstract

Genotyping-by-sequencing (GBS) is a method to discover and genotype simultaneous genome-wide high-throughput single nucleotide polymorphisms (SNPs). GBS is based on reducing genome complexity with restriction enzymes. Here we describe a method developed by Elshire et al. for constructing simplified GBS libraries and recent bioinformatic approaches developed to analyze the large volume of polymorphism data generated by this method. GBS approach is suitable for population studies, taxonomic and phylogenic studies, germplasm characterization, and breeding and trait mapping for a wide range of organisms, including plants with complex genomes.

Key words Genotyping-by-sequencing (GBS), Next-generation sequencing (NGS), High-throughput single nucleotide polymorphism (SNP), Plant diversity markers

1 Introduction

1.1 Genotyping-by-Sequencing

Nowadays, the exploration of genetic diversity of plant is enhanced by advanced high-throughput sequencing (HTS) technologies, which provide the opportunity to simultaneously discover a high number of molecular markers at relatively low cost. In particular, genotyping-by-sequencing (GBS) is a method to discover and genotype genome-wide high-throughput single nucleotide polymorphisms (SNPs) in a large number of individuals at the same time. Plant genomes are complex, and sequencing more than one entire genome would be expensive and time-consuming. GBS in any large genome species requires reduction of genome complexity, which can be done by different approaches. The target enrichment approaches can use long-range PCR of specific genes or genomic subsets, molecular inversion probes, or hybridization-based sequence capture methods such as microarrays [1]. However, these methods require invariant primer binding site and remain technically difficult and expensive for assaying many samples at the same time. GBS is associated with restriction-site-reduced

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_10,

complexity approach (RAD) [2]. The concept is based on acquiring the sequence adjacent to a set of particular restriction enzyme (RE) recognition sites. Large volumes of polymorphism data can be generated by applying massively parallel sequencing and multiplexing RAD with RAD tag libraries. Herein we report a method for constructing GBS libraries based on reducing genome complexity with REs [3]. This approach is simple, quick, extremely specific, and highly reproducible and may reach important regions of the genome that are inaccessible to sequence capture approaches. A large number of methylation-sensitive REs with different size recognition sites can be chosen. Methylation-sensitive REs are not able to cleave methylated cytosine residues; thus they target gene regions and filter out repetitive genomic regions. Thousands of genome-wide markers can be identified with better chance to get something linked to the cause of the polymorphism. GBS focuses on next-generation sequencing (NGS) power to sequence the end of restriction fragments. Advances in NGS throughout the last decade have enabled GBS to be used for high diversity and large genome species. The method is based on a multiplex sequencing strategy that uses an inexpensive barcoding system. Barcodes are included in one of the adapter sequence and located just upstream of the RE cut site in genomic DNA. This procedure generates restriction fragments with appropriate adapter, limits the sample handling, and facilitates the association of fragments to the sample. GBS was applied initially to maize and barley mapping populations but provides results independently of the target species or population and does not require having previous available genomic information. Recent advances in bioinformatics and development of new software programs such as STACKS [4] are able to overcome the lack of reference genome, by using de novo assembly of short sequenced reads. GBS was already used in a large amount of studies in recent years. For instance, an analysis of genetic diversity of European blueberry cultivars by GBS has allowed to better define phylogeny and adaptation of plants to their environment in terms of flowering and fruit ripening [5]. These results should help the preservation of genetic resources and contribute to further breeding programs. GBS is also useful to explore the genetic structure of populations, such as in *Cynara cardunculus*, showing subpopulations within artichokes and cultivated cardoon [6]. Molecular markers identified by GBS are particularly useful for marker-assisted selection (MAS) to enhance genomic selection in plant breeding programs in wheat [7]. GBS has been successfully used in pepper with a wide range of applications. An important amount of informative genome-wide SNPs were identified and enabled to analyze germplasm diversity and population structure as a result of domestication or local adaptation [8, 9]. GBS-generated SNP markers have been also useful in the detection of trait-associated quantitative trait loci (QTLs) for both *Capsicum annuum and Capsicum*

baccatum and will support genome-wide association mapping studies and marker-assisted selection programs [10, 11]. This approach is also particularly efficient to identify QTLs or genes of interest involved in resistance to plant disease [12], in plant architecture [13], and in plant metabolite content [14]. Here we describe the highly multiplexed system developed by Elshire et al. for constructing GBS libraries for Illumina sequencing. Then we describe the de novo assembly using STACKS and a bioinformatic way to identify SNPs.

1.2 Some Limitations to the GBS Method

1. GBS sequencing produces a lot of missing data [15]. This may be partly explained by three main reasons: (a) the lack of the restriction site in particular samples, (b) polymorphism in restriction site, and (c) a low sequence coverage rate. Simulations showed that locus identification was highly reproducible with a sequence coverage somewhere between 20 and 40X [4].
2. This approach gives a random access to genomic regions, because of structure variations and repeated sequences, which are different in each individual.
3. The larger the library is, the more missing date is generated (*see* **Note 1**).
4. The most important under the GBS approach is to obtain enough high-quality molecular markers to answer to our questions. The GBS protocol can be modified to be used with new species or different enzymes, mainly to obtain more markers or fewer markers but with a deeper sequence coverage per locus, to increase multiplexing, to avoid more repetitive DNA classes, or for novel applications.

2 Materials

2.1 DNA Extraction and Quantification

1. DNA spin columns-based commercial kit such as DNeasy Plant Mini Kit (Qiagen, Hilden, Germany).
2. Qubit 4 Fluorometer and Qubit assays for DNA quantification (Invitrogen, Carlsbad, CA, USA).
3. *Hin*dIII or *Eco*RI restriction endonuclease (New England Biolabs, Ipswich, Massachusetts, USA). The nucleic acid recognition sequences where the enzymes cut are, respectively, 5′-A/AGCTT-3′ and 5′-G/AATTC-3′.
4. TAE buffer (1X): 0.04 M Tris-acetate and 0.001 M EDTA pH 8.0.
5. Agarose gel 2%.

2.2 GBS Library Construction

1. Sequences of double-stranded barcode adapter:
   5′-ACACTCTTTCCCTACACGACGCTCTTCCGATCT xxxx-3′ and

   5′-CWGyyyyAGATCGGAAGAGCGTCGTGTAGGGAAAG AGTGT-3′,

   where "xxxx" and "yyyy" indicate, respectively, the barcode and barcode complement end sequences.
2. Sequence of double-stranded common adapter:
   5′-CWGAGATCGGAAGAGCGGTTCAGCAGGAATGCCG AG-3′ and

   5′-CTCGGCATTCCTGCTGAACCGCTCTTCCGATCT-3′.
3. TE buffer (1X): 10 mM Tris–HCl and 1 mM EDTA-NaOH, pH 8.0.
4. Thermocycler.
5. PicoGreen (Invitrogen, Carlsbad, CA, USA) or similar instrument for quantification of the adapter.
6. PCR 96-well plate.
7. *Ape*KI that recognizes the sequence 5′-G/CWGC-3′ (New England Biolabs, Ipswich, Massachusetts, USA) or appropriate RE.
8. NEB Buffer 3 (1X): 50 mM Tris–HCl, 10 mM $MgCl_2$, 100 mM NaCl, 1 mM DTT, pH 7.9 at 25 °C (New England Biolabs, Ipswich, Massachusetts).
9. Ligase buffer with ATP and T4 ligase (New England Biolabs, Ipswich, Massachusetts, USA).
10. QIAquick PCR Purification Kit (Qiagen, Hilden, Germany).
11. PCR primer 1: 5′-AATGATACGGCGACCACCGAGATCTA CACTCTTTCCCTACACGACGCTCTTCCGATCT-3′.
12. PCR primer 2: 5′-CAAGCAGAAGACGGCATACGAG ATCGGTCTCGGCATTCCTGCTGAACCGCTCTTCCG ATCT-3′.
13. *Taq* Master Mix (1X) (New England Biolabs, Ipswich, Massachusetts, USA).
14. Bio-Rad Experion (Bio-Rad, Hercules, California, USA) or similar instrument.

2.3 Illumina Workflow

The protocol should be optimized depending on the sequencer used. Here, we describe the main steps of Illumina sequencing (Illumina Inc., San Diego, California, USA).

2.4 Data Analysis Equipment and Softwares

Data must be processed within a high-performance computing cluster. The following softwares and tools are used:

1. FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/).
2. Cutadapt [16].
3. Trimmomatic [17], the FASTX-Toolkit (http://hannonlab.cshl.edu/fastx_toolkit/), or similar tool.
4. GBS barcode splitter (https://sourceforge.net/projects/gbsbarcode/) and FASTQ/A Barcode splitter from the FASTX-Toolkit.
5. STACKS [4].
6. South Green bioinformatics platform (https://www.southgreen.fr).
7. R package "pegas" [18].
8. STRUCTURE [19].

3 Methods

3.1 DNA Extraction and Quantification

1. Extract high molecular weight DNAs from young leaves using a standard CTAB protocol or using DNA spin column-based commercial kits according the manufacturer's instructions (*see* Chapter 3).
2. Quantify genomic DNA by fluorimetric assays, and normalize gDNA concentrations at 50 ng/μL (*see* **Note 2**).
3. To test DNA homogeneity, mix 1 μL of uncut sample DNA with 4 μL of loading dye, load in a 2% agarose gel, and run at 110 V for 2 h. The gel must reveal one clear band for each sample.
4. Test whether DNA extractions are of sufficient quality by enzymatic digestion. The digestion test doesn't need to be done with methylation-sensitive enzymes that cleave only at unmethylated recognition sites. Cheaper RE that is not methylation-sensitive such as *Hin*dIII or *Eco*RI should be chosen. Pool some DNAs from the same extraction series to have 500 ng of DNA. Mix 10 μL of DNA (50 ng/μL) with 7.3 μL of ultrapure water, 2 μL of RE 10X, 0.2 μL of BSA, and 0.5 μL of enzyme. The mix is incubated first at 37 °C for 4 h and at 80 °C for 20 min, loaded in a 2% agarose gel, and run at 110 V for 2 h. The gel must reveal a regular smear without band.

3.2 GBS Library Construction

1. Two kinds of adapters are used for constructing GBS libraries, a barcode adapter and a common adapter. Adapters are designed to fit with Illumina sequencing (*see* **Note 3**). Dilute

oligonucleotides of each barcode and common adapters separately in TE buffer (50 μM each), and anneal them in a thermocycler at 95 °C during 2 min. Then decrease the temperature to 25 °C by 0.1 °C/s and wait 30 min at 25 °C. Hold the adapters at 4 °C.

2. Quantify the adapters with an intercalating dye (PicoGreen (Invitrogen, Carlsbad, CA, USA) or similar instrument) diluted in water to 0.6 ng/μL. Mix then barcode and common adapters together in a 1:1 ratio; plate the mix into a 96-well plate, each well with a different barcode adapter; and dry the plate (Fig. 1, **step 1**, *see* **Note 4**).
3. Add 100 ng of DNA in a volume of 10 μL into each well (96 different DNA samples can be studied on one plate). Dry the plate again (Fig. 1, **step 1**).
4. In each well, digest DNA for 2 h at 75 °C with *Ape*KI, in 20 μL volumes containing 1X NEB Buffer 3 and 3.6 U *Ape*KI (Fig. 1, **step 2**). This step should be optimized depending on the RE that is used (*see* **Note 5**).
5. Ligate adapters to the ends of the genomic DNA inserts: add 30 μL of a solution containing 1.66X ligase buffer with ATP and T4 ligase into each well. (Fig. 1, **step 3**).
6. T4 ligase inactivation: incubate samples at 22 °C for 1 h and then heat them to 65° for 30 min.
7. Pool an aliquot of each sample (5 μL) into an Eppendorf tube, and apply it to a size exclusion column to remove unreacted adapters. Purify samples using a commercial kit (QIAquick PCR Purification Kit, Qiagen, Hilden, Germany). DNA samples are then eluted in a final volume of 50 μL (Fig. 1, **step 4**).
8. Perform a PCR to amplify the fragment pool in 50 μL volumes containing 2 μL pooled DNA fragments from **step 7**, 1X *Taq* Master Mix, and 25 pmol of each PCR primers 1 and 2 (*see* **Note 6**). Use the following PCR temperature cycling: 72 °C for 5 min, 98 °C for 30 s; 18 cycles of 98 °C for 30 s, 65 °C for 30 s, and 72 °C for 30 s with a final *Taq* extension step at 72 °C for 5 min (Fig. 1, **step 5**).
9. Clean up PCR products, and evaluate fragment sizes of the resulting library on a DNA analyzer (Bio-Rad Experion or similar instrument) (Fig. 1, **step 6**). Libraries without adapter dimer are retained for DNA sequencing (*see* **Note 7**).

3.3 Illumina Sequencing Workflow

Illumina sequencing being most often outsourced to private companies, we simply propose here a step-by-step workflow rather than a classical wetlab protocol.

1. Perform single-end sequencing or paired-end sequencing of the library in a flow cell channel using the HiSeq 3000/

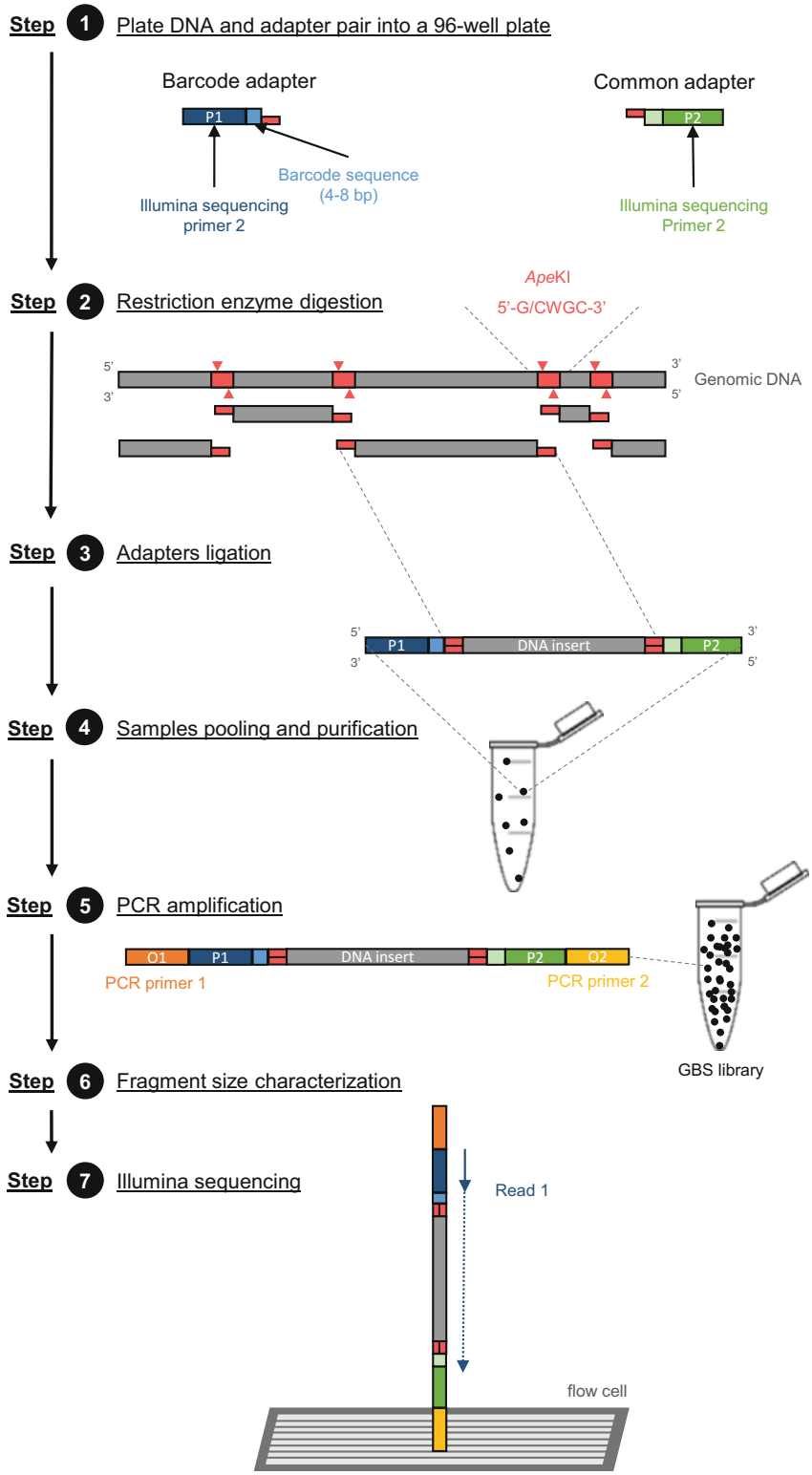


Fig. 1 Diagram of the genomic library construction method based on reducing genome complexity with restriction enzymes (REs) for Illumina sequencing

HiSeq 4000 Systems (Illumina Inc., San Diego, California, USA) (*see* **Note 8**).

2. Drop PCR products off at an Illumina flow cell (Fig. 1, **step 7**).
3. Bridge amplification of DNA fragments in cluster. The amplification is based on solid-phase PCR. Cluster formation amplifies sequencing signal.
   (a) Bind single-stranded fragments randomly to the inside surface of the flow cell channels.
   (b) Add unlabeled nucleotides and enzyme to initiate solid-phase bridge amplification.
   (c) First amplification: the enzyme incorporates nucleotides to build double-stranded bridges on the solid-phase substrate.
   (d) Linearization: denature the double-stranded DNA to leave single-stranded DNA anchored to the substrate.
   (e) Complete amplification: several million dense clusters of same single-stranded DNA are generated in each channel of the flow cell.
4. Sequencing by synthesis:
   (a) To initiate the first sequencing cycle, add Illumina sequencing primer P1, DNA polymerase enzyme, and all four nucleotides, each labeled with a different dye, to the flow cell.
   (b) After laser excitation, capture the image of emitted fluorescence from each cluster on the flow cell. Record the identity of the first base for each cluster. Cleave dye and terminating groups and wash.
   (c) To initiate the next sequencing cycle, add all four labeled nucleotides and enzyme to the flow cell. After laser excitation, collect the image data as previously described. Record the identity of the second base for each cluster.
   (d) Repeat cycles of sequencing to determine the sequence of bases. This entire process generates millions of reads, representing all the genomic fragments (*see* **Note 9**).

3.4 GBS Data Processing

3.4.1 Filtering Raw Sequence Data

1. Illumina sequencing produces a large amount of data. Outputs are "fastq" files with four lines per sequence with (a) sequence name; (b) DNA sequence; (c) metadata with sample information such as plant line, location, and year; and (d) quality score of each base of the sequence.
2. View sequences with the quality control tool for high-throughput sequence fastQC.

3. Remove adapter sequences at the end of reads and low-quality extremities. Some tools can be used such as Cutadapt [16] (*see* **Note 10**).
4. Remove low-quality reads based both of quality score, read length, low complexity, and N (unsequenced) bases using Trimmomatic [17], the FASTX-Toolkit, or similar tool.
5. Reads are then demultiplexed and assigned to each sample (*see* **Note 11**). GBS barcode splitter and FASTQ/A Barcode splitter from the FASTX-Toolkit are tools to split GBS reads by barcode.

3.4.2 Mapping

1. Analyze assigned reads with STACKS to identify all the SNPs [4].
2. Identified SNPs are conserved in a Variant Call Format (VCF) file for analysis [20].
3. Screen SNPs according to several criteria: sequencing depth, missing data, and gene frequency (*see* **Note 12**). VCF files derived from SNP calling can be filtered directly by some tools such as SNiPlay which is part of the South Green bioinformatics platform.

3.4.3 Analysis

The identified SNPs can be used in phylogenetic studies. SNiPlay computes on the web series of tools for analyses at a whole-genome scale (general statistics, polyploid analysis, chromosome viewer, SNP density, diversity analysis, association studies, etc.). VCF files can also be analyzed using the R package "pegas" to calculate similarity and construct phylogenetic trees [18]. To analyze the genetic structure of a population and identify groups that are genetically linked, the Bayesian method of the software STRUCTURE can be used [19]. Many other tools are efficient to analyze a VCF files.

4 Notes

1. The technical option to limit the missing data is to reduce the multiplexing level or sequence the same library several times, and the molecular option is to choose less frequently cutting enzymes.
2. Leaves can be first lyophilized. A quantity of 20–30 mg of lyophilized leaves should be enough for DNA extraction. A great amount of DNA is not needed (50–100 ng/sample). However, quality and quantity of the DNA must be homogeneous. Use preferably fluorimetric method for quantification than spectrophotometric method which could overestimate DNA concentrations. DNA purity is crucial for complete enzyme digestion.

3. The DNA fragment is ligated to a barcode adapter and a common adapter. The barcode adapter contains an Illumina sequencing primer 1 and a barcode which is a 4–8 bp sequence used to identify a sample. The barcode is on the 3′ end of the top strand of the adapter. The 5′ end of its bottom strand is terminated by a 3 bp overhang which is complementary to the end genomic DNA fragments generated by the RE. The common adapter is complementary to the other end and is only containing an Illumina sequencing primer 2 end. The adapters are designed for either single-end or paired-end sequencing on the Illumina (Fig. 1, **step 1**). Barcodes are enzyme specific: they must not recreate the enzyme recognition site to avoid being cut and must have complementary overhangs. Barcodes must be of variable length and different enough from each other to avoid confusion if there is a sequencing error (at least 3 bp differences among barcodes).
4. Up to 96 DNA samples can be processed simultaneously (48/96/384-well plate).
5. Choose methylation-sensitive REs to avoid repetitive regions of genomes and target lower copy regions. Select REs that leave 2 to 3 bp overhangs for efficient adapter ligation to fragments of DNA. *ApeKI* (New England Biolabs) is often used and suitable for maize because it is known to have low recognition sites in maize retrotransposons [3]. *ApeKI* recognizes the sequence 5′-GCWGC-3′ (with W is A or T) and leave a 5′ overhang with 3 bp. *PstI* (New England Biolabs) and *EcoT22I* can also be used and recognize, respectively, the sequences 5′-CTGCA/G-3′ and 5′-ATGCA/T-3′. *PstI* was used in artichoke GBS libraries [6].
6. PCR primer 1 is designed to bind, on one hand to 3′ strand of barcode adapter and on the other hand to flow cell oligonucleotide 1 for Illumina sequencing. PCR primer 2 binds to 3′ strand of common adapter and to flow cell oligonucleotide 2. The PCR primers 1 and 2 contain sequences for amplifying restriction fragments with ligated adapters and binding PCR products to oligonucleotides contained in Illumina flow cell (Fig. 1, **step 5**).
7. Libraries were considered suitable for sequencing if adapter dimers (around 128 bp) are minimal or absent and the majority of other DNA fragments are between 170 and 350 bp. Do the protocol again and adapt and decrease adapter amounts if adapter dimers are present in excess of 0.5%.
8. Two kinds of sequencing can be used for GBS. The single-end sequencing produces reads up to 300 bp and is better for species without reference genome. The paired-end sequencing from both fragment ends generates longer reads from 300 to

500 bp and should be used preferably in species with high-quality reference genome.

9. GBS captures barcode and insert DNA sequence in single read. It ensures that the barcode fits well with its sample because they are physically attached.
10. Different parameters and tools should be tested to removes adapters and low-quality extremities. Reads must be verified each time with fastQC.
11. A table of correspondence between barcode sequences and samples would be very useful for demultiplexing.
12. At least five reads at each locus for each sample are recommended. The minimal gene frequency recommended is 30% to avoid keeping sequencing error. However, gene frequency can be adjusted and reduced depending on the sampling. If the sampling is unbalanced, a low frequency provides access to alleles that are associated with underrepresented individuals in the dataset.

References

1. Mamanova L, Coffey AJ, Scott CE et al (2010) Target-enrichment strategies for next-generation sequencing. Nat Methods 7:111–118
2. Miller MR, Dunham JP, Amores A et al (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res 17:240–248
3. Elshire RJ, Glaubitz JC, Sun Q et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379
4. Catchen J, Hohenlohe PA, Bassham S et al (2013) Stacks: an analysis tool set for population genomics. Mol Ecol 22:3124–3140
5. Campa A, Ferreira JJ (2018) Genetic diversity assessed by genotyping by sequencing (GBS) and for phenological traits in blueberry cultivars. PLoS One 13:e0206361
6. Pavan S, Curci PL, Zuluaga DL et al (2018) Genotyping-by-sequencing highlights patterns of genetic structure and domestication in artichoke and cardoon. PLoS One 13:e0205988
7. Poland J, Endelman J, Dawson J et al (2012) Genomic selection in wheat breeding using genotyping-by-sequencing. Plant Genome 5:103–113
8. Taranto F, D'Agostino N, Greco B et al (2016) Genome-wide SNP discovery and population structure analysis in pepper (*Capsicum annuum*) using genotyping by sequencing. BMC Genomics 17:943
9. Taitano N, Bernau V, Jardón-Barbolla L et al (2019) Genome-wide genotyping of a novel Mexican Chile pepper collection illuminates the history of landrace differentiation after *Capsicum annuum* L. domestication. Evol Appl 12:78–92
10. Nimmakayala P, Abburi VL, Saminathan T et al (2016) Genome-wide divergence and linkage disequilibrium analyses for *Capsicum baccatum* revealed by genome-anchored single nucleotide polymorphisms. Front Plant Sci 7:1646
11. Nimmakayala P, Abburi VL, Saminathan T et al (2016) Genome-wide diversity and association mapping for Capsaicinoids and fruit weight in *Capsicum annuum* L. Sci Rep 6:38081
12. Salgon S, Raynal M, Lebon S et al (2018) Genotyping by sequencing highlights a polygenic resistance to *Ralstonia pseudosolanacearum* in eggplant (*Solanum melongena* L.). Int J Mol Sci 19:357
13. Gabay G, Dahan Y, Izhaki Y et al (2018) High-resolution genetic linkage map of European pear (*Pyrus communis*) and QTL fine-mapping of vegetative budbreak time. BMC Plant Biol 18:175
14. Gonda I, Ashrafi H, Lyon DA et al (2019) Sequencing-based bin map construction of a tomato mapping population, facilitating high-resolution quantitative trait loci detection. Plant Genome 12:1–14
15. Scheben A, Batley J, Edwards D (2017) Genotyping-by-sequencing approaches to

characterize crop genomes: choosing the right tool for the right application. Plant Biotechnol J 15:149–161

16. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10
17. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120
18. Paradis E (2010) Pegas: an R package for population genetics with an integrated-modular approach. Bioinformatics 26:419–420
19. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
20. Danecek P, Auton A, Abecasis G et al (2011) The variant call format and VCFtools. Bioinformatics 27:2156–2158

Chapter 11

Development of Microsatellite Markers Using Next-Generation Sequencing

Hélène Vignes and Ronan Rivallan

Abstract

Among the molecular markers used for plant genetic studies, microsatellite markers are easy to implement and can provide suitable codominant markers for molecular taxonomy.

Here we describe a method to obtain microsatellite primers from genomic DNA using a next-generation sequencer.

Key words Microsatellite, Next-generation sequencing, Bioinformatic analysis

1 Introduction

Microsatellites, also called simple sequence repeats (SSRs) [1], are small repeats of one, two, three, or four tandemly arranged nucleotides that are ubiquitous components of eukaryotic genomes. They have a high level of polymorphism due to mutation affecting the number of repeat units. Their variable length polymorphism can be revealed by polymerase chain reaction (PCR) [2] with unique flanking primers [3] that generate codominant markers. Microsatellites have a Mendelian heritability [4] and have a potential advantage of reliability, reproducibility, discrimination, standardization, and cost-effectiveness [5]. All these characteristics make them a suitable tool for genetic analysis, diversity analysis, population structure studies, genetic mapping, and quantitative trait analysis.

The precedent method for microsatellite discovery was an expensive and time-consuming task, through the construction of microsatellite-enriched genome libraries, cloning, and sequencing by Sanger method.

Today, the method using NGS [6] and bioinformatics makes it possible to obtain a very large number of microsatellite markers quickly and easily. Microsatellite markers discovered represent a

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_11,

non-exhaustive part depending on the sequencing step. The search for microsatellite motifs is done after high-throughput sequencing using a specific pipeline or Galaxy.

2 Materials

2.1 Library DNA Construction

Use the Nextera kit (FC-121-1030), Illumina. This kit allows the construction of the DNA bank, from fragmentation to amplification.

2.1.1 DNA Fragmentation

1. Total genomic DNA (2.5 ng/μL).
2. TD Tagment DNA Buffer (Nextera kit).
3. TDE1 Tagment DNA Enzyme (Nextera kit).
4. Mastercycler® nexus PCR thermal cycler, Eppendorf AG, Germany.

2.1.2 Fragmented DNA Purification

1. DNA Clean & Concentrator (ZD4013) kit, Zymo (as recommended for the Nextera kit).
2. RSB Resuspension Buffer (Nextera kit).

2.1.3 DNA Amplification

1. NPM Nextera PCR Master Mix (Nextera kit).
2. Nextera Index Kit (N501, N502, N503, N504, N701, N702, N703, N704, N705, N706).
3. PPC PCR Primer Cocktail (Nextera kit).
4. Mastercycler® nexus PCR thermal cycler, Eppendorf AG, Germany.

2.1.4 DNA Purification

1. Agencourt AMPure XP beads (A63881), Beckman Coulter.
2. Agencourt SPRIPlate 96 Ring Super Magnet Plate (A32782) Beckman Coulter.
3. 80% ethanol, fresh.

2.2 Library Verification

1. Agilent 4200 TapeStation.
2. ScreenTape D5000, Agilent (5067-5588) (*see* **Note 1**).
3. LightCycler® 480 Real-Time PCR System, Roche Life Science.
4. Takara Library Quantification Kit (638324) (*see* **Note 2**).

2.3 Illumina Sequencing

1. MiSeq System, Illumina.
2. 500 cycles V2 cartridge (MS-102-2003) or 600 cycles V3 cartridge (MS-102-3003), Illumina.

2.4 Selection and SSR Screening

1. Galaxy platform for bioinformatic analysis (*see* **Note 3**).
2. Mastercycler® nexus PCR thermal cycler, Eppendorf AG, Germany.
3. ABI 3500xl sequencer, Life Technologies, Carlsbad, California, USA.
4. GeneMapper® Software, Life Technologies, Carlsbad, California, USA.
5. Zymo DNA Binding buffer.

2.5 PCR Amplification

1. Buffer 10 × (100 mM Tris–HCl, pH 9.0; 100 mM KCl 80 mM (NH4)2SO4; X-100 1% Triton).
2. 2 mM dNTP.
3. 50 mM $MgCl_2$.
4. 5U/μL Taq DNA Polymerase.
5. Template DNA (1 ng/μL).
6. Water, Milli-Q.

2.6 Sequencer Revelation and Scoring Analysis

1. Hi-Di formamide.
2. 0.12 μL of GeneScan 600 LIZ size standard (Applied Biosystems).

3 Methods

3.1 DNA Fragmentation

1. Put 20 μL of 50 ng total genomic DNA (2.5 ng/μL) into a 0.5 mL Eppendorf microtube.
2. Mix with 25 μL of TD buffer and 5 μL of TDE1.
3. Incubate the tube in the Mastercycler® nexus PCR thermal cycler for 15 min at 55 °C.

3.2 DNA Purification with the Zymo Clean-Up Kit

All the centrifugation must be performed at 10,000 × *g*.

1. Add 180 μL of Zymo kit DNA binding buffer in a 1.5 mL Eppendorf tube.
2. Add 50 μL of DNA and mix by up and down.
3. Transfer the mix in a Zymo kit spin column and centrifuge for 30 s.
4. Add 200 μL of Zymo kit wash buffer and centrifuge for 30 s (repeat this step a second time).
5. Place the column on a new 1.5 mL Eppendorf tube, and add 25 μL of RSB buffer to the column.
6. Wait 2 min at room temperature and centrifuge for 1 min.
7. Recover the eluate and transfer it to a new 1.5 mL Eppendorf tube.

3.3 PCR Amplification with Adding Indexes

1. For one library, one index N7 and one index S5 must be added with purified DNA fragments. In a new 1.5 mL Eppendorf tube, add 15 µl of NPM, 5 µL of index 1 (e.g., N701), 5 µL of index 2 (e.g., S501), and 5 µL of PPC. The indexes are used to label the samples. It is therefore possible to multiplex several samples during sequencing. At the output of sequencing, there is a sequence file for each sample.
2. Transfer 20 µL of DNA. Mix well and do a pulse.
3. Place the tube in the Mastercycler® nexus PCR thermocycler with the following program: 72 °C 3 min, 95 °C 30 s, five cycles (95 °C 10 s, 63 °C 30 s, 72 °C 30 s), and hold at 10 °C.

3.4 Amplified DNA Purification

1. Add 30µL of AMPure XP beads in the tube containing 50 µL of amplification product. Mix gently and incubate for 5 min at room temperature.
2. Place the tube on SPRIPlate magnetic support for 5 min and then discard the supernatant.
3. Add 200 µL of 80 % ethanol, incubate for 30 s, and discard the supernatant. Repeat this step a second time.
4. Dry beads for 15 min on the SPRIPlate magnetic plate. Remove the tube and add 32.5 µL of RSB. Mix by up and down, and incubate for 2 min at room temperature.
5. Put the tube on SPRIPlate magnetic support for 5 min. Transfer 30 µL of supernatant in a new 1.5 mL Eppendorf tube.

   The library is ready.

3.5 Library Verification

1. The quality is checked using an Agilent 4200 TapeStation with a ScreenTape D5000. The size of the fragments must be between 100 and 600 pb (Fig. 1).
2. The library is quantified using the Takara kit on real-time PCR system.

3.6 Illumina Sequencing

The sequencing is performed on MiSeq system Illumina sequencer, using a 500-cycle V2 cartridge Illumina (2 × 250 pb) or a 600-cycle V3 cartridge Illumina (2 × 300 pb).

3.7 Bioinformatic Analysis and Primer Design

Among all the sequences obtained, it is now necessary to sort and recover the sequences containing microsatellite motifs via bioinformatic analysis. It can be performed directly on the command line. Uninitiated people can use the Galaxy platform [7]. Galaxy is a workflow manager, which permits to run several bioinformatic analysis using a simple web interface. The raw sequences (fast gz format) are processed in Galaxy using several tools to obtain the data matrix containing the microsatellite primers.

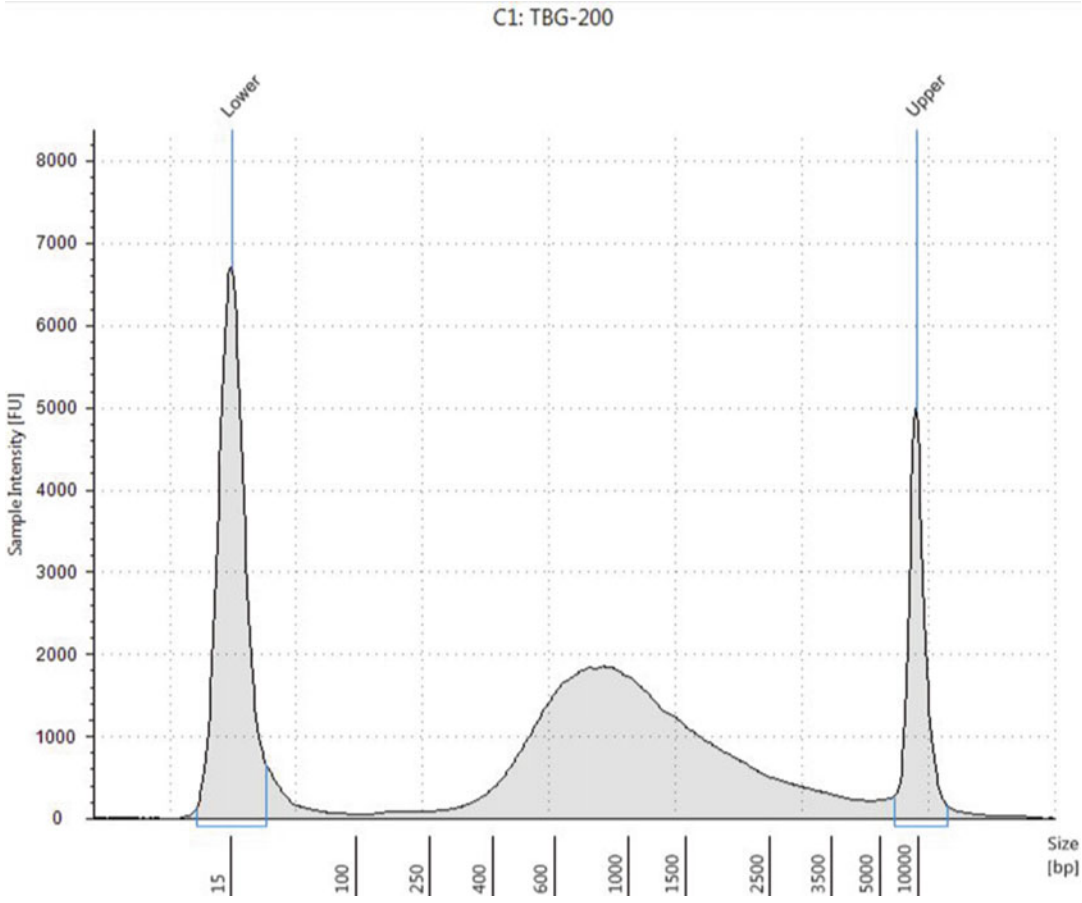


Fig. 1 Library profile on ScreenTape D5000

The different tools used on Galaxy pipeline are:

1. **FASTQ Groomer** [8]**:** offers several conversion options relating to the FASTQ format.
2. **Filter FASTQ:** reads by quality score and length tool allows filtering by minimum and maximum read lengths and quality score values.
3. **ABySS** [9]**:** a de novo sequence assembler intended for short paired-end reads and large genomes.
4. **MISA** [10] **+ Primer3:** search for microsatellites and design primers.

   This tool allows the identification and the localization of perfect microsatellites as well as compound microsatellites, which are interrupted by a certain number of bases. In order to design primers flanking the microsatellite loci, two perl scripts serve as interface modules for the program-to-program data interchange between MISA and the primer modelling software Primer3 (Whitehead Institute).

ID	SSR nr.	SSR type	SSR	size	start	end	FORWARD PRIMER1 (5'-3')	Tm(°C)	size	REVERSE PRIMER1 (5'-3')	Tm(°C)	size	PRODUCT1 size (bp)	start (bp)	end (bp)
753497_199_477	1	p2	(AG)11	22	152	173	AAAACAAAAGTGCAGAGG	51.160	18	ATAAACAAACACATGCTGAC	50.964	20	117	80	196
408902_221_533	1	p2	(TA)8	16	28	43	AAAACAAAGGAACGGTGAA	55.703	19	AAGAGTGAGGGGAGCAAA	55.787	18	213	6	218
666463_210_340	1	p2	(GA)7	14	87	100	AAAACAAGGAAGAGGGCA	55.781	18	CTCTCTATCGTCTCATGTCTCC	55.640	22	162	31	192
470025_172_396	1	p2	(AT)7	14	86	99	AAAACACAGGCATCACACA	55.862	19	CCCGATTCCTAAACTTCCT	55.811	19	146	20	165
187452_249_801	1	p2	(AT)5	10	44	53	AAAACAGGTTCTAGCAATTC	51.342	20	GCTTCTGGGGTTACTTG	51.211	17	162	21	182
255156_327_790	1	p2	(CT)9	18	156	173	AAAACCCGCTTATCTCTTT	53.335	19	AGGTGGATGATTGATTTCTT	53.559	20	244	50	293
199843_250_420	1	p2	(GA)14	28	68	95	AAAACCGATGGGGAGAG	55.830	17	GCCGCCTCCTGTTTC	55.914	15	191	34	224
168890_420_2205	1	p2	(AC)6	12	311	322	AAAACGCCCCTCTCAC	54.236	16	TGTTGGGGAGTGTATCAAG	54.271	19	142	271	412
493489_220_360	1	p2	(GA)5	10	61	70	AAAACGGAAACGGGG	55.066	15	TCTCATCCCTTCCTGTCC	55.813	18	102	41	142
711635_190_424	1	p2	(TA)5	10	92	101	AAAACTGTTCTGTGGTTGT	50.171	19	TTCTCAATCAGATAATGCAC	51.076	20	146	12	157
494465_249_580	1	p2	(GA)11	22	21	42	AAAAGAAAGCAACACTGC	51.326	18	AGTCACCTTCATCTCCCT	51.518	18	196	0	195
111087_142_306	1	p3	(GAA)6	18	26	43	AAAGAAGAGAGAGAAAGGAAC	51.326	21	GATTTGCACAACACAAATAC	51.442	20	126	0	125
897050_124_252	1	p3	(TTG)4	12	51	62	AAAGAAGAGCAGATTTGGATG	56.187	21	CTTCTCCCACCAATCCTTT	56.587	19	120	4	123
593228_251_608	1	p3	(AAT)6	18	137	154	AAAGAAGCCGCCAAC	52.687	15	ACGTGTATGGTTGATATGTG	51.521	20	139	53	191
790116_143_310	1	p3	(GGT)4	12	72	83	AAAGACCCGAAGATGAGG	55.075	18	TCCTCCCACTTCTCTCTAAAC	55.183	21	117	25	141
222270_308_943	1	p3	(GAG)4	12	165	176	AAAGACGACACAGGCGA	56.132	17	ACTTCCAATGCTTCTTCCTC	56.025	20	241	33	273
676_250_530	1	p3	(TAT)4	12	166	177	AAAGAGAAAGCACTGGAGGT	55.684	20	AAAACAGAGAAATGCCATCA	55.801	20	104	124	227
205901_159_423	1	p3	(ATA)4	12	99	110	AAAGAGGAGCAACTGTGAA	53.370	19	GATGCTACTTGAAATTGGG	53.177	19	107	46	152
389478_250_454	1	p3	(TGA)4	12	62	73	AAAGATGAAATGGTGTGGG	55.758	19	GACGCAGCAGAGGAAAA	55.882	17	195	37	231

Fig. 2 Data matrix (Galaxy)

In output, a data matrix containing all the microsatellite primers is obtained, according to the parameters indicated for the different tools used (Fig. 2).

3.8 Selection and SSR Screening

Validation and verification of microsatellite markers by PCR amplification.

3.8.1 Selection of Markers and Primers

1. Select only the primer pairs flanking for di- and trinucleotide SSR motifs. It is best to choose a minimum of eight repetitions for dinucleotides and six repetitions for trinucleotides. The amplification fragment size must be between 100 and 400 bp. You can select 40 SSR markers for a first serial of screening (20 markers for dinucleotide SSR motifs and 20 markers for trinucleotide SSR motifs).
2. The forward and reverse primers are synthesized. It is recommended to use a M13 tail (5′-CACGACGTTGTAAAACGAC-3') and adding it to the forward primer to lower the costs or use directly one-labelled primer on both.

3.8.2 PCR Amplification

1. PCR reactions are performed as simplex experiments in a volume of 10 μL containing 1 μL of reaction buffer 10×, 1 μL of 2 mM dNTP, 0.3 μL of 50 mM MgCl2, 0.08 μL of 10 μM forward primer with a M13 tail at the 5′-end, 0.1 μL of 10 μM reverse primer, 0.1 μL of fluorescently labelled M13-tail (6-FAM, NED, VIC, or PET from Applied Biosystems, Foster City, California, USA), 0.12 μL of 5U/μL Taq DNA Polymerase, 5 μL of template DNA (1 ng/μL), and 2.3 μL of water.
2. Use a touchdown cycling program with an initial denaturation at 94 °C for 5 min; followed by ten cycles at 94 °C for 30 s, 55 °C for 60 s (0.5 °C decrease at each cycle), and 72 °C for 1 min; followed by 25 cycles at 94°C for 45 s, 50 °C for 1 min, 72 °C for 1 min; and a final extension at 72 °C for 30 min.

3.8.3 Sequencer Revelation and Scoring Analysis

1. PCR products can be multiplexed according to dye and expected sizes (between 100 and 400 bp). Fluorescently labelled PCR products are organized in several SSR multiplexes for electrophoresis, using, respectively, 2 μL of products labelled with 6-FAM, 2 μL of those with VIC, 2.5 μL of those with NED, 3.5 μL of those with PET, and completed at 20 μl with high purity water.
2. Take 2 μL of PCR pool and add to 10 μL of Hi-Di formamide and 0.12 μL of GeneScan 600 LIZ size standard. The PCR products are revealed on ABI 3500xL Genetic Analyzer.
3. ABI electropherograms are analyzed with GeneMapper Software. Allele calling is obtained by checking for each data point in the amplification peaks. The markers selected must be polymorphic and easy to score.

4 Notes

1. We use the ScreenTape D5000 (5067-5588) Agilent, but we could also use a ScreenTape D1000 (5067-5582) or a ScreenTape genomic DNA (5067-5365). This step allows you to visualize the quality and size of the obtained bank.
2. During sequencing, the flowcell must be loaded correctly. The kit permits to quantify library with a method a highly sensitive by quantitative PCR.
3. We use the Galaxy platform and its various tools. However, if you have the skills in bioinformatics, then you can create your own analysis pipeline and work on the command line.

References

1. Tautz D, Rentz P (1984) Simple sequences are ubiquitous repetitive components of eukaryotic genome. Nucleic Acids Res 12:4127–4138
2. Mullis K, Faloona S, Scharf R, Saiki R, Horn G, Erlich H (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harbor Symp Quant Biol 51:263–273
3. Beckmann JS, Soller M (1990) Toward a unified approach to genetic mapping of eukaryotes based on sequence tagged microsatellite sites. Biotechnology 8:930–932
4. Weissenbach J, Gyapay G, Dib C, Vignal A, Morissette J, Millaseau P, Vaysseix G, Lathrop M (1992) A second generation map of the human genome. Nature 359:794–801
5. Smith JSC, Chin ECL, Shu H, Smith OS, Wall SJ, Senior ML, Mitchell SE, Kresovich S, Ziegle J (1997) An evaluation of the utility of SSR loci as molecular markers in maize (Zea mays L.), comparisons with data from RFLPs and pedigree. Theor Appl Genet 95:163–173
6. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nature 12(7):499–510
7. Afgan E, Baker D, Batut B, van den Beek M, Bouvier D, Čech M, Chilton J, Clements D, Coraor N, Grüning B, Guerler A, Hillman-Jackson J, Jalili V, Rasche H, Soranzo N, Goecks J, Taylor J, Nekrutenko A, Blankenberg D (2018) The Galaxy platform for accessible, reproducible and collaborative biomedical analyses: 2018 update. Nucleic

Acids Res 46(W1):W537–W544. https://doi.org/10.1093/nar/gky379

8. Blankenberg D, Gordon A, Von Kuster G, Coraor N, Taylor J, Nekrutenko A, Galaxy Team (2010) Manipulation of FASTQ data with Galaxy. Bioinformatics 26(14):1783–1785
9. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol I (2009) ABySS: A parallel assembler for short read sequence data. Genome Res 19(6):1117–1123. https://doi.org/10.1101/gr.089532.108
10. Beier S, Thiel T, Münch T, Scholz U, Mascher M (2017) MISA-web: a web server for microsatellite prediction. Bioinformatics 33:2583–2585. https://doi.org/10.1093/bioinformatics/btx198

Chapter 12

Amplified Fragment Length Polymorphism: Applications and Recent Developments

Thotten Elampilay Sheeja, Illathidath Payatatti Vijesh Kumar, Ananduchandra Giridhari, Divakaran Minoo, Muliyar Krishna Rajesh, and Kantipudi Nirmal Babu

Abstract

AFLP or amplified fragment length polymorphism is a PCR-based molecular technique that uses selective amplification of a subset of digested DNA fragments from any source to generate and compare unique fingerprints of genomes. It is more efficient in terms of time, economy, reproducibility, informativeness, resolution, and sensitivity, compared to other popular DNA markers. Besides, it requires very small quantities of DNA and no prior genome information. This technique is widely used in plants for taxonomy, genetic diversity, phylogenetic analysis, construction of high-resolution genetic maps, and positional cloning of genes, to determine relatedness among cultivars and varietal identity, etc. The review encompasses in detail the various applications of AFLP in plants and the major advantages and disadvantages. The review also considers various modifications of this technique and novel developments in detection of polymorphism. A wet-lab protocol is also provided.

Key words AFLP , cDNA, Epigenetics, Genetic diversity, Transcriptomics, MSAP , Restriction enzymes

1 Introduction

The AFLP technique is a patented technology first described by [1] and is applied widely in monitoring inheritance of agronomic traits in plants, pedigree analysis, parentage analysis, screening of DNA markers linked to genetic traits and genes of interest, etc. AFLP technique uses the entire genome for polymorphism and reproducibility and is recognized as a universal DNA fingerprinting system, universally accepted regarding origin and complexity of DNA samples and even small sequence variations that can be identified using a small quantity of DNA as low as 0.05 μg. A large number of

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_12,

fragments are detected on a gel that allows evaluation of a large number of loci at a time. It is much advantageous in terms of number of polymorphisms identified per reaction, reproducibility, ease, and cost of analysis.

1.1 Principle of AFLP

AFLP employs selective amplification of restriction fragments from a digested total genomic DNA using PCR. Genomic DNA is first digested by two restriction enzymes that cut the big molecules into a mixture of fragments enabling amplification by PCR. Usually in AFLP two restriction enzymes, a rare cutter like *Eco*RI (6-bp restriction site) and a frequent cutter like *Mse*I (4-bp restriction site), are used for restriction. Double-stranded oligonucleotide adapters consisting of a core sequence and a restriction enzyme-specific sequence homologous to one 5′ or 3′ end are then ligated to the DNA fragments using T4 DNA ligase. The ligated DNA fragments are amplified by PCR using primers complementary to the adapter and restriction site sequence with additional selective nucleotides at their 3′ end. Using selective primers reduces the complexity of the mixture, and those fragments complementary to nucleotides beyond restriction site will be amplified by these selective primers under stringent annealing conditions. Later, the polymorphisms are identified by a denaturing polyacrylamide gel electrophoresis and patterns between individuals are compared. In AFLP polymorphisms observed arise due to a mutation in the restriction site, a mutation in the regions complementary to primer extensions and adjacent to restriction site, or a deletion/insertion within the amplified region. Molecular polymorphisms are identified based on the presence or absence of particular DNA fragments of a given size among individuals (Fig. 1).

1.2 Basic Steps Involved in AFLP Analysis

A suitable DNA extraction protocol that yields good quality DNA without degradation may be employed for AFLP analysis. Quality of DNA needs to be ensured by an extra purification step; in case if the DNA extracts contain restriction or PCR inhibitors, an extra purification step may be incorporated. In AFLP, it is required to optimize the quantity of DNA for generating clear, intense AFLP patterns. These patterns may vary from species to species and depend on the genome size. Restriction fragments are generated using two restriction endonucleases, a rare cutting enzyme with 6–8 base recognition, in combination with a frequent cutting enzyme of four-base recognition. Enzymes are chosen based on the genome complexity and methylation status of the DNA. Complete digestion is to be ensured in order to avoid false polymorphisms due to amplification of fragments that are not fully digested. The AFLP protocol is designed to amplify and preferentially detect fragments with EcoRI cut at one end and MseI cut at the other. In AFLP different combinations of enzymes and multiple combinations of primers can be used for accessing hundreds of polymorphic markers.

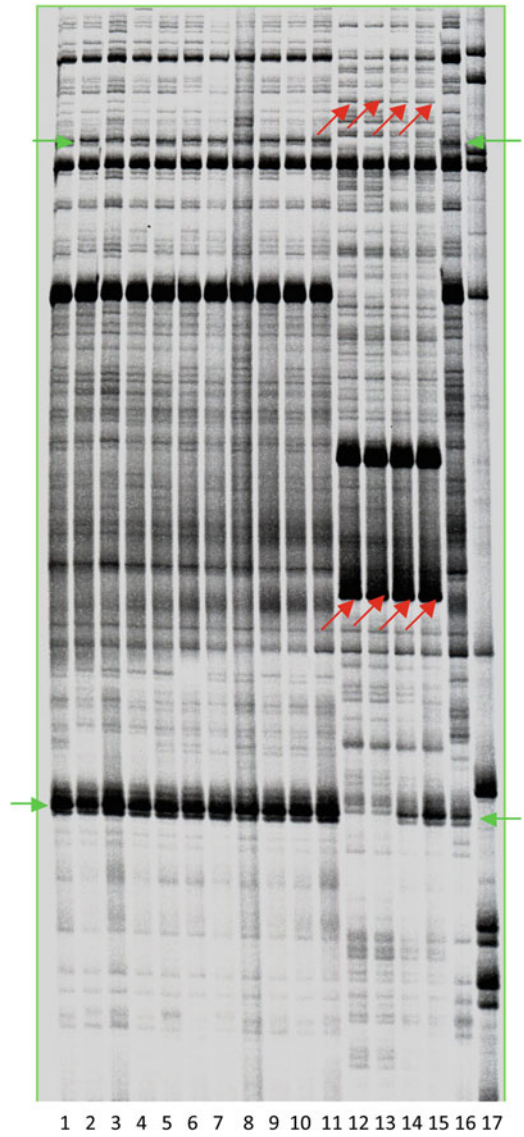


Fig. 1 AFLP profiles of *Vanilla* spp., seedling progenies, and interspecific hybrids, developed using primer combination EGG-MTG. Lanes 1–10: Seedling progenies of *V. planifolia* (1, V1; 2, V2; 3, V4; 4, V6; 5, V7; 6, V8; 7, V10; 8, V11; 9, V12; 10; V24); 11, *V. planifolia*; 12, *V. aphylla1*; 13, *V. aphylla2*; 14–16: interspecific hybrids of *V. planifolia* and *V. aphylla* (14, VH1; 15, VH4; 16, VH5); 17, Water Control. Arrows indicate species-specific bands (*V. planifolia* in green, *V. aphylla* in red)

1.3 AFLP Advantages and Applications

The AFLP technique is a robust tool due to the ability to generate quickly a large number of marker fragments without prior knowledge of the genomic sequence and can be multiplexed for analysis of hundreds of individuals at a time. It requires only a small quantity of DNA and is highly reproducible. Due to this reason, it is used in DNA fingerprinting of non-model organisms where no prior

sequence information is available. AFLP can be used for samples of any origin and complexity to detect sequence variations. Commercial AFLP primer sets are available which work on most organisms making this technique versatile. In-depth coverage of the genome is possible since large numbers of AFLP markers can be typed rapidly at a low cost. AFLP markers are largely independent since 90% of these reflect point mutations in the restriction sites. The co-migrating markers in AFLP are mostly homologous and locus specific and follow a Mendelian inheritance in plants [2, 3].

AFLP markers reveal a greater amount of diversity compared to other popular markers like RAPD, ISSR, SSR, RFLP, etc. (Chapters 11, 13, 14) and are highly reproducible and reliable due to the stringent hybridization conditions employed [4–6]. Due to these reasons, it can be upscaled, reproduced between different laboratories and conditions. These methods require very small quantity of DNA to generate huge amount of data.

AFLP differs from RFLP in that it employs PCR amplification to detect the polymorphisms on a denaturing PAGE while RFLP employs agarose or PAGE gels followed by hybridization. AFLP provides additional possibilities of detecting polymorphisms beyond the restriction site in comparison to RFLP wherein only the length variation within restriction sites is available and detects more point mutations, insertions, and deletions than RFLP to the tune of about 100–200 loci at a time. There is a scope of detecting unlimited polymorphisms by simply varying the restriction enzymes and the nature and number of selective nucleotides. AFLP fragments are mostly homologous and locus specific [7] except in polyploid species. Due to the above advantages, AFLP markers have proved effective in determining genetic differences among individuals, populations, and species. AFLP markers unravel cryptic genetic variation of closely related species which cannot be distinguished using conventional strategies. AFLP markers have the widest application in genetic variation analysis below species level for investigating population structure and differentiation and phylogenetic relationships based on genetic distances. They are highly instrumental in characterization of gene banks, fingerprinting, and estimation of genetic diversity for gene bank management. AFLP markers have been applied to evaluate gene flow and dispersal, outcrossing, introgression, and hybridization. The different applications of this versatile technique are detailed below.

1.3.1 Genetic Diversity Studies Using AFLP Markers

Analysis of genetic diversity and phylogenetic relationships is an important prerequisite for future breeding programs and conservation. It helps to understand evolutionary history of a species and the future risks to diversity. Evaluation of interpopulation variations indicates scope of geographic origin, dispersal of plant material, and gene flow between populations. Intraspecific genetic variability in

natural populations is an indicator of the potential to cope with changing environmental conditions and provides valuable inputs with respect to conservation and management of endangered and endemic plant taxa [8]. Diversity studies based on molecular markers are found to be more informative and reliable than that based on morphological and phenotypic traits. AFLP requires no prior sequence information and has a multi-locus and genome-wide nature, which makes it more popular than other molecular markers in DNA fingerprinting and genetic diversity analysis [6, 8–17].

In genus *Brassica*, several reports [4, 18–25] demonstrate the utility of AFLP in addressing important phylogenetic questions within the species and provide new insights for future breeding programs. In rice, AFLP analysis in four populations provided valuable insights regarding unique genes in Iranian native varieties, which will be useful for future breeding programs and stresses upon the need for conserving this unique diversity [26]. In *Jatropha*, AFLP analysis of five populations showed high intrapopulation variability, and this could identify promising genetic resources to be included in breeding programs [27]. Distribution of genetic variation in Illinois bundle flower was detected using AFLP markers, with a view to increase the efficiency of germplasm preservation and expedited plant breeding programs [28]. AFLP-based genetic diversity studies in *Pinus pinaster* populations provided important information on organization and subdivision of diversity, the genetic mechanisms underlying it, and sampling strategies to be adopted for species conservation [29].

Evidence for maintenance of genetic variability in Italian and Spanish durum wheat over the last century was revealed through AFLP marker-based analysis [30], which showed an enrichment of diversity in the cultivated pool and broadening of genetic background. In snap bean, AFLP-based genetic variability analysis exhibited a good level of variability and a possible relationship between bean growth habitat and the gene pool, which can be exploited for future breeding programs [31]. AFLP-based fingerprinting is a suitable technology for discovering genetic diversity in banana [32–36], and it also has an impact on conservation strategies and breeding ventures in banana. Phylogenetic and genetic diversity analysis of conserved endangered plant species has been successfully done through AFLP [37–40]. Diversity study within population and subpopulation of endangered sentry milk vetch (*Astragalus cremnophylax var. cremnophylax*) [41] through AFLP could estimate their adaptability to alien environments and also provides strategically important inputs for their conservation.

Germplasm collections have been characterized in *Jatropha curcas* [42] and *Rhodiola rosea* [43] using AFLP. Genetic diversity studies in natural populations of *Dendrobium thyrsiflorum* and radish [44] showed high interpopulation variations and correlation

of a few AFLP markers with the antioxidant activity [45] in case of the former. In teak, high genetic diversity could be observed within locations indicating importance of intensive location-wise collection of diverse superior genotypes for conservation and genetic improvement [46]. In lentil accessions genetic diversity and phylogenetic studies were conducted, and intraspecific genetic variability at high levels could be detected. An important outcome of this study was information on progenitor species of cultivated lentils [47]. Genetic diversity analysis in *Microlaena stipoides* using AFLP showed outcrossing and significant amount of variation within populations which can be used as a probable strategy for its propagation and for making microlaena more resilient in the long term [48].

Genetic relationships among different species of *Solanum* gave leads into the taxonomic resolution of this complex species and also provided insights into the origins/introductions of some of the important species [49]. Several other studies also have utilized AFLP for *Solanum* taxonomy [50–57].

In many cases AFLP analysis showed limited genetic diversity existing within germplasm collections, which indicates the need for conservation and also suggests that new accessions should be obtained from the center of origin of the species [58]. Intra-accession diversity studies in potato population showed lower levels of polymorphism within accessions of self-compatible when compared to self-incompatible taxa, thereby showing the high suitability of AFLP makers for evaluation of diversity between accessions in gene banks [59].

In many of the genetic diversity and phylogenetic studies, grouping of individuals showed high correlation with taxonomic and molecular classifications, indicating that the observed variations could be due to genetic factors. However, in some cases morphological and agronomic traits did not correlate well with molecular classification due to genotype × environment interaction and polygenic nature of the traits [60].

Using AFLP markers genetic variation was detected among tea genotypes [61] that could not be distinguished using morphological and phenotypic markers. The grouping of populations in a dendrogram was consistent with the taxonomy, known pedigree of genotypes, and geographical origin. Valuable observations could be made regarding the origin/ancestry and genetic diversity of tea from this study. Analysis of genetic diversity using AFLP markers in jackfruit [58] showed that grouping of accessions correlated well with the taxonomic classifications. Through this study incorrect classifications could be rectified, and self-fertilization of clones in a hybridization material could also be detected. Genetic diversity studies using AFLP assigned genotypes into groups corresponding to origin and lineage relationships in cotton which can be exploited

in marker-assisted parental selection tool for plant breeders [62]. A study involving three species of Malvaceae depicted good congruence of AFLP-based clustering with earlier morphological and molecular investigations [63]. In pineapple cultivars from Thailand, AFLP-based clustering revealed moderate genetic diversity and congruence with earlier morphological characterization [64]. Phylogenetic relationship studies indicated that AFLP data correlated well with the taxonomic relationships among the cultivated lettuce and wild species, and the dendrogram generated was similar to the phenetic tree constructed using RFLP data [65]. In *Triticum aestivum* genotypes, a moderate correlation between AFLP and morphological markers was observed [66], while in olive cultivars, AFLP fingerprinting of core collection discriminated different cultivars, but clustering based on AFLP and fruit traits did not show significant correlation [67]. In azalea [68] and banana [32, 33] cultivars, no correlation between AFLP data and morphological traits existed, indicating that the majority of the polymorphisms did not contribute to phenotypic variation.

Genetic diversity and influence by environment could provide a better understanding of the natural variation and gene exchange that existed in a species with respect to its geographical location. This can help in preservation and development of germplasm resources especially in case of endangered species. In some studies a good correlation of AFLP data with the geographical origins and distance could be observed. In *Vigna* sp. [69], *Triticum* landraces [70], and banana [71], significant association was observed between AFLP data and geographic location. In *Hibiscus tiliaceus*, estimates of genetic diversity using AFLPs agreed well with the geographical distribution and life history traits [72]. AFLP analysis of Iranian potato germplasm [73] and *Lactuca* species [74] showed a high level of genetic diversity and clustering corresponding to the geographical origin of these varieties. In cowpea genetic distances were estimated in wild, weedy annuals, domesticated cowpea, perennial accessions, and wild subspecies, and AFLP markers could successfully uncover variation within both domesticated and wild accessions [75].

In alfalfa [76], soybean [77], and *Croton* sp. [78], AFLP was used to study genetic diversity of cultivated and natural populations, which showed no correlation between genetic and geographic distances. In betel vine cultivars, cluster analysis based on AFLP data showed that grouping of individuals was based on their genetic relatedness rather than place of collection [79]. In kale, landraces, cultivars, and wild populations exhibited higher levels of diversity among wild populations. The study indicated that genetic distance was not related to geographical distance and provided inputs on conservation strategies to be adopted [80]. Wild

populations of *Agave angustifolia* fingerprinted using AFLP showed a partial correlation with geographical distribution and variation between mother plants and vegetatively propagated mother rhizomes [81]. In the endangered *Glehnia littoralis*, AFLP analysis showed no obvious correlation between genetic and geographic distances, and the endangered status was attributed to the loss of wild habitats calling for ecological conservation strategies [16]. In black gram AFLP-based clustering of landraces indicated influence of soil pattern and topography in the genetic makeup and genetic distinctness [82].

1.3.2 Variety/Cultivar Fingerprinting, Kinship, and Genetic Fidelity

Lack of genetic identity is a serious problem in plant propagation and seed production of elite genotypes. For certification purpose, genotypes need to be characterized both at phenotypic and molecular level for identifying promising ones with outstanding agronomic, nutraceutical, and nutritional characteristics. Availability of informative molecular markers is an essential prerequisite for proprietary protection, establishing identity, early detection of seedlings in the nursery, and monitoring trade. AFLP being a dominant marker system and the availability of multi-locus and genome-wide marker profiles are the reasons that make it a preferred method for DNA fingerprinting [42]. Several studies endorse the utility of AFLP markers for discriminating between closely related individuals when compared to nuclear and chloroplast DNA markers [83, 84]. AFLPs are also the preferred method for establishing genetic fidelity in in vitro culture systems especially in commercial propagation [85] where soma clonal variation is a problem.

Along with genetic variability estimations in selected cultivars and lines of *Cornus florida,* a dichotomous key using specific AFLP markers was constructed to distinguish some of the popular cultivars and breeding lines [86]. Genomic fingerprints of elite genotypes of farmers were done using AFLP markers for the purpose of variety protection, seed certification, and future support to breeding programs in blackberry [87] and for detection of duplicates in germplasm collections of yam [88]. AFLP markers have the potential to resolve genetic differences at the level of "DNA fingerprints" for individual identification and parentage analysis [89].

In case of identification of clonally identical individuals, a large number of markers need to be screened to uncover existing genetic differences due to their extremely close nature. Clonally derived individuals in several plants could be delineated by AFLP making them suitable for analysis of relatedness, parentage, mating frequency, etc. due to low levels of co-migration of non-allelic fragments. AFLPs clearly established their utility for clonal differentiation and/or identification in *Vitis vinifera* ecotypes [90], and the profiles were well in congruence with those generated by ISTR (inverse sequence tagged repeat) markers. However, in

certain populations, ISTR revealed more polymorphism. Differences at the molecular level were identified between agave offsets and bulbils produced asexually from the same mother plant from different tissues using AFLP depicting the great potential of this method in plant cultivar identification [91]. In near-isogenic lines of soybean, distinguishing between individuals that differ at only a single small region in the entire genome was possible [9]. AFLP markers also enable testing of clonal identity between individuals and thus permit to make inferences about the sexual versus asexual reproduction modes [92].

AFLP markers have also been used to establish genetic fidelity in in vitro derived plants in several crops for confirming the commercial-scale plant production protocol [93, 94]. Clonal fidelity of micropropagated plants was established through AFLP in endangered *Arachis retusa* for germplasm storage and in *Dendrocalamus hamiltonii* [95]. In *Bambusa nutans*, AFLP revealed a high level of genetic stability in somatic embryo-derived plantlets [96]. AFLP successfully identified variations in cryopreserved in vitro shoot tips in *Rubus* [97].

1.3.3 QTL Mapping

AFLP markers have been used extensively for constructing linkage maps for QTL analysis of agronomic traits including disease resistance and salt tolerance [98–123]. AFLP markers have been widely used for map-based cloning of target genes linked to them, and SCAR markers for quality traits were developed in asparagus bean [124], alfalfa [125], tomato [126], eggplant [127], and maize [128].

1.3.4 Other Specific Applications of AFLP Marker Systems

In barley, AFLP assay and bulked segregant analysis involving selected individuals of a cross between water stress-tolerant and stress-sensitive genotypes identified a marker that was present only in the tolerant parent and tolerant bulk of F2 individuals [129]. In *Salvia miltiorrhiza* segregating sterile and fertile populations when subjected to bulked segregant analysis and AFLP marker analysis indicated several markers tightly linked to the drought stress genes. One of the markers was found to be identical to another marker tightly linked to male sterile gene with 95% identity [130]. Molecular tagging of male sterility locus was done using AFLP technique in a BC1 mapping population segregating for male sterility/fertility. Markers were identified for marker-assisted selection and genetic map constructed for the male sterility gene [131]. In *Piper betle*, a combination of bulked segregant analysis and AFLP screening identified two male sex-specific markers [80]. Bulked segregant analysis combined with AFLP identified markers linked to resistance to yellow rust disease in *Triticum aestivum* L [132]. AFLP coupled with bulk segregant analysis could identify markers linked to virus disease in tomato [133].

Species-specific AFLP fingerprints were generated and used for authentication in three species of *Zingiber*, which is proposed to help in resolving adulteration-related problems faced by commercial users [134]. In Andean blackberry, attempt was made to generate genomic fingerprints that will enable protection, seed certification, and future support to breeding programs [88]. In a study, AFLP genome scan was combined with environmental analysis for testing natural populations of *Liriodendron chinense* for signals of natural selection, and it identified a few outlier locus strongly associated with climatic factors [135]. AFLP investigation of 14 wild *D. glomerata* indicated that the genetic diversity and structure pattern of populations could be influenced by environmental factors like altitude, precipitation, latitude, and longitude [136]. In *Lactuca* sp. studies indicate that ecogeographical conditions can influence the genetic background of populations originating from them [137], and influence of biotic and abiotic stresses in the center of origin regions can lead to high genome-wide diversity in populations [138]. In rice, several high temperature responsive transcript-derived fragments (TDFs) were identified employing differential gene expression analysis coupled with AFLP [139]. Similar strategy in sugarcane identified several induced and repressed TDFs in response to infection by Sugarcane Mosaic Virus [140].

Isolation and characterization of differential genes in *Capsicum annuum* L. using AFLP indicated that space flight influenced main quality characters at genetic level, and induction of several novel genes was observed [141]. In *Spondias tuberosa* [142], outcrossing rates estimated using AFLP in a large population involving 12 families exhibited the open pollinated nature of the species and provided valuable inputs on strategies for conservation and breeding.

In Oregano, a high correlation between key chemotypic traits and AFLP markers could be established [143]. Genetic diversity assessments by AFLP markers in populations of *Amaranthus palmeri* was done to understand the distribution and development of herbicide resistance to glyphosate [144]. AFLP also helps to target other levels of diversity especially DNA methylation polymorphism and transcriptomic variation [145].

1.4 AFLP Versus Other Popular DNA Markers

In several species a greater degree of polymorphism was observed in AFLP-based diversity analysis compared to other popular markers like SNP, SSR, ISSR, and RAPD [146–151]. In vanilla RAPD and AFLP profiles coupled with morphological characters could successfully assess variability of genotypes and of successful interspecific hybridization and production of hybrids [152]. Genetic relationship studies in soybean genotypes [153] indicated a lower level of expected heterozygosity in case of AFLP markers in comparison with microsatellites and RAPD, in spite of the fact that AFLP generated the highest effective multiplex ratio as in other

studies. However, the marker index, a parameter involving expected heterozygosity and multiplex ratio, was much higher for AFLP markers indicating its superiority for detecting polymorphisms. The RFLP, AFLP, and microsatellite marker systems showed a good correlation in the present study. In *Brassica napus* hybrids, SSR was found to be more efficient than AFLP in evaluating genetic diversity, while AFLP was better for varietal identification and DNA fingerprinting [154]. In common bean SSR and AFLP showed a comparable accuracy in grouping genotypes according to their gene pool of origin [155]. AFLP was found to be the best molecular marker for fingerprinting and assessing genetic relationship among genotypes of *Dactylis glomerata* when compared to other markers like RAPD and ISSR [156].

In brinjal [157], *Jatropha* [158, 159], sugarcane [160], and *Miscanthus* sp. [161], the superiority of AFLP over RAPD in discriminating genotypes and estimation of genetic diversity was reported. In yet another study on *Aegilops* species, 50 populations analyzed using AFLP showed superiority of AFLP markers over RAPD as a tool for molecular variability studies in plant breeding programs [162]. AFLP turned out to be a better method for obtaining a more definitive grouping for study of genetic relationships both at species and cultivar level [35] in banana. AFLP was more efficient compared to SSR markers for detecting genetic variation among Ethiopian Arabica coffee genotypes [163], and on a small spatial scale, AFLPs outperformed SSRs in discriminating individuals and assigning them to population of origin [164] in *Eryngium*. In banana [36] estimates of genetic diversity did not show any significant correlation between microsatellite and AFLP markers. In maize [165], SSR and AFLPs were found to be equally suitable for genetic diversity studies. However, intrapopulation diversity studies in neem indicate a better efficiency of SAMPL markers over AFLPs in resolving differences between closely related accessions [166]. SRAP markers were found to be more informative than AFLP in giving high number of unique markers for identification of banana genotypes [167].

However, in the genus *Ocimum*, a combined analysis of morphological traits, volatile oil composition, and molecular markers is found to be an ideal strategy for taxonomical classification [168]. Genetic relationship study showed good correlation between AFLPs and RAPDs in potato and endorsed the application of a combination of marker systems like AFLP, SSR, and RAPD for better understanding of genetic relationship [169].

1.5 Disadvantages of AFLP Technique

AFLP is a cumbersome process involving several steps and requires reasonably large quantity (300–1000 ng per reaction) of good quality DNA and is a technically complicated procedure than simple markers like RAPD. AFLP employs polyacrylamide gels and silver staining and radioactivity of fluorescent probes for detection that

are laborious and expensive compared to agarose gels. It requires ligation and restriction enzymes and adapters, which adds to the extra cost compared to techniques like RAPD (Chapter 13). Post-run data analysis is lengthy and complex compared to RAPD. However, recently available kits and automation have made it more user-friendly. AFLP markers are dominant biallelic markers and polymorphic information content is low (maximum is 0.5). It is difficult to distinguish between heterozygous and homozygous individuals for the presence of allele, and precise estimation of heterozygosity is not possible, which limits its usage in population genetic analysis, genetic mapping, and marker-assisted selection. AFLP technique can produce artifacts in degraded samples like herbarium specimens, and to overcome this, fresh samples were included for comparison, thereby ensuring the presence of monomorphic fragments in the fresh samples as well as herbarium AFLPs [170].

1.6 Modifications of AFLP

1.6.1 SAMPL

The selectively amplified microsatellite polymorphic loci (SAMPL) marker technique may be employed to detect higher levels of genetic variation within genotypes. SAMPL is a microsatellite-based modification of the AFLP assay and has all the advantages of the latter [171]. Due to its association with the hypervariable microsatellite region, this assay can detect high levels of polymorphism between closely related genotypes. Due to its ability to survey the hypervariable microsatellite region in the genome, it can detect higher levels of polymorphism per locus compared to AFLP. The SAMPL assay has been employed for analysis of genetic diversity in lettuce [172] and sweet potato [173] among other crops [174, 175]. The SAMPL assay revealed higher levels of polymorphism among *Withania somnifera* genotypes compared to the use of standard AFLP in all the genotypes tested. The AFLP markers and their modifications such as SAMPL are generally expensive to generate, technically tedious, and dominant in nature. This limits their large-scale application as diagnostic markers for species, cultivar, or varietal identification. For practical applications, these markers need to be converted to rapid, technically simple assays that can be used on crude DNA preparation. A fruitful attempt at converting SAMPL markers to useful diagnostic markers was one where *W. somnifera*-specific bands generated with SAMPL were used to develop a simple PCR-based assay [174]. All the tested genotypes can be distinguished at the seedling stage by the diagnostic markers generated.

1.6.2 M-AFLP

Microsatellite-amplified fragment length polymorphism (M-AFLP) is a modification of AFLP to detect intravarietal genetic differences and is known to be the most efficient system and generates the highest number of polymorphic bands compared to SSR, AFLP, and SAMPL [176]. Markers are anchored to the 5′-end of

microsatellite (e.g., SSR) loci in this new AFLP-derived marker system. M-AFLP combines the high heterozygosity of microsatellites (SSRs) with high multiplex ratio of AFLP-derived markers. Variation in the number of repeat units is the source of polymorphisms detected by the M-AFLP, and it is employed to develop SSR-type codominant markers from polymorphic M-AFLP bands. The technique does not require hybridization enrichment steps and provides substantial efficiency of SSR identification compared with conventional library procedures [177]. M-AFLP has been employed in cassava for genetic diversity analysis of cassava and other *Manihot* species [178], in grapevine for clone differentiation and varietal identification [176], in *Cynara cardunculus* for microsatellite locus identification [179], in *Poa pratensis* L. for genetic mapping of complex polyploids [177], and in *Lupinus angustifolius* L. for the isolation of sequence-specific PCR markers [180].

1.6.3 SSAP

Sequence-specific amplified polymorphism (SSAP) analysis [181] was one of the first retrotransposon-based barcoding methods based on AFLP. The BARE-1 LTR-RT is utilized by SSAP technique for molecular barcoding [181] using one primer complementary to an RT (e.g., 3′ LTR) and the other primer complimentary to the AFLP-like restriction site (usually MseI or PstI) adaptor. Primer pairs contain two or three selective nucleotides of MseI or PstI (or any restriction enzyme) adaptor primers and one selective nucleotide of either ^{32}P or fluorescently labeled retrotransposon-specific primers [179]. The primers in SSAP technique are designed based on the LTR region, but could also match to an internal sequence of the RT, like the polypurine tract (PPT), which is found internal to the 3′-LTR of retrotransposons [179]. When restriction enzymes have a long recognition site sequence, nonselective primers could also be used or when the copy number of the RTs is low. The type of SSAP primers used determines the quality of the SSAP pattern. SSAP usually exhibits higher level of polymorphism compared to AFLP and has been extensively used for diversity analysis studies in *Triticum* spp. [182], *Hordeum vulgare* [183], *Avena sativa* [184], *Aegilops* spp. [185], *Malus domestica* [186], *Cynara cardunculus* [187], *Lactuca sativa* [188], *Pisum sativum* and other Fabaceae species [179, 189], *Capsicum annuum*, *Solanum lycopersicum* [190], and *Ipomoea batatas* [191]. SSAP was also used for cladistic molecular barcodes to resolve evolutionary history in *Nicotiana* [192], *Vicia* [193], *Oryza* [194], *Triticum* [182], and *Zea* [195].

1.6.4 AIMS

The amplification of insertion mutagenized sites (AIMS) technique is mainly based on reducing the band complexity by specific PCR amplification of insertion mutagenized sites, by using a primer that is specific to *Mutator* transposon flanking sequences [196]. AIMS

procedure delivers possible gene candidates, but isolation of the gene has to be verified by another method. MuAFLP, another variant of AFLP, is similar to AIMS, and it targets amplification of *Mutator* transposon regions [197].

1.6.5 MSAP

The methylation-sensitive amplified polymorphism (MSAP) technique mainly involves cleavage with the methylation-sensitive restriction enzymes *HpaII* or *MspI*, followed by adapter ligation, amplification, and gel-based visualization [198, 199]. The methylation state of the external and internal cytosine residues strongly affects the cleavage capacities of *HpaII* and *MspI* within the recognized 5′-CCGG-3′ sequences. Thus, the methylation state is determined based on the ability of each enzyme to cleave the restriction site, for each of the specific bands. MSAP-based analyses can be performed for a range of species regardless of their genome size and availability of reference genome. Established in 1997 [198], MSAP has been effective in analyses of DNA methylation in various plant species [200–209]. This technique is widely used in non-model and model plants [210–214]. Being simple and useful, MSAP only provides a general overview of the methylation state and does not provide a specific sequence context. A novel technique called Methylation Sensitive Amplification Polymorphism Sequencing (MSAP-Seq) for the analysis of DNA methylation patterns in *Hordeum vulgare* based on the conventional MSAP analysis, with direct high-throughput sequencing using next-generation sequencing (NGS) and automated data analysis, was introduced [215]. MSAP-Seq allows for the global and direct identification of a large set of sequences that undergo DNA methylation changes without laborious band excisions, re-amplification, and subcloning, which are required for MSAP analysis.

1.6.6 AFLP-RGA

Resistance gene analog-anchored amplified fragment length polymorphism (AFLP-RGA) is a modified AFLP procedure first proposed in soybean (*Glycine max* L.) [216]. Here the degenerate RGA primers are used in combination with selective AFLP primer in the second round of amplification. The AFLP-RGA method combines the approach of AFLP with gene-anchored amplification and can provide more functional markers that are possibly distributed in other regions of the genome, thereby increasing the genome coverage.

1.6.7 TE-AFLP

The three endonuclease AFLP (TE-AFLP) technique reduces the number of amplified fragments not only by primer extension but also by selective ligation. Three endonucleases and two sets of adapters are used in a single reaction. As a consequence, the reduced number of potential amplifiable fragments diminishes competition during PCR, permitting stringent reaction conditions

and thus eliminating the need for a two-step amplification in fingerprinting complex genomes. TE-AFLP primer combinations generated a total of 12 and 48 polymorphic bands in 12 *Pongamia* accessions from different regions of Delhi [217].

1.6.8 SDAFLP

The secondary digest AFLP (SDAFLP) is a variation of MSAP technique wherein a restriction endonuclease site-specific single primer is used to amplify the digested template DNA and later digested with a methylation-sensitive enzyme. The fragments are re-amplified using a primer from previous amplification and a second primer specific to cleavage sites of methylation-sensitive primer [218].

1.6.9 MITE-AFLP

Miniature Inverted-repeat Transposable Elements (MITEs) were transposon elements discovered in plant genomes [219]. A successful application of conserved motif of a *Mite* element as a molecular marker in maize was demonstrated [220] with minor modifications of AFLP protocol.

1.6.10 RNA Fingerprinting Using cDNA-AFLP

cDNA-AFLP is a variation that combines RNA fingerprinting technique and AFLP wherein the standard AFLP protocol is applied on a cDNA template. This method is comparable with the northern blot analysis in studying gene expression [221]. This method is a useful modification to the RNA fingerprinting since it is possible to eliminate all nontarget bands. This modified method can be utilized in gene expression studies vis-a-vis biological pathways in plants. AFLP has also been used to generate mRNA fingerprints in hexaploid wheat and one of its deletion mutants, and the method was found useful for isolating sequences mapping to deleted chromosome segments in hexaploid wheat [222].

1.6.11 Nonradioactive DD-AFLP

It is a method of coupling differential display (DD) and AFLP for monitoring differentially expressed genes. Here double-stranded cDNA molecules are restricted and ligated to the defined adaptor sequences followed by amplification of a subset of ligation products with adaptor-specific primers carrying two or more arbitrary nucleotides and detection of bands representing gene of interest on a polyacrylamide gel. It is considered as a high-throughput method in functional genomics, and DD-AFLP patterns can be simulated for sequenced genomes by computer softwares, and information on undetermined genomes can be retrieved. Several modified methods that avoid use of radioisotopes were optimized and were widely used for detection of responsive genes in plants and tissues subjected to elicitors [223].

1.7 Patents and IPR Protection

Two patents regarding AFLP technology have been filed in the year 2018 and 2019. One patent is concerned with high-throughput detection of molecular markers based on AFLP and high-throughput sequencing. The invention relates to a high-

throughput method for the identification and detection of molecular markers wherein restriction fragments are generated and suitable adaptors comprising (sample-specific) identifiers are ligated. The restriction fragments which are adapter-ligated may be selectively amplified with adaptor-compatible primers carrying selective nucleotides at their 3′ end. The resulting fragments are sequenced at least partly using high-throughput sequencing methods, and the sequence parts of the restriction fragments together with the sample-specific identifiers serve as molecular marker. The other patent is titled as method for high-throughput AFLP-based polymorphism detection. The invention is mainly intended for discovery, detection, and genotyping of one or more genetic markers in one or more samples, comprising the steps of restriction endonuclease digest of DNA, adaptor ligation, optional pre-amplification, selective amplification, pooling of the amplified products, sequencing the libraries with sufficient redundancy, clustering followed by identification of the genetic markers within the library and/or between libraries, and determination of codominant genotypes of the genetic markers [224].

1.8 Conclusions

The wide popularity of AFLP technology is evident from the available literature. It has immense future prospects due to the versatility and flexibility especially in situations where no genomic information is available. The method is reliable both under sophisticated and ordinary conditions of processing and detection. While choosing an appropriate method for molecular marker analysis, the important factors into consideration are low cost, good throughput, convenience, and ease of operation and automation. RAPD, RFLP, SSR, etc. are popularly used markers and each one has its own advantage. However, many studies that we have mentioned in this chapter endorse the superiority of AFLP in diversity analysis, phylogenetic characterization, fingerprinting, etc. Despite the fact that AFLP provides a better coverage and estimate of genetic diversity, it is prudent to consider markers like SSR that are codominant and enable discrimination of heterozygous and homozygous individuals. Dominant AFLPs cannot be used to study heterozygosity. An integrated marker approach was found to be better in many studies for more accurate genotype characterization and taxonomy. It is prudent to use an appropriate marker considering the biological question and geographical scale investigated, last but not least the financial and resource constraints prevailing. More importantly results from molecular studies need to be integrated with knowledge on the morphological characteristics for a better understanding toward genetic improvement as well as germplasm conservation programs.

2 Materials

In case of AFLP ready-made chemicals are generally used. Uniformity in terms of chemical concentration needs to be maintained for all individuals to be analyzed in the AFLP experiments. All the reagents need to be stored at −20 °C. Some AFLP kits are currently available (*see* **Note 1**).

2.1 DNA Template Preparation

1. TE buffer (1): Dissolve 10 mM Tris–HCl and 1 mM EDTA in 1 L ddH_2O, and adjust to pH 8. Store at room temperature.

2.2 Restriction-Ligation (RL)

1. MseI restriction endonuclease (the "frequent cutter"—recognizes a four-base motif, i.e., 5′-TTAA). 1 U MseI is required for one reaction.
2. EcoRI restriction endonuclease (the "rare cutter"—recognizes a six-base motif, i.e., 5′-GAATTC). 5 U EcoRI is required for one reaction.
3. MseI-adaptor pair: 5′-GACGATGAGTCCTGAG and 5′-TAC TCAGGACTCAT. Stored at −20 °C as stock with concentration of 100 μM. Immediately prior to adding to the RL reaction, mix in proportion 1:1 (to obtain a concentration of 50 μM for each), then denature (i.e., heat up at 95 °C for 5 min) the required amount of combined MseI adaptors, and allow slow renature (let them cool slowly at room temperature for 10 min) to form double-stranded adaptor. Spin briefly.
4. EcoRI-adaptor pair: 5′-CTCGTAGACTGCGTACC and 5′-AAT TGGTACGCAGTCTAC. Store each adaptor primer individually at −20 °C as stock with concentration of 100 μM. Immediately prior to adding to the RL reaction, mix in proportion 1:1 (to obtain a concentration of 50 μM for each), then denature (i.e., heat up at 95 °C for 5 min) the required amount of combined EcoRI adaptors, and allow slow renature (let them cool slowly at room temperature for 10 min) to form double-stranded adaptor. Spin briefly.
5. T4 DNA ligase: 0.6 U T4 DNA ligase is required per ligation reaction.
6. T4 DNA ligase buffer.
7. BSA (bovine serum albumin). Stock solution of 10 mg/mL. Dilute prior to use (1 mg/mL).
8. 0.5 M NaCl.
9. TE 0.1 M buffer (1×): Dissolve 20 mM Tris–HCl and 0.1 mM EDTA in 1 L ddH_2O, and adjust to pH 8. Store at room temperature.
10. TBE buffer (stock solution 10×): Dissolve 108 g Tris base, 55 g boric acid, and 8.1 g Na_2EDTA in 1 L ddH_2O. Make up the pH to 8.2–8.3.

11. Size ladder of 1500 bp.
12. Loading buffer for electrophoresis.

2.3 Pre-Selective PCR Amplification (See Note 1)

1. AmpliTaq or RedTaq.
2. Taq DNA polymerase buffer.
3. Deoxynucleotide mix (dNTPs) in concentration 10 mM each dATP, dCTP, dGTP, dTTP. Ready-made mix (e.g., GeneAmp dNTP Blend, 10 mM, from Life Technologies) is recommended.
4. EcoRI primer: 5′-GACTGCGTACCAATTCA. Store as stock solution at 100 μM.
5. MseI primer: 5′-GATGAGTCCTGAGTAAC. Store as stock solution at 100 μM.
6. TE 0.1 M buffer (1×) (prepared as above).
7. 1000 bp ladder.

2.4 Selective PCR Amplification

1. RedTaq (1 unit).
2. RedTaq buffer (10×).
3. dNTPs (10 mM).
4. EcoRI primers: 5-GACTGCGTACCAATTCXXX where X stands for selective nucleotides. These primers are fluorescently labeled, and the working concentration of the EcoRI primer is 1 μM. Store as stock solution (100 μM) for several years and as working solution (1 μM) for several months (*see* **Note 2**).
5. MseI primers: 5-GATGAGTCCTGAGTAAXXX where X stands for selective nucleotides. The working concentration of the MseI selective primer is 5 μM. Store as stock solution (100 μM) and as working solution (5 μM) (*see* **Note 2**).
6. Thermal cycler.

2.5 Separation and Visualization of Fragments (See Note 3)

1. Sephadex G-50 Fine or Superfine. Weigh 10 g of the powder and mix with 120 mL ddH_2O and 100 μL 100× TE buffer. Let it stand for a couple of hours. Store at room temperature and use within 1 week. The solution of Sephadex settles out, and it should be resuspended before using.
2. MultiScreen HV plates. Store at room temperature.
3. GeneScan ROX or another fluorescently labeled, internal ladder suitable for sequencers. Store at 4 °C.
4. Hi-Di formamide.
5. LI-COR DNA Analyzer used for visualization of fragments.
6. Polymer and buffers, specific for the type of sequencer used. Usually stored at 4 °C.

3 Methods

In addition to the described methods, recent AFLP modifications in procedure and detection are also available (*see* **Note 3**).

3.1 DNA Template Preparation

The AFLP procedure requires genomic DNA stored in 1× TE buffer.

3.2 RL (Restriction-Ligation)

1. Heat the required amount of MseI (50 μM) and EcoRI (5 μM) of each adaptor pairs at 95 °C for 5 min, each pair in a separate vial. Allow them to cool gradually to room temperature for 10 min. Spin briefly in a microcentrifuge for 10 s (*see* Subheading 2.2, **item 3**).
2. Master mix for all samples is to be prepared, which is planned to be analyzed in one batch, starting with ddH_2O, T4 ligase buffer (contains 50 mM Tris–HCl, 10 mM $MgCl_2$, 1 mM ATP, and 10 mM dithiothreitol in a solution of pH 7.5 at room temperature), T4 ligase (0.6 units), NaCl (0.5 M), BSA (1 mg/mL), both adaptor pairs, and finishing with the three enzymes. Spin briefly (*see* Subheading 2.2, **item 4**).
3. Aliquot 5.5 μL of the master mix in individual tubes.
4. For each sample, add 5.5 μL DNA in one tube. The final reaction volume will be 11 μL. Vortex and centrifuge briefly.
5. The reaction is incubated at 37 °C for at least 3 h in a thermal cycler with a heated cover. The incubation is continued at 17 °C overnight, or at least for 3 h (17 °C is the optimum temperature for ligation activity).
6. The efficiency of the restriction reaction can be tested by running 5 μL of several of the reactions on 1.5% agarose gel prepared in 1× TBE buffer for 20 min at 90 V (*see* **Note 4**).
7. The reaction is stopped by diluting it 20-fold with 1× TE 0.1 M buffer.
8. The RL reactions can be stored for longer periods at −20 °C.

3.3 Pre-selective PCR Amplification

1. Dilute and mix pre-selective primers in proportion of 1:1:18 with ddH_2O to result in a working concentration of 5 μM each primer (*see* Subheading 2.3, **items 4** and **5**).
2. Prepare a master mix for all samples that you plan to analyze in one batch, starting with ddH_2O, 10× Taq buffer (2.5 μL for each reaction), dNTPs (10 mM), primers (5 μM each), and Taq polymerase (1 unit for each reaction). The quantities of various components are according to manufacturer's instruction.
3. Aliquot 8 μL of the master mix in individual 1.5 mL Eppendorf tubes.

4. Add 2 μL of the diluted RL product to each tube. The final reaction volume will be 10 μL. Vortex and centrifuge (1500 × *g*) briefly.
5. Use a thermal cycler with heated cover and run the following program: one hold of 72 °C for 2 min; 20 cycles of 94 °C for 1 s, 56 °C for 30 s, and 72 °C for 2 min; and finish with a hold of 60 °C for 30 min.
6. The efficiency of the pre-selective amplification can be tested by running 5 μL of several of the reactions on a 1.5% agarose gel in 1× TBE buffer, for 20 min at 90 V. If the RedTaq polymerase is used, no loading buffer is to be used. A smear product with few brighter bands in the 100–1500 base pair range should be visible (*see* **Note 4**).
7. Dilute the pre-selective reactions 20-fold with 1× TE 0.1 M buffer. Mix thoroughly. For the samples for which an aliquot of the PCR product has been run on agarose gel, reduce the dilution volume.
8. Store the diluted pre-selective reactions in the fridge for 1 day and at −20 °C for months.

3.4 Selective PCR Amplification (See Note 5)

1. Prepare a master mix for all samples that is planned to be analyzed in one batch, starting with ddH_2O, 10× Taq buffer (2.5 μL for each reaction), dNTPs (10 mM), primers (EcoRI primer 1 μM and Mse I primer 5 μM), and finishing with the Taq (1 unit). The components were added based upon manufacturer's instruction. Spin briefly.
2. Aliquot 8 μL of the master mix in individual 1.5 mL Eppendorf tubes.
3. Add 2 μL of the diluted pre-selective product to each tube. The final reaction volume will be 10 μL. Vortex and centrifuge (1500 × *g*) briefly.
4. Use a thermal cycler with heated cover and run the following program (90% ramp time): one hold of 94 °C for 2 min; nine cycles of 94 °C for 1 s, 65 °C—1 °C every cycle for 30 s, and 72 °C for 2 min; followed by 23 cycles of 94 °C for 1 s, 56 °C for 30 s, and 72 °C for 2 min; and finish with a hold of 60 °C for 30 min. Program the cycler to keep the reactions at 4 °C until they are removed.
5. Freezing the selective reactions is recommended as soon as possible. They can, however, be kept for 1 day in the fridge.

3.5 Separation and Visualization of Fragments (See Note 6)

1. Apply 200 μL of mixed Sephadex solution to each well of a MultiScreen (MS) HV plate. Place the MS plate on top of a microtiter plate to collect water. Pack the Sephadex by spinning at 600 × *g* for 1 min. Discard water that has been collected in the microtiter plate.

2. Repeat **step 1**.
3. Repeat **step 1** by packing the Sephadex by centrifuging at 600 × *g* for 5 min.
4. The MS plate is placed along with the Sephadex filter on top of a fresh microtiter plate to collect the filtered selective product.
5. Mix together the selective reactions of up to three primer combinations corresponding to one individual sample, by applying 5 μL of each selective PCR product, and the PCR product was labeled separately for easy identification (e.g., labeled green, yellow, and blue). Spin the MS plate (on top of the clean microtiter plate) at 600 × *g* for 5 min (*see* **Note 6**).
6. Discard the Sephadex filter. The HV plate can be reused for up to ten times after washing.
7. Make up the loading mixture for the number of samples to be loaded on the sequencer using 9.8 μL Hi-Di formamide and 0.2 μL of GeneScan ROX per sample. Do not forget to account also for two more samples as a tolerance for potential pipetting inaccuracies.
8. Aliquot 10 μL of loading mixture to each well of a clean microtiter plate.
9. Add 1.2 μL of the filtered, combined selective products to each well. Vortex and centrifuge briefly.
10. Cover the microtiter plate containing loading mixture and sample; heat it up at 95 °C for 5 min and cool the plate on ice immediately to denature the AFLP fragments.
11. Load the plate containing the denatured samples onto the sequencer.

4 Notes

1. PE Applied Biosystems (Foster City, CA, USA) has developed an AFLP™ Plant Mapping Kit based on the AFLP procedure patented by Keygene NV (Wageningen, The Netherlands). Two modules are available depending on the genome size. The Small Plant Genome Kit is used for genomes ranging from 50 to 500 megabases, and the Regular Plant Genome Kit is for genomes of 500–5000 megabases. Restriction fragments are generated using EcoRI and MseI restriction enzymes. For pre-amplification, both pre-selective primers in the Regular Plant Genome Kit have an additional selective nucleotide at the 3′-end. However, only the MseI pre-selective primer has a selective base in the Small Plant Genome Kit. AFLP Analysis System II, a kit developed by Thermo Fisher Scientific, is designed for use with plants having genomes ranging in size from 1×10^8 to 5×10^8 bp. The

AFLP Analysis System I is designed for plants having genome size of 5×10^8 to 6×10^9 bp range. AFLP kits developed by Li-COR Biotechnologies also helps to genotype individuals in certain populations with less genetic variability.

2. The number of selective nucleotides of the primers can be increased or decreased based on the genome size and the availability of restriction sites in the genomes that are to be analyzed. Longer pre-selective and selective primers are used for large genomes and shorter selective primers, with only two selective nucleotides for smaller genomes. The use of a different combination of restriction enzymes results in fine-tuning of the number of AFLP fragments generated as a result.
3. Since the original AFLP protocol was published (1), numerous variants have been introduced. The major improvements in the main protocol include (1) the use of IRDye® infrared dye (IRD) or other fluorescently labeled oligonucleotide primers instead of radioactive ones and (2) fragment analysis with an automated DNA sequencer instead of polyacrylamide gel electrophoresis. AFLP markers generated using IRD primers and visualization of fragments by a gel-based sequencer such as a LI-COR DNA Analyzer produced successful results for plant species with genomes of varying complexities [225–227].
4. A smear product in the 100–1500 base pair range should be visible. Make sure the genomic DNA is fully restricted, so no high-weight DNA molecules are present.
5. Another modified protocol wherein which genomic DNA was digested with 5 units of EcoRI and 5 units of TruI (an isoschizomer of MseI). Selective PCR reaction was done with fluorescently labeled EcoRI+NNN and 1 mM un-labeled MseI+CTT [228].
6. A modified protocol in amaranth [229] involved the analysis of AFLP products in ABI PRISM 310 Genetic analyzer (Applied Biosystems), and GeneScan software program was also used in the analysis. Modification in the analysis of AFLP fragments was introduced for AFLP marker study of the wild species of lettuce crop, *Lactuca aculeata*, resistant against downy mildew pathogen [230]. AFLP analyses were performed using the commercial IRDye® Fluorescent AFLP® Kit designed for large plant genome analysis. The results were visualized using an automated AFLP analysis program (LI-COR SAGAMX v.3.3) [78]. In another modification of the protocol (1), Pst1 and EcoR1 and the 4-bp cutting enzyme Mse1 were used. PCR reactions were set up in Beckman Biomek 2000 liquid handling device. Electrophoresis was carried out on the Bio-Rad Sequi-Gen GT system. A Promega *fmol* DNA Cycle Sequencing System marker (Promega Q4100) was run to estimate the

product size and "control lanes" of standard potato genotypes. Gels were dried onto paper exposed to X-ray film which was then developed using a Konica Minolta film processor (SRX-101A 2006) [59]. A modified protocol was followed for AFLP fingerprinting [231], wherein which the primer combinations with highest polymorphic index were selected to investigate the genetic variability in separate sets of analysis for wild population of two important medicinal plant species. In a modified protocol developed [232], fluorescently labeled AFLP primer combinations were used, and PCR products were separated using capillary electrophoresis.

References

1. Vos P, Hogers R, Bleeker M et al (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23(21):4407–4414
2. Ritland C, Ritland K (2000) DNA-fragment markers in plants. In: Baker AJ (ed) Molecular methods in ecology, 6th edn. Blackwell Science, Oxford, London
3. Savelkoul PH, Aarts HJ, de Haas J et al (1999) Amplified-fragment length polymorphism analysis: the state of an art. J Clin Microbiol 37(10):3083–3091
4. Das S, Rajagopal J, Bhatia S, Srivastava PS et al (1999) Assessment of genetic variation within *Brassica campestris* cultivars using amplified fragment length polymorphism and random amplification of polymorphic DNA markers. J Biosci 24:433–440
5. Lombard V, Baril CP, Dubreuil P et al (2000) Genetic relationship and fingerprinting of rapeseed cultivars by AFLP: consequences for varietal registration. Crop Sci 40:1417–1425
6. El-Esawi MA, Germaine K, Bourke P et al (2016) AFLP analysis of genetic diversity and phylogenetic relationships of *Brassica oleracea* in Ireland. C R Biol 339:163–170
7. Qi X, Stam P, Lindhout P et al (1998) Use of locus-specific AFLP markers to construct a high-density molecular map in barley. Theor Appl Genet 96:376–384
8. Nag A, Ahuja PS, Sharma RK et al (2014) Genetic diversity of high-elevation populations of an endangered medicinal plant. AoB Plants 7
9. Maughan PJ, Saghai MMA, Buss GR (1996) Amplified fragment length polymorphism (AFLP) in soybean: species diversity, inheritance, and near-isogenic line analysis. Theor Appl Genet 93(3):392–401
10. Polanco C, Ruiz ML (2002) AFLP analysis of somaclonal variation in *Arabidopsis thaliana* regenerated plants. Plant Sci 162:817–824
11. Peng M, Zong X, Wang C et al (2015) Genetic diversity of strawberry (*Fragaria ananassa* Duch.) from the Motuo County of the Tibet plateau determined by AFLP markers. Biotechnol Biotechnol Equip 29 (5):876–881
12. Bian F, Pang Y, Wang Z et al (2015) Genetic diversity of the rare plant *Anemone shikokiana* (Makino) Makino (Ranunculaceae) inferred from AFLP markers. Plant Syst Evol 301 (2):677–684
13. Singh SK, Katoch R, Kapila RK et al (2015) Genetic and biochemical diversity among *Valerian jatamansi* populations from Himachal Pradesh. Sci World J 863913:10
14. Krishnamurthy SL, Prashanth Y, Rao AM (2015) Assessment of AFLP marker based genetic diversity in chilli (*Capsicum annuum* L and *C. Baccatum* L.). Indian J Biotech 14:49–54
15. Wu FQ, Shen SK, Zhang XJ et al (2015) Genetic diversity and population structure of an extremely endangered species: the world's largest rhododendron. Ecol Evol 5 (15):3003–3022
16. Li B, Wang A, Zhang P et al (2019) Genetic diversity and population structure of endangered *Glehnia littoralis* (Apiaceae) in China based on AFLP analysis. Biotechnol Biotechnol Equip 33(1):331–337
17. Divakaran M, Babu KN, Ravindran PN et al (2006) Interspecific hybridization in vanilla and molecular characterization of hybrids and selfed progenies using RAPD and AFLP markers. Sci Horti 108(4):414–422
18. Huh MK, Huh HW (2001) AFLP fingerprinting of *Brassica campestris* L. ssp. *napus*

var. *nippo-oleifera* Makino from Korea. Korean J Biol Sci 5:101–106

19. Srivastava A, Gupta V, Pental D et al (2001) AFLP-based genetic diversity assessment amongst agronomically important natural and some newly synthesized lines of *Brassica juncea*. Theor Appl Genet 102:193–199
20. Negi MS, Sabharwal V, Bhat SR (2004) Utility of AFLP markers for the assessment of genetic diversity within *Brassica nigra* germplasm. Plant Breed 123:13–16
21. Warwick SI, James T, Falk KC et al (2008) AFLP-based molecular characterization of *Brassica rapa* and diversity in Canadian spring turnip rape cultivars. Plant Genet Resour 6:11–21
22. Liu RH, Meng JL et al (2006) RFLP and AFLP analysis of inter-and intraspecific variation of *Brassica rapa* and *B. napus* shows that *B. rapa* is an important genetic resource for *B. napus* improvement. Acta Genet Sin 33(9): 814–823
23. Jiang Y, Tian E, Li R et al (2007) Genetic diversity of *Brassica carinata* with emphasis on the interspecific crossability with *B. rapa*. Plant Breed 126:487–491
24. Takuno S, Kawahara T, Ohnishi O et al (2007) Phylogenetic relationships among cultivated types of *Brassica rapa L. em. Metzgas* revealed by AFLP analysis. Genet Resour Crop Evol 54:279–285
25. Zhao J, Wang X, Deng B et al (2005) Genetic relationships within *Brassica rapa* as inferred from AFLP fingerprints. Theor Appl Genet 110(7):1301–1314
26. Sorkheh K, Masaeli M, Chaleshtori MH (2016) AFLP-based analysis of genetic diversity, population structure, and relationships with agronomic traits in rice germplasm from north region of Iran and world core germplasm set. Biochem Genet 54(2): 177–193
27. Pioto F, Costa R, França S et al (2015) Genetic diversity by AFLP analysis within *Jatropha curcas* L. populations in the state of São Paulo, Brazil. Biomass Bioenergy 80:316–320
28. DeHaan LR, Ehlke NJ, Sheaffer CC (2003) Illinois bundle flower genetic diversity determined by AFLP analysis. Crop Sci 43:402–408
29. Mariette S, Chagne D, Lezier C et al (2001) Genetic diversity within and among *Pinus pinaster* populations: comparison between AFLP and microsatellite markers. Heredity 86:469–479
30. Martos V, Royo C, Rharrabti Y et al (2005) Using AFLPs to determine phylogenetic relationships and genetic erosion in durum wheat cultivars released in Italy and Spain throughout the 20th century. Field Crops Res 91:107–116
31. Andrade F, Gonçalves L, Miglioranza E (2016) AFLP analysis of genetic diversity in determinate and indeterminate snap bean accessions. Acta Sci Agron 38:29
32. Opara UL, Jacobson D, Al-Saady NA (2010) Analysis of genetic diversity in banana cultivars (*Musa* cvs.) from the south of Oman using AFLP markers and classification by phylogenetic, hierarchical clustering and principal component analyses. J Zhejiang Univ 11:332–341
33. Wong C, Kiew R, Loj JP et al (2001) Genetic diversity of the wild banana *Musa acuminata* Colla in Malaysia as evidenced by AFLP. Annals Bot 88:1017–1025
34. Ude G, Pillay M, Ogundiwin E et al (2003) Genetic diversity in an African plantain core collection using AFLP and RAPD markers. Theor Appl Genet 107:248–255
35. Wang XL, Chiang T, Roux N et al (2007) Genetic diversity of wild banana (*Musa balbisiana* Colla) in China as revealed by AFLP markers. Genet Res Crop Evol 54:1125–1132
36. Ahmad F, Megia R, Poerba Y et al (2014) Genetic diversity of *Musa balbisiana* Colla in Indonesia based on AFLP marker. HAYATI J Biosci 21:39–47
37. Zawko G, Krauss SL, Dixon KW et al (2001) Conservation genetics of the rare and endangered *Leucopogon obtectus* (Ericaceae). Mol Ecol 10:2389–2396
38. Van Ee BW, Jelinski N, Berry PE et al (2006) Phylogeny and biogeography of *Croton alabamensis* (Euphorbiaceae), a rare shrub from Texas and Alabama, using DNA sequence and AFLP data. Mol Ecol 15:2735–2751
39. Ronikier M (2002) The use of AFLP markers in conservation genetics-a case study on *Pulsatilla vernalis* in the polish lowlands. Cell Mol Biol Lett 7:677–684
40. Li X, Ding X, Chu B et al (2008) Genetic diversity analysis and conservation of the endangered Chinese endemic herb *Dendrobium officinale* Kimura et Migo (Orchidaceae) based on AFLP. Genetica 133:159–166
41. Travis SE, Maschinski J, Keim P et al (1996) An analysis of genetic variation in *Astragalus cremnophylax var cremnophylax*, a critically endangered plant, using AFLP markers. Mol Ecol 5:735–745

42. Tatikonda L, Wani SP, Kannan S et al (2009) AFLP-based molecular characterization of an elite germplasm collection of *Jatropha curcas* L., a biofuel plant. Plant Sci 176:505–513
43. Elameen A, Klemsdal SS, Dragland S (2008) Genetic diversity in a germplasm collection of Roseroot (*Rhodiola rosea*) in Norway studied by AFLP. Biochem Syst Ecol 36:706–715
44. Huh MK, Ohnishi O (2002) Genetic diversity and genetic relationships of east Asian natural populations of wild radish revealed by AFLP. Breeding Sci 52:79–88
45. Bhattacharyya P, Ghosh S, Mandi SS et al (2017) Genetic variability and association of AFLP markers with some important biochemical traits in *Dendrobium thyrsiflorum*, a threatened medicinal orchid. S Afr J Bot 109:214–222
46. Vaishnav V, Wali SA, Tripathi SB et al (2018) Preliminary investigation on AFLP marker-wood density trait association in teak (*Tectona grandis* L. f.). Ann For Res 61(1):1–15
47. Sharma SK, Knox MR, Ellis THN (1996) AFLP analysis of the diversity and phylogeny of lens and its comparison with RAPD analysis. Theor Appl Genet 93:751–758
48. Mitchell ML, Stodart BJ, Virgona JM et al (2015) Genetic diversity within a population of *Microlaena stipoides*, as revealed by AFLP markers. Aust J Bot 62:580–586
49. Olet EA, Lye KA, Heun M et al (2011) Amplified fragment length polymorphisms (AFLPs) analysis of species of solanum section solanum (Solanaceae) from Uganda. Afr Biotech 10:6387–6395
50. Kardolus JP, Van Eck HJ, Van den Berg RG (1998) The potential of AFLPs in biosystematics: a first application in solanum taxonomy (Solanaceae). Plant Syst Evol 210:87–103
51. Lara-Cabrera SI, Spooner DM (2004) Taxonomy of north and central American diploid wild potato (solanum sect. Petota) species: AFLP data. Plant Syst Evol 248:129–142
52. Mace ES, Lester RN, Gebhardt CG (1999) AFLP analysis of genetic relationships among the cultivated eggplant, *Solanum melongena* L., and wild relatives (Solanaceae). Theor Appl Genet 99:626–633
53. Nuez F, Prohens J, Blanca JM (2004) Relationships, origin, and diversity of Galapagos tomatoes: implications for the conservation of natural populations. Am J Bot 91:86–99
54. Olet EA, Heun M, Lye KA (2005) African crop or poisonous nightshade; the enigma of poisonous or edible black nightshade solved. Afr J Ecol 43:158–161
55. Manoko ML, Van den Berg RG, Feron RM et al (2008) Genetic diversity of the African hexaploid species *Solanum scabrum* mill. And *Solanum nigrum* L. (Solanaceae). Genet Resour Crop Evol 55:409–418
56. Spooner DM, McLean K, Ramsay G et al (2005) A single domestication for potato based on multilocus amplified fragment length polymorphism genotyping. PNAS 102:14694–14699
57. Spooner DM, Peralta IE, Knapp S (2005) Comparison of AFLPs with other markers for phylogenetic inference in wild tomatoes solanum L. section *Lycopersicon* (mill.) Wettst. Taxon 54:43–61
58. Schnell RJ, Olano CT, Campbell RJ et al (2002) AFLP analysis of genetic diversity within a jackfruit germplasm collection. Euphytica 125(1):89–102
59. Bryan GJ, McLean K, Waugh R et al (2017) Levels of intra-specific AFLP diversity in tuber-bearing potato species with different breeding systems and ploidy levels. Front Genet 8:119
60. Smith JSC, Smith OS (1992) Fingerprinting crop varieties. Adv Agron 47:85–140
61. Paul S, Wachira FN, Powell W (1997) Diversity and genetic differentiation among populations of Indian and Kenyan tea (*Camellia sinensis* (L.) O. Kuntze) revealed by AFLP markers. Theor Appl Genet 94:255–263
62. Murtaza N (2006) Cotton genetic diversity study by AFLP markers. Electron J Biotechnol 9
63. Shaheen N, Pearce SR, Khan MA et al (2010) AFLP mediated genetic diversity of malvaceae species. J Med Plant Res 4(2):148–154
64. Rattanathawornkiti K, Kanchanaketu T, Suwanagul A et al (2016) Genetic relationship assessment of pineapple germplasm in Thailand revealed by AFLP markers. Genomics Genet 9(2):1–10
65. Hill M, Witsenboer H, Zabeau M (1996) PCR-based fingerprinting using AFLPs as a tool for studying genetic relationships in *Lactuca* spp. Theor Appl Genet 93:1202–1210
66. Vieira EA, Carvalho FIF, Oliveira AC et al (2007) Path analysis among primary and secondary yield components in wheat. Rev Bras Fisioter 13(2):169–174
67. Ipek M, Seker M, Ipek A et al (2015) Identification of molecular markers associated with fruit traits in olive and assessment of olive core collection with AFLP markers and fruit traits. Genet Mol Res 14:2762–2774
68. Zhou H, Liao J, Xia YP et al (2013) Determination of genetic relationships between

evergreen azalea cultivars in China using AFLP markers. J Zhejiang Univ Sci B 14:299–308

69. Xu RQ, Tomooka N, Vaughan DA (2000) AFLP markers for characterizing the azuki bean complex. Crop Sci 40(3):808–815
70. Stodart BJ, Mackay M, Raman H (2005) AFLP and SSR analysis of genetic diversity among landraces of bread wheat (*Triticum aestivum* L. em. Thell) from different geographic regions. Aust J Agric Res 56:691–697
71. Al-Saady NA, Al-Lawati AH, Al-Subhi AM et al (2010) Evaluation of genetic diversity in Omani banana cultivars (Musa cvs.) using AFLP markers. J Plant Sci 5(4):402–413
72. Tang T, Zhong Y, Jian S et al (2003) Genetic diversity of *Hibiscus tiliaceus* (Malvaceae) in China assessed using AFLP markers. Ann Bot 92:409–414
73. Esfahani ST, Shiran B, Balali G (2009) AFLP markers for the assessment of genetic diversity in European and North American potato varieties cultivated in Iran. Crop Breed Appl Biotech:9
74. Lebeda A, Doležalová I, Křístková E et al (2009) Wild lactuca germplasm for lettuce breeding: current status, gaps and challenges. Euphytica 170:15–34
75. Coulibaly S, Pasquet RS, Papa R et al (2002) AFLP analysis of the phenetic organization and genetic diversity of *Vigna unguiculata* L. Walp. reveals extensive gene flow between wild and domesticated types. Theor Appl Genet 104(2-3):358–366
76. Keivani M, Ramezanpour SS, Soltanloo H et al (2010) Genetic diversity assessment of alfalfa (*Medicago sativa* L.) populations using AFLP markers. Aust J Crop Sci 4:491–497
77. Zargar M, Romanova E, Trifonova A et al (2017) AFLP-analysis of genetic diversity in soybean (Glycine max L. Merr.) cultivars of Russian and foreign selection. Agron Res 15:2217–2225
78. Oliveira TG, Pereira AMS, Coppede JS et al (2016) Genetic diversity analysis of *Croton antisyphiliticus* Mart. Using AFLP molecular markers. Genet Mol Res 15(1):1–8
79. Goyat S, Grewal A, Singh D (2019) Sex-linked AFLP marker identification in dioecious Betelvine (*Piper betle* L.). J Horti Sci Biotech 94(4):422–427
80. Christensen S, Von Bothmer R, Poulsen G (2011) AFLP analysis of genetic diversity in leafy kale (*Brassica oleracea* L. convar. Acephala (DC.) Alef.) land races, cultivars and wild populations in Europe. Genet Resour Crop Evol 58:657–666
81. Teyer FS, Salazar MS, Esqueda M et al (2009) Genetic variability of wild *Agave angustifolia* populations based on AFLP: a basic study for conservation. J Arid Environ 73:611–616
82. Sivaprakash KR, Prashanth SR, Mohanty BP et al (2004) Genetic diversity of black gram (*Vigna mungo*) landraces as evaluated by amplified fragment length polymorphism markers. Curr Sci 86:1411–1416
83. Koopman WJM, Zevenbergen MJ, Van den Berg RG et al (2001) Species relationships in Lactuca sp (Lactuceae, Asteraceae) inferred form AFLP fingerprints. Am J Bot 88:1881–1887
84. Guo YP, Sauke J, Mittermayr R et al (2005) AFLP analyses demonstrate genetic divergence, hybridization, and multiple polyploidization in the evolution of Achillea (Asteraceae-anthemideae). New Phytol 166:273–290
85. Tiwari JK, Chandel P, Gupta S (2013) Analysis of genetic stability of in vitro propagated potato micro tubers using DNA markers. Physiol Mol Biol Plants 19(4):587–595
86. Smith NR, Trigiano RN, Windham MT et al (2007) AFLP markers identify *Cornus florida* cultivars and lines. J Am Soc Hortic Sci 132:90–96
87. Duarte-Delgado D, Chacón MI, Núñez V et al (2011) Preliminary assessment of AFLP fingerprinting of *Rubus glaucus* Benth. Elite genotypes. Agron Colomb 29:7–16
88. Mignouna HD, Abang MM, Fagbemi SA (2003) A comparative assessment of molecular marker assays (AFLP, RAPD and SSR) for white yam (*Dioscorea rotundata* Poir) germplasm characterisation. Ann Appl Biol 142:269–276
89. Krauss SL (1999) Complete exclusion of nonsires in an analysis of paternity in a natural plant population using amplified fragment length polymorphism (AFLP). Mol Ecol 8:217–226
90. Sensi E, Vignani R, Rohde W et al (1996) Characterization of genetic biodiversity with *Vitis vinifera* L. Sangiovese and Colorino genotypes by AFLP and ISTR DNA marker technology. Vitis 35:183–188
91. Juárez AMJ, Ramírez-Malagón R, Gil-Vega KDC et al (2009) AFLP analysis of genetic variability in three reproductive forms of *Agave tequilana*. Rev Fitotecnia Mexicana 32:171–175
92. Mueller UG, Wolfenbarger LL (1999) AFLP genotyping and fingerprinting. Trees 10:389–394
93. Alizadeh M, Krishna H, Eftekhari M et al (2015) Assessment of clonal fidelity in micro

propagated horticultural plants. J Chem Pharm Res 7(12):977–990

94. Chittora M, Sharma D, Veer C (2015) Molecular markers: an important tool to assess genetic fidelity in tissue culture grown long-term cultures of economically important fruit plants. Asian J Bio Sci 10(1):101–105
95. Singh SR, Dalal S, Singh R, Dhawan AK et al (2013) Ascertaining clonal fidelity of micro propagated plants of Dendrocalamus hamiltonii Nees et Arn. Ex Munro using molecular markers. In Vitro Cell Dev Plant 49 (5):572–583
96. Mehta R, Sharma V, Sood A et al (2011) Induction of somatic embryogenesis and analysis of genetic fidelity of in vitro-derived plantlets of *Bambusa nutans* wall, using AFLP markers. Eur J For Res 130:1–10
97. Castillo N, Bassil N, Wada S et al (2010) Genetic stability of cryopreserved shoot tips of Rubus germplasm. In Vitro Cell Dev Biol Plant 46:246–256
98. Mignouna D, Mank R, Ellis T et al (2002) A genetic linkage map of Guinea yam (*Dioscorea rotundata* Poir.) based on AFLP markers. Theor Appl Genet 105:716–725
99. Terashima K, Matsumoto T, Hayashi E et al (2002) A genetic linkage map of *Lentinula edodes* (shiitake) based on AFLP markers. Mycol Res 106:911–917
100. Rouppe Van der Voort J, Wolters JP, Folkertsma R et al (1997) Mapping of the cyst nematode resistance locus Gpa2 in potato using a strategy based on co-migrating AFLP markers. Theor Appl Genet 95:874–880
101. Quarrie S, Laurie D, Zhu J (1997) QTL analysis to study the association between leaf size and abscisic acid accumulation in droughted rice leaves and comparisons across cereals. Plant Mol Biol 35:155–165
102. Voorrips RE, Jongerius MC, Kanne HJ (1997) Mapping of two genes for resistance to club root (*Plasmodiophora brassicae*) in a population of doubled haploid lines of Brassica oleracea by means of RFLP and AFLP markers. Theor Appl Gen 94:75–82
103. Jin H, Domier L, Kolb F et al (1998) Identification of quantitative loci for tolerance to barley yellow dwarf virus in oat. Phytopathology 88:410–415
104. Jin H, Domier L, Shen X (2000) Combined AFLP and RFLP mapping in two hexaploid oat recombinant inbred populations. Genome 43:94–101
105. Cho YG, McCouch SR, Kuiper M et al (1998) Integrated map of AFLP, SSLP and RFLP markers using a recombinant inbred population of rice (*Oryza sativa* L.). Theor Appl Genet 97:370–380
106. Becker J, Vos P, Kuiper M (1995) Combined mapping of AFLP and RFLP markers in barley. Mol Gen Genet 249:65–73
107. Peters J, Cnops G, Neyt P (2004) An AFLP-based genome-wide mapping strategy. Theor Appl Genet 108:321–327
108. Mackill D, Zhang Z, Redona E et al (1996) Level of polymorphism and genetic mapping of AFLP markers in rice. Genome 39:969–977
109. Ballvora A, Hesselbach J, Niewhner J et al (1995) Marker enrichment and high-resolution map of the segment of potato chromosome VII harbouring the nematode resistance gene Gro1. Mol Gen Genet 249:82–90
110. Brigneti G, Garcia-Mas J, Baulcombe DC (1997) Molecular mapping of the potato virus Y resistance gene Rysto in potato. Theor Appl Genet 94:198–203
111. Meksem K, Leister D, Peleman J et al (1995) A high resolution map of the vicinity of the R1 locus on chromosome V of potato based on RFLP and AFLP markers. Mol Gen Genet 249:74–81
112. van Eck HJ, van der Voort JR, Draaistra J et al (1995) The inheritance and chromosomal localization of AFLP markers in a non-inbred potato offspring. Mol Breed 1:397–410
113. Thomas CM, Vos P, Zabeau M (1995) Identification of amplified restriction fragment polymorphism (AFLP) markers tightly linked to the tomato Cf-9 gene for resistance to *Cladosporium fulvum*. Plant J 8:785–794
114. Cnops G, Denboer B, Gerats A (1996) Chromosome landing at the Arabidopsis TORNADO1 locus using an AFLP-based strategy. Mol Gen Genet 253:32–41
115. Jeuken M, van Wijk R, Peleman J et al (2001) An integrated interspecific AFLP map of lettuce (Lactuca) based on two *L. sativa* × *L. saligna* F2 populations. Theor Appl Genet 103:638–647
116. Jeuken M, Lindhout P (2002) *Lactuca saligna*, a non-host for lettuce downy mildew (Bremia lactucae), harbors a new race-specific Dm gene and three QTLs for resistance. Theor Appl Genet 105:384–391
117. Johnson WC, Jackson LE, Ochoa O et al (2000) Lettuce, a shallow-rooted crop, and *Lactuca serriola*, its wild progenitor, differ at QTL determining root architecture and deep soil water exploitation. Theor Appl Genet 101:1066–1073

118. Cervera MT, Gusmao J, Steenackers M et al (1996) Identification of AFLP molecular markers for resistance against *Melampsora larici-populina* in Populus. Theor Appl Genet 93:733–737
119. Mukeshimana G, Paneda A, Rodríguez-Suárez C (2005) Markers linked to the bc-3 gene conditioning resistance to bean common mosaic potyviruses in common bean. Euphytica 144:291–299
120. Qi X, Lindhout P et al (1997) Development of AFLP markers in barley. Mol Gen Genet 254:330–336
121. Castiglioni P, Pozzi C, Heun M et al (1998) An AFLP-based procedure for the efficient mapping of mutations and DNA probes in barley. Genetics 149:2039–2056
122. Keim P, Schupp JM, Travis SE et al (1997) A high-density soybean genetic map based on AFLP markers. Crop Sci 37:537–543
123. Hazen SP, Leroy P, Ward RW (2002) AFLP in *Triticum aestivum* L. patterns of genetic diversity and genome distribution. Euphytica 125:89–102
124. Li G, Liu Y, Ehlers JD et al (2007) Identification of an AFLP fragment linked to rust resistance in asparagus bean and its conversion to a SCAR marker. Hort Sci 42:1153–1156
125. Wang Y, Bi B, Yuan QH et al (2012) Association of AFLP and SCAR markers with common leaf spot resistance in auto tetraploid alfalfa (*Medicago sativa*). Genet Mol Res 11:606–616
126. Miao L, Shou S, Cai J (2009) Identification of two AFLP markers linked to bacterial wilt resistance in tomato and conversion to SCAR markers. Mol Biol Rep 36:479–486
127. Liao Y, Sun B, Sun G et al (2009) AFLP and SCAR markers associated with peel color in eggplant. Sci Agric Sin 42:3996–4003
128. Peng SF, Lin YP, Lin BY (2005) Characterization of AFLP sequences from regions of maize B chromosome defined by 12 B-10L translocations. Genetics 169:375–388
129. Altinkut A, Kazan K, Gozukirmizi N et al (2003) AFLP marker linked to water-stress-tolerant bulks in barley (*Hordeum vulgare* L.). Genet Mol Biol 26:77–82
130. Zhang Y, Guo L, Shu Z et al (2013) Identification of amplified fragment length polymorphism (AFLP) markers tightly associated with drought stress gene in male sterile and fertile *Salvia miltiorrhiza* Bunge. Int J Mol Sci 14:6518–6528
131. Wei P, Feng H, Piao Z et al (2009) Identification of AFLP markers linked to Ms, a genic multiple allele inherited male-sterile gene in Chinese cabbage. Breed Sci 59(4):333–339
132. Balta H, Karakas MO, Sentürk AF et al (2014) Identification of an AFLP marker linked with yellow rust resistance in wheat (*Triticum aestivum* L.). Turk J Biol 38:371–379
133. Moon H (2006) Identification of AFLP markers linked to tomato spotted wilt virus resistance in tobacco. Dissertation, North Carolina State University
134. Ghosh S, Majumder PB, Mandi SS et al (2011) Species-specific AFLP markers for identification of *Zingiber officinale*, *Z. montanumand* and *Z. zerumbet* (Zingiberaceae). Genet Mol Res 10:218–229
135. Yang AH, Wei N, Fritsch PW et al (2016) AFLP genome scanning reveals divergent selection in natural populations of *Liriodendron chinense* (Magnoliaceae) along a latitudinal transect. Front Plant Sci 7:698
136. Zhang C, Sun M, Zhang X et al (2018) AFLP-based genetic diversity of wild orchard grass germplasm collections from Central Asia and Western China, and the relation to environmental factors. PLoS One 13:0195273
137. Jemelkova M, Kitner M, Křístková E et al (2018) Genetic variability and distance between *Lactuca serriola* L. populations from Sweden and Slovenia assessed by SSR and AFLP markers. Acta Bot Croatica 77:172–180
138. Kuang H, van Eck HJ, Sicard D (2008) Evolution and genetic population structure of prickly lettuce (*Lactuca serriola*) and its RGC2 resistance gene cluster. Genetics 178 (3):1547–1558
139. Cao Y, Zhang Q, Chen Y et al (2013) Identification of differential expression genes in leaves of rice (Oryza sativa L.) in response to heat stress by cDNA-AFLP analysis. Bio med res Int. 576189
140. Medeiros CN, Gonçalves MC, Harakava R et al (2014) Sugarcane transcript profiling assessed by cDNA-AFLP analysis during the interaction with sugarcane mosaic virus. Adv Microbiol 4:511
141. Xie L, Wang X, Peng M et al (2014) Isolation and detection of differential genes in hot pepper (*Capsicum annuum* L.) after space flight using AFLP markers. Biochem Syst Ecol 57:27–32
142. Santos CAF, Gama RNC (2013) An AFLP estimation of the outcrossing rate of *Spondias tuberosa* (Anacardiaceae), an endemic species to the Brazilian semiarid region. Rev Biol Trop 61:577–582

143. Azizi A, Ardalani H, Honermeier B et al (2016) Statistical analysis of the associations between phenolic monoterpenes and molecular markers, AFLPs and SAMPLs in the spice plant oregano. Herba Pol 62:42–56
144. Chandi A, Milla-Lewis S, Jordan D et al (2013) Use of AFLP markers to assess genetic diversity in palmer amaranth (*Amaranthus palmeri*) populations from North Carolina and Georgia. Weed Sci 61(1):136–145
145. Paun O, Schonswetter P (2012) Amplified fragment length polymorphism (AFLP)–an invaluable finger printing technique for genomic, transcriptomic and epigenetic studies. Methods Mol Biol 862:75–87
146. Haghpanah M, Kazemitabar SK, Hashemi SH et al (2016) Comparison of ISSR and AFLP markers in assessing genetic diversity among nettle (*Urtica dioica* L.) populations. J Plant Mol Breed 4:10–16
147. Sarwat M, Das S, Srivastava PS (2008) Analysis of genetic diversity through AFLP, SAMPL, ISSR and RAPD markers in *Tribulus terrestris*, a medicinal herb. Plant Cell Rep 27:519–528
148. Ci X-q, Jun-qiu C, Qiao-ming L et al (2008) AFLP and ISSR analysis reveals high genetic variation and inter-population differentiation in fragmented populations of the endangered *Litsea szemaois* (Lauraceae) from south-West China. Plant Syst Evol 273:237–246
149. Roy JK, Lakshmikumaran MS, Balyan HS et al (2004) AFLP-based genetic diversity and its comparison with diversity based on SSR, SAMPL, and phenotypic traits in bread wheat. Biochem Genet 42:43–59
150. Garcia AAF, Benchimol LL, Barbosa AMM et al (2004) Comparison of RAPD, RFLP, AFLP and SSR markers for diversity studies in tropical maize inbred lines. Genet Mol Biol 27(4):579–588
151. Abdelhamid S, Le CL, Conedera M et al (2014) The assessment of genetic diversity of Castanea species by RAPD, AFLP, ISSR, and SSR markers. Turk J Botany 38:835–850
152. Minoo D, Babu KN, Ravindran PN et al (2006) Inter specific hybridization in vanilla and molecular characterization of hybrids and selfed progenies using RAPD and AFLP markers. Sci Horti 108:414–422
153. Powell W, Morgante M, Andre C et al (1996) The comparison of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol Breed 2:225–238
154. Li L, Wanapu C, Huang X et al (2011) Comparison of AFLP and SSR for genetic diversity analysis of *Brassica napus* hybrids. J Agri Sci 3 (3):101–110
155. Maras M, Sustar-Vozlic J, Javornik B et al (2008) The efficiency of AFLP and SSR markers in genetic diversity estimation and gene pool classification of common bean (*Phaseolus vulgaris* L.). Acta Agric Slov 91:87–96
156. Costa R, Pereira G, Garrido I et al (2016) Comparison of RAPD, ISSR, and AFLP molecular markers to reveal and classify orchard grass (*Dactylis glomerata* L.) germplasm variations. PLoS One 11(4):e0152972
157. Koundal M, Sharma DR, Mohapatra T et al (2006) Comparative evaluation of RAPD and AFLP based genetic diversity in Brinjal (*Solanum melongena*). J Plant Biochem Biotechnol 15(1):15–19
158. Pamidimarri DS, Singh S, Mastan SG et al (2009) Molecular characterization and identification of markers for toxic and non-toxic varieties of *Jatropha curcas* L. using RAPD, AFLP and SSR markers. Mol Biol Rep 36:1357–1364
159. Avendaño R, José S, Rica C et al (2015) Genetic diversity analysis of Jatropha species from Costa Rica using AFLP markers. Am J Plant Sci 6:2426
160. Poeaim S, Chaiyabut A, Poeaim A et al (2017) Genetic diversity and relationships among sugarcane (Saccharum sp.) from Thailand revealed by RAPD and AFLP markers. Indian J Sci Technol 10(28):1–9
161. Qin Y, Kabir MA, Wang HW et al (2013) Assessment of genetic diversity and relationships based on RAPD and AFLP analyses in Miscanthus genera landraces. Can J Plant Sci 93:171–182
162. Monte JV, De Nova PJ, Soler C et al (2001) AFLP-based analysis to study genetic variability and relationships in the Spanish species of the genus Aegilops. Hereditas 135(2-3):233–238
163. Dessalegn Y, Liezel H, Maryke L et al (2009) Comparison of SSR and AFLP analysis for genetic diversity assessment of Ethiopian arabica coffee genotypes. S Afr J Plant Soil 26:119–125
164. Gaudeul M, Till-Bottraud I, Barjon F et al (2004) Genetic diversity and differentiation in *Eryngium alpinum* L. (Apiaceae): comparison of AFLP and microsatellite markers. Heredity 92(6):508–518
165. Beyene Y, Botha AM, Myburg AA et al (2005) A comparative study of molecular and morphological methods of describing genetic relationships in traditional Ethiopian highland maize. Afr J Biotechnol 4:586–595

166. Singh A, Negi MS, Moses VK et al (2002) Molecular analysis of micropropagated neem plants using AFLP markers for ascertaining clonal fidelity. In Vitro Cell Dev Biol Plant 38:519–524
167. Youssef M, James AC, Rivera-Madrid R et al (2011) Musa genetic diversity revealed by SRAP and AFLP. Mol Biotechnol 47:189–199
168. Labra M, Miele M, Ledda B et al (2004) Morphological characterization, essential oil composition and DNA genotyping of *Ocimum basilicum* L. cultivars. Plant Sci 167:725–731
169. Milbourne D, Meyer R, Bradshaw JE et al (1997) Comparison of PCR-based marker systems for the analysis of genetic relationships in cultivated potato. Mol Breed 3:127–136
170. Lambertini C, Frydenberg J, Gustafsson MHG et al (2008) Herbarium specimens as a source of DNA for AFLP fingerprinting of Phragmites (Poaceae): possibilities and limitations. Plant Syst Evol 272:223–231
171. Morgante M, Vogel J (1994) Compound microsatellite primers for the detection of genetic polymorphisms. U.S. patent, 08/326456, 1994
172. Witsenboer H, Michelmore RW, Vogel J et al (1997) Identification, genetic localization, and allelic diversity of selectively amplified microsatellite polymorphic loci in lettuce and wild relatives (*Lactuca* spp.). Genome 40:923–936
173. Tseng YT, Lo HF, Hwang SY (2002) Genotyping and assessment of genetic relationships in elite polycross breeding cultivars of sweet potato in Taiwan based on SAMPL polymorphisms. Bot Bull Acad Sinica 43
174. Negi MS, Sabharwal V, Wilson N et al (2006) Comparative analysis of the efficiency of SAMPL and AFLP in assessing genetic relationships among *Withania somnifera* genotypes. Curr Sci 91:464–471
175. Masiga DK, Turner CMR (2004) Amplified (restriction) fragment length polymorphism (AFLP) analysis. In: Parasite Genomics Protocols. Humana Press, New York
176. Cretazzo E, Meneghetti S, De Andrés MT et al (2010) Clone differentiation and varietal identification by means of SSR, AFLP, SAMPL and M-AFLP in order to assess the clonal selection of grapevine: the case study of Manto negro, Callet and moll, autochthonous cultivars of Majorca. Ann Appl Biol 157:213–227
177. Albertini E, Porceddu A, Marconi G et al (2003) Microsatellite-AFLP for genetic mapping of complex polyploids. Genome 46:824–832
178. Whankaew S, Sraphet S, Thaikert R et al (2012) Characterization of microsatellite markers in cassava based on microsatellite-AFLP technique. Genet Mol Res 11:1319–1326
179. Ellis THN, Poyser SJ, Knox MR et al (1998) Polymorphism of insertion sites of Ty1-copia class retro transposons and its use for linkage and diversity analysis in pea. Mol Gen Genet 260:9–19
180. Yang H, Shankar M, Buirchell B et al (2002) Development of molecular markers using MFLP linked to a gene conferring resistance to *Diaporthe toxica* in narrow-leafed lupin (*Lupinus angustifolius* L.). Theor Appl Genet 105:265–270
181. Waugh R, McLean K, Flavell AJ et al (1997) Genetic distribution of Bare–1-like retro transposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Mol Gen Genet 253:687–694
182. Queen RA, Gribbon BM, James C et al (2004) Retro transposon-based molecular markers for linkage and genetic diversity analysis in wheat. Mol Gen Genomics 271:91–97
183. Leigh F, Kalendar R, Lea V et al (2003) Comparison of the utility of barley retro transposon families for genetic analysis by molecular marker techniques. Mol Gen Genomics 269:464–474
184. Yu GX, Wise RP (2000) An anchored AFLP- and retro transposon-based map of diploid Avena. Genome 43:736–749
185. Nagy ED, Molnar I, Schneider A et al (2006) Characterization of chromosome-specific S-SAP markers and their use in studying genetic diversity in Aegilops species. Genome 49:289–296
186. Venturi S, Dondini L, Donini P et al (2006) Retro transposon characterisation and fingerprinting of apple clones by S-SAP markers. Theor Appl Genet 112:440–444
187. Lanteri S, Acquadro A, Comino C et al (2006) A first linkage map of globe artichoke (*Cynara cardunculus* var. scolymus L.) based on AFLP, S-SAP, M-AFLP and microsatellite markers. Theor Appl Genet 112:1532–1542
188. Syed NH, Sørensen AP, Antonise R et al (2006) A detailed linkage map of lettuce based on SSAP, AFLP and NBS markers. Theor Appl Genet 112:517–527
189. Jing R, Knox MR, Lee JM et al (2005) Insertional polymorphism and antiquity of PDR1 retro transposon insertions in *Pisum* species. Genetics 171:741–752

190. Tam SM, Mhiri C, Vogelaar A et al (2005) Comparative analyses of genetic diversities within tomato and pepper collections detected by retrotransposon-based SSAP, AFLP and SSR. Theor Appl Genet 110:819–831
191. Tahara M, Aoki T, Suzuka S et al (2004) Isolation of an active element from a high-copy-number family of retro transposons in the sweet potato genome. Mol Gen Genomics 272:116–127
192. Petit M, Lim KY, Julio E et al (2007) Differential impact of retro transposon populations on the genome of allotetraploid tobacco (*Nicotiana tabacum*). Mol Gen Genomics 278:1–15
193. Sanz AM, Gonzalez SG, Syed NH et al (2007) Genetic diversity analysis in Vicia species using retro transposon-based SSAP markers. Mol Gen Genomics 278:433–441
194. Gao L, McCarthy EM, Ganko EW et al (2004) Evolutionary history of *Oryza sativa* LTR retro transposons: a preliminary survey of the rice genome sequences. BMC Genomics 5:18
195. García-Martínez J, Martínez-Izquierdo JA (2003) Study on the evolution of the Grande retro transposon in the Zea genus. Mol Biol Evol 20:831–841
196. Frey M, Stettner C, Gierl A (1998) A general method for gene isolation in tagging approaches: amplification of insertion mutagenised sites (AIMS). Plant J 13:17–721
197. Edwards D, Coghill J, Batley J et al (2002) Amplification and detection of transposon insertion flanking sequences using fluorescent mu AFLP. BioTechniques 32:1090–1097
198. Reyna-Lopez GE, Simpson J, Ruiz-Herrera J (1997) Differences in DNA methylation patterns are detectable during the dimorphic transition of fungi by amplification of restriction polymorphisms. Mol Gen Genet 253:703–710
199. Xiong LZ, Xu CG, Maroof MS et al (1999) Patterns of cytosine methylation in an elite rice hybrid and its parental lines, detected by a methylation-sensitive amplification polymorphism technique. Mol Gen Genet 261:439–446
200. Peraza-Echeverria S, Herrera-Valencia VA, Kay AJ (2001) Detection of DNA methylation changes in micro propagated banana plants using methylation-sensitive amplification polymorphism (MSAP). Plant Sci 161:359–367
201. Chakrabarty D, Yu KW, Paek KY (2003) Detection of DNA methylation changes during somatic embryogenesis of Siberian ginseng (*Eleuterococcus senticosus*). Plant Sci 165:61–68
202. Portis E, Acquadro A, Comino C et al (2004) Analysis of DNA methylation during germination of pepper (*Capsicum annuum* L.) seeds using methylation-sensitive amplification polymorphism (MSAP). Plant Sci 166:169–178
203. Filek M, Janiak A, Szarejko I et al (2006) Does DNA methylation pattern mark generative development in winter rape? Zeitschrift für Naturforschung C 61:387–396
204. Salmon A, Clotault J, Jenczewski E et al (2008) *Brassica oleracea* displays a high level of DNA methylation polymorphism. Plant Sci 174:61–70
205. Tan MP (2010) Analysis of DNA methylation of maize in response to osmotic and salt stress based on methylation-sensitive amplified polymorphism. Plant Physiol Biochem 48:21–26
206. Li A, Hu BQ, Xue ZY et al (2011) DNA methylation in genomes of several annual herbaceous and woody perennial plants of varying ploidy as detected by MSAP. Plant Mol Biol Rep 29:784–793
207. Guzy-Wrobelska J, Filek M, Kaliciak A et al (2013) Vernalization and photoperiod-related changes in the DNA methylation state in winter and spring rapeseed. Acta Physiol Plant 35:817–827
208. Marconi G, Pace R, Traini A et al (2013) Use of MSAP markers to analyse the effects of salt stress on DNA methylation in rapeseed (*Brassica napus var. oleifera*). PLoS One 8 (9):75597
209. Tang XM, Tao X, Wang Y et al (2014) Analysis of DNA methylation of perennial rye grass under drought using the methylation-sensitive amplification polymorphism (MSAP) technique. Mol Gen Genomics 289:1075–1084
210. Li Z, Liu Z, Chen R et al (2015) DNA damage and genetic methylation changes caused by cd in Arabidopsis thaliana seedlings. Environ Toxicol Chem 34:2095–2103
211. Gimenez MD, Yañez-Santos AM, Paz RC et al (2016) Assessment of genetic and epigenetic changes in virus-free garlic (*Allium sativum* L.) plants obtained by meristem culture followed by in vitro propagation. Plant Cell Rep 35(1):129–141
212. Gautam M, Dang Y, Ge X et al (2016) Genetic and epigenetic changes in oilseed rape (*Brassica napus* L.) extracted from inter generic allopolyploid and additions with Orychophragmus. Front Plant Sci 7:–438

213. Wang B, Liu L, Zhang D et al (2016) Genetic map between Gossypium hirsutum and the Brazilian endemic *G. mustelinum* and its application to QTL mapping. G3 (Bethesda) 6(6):1673–1685
214. Abid G, Kamel H, Marwa A et al (2017) Agro-physiological and biochemical responses of faba bean (*Vicia faba* L. var. 'minor') genotypes to water deficit stress. Biotech Agron Soc Environ 21
215. Chwialkowska K, Nowakowska U, Mroziewicz A et al (2016) Water-deficiency conditions differently modulate the methylome of roots and leaves in barley (*Hordeum vulgare* L.). J Exp Bot 67:1109–1121
216. Hayes A, Saghai MM (2000) Targeted resistance gene mapping in soybean using modified AFLPs. Theor Appl Genet 100:1279
217. Sharma SS, Aadil K, Negi MS et al (2014) Efficacy of two dominant marker systems, ISSR and TE-AFLP for assessment of genetic diversity in biodiesel species *Pongamia pinnata*. Curr Sci 106:1576–1580
218. Knox MR, Ellis THN (2001) Stability and inheritance of methylation states at PstI sites in Pisum. Mol Gen Genet 265:497–507
219. Wessler SR, Bureau TE, White SE et al (1995) LTR-retro transposons and MITEs: important players in the evolution of plant genomes. Curr Opin Genet Dev 5:814–821
220. Casa AM, Brouwer C, Nagel A et al (2000) The MITE family heartbreaker (Hbr) molecular markers in maize. PNAS 97:10083–10089
221. Bachem CW, Van Der Hoeven RS, De Bruijn SM et al (1996) Visualization of differential gene expression using a novel method of RNA fingerprinting based on AFLP: analysis of gene expression during potato tuber development. Plant J 9:745–753
222. Money T, Reader S, Qu LJ et al (1996) AFLP based mRNA fingerprinting. Nucleic Acids Res 24:2616–2617
223. Razavi K, Mohsenzadeh S, Malboobi M et al (2014) The application of a non-radioactive DD-AFLP method for profiling of Aeluropus lagopoides differentially expressed transcripts under salinity or drought conditions. Iranian J Biotech 12(4):47–57
224. Van Eijk MJT, Preben A, Marco S et al (2018) Method for high-throughput AFLP-based polymorphism detection. US patent 8.481.257 B2, 2018
225. Remington DL, Whetten RW, Liu BH et al (1999) Construction of an AFLP genetic map with nearly complete genome coverage in *Pinus taeda*. Theor Appl Genet 98:1279–1292
226. Klein PE, Klein RR, Cartinhour SW et al (2000) A high-throughput AFLP-based method for constructing integrated genetic and physical maps: progress toward a sorghum genome map. Genome Res 10:789–807
227. Ukrainetz NK, Ritland K, Mansfield SD et al (2008) An AFLP linkage map for Douglas fir based upon multiple full-sib families. Tree Genet Genom 2:181–191
228. Blignaut M, Ellis AG, Le Roux JJ et al (2013) Towards a transferable and cost-effective plant AFLP protocol. PLoS One 8(4):61704
229. Oduwaye O, Baránek M, Cechová J et al (2014) Reliability and comparison of the polymorphism revealed in amaranth by amplified fragment length polymorphism (AFLPs) and inters simple sequence repeats (ISSRs). J Plant Breed Crop Sci 6(4):48–56
230. Jemelkov M, Kitnera M, Krístkov E et al (2015) Biodiversity of *Lactuca aculeata* germplasm assessed by SSR and AFLP markers, and resistance variation to *Bremia lactucae*. Biochem Syst Ecol 61:344–356
231. Varma A, Shrivastava N (2018) Genetic structuring in wild populations of two important medicinal plant species as inferred from AFLP markers. Plant Biosyst 152(5):1088–1100
232. Liersch A, Bocianowski J, Popławska W et al (2019) Creation of gene pools with amplified fragment length polymorphism markers for development of winter oilseed rape (*Brassica napus* L.) hybrid cultivars. Euphytica 215:22

Chapter 13

Random Amplified Polymorphic DNA (RAPD) and Derived Techniques

Kantipudi Nirmal Babu, Thotten Elampilay Sheeja, Divakaran Minoo, Muliyar Krishna Rajesh, Kukkamgai Samsudeen, Erinjery Jose Suraby, and Illathidath Payatatti Vijesh Kumar

Abstract

Understanding biology and genetics at molecular level has become very important for dissection and manipulation of genome architecture for addressing evolutionary and taxonomic questions. Knowledge of genetic variation and genetic relationship among genotypes is an important consideration for classification, utilization of germplasm resources, and breeding. Molecular markers have contributed significantly in this respect and have been widely used in plant science in a number of ways, including genetic fingerprinting, diagnostics, identification of duplicates and selection of core collections, determination of genetic distances, genome analysis, development of molecular maps, and identification of markers associated with desirable breeding traits. The application of molecular markers largely depends on the type of markers employed, distribution of markers in the genome, type of loci they amplify, level of polymorphism, and reproducibility of products. Among many DNA markers available, random amplified polymorphic DNA (RAPD) is the simplest, is cost-effective, and can be performed in a moderate laboratory for most of its applications. In addition, RAPDs can touch much of the genome and has the advantage that no prior knowledge of the genome under research is necessary. The recent improvements in the RAPD technique like arbitrarily primed polymerase chain reaction (AP-PCR), sequence characterized amplified region (SCAR), DNA amplification fingerprinting (DAF), sequence-related amplified polymorphism (SRAP), cleaved amplified polymorphic sequences (CAPS), random amplified microsatellite polymorphism (RAMPO), and random amplified hybridization microsatellites (RAHM) can complement the shortcomings of RAPDs and have enhanced the utility of this simple technique for specific applications. Simple protocols for these techniques are presented along with the applications of RAPD in genetic diversity analysis, mapping, varietal identification, genetic fidelity testing, etc.

Key words AP-PCR, SCAR, DAF, SRAP, CAPS, RAMPO, RAHM, DNA fingerprinting, Genetic diversity, Population and evolutionary genetics, Mapping, Genetic fidelity, Cultivar identification, Bulked segregant analysis

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222,
https://doi.org/10.1007/978-1-0716-0997-2_13,

1 Introduction

1.1 RAPD Technique

The advent of polymerase chain reaction (PCR) and subsequent emergence of DNA-based markers have provided plant taxonomists easy and reliable techniques to study the extent and distribution of variation in species gene pools and to answer typical evolutionary and taxonomic questions which were not previously possible with only phenotypic methods. Properties desirable for ideal DNA markers include highly polymorphic nature, codominant inheritance, and frequent occurrence in the genome, easy access, easy and fast assay, and high reproducibility. DNA marker systems based on PCR include random amplified polymorphic DNAs (RAPDs) [1], amplified fragment length polymorphisms (AFLPs) [2] (Chapter 12), microsatellites/simple sequence repeats (SSRs) [3] (Chapter 11), and single nucleotide polymorphisms (SNPs) [4] (Chapters 9 and 10). Although the sequencing-based molecular techniques provide better resolution at intra-genus and above level [5], they are expensive and laborious. Frequency data from markers such as random amplified polymorphic DNA (RAPD), amplified fragment length polymorphism (AFLP), and microsatellites provide the means to classify individuals into nominal genotypic categories and are mostly suitable for intraspecies genotypic variation study. Compared to other PCR-based techniques which vary in detecting genetic differences and applicability to particular taxonomic levels, RAPD is a cost-effective tool for taxonomic studies.

RAPD is an adaptation of the PCR which relies on the rationale that at low stringency, a given synthetic oligonucleotide primer is likely to find a number of sequences in the template DNA to which it can anneal when these sites are close to each other and lie in opposite orientations and the DNA sequence between the sites will be amplified to produce a DNA fragment characteristic of that genome. Multiple bands of different sizes produced from the same genomic DNA constitute a "fingerprint" of that genome [1]. Patterns from different individuals and species will vary as a function of how similar the genomic DNA sequences are between samples. RAPD polymorphisms result from either chromosomal changes in the amplified regions or base changes that alter primer binding. This assay has the advantage of being readily employed, requiring very small amounts of genomic DNA, and eliminating the need for blotting and radio-active detection. As RAPD requires initial genome information, it provides markers in regions of the genome previously inaccessible to analysis. RAPD-derived estimates of genetic relationships are in good agreement with pedigree, RFLP, and isozyme data [6, 7].

1.2 Recent Applications of RAPD and Its Derived Techniques

DNA fingerprinting for cultivar or varietal identification has become an important tool for estimating genetic diversity for plant breeding, germplasm management, utilization [8], monitoring genetic erosion, and removing duplicates from germplasm collections [9]. As RAPD markers could gain information about genetic similarities or differences that are not expressed in phenotypic information, RAPD analysis becomes an inexpensive tool to characterize germplasm collections [10], to understand the pattern of evolution from wild progenitors, and to classify them into appropriate groups.

RAPDs have been successfully applied in estimation of varietal distinctiveness and relatedness of commercially important crops and registration activities like cultivar identification [11] and hybrid verification [12]. The potential of RAPD for varietal identification has been used to know about the variety being exported or sold under various trade names, for settling a lawsuit involving unauthorized commercialization of patented varieties [13], and to identify the cases of adulteration and even the level of adulteration [14].

As RAPDs make use of arbitrary primers, some of them amplify DNA at highly conserved region, leading to generate polymorphisms at a high level of classification, whereas some will amplify at highly variable region, useful for classification and analyses at and below the species level. This property of RAPD is taxonomically useful at subgeneric level [15] and species level [16] and for the analysis of geographic variation. Another application of RAPD is for evaluation of the genetic integrity of somatic embryo-derived plants [17].

RAPDs have significant use in ecology in studying mating systems and assigning paternity. In plants, insect pollination might be studied by fingerprinting all the potential pollen sources by RAPDs and comparing the dominant RAPD bands seen in the resulting seeds [18]. RAPDs are useful in hybridization studies to document intergeneric hybridization [19] to identify species specific bands as well as interspecific hybridization and detection of introgression in both natural and cultivated plant populations [20]. RAPDs may provide insights into organismal evolutions that are overlooked by single-gene comparisons [21].

The RAPD technique has received a great deal of attention from population geneticists [22] because of its simplicity and rapidity in revealing DNA-level genetic variation.

The RAPD protocol is refined to techniques like sequence characterized amplified region (SCAR), arbitrarily primed polymerase chain reaction (AP-PCR), DNA amplification fingerprinting (DAF), sequence-related amplified polymorphism (SRAP), cleaved amplified polymorphic sequences (CAPS), random amplified microsatellite polymorphism (RAMPO), and random amplified hybridization microsatellites (RAHM) so that some of the current problems such as lack of reproducibility and codominant nature of

inheritance will be overcome. Using several strategies, various modifications have been developed in conjunction with RAPD to enhance the ability to detect polymorphism either by using more than one arbitrary primer [23] or by using a degenerate primer in the amplification reaction [24].

Sequence characterized amplified region (SCAR) markers are generated by sequencing RAPD marker termini and designing longer primers (22–24 nucleotide bases long) for specific amplification of a particular locus [25, 26]. SCARs are usually dominant markers; however, some of them can be converted into codominant markers by digesting them with tetra cutting restriction enzymes, and polymorphism can be deduced by either denaturing gel electrophoresis or single-strand conformation polymorphism (SSCP) [27]. Besides higher specificity, it is based on the presence/absence of a single specific amplicon, considerably simplifying the interpretation of the results, especially when a large number of samples are checked. SCARs also allow comparative mapping or homology studies among related species, thus making them an extremely adaptable concept in the near future.

Arbitrarily primed polymerase chain reaction (AP-PCR) is a special case of RAPD, wherein discrete amplification patterns are generated by employing single primers of 10–50 bases in length in PCR of genomic DNA. Unlike RAPDs, the oligonucleotide length and primer concentrations are tenfold higher [28], and two cycles of low-stringency annealing conditions to allow mismatches followed by PCR at high stringency and the newly synthesized fragments are radiolabeled using dCTP. AP-PCR generated fragments are analyzed as plus/minus DNA amplification-based polymorphism [29] due to either sequence divergence at one of the priming sites or insertion/deletion within the amplification region.

DNA amplification fingerprinting (DAF) uses single arbitrary primers as short as five bases to amplify DNA using polymerase chain reaction with high multiplex ratio [30]. This marker shares those features common to AP-PCR and RAPDs—namely, it results in plus/minus heritable amplification polymorphism, a preponderance of dominant marker loci, and unknown allelism between fragments of equivalent molecular weight. DAF bands contain many more bands than AP-PCR and RAPD patterns, and the likelihood is increased for observing polymorphism between samples. DNA amplification fingerprinting (DAF) has been found to be promising in many plants for cultivar identification and sex determination [31] and for determination of genetic origin and diversity analysis [32].

The sequence-related amplified polymorphism (SRAP) technique, a variation of RAPD, also uses arbitrary primers of 17–21 nucleotides to generate a specific banding pattern aimed to amplify coding sequences (open reading frames (ORFs)) in the genome [33] and results in a moderate number of codominant markers.

SRAP results from two events: fragment size changes due to insertions and deletions, which could lead to codominant markers, and nucleotide changes leading to dominant markers. It has several advantages over other systems: simplicity, reasonable throughput rate, and it allows easy isolation of bands for sequencing, discloses numerous codominant markers, and allows screening thousands of loci shortly to pinpoint the genetic position underlying the trait of interest. The primers and primer concentration vary for each RAPD derived technique which increases its utility in various applications (*see* **Note 1**).

To derive greater information from RAPD patterns, the strategy of hybridizing SSR repeat primers to RAPD amplification patterns has been described. The method has been called either random amplified hybridization microsatellites (RAHM) [34] or random amplified microsatellite polymorphism (RAMPO) [35]. In RAHM, RAPD amplification and oligonucleotide screening are combined for detection of microsatellites to provide more information from RAPD gels and also help to reveal microsatellite genomic clones without the time-consuming screening of genomic libraries [34] (Chapter 9). RAMPO combines arbitrarily or semi-specifically primed PCR with microsatellite hybridization to produce several independent and polymorphic genetic fingerprints per electrophoretic gel. In this approach, the amplified products resolve length polymorphism that may be present either at the SSR target site itself or at the associated sequence between the binding sites of the primers [35]. The RAPD binding site actually serves as an arbitrary end point for the SSR-based amplification product, and therefore, the products obtained are not as restricted by the relative genomic positions of a specific SSR.

Another strategy is referred to as cleaved amplified polymorphic sequences (CAPS), in which sequence information from cloned RAPD bands can be used for analyzing nucleotide polymorphisms. CAPS markers rely on differences in restriction enzyme digestion patterns of PCR fragments caused by nucleotide polymorphism between ecotypes. Sequence information available in databank of genomic DNA or cDNA sequences or cloned RAPD bands can be used for designing PCR primers for this process. Cleaved amplified polymorphic sequences (CAPS) [36] are analogous to RFLP markers in that a region of DNA containing a restriction enzyme site unique to an allele is amplified, cleaved, and compared for their differential migration [36, 37]. The sizes of the cleaved and uncleaved amplification products can be adjusted arbitrarily by the appropriate placement of the PCR primers. Critical steps in the CAPS marker approach include DNA extraction, PCR conditions, and the number or distribution of polymorphic sites.

RAPD has gained a lot of popularity over the last decades due to its ease of operation, low cost, and versatility. It has been extensively used in cultivar identification, genetic diversity analysis, population studies, mapping, molecular breeding and gene tagging, genetic fidelity establishment, etc. RAPD-based identification and characterization of plant genetic resources have helped in attaining goals of conservation of plant resources and in understanding extent and distribution of variation in species gene pools to sort out evolutionary and taxonomic ambiguities. Frequency data from RAPD helps to classify individual into genotypic classes and thus is appropriate for intraspecies genotypic variation studies. RAPD either alone or in combination with other markers like RFLP and SSR provides essential start points for physical isolation of genes of interest, which may further be exploited through marker- assisted selection, gene pyramiding, and transfer to other species. Especially in gene tagging, RAPDs are a preferred method in self-pollinated crops wherein variations between individuals within a species or related breeding material is sought [38]. RAPD is a preferred method for detecting genetic variations induced by somaclonal variation in micro-propagated as well as cryopreserved plants [39]. However, the usage of RAPD has shown a decline in the past few years owing to several factors including the lack of reliability and reproducibility of the technique, advent of novel and derived strategies, and cost-effective means of next-generation sequencing methods. Hence, in the recent references, we could find a trend wherein RAPD analysis was done using very high number of primers [40] or was used along with other markers like ISSR (inter-simple sequence repeat), SSR, AFLP [41, 42], etc., for improving reliability of results. The various applications of RAPD and its derived techniques in plants are extensively dealt in earlier reviews [43–52]. Here, we have compiled only the recent important references on applications of RAPD and its derived techniques as detailed below.

1.2.1 Cultivar Identification

Traditionally, grapevine cultivars have been identified based on the morphological characteristics, but because of the similar pedigree backgrounds, the identification of closely related cultivars has been difficult. Identification of 37 different grapevine cultivars was done using 16 SCAR markers developed from RAPD marker [53]. For identifying cultivars based on random amplified polymorphic DNA (RAPD) markers, cultivar identification diagrams (CIDs) provide a rapid and efficient approach. About 64 tomato cultivars were identified using CID [54]. About 22 onion cultivars were identified using RAPD markers. The cultivars could be easily distinguished based on the polymorphic bands produced by various RAPD primers [55]. Ten autochthonous cultivars of sweet cherry (*Prunus avium*) were validated using 30 RAPD markers. It was also possible to distinguish two important cultivars of tremendous market value

based on the markers [56]. In olive, cultivars sampled from different countries in the Mediterranean region exhibited high resolving power for cultivar identification using RAPD [57]. RAPD technique was used for rapid characterization of Indian medicinal plant *Strychnos minor* Dennst of 16 different localities of Coromandel Coast of Tamil Nadu [58].

SCAR markers based on species-specific RAPD amplicons were developed in four species of the medicinal tuber, *Pinellia ternata*, *Pinellia tripartita*, *Pinellia pedatisecta*, and *Typhonium flagelliforme*, for verification through multiplexing [59]. RAPD-PCR-amplified fragments were used to develop SCAR markers for identification of medicinal plant *Lonicera japonica* [60] and in longan fruits [61]. RAPD fragments from *Litchi chinensis* were cloned, sequenced, and converted into stable SCAR markers for authentication and validation of *L. chinensis* cultivars [62]. Certification of the two maple species, red maple (*Acer rubrum*) and silver maple (*A. saccharinum*), and their hybrids was done through the development of SCAR markers. The information obtained can be used for tracking the introgression of *A. rubrum* and *A. saccharinum* DNA in other hybrid trees or their populations [63]. RAPD-DAF markers were used to discriminate between jalapeño peppers with little phenotypic difference [64]. In yet another study, RAMP-PCR-amplified fragments were used to develop four novel SCAR markers for the genetic authentication of *L. japonica* from its substitutes [65]. RAMP-PCR was found to be better than traditional RAPD-PCR when employed to study genetic diversity and varietal authentication of the herb *Angelica sinensis* (Oliv) [66].

1.2.2 Genetic Mapping and Tagging

For genetic mapping applications, RAPD has been known as a non-biased and neutral marker. It does not require information about a particular sequence in the genome [67]. In RAPD analysis, the entire plant genome is targeted for primer annealing which facilitates development of a higher density map. RAPD does not require DNA probes, blotting and hybridization, and primer designing procedures. Small amounts of DNA are required, and high-throughput sampling can be obtained. RAPD generated DNA fragments possessed many of the DNA sequences that are related to chromosome size changes as it is reported in many studies that the amplified fragments in an RAPD reaction were preferentially amplified from species containing a common genome consisting of large chromosomes [68]. The above advantages make RAPD a preferred choice in gene tagging involving several different types of populations like backcross selection progenies, recombinant inbred lines, near-isogenic lines, etc. Bulked segregant analysis was also employed to tag traits from populations having contrasting characters [38].

In *Saccharum officinarum* L., an RAPD marker was found to be linked to eyespot susceptibility, and it also helped to identify additional linkage groups. This particular work showed that linkages identified in this map could potentially be used for marker-assisted selection [69]. Molecular evaluation of two guava mapping populations (MP), MPI comprising 94 F1 progenies and MPII comprising 46 F1 progenies, was carried out using random amplified polymorphic DNA (RAPD) markers. Genotypic data thus generated can be further exploited for constructing genetic linkage maps and mapping complex Quantitative Trait Loci (QTLs) governing fruit quality traits in guava [70]. A reference genetic map for *Capsicum baccatum* was constructed based on RAPD molecular markers [71]. Using SRAP markers, a molecular genetic map for hawthorn, a medicinal plant, was constructed which can be used for marker-assisted selection in the particular plant species [72].

1.2.3 Assessment of Outcrossing Rates

Outcrossing rates in sweet passion fruit were assessed using RAPD molecular markers. The results showed that all the progenies assessed were derived as a result of outcrossing [73]. RAPD was used to study outcrossing in *Agave schottii*, and it was found that RAPD markers are useful tools for assessing ecological phenomena like outcrossing [74]. RAPD markers were used to estimate the outcrossing rate in Ethiopian mustard (*Brassica carinata*). It was analyzed by looking into the banding pattern of offsprings of two parental lines grown in open pollinated isolation lines [75]. The rate of outcrossing in orchards containing 'Hass' avocado (*Persea americana* Mill.) was determined using RAPD markers. The data included 2393 fertilization events taken from two areas of southern California of different climate over a period of 4 years. Three potential pollen sources were also investigated using RAPD markers specific to each pollen source [76]. RAPD markers were found to be useful in understanding breeding patterns in faba beans [77]. In *B. carinata*, RAPD markers helped in estimating outcrossing rate and the opportunity for exploiting heterosis through synthetic and/or hybrid cultivar breeding [75].

1.2.4 Genetic Fidelity Testing

Genetic fidelity testing of in vitro propagated *Araucaria excelsa* R. Br. var. *glauca* plantlets was done using RAPD technique. A total of 1676 fragments were generated with 12 RAPD primers in micro-propagated plants and mother plants [78]. RAPD was employed to test the genetic fidelity among the regenerants in *Spilanthes calva* DC [79]. Genetic fidelity was confirmed in micro-propagated *Drosera* plantlets using RAPD [80]. Assessment of genetic fidelity through RAPD analysis was done in in vitro raised plants (*Swertia chirayita*), and the plants showed high clonal fidelity [81]. In vitro regeneration of *Guizotia abyssinica* Cass and

evaluation of genetic fidelity through RAPD markers showed the presence of somaclonal variation in the plantlets arising from direct regeneration as well as from indirect regeneration [82]. Some studies endorse utilizing one more marker like ISSR in conjunction with RAPD for better analysis of genetic fidelity in banana [83], grapes [84], and mango ginger [85]. Genetic stability of in vitro propagated potato micro-tubers examined using AFLP, SSR, and ISSR indicated them to be superior to RAPD [86]. In endemic medicinal plants *Pittosporum eriocarpum* Royle [87] and *Rauvolfia tetraphylla* L., [41], RAPD was used to validate the genetic homogeneity of in vitro raised plantlets in conjunction with SCoT and ISSR markers. In *Salvia hispanica* L., a reasonably good number of RAPD and ISSR primers were employed for confirming genetic fidelity of in vitro regenerated plantlets [88]. The genetic uniformity of blackberry plants (*Rubus fruticosus* L.) obtained by micropropagation was analyzed by RAPD and SRAP markers [89]. ISSR and RAPD analysis was used to assess genetic uniformity of transgenic cotton containing Bt and chitinase genes [42].

1.2.5 Inter and Intraspecies Variations and Genetic Diversity

RAPD is found to be more suitable in large-scale screening of closer populations found in similar habitats. However, the discrimination capacity decreases relatively when populations from distant locations are analyzed. RAPD may not be much suitable for genetic diversity analyses of populations in wide geographic areas. RAPD includes some deflections in the genetic discrimination of populations having high genetic diversity in different habitats. Combining RAPD and SCAR markers provides a simple and reliable tool for genetic characterization of plant species. Genetic diversity of 21 aromatic rice genotypes (*Oryza sativa* L.) was assessed using about 38 RAPD primers [90]. The RAPD profile helps to identify variations of the diagnostic markers on aromatic rice genotypes [91], identification of rice at the level below species [92]. For the identification and protection of natural resources, genetic tracking of aromatic rice germplasm is essential. Genetic variation in *Ocimum* species was studied using RAPD markers. Many unique species-specific alleles were amplified by RAPD in *Ocimum* species [93]. In bamboo, RAPD-RFLP analysis was able to generate a low-cost and fast screening method for genetic characterization of genera and species of bamboo [94]. In *Miscanthus* spp., genetic diversity and relationships based on RAPD and AFLP indicated significant genetic differentiation among accessions due to geographic distance [95].

Genetic diversity analysis in sweet potato [96] and *Elymus* spp. [97] indicated a close correspondence of RAPD and ISSR markers in detecting variability. Genetic diversity studies in *Harpagophytum* species using ISSR and RAPD markers indicated evidences of introgression and interspecific gene flow [98]. Genetic diversity analysis

of cumin genotypes based on sequence-related amplified polymorphism (SRAP) markers was conducted, and it was found that there is a need for enhancing the genetic base of cumin germplasm using different breeding approaches, viz., mutagenesis, wide hybridization or somaclonal variation, and germplasm introduction [99]. Genetic diversity and population structure study within and among six natural populations of *Limonium sinense*, a plant which has medicinal and ornamental values, was conducted using SRAP markers, which could develop insight and useful strategies for its conservation [100]. A highly efficient and economical technology of sequence-related amplified polymorphism (SRAP) molecular markers with an automated fragment analyzer ABI 3500xL was developed, to detect genetic diversity in upland cotton [101]. Genetic diversity studies in strawberry cultivars in Indonesia using CAPS molecular markers resulted in the grouping of the cultivars into four clusters [102].

Cleaved amplified polymorphic sequence (CAPS) marker analysis of four chloroplast DNA regions, rbcL-ORF106, trnF-trnV, trnV-rbcL, and trnK2-trnQ, in 42 citrus accessions including mandarins and their close relatives showed their close relationship and low variation in chloroplast DNA of mandarins [103].

1.2.6 Others

RAPD was used to evaluate genotoxic effects in many studies to identify DNA damage induced due to harmful agents like heavy metals [104–106]. RAPD was successfully applied to whole germplasm collections of flax to identify redundant and distinct accessions and associated traits useful in future breeding programs [107] and to identify duplicates in germplasm collections of rice at International Rice Research Institute, Philippines [108]. RAPD is a preferred choice for the detection of adulteration in medicinal plants and successfully used especially when the adulterant is a different species [109, 110]. An interesting study has been reported that utilizes commercial RAPD analysis beads in differentiating about 63 different food and feed legume species for establishing authenticity and correct labeling of raw material used in food or feed samples [111]. Similarly in medicinally important *Ocimum* spp., diagnostic RAPD markers were useful in identifying raw materials for herbal drugs [112]. RAPD markers linked to disease-resistant genes in plants like the rpg4 gene responsible for stem rust resistance in barley [113] and heat smut resistance [114] have been identified. Dwarfism gene has been located by an RAPD marker in barley [115]. RAPD markers were exploited in identifying somatic hybrids [116]. RAPD was successfully used to reveal polymorphism in mutant potato [117] and chrysanthemum [118] obtained via gamma irradiation.

1.3 Disadvantages of RAPD Technique and Solutions

The main concern about RAPD is its lack of reproducibility within and between laboratories. Differences in amplification patterns based on type of thermocycler and primers used and also concentration of Taq polymerase and amplification conditions are the commonly reported issue. The most important factor affecting reproducibility is the low quality of DNA template [28]. Differences between template DNA concentrations of individual samples can also affect the amplification profile [45]. It is a dominant marker and presence of a band of apparently identical molecular weight in two different individuals cannot be considered as identical loci and thus gives more accurate estimates between closely related populations than the distant ones (1). A single RAPD band can be comprised of a number of co-migrating amplification products. However, it is suggested that RAPD polymorphisms can be successfully reproduced among laboratories when standard reaction conditions are used and similar temperature profiles in tubes are followed [119]. Some authors also report that when more samples and primers are included in the study, the fingerprint and phylogeny are more accurate [120]. A preliminary pedigree analysis is a prerequisite to assign markers to specific loci. To get comparable results with other codominant markers, two to ten times more individuals need to be sampled per locus, and marker alleles for these loci should be in low frequencies [121]. Many studies indicate that RAPD shows significant difficulties in cultivar characterization due to low polymorphism, irreproducibility, and weak grouping due to artifacts [122, 123]. RAPD marker identity might be established by fingerprinting a set of standard genotypes by RAPD to facilitate communication and the reproducibility among laboratories. In cases where a single primer is unable to distinguish all cultivars in a study, a combination of polymorphic bands generated by various primers can be utilized. Converting RAPD markers to

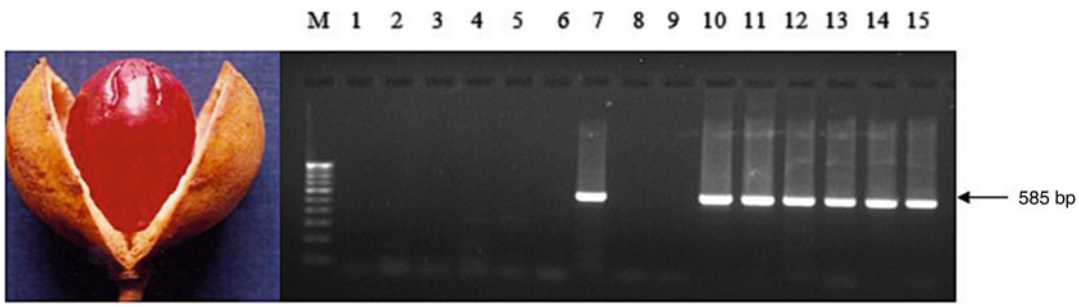


Fig. 1 Unique RAPD-derived SCAR marker for identification of an endangered and endemic species of *Myristica*, viz., *Knema andamanica*. (**a**) Fruits of *K. andamanica* with unique fused mace. (**b**) RAPD derived SCAR marker showing amplification of a marker of 585 bp in *K. andamanica* accessions absent in other wild and related genera of *Myristica*. Lanes M- 100 bp marker. 1–8: *M. fragrans*, *M. beddomei*, *M. malabarica*, *M. prainii*, *M. fatua*, *M. andamanica*, *K. andamanica*, *M. amygdalina*, 9: Control, lanes 10–15: Different germplasm accessions of *K. andamanica* from the repository at ICAR-IISR

more reliable SCAR markers and also using one or two other marker methods in conjunction with RAPD are some useful tips to improve reliability and reproducibility of results (Fig. 1).

2 Materials

2.1 Genomic DNA Isolation and Quantification

1. Extraction buffer (2×): 2% cetyltrimethylammonium bromide (CTAB), 100 mM Tris HCl, pH 8, 20 mM ethylenediaminetetraacetic acid (EDTA), pH 8, 1.4 M NaCl, 1% polyvinyl polypyrrolidone (PVPP).
2. Chloroform: isoamyl alcohol (24:1).
3. 100% Ethanol or isopropanol.
4. 70% Alcohol.
5. TE buffer (10 mM Tris, 0.1 mM EDTA, pH 8).
6. RNase A (10 mg/mL).
7. Tris-acetate-EDTA (TAE) buffer (pH 8) (50×).
8. Agarose.
9. Ethidium bromide (10 mg/mL).
10. Loading dye (6×): 30% glycerol, 5 mM EDTA, 0.15% bromophenol blue, 0.15% xylene cyanol.
11. MassRuler 1000 bp DNA ladder.

2.2 Reagents Used for RAPD-PCR

1. Taq DNA polymerase with 10× buffer.
2. 10 mM dNTPs: 10 mM each of dATP, dCTP, dGTP, and dTTP.
3. 25 mM $MgCl_2$.
4. 10 μM Primers (operon primers are the most commonly used RAPD primers) (*see* **Notes 1** and **2**).
5. Milli-Q water.

2.3 Sequence Characterized Amplified Region (SCAR)

2.3.1 Genomic DNA Isolation and Quantification

(*See* Subheading 2.1).

2.3.2 Reagents for PCR

(*See* Subheading 2.2).

2.3.3 Gel Extraction

1. QIAquick gel extraction kit, Qiagen, Germany.

2.3.4 Cloning of PCR Amplified Gene

1. PCR amplified and purified product.
2. PCR cloning vector.
3. T4 DNA ligase.
4. Ligation buffer (5×).
5. Sterile deionized water.
6. Overnight culture of *E. coli* DH5∝.
7. $CaCl_2$ (100 mM).
8. Mg Cl_2 (25 mM).
9. LB medium.
10. Sterile microcentrifuge tubes and tips.
11. Sterile glycerol (80%).
12. LB agar with ampicillin (100 µg/mL), X gal (20 µg/mL), and IPTG (40 µg/mL).

2.4 Arbitrarily Primed Polymerase Chain Reaction (AP-PCR)

(*See* Subheading 2.1).

2.4.1 Genomic DNA Isolation and Quantification

2.4.2 Reagents for PCR

1. Taq polymerase.
2. PCR buffer (10×).
3. 25 mM $MgCl_2$.
4. 10 mM each of dNTPs.
5. 50 µCi α-[^{32}P] dCTP.
6. 10 µM of each primer.

2.4.3 Electrophoresis

1. 40% Acrylamide-bis-acrylamide.
2. 7.5 M Urea.
3. Tris-borate-EDTA (TBE) buffer, pH 8 (10×).

2.5 DNA Amplification Fingerprinting (DAF)

(*See* Subheading 2.1).

2.5.1 Genomic DNA Isolation and Quantification

2.5.2 Reagents for PCR

(*See* Subheading 2.2).

2.5.3 PAGE Reagents

1. 40% Acrylamide-bis-acrylamide.
2. 7.5 M Urea.
3. Tris-borate-EDTA (TBE) buffer, pH 8 (10×).
   Cover the bottle with aluminum foil and store at 4 °C and use before 1 month.
4. 10 bp MassRuler.
5. 100 bp MassRuler.

2.5.4 Silver Staining Reagents

1. Acetic acid, glacial.
2. Silver nitrate crystal, AR (ACS) ($AgNO_3$).
3. Formaldehyde solution, AR (ACS) (HCHO).
4. Sodium thiosulfate (Na_2S_2O).
5. Sodium carbonate powder, ACS reagent (Na_2CO_3).
6. Ethanol.
7. Silver staining solution (250 mg silver nitrate and 375 μL formaldehyde and 50 μL sodium thiosulfate).
8. Ice-cold developer solution (10 °C) (7.5 g sodium carbonate, 375 μL formaldehyde, and 50 μL sodium thiosulfate (10 mg in 1 mL water) in 250 mL water).
9. Formamide loading dye: 80% formamide, 10 mM EDTA, pH 8.0, 1 mg/mL xylene cyanol, 1 mg/mL bromophenol blue, 50% glycerol in a final volume of 10 mL.

2.6 The Sequence-Related Amplified Polymorphism (SRAP) Technique

2.6.1 Genomic DNA Isolation and Quantification

(*See* Subheading 2.1).

2.6.2 Reagents for PCR Conditions

(*See* Subheading 2.2 but using different primers in **step 4**).

1. *Primers*: The arbitrary primers consist of the following elements: core sequences, which are 13 to 14 bases long, where the first ten or 11 bases starting at the 5′ end are sequences of no specific constitution ("filler" sequences), followed by the sequence CCGG in the forward primer and AATT in the reverse primer. The purpose of using the "CCGG" sequence in the core of the first set of SRAP primers was to target exons to open reading frame (ORF) regions.

2.6.3 PAGE Electrophoresis

(*See* Subheadings 2.5.3 and 2.5.4).

2.7 Random Amplified Microsatellite Polymorphism (RAMPO)

(*See* Subheading 2.1).

2.7.1 Genomic DNA Isolation and Quantification

2.7.2 Reagents Used for RAPD and Microsatellite-Primed PCR (MP-PCR)

(*See* Subheading 2.2).

2.7.3 Hybridization with Microsatellite-Complementary Probes

1. Nylon membrane (Hybond, Amersham).
2. ^{32}P-labeled microsatellite-complementary oligonucleotide probes.
3. 5 mM EDTA.

2.8 Random Amplified Hybridization Microsatellites (RAHM)

(*See* Subheading 2.1).

2.8.1 Genomic DNA Isolation and Quantification

2.8.2 Reagents Used for RAPD-PCR

(*See* Subheading 2.2).

2.8.3 Hybridization with Microsatellite-Complementary Probes

(*See* Subheading 2.7.3).

2.9 Cleaved Amplified Polymorphic Sequences (CAPS)

(*See* Subheading 2.1).

2.9.1 Genomic DNA Isolation and Quantification

2.9.2 Reagents for PCR Conditions

(*See* Subheading 2.2).

2.9.3 Restriction Enzyme Digestion

1. Restriction enzymes: Mse I, Alu I, Mbo I, Hae III.
2. Buffer 2 (New England Biolabs (NEB), UK)—supplied at 10× concentration.
3. NEB buffer 2 (1×).

4. 50 mM NaCl.
5. 10 mM Tris–HCl.
6. 10 mM $MgCl_2$.
7. 1 mM DTT, pH 7.9 at 25 °C.
8. 100× BSA (10 mg/mL)—use at 1×.

2.9.4 PAGE Reagents (*See* Subheading 2.5.3).

2.9.5 Silver Staining Reagents (*See* Subheading 2.5.4).

3 Methods

3.1 Isolation of Genomic DNA (Modified Doyle and Doyle, 1990) [124]

1. Grind 2 g of clean young leaf tissue to fine powder with a pestle and mortar after freezing in liquid nitrogen; transfer it to 10 mL CTAB extraction buffer and incubate at 60 °C for 1 h.
2. Extract the supernatant with chloroform: isoamyl alcohol (24:1) and centrifuge at 12,378 × *g* for 10 min at room temperature.
3. Precipitate the DNA with 100% ethanol or isopropanol; centrifuge at 19,341 × *g* for 10 min at 4 °C.
4. Wash the DNA with 70% ethanol; centrifuge at 19,341 × *g* for 5 min at 4 °C.
5. Dry the pellet and dissolve the DNA in 1× TE buffer.
6. Treat the DNA in solution with RNase (10 μg/mL) at 37 °C for 30 min.
7. Wash with chloroform: isoamyl alcohol (24:1) and centrifuge at 12,378 × *g* for 10 min at room temperature.
8. Precipitate with 100% ethanol and dissolve in 1× TE buffer. Store frozen at −20 °C.

3.2 DNA Quantification

It is an essential step in many procedures where it is necessary to know the amount of DNA that is present when performing techniques such as PCR and RAPDs.

3.2.1 By Gel Electrophoresis

The comparison of an aliquot of the extracted sample with standard DNAs of known concentration (Lambda *Hind* III) can be done using gel electrophoresis.

1. 5 μL of the DNA is mixed with 1 μL of 6× loading dye and loaded onto a 0.8–1% agarose gel along with 500 ng of Lambda *Hind* III digest marker and electrophoresed at 90 V for 30 min.
2. The quantity of extracted DNA is estimated based on the intensity of Lambda *Hind* III digest marker bands as the top

bands account for half amount (250 ng) of total loaded amount.

3. The quality of genomic DNA is confirmed for its integrity.

3.2.2 Using UV Spectrophotometer

1. Take 1 mL of TE buffer in a cuvette and calibrate the spectrophotometer at 260 nm and 280 nm wavelength.
2. Add 2 to 5 μL of DNA, mix properly, and record the optical density at both 260 nm and 280 nm.
3. Estimate the DNA concentration employing the following formula:

$$\text{Amount of DNA } (\mu g/\mu L) = (\text{OD})\ 260 \times 50 \times \text{dilution factor}/1000$$

4. Judge the quality of DNA from the ratio of OD values recorded at 260 and 280 nm. Pure DNA has values close to 1.8.
5. Dilute the DNA sample to get 20 ng/μL.

3.3 RAPD

3.3.1 PCR Amplification of Genomic DNA with Primers

Amplify 20–50 ng of genomic DNA in a reaction mix containing 1.0 U *Taq* DNA polymerase, 1 μM primer, 1.5–2.0 mM $MgCl_2$, 0.125 mM each of dNTPs, and 1× *Taq* DNA polymerase buffer (*see* **Note 1**).

1. The amplification profile consists of an initial denaturation of 3 min at 94 °C followed by 35–40 cycles of denaturation for 1 min at 94 °C, annealing for 37 °C for 1 min and extension at 72 °C for 2 min and final extension for 6 min at 72 °C (*see* **Note 2**).

3.3.2 Gel Electrophoresis

1. Amplified RAPD products are separated by horizontal electrophoresis in 1.5% (w/v) agarose gel, with 1× TAE buffer, stained with ethidium bromide (0.5 μg/mL) and analyzed under ultraviolet (UV) light. The length of the DNA fragments is estimated by comparison with DNA ladder.

*3.3.3 Scoring and Interpretation of RAPD Banding Patterns (See **Note 3**)*

Variability is then scored as the presence or absence of a specific amplification product.

Polymorphism usually results from mutations or rearrangements either at or between the primer binding sites, due to appearance of a new primer site, mismatches at the primer site, and difference in the length of the amplified region between the primer sites due to deletions or insertions in the DNA.

1. Each gel is analyzed by scoring the present (1) or absent (0) polymorphic bands in individual lanes. The scoring procedure is based on the banding profiles which are clear, transparent, and repeatable (*see* **Note 4**).

   The RAPD profiles are compared between the genotypes to estimate the similarity index. Studies are initiated to assess the similarity/differences between the genotypes using RAPD polymorphism as estimated by Paired Affinity Indices (PAIs).

   PAI is calculated by the formula PAI
   = no.of similar bands/total no.of bands

   The PAIs expressed as percentage indicate the similarity (%) between any two genotypes.
2. The binary matrix is transformed into similarity matrix using Dice similarity (NTSYS-PC 2.01; Numerical Taxonomy System of Multivariate Programs) [125]. The Dice coefficient is preferred to the Jaccard coefficient because it assigns weights to matches rather than to mismatches and does take shared absences of bands into account (*see* **Notes 5** and **6**).
3. The similarity matrix is subjected to a clustering analysis using the unweighted pair group method with arithmetic means (UPGMA; NTSYS-PC 2.0) [125].
4. The RAPD matrix can also be analyzed using the neighbor-joining (N-J) method. Evaluate statistical support for the clusters recovered both in the UPGMA and N-J trees by generating 1000 bootstrap pseudoreplicates (*see* **Note 7**) (Fig. 2).

3.4 Sequence Characterized Amplified Region (SCAR)

1. Genomic DNA is isolated, quantified, and diluted (*see* Subheading 3.1).
2. 20–50 ng of genomic DNA is amplified using random primers (*see* Subheading 3.3.1).

3.4.1 Amplification

3. Aliquots (5.0 μL) of RAPD products are separated by horizontal electrophoresis in 1.5% (w:v) agarose gel, with 1× TAE buffer, stained with ethidium bromide (0.5 μg/mL) and analyzed under ultraviolet (UV) light. The length of the DNA fragments is estimated by comparison with DNA ladder.

3.4.2 RAPD Fragment Selection and Cloning

1. From obtained RAPD fingerprints, the polymorphic RAPD marker bands are selected.
2. These bands are cut, eluted, and purified using QIAquick gel extraction kit, cloned and sequenced.
4. Primer design: New longer and specific primers of 15–30 bp are designed for the DNA sequence, which is called the SCAR (*see* **Note 8**).

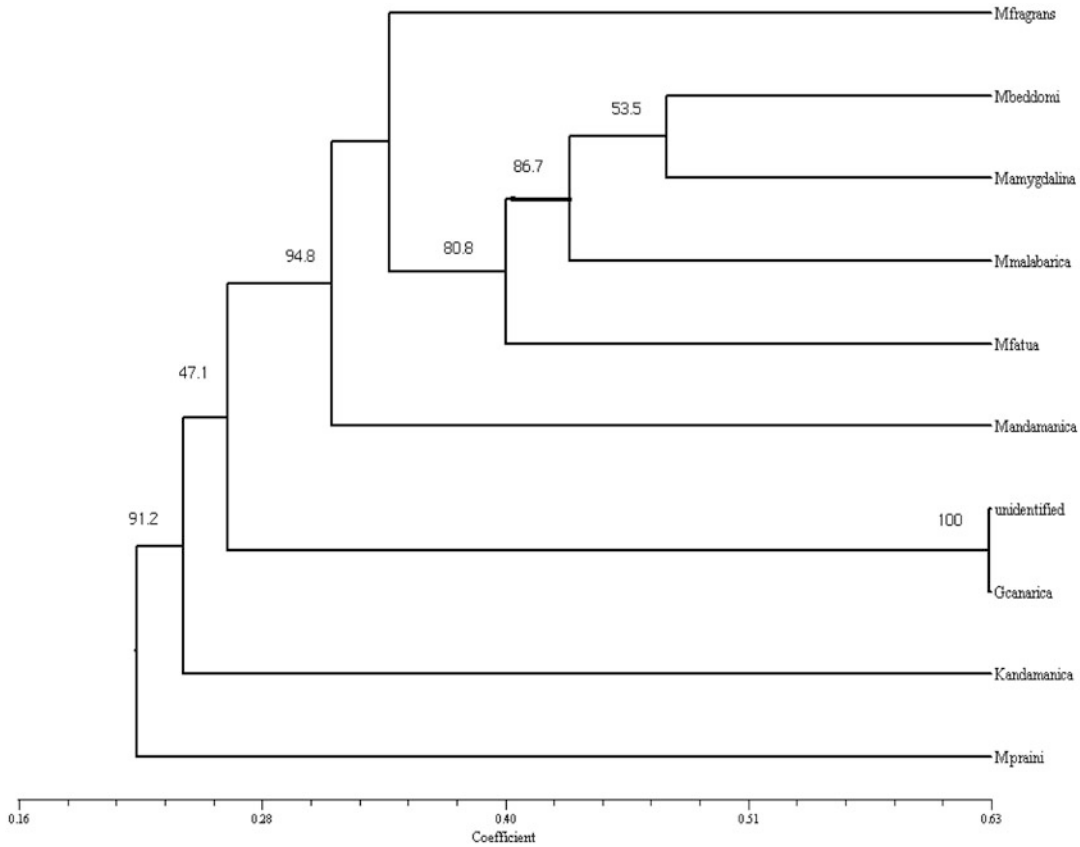


Fig. 2 Dendrogram generated using UPGMA using RAPD marker data in wild and related genera of *Myristica*. Number of forks indicates confidence limits for grouping of those species in a branch occurred, based on 2000 cycles in bootstrap analysis, using Winboot program

3. PCR amplification: For the verification of primers ability to amplify predicted fragment length, primers are tested with isolated DNA.

3.5 Arbitrarily Primed Polymerase Chain Reaction (AP-PCR)

3.5.1 Amplification

1. Amplify 20 ng genomic DNA in a PCR reaction mix containing 0.025 U Taq polymerase and 1× buffer (Stratagene) with 4 mM $MgCl_2$, 0.2 mM of each dNTP, and 10 μM primer.
2. Amplification profile consists of an initial denaturation of 94 °C for 5 min followed by 40 °C for 5 min for low stringency annealing of primer and 72 °C for 5 min for extension for two cycles. This temperature profile is followed by ten high stringency cycles: 94 °C for 1 min, 60 °C for 1 min, and 72 °C for 2 min for ten cycles.
3. At the end of this reaction, add 90 μL of a solution containing 2.25 U Taq polymerase in 1× buffer, 0.2 mM dNTPs, and 50 μCi α-[^{32}P] dCTP, and the high stringency cycles are continued for an additional 20 or 30 rounds.

3.5.2 Electrophoresis

1. Prepare the 40% stock 19:1 acrylamide bis-acrylamide solution and store it in dark bottles at 4 °C.
2. Prepare 5% working solution containing 7.5 M urea, 40% acrylamide bis-acrylamide. Assemble electrophoresis unit by adding 0.5× TBE buffer to upper tank and lower tank.
3. Add 4 μL of the loading buffer to 8 μL of the final amplified reaction mix.
4. Load this sample into the gel and conduct electrophoresis at 200 V for 55 min.
5. The AP-PCR generated fragments are size separated on polyacrylamide and visualized via radiography.

3.6 DNA Amplification Fingerprinting (DAF)

3.6.1 Amplification

1. Amplify 20 ng of genomic DNA in a 10 μL PCR reaction mix containing 0.5 U of Taq polymerase, 200 μM each dNTP, 0.5 μM primer, and 1× PCR buffer with 2 mM $MgCl_2$ overlaid with a drop of mineral oil.
2. The amplification profile consists of an initial denaturation at 5 min of 94 °C followed by 40 cycles of denaturation for 5 s at 94 °C, annealing at either 35 °C or 45 °C and 30 s at 72 °C.
3. The amplification products are separated in a vertical electrophoresis system using 5% non-denaturing polyacrylamide gel of 0.5 mm thickness to separate DNA fragments according to their molecular weight.
4. Gel preparation (*see* Subheading 3.5.2).

3.6.2 Silver Staining for DNA Visualization

1. Gently place the gel in 10% (v/v) glacial acetic acid for 30 min at room temperature.
2. Rinse the gel in deionized water twice for about 2 min each.
3. Immerse the gel in silver staining solution for 20 min.
4. Pour out the silver stain solution and wash the gel quickly with deionized water within 10 s.
5. Immerse the gel in an ice-cold developer solution (10 °C) until optimal image intensity is obtained. Stop the developing process by immersing the gel in 7.5% ice-cold glacial acetic acid.
6. Transfer gel onto the Whatman paper.
7. Air-dry the gel or dry using gel drier at 70 °C for 30 min.

3.6.3 Gel Interpretation

Scoring can be done by the presence or absence of band. Bands are sized and matched directly on gels, autoradiographic or photographic films, or photocopies on transparency overlays.

3.7 Sequence-Related Amplified Polymorphism (SRAP) (See Note 9)

3.7.1 Amplification

1. Amplify 20 ng of genomic DNA in a PCR reaction mix containing 1 U of *Taq* polymerase, 200 μM each dNTP, 0.1 mM each forward and reverse primer, and 1× PCR buffer with 1.5 mM $MgCl_2$.
2. The amplification profile consists of an initial denaturation at 2 min of 94 °C followed by five cycles of denaturation for 1 min at 94 °C, annealing at 35 °C for 1 min and 72 °C for 1 min; followed by 35 cycles of 94 °C for 1 min, 50 °C for 1 min, and 72 °C for 1 min; followed by 7 min at 72 °C.
3. Polyacrylamide gel electrophoresis (*see* Subheading 3.5.2).
4. Marker analysis: Each polymorphic band can be scored as a single dominant marker.

3.7.2 Sequencing of SRAP Marker Bands

1. After electrophoresis, the gel is exposed overnight to a high-sensitivity film (Kodak BioMax).
2. Using the exposed film as a blueprint, the gel pieces containing the polymorphic bands are cut and introduced into a dialysis tube.
3. The dialysis tube is placed into the buffer tank of a sequencing-gel apparatus, and the DNA is electro-eluted in 1× TBE buffer. The application of 2000 V, which is the same voltage used for running sequencing gels, results in the complete electro-elution of DNA into buffer from the gel fragment.
4. After ethanol precipitation and TE buffer suspension, the DNA can be used for direct sequencing.

3.8 Random Amplified Microsatellite Polymorphisms (RAMPO)

(*See* Subheadings 3.1 and 3.2).

3.8.1 Genomic DNA Isolation

3.8.2 Amplification of Genomic DNA with RAPD Primers/Microsatellite Primers

1. The DNA is first amplified with a single arbitrary (*see* Subheading 3.3.1) or microsatellite-complementary PCR primer (MP-PCR) (*see* **Note 10**).
2. The products are separated on agarose gel (1.4%), stained with ethidium bromide, and photographed.

3.8.3 Hybridization with Microsatellite-Complementary Probes

1. The gel is either dried or blotted onto a nylon membrane.
2. Hybridize to a [^{32}P]-labelled, microsatellite-complementary oligonucleotide probe.
3. Hybridization was done overnight at 42 °C containing 20–40 ng/mL of the probe.

4. Filters are washed twice for 5 min at room temperature in 2× SSC, 0.1% SDS followed by two final washing steps (2× 15 min) at different stringencies.
5. The stringency can be varied through temperature (50–65 °C) and salt concentration (1× SSC; 0.1% SDS to 0.1× SSC; 0.1% SDS).
6. Positive signals are detected by chemiluminescence system and documented by exposure to X-ray film for 1–2 h.

3.9 Random Amplified Hybridization Microsatellites (RAHM)

1. Amplify the DNA using RAPD primers (*see* Subheading 3.3.1).
2. The amplified products are separated by gel electrophoresis (*see* Subheading 3.3.2).
3. The polymorphisms on the agarose gel are identified and scored (*see* Subheading 3.3.3).
4. The amplified DNA is then transferred onto Hybond-N+ filters using Southern blot procedures.
5. The filters are then hybridized with radiolabeled oligonucleotide probes carrying simple sequence repeats (SSR).
6. The luminescent signals produced are detected by autoradiography. Hybridizing bands are named random amplified hybridization microsatellites (RAHM).

3.10 Cleaved Amplified Polymorphic Sequences (CAPS)

1. Genomic DNA is isolated (*see* Subheadings 3.1 and 3.2).
2. Amplify the different CAPS marker locus by PCR (*see* Subheading 3.3.1).
3. Analyze the PCR by gel electrophoresis to confirm amplification of DNA and the yield.
4. Mix 5 μL PCR reaction and 10 μL digest mix. The reaction mixture for the enzyme digestion contained 5 μL PCR product, 9 μL ddH_2O, and 0.3 μL restriction enzyme (10 U/μL), which were then incubated at 37 °C for 5 h and then heated to 65 °C for 5 min.
5. Mix equal parts of digest mix and formamide loading dye. Denature sample by heating at 94 °C for 5 min and then placing tube on ice.
6. Resolve restriction fragments using 1× TBE, 8.25% polyacrylamide gel.
7. Load 2.5 μL of the denatured sample per lane.
8. Denature by heating at 94 °C for 5 min and then placing tube on ice.
9. Load 3.5 μL of the denatured ladder per lane, equivalent to 117 ng DNA.

10. Run gel at 80 W for approximately 80 min or until the bromophenol blue dye front has reached the bottom of the gel.
11. Follow usual silver staining protocol to stain gel (*see* Subheading 3.6.2).

4 Notes

1. RAPD reaction is far more sensitive than conventional PCR because of the length of a single and arbitrary primer used to amplify anonymous regions of a given genome. Optimization of reaction conditions should precede the actual RAPD analysis to get consistent and reproducible results. The following optimizations are essential: template DNA concentration and quality, *Taq* DNA polymerase concentration, Mg^{2+} ion concentration, primer concentration and annealing temperature, and primers suitable for detection of polymorphic loci in the taxa to be analyzed [126].
2. Too many RAPD cycles can increase the amount and complexity of nonspecific background products, while too few cycles give low product yield. The optimum number of cycles will depend mainly upon the starting concentration of target DNA when other parameters are optimized. Although the sequences of RAPD primers are arbitrarily chosen, two basic criteria must be met: a minimum of 40% GC content (50–80% GC content is generally used) and the absence of palindromic sequence (a base sequence that reads exactly the same from right to left as from left to right). Because G-C bond consists of three hydrogen bridges and A-T bond consists of only two, a primer-DNA hybrid with less than 50% GC will probably not withstand the 72 °C temperature at which DNA elongation takes place by DNA polymerase [1].
3. Data from at least ten primers with a total of 100 RAPD bands are needed to produce a stable classification [127].
4. The probability of a scored RAPD band being scored in replicate data is strongly dependent on the uniformity of amplification conditions between experiments, as well as relative amplification strength of the RAPD band [128]. The criteria for selecting scoring bands include reproducibility and consistency (the experiments need to be repeated to achieve reproducible results) and thickness and size of the bands.
5. Deleting inconsistent or faint bands or using only those bands that are reproducible introduces false negatives, and simply ignoring RAPD artifacts and using all bands introduces false positive into RAPD data [129].

6. If estimates of the percent of false-positive and false-negative bands in the RAPD data are available (such as when replicate runs have been made), equations described earlier [130] can be used to determine the actual bias by subtracting the true value from the estimated value. Once the bias is known, it can be used to determine whether the RAPD protocol has been optimized sufficiently to provide accurate enough estimates of the similarities.
7. Other softwares like PAUP, PHYLIP, CLINCH, MaClade, PopGene, and Arlequin can also be used to accomplish the cluster algorithms and for phylogenetic analysis.
8. In SCAR, the longer primer sequence increases the specificity of the PCR reaction and produces results less sensitive to changes in reaction conditions. SCAR is thus more reproducible than RAPD [131].
9. The rationale behind primer designing in SRAP is based on the fact that exons are normally in GC-rich regions. The core is followed by three selective nucleotides at the 3′ end. The filler sequences of the forward and reverse primers must be different from each other and can be 10 or 11 bases long.
10. If RAPD gels were used for RAMPO analysis, banding patterns are generally less complex, less variable, and easier to interpret than those derived from MP-PCR gels [132].

References

1. Williams JG, Kubelik AR, Livak KJ et al (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535
2. Vos P, Hogers R, Bleaker M et al (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23:4407–4414
3. Powell W, Machray GC, Provan J (1996) Polymorphism revealed by simple sequence repeats. Trends Plant Sci 1:215–222
4. Gupta PK, Roy JK, Prasad M (2001) Single nucleotide polymorphisms: a new paradigm for molecular marker technology and DNA polymorphism detection with emphasis on their use in plants. Curr Sci 80:524–535
5. Robinson JP, Harris SA (1999) Which DNA marker for which purpose. In: Gillet EM (ed) . Institut für Forstgenetik und Forstpflanzenzüchtung, Universität Göttingen, Göttingen, Germany
6. Vierling RA, Nguyen HT (1992) Use of RAPD markers to determine the genetic diversity of diploid, wheat genotypes. Theor Appl Genet 84:835–838
7. dos Santos JB, Nienhuis J, Skroch P et al (1994) Comparison of RAPD and RFLP genetic markers in determining genetic similarity among *Brassica oleracea* L. genotypes. Theor Appl Genet 87:909–915
8. Maria D, Angela P, Chialexei L et al (2008) Characteristics of RAPD markers inbreeding of *Cucumis sativus* L. Roum. Biotechnol Lett 13:3843–3850
9. Khadari B, Breton C, Moutier N et al (2003) The use of molecular markers for germplasm management in a French olive collection. Theor Appl Genet 106:521–529
10. Tinker NA, Fortin MG, Mather DE et al (1993) Random amplified polymorphic DNA and pedigree relationships in spring barley. Theor Appl Genet 85:976
11. Mailer RJ, Scarth R, Fristensk B et al (1994) Discrimination among cultivars of rapeseed (*Brassica napus* L.) using DNA polymorphism amplified from arbitrary primers. Theor Appl Genet 87:697–704
12. Rajesh MK, Jerard BA, Preethi P et al (2014) Application of RAPD markers in hybrid

verification in coconut. Crop Breed Applied Biotechnol 14(1):36–41

13. Congiu L, Chicca M, Cella R et al (2000) The use of randomly amplified polymorphic DNA (RAPD) markers to identify strawberry varieties: a forensic application. Mol Ecol 9:229–232
14. Bligh HFJ (2000) Detection of adulteration of basmati rice with non-premium long grain rice. Int J Food Sci Technol 35:257–265
15. Adams RP, Demeke T (1993) Systematic relationships in junipers based on random amplified polymorphic DNA. Taxon 42:553–571
16. Wilkie SE, Issac PG, Slater RJ et al (1993) Random amplified polymorphic DNA (RAPD) markers for genetic analysis in allium. Theor Appl Genet 86:497–504
17. Isabel N, Tremblay L, Michaud M et al (1993) RAPDs as an aid to evaluate the genetic integrity of somatic embryogenesis-derived populations of *Picea mariana* (Mill.) B.S.P. Theor Appl Genet 86:81–87
18. Lewis PO, Snow AA (1992) Deterministic paternity exclusion using RAPD markers. Mol Ecol 1:155–160
19. Crawford DJ, Brauner S, Cosner MB et al (1993) Use of RAPD markers to document the origin of inter generic hybrid *Margyracaena skottsbergii* (Rosaceae) on the Juan Fernandez Islands. Am J Bot 80:89–92
20. Waugh R, Baird E, Powell W (1992) The use of RAPD markers for the detection of gene introgression in potato. Plant Cell Rep 11:466–469
21. Halima HS, Bahy AA, Tian-Hua H et al (2007) Use of random amplified polymorphic DNA analysis for economically important food crops. J Integr Plant Biol 49(12): 1670–1680
22. Hedrick P (1992) Shooting the RAPDs. Nature 355:679–680
23. Challahan LM, Weaver KR, Caetano-Anolles G et al (1993) DNA fingerprinting of turf grass. Int Turfgrass Soc Res J 7:761–767
24. Caetano-Anollés G, Gresshoff PM (1994) DNA amplification fingerprinting using arbitrary mini-hairpin oligonucleotide primers. Biotech 12:619–623
25. Michelmore RW, Paran I, Kesseli RV et al (1991) Identification of markers linked to disease-resistance genes by bulked segregant analysis: a rapid method to detect markers in specific genomic regions by using segregating populations. Proc Natl Acad Sci U S A 88:9828–9832
26. Martin GB, Williams JGK, Tanksley SD et al (1991) Rapid identification of markers linked to a pseudomonas resistance gene in tomato by using random primers and near-isogenic lines. Proc Natl Acad Sci U S A 88:2336–2340
27. Rafalski JA, Tingey SV (1993) Genetic diagnostics in plant breeding: RAPDs, microsatellites and machines. Trends Genet 9:275–280
28. Welsh J, McClelland M (1990) Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res 18:7213–7218
29. Welsh J, Honeycutt RS, McClelland M et al (1991) Parentage determination in maize hybrids using the arbitrarily primed polymerase chain reaction (AP-PCR). Theor Appl Genet 82:473–476
30. Caetano-Anollés G, Bassam BJ, Gresshoff PM et al (1991) DNA amplification finger printing using short arbitrary oligonucleotide primers. Biotech 9:553–557
31. Somsri S, Bussabakornkul S (2008) Identification of certain papaya cultivars and sex identification in papaya by DNA amplification fingerprinting (DAF). Acta Hort (ISHS) 787:197–206
32. Luro S (1995) DNA amplified fingerprinting, a useful tool for determination of genetic origin and diversity analysis in citrus. HortScience 30(5):1063–1067
33. Li G, Quiros CF (2001) Sequence-related amplified polymorphism (SRAP), a new marker system based on a simple PCR reaction: its application to mapping and gene tagging in brassica. Theor Appl Genet 103:455–546
34. Cifarelli RA, Gallitelli M, Cellini F et al (1995) Random amplified hybridization microsatellites (RAHM): isolation of a new class of microsatellite containing DNA clones. Nucleic Acid Res 23:3802–3803
35. Richardson T, Cato S, Ramser J et al (1995) Hybridization of microsatellites to RAPD: a new source of polymorphic markers. Nucleic Acids Res 23:3798–3799
36. Koniecyzn A, Ausubel FM (1993) A procedure for mapping *Arabidopsis* mutations using co-dominant ecotype-specific PCR-economically important pathogen based markers. Plant J 4:403–410
37. Jarvis P, Lister C, Szabo V et al (1994) Integration of CAPS markers into the RFLP map generated using recombinant inbred lines of *Arabidopsis thaliana*. Plant Mol Biol 24:685–687
38. Ranade SA, Farooqui N, Bhattacharya E et al (2001) Gene tagging with random amplified polymorphic DNA (RAPD) markers for molecular breeding in plants. Crit Rev Plant Sci 20(3):251–275
39. Gould AR (1986) Factors controlling generation of variability in vitro in: Vasil IK (ed) cell

culture and somatic cell genetics in plants, plant regeneration and genetic variability, 3rd edn. Academic Press, Orlando

40. Wang S, Chen X, Han F et al (2016) Genetic diversity and population structure of ginseng in China based on RAPD analysis. Open Life Sci 11(1):387–390
41. Rohela GK, Jogam P, Bylla P et al (2019) Indirect regeneration and assessment of genetic fidelity of acclimated plantlets by SCoT, ISSR, and RAPD markers in *Rauwolfia tetraphylla* L.: an endangered medicinal plant. Biomed Res Int 2019:3698742
42. Ali EM, Tohidfar M, Karimi M et al (2015) Determination of genetic uniformity in transgenic cotton plants using DNA markers (RAPD and ISSR) and SDS-PAGE. J Plant Mol Breed 3(2):36–43
43. Tingey SV, del Tufo JP (1993) Genetic analysis with random amplified polymorphic DNA markers. Plant Physiol 101:349–352
44. Powell W, Morgante M, Andre C et al (1996) The unity of RFLP, RAPD, AFLP and SSR (microsatellite) markers for germplasm analysis. Mol Breed 2:225–238
45. Bardacki F (2001) Random amplified polymorphic DNA (RAPD) markers. Turk J Biol 25:185–196
46. Chao S (2006) Application of molecular marker technologies on cereal crops improvement. Paper presented at the American oat workers conference, Fargo, ND
47. Jiang GL (2013) Molecular markers and marker-assisted breeding in plants. In: Sven BA (ed) Plant breeding from laboratories to fields. Intech, London
48. Shivashankar M (2014) Random amplified polymorphic DNA (RAPD) markers in anticancer drug plants. Int J Curr Microbiol App Sci 3(7):1091–1101
49. Kordrostami M, Rahimi M (2015) Molecular markers in plants: concepts and applications. Paper presented at conference on Genetics in the Third Millennium Vol. 13, pp 4024–4031
50. Selvakumari E, Jenifer J, Priyadharshini S et al (2017) Application of DNA fingerprinting for plant identification. J Acad Ind Res 5(10)
51. Nadeem MA, Nawaz MA, Shahid MQ et al (2018) DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. Biotechnol Biotechnol Equip 32(2):261–285
52. Arul S, Selvakumar R (2019) Genetic diversity and application of DNA markers in garden pea-review. Acta Sci Agric 3(2):153–161
53. Cho KH, Noh JH, Park SJ et al (2015) Development of sequence characterized amplified region markers for the identification of grapevine cultivars. Hort Sci 50(12):1744–1750
54. Cao X, Wu Z, Zhou R et al (2015) A novel random amplified polymorphic DNA-based strategy for genetic diversity analysis and identification of tomatoes. Genet Mol Res 14(1): 1650–1661
55. Tanikawa T, Takagi M, Ichii M et al (2002) Cultivar identification and genetic diversity in onion (*Allium cepa* L.) as evaluated by random amplified polymorphic DNA (RAPD) analysis. J Japan Soc Hort Sci 71(2):249–251
56. Vaio CD, Villano C, Marallo N et al (2015) Molecular analysis of native cultivars of sweet cherry in southern Italy. Hort Sci 42(3): 114–118
57. Besnard G, Breton C, Baradat P et al (2001) Cultivar identification in olive based on RAPD markers. J Amer Soc Hort Sci 126(6): 668–675
58. Arumugam T, Jayapriya G, Sekar T et al (2019) Molecular fingerprinting of the Indian medicinal plant *Strychnos* minor Dennst. Biotechnol Rep 21:00318
59. Lee YM, Ji Y, Kang YM et al (2016) Molecular authentication of Pinelliae tuber and its common adulterants using RAPD-derived multiplex sequence characterized amplified region (multiplex-SCAR) markers. Int J Clin Exp Med 9(1):40–50
60. Yang L, Khan MA, Mei Z et al (2014) Development of RAPD-SCAR markers for *Lonicera japonica* (*Caprifoliaceae*) variety authentication by improved RAPD and DNA cloning. Rev Biol Trop 62(4):1649–1657
61. Yang L, Fu S, Khan MA et al (2013) Molecular cloning and development of RAPD-SCAR markers for *Dimocarpus longan* variety authentication. Springerplus 2:501
62. Cheng J, Long Y, Khan MA et al (2015) Development and significance of RAPD-SCAR markers for the identification of *Litchi chinensis* Sonn. By improved RAPD amplification and molecular cloning. Electron J Biotechnol 18:35–39
63. Boyd M, Panoyan MA, Michael P et al (2019) Development and characterization of species-diagnostic ISSR and SCAR DNA markers for differentiating red maple (Acer rubrum) and silver maple (*A. saccharinum*). Genome 62:527–535
64. Moctezuma VE, Lopez AL, Pardo CVT et al (2018) Usefulness of three DNA-PCR techniques to differentiate Jalapeño pepper varieties. Indian J Biotechnol 17:527–532
65. Cheng JL, Li J, Qi YM et al (2016) Development of novel SCAR markers for genetic

characterization of *Lonicera japonica* from high GC-RAMP-PCR and DNA cloning. Genet Mol Res 15:10.4238

66. Mei Z, Zhang C, Khan AM et al (2015) Efficiency of improved RAPD and ISSR markers in assessing genetic diversity and relationships in *Angelica sinensis* (Oliv.) Diels varieties of China. Electron J Biotechnol 18(2):96–102
67. Paran I, Kesseli R, Michelmore R et al (1991) Identification of restriction fragment-length-polymorphism and random amplified polymorphic DNA markers linked to downy mildew resistance genes in lettuce, using near isogenic lines. Genome 34:1021–1027
68. Hoshi Y, Shirakawa J, Takeo M et al (2010) A molecular genetics of *Drosera spatulata* complex by using RAPD analysis. Chromosome Bot 5:23–26
69. Mudge J, Andersen WR, Kehrer RL et al (1996) A RAPD genetic map of *Saccharum officinarum*. Crop Sci 36(5):1362–1366
70. Padmakar B, Sailaja D, Aswath C et al (2015) Molecular exploration of guava (*Psidium guajava L.*) genome using SSR and RAPD markers: a step towards establishing linkage map. J Hort Sci 10(2):130–135
71. Moulin MM, Rodrigues R, Ramos HCC et al (2015) Construction of an integrated genetic map for *Capsicum baccatum* L. Genet Mol Res 14(2):6683–6694
72. Wanga G, Guoa Y, Zhaoa Y et al (2015) Construction of a molecular genetic map for hawthorn based on SRAP markers. Biotechnol Biotechnol Equip 29(3):441–447
73. Ferreira TGT, Penha HA, Zuchhi MI et al (2010) Outcrossing rate in sweet passion fruit based on molecular markers. Plant Breed 129:727–730
74. Trame AM, Coddington AJ, Paige KN et al (1995) Field and genetic studies testing optimal outcrossing in *Agave schottii*, a long-lived clonal plant. Oecologia 104(1):93–100
75. Teklewold A, Velaso L, Becker HC (2013) Estimation of outcrossing in Ethiopian mustard (*B. carinata*) using RAPD markers. Int J Plant Breed 7(1):1–11
76. Kobayashi M, Lin J, Davis J et al (2000) Quantitative analysis of avocado outcrossing and yield in California using RAPD markers. Sci Hortic 86:135–149
77. Hazem AO, Naheif EMM, Khaled AGA et al (2015) Inbreeding, outbreeding and RAPD markers studies of faba bean (*Vicia faba* L.) crop. J Adv Res 6:859–868
78. Sarmast MK, Salehi H, Ramezani A et al (2012) RAPD fingerprint to appraise the genetic fidelity of in vitro propagated *Araucaria excelsa* R. Br. var. glauca plantlets. Mol Biotechnol 50(3):181–188
79. Razaq M, Heikrujam M, Chetri SK et al (2013) *In vitro* clonal propagation and genetic fidelity of the regenerants of *Spilanthes calva* DC. Using RAPD and ISSR marker. Physiol Mol Biol Plants 19(2): 251–260
80. Kawiak A, Lojkowska E (2004) Application of RAPD in the determination of genetic fidelity in micro propagated *Drosera* plantlets. In Vitro Cell Dev Biol Plant 40(6):592–595
81. Sharma V, Belwal N, Kamal B et al (2016) Assessment of genetic Fidelity of in vitro raised plants in *Swertia chirayita* through ISSR, RAPD analysis and peroxidase profiling during organogenesis. Braz Arch Biol Technol 59:16160389
82. Baghel S, Bansal YK (2017) *In vitro* regeneration of *Guizotia abyssinica* Cass. And evaluation of genetic fidelity through RAPD markers. S Afr J Bot 109:294–307
83. Venkatachalam L, Sreedhar RV, Neelwarne B et al (2007) Micro propagation in banana using high levels of cytokinins does not involve any genetic changes as revealed by RAPD and ISSR markers. Plant Growth Regul 51:193–205
84. Alizadeh M, Singh S (2009) Molecular assessment of clonal fidelity in micro propagated grape (*Vitis* spp.) rootstock genotypes using RAPD and ISSR markers. Iranian J Biotechnol 7(1):37–44
85. Mohanty S, Joshi RS, Subudhi E et al (2012) Genetic stability assessment of micro propagated mango ginger (*Curcuma amada* Roxb.) through RAPD and ISSR markers. Res J Med Plants 6:529–536
86. Tiwari JK, Chandel P, Gupta S et al (2013) Analysis of genetic stability of *in vitro* propagated potato micro tubers using DNA markers. Physiol Mol Biol Plants 19(4):587–595
87. Thakur J, Dwivedi MD, Sourabh P et al (2016) Genetic homogeneity revealed using SCoT, ISSR and RAPD markers in micro propagated *Pittosporum eriocarpum* Royle- an endemic and endangered medicinal plant. PLoS One 11(7):0159050
88. Yadav A, Kothari SL, Kachhwaha S et al (2019) In vitro propagation of chia (*Salvia hispanica* L.) and assessment of fidelity using random amplified polymorphic DNA and inter simple sequence repeat molecular markers. J Appl Biol Biotechnol 7(1):42–47
89. Borsai O, Clapa D, Fira A et al (2018). Evaluation of the genetic fidelity of in vitro-propagated blackberry plants (*Rubus*

fruticosus L.) using molecular markers. Paper presented at XXX international horticultural congress, Istanbul, Turkey. 12–16 August, 2018

90. Zakiyah N, Handoyo T, Kim KM et al (2019) Genetic diversity analysis of Indonesian aromatic rice varieties (*Oryza sativa* L.) using RAPD. J Crop Sci Biotechnol 22:55–63
91. Patwardhan A, Ray S, Roy A et al (2014) Phylogenetics and evolutionary biology molecular markers in phylogenetic studies - a review. Phylogenetics Evol Biol 57
92. Kibria K, Begum S, Islam M et al (2009) Molecular marker based genetic diversity analysis in aromatic rice genotypes using SSR and RAPD markers. Int J Sustain Crop Prod 4
93. Patel HK, Fougat RS, Kumar S et al (2015) Detection of genetic variation in *Ocimum* species using RAPD and ISSR markers. 3. Biotech 5:697
94. Konzen ER, Peron R, Ito MA et al (2017) Molecular identification of bamboo genera and species based on RAPD-RFLP markers. Silva Fennica 51(4):1691
95. Qin J, Yang Y, Jiang J et al (2012) Comparison of lignocellulose composition in four major species of *Miscanthus*. Afr J Biotechnol 11
96. Moulin MM, Rodrigues R, Gonçalves LSA et al (2012) A comparison of RAPD and ISSR markers reveals genetic diversity among sweet potato landraces (*Ipomoea batatas* (L.) lam.). Acta Sci Agron 34(2):139–147
97. Ma X, Chen SY, Bai SQ et al (2012) RAPD analysis of genetic diversity and population structure of *Elymus sibiricus* (Poaceae) native to the southeastern Qinghai-Tibet plateau, China. Genet Mol Res 11(3):2708–2718
98. Muzila M, Werlemark G, Ortiz R et al (2014) Assessment of diversity in *Harpagophytum* with RAPD and ISSR markers provides evidence of introgression. Hereditas 151(4-5): 91–101
99. Bhatt J, Kumar S, Patel S et al (2017) Sequence-related amplified polymorphism (SRAP) markers based genetic diversity analysis of cumin genotypes. Ann Agrar Sci 15:434–438
100. Ge D, Daizhen Z (2015) Application of sequence-related amplified polymorphism to genetic diversity analysis in *Limonium sinense*. J Genet 94:35–38
101. Hou S, Zhu GZ, Li Y, Li WX et al (2018) Genome-wide association studies reveal genetic variation and candidate genes of drought stress related traits in cotton (*Gossypium hirsutum* L.). Front Plant Sci 9:1276
102. Arif M, Aristya G, Kasiamdari R (2019) Genetic diversity of strawberry cultivars in Banyuroto, Magelang, Indonesia based on cleaved amplified polymorphic sequence. 10:13057
103. Sharafi A, Abkenar A, Sharafi A (2017) Molecular genetic diversity assessment of citrus species grown in Iran revealed by SSR, ISSR and CAPS molecular markers. J Sci Res 2(22):22–27
104. Taspinar MS, Guleray A, Nalan Y et al (2009) Evaluation of selenium effect on cadmium genotoxicity in *Vicia faba* using RAPD. J Food Agric Environ 7(3&4):857–860
105. Rai P, Dayal S (2009) RAPD-PCR based analysis of genetic variation induced in *Triticum aestivum* under chromium stress. Int J Adv Sci Eng Inf Technol 4(4):117–120
106. Sameer H, Qari M (2010) DNA-RAPD fingerprinting and cytogenetic screening of genotoxic and anti-genotoxic effects of aqueous extracts of *Costus speciosus* (Koen.). JKAU Sci 22(1):133–152
107. Fu Y (2006) Redundancy and distinctness in flax germplasm as revealed by RAPD dissimilarity. Plant Genet Res 4(2):117–124
108. Virk PS, Newbury HJ, Jackson MT et al (1995) The identification of duplicate accessions within a rice germplasm collection using RAPD analysis. Theor Appl Genet 90:1049
109. Vekariya S, Taviad K, Acharya RN et al (2017) Development of random amplified polymorphic DNA markers for authentication of *Croton tiglium* Linn. J Phytopharmacol 6(3): 164–166
110. Shinde VM, Dhalwal K, Mahadik KR et al (2007) RAPD analysis for determination of components in herbal medicine. Evid Based Complement Alternat Med 4:21–23
111. Weder JK (2002) Identification of plant food raw material by RAPD-PCR: legumes. J Agric Food Chem 50(16):4456–4463
112. Sarwat M, Srivastava S, Khan TH et al (2016) RAPD and ISSR polymorphism in the medicinal plants: *Ocimum sanctum, O basilicum* and *O gratissimum*. IJPPR 8(8):1417–1424
113. Solanki S, Richards J, Ameen G et al (2019) Characterization of genes required for both Rpg1 and rpg4-mediated wheat stem rust resistance in barley. BMC Genomics 20:495
114. Li Y, Zou J, Ma L et al (2012) Development of head smut resistance-linked sequence characterized amplified regions markers in sorghum. Int J Agric Biol:14
115. Barua UM, Chalmers KJ, Thomas WT et al (1993) Molecular mapping of genes determining height, time to heading, and growth

habit in barley (*Hordeum vulgare*). Genome 36(6):1080–1087

116. Baird E, Cooper-Bland S, Waugh R et al (1992) Molecular characterization of inter- and intra-specific somatic hybrids of potato using randomly amplified polymorphic DNA (RAPD) markers. Mol Gen Genet 233(3): 469–475
117. Yaycili O, Alikamanoglu S (2012) Induction of salt-tolerant potato (*Solanum tuberosum* L.) mutants with gamma irradiation and characterization of genetic variations via RAPD-PCR analysis. Turk J Biol 36:405–412
118. Barakat MN, Abdel Fattah RS, Badr M (2010) In vitro mutagenesis and identification of new variants via RAPD markers for improving *Chrysanthemum morifolium*. African J Agric Res 5(8):748–757
119. Penner GA, Bush A, Wise R (1993) Reproducibility of random amplified polymorphic DNA (RAPD) analysis among laboratories. PCR Methods Appl 2:341–345
120. Aly MAM, El-Hewiety AY (2009) DNA fingerprint of UAE grown date palm varieties. In: proc. 10th annual UAE university research conference. United Arab Emirates University Al-Ain, UAE
121. Garcia AAF, Benchimol LL, Barbosa AMM (2004) Comparison of RAPD, RFLP, AFLP and SSR markers for diversity studies in tropical maize inbred lines. Genet Mol Biol 27:579–588
122. Sedra MH, Lashermes P, Trouslot P et al (1998) Identification and genetic diversity analysis of date palm (*Phoenix dactylifera* L.) varieties from Morocco using RAPD markers. Euphytica 103:75
123. Trifi M, Rhouma A, Marrakchi M et al (2000) Phylogenetic relationships in Tunisian date-palms (*Phoenix dactylifera* L.) germplasm collection using DNA amplification fingerprinting. Agronomie 20:665–671
124. Doyle JJ, Doyle LJ (1990) Isolation of plant DNA from fresh tissue. Focus 12:13–15
125. Rohlf FJ (1998) NTSYS-pc numerical taxonomy and multivariate analysis system. Version 2.02. Exeter publications Setauket, New York
126. Wolff K, Schoen ED, Peters-Van Rijn J (1993) Optimizing the generation of random amplified polymorphic DNA in chrysanthemum. Theor Appl Genet 86:1033–1037
127. Demeke T, Adams RP (1994) The use of RAPD-PCR analysis in plant taxonomy and evolution. In: Griffin HG, Griffin AM (eds) PCR technology: current innovations. CRC Press, Boca Raton, FL
128. Skroch P, Nienhuis J (1995) Qualitative and quantitative characterization of RAPD variation among snap bean (*Phaseolus vulgaris*) genotypes. Theor Appl Genet 91:1078–1085
129. Lamboy WF (1994a) Computing genetic similarity coefficients from RAPD data: the effects of PCR artifacts. PCR Methods Appl 4:31–37
130. Lamboy WF (1994b) Computing genetic similarity coefficients from RAPD data: correcting for the effects of PCR artifacts caused by variation in experimental conditions. PCR Methods Appl 4:38–43
131. Hernandez P, Martin A, Dorado G (1999) Development of SCARs by direct sequencing of RAPD products: a practical tool for introgression and marker- assisted selection of wheat. Mol Breed 5:245–253
132. Davis MJJ, Bailey CS, Smith CK (1997) Increased informativeness of RAPD analysis by detection of microsatellite motifs. BioTechniques 23:285–290

Chapter 14

Inter-Simple Sequence Repeats (ISSR), Microsatellite-Primed Genomic Profiling Using Universal Primers

Chrissen E. C. Gemmill and Ella R. P. Grierson

Abstract

Inter-simple sequence repeat (ISSR) markers are highly polymorphic, relatively easy to develop, and inexpensive compared to other methods and have numerous applications. Importantly, the same ISSR primers can potentially be used universally across plant phylogenetic diversity. The basic technique of ISSRs is flexible and can be modified with options for implementation for a broad range of projects and budgets. Ranked in increasing order of technical demand and costs, these are manual agarose and manual polyacrylamide with silver staining and automated using fluorescently labeled primers and capillary electrophoresis. Overall manual agarose-based ISSRs are a sound, safe, easy, and low-cost method for reliably inferring plant genetic diversity. Here, we provide detailed protocols to undertake this fingerprinting method and provide guidance to the literature for the many options available for this technique.

Key words Conservation, Cultivar, Dominant marker, Genetic diversity, ISSR, Molecular identification, Phylogenetic relationships, Species delimitations, Taxonomy

1 Introduction

The term DNA fingerprinting was coined by Jeffreys [1] to reflect the unique multilocus genotype profiles of minisatellites observed via DNA hybridization analyses. This method was quickly adapted [2, 3] to take advantage of the polymerase chain reaction (PCR) [4, 5] and has catalyzed a diverse range of fingerprinting techniques including randomly amplified polymorphic DNA (RAPDs) [6] (Chapter 13), amplified fragment length polymorphisms (AFLPs) [7] (Chapter 12), and inter-simple sequence repeats (ISSRs) [8, 9]. ISSRs are a microsatellite-directed polymerase chain reaction technique and are also called anchored microsatellite-primed PCR (AMP-PCR). This approach takes advantage of the ubiquitous microsatellite loci distributed throughout eukaryotic genomes (see review of Bruford and Wayne [10]). These markers are hypervariable, arbitrarily amplified dominant markers that randomly

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_14,

target multiple regions within the genome simultaneously. The targets for amplification are the variable regions between identical inverted microsatellites. Multiple priming sites occur in the genomes of most organisms; hence, the production of multiple amplicons of varying lengths for a given reaction produces highly variable multilocus DNA fingerprints. By necessity, the primer must anneal to the repetitive elements on each of the two complementary strands of DNA within about 1 kB of each other. Lack of polymerization results from either a lack of one priming site or both priming sites; it is not possible to determine which. The resulting fragments have been shown to be inherited in a Mendelian fashion [8]; however, this is not likely the case for all loci. One issue inherent in DNA fingerprinting analyses is that co-migrating fragments are assumed to be homologous. Homology of fragments is assumed, but amplicons can co-migrate due to convergence [11, 12]. Furthermore, since these markers are dominant and genotypes cannot be inferred as in codominant markers, applying standard population genetic analyses is problematic as allele frequencies cannot be estimated [13].

ISSRs have become a very popular method globally to probe organisms for hidden genetic diversity without any prior knowledge of the genome. A search on the Scopus database on December 2, 2019, for "ISSR AND plant" returned 2502 articles for the years 1996–2019 and is on the rise. ISSRs have been applied to a multitude of questions and have made detailed comparisons to other methods [14–23]. Some applications include identification of plant cultivars, medicinal plants and constituents of products, and invasive plants; taxonomic identification particularly between closely related taxa and cryptic taxa; genetic mapping and fidelity; comparison of levels of genetic variation of in situ versus ex situ plants and populations for conservation and restoration management; and even probe genetic variability of plants raised in space [24]. Kumar et al. [25] suggest that ISSRs may be useful to delimit species where other markers, such as those used in DNA barcoding, have failed to provide adequate resolution.

As with other molecular techniques, there are advantages and disadvantages associated with ISSRs, which need to be weighed by the researcher on a case-by-case basis. Other considerations will include taxonomic range, discriminatory power required, reproducibility, technical difficulty, budget, and ease of interpretation (see also [14, 26]). RAPDs, AFLPs (Chapters 12 and 13), and ISSRs do not require any previous knowledge of the genome and hence can be applied to non-model systems easily and require only small amounts of DNA. ISSRs are on par with RAPDs with a low level of difficulty and low cost yet are more highly reproducible [27]. ISSRs tend to exhibit fewer loci than the technically more demanding AFLPs yet often produce congruent results with much less cost [21]. However, AFLPs may provide finer resolution of the

population genetic structure than ISSRs; this hurdle may be able to be overcome by adding more ISSR primers to an analysis. For many projects, ISSRs will provide appropriate levels of polymorphism and resolution, allowing for more primers to be assessed and/or more individuals to be included for a given budget. The number of primers used, usually three to ten, will depend on the level of polymorphism and the number of amplicons resolved per primer. Overall manual agarose-based ISSR is a sound, safe, easy, and low-cost method for reliably inferring genetic diversity across the diversity of plants.

Potentially, a number of different materials (e.g., fresh, dried in silica, frozen −80 °C, herbarium specimens) could be used, but this will vary with the plant group as well as collection, preservation, and storage methods (Chapters 3 and 4). Conducting preliminary analyses is key to sorting out issues early on in this process. The ideal material to use for ISSRs is fresh, clean, and young leaves; however, this is not always logistically feasible. The next best materials would be those collected into silica gel. As with all other molecular studies when collecting field and/or botanical garden specimens, lodging herbarium vouchers and/or obtaining accession information, respectively, is essential. Herbarium and botanic garden accession information should be reported along with the sample information. Acknowledging all permitting agencies is recommended.

Many studies report using CTAB (hexadecyltrimethylammonium bromide) [28, 29] or modifications thereof, such as treatment with RNase. Deng et al. [30] recently compared CTAB DNA extraction methods. Today, commercially available kits provide high quality and reasonable, and possibly consistent, quantities of DNA and are much easier, faster, and technically less demanding than CTAB extractions. Quality and purity of DNA can be assessed qualitatively via gel electrophoresis and/or quantitatively through spectrophotometry. Some authors suggest adjusting all DNA extractions to a concentration of 50 ng [31].

One important but often neglected step is screening the DNA extracts for contamination by nontarget epiphytes [32] and endophytes [33, 34]. To check DNA purity, amplify the extracts for the internal transcribed spacer (ITS) region [35, 36] using relaxed protocols (e.g., lower than normal annealing temperatures to encourage amplification on nontarget DNA); do not use a cpDNA maker for screening as you may miss fungal contaminants. A single strong band is indicative of non-contaminated DNA, whereas multiple bands indicate contamination. These contaminated samples cannot be used for ISSRs. An alternative to this is to check only the aberrant samples that appear to have more fragments than other conspecific samples. In our experience, the multilocus profiles of contaminated DNA are not completely additive but will differ from non-contaminated samples.

A single oligonucleotide primer, usually ≤25-mer, containing a microsatellite repeat is used in the PCR. The primers can be anchored at the 5′ or 3′ end with a single nucleotide or a short degenerate sequence. Archibald et al. [37] recommend using a 3′ anchor and avoiding redundancy of primer selection [38] or primers that may self-anneal such as CG and AT. Arens et al. [39] suggest avoiding tetra-nucleotide primers as they may not be evenly distributed within the genome. Many papers refer to the University of British Columbia's (UBC) Primer Set 9, and while UBC no longer synthesizes these primers nor maintains a list of the primers, Prince [40] included a full table of these primers along with melting temperatures and comments.

Use of negative controls to check for PCR contamination is standard in PCR. Positive controls are used to check reproducibility between runs, and independent repeat runs to check reproducibility overall. PCR can be optimized in a number of ways. PCR enhancers commonly used include bovine serum albumin (BSA) and dimethyl sulfoxide (DMSO) or 1,2-propanediol. Different concentrations of $MgCl_2$ can be trialed during PCR optimization. Testing primers with a temperature gradient is also important. Using a touchdown protocol may enhance clarity and reproducibility [41].

ISSRs can be conducted manually with agarose or polyacrylamide gels or automated with capillary gel electrophoresis and fluorescently labeled primers. Agarose gels are the least technically demanding and also likely the safest when employing a nontoxic nucleic acid staining solution in lieu of ethidium bromide. However, resolution is lowest among the methods. Goulao and Oliveira [42] provided detailed protocols for polyacrylamide gel electrophoresis and detection by silver staining. Increasingly, authors [37, 38, 43–45] have used fluorescently labeled primers and separated the fragments via capillary electrophoresis on automated DNA sequencing instruments, which will generate the most accurate fragments sizes. This method is the most sensitive, producing more fragments [16, 46, 47], but will be costlier (fluorescently labeled primers, cycle sequencing, running on automated sequencer). These costs may be somewhat offset by the faster scoring and analysis. As with all ISSR protocols, nonspecific amplification can be an issue. Note also that fingerprinting patterns can differ between manual and automated methods and between labeled and unlabeled products. Regardless of final method chosen, initial screening of primers can be done with standard oligonucleotides and agarose gels to keep costs down. Screen as many primers as possible and assess clarity, reproducibility, and number of polymorphisms. As an example, Grierson [48] assessed species delimitations and genetic variation within and among three endemic species of New Zealand *Sophora* L. (Fabaceae), *Sophora microphylla* Aiton, *Sophora prostrata* Buchanan, and *Sophora tetraptera* JF Mill.

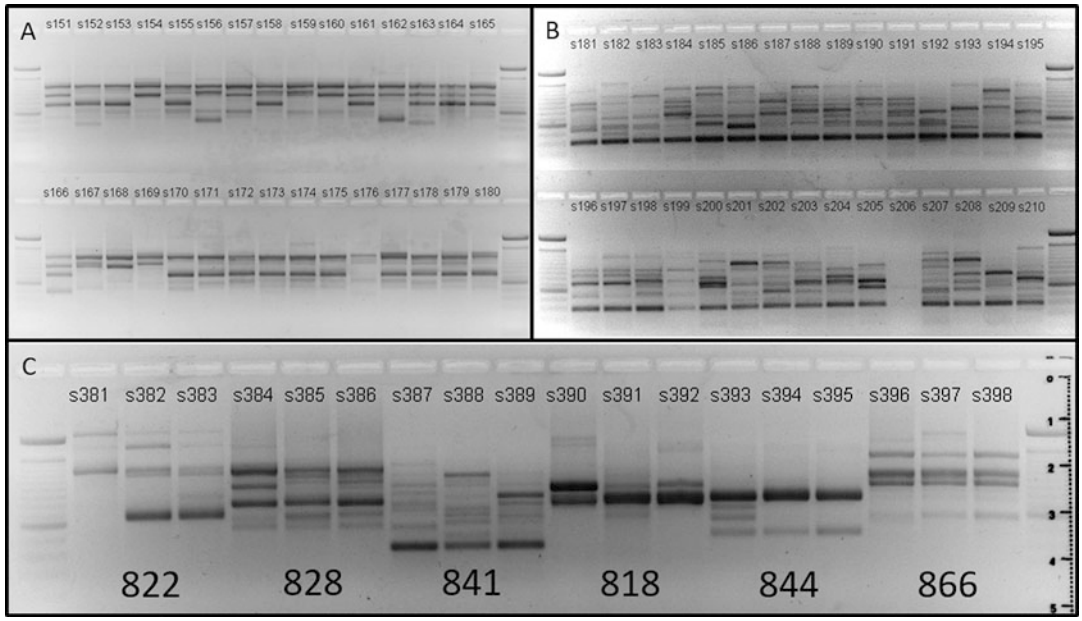


Fig. 1 ISSR products of endemic New Zealand *Sophora* species visualized using 2% agarose gels stained with ethidium bromide. All primers are from UBC set number 9. Each gel has two origins and the first and last lanes of all gels are a 100-kb ladder. The numbers above each lane are the reaction tube numbers. (**a**) primer 828 $((TG)_8A)$. (**b**) primer UBC841 $((GA)_8YC$, where Y is C or T). For (**a**) and (**b**), *S. prostrata*, origin 1, lanes 2–16 and origin 2, 19–22; *S. tetraptera*, origin 2, lanes 23–27; *S. microphylla,* origin 2, lanes 28–33. (**c**) test of reproducibility using one sample for each of *S. prostrata*, *S. tetraptera*, and *S. microphylla*, respectively, with primers 822 $((TC)_6A)$, 828, 841, 818 $((CG)_8A)$, 844 $((CT)_8RA$, where R is A or T), and 866 $((CTC)_6)$

Primers used included three 3′ anchored dinucleotide repeats (UBC818, 822, 828), two 3′ anchored degenerate dinucleotide repeats (UBC841, 844), and one nonanchored trinucleotide repeat (UBC866). To evaluate the performance of each primer at both inter- and intraspecific levels, the strategy employed a subset of the taxa that represented the broad diversity of these taxa and included multiple individuals per species (Fig. 1). Amplicons resulting from 3′-anchored dinucleotide and unanchored trinucleotide primers were clear, reproducible, and scorable from 2% agarose gels.

Scoring of manual ISSR profiles can be done by eye or using an automated scoring program; see Crawford and Mort [49] for discussion. Fragments are scored as present (1) or absent (0) producing a diallelic data matrix representing each fragment scored for each locus, across all loci, for each sample. Consistency of approach is key, including limiting scoring to fragments within a certain size range and of a certain intensity of brightness. Select an exclusion threshold for using missing data, e.g., samples or primers with >5% missing data are excluded from the analysis. The resulting binary/diallelic matrix is then subjected to a variety of analyses including estimates of common genetic parameters (percent of loci

polymorphic, genetic diversity, etc.), analysis of molecular variance (AMOVA), principal component analysis (PCA), and clustering algorithms such as neighbor joining (NJ) and the unweighted pair group method (UPGMA). Methods that avoid shared absences [50, 51] may be best, as shared absences are more likely to be homoplastic than shared presences. Early on, Lynch and Milligan [13] outlined the issues of using dominant markers for estimating traditional population genetic parameters. Hollingsworth and Ennos [52] examined issues with the analysis of dominant markers and importantly the effect of the number of loci on the topology of NJ analyses. Nelson and Anderson [53] recently discussed the issues related to the number of loci needed for AMOVA as that implemented in the software Arlequin [54, 55] was moderate, e.g., 30, while STRUCTURE, a Bayesian clustering algorithm [56–58], required >90 and is more sensitive to unequal sampling. Hence, it is important to determine the type of analyses that will be conducted before sampling. Increasing the number of ISSR loci and/or samples is likely to increase the resolution of genetic structure.

Below, we describe the implementation of manual agarose ISSR analyses as this method can be implemented in most general molecular biology laboratories. The major steps are experimental design including sampling, collection of vouchers and curation and storage of materials, DNA extraction, screening of primers, PCR of all primers using a set regime, scoring, and analysis. Detailed protocols for polyacrylamide gels with silver staining have been presented by Goulao and Oliveira [42] and automated fluorescently labeled primers with capillary electrophoresis have presented by Prince [40].

2 Materials

All solutions are made with Milli-Q water or molecular grade water, and all chemicals are analytical reagent grade.

2.1 Reagents

1. Plant DNA isolation kit such as Bioline ISOLATE II Plant DNA Kit. Alternatively, CTAB extraction as described in Doyle and Doyle [29]. CTAB buffer: 100 mM Tris–HCl, 1.4 M NaCl, 30 mM EDTA, 2% (w/v) hexadecyltrimethylammonium bromide.
2. Liquid nitrogen or sterile fine-grained sand for homogenization of plant tissues.
3. RNase if using CTAB method.
4. ISSR primers diluted in water or TE buffer to a 10-μM concentration (*see* **Note 1**).

5. TE buffer: 10 mM Tris–HCl, 1 mM EDTA.
6. *Taq* polymerase with PCR buffer (*see* **Note 2**).
7. Electrophoresis buffer: 1x TAE (40 mM Tris, 20 mM acetic acid, 1 mM EDTA), 1x TBE (89 mM Tris, 89 mM boric acid, 2 mM EDTA), or 1x SB (5 mM sodium borate) (*see* **Note 3**).
8. Milli-Q water.
9. Agarose.
10. Nucleic acid stain, such as RedSafe™ (*see* **Note 4**).
11. 100 bp DNA ladder, such as Invitrogen™ TrackIt DNA ladder.
12. Loading buffer (6×): 30% glycerol, 5 mM EDTA, 0.15% bromophenol blue, 0.15% xylene cyanol. Use at final 1x concentration (use only if *Taq* polymerase mix does not contain some already).

2.2 Equipment

1. Mortars and pestles for homogenization of leaf tissue or other methods of tissue disruption.
2. Bench microcentrifuge.
3. Thermo-mixer (kit) or water bath (CTAB).
4. Vortex mixer.
5. Optional: spectrophotometer or Qubit for quantification of DNA and required consumables/standards.
6. Laminar flow hood to prepare the PCR.
7. Programmable thermal cycler such as Eppendorf Gradient with 96 tube capacity.
8. Microwave.
9. Standard agarose gel electrophoresis apparatus with power supply.
10. UV imaging equipment such as Alphaimager.
11. Optional: software to automatically score bands.

3 Methods

3.1 DNA Extraction

Follow the manufacturer's directions for kit extractions; we use ca. 5 mm × 5 mm fresh or dried leaf tissue. For CTAB extraction, use up to 5 g of fresh leaf material and grind to a powder in liquid nitrogen. Samples extracted with CTAB should be treated with RNase. The quality and quantity of the DNA extracts can be assessed qualitatively via agarose gel electrophoresis or quantitatively via spectrophotometry (*see* **Notes 5–8** for DNA extraction tips).

3.2 PCR

The following method uses a *Taq* polymerase mix that includes all reagents (dNTPs, $MgCl_2$, loading buffer) except primer, PCR enhancers, and DNA. The PCR is set up at room temperature in a laminar flow hood (*see* **Notes 9** and **10** for tips on optimizing your PCR).

1. As a guide, use a total volume of 15 μL for the PCR, in 0.2 mL PCR tubes with final concentrations as follows: 0.25 μM of primer, 1× PCR mix, 1.0 μL of DNA (approx. 5–10 ng/μL), 0.05 U *Taq* polymerase (*see* **Notes 11**–**13**).
2. Include a negative control containing no DNA in each PCR run to check for contamination. You can also include a positive control for reproducibility in each run—a sample known to amplify well. Also *see* **Note 14** for other reproducibility considerations.
3. Transfer the tubes to a thermocycler using a program of 5 min at 94 °C for initial denaturation; 35–40 cycles of denaturation 45 s at 94 °C, annealing for 45 s at annealing temperature for specific primer, and extension for 90 s at 72 °C, followed by 5 min at 72 °C for the final extension. Keep tubes at 4 °C until the initiation of gel electrophoresis (*see* **Notes 15** and **16**).

3.3 Gel Electrophoresis

This method uses a standard agarose gel, but other options can be explored for increased resolution (*see* **Note 17**). Standard oligos can be used and visualized on 2–3% agarose gels. The following is to prepare 240 mL of 2% agarose gel; calculate the volume needed by multiplying the surface area of the gel tray by the desired thickness (approx. 3–5 mm).

1. Set up gel mold with comb(s), level on surface. Depending on the specific apparatus, two origins may be used on a large gel (*see* **Note 18**).
2. Prepare a 2% (w/v) gel, weigh out 4.8 g of agarose into a flask, and add 240 mL of electrophoresis buffer. Dissolve agarose completely using a microwave (*see* **Note 19**).
3. Let solution cool until temperature is below 60 °C; then add nucleic acid stain as per manufacturer's directions. Swirl gently to mix.
4. Pour the gel into the gel tray, add combs, and let set for at least 30 min (*see* **Note 20**).
5. Place gel into electrophoresis tank, and cover by 5 mm with electrophoresis buffer.
6. Load appropriate ladder into the first and last lanes as a size reference; 3 μL of Invitrogen 100 bp ladder is sufficient in most cases (*see* **Notes 21** and **22**).

7. Add loading buffer to PCR product if needed (to a final concentration of 1x).
8. Load 15 μL PCR product into each well (*see* **Note 23**).
9. Run the gel at approx. 125 V for 2 h. Running time will depend on agarose concentration and size of gel (*see* **Note 24**).
10. View and photograph the gel using an imager. Save the file at highest resolution possible.

3.4 Scoring

A binary data matrix is created for the sample set by scoring presence (1) and absence (0) of specific bands across samples for each primer using the ladder to determine size. Bands that fail reproducibility tests, or are very faint compared to others, should be scored as absent. Subtle differences in band intensity are not usually considered. Gels can be rescored at least twice to test the consistency of the researcher. A subset of samples can be reamplified and run on a gel next to the original products to ensure reproducibility (*see* **Note 14**).

4 Notes

1. See Prince [40] for list of UBC primers or Wolfe [59]. Once rehydrated, aliquot primers into multiple tubes to prevent cross contamination of primers. For all steps, employ good molecular biology techniques and anti-contamination protocols used in standard PCR.
2. For ease, consistency, and minimization of errors, select a *Taq* polymerase that includes nucleotides, $MgCl_2$, and loading buffer such as MyTaq™ Red Mix. Each *Taq* may require different concentrations of $MgCl_2$; hence, this is one reagent that might be optimized during trials.
3. We have been able to use 0.5× TBE to save on costs. SB may be a less expensive option and is easier to make and has provided excellent results. We make up 10× SB (wearing a mask) and dilute to 1× as needed.
4. Ethidium bromide can also be used, but it is a known mutagen and carcinogen and has largely been replaced in most labs by less toxic nucleic acid stains such as RedSafe™. If you do use ethidium bromide, be sure to observe appropriate safety protocols.
5. If there are any concerns of epiphytes, carefully wash leaves in distilled water and dry prior to extraction or for storage in silica or at −30 to −80 °C.
6. Ideally, all extracts would be checked for nontarget DNA contamination; however, this adds time and cost to the analyses.

An alternative would be to check any anomalous samples with additional or errant banding patterns for contamination via PCR for ITS. We use ITS4 [35] and the higher-plant primer ITS5 HP [60] to amplify the entire ITS region as detailed in Carter et al. [61]. ITS fragments vary in size plant species to species, as do fungal contaminants, which are usually smaller than those of plants. An ITS product with two bands or a single smaller band indicates fungal contamination, and hence, the sample cannot be used for ISSRs.

7. For DNA extraction from dried material, the lysis period can be increased; we generally use 3 h for samples stored in silica gel or herbarium specimens when using the Bioline ISOLATE II Plant DNA Kit. Before sampling from herbarium specimens, be sure to obtain proper permission from the curator.
8. DNAs can be quantified using spectrophotometry or "by eye"; in this case, adjust DNAs to have similar fluorescence intensities under standardized conditions (*see* [58]). Too much DNA can inhibit PCR, as can secondary compounds that coprecipitate with the DNA; here, try diluting an aliquot of DNA 1:10 or use a standard protocol, e.g., ethanol precipitation, phenol:chloroform:isopropanol 25:24:1 to clean the DNA. Use prescribed safety protocols when working with phenol.
9. As a first step, conduct a comprehensive survey of the literature for recent journal articles on ISSRs, particularly those of the same genus or family if possible to get a start on which primers to screen first. It is imperative to spend time screening numerous primers and optimizing the PCR.
10. Once you have begun the full-scale analyses, you will not be able to make any further changes to the established protocol as they may cause deviations in the PCR results; hence, diligence in preliminary optimization and screening is critical. Minimize changes within the laboratory such as reagents and thermal cyclers. Ideally, it is best to conduct all of the PCRs within the shortest time frame possible. Make sure that you use the same reagents, even the same lots if possible. In some labs, water quality can change over the season, as can other environmental factors.
11. Always include a few additional reactions in your master mix calculations to account for pipetting errors. We use the guide of one extra sample per ten samples.
12. The volume/quantity of DNA in the PCR can be reduced if the fragments are overwhelmingly bright and hence possibly obscuring other bands.
13. Additives that can be used to improve your PCR reaction can include 0.1% BSA and/or up to 5% DMSO, or 0.86 M 1,2-propanediol.

14. Reproducibility
    (a) Repeat PCRs in duplicate or triplicate [40] applying a majority rule approach when scoring individual bands. Run products side by side to assess reproducibility and/or include PCRs run previously on each run for each primer to make sure the banding patterns for each primer are consistent run to run.
    (b) To have ample sample for loading on multiple gels as controls, the total volume of the PCR can be increased. Consider including a sample that is not closely related to your focus group, as another method to monitor reproducibility.
    (c) Only individuals that can be scored with low percent of missing data, for example, <5% missing data per individual, should be included in the analyses.
15. You may need to use different annealing temperatures for different primers. Conduct a gradient PCR to determine the optimal annealing temperatures that give clear, strong, and well-separated fragments. The annealing temperature may be above the calculated T_m. See Bornet and Branchard [27].
16. Optimization of the thermal cycling may include changes in the length of each step (denature, anneal, polymerization) and/or the number of cycles. A touchdown protocol can also be trialed.
17. Agarose gels (2–3%) will likely provide adequate resolution in most systems, but polyacrylamide or capillary electrophoresis will provide higher levels of resolution. Agarose gels stained with alternatives to ethidium bromide are the safest. Costs, as well as health risks and impacts on the environment, may be increased with silver-stained polyacrylamide gels. Capillary electrophoresis will provide the highest level of detail but is also the most expensive option.
18. Trial use of wide/broad thin (e.g., 25 wells, 1 mm thick) versus narrow thick gel combs (e.g., 50 wells, 1.5 mm thick) to assess method for best visualization of fragments.
19. When dissolving agarose gel, the liquid can superheat. It is best to heat in short bursts followed by gentle swirling with flask pointed away from face, wearing appropriate protection.
20. We use large gels (230 mm × 210 mm, CLP System) and two combs with fine but thick teeth (50 wells, 1.5 mm thick) that produce 100 wells per gel. This allows us to run up to 94 samples, a negative control, and two ladders per origin at a time. This works particularly well with a 96-well thermal cycler.
21. You can also include a ladder in the center lane of the gel to aid sizing of fragments if your gel is very wide.

22. To differentiate bands of similar molecular weight, gels can be run longer or rerun with a higher-resolution ladder or higher percentage of agarose.
23. Do not overload the wells of the gel, as this could result in fragments being overwhelmingly bright and possibly obscuring other bands. Background smear may also be reduced by loading less product into the wells.
24. Time and voltage required can be varied—keep an eye on the loading dye to ensure your fragments do not run off the gel and that your gels are not running too hot. For long runs, it is best to refresh at least 50% of the electrophoresis buffer each run.

References

1. Jeffreys AJ, Wilson V, Thein S (1985) Hypervariable "minisatellite" regions in human DNA. Nature 314:67–73
2. Welsh J, McClelland M (1990) Fingerprinting genomes using PCR with arbitrary primers. Nucleic Acids Res 17:7213–7218
3. Meyer W, Mitchell TG, Freedman EZ et al (1993) Hybridization probes for conventional DNA fingerprinting used as single primers in the polymerase chain reaction to distinguish strains of *Cryptococcus neoformans*. J Clin Microbiol 31:2274–2280
4. Mullis KF, Faloona F, Scharf S et al (1986) Specific enzymatic amplification of DNA in vitro: the polymerase chain reaction. Cold Spring Harb Symp Quant Biol 51:263–273
5. Saiki RK, Gelfand DH, Stoffel S et al (1998) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239:487–491
6. Williams JGK, Kubelik AR, Livak KJ et al (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18:6531–6535
7. Vos P, Hogers R, Bleeker M et al (1995) AFLP: a new technique for DNA finger- printing. Nucleic Acids Res 23:4407–4414
8. Zietkiewicz E, Rafalski A, Labuda D (1994) Genomic fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. Genomics 20:176–183
9. Gupta M, Chyi YS, Romero-Severson J et al (1994) Amplification of DNA markers from evolutionarily diverse genomes using single primers of simple-sequence repeats. Theor Appl Genet 89:998–1006
10. Bruford MW, Wayne RK (1993) Microsatellites and their application to population genetic studies. Curr Opin Genet Dev 3:939–943
11. Hillis DM (1994) Homology in molecular biology. In: Hall BK (ed) Homology: the hierarchical basis of comparative biology. Academic Press, London
12. Dowling TE, Moritz C, Palmer JD et al (1996) Nucleic acids III: analysis of fragments and restriction sites. In: Hillis DM, Mortiz C, Mable BK (eds) Molecular systematics. Sinauer Associates Inc, Sunderland
13. Lynch M, Milligan BG (1994) Analysis of population genetic structure with RAPD markers. Molec Ecol 3:91–99
14. Bussell JD, Waycott M, Chappill JA (2005) Arbitrarily amplified DNA markers as characters for phylogenetic analyses. Perspect Plant Ecol Evol Syst 7:3–16
15. Ganie SH, Upadhyay P, Das S et al (2015) Authentication of medicinal plants by DNA markers. Plant Gene 4:83–99
16. Goodwin ID, Aitkin EAB, Smith LW (1997) Application of inter simple sequence repeat markers (ISSR) to plant genetics. Electrophoresis 18:1524–1528
17. Grover A, Sharma PC (2016) Development and use of molecular markers: past and present. Crit Rev Biotechnol 36:290–302. https://doi.org/10.3109/07388551.2014.959891
18. Kuluev BR, AnKh B, Gerashchenkov GA et al (2018) Random priming PCR strategies for identification of multilocus DNA polymorphism in eukaryotes. Russ J Genet 54:499–513

19. Nadeem MA, Nawaz MA, Shahid MQ et al (2018) DNA molecular markers in plant breeding: current status and recent advancements in genomic selection and genome editing. Biotechnol Biotechnol Equip 32:261–285. https://doi.org/10.1080/13102818.2017.1400401
20. Reddy MP, Sarla N, Siddiq EA (2002) Inter simple sequence repeat (ISSR) polymorphism and its application in plant breeding. Euphytica 128:9–17
21. Stevens MI, Clarke AC, Clarkson FM et al (2015) Are current ecological restoration practices capturing natural levels of genetic diversity? A New Zealand case study using AFLP and ISSR data from māhoe (*Melicytus ramiflorus*). New Zeal J Ecol 39:190–197
22. Tamboli AS, Yadav PB, Gothe AA et al (2018) Molecular phylogeny and genetic diversity of genus *Capparis* (Capparaceae) based on plastid DNA sequences and ISSR markers. Plant Syst Evol 304:205–217
23. Wolfe A, Liston A (1998) Contributions of PCR-based methods to plant systematics and evolutionary biology. In: Soltis DE, Soltis PS, Doyle JJ (eds) Plant molecular systematics II. Kluwer Academic Publishers, Boston. https://doi.org/10.1007/978-1-4615-5419-6
24. Wu Y, Yang DY, Tu PF et al (2011) Genetic differentiation induced by spaceflight treatment of *Cistanche deserticola* and identification of inter-simple sequence repeat markers associated with its medicinal constituent contents. Adv Space Res 47:591–599
25. Kumar A, Mishra P, Baskaran K et al (2016) Higher efficiency of ISSR markers over plasmid psbA-trnH region in resolving taxonomical status of genus *Ocimum* L. Ecol Evol 6:7671–7682
26. Savelkoul PH, Aarts HJ, de Haas J et al (1999) Amplified-fragment length polymorphism analysis: the state of an art. J Clin Microbiol 37:3083–3091
27. Bornet B, Branchard M (2001) Nonanchored inter simple sequence repeat (ISSR) markers: reproducible and specific tools for genome fingerprinting. Plant Mol Biol Rep 19:209–215
28. Rogers SO, Bendich AJ (1985) Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. Plant Mol Biol 5:69–76
29. Doyle JJ, Doyle JL (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem Bull 19:11–15
30. Deng Q, Deng Q, Liu L et al (2019) DNA extraction and optimization of ISSR-PCR reaction system for *Pyracantha*. IOP Conf Ser Earth Environ Sci 237:052025. https://doi.org/10.1088/1755-1315/237/5/052025
31. Mondal A, Pal T, De KK (2018) Fluorescent inter simple sequence repeat (F-ISSR) markers and capillary electrophoresis to assess genetic diversity and relatedness within commercial sugarcane varieties. Int J Agric Technol 14:717–730
32. Stevens MI, Hunger SA, Hills SFK et al (2007) Phantom hitch-hikers mislead estimates of genetic variation in Antarctic mosses. Plant Syst Evol 263:191–201
33. Camacho FJ, Gernandt DS, Liston A et al (1997) Endophytic fungal DNA, the source of contamination in spruce needle DNA. Mol Ecol 6:983–987
34. Smith DE, Klein AS (1996) Erratum. Mol Phylogenet Evol 5:286–287
35. White TJ, Bruns T, Lee S et al (1990) Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics. In: Innis MA, Gelfand DH, Sninsky JJ et al (eds) PCR protocols: a guide to methods and applications. Academic Press, San Diego
36. Baldwin BG (1992) Phylogenetic utility of the internal transcribed spacers of nuclear ribosomal DNA in plants: an example from the Compositae. Mol Phylogenet Evol 1:3–16
37. Archibald JK, Mort ME, Crawford DJ et al (2006a) The utility of automated analysis of inter-simple sequence repeat (ISSR) loci for resolving relationships in the Canary Island species of *Tolpis* (Asteraceae). Am J Bot 93:1154–1162
38. Archibald JK, Mort ME, Crawford DJ et al (2006b) Evolutionary relationships within recently radiated taxa: comments on methodology and analysis of inter-simple sequence repeat data and other hypervariable, dominant markers. Taxon 55:747–756
39. Arens P, Odinot P, van Heusden AW et al (1995) GATA and GACA repeats are not evenly distributed throughout the tomato genome. Genome 38:84–90
40. Prince LM (2015) Plant genotyping using fluorescently tagged inter-simple sequence repeats (ISSRs): basic principles and methodology. In: Batley J (ed) Plant genotyping. Methods in molecular biology (methods and protocols), vol 1245. Humana Press, New York
41. Oliveira EC, Amaral Júnior AT, LSA G et al (2010) Optimizing the efficiency of the touchdown technique for detecting inter-simple sequence repeat markers in corn (*Zea mays*). Genet Mol Res 9:835–842
42. Goulao LF, Oliveira CM (2014) Multilocus profiling with AFLP, ISSR, and SAMPL. In:

Besse P (ed) Molecular plant taxonomy, methods and protocols. Humana Press, New York

43. Prince LM (2009) The relationship of Monardella viminea to closely related taxa based on analyses of ISSRs. USFWS Report P0750003
44. Applied Biosystems (2010) Application Note, ISSR Plant Genotyping. Publication 106AP31-01. Life Technologies Corporation, USA. http://tools.invitrogen.com/content/sfs/brochures/cms_079244.pdf
45. Somasundaram SM, Subbaraya U, Durairajan SG et al (2019) Comparison of two different electrophoretic methods in studying the genetic diversity among plantains (*Musa* spp.) using ISSR markers. Electrophoresis 40:1265–1272
46. Liu B, Wendel JF (2001) Inter simple sequence repeat (ISSR) polymorphisms as a genetic marker system in cot ton. Molec Ecol Notes 1:205–208
47. Barth S, Melchinger AE, Lübberstedt T (2002) Genetic diversity in *Arabidopsis thaliana* L. Heynh. Investigated by cleaved amplified polymorphic sequence (CAPS) and inter-simple sequence repeat (ISSR) markers. Mol Ecol 11:495–505
48. Grierson ERP (2014) The Development and Genetic Variation of Sophora prostrata—A New Zealand divaricating shrub. Unpublished MSc Thesis, University of Waikato
49. Crawford DJ, Mort ME (2004) Single-locus molecular markers for inferring relationships at lower taxonomic levels: observations and comments. Taxon 53:631–635
50. Dice LR (1945) Measures of the amount of ecological association between species. Ecology 26:297–302
51. Nei M, Li W-H (1979) Mathematical model for studying genetic variation in terms of restriction endonucleases. Proc Natl Acad Sci U S A 76:5269–5273
52. Hollingsworth PM, Ennos RA (2004) Neighbour joining trees, dominant markers and population genetic structure. Heredity 92:490–449
53. Nelson MF, Anderson NO (2013) How many marker loci are necessary? Analysis of dominant marker data sets using two popular population genetic algorithms. Ecol Evol 3:3455–3470. https://doi.org/10.1002/ece3.725
54. Excoffier L, Smouse PE, Quattro JM (1992) Analysis of molecular variance inferred from metric distances among DNA haplotypes: application to human mitochondrial DNA restriction data. Genetics 131:479–491
55. Excoffier L, Laval G, Schneider S (2005) Arlequin (version 3.0): an integrated software package for population genetics data analysis. Evol Bioinformatics Online 1:47–50
56. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
57. Falush D, Stephens M, Pritchard JK (2007) Inference of population structure using multilocus genotype data: dominant markers and null alleles. Mol Ecol Notes 7:574–578
58. Porras-Hurtado L, Ruiz Y, Santos C et al (2013) An overview of STRUCTURE: applications, parameter settings, and supporting software. Front Genet 4:98. https://doi.org/10.3389/fgene.2013.00098
59. Wolfe AD (2005) ISSR techniques for evolutionary biology. Methods Enzymol 395:134–144. https://doi.org/10.1016/S0076-6879(05)95009-X
60. Gemmill CEC, Allan GJ, Wagner WL et al (2002) Evolution of insular pacific *Pittosporum* (Pittosporaceae): origin of the Hawaiian radiation. Mol Phylogenet Evol 22:31–42
61. Carter SN, Miller S, Meyer SJ et al (2018) A new species of *Pittosporum* described from the poor Knights Islands, northland, Aotearoa/New Zealand. Syst Bot 43:633–643. https://doi.org/10.1600/036364418X697355

Chapter 15

Retrotransposable Elements: DNA Fingerprinting and the Assessment of Genetic Diversity

Ruslan Kalendar, Alexander Muterko, and Svetlana Boronnikova

Abstract

Retrotransposable elements (RTEs) are highly common mobile genetic elements that are composed of several classes and make up the majority of eukaryotic genomes. The "copy-out and paste-in" life cycle of replicative transposition in these dispersive and ubiquitous RTEs leads to new genome insertions without excision of the original element. RTEs are important drivers of species diversity; they exhibit great variety in structure, size, and mechanisms of transposition, making them important putative components in genome evolution. Accordingly, various applications have been developed to explore the polymorphisms in RTE insertion patterns. These applications include conventional or anchored polymerase chain reaction (PCR) and quantitative or digital PCR with primers designed for the 5′ or 3′ junction. Marker systems exploiting these PCR methods can be easily developed and are inexpensively used in the absence of extensive genome sequence data. The main inter-repeat amplification polymorphism techniques include inter-retrotransposon amplified polymorphism (IRAP), retrotransposon microsatellite amplified polymorphism (REMAP), and Inter-Primer Binding Site (iPBS) for PCR amplification with a single or two primers.

Key words Retrotransposon, Molecular marker, IRAP, REMAP, iPBS

1 Introduction

All eukaryotic genomes contain DNA sequences termed "repetitive elements" that are present in multiple copies throughout the genome [1–3]. These repetitive sequences can either be tandemly arrayed or interspersed throughout the genome [4]. Interspersed repetitive sequences comprise a large fraction of eukaryotic genomes and are predominantly comprised of retrotransposable elements (retrotransposons, or RTEs) [1, 5–9]. For example, retrotransposons can comprise up to 90% of the genome in some eukaryotes. In most of the species studied thus far, these interspersed repeats are distributed unevenly across the nuclear genome, with some repeats having a tendency to cluster around the centromeres or telomeres [10, 11]. Moreover, RTEs are predominantly located in heterochromatic regions of the genome. Cereals and

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_15,

citrus fruits often have retrotransposons locally nested within one another and in extensive domains, referred to as "retrotransposon seas," which surround gene islands. Nevertheless, the most prevalent retrotransposons are dispersed throughout the genome [3]. RTEs can be subdivided on the basis of their size; short interspersed elements are less than 1000 bp long, and the rest are considered to be long interspersed elements. Variation in the copy number of repeat elements and internal rearrangements on both homologous chromosomes can arise after the induction of recombinational processes during the meiotic prophase [12, 13].

It was previously shown that a specific retrotransposon is universally distributed among closely and distantly related species [14, 15]. Although there is no unique set of retrotransposons for a particular species [14, 16–19], related species have phylogenetically cognate (related) RTE sequences. As such, any high-copy mobile genetic element (MGE) shows phylogenetic similarity among related species. For example, both the long terminal repeats (LTRs) and the central part are conserved, consistent with their parent plant families. Generally, retrotransposons have not been extensively explored as phylogenetic markers, except in a few articles that have discussed the phylogenetic relationships among concrete retrotransposon sequences [14, 15, 17, 20]. Since high-copy RTEs are widely distributed and diverse in eukaryotes, they offer many advantages for their use in eukaryotic phylogenetic studies. Their features of abundance, general dispersion, and activity provide ideal conditions for developing molecular phylogenetic markers [21, 22].

1.1 LTR Retrotransposons

Transposable elements (TEs) are classified into two main groups in eukaryotic genomes, defined according to their mechanism of transposition [23]. Class I TEs transpose through an RNA intermediate, which class II transposons lack [24] (Fig. 1). The two classes can be further divided into two subclasses according to their structure and transposition cycle: LTR retrotransposons and non-LTR retrotransposons. Non-LTR retrotransposons can either be long interspersed nuclear elements (LINE) or short interspersed nuclear elements (SINE). All groups are complemented by degraded members of their nonautonomous forms, which lack genes that are essential for transposition. Specifically, miniature inverted-repeat transposable elements (MITEs) are the nonautonomous form of class II transposons, SINEs are the nonautonomous form of non-LTR retrotransposons, and terminal-repeat retrotransposons in miniature (TRIMs) and large retrotransposon derivatives (LARDs) are nonautonomous LTR retrotransposons [14, 15, 23, 25].

Class I transposable elements/retrotransposons replicate by a process of reverse transcription, as do lentiviruses such as HIV [1, 26, 27]. The retrotransposons themselves encode the proteins

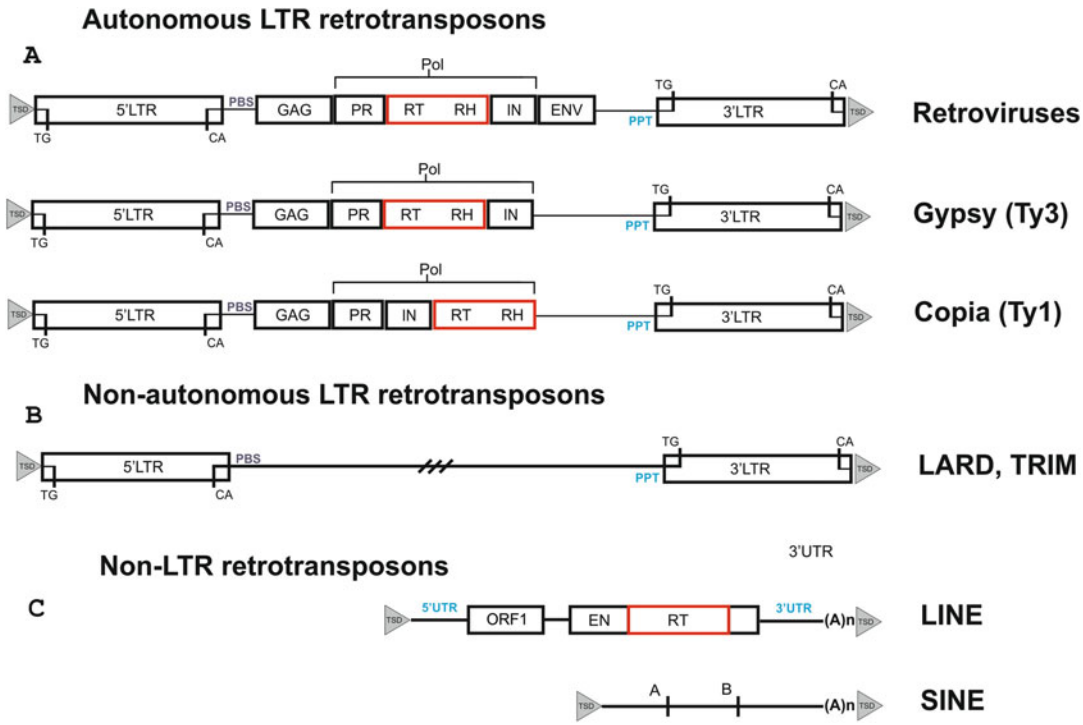


Fig. 1 Retrotransposon architecture: the main groups of autonomous and nonautonomous retrotransposons. (**a**) Retroviruses and autonomous LTR retrotransposons. Above, the basic structure of an LTR retrotransposon, comprising target site duplication (TSD); long terminal repeats (LTRs); the primer binding site (PBS), which is the (−)-strand priming site for reverse transcription; and the polypurine tract (PPT), which is the (+)-strand priming site for reverse transcription. The PBS and PPT are part of the internal domain, which in autonomous elements includes the protein-coding open reading frames (ORFs). The ORFs of the internal domain are *GAG* encoding the capsid protein Gag, *PR* proteinase, *RT-RH* reverse transcriptase-RNase H, *INT* integrase, and *ENV* envelope protein. (**b**) Nonautonomous retrotransposons. LARD elements have a long internal domain with a conserved structure but lack a coding capacity. TRIM elements have virtually no internal domain except for the PBS and PPT signals. (**c**) Autonomous and nonautonomous non-LTR retrotransposons. The autonomous order LINE of the L1 superfamily and the nonautonomous order SINE are shown

needed for their replication and integration back into the genome [28]. Their "copy-out and paste-in" life cycle means that they do not need to be excised in order for a copy to be inserted elsewhere in the genome. Hence, genomes diversify by the insertion of new copies, while the original copies persist. Their abundance in the genome is generally highly correlated with genome size. Indeed, large plant genomes contain hundreds of thousands of these elements, which together form the vast majority of the total DNA [6, 17].

Three basic types of LTR retrotransposon structures are illustrated in Fig. 1, each having two LTRs. An LTR varying in length from 100 bp to a few kb generally starts, and its inverted repeat sequence 5′-TG--CA-3′ ends. They tend to form direct repeats of

4–6 bp (target site duplications, TSDs) at both ends of the transposon upon insertion into the genome. An LTR retrotransposon is comprised of a gene encoding a variety of proteins, including the GAG (encoding structural proteins forming the shell, the synthesis of reverse transcription) and poly POL gene (encoding a series of reverse transcription enzymes). In addition, LTR retrotransposons contain transcription initiation and termination sequences related to a tRNA binding site (primer binding site, PBS) and a polypurine sequence (polypurine tract, PPT). Based on the similarity of the order and sequence of the enzyme transposase genes, LTR retrotransposons can be subdivided into the Tyl-*copia* type and Ty3-*gypsy* type.

Human and other mammalian genomes contain an abundance of retrotransposons. The majority of these, however, are not LTR retrotransposons. Rather, they are LINEs and SINEs, which replicate by a somewhat different copy-and-paste mechanism [2, 29, 30]. The L1 family of LINEs and the *Alu* family of SINEs together comprise roughly 30% of human genomic DNA and contain nearly two million copies [31]. Integrated retroviruses, which are remnants of ancient infections, are also abundant in mammalian genomes [32]. These elements, called "endogenous retroviruses" (ERVs, or HERVs in humans), are functionally equivalent to LTR retrotransposons. The features of integration activity, persistence, dispersion, high copy number, as well as conserved structure and sequence motifs together make retrotransposons well-suited to build molecular marker systems [18, 22, 33–41].

1.2 Retrotransposons as DNA Markers

Retrotransposable elements, which are among the MGEs that are abundantly present in the genomes of plants, are known to be excellent molecular genetic markers. The insertion of LTR retrotransposons is random, and it occurs during the transposition process in the continuous evolution of species. This can provide a wealth of information for the study of evolution, species diversification, and genomic differentiation. The transposition mechanism for the LTR-LTR retrotransposon sequence determines the ends after transposition and is completely consistent. Therefore, by comparing the sequence LTR ends of the complete transposon, the insertion time can be calculated based on their mutation rates.

RTE-based molecular genetic marker applications have become a key part of research on genetic variability and diversity [28, 37–39, 42–44]. The scope of their usage includes creating genetic maps and identifying individuals or lines that carry certain genetic polymorphic variation [45, 46]. The DNA marker system takes advantage of the developments in molecular genetics and biochemistry [47], using "fingerprints" (i.e., distinctive patterns of DNA fragments resolved by gel electrophoresis or next-generation sequencing (NGS)). Specifically, molecular genetic markers work by finding polymorphisms in a nucleotide sequence at a particular

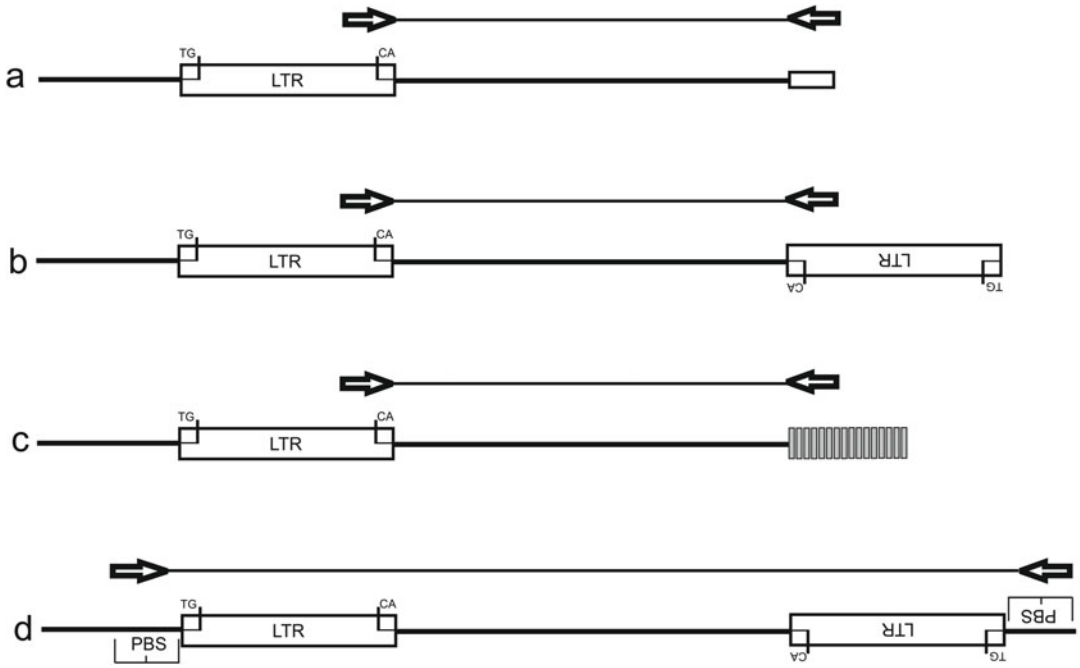


Fig. 2 Retrotransposon-based molecular marker methods. Multiplex products of various lengths from different loci are indicated by the bars above or beneath the diagrams for each reaction. Primers are indicated by arrows. (**a**) The SSAP method. The primers used for amplification match the adapter (empty box) and retrotransposon (LTR box). (**b**) The IRAP method. Amplification takes place between retrotransposons (left and right LTR boxes) near each other in the genome (open bar), using retrotransposon primers. The elements are shown oriented head-to-head, using a single primer. (**c**) The REMAP method. Amplification takes place between a microsatellite domain (vertical bars) and a retrotransposon, using a primer anchored to the proximal side of the microsatellite and a retrotransposon primer. (**d**) The inter-PBS (iPBS) amplification scheme and LTR retrotransposon structure. Two nested LTR retrotransposons in inverted orientations amplified from a single primer or two different primers from primer binding sites. The PCR product contains both LTRs and PBS sequences as PCR primers in the termini. In the figure, the general structure for PBS and LTR sequences and the several-nucleotide-long spacer between 5′-LTR and PBS are schematically shown

genomic location; when this nucleotide sequence varies between the parents of the chosen cross, it can be discernible between plant accessions; hence, its pattern of inheritance can be investigated.

Retrotransposon-based molecular genetic systems (Fig. 2) detect the insertion of elements hundreds to thousands of nucleotides long, although generally only the insertion joint itself is screened due to the impracticality of amplifying and resolving long fragments and discriminating their insertion sites. The LTRs that encompass a complete retrotransposon contain ends that are highly conserved in a given family of elements. Newly inserted retrotransposons, therefore, form a joint between the conserved LTR ends and flanking anonymous genomic DNA. Most retrotransposon-based marker systems use polymerase chain reaction (PCR) to amplify a segment of genomic DNA at this joint.

Generally, one primer is designed to match a segment of the LTR that is conserved within a given family of elements, but different in other families. The primer is oriented toward the LTR end. The second primer is designed to match some other features of the genome. The first described retrotransposon method was SSAP (sequence-specific amplified polymorphism; *see* Fig. 2) in barley, where one primer matched the end of the *BARE1* retrotransposon and the other matched an AFLP (amplified fragment length polymorphism)-like restriction site adapter [48].

1.3 Inter-Retrotransposon Amplified Polymorphism (IRAP) and Retrotransposon Microsatellite Amplified Polymorphism (REMAP)

Inter-repeat amplification polymorphism techniques such as inter-retrotransposon amplified polymorphism (IRAP), retrotransposon microsatellite amplified polymorphisms (REMAP), and inter-MITE amplification have been used to amplify abundantly dispersed repeats, including the LTRs of retrotransposons and SINE-like sequences (inter-SINE amplified polymorphisms) [49]. The IRAP and REMAP (Fig. 2) PCR methods represent a departure from SSAP (Chapter 12)—no restriction enzyme digestion or ligation step is needed, and the products can be resolved by conventional agarose gel electrophoresis without the need of a sequencing apparatus. The IRAP method detects retrotransposon insertional polymorphisms by amplifying the portion of DNA between two retroelements. It uses one or two primers pointing outward from an LTR and therefore amplifies the tract of DNA between two nearby retrotransposons. IRAP can be carried out with a single primer matching either the 5′ or 3′ end of the LTR (oriented away from the LTR itself) or with two or more primers. The two primers may be from the same retrotransposon element family or from different families. The PCR products—and therefore the fingerprint patterns—result from amplification of hundreds to thousands of target sites in the genome. LTR primers from one species can be used on other species because related species have phylogenetically related TE sequences. In such cases, primers designed for conservative TE sequences are advantageous.

The complexity of the pattern obtained will be influenced by the retrotransposon copy number, which mirrors genome size, as well as by their insertion pattern and the size of the retrotransposon families of interest. Furthermore, thousands of products can neither be simultaneously amplified to detectable levels nor resolved on a gel system. Hence, the pattern obtained represents the result of competition between the targets and products in the reaction. As a result, the products obtained with two primers do not represent the simple sum of the products obtained with the primers individually.

If retrotransposons were fully dispersed within the genome, IRAP would either produce products too large to provide a clear resolution on gels or target amplification sites too far apart to produce products with the available thermostable polymerases.

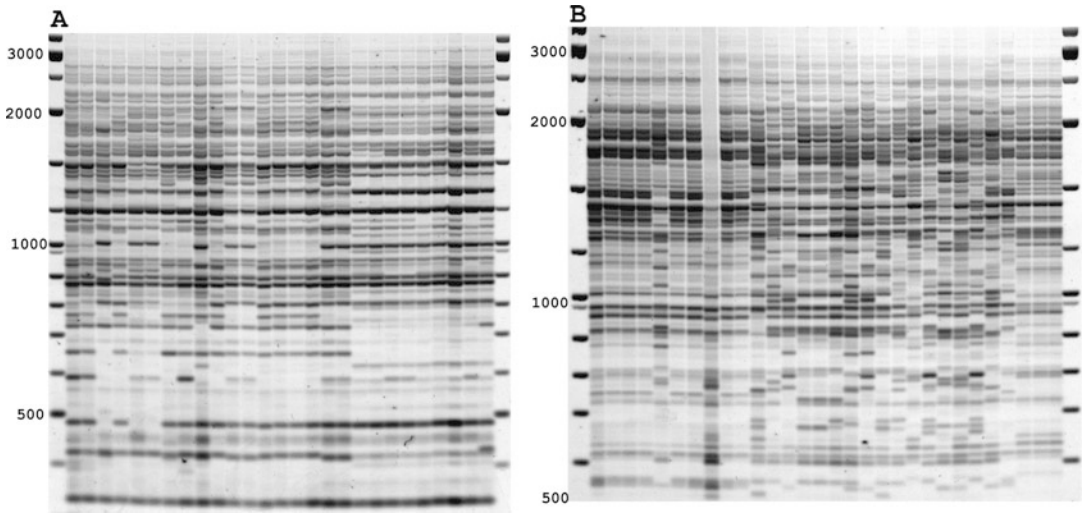


Fig. 3 The use of IRAP in the diversity analysis of 30 genotypes of populations of *Triticum dicoccoides*. (**a**) Results for the LTR retrotransposon *Wilma* (LTR primer 2108: 5′-AGAGCCTTCTGCTCCTCGTTGGGT-3′). (**b**) Results for the LTR retrotransposon *Wis2* (LTR primer 2106: 5′-TAATTTCTGCAACGTTCCCCAACA-3′). A size marker (Thermo Fisher Scientific GeneRuler DNA Ladder Mix (100 to 3000 bp)) is present on both sides, marked on the left in bp

This is because retrotransposons generally tend to cluster together and may even nest within each other. For example, the *Wis2* and *Wilma* retrotransposons from grasses, which are the average abundant superfamily *copia* elements, are present as roughly 20,000 full-length copies of about 8 kb in the wheat genome. IRAP with *Wilma* and *Wis2* primers displays a range of products from 500 bp to upward of 5 kb (Fig. 3) [50, 51].

The REMAP method is similar to IRAP, except that one of the two primers matches a simple sequence repeat (SSR) motif with one or more non-SSR anchor nucleotides present at the 3′ end of the primer. Microsatellites of the form $(NN)_n$, $(NNN)_n$, or $(NNNN)_n$ are found throughout plant and animal genomes. Furthermore, in cereals, they appear to be associated with retrotransposons [16]. Differences in the number of SSR units in a microsatellite are generally detected using primers designed for unique sequences flanking the microsatellite. Alternatively, the stretches of the genome that are present between two microsatellites may be amplified by inter-simple sequence repeats (ISSRs) (Chapter 14), in a way akin to IRAP. In REMAP, anchor nucleotides are used at the 3′ end of the SSR primer, in order to both avoid slippage of the primer within the SSR (which would produce a "stutter" pattern in the fingerprint) and avoid detection of variation in repeat numbers within the SSR. REMAP uses primer types that are shared by IRAP and ISSR. It is in theory possible that the SSR primers in REMAP could also yield ISSR products, and the LTR primers

could also yield IRAP products. However, in practice, this rarely occurs, probably due to a combination of factors including genome structure and competition within the PCR.

The generation of a virtually unlimited number of unique markers is possible through the combination of different LTR primers or using combinations with microsatellite primers (REMAP) [52]. The same primers produce completely different banding patterns depending on whether they are used alone or in combination, demonstrating that most of the IRAP/REMAP bands were derived from sequences flanked by an LTR or a microsatellite on one side and by another LTR on the other side [39]. In general, a more variable and stable pattern has been observed in IRAP than in ISSR/RAPD (random amplified polymorphic DNA) (Chapters 13 and 14); frequently, but not always, single priming PCR also shows less variability than the IRAP pattern generated when primer combinations are used, depending on the LTR sequence [53].

1.4 Inter-PBS (iPBS) Amplification: A Universal Method for Isolating and Displaying Retrotransposon Polymorphisms

A major disadvantage of all retrotransposon-based molecular-genetic marker techniques is the need for sequence information to design retrotransposon-specific primers. The primary requirement is the sequence of an LTR end, either mined from a database or produced by cloning and sequencing the genomic DNA that flanks the conserved segments of retrotransposons. Indeed, rapid retrotransposon isolation methods based on PCR with conserved primers for RTE have been designed. Nevertheless, it maybe still necessary to clone and sequence hundreds of clones to obtain only a few good primer sequences. The LTRs do not contain conserved motifs for distantly related species, which would allow their direct amplification by PCR. However, all reverse transcribing elements, including LTR and LINE retrotransposons, can be obtained by PCR with degenerate RT primers. Based on how conserved the reverse transcriptase domain is—particularly for the Ty1-*copia* type—a few restrictions and adapter-based methods for LTR cloning have been developed [54]. Major classes of retrotransposons include the Pseudoviridae (Ty1-*copia*), Metaviridae (Ty3-*gypsy*), and Retroposineae *LINE* (non-LTR) groups. PCR with degenerate RT primers can produce all reverse transcribing elements. For instance, two Ty1-*copia* degenerate primers have been designed for the RT domain encoding TAFLHG and the reverse site YVDDML and also encoding QMDVKT and the reverse YVDDML [55–57]. For the Ty3-*gypsy* element, degenerate primers have been designed for the RT domain encoding RMCVDYR, LSGYHQI, or YPLPRID and the reverse encoding sites YAKLSKC and LSGYHQI. The method based on reverse transcriptase can only be applied to the family of retrotransposons that contains this sequence. Therefore, for example, TRIM or LARDs and unknown classes of LTR retrotransposons cannot be found using this approach [14, 15, 58, 59].

The LTR retrotransposons and all retroviruses contain a conserved binding site for tRNA. Generally, $tRNAi^{Met}$ is the most common, but $tRNA^{Lys}$, $tRNA^{Pro}$, $tRNA^{Trp}$, $tRNA^{Asn}$, $tRNA^{Ser}$, $tRNA^{Arg}$, $tRNA^{Phe}$, $tRNA^{Leu}$, and $tRNA^{Gln}$ can also be found. Elongation from the 3′-terminal nucleotides of the respective tRNA results in the conversion of the retroviral or retrotransposon RNA genome to double-stranded DNA prior to its integration into the host DNA. While the process of reverse transcription is conserved among virtually all retroelements, the specific tRNA capture varies for different retroviruses and retroelements. The primer binding sites (PBS) are almost universally present in all LTR retrotransposon sequences. Hence, an isolation method for retrotransposon LTRs based on the PBS sequence has the potential for cloning all possible LTR retrotransposons.

The inter-PBS (iPBS) amplification technique has led to the development of a virtually universal and exceedingly efficient method, which utilizes the conserved parts of PBS sequences, for direct visualization of polymorphisms between individuals, polymorphisms in transcription profiles, fast cloning of LTR segments from genomic DNA, as well as for database searches of LTR retrotransposons (Fig. 2). Although many retrotransposons are nested, recombined, inverted, or truncated, they can still be easily amplified using conservative PBS primers in any plant species tested. Fragments of retrotransposons containing a 5′ LTR and part of the internal domain are often located near other entirely or similarly truncated retrotransposons. Therefore, PBS sequences are very often located sufficiently near to each other to allow amplification. This situation allows the use of PBS sequences for cloning LTRs. Where the retrotransposon density is high within a genome, PBS sequences can be exploited for detection of their chance association with other retrotransposons. When retrotransposon activity or recombination has led to new genome integration sites, the iPBS method can be used to distinguish reproductively isolated plant lines. In this case, amplified bands derived from a new insertion event or from recombination will be polymorphic, appearing only in plant lines in which the insertions or recombination has taken place.

The PBS primer(s) can amplify the sequences of nested inverted retrotransposons or related elements that are dispersed throughout genomic DNA. In this case, the PCR amplification occurs between two nested elements' PBS domains and produces fragments containing the insertion junction between the two nested LTRs. After retrieving LTR sequences from a selected family of retrotransposons, they are aligned to identify the most conserved region. Related plant species have conserved regions in LTRs for members of the same retrotransposon family. Thus, alignments of several LTR sequences from several species will identify these conserved regions. Subsequently, these conserved LTR domains can be used

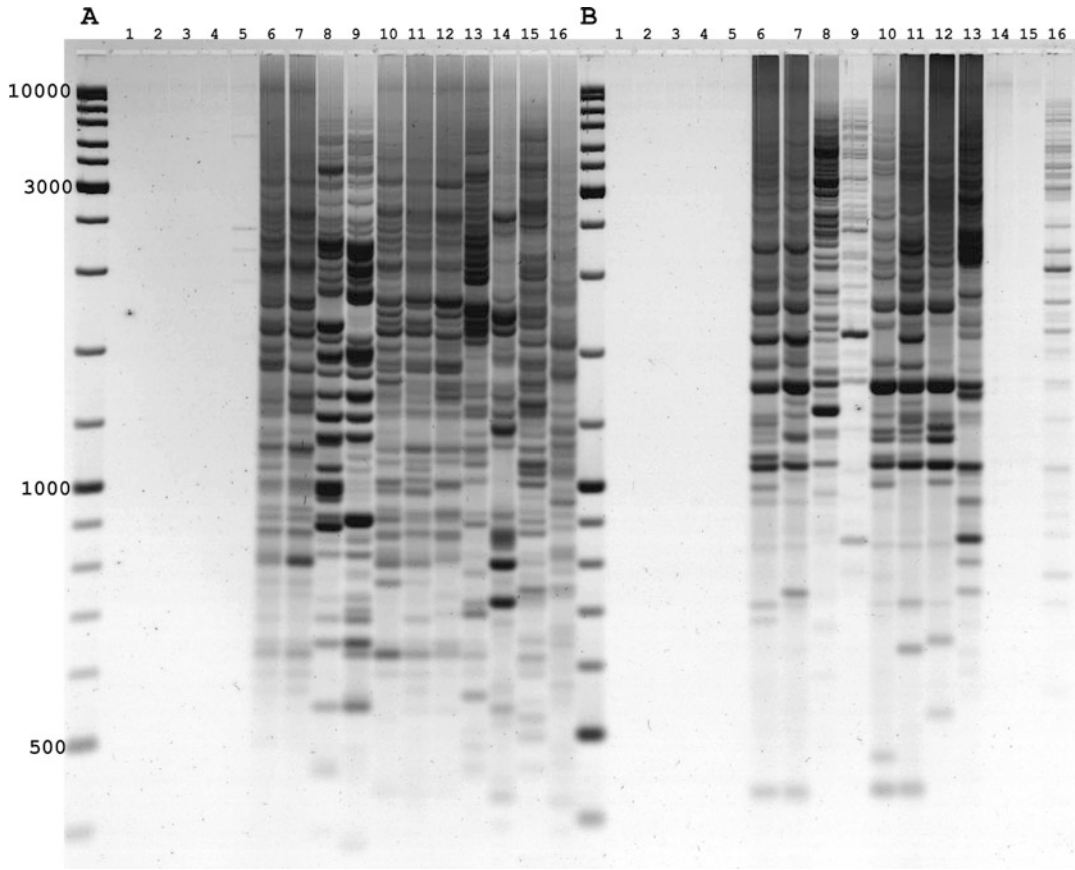


Fig. 4 The effectiveness of IRAP amplification according to genome size. An IRAP gel produced with LTR primers: (**a**) LTR retrotransposon (*gypsy*) *Bagy2* (primer 833: 5′-TGATCCCCTACACTTGTGGGTCA-3′). (**b**) LTR retrotransposon (TRIM) *Cassandra* (primer 2015: 5′-ACCTGGATGCAACAGAGGTCTATG-3′). A size marker (Thermo Fisher Scientific GeneRuler DNA Ladder Mix (100 to 10,000 bp)) is present on both sides, marked on the left in bp. DNA samples of *Triticeae* species with a small genome include *Brachypodium distachyon* (lanes 1–5); those with a large genome include *Triticum aestivum* (ABD; lane 6); *Triticum durum* (AB; lane 7); *Aegilops tauschii* (D; lanes 8–9); *Triticum dicoccoides* (AB; lanes 10–12); *Aegilops peregrina* (S; lane 13); *Phleum pratense* (lane 14); *Avena sativa* (lane 15); *Secale strictum* (H4342; lane 16). For the small genome of *Brachypodium distachyon*, there is no IRAP amplification, whereas for the large genomes of *Triticeae* species, multiple amplicons are observed

for inverted primers designed for long distance PCR, for cloning of whole retrotransposons, and also for the IRAP, REMAP, or SSAP marker techniques (Figs. 3 and 4). The iPBS amplification technique shows roughly the same level of polymorphism as IRAP and REMAP, and it is an efficient method for the detection of cDNA polymorphism and clonal differences resulting from retrotransposon activities or recombination [37, 60, 61]. In order to obtain a vigorous, rapid, and economical marker system for genotyping applications in plant breeding and marker-assisted selection, iPBS amplification was elaborated.

Further research on related varieties or breeding lines could be carried out through the development of a native RTE system, which then requires the cloning and sequencing of elements from new a species by using iPBS amplification or a technique based on the conservancy of the reverse transcriptase domain. The process is initiated by the amplification and cloning of segments between retrotransposon domains that are highly or universally conserved, the development of new primers specific for the retrotransposon families found, and the testing of these for their efficacy as markers.

Next-generation sequencing allows small-scale, inexpensive genome sequencing with a turnaround time measured in days [62, 63]. However, as NGS is generally performed and currently understood, all regions of the genome are sequenced with roughly equal probability, meaning that a large amount of a genomic sequence is collected and discarded to collect sequence information from the relatively low percentage of areas where the function is understood well enough to interpret potential mutations.

2 Materials

2.1 Reagents

All solutions should be prepared using Milli-Q or equivalent ultrapure water and analytical-grade reagents.

1. TE buffer (10×): 100 mM Tris–HCl, pH 7.5–8.0, 10 mM EDTA. DNA and primers should be stored in a 1× TE solution.
2. Electrophoresis buffer (10× TBE): 450 mM H_3BO_3, pH 8.8, 5 mM EDTA. Weigh 54.5 g Tris-base and 27.8 g H_3BO_3, add 10 mL 0.5 M EDTA, pH 8.0, dissolve in water, and bring final volume to 1 L. Store at room temperature.
3. Gel loading buffer (10×): 20% (w/w) Polysucrose 400, 100 mM Tris–HCl, pH 8.0, 10 mM EDTA, ~0.01% (w/w) Orange G, and ~0.01% Xylene Cyanol FF. Dissolve 10 g Polysucrose 400 (Ficoll 400) in 80 mL 10× TE buffer. Add Orange G and Xylene Cyanol FF according to the desired color intensity. Store at 4 °C.
4. Thermostable DNA polymerase: many types and sources of recombinant thermostable DNA polymerases are effective. Those that are most preferable for PCR use are the recombinant DNA polymerases from *Thermus aquaticus* (*Taq*) where applicable (*see* **Note 1**). A polymerase mix consisting of 50 units of *Taq* DNA polymerase and 1 unit of *Pfu* DNA polymerase improves amplification of long bands and the accuracy of the PCR.
5. PCR reaction buffers (1×): several PCR buffers for *Taq* polymerases are suitable for PCR: Buffer 1: 60 mM Tris-SO_4

(pH 9.0), 2 mM $MgSO_4$, 20 mM $(NH_4)_2SO_4$. Buffer 2: 10 mM Tris–HCl (pH 8.8), 2 mM $MgCl_2$, 50 mM KCl, 0.1% Triton X-100. Buffer 3: 50 mM Tris–HCl (pH 9.0), 1.5 mM $MgCl_2$, 15 mM $(NH_4)_2SO_4$, 0.1% Triton X-100. The PCR and its efficiency depend on which buffer and enzyme combination is used (*see* **Note 1**).

6. Ethidium bromide solution in water, 0.5 mg/mL. Store at room temperature.
7. DNA ladder for electrophoresis 100–10,000 base range. DNA ladder should be diluted with 1× gel loading buffer to a final concentration of 25 ng/μL.
8. Agarose wide range for DNA electrophoresis with gel strength >1700 g/cm^2 can be used.

2.2 Equipment

1. Thermal cycler for 0.2 mL tubes or plates (96-well), with a rapid heating and cooling capacity between 4 °C and 99 °C—the temperature should be able to change by 5–10 °C per second.
2. Power supply (minimum 300 V, 400 mA) for electrophoresis.
3. Horizontal electrophoresis apparatus without special cooling. Most commercially available medium- or large-scale horizontal DNA gel electrophoresis systems are suitable (*see* **Note 2**).
4. Gel comb with at least 36 wells, 1 mm thickness, forming 3–4 mm wide wells, with a 1 mm well spacing (*see* **Note 3**).
5. UV transilluminator, for visualization of ethidium bromide-stained or SYBR Green-stained nucleic acids, with a viewing area of 20 × 20 cm.
6. Imaging system. A digital gel electrophoresis scanner for detection of ethidium bromide-stained nucleic acids by fluorescence (532 nm green laser) with a resolution of 50–100 μm. Software such as ImageJ (https://imagej.nih.gov/ij/) is required for image analysis and manipulation.

2.3 DNA Template

The DNA template should be diluted with 1× TE solution to obtain the appropriate working concentration (5–10 ng/μL) and stored at 4 °C. Pure DNA can be stored at 4 °C for many years without showing any PCR inhibition or decrease in amplification efficiency for DNA fingerprinting.

The quality of genomic DNA plays an important role in the quality of the resulting fingerprint. In order to be used in DNA fingerprinting applications, the target DNA should be free of contaminants that inhibit PCR and other downstream applications (*see* **Note 4**).

Furthermore, contaminated DNA will decline in PCR performance after prolonged (e.g., 1 month or longer) periods of storage

due to chemical modification. Such DNA should be extracted, for example, with methods involving CTAB in weakly acidic conditions (pH < 6), followed by chloroform DNA extraction (*see* Chapter 3). Depending on the biological material, DNA can be precipitated by filtration through a column with a glass microfiber filter or through cellulose paper.

Pure DNA should exhibit an A_{260}/A_{230} ratio in the range of 1.8–2.0. Significantly lower values may indicate contamination with polysaccharides. The integrity of the genomic DNA samples extracted can be analyzed by electrophoresis on a 1% agarose gel for 1 h at 70 V, with a DNA ladder for scale.

2.4 Primer Design

PCR primers should be designed to match an LTR sequence close to either its 5′ or 3′ end and oriented so that the amplification direction is toward the nearest end of the LTR. Generally, the design should be based on a sequence alignment for the representative LTRs from a particular family of elements and placed within the most conserved region for that family. For LTRs, it is often useful to test primers at several locations within the LTR or internal part of the retrotransposon and in both orientations—particularly if there is evidence for nested insertions in the genome. Primers can be placed directly at the end of the LTR facing outward, as long as they do not form dimers or loops. For primers placed at the edge of the LTR, one or more additional selective bases can be added at the 3′ end in order to reduce the number of amplification targets. For example, if the initial primer yields amplification products containing too many weak individual bands for confident analysis by gel electrophoresis, a second round of primer design with additional bases should be included.

Database searches can be used to find unannotated, native LTR sequences that match the characterized retrotransposons from other species (Tables 1 and 2). However, care should be taken when defining the ends of the LTRs. Generally, mapping of the RT-primer binding sites PBS and PPT is needed in order to define the LTR ends with confidence. Microsatellite primers for REMAP or ISSR should be designed according to two principles: first, the primer length should be between 19 and 22 bases; second, the last base at the 3′ end of the primer should be designed as a selective base that is absent in the repeat unit itself. Examples of LTR conservation and consequent primer design for LTRs and microsatellites are provided in Tables 1, 2 and 3.

Our primers were designed using the FastPCR software or online Java Web tools [52, 64–67] (*see* **Note 5**).

Table 1
ISSR primers

ID	Sequence ISSR primer	T_m (°C)[a]	CG (%)	Linguistic complexity (%)[b]
	Dinucleotide microsatellites:			
$(CA)_{10}G$	CACACACACACACACACAG	56.8	52.4	23
$(CA)_{10}T$	CACACACACACACACACAT	55.8	47.6	23
$(CA)_{10}A$	CACACACACACACACACAA	56.0	47.6	21
$(TG)_{10}G$	TGTGTGTGTGTGTGTGTGG	57.7	52.4	21
$(TG)_{10}C$	TGTGTGTGTGTGTGTGTGC	58.3	52.4	23
$(TG)_{10}A$	TGTGTGTGTGTGTGTGTGA	56.5	47.6	23
$(AG)_{10}G$	AGAGAGAGAGAGAGAGAGG	52.4	52.4	21
$(AG)_{10}C$	AGAGAGAGAGAGAGAGAGC	53.1	52.4	23
$(AG)_{10}T$	AGAGAGAGAGAGAGAGAGT	51.5	47.6	23
$(AC)_{10}G$	ACACACACACACACACACG	58.1	52.4	23
$(AC)_{10}C$	ACACACACACACACACACC	57.7	52.4	21
$(AC)_{10}T$	ACACACACACACACACACT	56.4	47.6	23
$(GA)_{10}T$	GAGAGAGAGAGAGAGAGAT	50.4	47.6	23
$(GA)_{10}C$	GAGAGAGAGAGAGAGAGAC	51.6	52.4	23
$(GA)_{10}A$	GAGAGAGAGAGAGAGAGAA	50.6	47.6	21
$(GT)_{10}T$	GTGTGTGTGTGTGTGTGTT	56.0	47.6	21
$(GT)_{10}C$	GTGTGTGTGTGTGTGTGTC	56.8	52.4	23
$(GT)_{10}A$	GTGTGTGTGTGTGTGTGTA	55.2	47.6	23
	Tri-nucleotide microsatellites:			
$(CTC)_6G$	CTCCTCCTCCTCCTCCTCG	56.3	68.4	30
$(CTC)_6T$	CTCCTCCTCCTCCTCCTCT	54.4	63.2	24
$(CTC)_6A$	CTCCTCCTCCTCCTCCTCA	54.7	63.2	30
$(GAC)_6C$	GACGACGACGACGACGACC	59.7	68.4	30
$(GAC)_6T$	GACGACGACGACGACGACT	58.3	63.2	32
$(GAC)_6A$	GACGACGACGACGACGACA	58.6	63.2	30
$(CAC)_6G$	CACCACCACCACCACCACG	60.0	68.4	30
$(CAC)_6T$	CACCACCACCACCACCACT	58.1	63.2	30
$(CAC)_6A$	CACCACCACCACCACCACA	58.4	63.2	24
$(ACC)_6G$	ACCACCACCACCACCACCG	60.7	68.4	30
$(ACC)_6T$	ACCACCACCACCACCACCT	58.9	63.2	30

(continued)

Table 1
(continued)

ID	Sequence ISSR primer	T_m (°C)[a]	CG (%)	Linguistic complexity (%)[b]
$(ACC)_6C$	ACCACCACCACCACCACCC	60.4	68.4	24
$(ACA)_6G$	ACAACAACAACAACAACAG	48.5	36.8	30
$(ACA)_6T$	ACAACAACAACAACAACAT	47.6	31.6	30
$(ACA)_6C$	ACAACAACAACAACAACAC	48.9	36.8	24

[a]Oligonucleotide concentration is 200 nM, 0 mM Mg^{2+} [65]
[b]Sequence linguistic complexity, nucleotide arrangement, and composition [65]

3 Methods

3.1 PCR Protocol for IRAP, REMAP, and iPBS

The method described below applies to standard *Taq* DNA polymerase. PCR products can be separated using an agarose gel electrophoresis protocol. Alternatively, if fluorescent-labeled primers are used following *Taq*I digestion of PCR fragments, Fluorescent Image Analyzer gel systems may be employed. For separation on sequencing systems, fluorescent-labeled primers must be used; no special reaction conditions are needed.

1. The 25 μL reaction with *Taq* DNA polymerase should include the following: 25 ng DNA, 1× PCR buffer (containing 1.5–2.0 mM $MgCl_2$), 0.2–1 μM primer(s), 200 μM dNTPs, and 0.2 μL (1 U) *Taq* DNA polymerase (5 U/μL) (*see* **Note 6**).
2. Centrifuge all tubes or the plate before starting the PCR.
3. The PCR with *Taq* DNA polymerase (60 min total) should consist of the following steps: 3 min initial denaturation step at 95 °C; 30–32 cycles of 15 s at 95 °C, 20 s at 55°–72 °C, and 60 s at 72 °C; a 5 min final extension at 72 °C. The thermal cycling conditions can be varied without large effects on the resulting band pattern (*see* **Note 7**).
4. PCR product can be stored at 4 °C overnight.

3.2 Sample Preparation and Loading

1. Add an equal volume of 2× loading buffer to the completed PCR in tubes or plates and mix well.
2. Collect the mixture by a short centrifugation (turn a benchtop microcentrifuge on and immediately off).
3. Load the gels with a sample volume of 8–10 μL (*see* **Note 8**).

Table 2
Retrotransposon LTR primers

ID	Sequence LTR primer	TE name	T_m (°C)[a]	CG (%)	Linguistic complexity (%)
560	TTGCCTCTAGGGCATATTTCCAACA	*Wis2*	57.7	44.0	93
554	CCAACTAGAGGCTTGCTAGGGAC		58.6	56.5	80
2105	ACTCCATAGATGGATCTTGGTGA		54.1	43.5	88
2106	TAATTTCTGCAACGTTCCCCAACA		57.0	41.7	83
2107	AGCATGATGCAAAATGGACGTATCA	*Wilma*	57.2	40.0	84
833	TGATCCCCTACACTTGTGGGTCA		58.4	52.2	88
2108	AGAGCCTTCTGCTCCTCGTTGGGT		62.9	58.3	83
516	TCCTCGTTGGGATCGACACTCC		59.8	59.1	82
2109	TACCCCTACTTTAGTACACCGACA	*Daniela*	55.8	45.8	74
2110	TCGCTGCGACTGCCCGTGCACA		67.2	68.2	78
2111	CAGGAGTAGGGTTTTACGCATCC		57.3	52.2	88
2112	TGCTGCGACTGCCCGTGCACA		65.6	66.7	72
2113	TACGCATCCGTGCGGCCCGAAC		66.1	68.2	90
2114	GGACACCCCCTAATCCAGGACTCC	*Fatima*	61.8	62.5	76
2115	CAAGCTTGCCTTCCACGCCAAG		61.5	59.1	75
2116	CGAACCTGGGTAAAACTTCGTGTC		58.3	50.0	86
2117	AGATCCGCCGGTTTTGACACCGACA		63.9	56.0	81
432	GATAGGGTCGCATCTTGGGCGTGAC	*Sukkula*	63.5	60.0	93
480	GGAACGTCGGCATCGGGCTG		63.1	70.0	82
1319	TGTGACAGCCCGATGCCGACGTTCC		66.8	64.0	81
2123	GGAAAAGTAGATACGACGGAGACGT	*Wham*	58.0	48.0	70
483	TCTGCTGAAAACAACGTCAGTCC		57.5	47.8	80
1623	TGCGATCCCCTATACTTGTGGGT		58.6	52.2	90
552	CGATGTGTTACAGGCTGGATTCC	*Bagy1*	57.9	52.2	93
1369	TGCCTCTAGGGCATATTTCCAACAC	*BARE-1*	58.6	48.0	93
2015	ACCTGGATGCAACAGAGGTCTATG	*Cassandra*	57.7	50	93

[a]Oligonucleotide concentration is 200 nM, 0 mM Mg^{2+}

Table 3
PBS 18-mers primers

ID	Sequence	T_m (°C)[a]	CG (%)	Linguistic complexity (%)
2217	ACTTGGATGTCGATACCA	51.0	44.4	89
2218	CTCCAGCTCCGATTACCA	54.4	55.6	81
2219	GAACTTATGCCGATACCA	50.1	44.4	89
2220	ACCTGGCTCATGATGCCA	56.9	55.6	81
2221	ACCTAGCTCACGATGCCA	56.4	55.6	89
2222	ACTTGGATGCCGATACCA	54.0	50.0	86
2224	ATCCTGGCAATGGAACCA	54.4	50.0	83
2225	AGCATAGCTTTGATACCA	48.9	38.9	81
2226	CGGTGACCTTTGATACCA	52.6	50.0	83
2228	CATTGGCTCTTGATACCA	50.2	44.4	86
2229	CGACCTGTTCTGATACCA	52.0	50.0	83
2230	TCTAGGCGTCTGATACCA	52.4	50.0	92
2231	ACTTGGATGCTGATACCA	51.2	44.4	83
2232	AGAGAGGCTCGGATACCA	54.7	55.6	83
2237	CCCCTACCTGGCGTGCCA	62.8	72.2	78
2238	ACCTAGCTCATGATGCCA	53.6	50.0	83
2239	ACCTAGGCTCGGATGCCA	58.3	61.1	92
2240	AACCTGGCTCAGATGCCA	56.9	55.6	89
2241	ACCTAGCTCATCATGCCA	53.6	50.0	78
2242	GCCCCATGGTGGGCGCCA	67.0	77.8	67
2243	AGTCAGGCTCTGTTACCA	53.1	50.0	89
2244	GGAAGGCTCTGATTACCA	51.8	50.0	94
2245	GAGGTGGCTCTTATACCA	51.2	50.0	94
2246	ACTAGGCTCTGTATACCA	49.2	44.4	89
2249	AACCGACCTCTGATACCA	52.9	50.0	81
2251	GAACAGGCGATGATACCA	52.8	50.0	86
2252	TCATGGCTCATGATACCA	51.0	44.4	78
2253	TCGAGGCTCTAGATACCA	51.7	50.0	89
2255	GCGTGTGCTCTCATACCA	55.9	55.6	81
2256	GACCTAGCTCTAATACCA	47.8	44.4	81
2257	CTCTCAATGAAAGCACCA	50.8	44.4	83

(continued)

Table 3
(continued)

ID	Sequence	T_m (°C)[a]	CG (%)	Linguistic complexity (%)
2295	AGAACGGCTCTGATACCA	53.3	50.0	94
2298	AGAAGAGCTCTGATACCA	49.8	44.4	86
2373	GAACTTGCTCCGATGCCA	56.5	55.6	86
2395	TCCCCAGCGGAGTCGCCA	63.9	72.2	75
2398	GAACCCTTGCCGATACCA	55.4	55.6	86
2399	AAACTGGCAACGGCGCCA	61.8	61.1	75
2400	CCCCTCCTTCTAGCGCCA	59.5	66.7	75
2401	AGTTAAGCTTTGATACCA	46.3	33.3	92
2402	TCTAAGCTCTTGATACCA	47.5	38.9	89
2415	CATCGTAGGTGGGCGCCA	60.9	66.7	86

[a]Oligonucleotide concentration is 1000 nM, 0 mM Mg^{2+}

3.3 Casting the Agarose Gel

1. Prepare 200 mL of 1.2% (w/v) agarose containing 1× TBE buffer in a 500 mL bottle. This volume is required for one gel with the dimensions 0.4 cm × 20 cm × 20 cm.
2. Dissolve and melt the agarose in a microwave oven and then allow to slowly cool until its temperature drops to about 50–60 °C (*see* **Note 9**).
3. The ethidium bromide solution can be added at a rate of 50 μL per 200 mL, to bring the final concentration to 0.5 μg/mL. Alternatively, the gel can be stained at the end of the run (*see* **Note 7**).
4. Pour the agarose into the gel tray (20 × 20 cm) and set the gel combs. Allow the agarose to solidify at room temperature for at least 1 h.
5. Fill the chamber with 1× TBE running buffer until the gel is covered by about 3–5 mm of buffer.

3.4 Gel Electrophoresis

For a standard 20 × 20 cm gel, carry out electrophoresis at a constant 80–100 V for 5–9 h (in total, 700–900 volt-hours) (*see* **Note 10**).

3.5 DNA Visualization

DNA can be visualized directly by casting ethidium bromide into a gel as described above or by incubating in an ethidium bromide solution of equivalent strength following electrophoresis.

A high-quality gel scanner with good sensitivity and resolution is very important (*see* **Note 11**).

4 Notes

1. We have tested several *Taq* DNA polymerases, including DreamTaq DNA Polymerases (Thermo Fisher Scientific), One-Taq® (New England Biolabs), LongAmp® Taq DNA Polymerase (New England Biolabs), and FIREPol® (Solis BioDyne).
2. Small electrophoresis boxes and short gel trays are not suitable due to the large number of PCR products that need to be resolved. We routinely employ an apparatus with a run length of 20 cm.
3. This comb is ideal for analysis of any PCR amplification product or DNA restriction enzyme digest. The small space between the slots is important for analysis of banding patterns and for comparing lanes across the gel. Also, this comb thickness improves band resolution.
4. Such contaminants can chemically and mechanically block or inhibit chemical or enzymatic reactions, including denaturation and hybridization of nucleic acids; contaminants can also degrade or modify the nucleic acid. These contaminants include high-molecular-weight substances, such as polysaccharides and polyphenols, as well as substances of lower molecular weight, such as pigments, secondary metabolites, lipids, humic substances, and low-molecular-weight enzyme inhibitors or oligonucleotides. Therefore, in order to use the genomic DNA contained in biological materials, it is important that these substances are eliminated entirely from the sample.
5. It must be expected that not all primers (those derived from retrotransposons or ISSR primers) will work in the PCR. The genome may contain too few retrotransposon or microsatellite target sites, or these targets may be too dispersed for the generation of PCR products. Alternatively, sequence divergence in ancient retrotransposon insertions or polymorphisms between heterologous primers and native elements may lead to poor amplification. Some primers generate smears under all PCR conditions. Many sources can contribute to this problem, including primer structure, variability in the target site, and competition from other target sites. Generally, it is more efficient to design another primer than to try to identify the source of the problem. Furthermore, primers that produce a single, very strong band are not suitable for fingerprinting.
6. The PCR can be set up at room temperature. Prepare a master mix for the appropriate number of samples to be amplified, plus at least one additional sample [65]. After adding all components to the PCR master mix—with the DNA polymerase added last—mix well and then centrifuge. The reaction volume may vary from 10 to 25 μL; 10 μL is enough for running two

gels. The final primer concentration(s) in the reaction can vary from 200 to 500 nM for combined primers. For a single PCR primer, 400 nM should be used for IRAP and 1000 nM for iPBS amplification. Although higher primer concentrations increase PCR efficiency and the speed of DNA amplification, they also produce over-amplified products.

7. The time of the annealing step can vary from 10 to 30 s, and the annealing temperature depends on the melting temperature of the primer, which should be between 55 °C and 68 °C (60 °C is optimal for almost all primers and their combinations in IRAP and REMAP).
8. The DNA concentration plays an important role in gel resolution: overloaded lanes will result in poor resolution.
9. Note that the bottle should be closed, but the plastic cap must not be tightened. The agarose gel must be completely melted—small undissolved inclusions will severely hamper the quality of the results. Do not allow the gel to cool unevenly before casting, for example, by leaving it on the benchtop or in cool water. The best way to cool the agarose is by shaking it at 37 °C for 20 min. Careful casting of gels is critical for success. Small, undissolved agarose inclusions in the gels will result in bands with spiked smears. For optimal resolution, cast horizontal gels 3–4 mm thick. The volume of gel solution needed can be estimated by measuring the surface area of the casting chamber and multiplying it by gel thickness.
10. Select running conditions that are appropriate to the configuration of your electrophoresis box. Electrophoresis may cause the gels to deteriorate after several hours; their temperature should not exceed 30 °C, as electrophoretic resolution will be impaired at higher temperatures. The best results are obtained with a slower run. We routinely use 90 V for 7 h, or overnight (14 h) at 50 V (700 volt-hours). It is useful to check the run with a UV transilluminator toward the end of the run. For samples with many or large (> 500 bp) bands, the gel electrophoresis should be performed at a constant voltage of 50 V overnight (17 h).
11. Older video systems may be suitable for checking the success of restriction digests, cloning reactions, or simple PCR; however, they are not suitable for analysis of complex banding patterns. The gels can be scanned on an imaging system with a resolution of 50–100 μm; a digital gel electrophoresis scanner can detect ethidium bromide-stained nucleic acids by fluorescence using a green laser (532 nm).

Acknowledgments

This work was partially supported by the Government of Perm Krai, research project no. C-26/174.3 on January 31, 2019.

References

1. Kojima KK (2018) Structural and sequence diversity of eukaryotic transposable elements. Genes Genet Syst 94(6):233–252. https://doi.org/10.1266/ggs.18-00024
2. Naville M, Henriet S, Warren I, Sumic S, Reeve M, Volff JN, Chourrout D (2019) Massive changes of genome size driven by expansions of non-autonomous transposable elements. Curr Biol 29(7):1161–1168.e6. https://doi.org/10.1016/j.cub.2019.01.080
3. Neumann P, Navratilova A, Koblizkova A, Kejnovsky E, Hribova E, Hobza R, Widmer A, Dolezel J, Macas J (2011) Plant centromeric retrotransposons: a structural and cytogenetic perspective. Mob DNA 2(1):4. https://doi.org/10.1186/1759-8753-2-4
4. Belyayev A, Josefiová J, Jandová M, Kalendar R, Krak K, Mandák B (2019) Natural history of a satellite DNA family: from the ancestral genome component to species-specific sequences, concerted and non-concerted evolution. Int J Mol Sci 20(5). https://doi.org/10.3390/ijms20051201
5. Arkhipova IR, Yushenova IA (2019) Giant transposons in eukaryotes: is bigger better? Genome Biol Evol 11(3):906–918. https://doi.org/10.1093/gbe/evz041
6. Galindo-Gonzalez L, Mhiri C, Deyholos MK, Grandbastien MA (2017) LTR-retrotransposons in plants: engines of evolution. Gene 626:14–25. https://doi.org/10.1016/j.gene.2017.04.051
7. Serrato-Capuchina A, Matute DR (2018) The role of transposable elements in speciation. Genes (Basel) 9(5). https://doi.org/10.3390/genes9050254
8. Sharma A, Wolfgruber TK, Presting GG (2013) Tandem repeats derived from centromeric retrotransposons. BMC Genomics 14:142. https://doi.org/10.1186/1471-2164-14-142
9. Macas J, Novak P, Pellicer J, Cizkova J, Koblizkova A, Neumann P, Fukova I, Dolezel J, Kelly LJ, Leitch IJ (2015) In depth characterization of repetitive DNA in 23 plant genomes reveals sources of genome size variation in the legume tribe Fabeae. PLoS One 10(11):e0143424. https://doi.org/10.1371/journal.pone.0143424
10. Pollak Y, Zelinger E, Raskina O (2018) Repetitive DNA in the architecture, Repatterning, and diversification of the genome of *Aegilops speltoides* Tausch (Poaceae, Triticeae). Front Plant Sci 9:1779. https://doi.org/10.3389/fpls.2018.01779
11. Bilinski P, Han Y, Hufford MB, Lorant A, Zhang P, Estep MC, Jiang J, Ross-Ibarra J (2017) Genomic abundance is not predictive of tandem repeat localization in grass genomes. PLoS One 12(6):e0177896. https://doi.org/10.1371/journal.pone.0177896
12. Belyayev A, Kalendar R, Brodsky L, Nevo E, Schulman AH, Raskina O (2010) Transposable elements in a marginal plant population: temporal fluctuations provide new insights into genome evolution of wild diploid wheat. Mob DNA 1:6. https://doi.org/10.1186/1759-8753-1-6
13. Baumel A, Ainouche M, Kalendar R, Schulman AH (2002) Retrotransposons and genomic stability in populations of the young allopolyploid species Spartina anglica CE Hubbard (Poaceae). Mol Biol Evol 19(8):1218–1227. https://doi.org/10.1093/oxfordjournals.molbev.a004182
14. Kalendar R, Tanskanen J, Chang W, Antonius K, Sela H, Peleg O, Schulman AH (2008) Cassandra retrotransposons carry independently transcribed 5S RNA. Proc Natl Acad Sci U S A 105(15):5833–5838. https://doi.org/10.1073/pnas.0709698105
15. Kalendar R, Vicient CM, Peleg O, Anamthawat-Jonsson K, Bolshoy A, Schulman AH (2004) Large retrotransposon derivatives: abundant, conserved but nonautonomous retroelements of barley and related genomes. Genetics 166(3):1437–1450. https://doi.org/10.1534/genetics.166.3.1437
16. Smykal P, Kalendar R, Ford R, Macas J, Griga M (2009) Evolutionary conserved lineage of Angela-family retrotransposons as a genome-wide microsatellite repeat dispersal agent. Heredity 103(2):157–167. https://doi.org/10.1038/hdy.2009.45
17. Moisy C, Schulman AH, Kalendar R, Buchmann JP, Pelsy F (2014) The Tvv1 retrotransposon family is conserved between plant genomes separated by over 100 million years.

Theor Appl Genet 127(5):1223–1235. https://doi.org/10.1007/s00122-014-2293-z

18. Hosid E, Brodsky L, Kalendar R, Raskina O, Belyayev A (2012) Diversity of long terminal repeat retrotransposon genome distribution in natural populations of the wild diploid wheat *Aegilops speltoides*. Genetics 190(1): 263–U412. https://doi.org/10.1534/genetics.111.134643
19. Masuta Y, Kawabe A, Nozawa K, Naito K, Kato A, Ito H (2018) Characterization of a heat-activated retrotransposon in *Vigna angularis*. Breed Sci 68(2):168–176. https://doi.org/10.1270/jsbbs.17085
20. Ivancevic AM, Kortschak RD, Bertozzi T, Adelson DL (2016) LINEs between species: evolutionary dynamics of LINE-1 retrotransposons across the eukaryotic tree of life. Genome Biol Evol 8(11):3301–3322. https://doi.org/10.1093/gbe/evw243
21. Kalendar R, Schulman AH (2014) Transposon-based tagging: IRAP, REMAP, and iPBS. Methods Mol Biol 1115:233–255. https://doi.org/10.1007/978-1-62703-767-9_12
22. Kalendar R (2011) The use of retrotransposon-based molecular markers to analyze genetic diversity. Field Veg Crops Res 48(2): 261–274. https://doi.org/10.5937/ratpov1102261K
23. Piegu B, Bire S, Arensburger P, Bigot Y (2015) A survey of transposable element classification systems--a call for a fundamental update to meet the challenge of their diversity and complexity. Mol Phylogenet Evol 86:90–109. https://doi.org/10.1016/j.ympev.2015.03.009
24. Finnegan DJ (1990) Transposable elements and DNA transposition in eukaryotes. Curr Opin Cell Biol 2(3):471–477. https://doi.org/10.1016/0955-0674(90)90130-7
25. Kalendar R, Raskina O, Belyayev A, Schulman A (2020) Long tandem arrays of LTR retroelements in plants. Int J Mol Sci 21:2931. https://doi.org/10.3390/ijms21082931
26. Feschotte C, Jiang N, Wessler S (2002) Plant transposable elements: where genetics meets genomics. Nat Rev Genet 3(5):329–341. https://doi.org/10.1038/nrg793
27. Kapitonov VV, Jurka J (2008) A universal classification of eukaryotic transposable elements implemented in Repbase. Nat Rev Genet 9(5):411–412.; ; author reply 414. https://doi.org/10.1038/nrg2165-c1
28. Wu L, Gingery M, Abebe M, Arambula D, Czornyj E, Handa S, Khan H, Liu M, Pohlschroder M, Shaw KL, Du A, Guo H, Ghosh P, Miller JF, Zimmerly S (2018) Diversity-generating retroelements: natural variation, classification and evolution inferred from a large-scale genomic survey. Nucleic Acids Res 46(1):11–24. https://doi.org/10.1093/nar/gkx1150
29. Kapitonov V, Jurka J (1996) The age of Alu subfamilies. J Mol Evol 42. https://doi.org/10.1007/bf00163212
30. Kapitonov VV, Tempel S, Jurka J (2009) Simple and fast classification of non-LTR retrotransposons based on phylogeny of their RT domain protein sequences. Gene 448(2): 207–213. https://doi.org/10.1016/j.gene.2009.07.019
31. Kojima KK, Jurka J (2013) A superfamily of DNA transposons targeting multicopy small RNA genes. PLoS One 8. https://doi.org/10.1371/journal.pone.0068260
32. Bannert N, Kurth R (2004) Retroelements and the human genome: new perspectives on an old relation. Proc Natl Acad Sci U S A 101(Suppl 2):14572–14579. https://doi.org/10.1073/pnas.0404838101
33. Kalendar R, Amenov A, Daniyarov A (2019) Use of retrotransposon-derived genetic markers to analyse genomic variability in plants. Funct Plant Biol 46(1):15–29. https://doi.org/10.1071/fp18098
34. Kalendar RN, Aizharkyn KS, Khapilina ON, Amenov AA, Tagimanova DS (2017) Plant diversity and transcriptional variability assessed by retrotransposon-based molecular markers. Russ J Genet 21(1):128–134. https://doi.org/10.18699/vj17.231
35. Kalendar R, Flavell AJ, Ellis THN, Sjakste T, Moisy C, Schulman AH (2011) Analysis of plant diversity with retrotransposon-based molecular markers. Heredity 106(4): 520–530. https://doi.org/10.1038/hdy.2010.93
36. Smykal P, Bacova-Kerteszova N, Kalendar R, Corander J, Schulman AH, Pavelek M (2011) Genetic diversity of cultivated flax (*Linum usitatissimum* L.) germplasm assessed by retrotransposon-based markers. Theor Appl Genet 122(7):1385–1397. https://doi.org/10.1007/s00122-011-1539-2
37. Milovanov A, Zvyagin A, Daniyarov A, Kalendar R, Troshin L (2019) Genetic analysis of the grapevine genotypes of the Russian Vitis ampelographic collection using iPBS markers. Genetica 147(1):91–101. https://doi.org/10.1007/s10709-019-00055-5
38. Vuorinen A, Kalendar R, Fahima T, Korpelainen H, Nevo E, Schulman A (2018) Retrotransposon-based genetic diversity assessment in wild emmer wheat (*Triticum turgidum*

ssp. dicoccoides). Agronomy 8(7):107. https://doi.org/10.3390/agronomy8070107

39. Abdollahi Mandoulakani B, Yaniv E, Kalendar R, Raats D, Bariana HS, Bihamta MR, Schulman AH (2015) Development of IRAP- and REMAP-derived SCAR markers for marker-assisted selection of the stripe rust resistance gene Yr15 derived from wild emmer wheat. Theor Appl Genet 128(2):211–219. https://doi.org/10.1007/s00122-014-2422-8
40. Li S, Ramakrishnan M, Vinod KK, Kalendar R, Yrjälä K, Zhou M (2020) Development and deployment of high-throughput retrotransposon-based markers reveal genetic diversity and population structure of Asian bamboo. Forests 11(1):31. https://doi.org/10.3390/f11010031
41. Ghonaim M, Kalendar R, Barakat H, Elsherif N, Ashry N, Schulman AH (2020) High-throughput retrotransposon-based genetic diversity of maize germplasm assessment and analysis. Mol Biol Rep. https://doi.org/10.1007/s11033-020-05246-4
42. Tanhuanpää P, Erkkilä M, Kalendar R, Schulman AH, Manninen O (2016) Assessment of genetic diversity in Nordic timothy (*Phleum pratense* L.). Hereditas 153(1):5. https://doi.org/10.1186/s41065-016-0009-x
43. Tenhola-Roininen T, Kalendar R, Schulman AH, Tanhuanpaa P (2011) A doubled haploid rye linkage map with a QTL affecting alpha-amylase activity. J Appl Genet 52(3):299–304. https://doi.org/10.1007/s13353-011-0029-1
44. Tanhuanpaa P, Kalendar R, Laurila J, Schulman AH, Manninen O, Kiviharju E (2006) Generation of SNP markers for short straw in oat (*Avena sativa* L.). Genome 49(3):282–287. https://doi.org/10.1139/g05-100
45. Antonius-Klemola K, Kalendar R, Schulman AH (2006) TRIM retrotransposons occur in apple and are polymorphic between varieties but not sports. Theor Appl Genet 112(6):999–1008. https://doi.org/10.1007/s00122-005-0203-0
46. Tanhuanpaa P, Kalendar R, Schulman AH, Kiviharju E (2008) The first doubled haploid linkage map for cultivated oat. Genome 51(8):560–569. https://doi.org/10.1139/G08-040
47. Kan YW, Dozy AM (1978) Polymorphism of DNA sequence adjacent to human beta-globin structural gene: relationship to sickle mutation. Proc Natl Acad Sci U S A 75(11):5631–5635. https://doi.org/10.1073/pnas.75.11.5631
48. Waugh R, McLean K, Flavell AJ, Pearce SR, Kumar A, Thomas BB, Powell W (1997) Genetic distribution of Bare-1-like retrotransposable elements in the barley genome revealed by sequence-specific amplification polymorphisms (S-SAP). Mol Gen Genet 253(6):687–694. https://doi.org/10.1007/s004380050372
49. Kalendar R, Grob T, Regina M, Suoniemi A, Schulman A (1999) IRAP and REMAP: two new retrotransposon-based DNA fingerprinting techniques. Theor Appl Genet 98(5):704–711. https://doi.org/10.1007/s001220051124
50. Schulman AH, Kalendar R (2005) A movable feast: diverse retrotransposons and their contribution to barley genome dynamics. Cytogenet Genome Res 110(1-4):598–605. https://doi.org/10.1159/000084993
51. Vicient CM, Jaaskelainen MJ, Kalendar R, Schulman AH (2001) Active retrotransposons are a common feature of grass genomes. Plant Physiol 125(3):1283–1292. https://doi.org/10.1104/pp.125.3.1283
52. Kalendar R, Khassenov B, Ramanculov E, Samuilova O, Ivanov KI (2017) FastPCR: an in silico tool for fast primer and probe design and advanced sequence analysis. Genomics 109(3-4):312–319. https://doi.org/10.1016/j.ygeno.2017.05.005
53. Sorkheh K, Dehkordi MK, Ercisli S, Hegedus A, Halasz J (2017) Comparison of traditional and new generation DNA markers declares high genetic diversity and differentiated population structure of wild almond species. Sci Rep 7(1):5966. https://doi.org/10.1038/s41598-017-06084-4
54. Pearce SR, Stuart-Rogers C, Knox MR, Kumar A, Ellis TH, Flavell AJ (1999) Rapid isolation of plant Ty1-copia group retrotransposon LTR sequences for molecular marker studies. Plant J 19(6):711–717. https://doi.org/10.1046/j.1365-313x.1999.00556.x
55. Hirochika H, Hirochika R (1993) Ty1-copia group retrotransposons as ubiquitous components of plant genomes. Jpn J Genet 68(1):35–46. https://doi.org/10.1266/jjg.68.35
56. Flavell AJ, Dunbar E, Anderson R, Pearce SR, Hartley R, Kumar A (1992) Ty1-copia group retrotransposons are ubiquitous and heterogeneous in higher plants. Nucleic Acids Res 20(14):3639–3644. https://doi.org/10.1093/nar/20.14.3639
57. Ellis THN, Poyser SJ, Knox MR, Vershinin AV, Ambrose MJ (1998) Polymorphism of insertion sites of Ty1-copia class retrotransposons and its use for linkage and diversity analysis in pea. Mol Gen Genet 260(1):9–19. https://doi.org/10.1007/PL00008630
58. Witte CP, Le QH, Bureau T, Kumar A (2001) Terminal-repeat retrotransposons in miniature

(TRIM) are involved in restructuring plant genomes. Proc Natl Acad Sci U S A 98(24): 13778–13783. https://doi.org/10.1073/pnas.241341898

59. Kalendar R, Shustov AV, Seppänen MM, Schulman AH, Stoddard FL (2019) Palindromic sequence-targeted (PST) PCR: a rapid and efficient method for high-throughput gene characterization and genome walking. Sci Rep 9(1):17707. https://doi.org/10.1038/s41598-019-54168-0
60. Kalendar R, Antonius K, Smykal P, Schulman AH (2010) iPBS: a universal method for DNA fingerprinting and retrotransposon isolation. Theor Appl Genet 121(8):1419–1430. https://doi.org/10.1007/s00122-010-1398-2
61. Doungous O, Kalendar R, Filippova N, Ngane BK (2020) Utility of iPBS retrotransposons markers for molecular characterization of African Gnetum species. Plant Biosyst 154(5): 587–592. https://doi.org/10.1080/11263504.2019.1651782
62. Debladis E, Llauro C, Carpentier MC, Mirouze M, Panaud O (2017) Detection of active transposable elements in *Arabidopsis thaliana* using Oxford Nanopore sequencing technology. BMC Genomics 18(1):537. https://doi.org/10.1186/s12864-017-3753-z
63. Qiu F, Ungerer MC (2018) Genomic abundance and transcriptional activity of diverse gypsy and copia long terminal repeat retrotransposons in three wild sunflower species. BMC Plant Biol 18(1):6. https://doi.org/10.1186/s12870-017-1223-z
64. Kalendar R, Lee D, Schulman AH (2011) Java web tools for PCR, in silico PCR, and oligonucleotide assembly and analysis. Genomics 98(2):137–144. https://doi.org/10.1016/j.ygeno.2011.04.009
65. Kalendar R, Lee D, Schulman AH (2014) FastPCR software for PCR, in silico PCR, and oligonucleotide assembly and analysis. Methods Mol Biol 1116:271–302. https://doi.org/10.1007/978-1-62703-764-8_18
66. Kalendar R, Muterko A, Shamekova M, Zhambakin K (2017) In silico PCR tools for a fast primer, probe, and advanced searching. Methods Mol Biol 1620:1–31. https://doi.org/10.1007/978-1-4939-7060-5_1
67. Kalendar R, Tselykh TV, Khassenov B, Ramanculov EM (2017) Introduction on using the FastPCR software and the related Java web tools for PCR and oligonucleotide assembly and analysis. Methods Mol Biol 1620:33–64. https://doi.org/10.1007/978-1-4939-7060-5_2

Chapter 16

Introduction to Population Genomics Methods

Thibault Leroy and Quentin Rougemont

Abstract

High-throughput sequencing technologies have provided an unprecedented opportunity to study the different evolutionary forces that have shaped present-day patterns of genetic diversity, with important implications for many directions in plant biology research. To manage such massive quantities of sequencing data, biologists, however, need new additional skills in informatics and statistics. In this chapter, our objective is to introduce population genomics methods to beginners following a learning-by-doing strategy in order to help the reader to analyze the sequencing data by themselves. Conducted analyses cover several main areas of evolutionary biology, such as an initial description of the evolutionary history of a given species or the identification of genes targeted by natural or artificial selection. In addition to the practical advices, we performed re-analyses of two cases studies with different kind of data: a domesticated cereal (African rice) and a non-domesticated tree species (sessile oak). All the code needed to replicate this work is publicly available on github (https://github.com/ThibaultLeroyFr/Intro2PopGenomics/).

Key words Whole-genome sequencing, Pool-seq, Nucleotide diversity, Molecular evolution, Genome scans, Population structure, Admixture, Artificial and natural selection, Bioinformatics, Perseverance

1 Introduction

Population genetics is an increasingly important discipline at the interface between genetics and evolutionary biology focusing on the analysis of DNA variation and evolution across different loci and populations. Population genetic concepts help to understand the contribution of key evolutionary forces (mutation, migration, genetic drift, and natural selection) to the observable present-day distribution of genetic diversity. Prior to describe how various important and long-standing questions in plant biology can be addressed using population genetic concepts (for plant breeding, plant conservation biology, plant ecology for example), it is important to notice that a major shift occurred in this discipline. Over the last decade, cost-effective and high-throughput sequencing methods have accelerated and amplified the interest for population genetics by taking advantage of large-scale comparisons of DNA

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_16,

sequences or large sets of single nucleotide polymorphisms (SNPs) to better understand the contribution of the different evolutionary forces to the present-day DNA variation, leading to the emergence of a closely related field, called *population genomics* ([1] for a historical retrospective).

Biologists have now access to very large amounts of sequencing data. This change makes new investigations possible but also induces a considerable shift in the professional skills needed to generate (wet lab) or analyze the data (dry lab). Indeed, large-scale sequencing projects with several hundreds or thousands of samples sequenced have considerably shifted the limits in plant research (e.g., 3000 Rice Genomes Project [2]; *Arabidopsis thaliana* 1001 Genomes Project [3]). These new investigations require additional skills in biology, especially regarding the bioinformatic analysis of the sequencing data (e.g., a strong experience in using command-line versions and high-performance computing clusters, a proficiency in scripting or programming, a solid competence in statistical methods) to be able to handle such big genomic data projects. This greater transdisciplinarity between genetics, informatics, and statistics can make access to population genetics more difficult. In this chapter, our main objective is to tackle this issue by providing a simple and step-by-step guide. Unlike many great academic writings in the field (e.g., [4]), this chapter is not interested at covering the basis of the theory of evolution, but rather at *introducing population genomics methods to beginners* following a "learning-by-doing" strategy. All the genomic data we used are publicly available, as well as our scripts (*see* Subheading 2 below).

Population genetics is a broad discipline, and we do not claim to be exhaustive. Our objective is rather to introduce population genomics by focusing on some key analyses: the analysis of population structure, the inference of population splits and exchanges, and the detection of footprints of natural or artificial selection. We hope that some plant biologists, including students, will discover the benefits of population genomics analyses, including its applications for breeding and conservation, despite the fact that this discipline is, rightly or wrongly, reputed to be particularly difficult and demanding.

2 Materials

This tutorial requires the use of command-line software (preferentially on high-performance computing clusters) and some basic knowledge about Linux and bash commands (e.g., cd, mkdir, cp, paste, awk, grep). There are plenty of good tutorials available on the Internet to learn these aspects in a couple of hours, such as the Ryan Chadwick's website (https://ryanstutorials.net).

Due to space constraints, the code and commands are not described in this book chapter. However, all our scripts (bash, python, and R) are freely available on github: https://github.com/ThibaultLeroyFr/Intro2PopGenomics/.

This code repository is therefore an essential and complementary part of this chapter.

These scripts require different softwares:

1. BayPass: http://www1.montpellier.inra.fr/CBGP/software/baypass/download.html.
2. BWA mem: http://bio-bwa.sourceforge.net/.
3. Blast+: https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastDocs&DOC_TYPE=Download.
4. Bowtie 2: http://bowtie-bio.sourceforge.net/bowtie2/index.shtml.
5. FastStructure: https://rajanil.github.io/fastStructure/.
6. GATK: https://software.broadinstitute.org/gatk/download/.
7. Plink: https://www.cog-genomics.org/plink2/.
8. Picard: https://broadinstitute.github.io/picard/.
9. R: https://cran.r-project.org/
   (Rstudio is not mandatory but can be useful: https://www.rstudio.com/products/rstudio/download/)
   including R packages:
   (a) ape: https://cran.r-project.org/web/packages/ape/index.html.
   (b) circlize: https://cran.r-project.org/web/packages/circlize/index.html.
   (c) ggplot2: https://cran.r-project.org/web/packages/ggplot2/index.html.
   (d) pcadapt: https://cran.r-project.org/web/packages/pcadapt/index.html.
   (e) poolfstat: https://cran.r-project.org/web/packages/poolfstat/index.html.
   (f) reshape2: https://cran.r-project.org/web/packages/reshape2/index.html.
   (g) SNPRelate: https://bioconductor.org/packages/release/bioc/html/SNPRelate.html.
10. SAMtools: http://samtools.sourceforge.net/.
11. Seq_stat to compute nucleotide diversity and Tajima's D: https://tinyurl.com/yxurjgdx.
12. TreeMix: https://bitbucket.org/nygcresearch/treemix/downloads/.

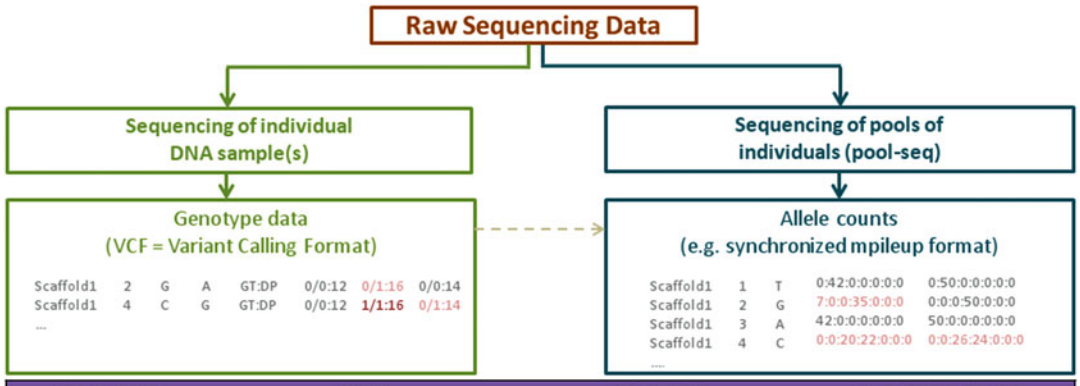
13. Trimmomatic: https://github.com/timflutre/trimmomatic.
14. VCFtools: http://vcftools.sourceforge.net/.
15. wget: https://www.gnu.org/software/wget/.

3 Methods

After introducing notions related to the handling of large sequencing data, we will provide guidelines to perform population genomic analyses based on two publicly available data from two different species: African rice from [5] and sessile oaks from [6]. These two examples were selected to cover broad plant biology-related issues, with both crop- and wild flora-associated topics. In addition, these two studies used different kind of sequencing data: individual-based genotypes vs. pooled DNA samples (a mixture of the DNA from several individuals prior to sequencing, hereafter *pool-seq*). As shown in Fig. 1, all analyses described in the analyses of the pool-seq data are based on the allele frequencies and can also be performed for individual-based data, at least when a minimum of 12–15 individuals were sequenced per population. In other words, analyses based on pool-seq data are far more limited than individual-based sequencing data, but pool-seq represents a cheaper strategy than the sequencing of individuals (*see* Subheading 3.3.1). Our analyses focus on plant species, but it has to be noted that such analyses can also be used to analyze various non-plant datasets, at least for diploid eukaryotic species.

3.1 From Raw DNA Data to Genetic Variants

1. Reads: All genomic projects start from the sequencing of very small pieces of DNA generated by a DNA sequencer, called *reads*. Despite recent advances in sequencing technologies (hereafter NGS, "for next-generation sequencing") to generate long fragments (up to 100,000 bases or more, e.g., Oxford Nanopore or PacBio technologies), these technologies remain, at the time of writing, too expensive to sequence multiple individuals of a given population in order to describe the genetic variation observed within this population. Moreover, long read technologies typically have a high error rate, that can negatively affect the accuracy of some population genomic analyses. Such new technologies therefore remain little used in population genomics projects. Most population genomicists rather use huge quantities of very short—but affordable—sequencing reads (e.g., Illumina sequencing of both ends of a short DNA fragment, so-called *paired-end reads*, generating 100–300 bases of known sequence for each end).
2. FASTQ file structure: High-throughput sequencing instruments generally output sequences under a FASTQ format. A FASTQ file is a text file with n repeats of four lines, with



MAIN GOALS	TASKS	GENOTYPES (INDIVIDUAL-BASED)	ALLELE FREQUENCIES (POOL-SEQ BASED)
Reconstruct the evolutionary history of a given species	Population structure	PCA, fastSTRUCTURE	Covariance Matrix
	Population splits & history of gene flow	TreeMix dadi/ABC	TreeMix dadi/ABC
	Historical changes in population sizes	PSMC/MSMC/SMC++ dadi/ABC	dadi/ABC
Detect key genes involved in: - plant adaptation - speciation - domestication - crop improvement	Detection of local adaptation	PCAdapt	BayPass
	Selective sweeps	Tajima's D/SweeD/RAiSD	Tajima's D (Popoolation suite)
	Genotype-Environment or Phenotype Association (GEA/GPA)	LFMM2/BAYPASS/GWAS/ Redundancy analysis (RDA)	BayPass, RDA

Fig. 1 Data format and analyses using individual versus pooled samples (i.e., DNA of several individuals mixed prior to sequencing, hereafter pool-seq). All analyses can be performed with individual data (dotted arrow), but the pool-seq data have limitations (*see* also Subheading 3.3.1 for the advantages and disadvantages of pool-seq). Methods or programs shown in green are those used in the following sections

n depending on the total number of generated reads. The first line begins with a "@" (equivalent of a ">" for a FASTA sequence) which indicates a new sequence. This line then contains a unique sequence identifier. The second line corresponds to the sequencing read itself, i.e., the succession of the different DNA bases read by the sequencer instrument. The third line generally only contains a "+" character. The fourth line corresponds to the quality values for the corresponding bases in second line, in the exact same order. In other words, the DNA sequencer provides a confidence score in the assignment of the corresponding base call. The very first step of a population genomic project is therefore to exclude low-quality reads and bases from these raw FASTQ files, in order to eliminate the majority of sequencing errors, a process commonly referred to as *read trimming*.

3. Read mapping to reference genome: All along this chapter, we assume that a reference genome is already available for the species you are interested in (or at least a closely related one). If not, the best solution is to start by generating a high-quality de novo genome assembly (this step ideally requires to establish a close collaboration with an experienced bioinformatician). If so, trimmed reads are then "mapped" against a reference genome in order to find the most likely genomic location for a read sequence, a process hereafter referred to as *read mapping*. A read mapper is not strictly speaking a read alignment software. The read mapper tries to find the best location (s) for a given read, but without establishing the base-to-base correspondence with the reference sequence. It might seem surprising but can be explained by a complex time-sensitivity trade-off. Any increase in the sensitivity of the mapping heavily slows down the speed of execution. To remain computationally efficient, particularly with extremely high volumes of sequence data, the two most commonly used read mappers, BWA [7, 8] and Bowtie 2 [9], identify the potential loci of origin of a sequencing read, but without performing precise local alignments. For short read data, these softwares remain fast and accurate methods, but it remains important to bear this limit in mind, especially in the future when reads will increase in length.
4. Variant calling: The identification of genetic variants from NGS data, hereafter *variant calling*, requires the accumulation of several reads at the same location, to increase the confidence in the identification of polymorphisms. Such methods generally predict the likelihood of variation at each locus to take into account some sequencing or mapping biases. Current population genomic studies are generally based on short polymorphisms, either SNP or short indels (insertions and deletions). Large structural variations (e.g., large indels, translocations, duplications) represent a non-negligible part of the genetic variation but remain quite difficult to access with the commonly used short-read data. This specific genetic diversity is therefore not addressed in the following sections.

3.2 Case Study 1: Individual-Based Genotyping

3.2.1 African Rice

Plant domestication might appear at first sight to be a simple and abrupt transition from a wild ancestor to a domesticated species. Following this view, it is generally assumed that only a part of the phenotypic (and genetic) diversity of the ancestral species has been used by the early farmers and therefore has contributed to the newly domesticated one, generating a so-called *domestication bottleneck*. As a consequence, theoretical work predicts that domestication is associated with a reduction of the genetic variation and a higher *mutation load*, i.e., an increase in the number of deleterious

alleles. This prediction is empirically supported in several plant or animal species [10]. For most domesticated species, domestication can be viewed as a long transitional process over millennia rather than a sudden event. This induces several other layers of complexity (reviewed in [11]), such as the possibility for (1) past and/or contemporary gene flow between wild and domesticated species, (2) several wild contributors, (3) several centers of domestication, and (4) massive changes in census and *effective population sizes* (*Ne*) of either the wild, the domesticated, or both species. All these situations are expected to have substantial impacts on neutral diversity and can generate confounding patterns leading to inappropriate conclusions.

In this section, we decided to use huge NGS data from the domesticated African rice (*Oryza glaberrima*). This species is characterized by a small genome (<350 Mb) and a simple organization (diploid), at least for a plant species. In addition, Cubry et al. [5] recently investigated the evolutionary history of this species through a large sequencing projects of 83 wild (*Oryza barthii*) and 163 domesticated individuals. This study represents an excellent and detailed piece of work. To speed up computations and help the reader to replicate this work, we have focused on a subset of 23 wild and 25 domesticated individuals from the center of domestication (as identified by Cubry et al. [5], corresponding to present-day Mali, Ghana, Niger, Nigeria, Benin, and Togo).

3.2.2 Variant Discovery from Publicly Available Data

1. Databases: Before downloading publicly available sequence from the Sequence Read Archive (SRA) or the European Bioinformatics Institute (EMBL-EBI), a close reading of the webpage associated to the project can provide considerable useful information about the data. Both the SRA and the EMBL-EBI website give relatively similar information, but from our perspective, the EMBL-EBI website is more user-friendly (Fig. 2). In the search bar, enter the ID of a project (e.g., ERP023549 for the African rice). To have an overview of the data, click on the associated project (for the African rice project: IRIGIN for International RIce Genome INitiative). The webpage contains a table with several fields by default: sample accession ID, species name, and some information relative to the sequencing instrument or the library protocol or different URL to download the data (Fig. 2). By selecting some additional columns, further information is available such as the number of reads or the sizes of the gzipped FASTQ files.

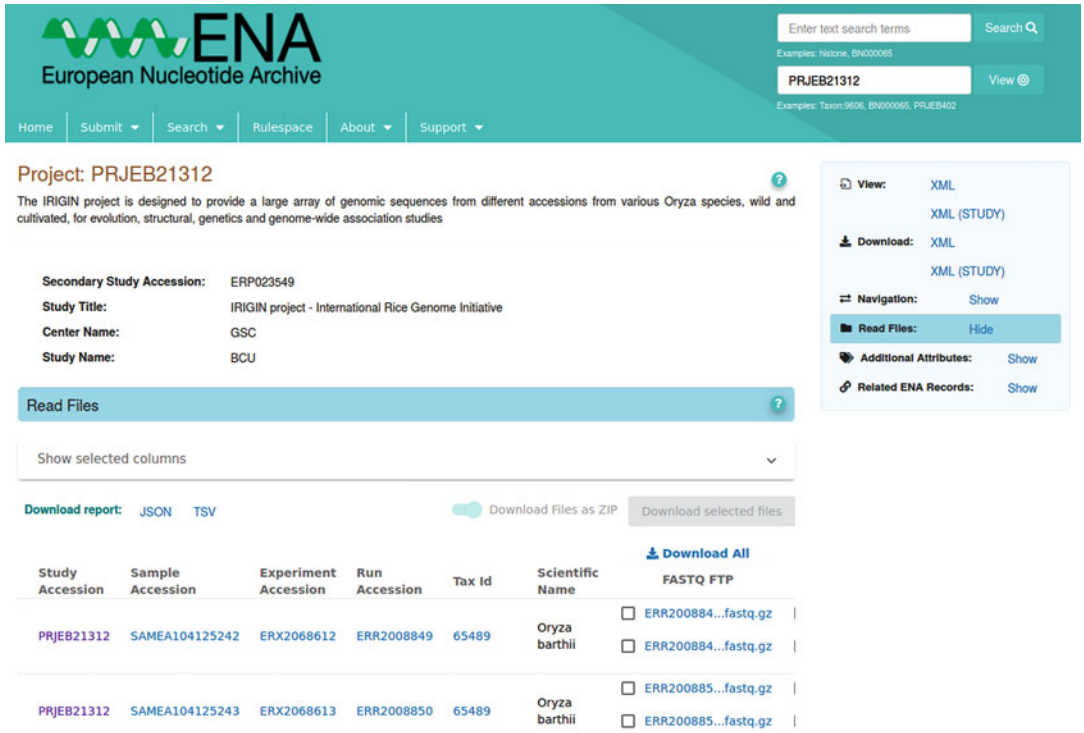


Fig. 2 Screenshot of the EMBL-EBI webpage for the African rice sequencing project described in Cubry et al. [5]

2. Data downloading to SNP dataset: To download the data, the best solution is to use a shell File Transfer Protocol (FTP) client such as wget. For example, the accession ERR2008855 can be downloaded from SRA servers using the following command in a terminal emulator: `wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR200/005/ERR2008855/ERR2008855_1.fastq.gz`

   `wget ftp://ftp.sra.ebi.ac.uk/vol1/fastq/ERR200/005/ERR2008855/ERR2008855_2.fastq.gz`

   And so on for all individuals you want to download.

   All our scripts are available to download and replicate all steps (including trimming, read mapping, and variant calling; *see* https://github.com/ThibaultLeroyFr/Intro2PopGenomics/tree/master/3.2.2/). In a nutshell, we use Trimmomatic to remove low-quality bases using a window computing average quality and sliding along the read, excluding all remaining bases of the read, if the average quality over four successive bases drops below 15. After excluding low-quality bases, reads with less than 50 remaining nucleotides are discarded. Then, we map all the remaining reads using BWA, remove duplicates with Picard, and perform the variant calling under GATK. We use the GATK HaplotypeCaller to first generate individual VCF file (gVCF for

genome Variant Call Format) and then perform the joint genotyping of the 48 individuals (joint VCF in Fig. 1). Low-quality SNPs are excluded, generating a set of 6,150,642 filtered SNPs (i.e., with a "PASS" label in the final VCF file).

3.2.3 Population Structure

Genetic differences between populations can be investigated by examining population structure—sometimes referred to as population stratification—which represents genome-wide differences in allele frequencies. Such a difference in genetic ancestry among individuals is possible because the samples can be derived from several populations that have experienced different demographic histories. As a consequence, all population genomics project first assess population genetic structure in order to take it into account in the downstream analyses. Aside from biological reasons, analyses of population structure allow to identify errors such as the accidental misidentification of some individuals arising during sample preparation, sequencing, or bioinformatics phases.

Given that this population structure represents a systematic shift in allele frequencies, a very large set of SNPs is unnecessary to investigate population structure patterns. A limited number of unlinked SNPs randomly selected across the entire genome (e.g., few thousands of SNPs with a low proportion of missing data) is sufficient to get an accurate picture of the population structure. Such genome complexity reduction is also more computationally efficient and reduces the number of variants in strong linkage disequilibrium (LD). LD represents a deviation from the hypothesis of random association of alleles within a genome and may impact the inferred population structure (*see* **Note 1**). Indeed, most popular population genetic tools use models assuming no or weak linkage disequilibrium within populations, including the most widely used model-based population genetics program STRUCTURE [12–14].

1. Principal component analysis (PCA): PCA is a commonly used exploratory analysis to infer population structure among individuals [15]. PCA helps to visualize genetic distance and relatedness between individuals by calculating principal components, with the top components explaining most of the differences among samples. In practice, PCA is sensitive to missing data. As a consequence, depending on the proportion of missing data in the VCF file (i.e., individuals with an unknown genotype: "./."), population geneticists either exclude all SNPs with missing data or replace missing values by the mean of the values based on the individuals with known calls. As a general rule, it is better to investigate population structure with few highly informative SNPs than using large proportion of poorly genotyped SNPs. This warning is

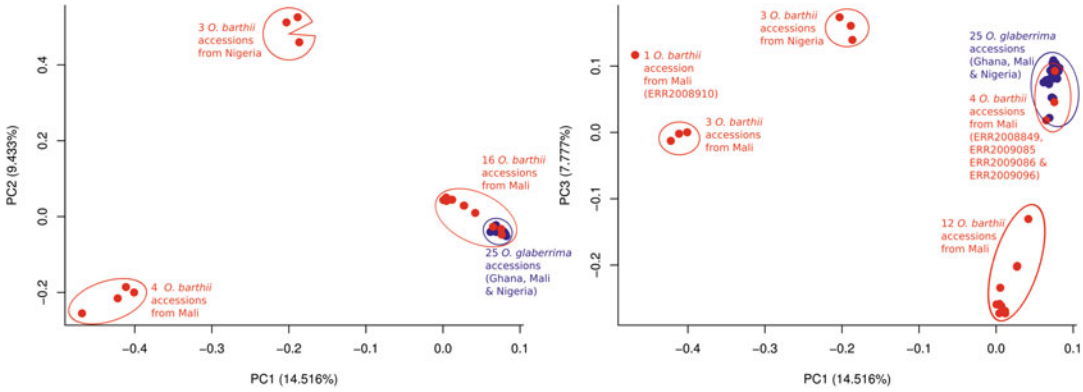


Fig. 3 Principal component analysis of the 48 investigated samples represented by dots (left: PC1 and PC2, right: PC1 and PC3). Geographical location and species labels are based on the information provided in Table S1 of Cubry et al. [5]

especially important for SNP set derived from Restriction site-Associated DNA data (RAD-seq data, [16]) which generally contain a large proportion of missing data.

The African rice project is based on massive Illumina sequencing data, leading to a VCF having very little missing data. As a consequence, we have chosen to remove all SNPs with missing data before performing PCA (i.e., *grep -v "\./\." [VCFfile]*). An example of PCA based on the 48 African rice samples is shown in Fig. 3. The first axis of the PCA accounts for 14.5% of the total variance and separates four wild individuals from present-day Mali and three wild individuals from Nigeria from all other samples. The second axis separates wild Nigerian samples from all other Malian samples. The third axis mostly separates 12 *O. barthii* samples from Mali. In summary, the PCA indicates different outcomes in the two species, with distinct population clusters observed in the wild species, while the domesticated species forms a single, relatively homogeneous, genetic group.

2. Bayesian clustering: In addition to PCA, *Bayesian clustering programs* assigning individuals to ancestral populations such as Structure [12–14], TESS 2 [17], and BAPS [18] are very popular tools to infer population structure. Some more recent methods used roughly similar method approach but are more adapted to large set of SNPs, e.g., FastStructure [19], LEA [20, 21], or TESS3 [22]. These methods *infer the admixture proportion* of each individual (Q-value) for a given number of ancestral populations ("K"). After a Plink transformation of the input file, we use the method implemented in FastStructure to provide an example based on the African rice data (Fig. 4). Assuming two ancestral populations ($K = 2$), FastStructure partially excludes seven wild *O. barthii* samples, including

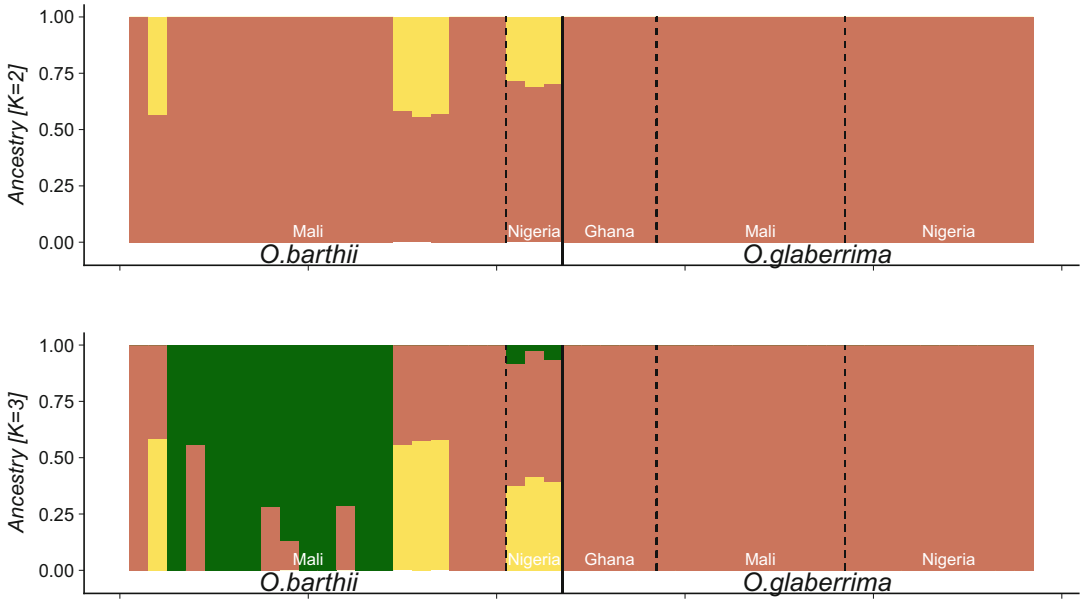


Fig. 4 Individual assignment to two (top) or three (below) genetic clusters by FastStructure. Each bar represents a single individual, with portions of the bar colored depending on the ancestry proportions estimated assuming $K = 2$ or $K = 3$. The number of subpopulations that maximize the marginal likelihood is 2 (*see* FastStructure manual for details). Geographical location and species labels are based on the information provided in Table S1 of Cubry et al. [5]

four from present-day Mali and three from Nigeria, from all other samples. The individual assignment of these seven samples suggests that these samples are admixed between the genetic cluster observed in all investigated cultivated samples (maroon) and an unknown genetic cluster (yellow). At $K = 3$, FastStructure infers a third group containing 12 samples from present-day Mali. PCA and FastStructure have generated very concordant results concerning these 48 African rice samples. Both analyses already suggest some complexity in the evolutionary history of the African rice.

3.2.4 *Diversity*

Nucleotide diversity greatly varies along the genome, with more genetic variation in intergenic regions than in genes. This general pattern reflects varying degrees of natural selection acting on the genome, from neutral regions that do not positively or negatively affect the organism's ability to survive and reproduce (i.e., fitness) to genes under strong negative or positive selection. Negative selection refers to the purging of *deleterious alleles* at functionally constrained genes because individuals with deleterious alleles are selected against and therefore contribute less to the next generation than the average of the population. Reciprocally, positive selection refers to the rapid fixation of advantageous mutation because

individuals carrying this advantageous allele are expected to contribute more to the next generation. In both cases, it is important to keep in mind that the footprints of natural selection can extend to the vicinity of these regions because of linkage disequilibrium, over relatively long distance in regions of low recombination (generating so-called linked selection [23]).

Two important measures of nucleotide diversity are generally used in population genomics, the number of polymorphic sites (θ) and the mean proportion of nucleotide differences between different pairs of sequences randomly sampled in a population (π). The diversity of different groups of samples can then be compared. For example, a reduction of diversity (ROD) index can be estimated by computing $1 - \frac{\pi\text{Group1(e.g.domesticated)}}{\pi\text{Group2(e.g.wild)}}$. Such ratios are particularly meaningful for different research questions associated to plant conservation or plant breeding. For instance, the total genetic diversity loss since the onset of plant domestication (or along a plant breeding program) can be investigated by comparing wild and domesticated species (e.g., wheat [24]). Based on a comparison of the 23 *Q. barthii* and 25 *Q. glaberrima* samples, an overall ROD of 0.327 is estimated, suggesting that 32.7% of the *Q. barthii* diversity was lost during the domestication or breeding process. Genomic heterogeneity in ROD is also informative to identify important genes, particularly regions with very high ROD estimates (ROD exceeding 0.8 in red, Fig. 5). Those regions with remarkably reduced levels of nucleotide diversity in the domesticated species as compared to the wild progenitor species can be informative about candidate genomic regions (including genes) that have been subjected to strong artificial selection during domestication or breeding.

A great statistical property of π and θ (to be strictly accurate, π and $\frac{\theta}{a1}$, where $a_1 = \sum_{i=1}^{n-1} \frac{1}{i}$) is that these two statistics are equal in values assuming mutation-drift equilibrium and constant population size ($d = \pi - \frac{\theta}{a1} = 0$, *see* [25]). Any excess or lack of rare alleles in the population, however, creates deviations from zero because π tends to underestimate the number of mutations that are rare in the population. As a consequence, the difference between the two estimators is a commonly used measure to evaluate nonequilibrium demographic situation such as population expansion (generating an excess of rare alleles, overall negative Tajima's *D* value) or population contraction (generating a lack of rare alleles, overall positive *D* value).

By observing the genomic heterogeneity in Tajima's *D*, the footprints of natural and artificial selection can also be revealed in some specific regions of the genome. Positive values can be observed if selection maintains variation in some specific regions (balancing selection). Strongly negative values are informative about recent selection that has removed neutral variation

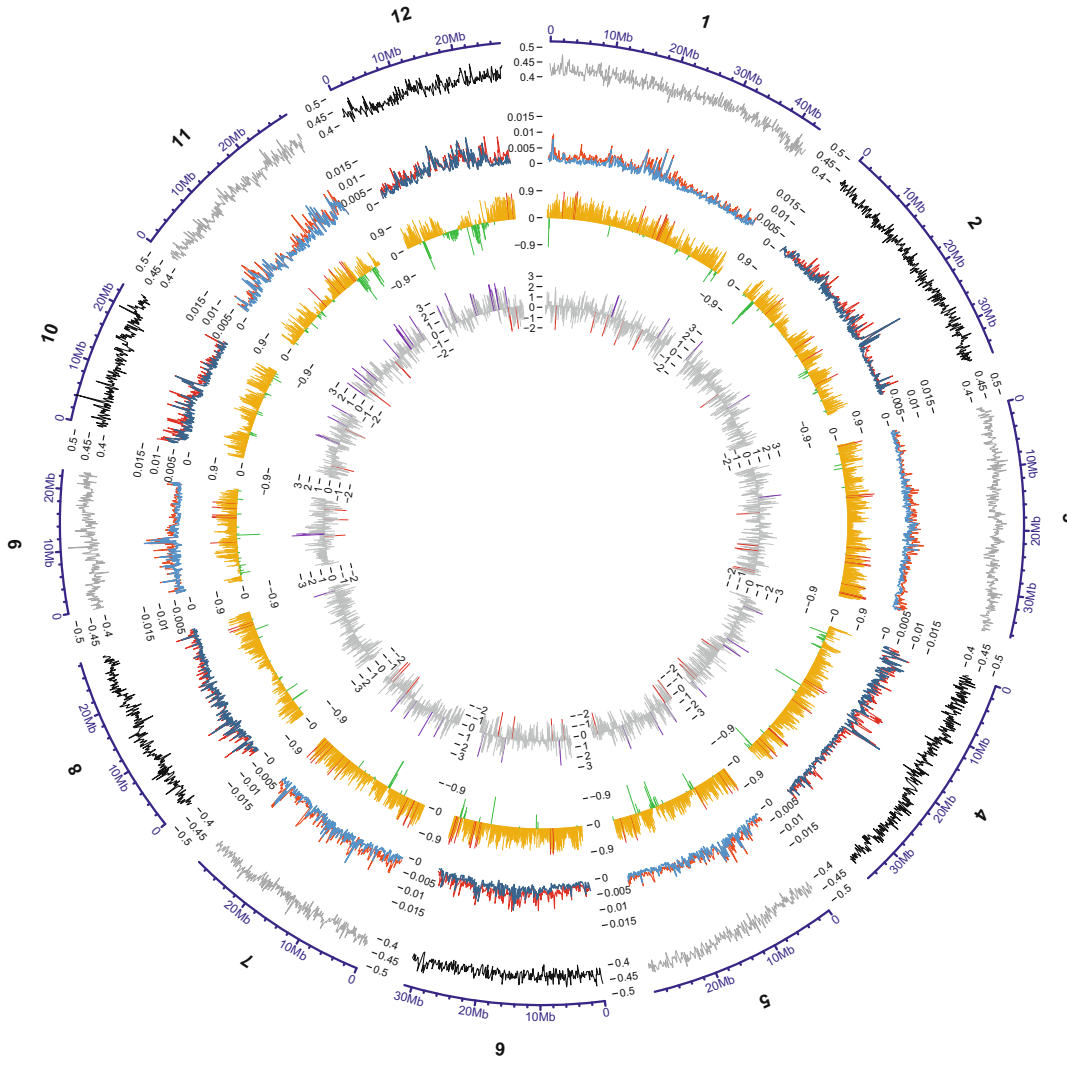


Fig. 5 Circular diagram showing different nucleotide diversity estimates for the two African rice species along the 12 chromosomes. From external to internal: GC content; π estimates (red = *O. barthii*, blue = *O. glaberrima*); reduction of diversity (ROD) to evaluate the difference in the domesticated *O. glaberrima* species as compared to the wild *O. barthii* (green = negative ROD values, orange = positive ROD values, red = positive ROD values exceeding 0.8); observed Tajima's *D* values for *O. glaberrima*. Tajima's *D* values are represented as a deviation from the median Tajima's *D* values observed over all sliding windows ($D = 0.171$). Values lower than −1.83 or greater than 2.17 are shown in red and purple, respectively. These threshold values correspond to the −2/+2 decision rule, which is a simple rule of thumb, but remain commonly used in practice to find some candidate regions under selection. All estimates are based on nonoverlapping 100-kb sliding windows

surrounding a selected site (i.e., a selective sweep). Negative Tajima's *D* values found in a domesticated species can therefore be informative about footprints of domestication and human selection (e.g., Fig. 5 for the case study on African rice).

3.2.5 Inferring Population Size History

Whole-genome sequence data are increasingly used to infer the history of a population, such as the historical changes in effective population sizes (*Ne*). *Ne* represents the number of breeding individuals in an idealized Wright-Fisher population that experiences similar amount of genetic drift than the real population (*see* [26] for a review). It may seem like an abstract concept, but the study of the evolution of *Ne* is particularly important in population genomics because *Ne* variation explains the dynamic of genetic diversity within a population (loss or gain) or the fixation of deleterious alleles (Subheading 3.2.6). Following the nearly neutral theory, the genetic diversity θ equals $4 \times Ne \times \mu$ (for a diploid species, where μ is the per-generation mutation rate). Assuming that μ remains constant over quite long periods of time, recent variation of θ only depends on the effective population size (*Ne*)—which captures the effect of genetic drift—with more chance for variants to be fixed by drift in small *Ne* as compared to large *Ne* populations.

To investigate this variation, many methods based on the coalescent theory are now available. Without going into details, a coalescence event occurs when two alleles merged into a single ancestral copy (i.e., the most recent common ancestor), when looking backward in time starting from the present. In other words, the coalescent theory models how genetic variants sampled from a given population may have originated from a common ancestor (*see* [27] for an introduction). By estimating the rate of coalescence during any period of time, it is therefore possible to infer population size changes. Over the last decade, these new methods have rapidly become popular to provide information about the factors driving genetic diversity of a given species, which is especially crucial for conservation-related issues. Major shifts in the evolutionary trajectories can be identified and potentially be correlated with the major climate change periods or with geological and anthropogenic disturbances.

The coalescent-based method implemented in SMC++ [28] is a good method currently available to reconstruct the history of *Ne*. This method is fast, easy to use, and efficient, even for analyzing tens or hundreds of unphased whole-genome sequences. We therefore performed a simple test based on the African rice dataset and observed considerable changes in past effective population size (Fig. 6).

As a limited number of individuals of the progenitor species had presumably been used by the early farmers and therefore contributed to the domesticated species, a drastic reduction in effective population size (*Ne*) at the onset of the domestication is generally assumed, which is commonly referred to as the domestication bottleneck. Similarly to the study of Cubry et al. [5], we inferred substantial changes in effective population size of the African rice over the last 100,000 years. Surprisingly, we were, however, unable to infer the expected reduction of *Ne* at the onset of the African rice

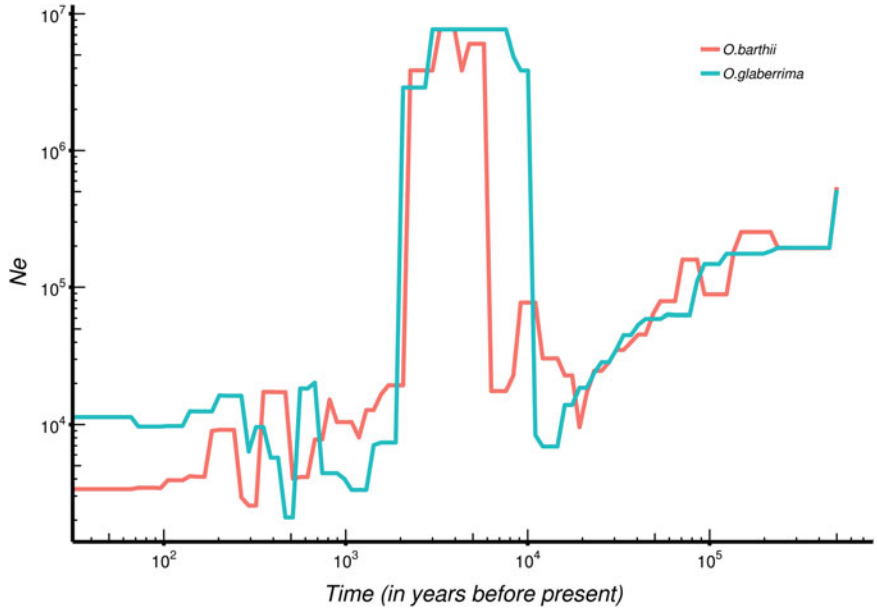


Fig. 6 Estimated changes in past effective population sizes (Ne) for *O. barthii* (red) and *O. glaberrima* (blue) inferred using the coalescent-based method SMC++

domestication, but rather we inferred an expansion between 2000 and 10,000 years ago. This lack of support for the domestication bottleneck can be due to a series of factors such as the reduced number of genomes used and the existence of long runs of homozygosity (masked in Cubry et al. [5]). As a consequence, the pattern we have recovered over the last 10,000 years should be interpreted with caution. This result is illustrative of the importance of remaining prudent when interpreting such inferences. Violations of some assumptions can substantially distort the inference of effective population size changes. SMC++, as well as similar methods (e.g., PSMC [29]; MSMC [30]), relies on the assumption of no external gene flow (originating from another population or species). This assumption is one of the most frequently violated. They also require high quality data (e.g. 30X sequencing depth or more). Some more advanced methods available to decipher more complex evolutionary histories (*see* **Note 2**), including several closely related species that have experienced different periods of gene flow, can also be helpful to provide additional statistical support for historical changes in *Ne* [31–33].

3.2.6 Deleterious Mutation Load

A downstream consequence of the domestication bottleneck is the higher load of deleterious mutations in the domesticated species as compared to the wild counterpart. Following the nearly neutral theory, neutral nucleotide diversity is expected to be reduced proportionally to the reduction in *Ne* because neutral variants have

more chance to be fixed by drift in small *Ne* as compared to large *Ne* populations (Subheading 3.2.5 above). For non-neutral variants (i.e., $s \neq 1$), fixation probabilities depend on the strength of selection and effective population size ($N_e s$, e.g., [34]). A domestication bottleneck is therefore expected to induce a shift in the balance between selection and drift, with drift playing a greater role after the bottleneck. This also holds true for deleterious mutations, particularly slightly deleterious mutations, which are therefore expected to accumulate more easily. In other words, the domestication bottleneck reduces the efficacy of purifying selection, the force which tends to remove harmful mutations. Domesticated plants are therefore expected to have an increased mutation load as compared to their wild progenitor species. This hypothesis is often referred to as the "cost of domestication" [35]. Some recent studies have provided considerable empirical support for this hypothesis, e.g., in maize [36], Asian rice [37], cassava [38], or wine [39].

In addition to the 23 and 25 WGS of *O. barthii* and *O. glaberrima*, we use sequencing data of three *Oryza meridionalis* (Australian wild rice individuals from [40]) and three *Oryza sativa* individuals (domesticated Asian rice individuals from [2]) to infer the ancestral allele of each SNP (the original non-mutated allele). In short, the recent phylogeny of the *Oryza* species based on the WGS data suggests that *O. meridionalis* had diverged from the common ancestor of African and Asian rice 2.4 million years ago. The divergence of African and Asian rice lineages is more recent (<1 million years ago; *see* [40] for details). Australian rice and Asian rice are used to infer the ancestral allele in order to count the number of derived alleles in the wild and the domesticated African rice. Based on the SNPs for which the ancestral allele was unambiguously determined, we identify more fixed derived alleles in *O. glaberrima*, as compared to *O. barthii* (1,050,545 and 825,826, respectively), which can be considered as another piece of proof supporting the hypothesis of a domestication bottleneck.

To look into more details the burden of deleterious genetic mutations, various methods are available. Simple methods such as the comparisons of ratios of the nucleotide diversity (or heterozygosity) at non-synonymous as compared to synonymous polymorphisms can be very relevant (e.g., between a wild progenitor and a domesticated species [41]). Indeed, most within-gene mutations changing the amino acid sequence are expected to be slightly or strongly disadvantageous (i.e., deleterious). Higher ratios of non-synonymous to synonymous polymorphisms are therefore informative of higher deleterious loads. In silico methods predicting the potential deleterious effects of mutations are more and more popular (e.g., SIFT [42]). Subsequently, we use the software PROVEAN (Protein Variation Effect Analyzer [43])

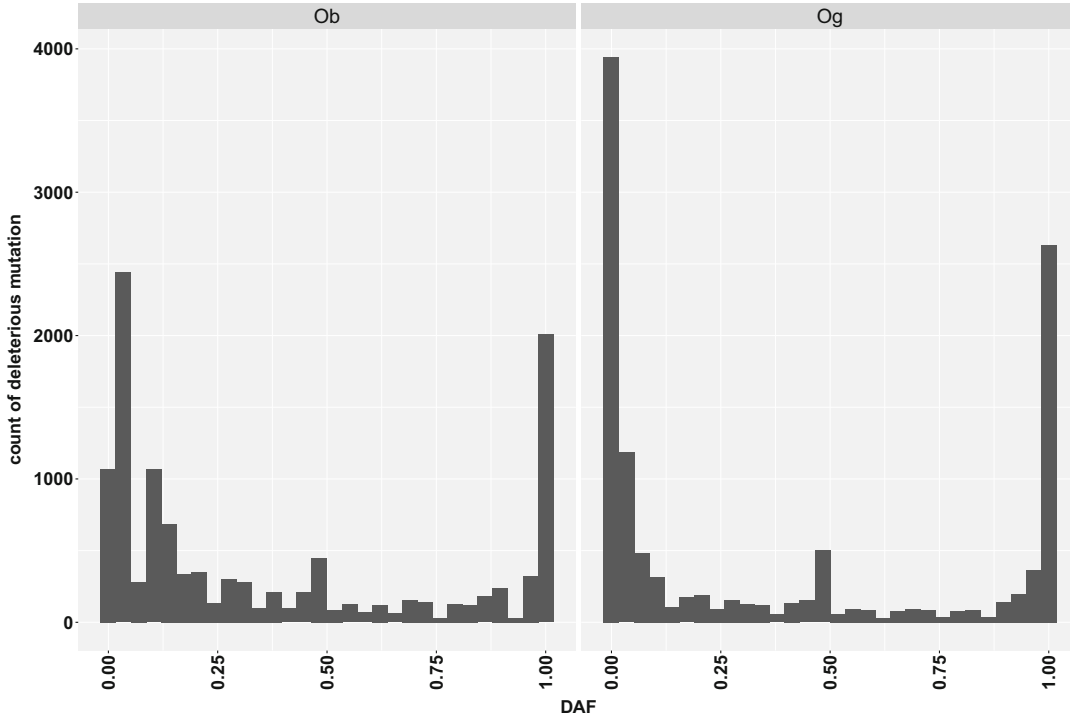


Fig. 7 Deleterious mutation loads in the wild *O. barthii* and the domesticated *O. glaberrima* species, as estimated using proteins of the African rice. DAF = deleterious allele frequencies

which performs local alignments (BLAST) against a protein database to predict whether an amino acid change in a given protein affects its function. A score is then computed based on the 30 best cluster hits. A negative PROVEAN score is indicative of a deleterious mutation.

This analysis requires different steps, which are detailed on github (https://github.com/ThibaultLeroyFr/Intro2PopGenomics/tree/master/3.2.6/Scripts_provean/). Before running PROVEAN, we have built an NCBI "nonredundant" (nr) database containing only proteins corresponding to monocot species. By limiting to monocotyledon species, our objective is to avoid spurious BLAST alignments against evolutionary distant species. Among a total of 120,324 candidate non-synonymous mutations passing PROVEAN filtering criteria, 18,369 mutations are predicted to be putatively deleterious mutations (score < -2.5). Among these 18,369 SNPs, the ancestral state is unambiguously determined for 11,829 variants (see above). Deleterious allele frequency spectra at these 11,829 putatively derived deleterious SNPs are generated for both the wild and domesticated species (Fig. 7). Interestingly, a higher mutation load in *O. glaberrima* as compared to *O. barthii* is identified, but this difference is relatively small. Our analyses are rather consistent

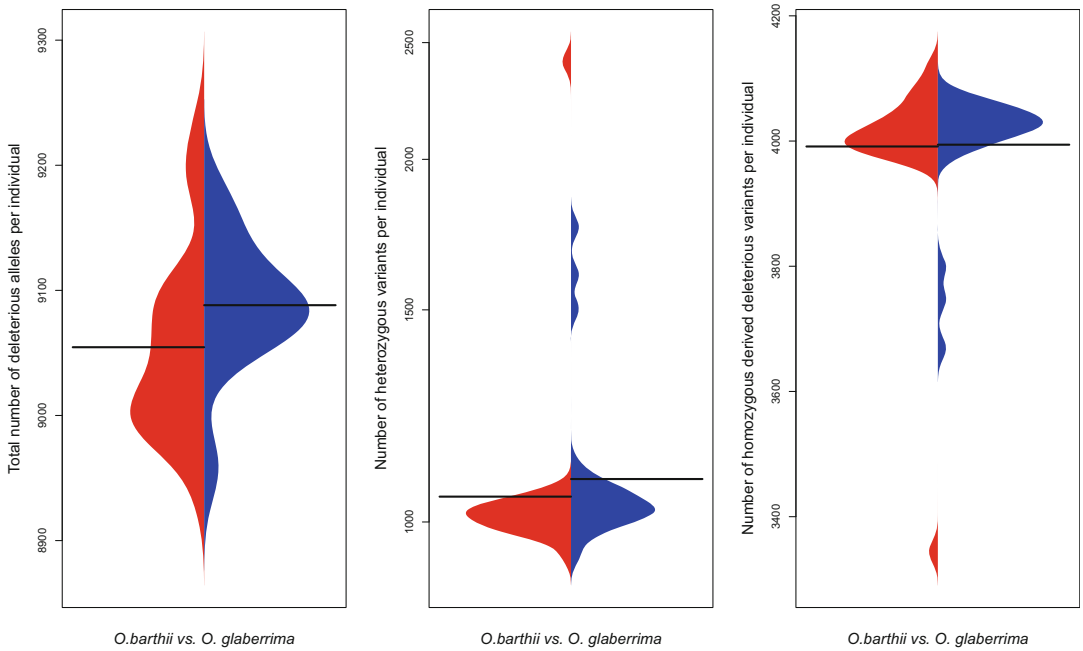


Fig. 8 Total numbers of deleterious alleles (left), heterozygous calls (center), and homozygous derived alleles per individual for *O. barthii* (red) and *O. glaberrima* (blue). The black bar indicates average per species

with a substantial deleterious mutation load in *O. barthii* and a slight increase in *O. glaberrima*, which can be compatible with the African rice domestication.

Looking at this difference more carefully, the number of deleterious alleles per individual is slightly higher in *O. glaberrima* (Fig. 8), but this difference seems to be more explained by a difference in heterozygous sites than by a strong difference in the number of homozygous deleterious variants. Because deleterious mutations tend to be recessive [44], such a limited difference in the number of homozygous variants therefore suggests that this higher mutation load may only induce a marginal fitness difference between the two species. This first investigation already gives an overview of the accumulation of deleterious variants, but some analyses are available to conduct more precise measurements [45–47].

3.2.7 F_{ST} and Genome Scans for Selection

The fixation index F_{ST} is probably the most widely used population genetic statistics. F_{ST} measures the differentiation between populations and ranges from 0 to 1. However, some slightly negative values can be observed in the case of uneven sample sizes and should be interpreted as a zero value. A value of zero indicates complete panmixia, i.e., free interbreeding between the two assumed populations resulting in no population structure or

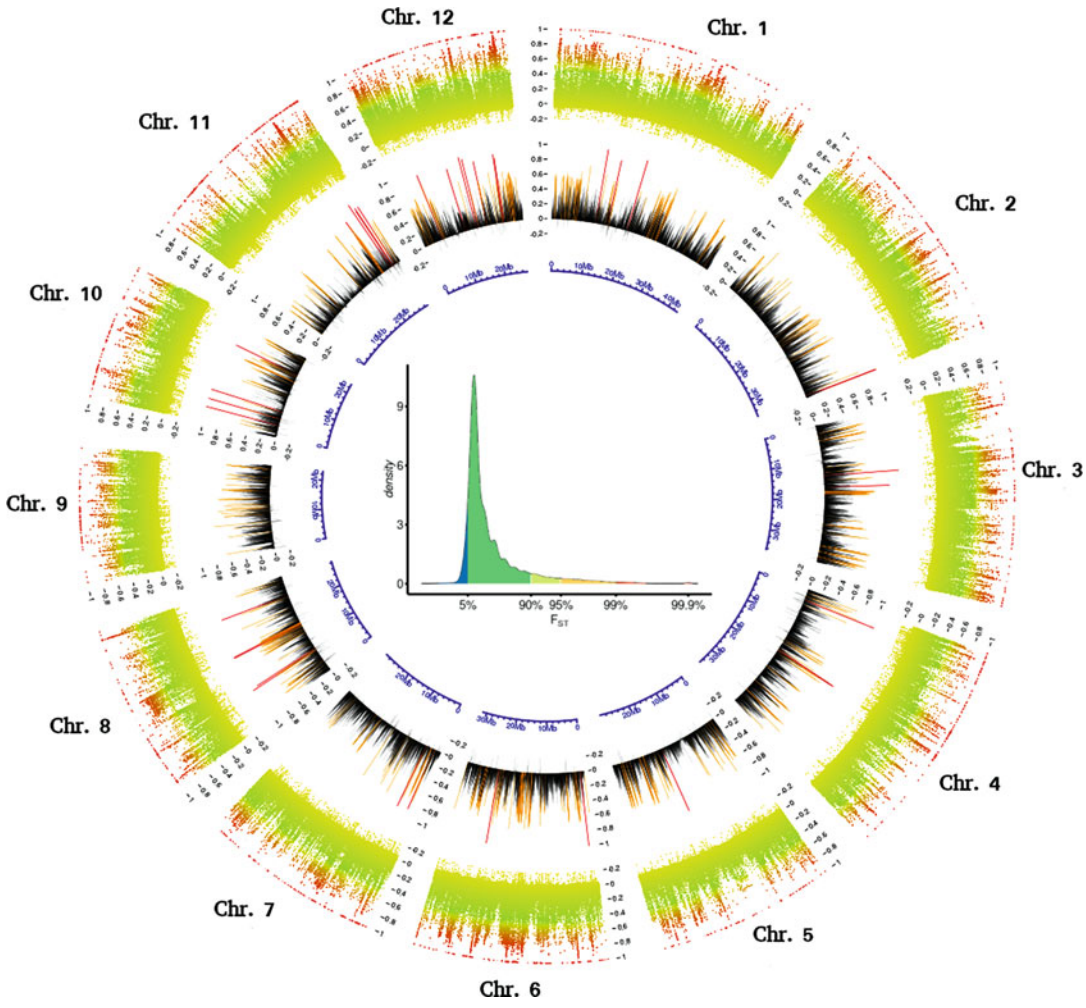


Fig. 9 Fixation index (F_{ST}) values as computed using VFCtools and estimated for each SNP (external circle) or for nonoverlapping 10-kb sliding windows (internal circle) over the 12 rice chromosomes. A color scale from yellow ($F_{ST} = 0$) to red ($F_{ST} = 1$) is used for the SNP-by-SNP F_{ST} estimates to illustrate the continuous variation in F_{ST} values. Empirical distribution of the observed F_{ST} values across all SNPs is shown in the center of this circular graph (corresponding F_{ST} values for the different quantiles: 5% = −0.03; 90% = 0.27; 95% = 0.41; 99% = 0.66; and 99.9% = 1.00)

subdivision. On the contrary, a value of 1 indicates that the two populations are homozygous for two different alleles (e.g., an SNP with genotypes A/A observed in all individuals of the first population and genotypes C/C for all individuals of the second population). In other words, the higher the F_{ST} value, the more different the allele frequencies in the two or more populations.

To give a better idea of how useful report of F_{ST} values can be, we computed F_{ST} between samples of *O. barthii* and *O. glaberrima* at two different genomic scales: on an SNP-by-SNP basis and using 10-kb sliding windows (Fig. 9).

The use of the empirical distribution of the among-locus variation in F_{ST} (Fig. 9) to identify loci that deviate from neutral expectations—and therefore representing candidate footprints for natural or artificial selection—is inspired by the seminal study of Lewontin and Krakauer [48]. Indeed, loci under balancing selection in the two populations are expected to exhibit lower F_{ST} values, while regions under diversifying selection are expected to exhibit larger differences in F_{ST} as compared to selectively neutral loci. Diversifying selection indeed triggers allele frequency changes over time in such a way of generating and maintaining high genetic differences in the two populations. In practice, identifying loci under balancing selection is a near-impossible task to achieve. Identifying *diversifying selection* remains a complex issue. The difficulty comes from the fact that the among-loci variation in F_{ST} is highly dependent on the demography of the investigated populations [49–53]. Over the last 20 years, considerable attention has been devoted to develop statistical approaches that partially address this challenge (e.g., [54, 55]; hereafter referred to as genome scans for selection). In this section, we introduce the use of pcadapt [56], an R package that is well suited to identify variants with large differences in allele frequencies between clusters of individuals. This package has several advantages. From the user's perspective, this solution is easy to use under an *R* environment, especially with the detailed tutorial available for this package. From a more computational and biological perspective, pcadapt is computationally efficient, and the analyses do not require to group individuals into populations—i.e., no prior information about the two or more populations, which can be a difficult task to achieve (e.g., Subheading 3.2.3 for the African rice). In addition, pcadapt can handle very large datasets and reports summary statistics in a reasonable computational time, offering an alternative to the genome scan methods based on a Bayesian framework, which are several orders of magnitude longer (*see* Subheading 3.3.6 for the use of a Bayesian method).

After a preliminary analysis revealing some regions of strong linkage disequilibrium (LD) in the African rice dataset, this dataset was pruned to remove SNPs in strong LD. Indeed, such an extent of LD is expected to have a considerable impact on the analysis (*see* **Note 1**). As a consequence, the dataset is first "pruned" to remove SNPs in strong LD, before computing the principal components and performing the outlier detection. Coordinates of individuals on the two principal components (PC) (Fig. 10, as compared to Fig. 3) are different after SNP pruning. This reduction of the LD likely improves the ability of the PCs to capture the genome-wide patterns reflecting ancestry differences, as commonly assumed [57]. The first PC mostly isolates samples from *O. barthii* and *O. glaberrima*, with the notable exception of four *O. barthii*

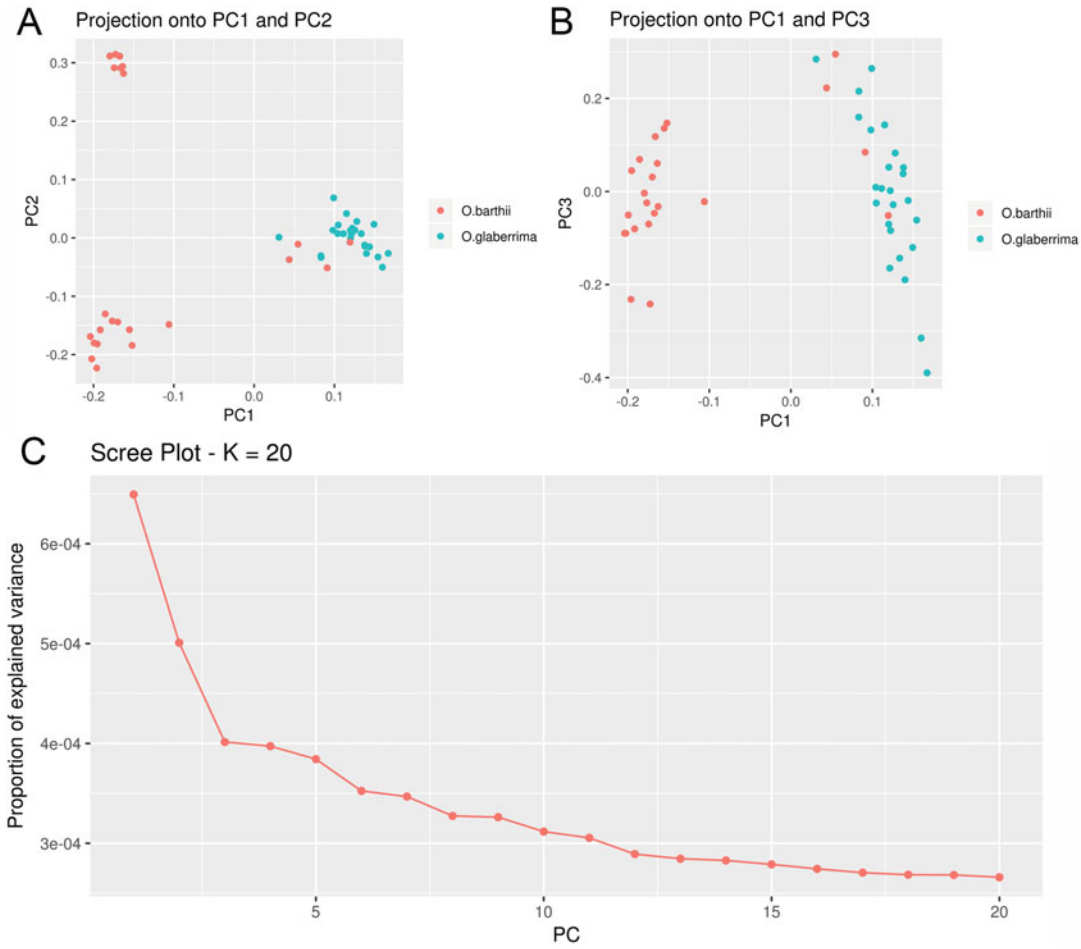


Fig. 10 Individual PCA and scree plot after LD thinning. (**a**) Coordinates of individuals on the two principal components. (**b**) Coordinates of individuals on the PC1 and PC3 (**c**) Scree plot (proportion of explained variance) for the 20 first PCs after LD thinning. Based on this scree plot, $K = 3$ was preferred

samples from *Mali*. Visual evaluation of the so-called scree plot [58] for PC1 to PC20 suggests that the three first components explain a substantial fraction of the total variance in the data, as compared to the 17 additional components that were also investigated (Fig. 10). As a consequence, we use the implemented method in pcadapt to scan genomes assuming these three components.

The genome positions of all outliers as shown in the so-called Manhattan plots (Fig. 11) reveal that they are distributed throughout the genome. SNPs deviating from neutral expectation and therefore potentially under selection are unexpected to have this distribution, since selection is unlikely to impact all the genome. These outputs are more consistent with a substantial background noise generating an excess of outliers. However, some genomic regions exhibiting hundreds of variants in several narrow genomic

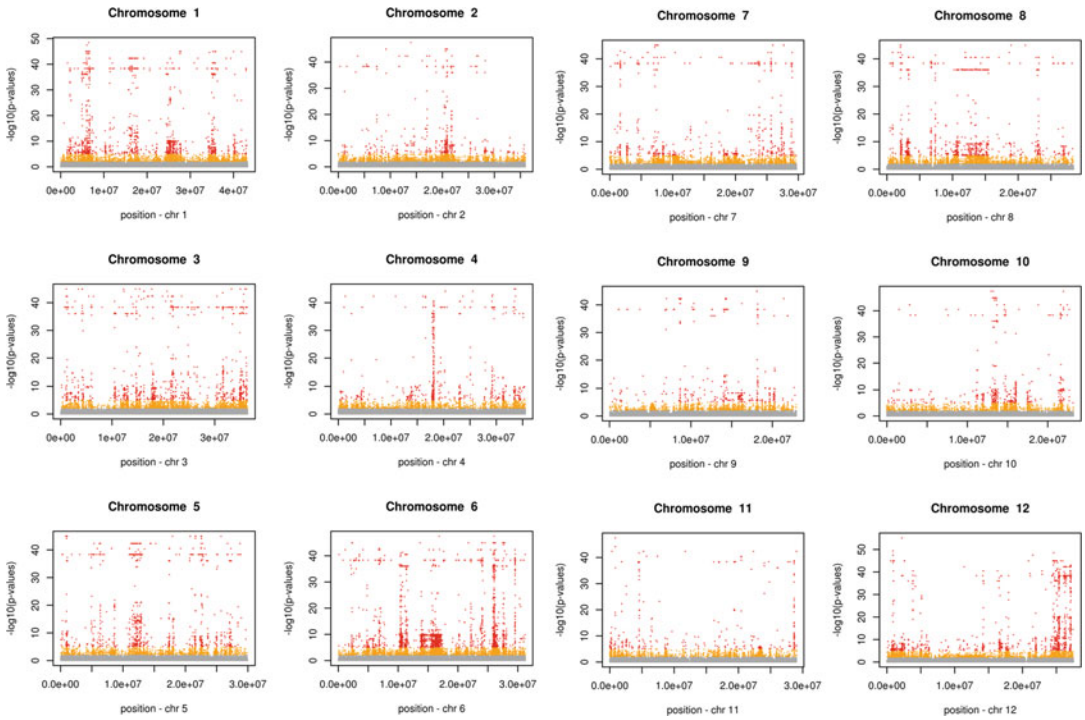


Fig. 11 Manhattan plots showing the chromosome position of each outlier detected using pcadapt and assuming $K = 3$. Score is expressed as $-\log_{10}(p\text{-values})$. SNPs with p-value <0.01 (i.e., $-\log_{10}(p\text{-values}) = 2$) are shown in orange, and those with p-value <0.00001 ($-\log_{10}(p\text{-values}) = 5$) are shown in red

regions, e.g., on chromosome 4 or 6 (Fig. 11), are more convincing. These regions therefore represent excellent candidate regions to identify the African rice domestication genes.

3.3 Case Study 2: Sessile Oak Populations

3.3.1 Pool-seq as a Cost-Efficient Method

For many plant species, the sequencing of hundreds or more individuals using an individual-based strategy represents a too expensive option. Consider, for example, the sequencing of 50 diploid individuals at reasonable sequencing coverage (20×)—the total sequencing effort would be around 1000×—in order to ensure accurate individual calls for all individuals. For some biological questions, the genotypes of all individuals are not truly necessary. Instead, accurate population estimates of the frequency of each allele along the genome can be sufficient [59]. In this case, a cost-effective alternative remains possible. The strategy is to first equimolarly mix the DNA of these 50 individuals prior to sequencing in order to sequence the pool at a lower coverage. Assuming that the pool is sequenced at 100× (so resulting in a tenfold drop in the sequencing cost), each chromosome is therefore expected to be sequenced only once, on average, which is low. But, given the total number of chromosome sequenced in the pool, the allele frequency estimated for the whole population is expected to be

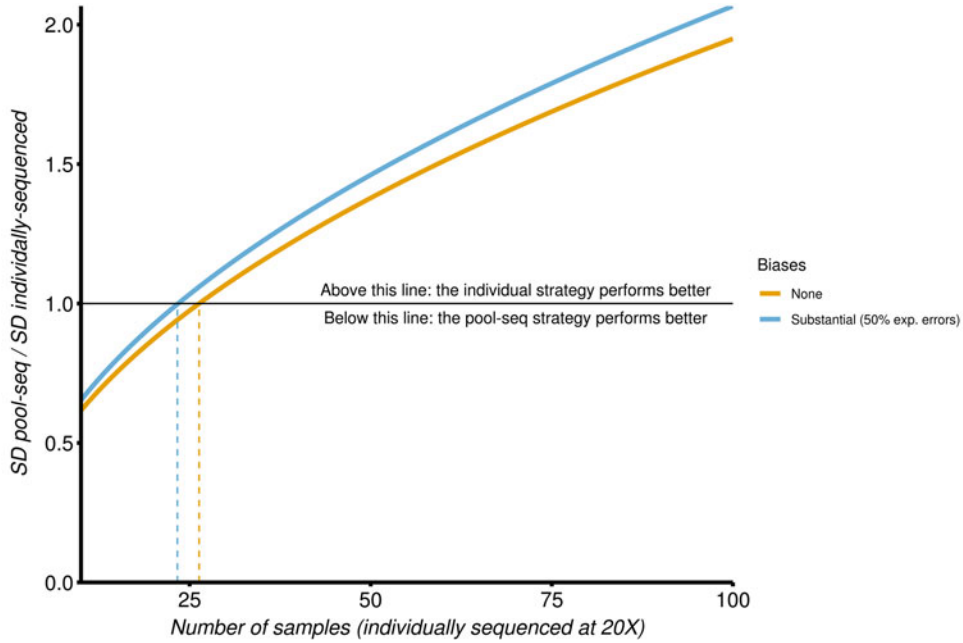


Fig. 12 Comparison of the accuracy in the allele frequency estimation between two strategies, as performed using PIF [60]: a pool-seq strategy of 50 individuals sequenced at a mean pool coverage of 100× and an individual-based genotyping strategy with a growing number of individuals sequenced at 20×. The tipping point is 26 individuals assuming no experimental biases. Even after considering some experimental biases, a pool-seq strategy of 50 individuals sequenced at a pool coverage of 100× is expected to outperform a design with 20 individuals sequenced at 20× (the equivalent of 400× of sequencing data; for details, *see* [60])

accurate. Based on mathematical derivations, Gautier et al. [60] provided theoretical support for this accuracy. These authors showed that the sequencing of DNA pools remains an efficient strategy under various realistic experimental designs. They also provide an easy-to-use tool (PIFs [60]) to optimize the experimental pool-seq design considering several parameters or experimental errors (e.g., pipetting biases).

Based on a rapid simulation using this tool and the number of individuals previously assumed (Fig. 12), it indicates that the sequencing of a pool of 50 individuals with a mean pool coverage of 100× is expected to generate as accurate allele frequency estimates as 26 individuals separately sequenced with a depth of coverage of 20× (the pool-seq strategy therefore reduces by five the sequencing costs). Even assuming substantial experimental error (50%) generating departure from equimolarity (i.e., a dispersion of individual contributions around the expected mean value assuming equal DNA quantities), the allele frequency estimates are expected to be roughly similar to those of 23 individuals separately sequenced, each with a depth of coverage of 20× (Fig. 12).

Fig. 13 Sessile oak distribution and climate variation. Left: European distribution map of *Q. petraea* created with QGIS from data made available by the European Forest Genetic Resources Programme (EUFORGEN [63]). *Right*: Sessile oak trees in the snow. Photo taken by T. Leroy on November 22, 2015, at an elevation of 1200 m in one of the French Pyrenees forests investigated in Leroy et al. [6] ("O12" population)

3.3.2 Population Genomics in Wild Sessile Oaks

The sessile oak (*Quercus petraea*), a species belonging to the European white oaks complex, is an example of plant species with an impressive amount of genomic resources, including huge pool-seq data [6, 61, 62]. Sessile oaks extend from Northern Spain to Southern Scandinavia, thus representing a large diversity of climatic conditions (Fig. 13). In South-West French Pyrenees, some sessile oak populations occur from lowlands to middle elevations (up to 1600 m, Fig. 13), with substantial differences in mean annual temperature (up to 7 °C) or in precipitation sums (a difference of up to 250 mm/year, [6] for details). In the subsequent sections, we perform a step-by-step reanalysis of the data used in Leroy et al. [6] to illustrate the possibilities of the pool-seq data. In this study, 18 pools were sequenced: ten sessile oak populations collected on a latitudinal gradient in Europe (including seven populations from France, two from Germany, and a population from Ireland) and eight sessile oak populations from an altitudinal gradient in the French Pyrenees (collected along two close valleys, with four populations per valley (100 m, 800 m, 1200 m, and 1600 m). The DNA of 20–25 individuals were equimolarly mixed prior to sequencing, except for the two populations at 1600 m for which only ten to 18 individuals were used (for details, *see* [6]). Analyses performed in this section are basically performed following the same strategy than in the original paper, but the analyses are simplified.

3.3.3 From Raw Sequencing Data to Allele Counts

The Illumina data can be downloaded from SRA or EMBL-EBI using the project ID PRJEB32209. We make available on github all the scripts used to download and perform the trimming and read mapping and to identify variants. The pipeline is roughly similar to

those used for the African rice data, at least for read trimming and mapping. A notable exception is the way in which variants are identified. As previously described (Subheading 3.1), variant calling methods have been developed to minimize the number of false-positive variants (e.g., sequencing errors). Indeed, each diploid individual possesses either two copies of the reference allele (homozygous for the same allele than the reference genome), one copy (heterozygous, with both a reference and an alternative allele), or none (homozygous for the alternative allele). In other words, the frequency of the reference allele estimated for each individual is expected to be close to 1, 0.5, or 0. When the coverage is high enough (>20), deviations from these situations can be informative of false-positive SNPs. In contrast, such investigations are impossible to perform with pool-seq data because DNA from several individuals are mixed prior to sequencing. As a consequence, only few parameters can be used to exclude false-positive SNPs, i.e., the minor allele frequency (MAF) and the depth of coverage at each position. Illumina sequencing errors are expected to be about 1% or less, so it is generally recommended to use a MAF that exceeds this value (e.g., 2% or more). Similarly, coverage is expected to vary across the genome following a Poisson distribution [64]. Extreme values in the observed distribution of coverage depth are also informative from some read-mapping biases inducing an excess or deficit of coverage compared to the expectations assuming this distribution. For example, highly covered regions can be due to reads corresponding to two genomic loci with almost similar sequences (e.g., recent duplications) aligning to a unique location of the reference sequence. Such regions therefore present a high risk of identifying false-positive SNPs. In practice, a matrix of allele counts (Fig. 1 and Table 1) contains both allele frequencies and coverages that can be used to filter variants.

One thing must be kept in mind, however: errors in pool-seq data are necessarily more numerous than in individual-based sequencing. Even after using some MAF or coverage thresholds,

Table 1
A hypothetical example of a read count matrix with two SNPs in rows

Chromosome	Position	Ref allele	Major allele (all populations)	Minor allele (all populations)	Major allele counts (pop1)	Minor allele counts (pop1)	Major allele counts (pop2)	Minor allele counts (pop2)
Chr1	47	G	G	C	75	30	49	55
Chr1	112	T	A	T	68	20	79	14

The two pools are assumed to be sequenced at a mean pool coverage of 100×. Allele frequencies can be easily derived from this matrix (e.g., $30/(75 + 30) = 0.29$ for the pop1 of the SNP Chr1:47)

the number of false-positive SNPs can remain substantial. Population-level estimates of nucleotide diversity can be greatly inflated, especially for species with low to extremely low genetic diversity, for which the noise-to-signal ratio can be high. In this section, we choose not to cover diversity-related analyses (including comparisons of estimators, e.g., Tajima's *D*) based on pool-seq data to call for caution. It must, however, be noted that some methods already exist (e.g., PoPoolation [65]) and some studies successfully reported similar range of estimates based both on individual and pool-seq datasets (e.g., oaks [62]).

3.3.4 Inferring the History of a Set of Populations

Allele frequencies are expected to be very informative about historical relatedness between populations. Indeed, two populations that have a recently shared history are expected to exhibit more similarities in allele frequencies because of a low influence of genetic drift, as compared to two genetically distant populations. As a consequence, inferring the history of a set of populations based on allele frequencies is expected to be possible. This is exactly what TreeMix [66] aims to do. This genetic tool infers the relationships among populations as a bifurcating tree, which can therefore be considered as an analogous to phylogenetic trees. To do so, the software first infers the variance-covariance matrix of allele frequencies between populations based on a large set of variants and then finds the maximum likelihood tree explaining most of the observed variance in relatedness between populations.

In the case of sessile oak, TreeMix computes the 18 × 18 variance-covariance matrix using a huge set of SNPs (37 million SNPs). Because the allele frequencies at nearby SNPs are expected to be highly correlated due to linkage disequilibrium (*see* **Note 1**), we set the parameter k to 1000 (blocks of 1000 SNPs) to take into account this bias. TreeMix therefore first estimates the variance-covariance matrix based on 37,062 blocks of 1000 SNPs.

Using the R scripts from the TreeMix suite, the total variance explained by a simple bifurcating tree can be estimated. Applied to the sessile oak dataset, drift alone accounted for more than 89% of the total variance in allele frequencies among populations. An example of phylogenetic visualization of the inferred best likelihood tree is shown in Fig. 14. As a first step, it provides a lot of information regarding the relatedness of populations. For example, sessile oak populations from the latitudinal gradient are genetically different from the populations from the altitudinal gradient, especially the six populations at the highest elevation (Fig. 14). The population from Ireland, however, departs from this general pattern, since this population is more related to Pyrenean populations at high elevation.

In the great majority of cases, a simple bifurcating tree cannot explain all the genetic variation observed in the variance-covariance matrix. TreeMix allows adding some additional edges connecting

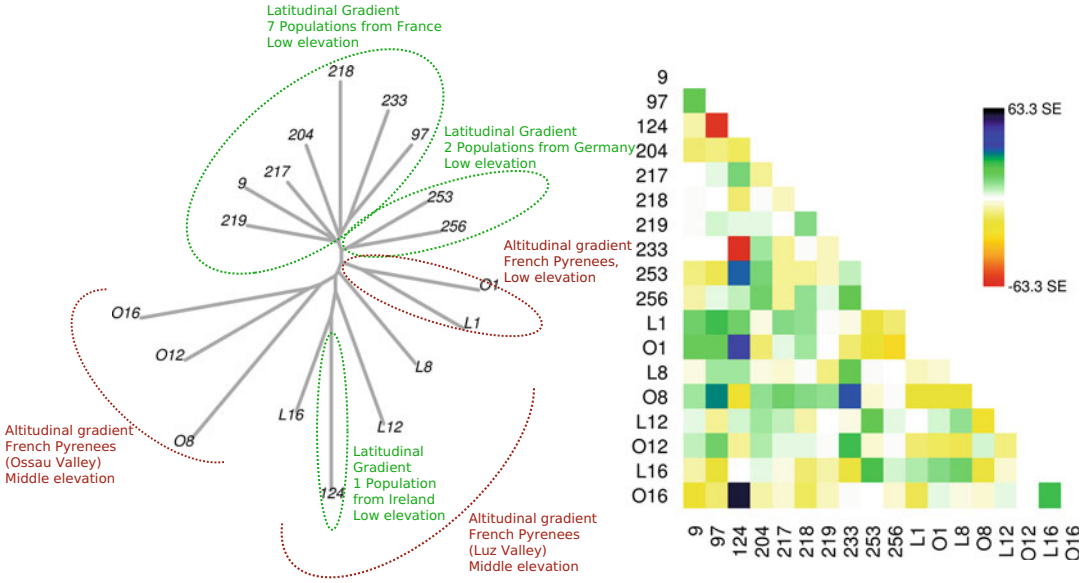


Fig. 14 Population splits inferences under TreeMix assuming a simple bifurcating tree (no migration nodes). *Left*: Unrooted visualization of the best likelihood tree. Unlike in the study of Leroy et al. [6], we do not use additional species to root the tree, i.e., to find the most basal ancestor of the tree, but only perform the inference based on the 18 sessile oak populations. *Right*: Visualization of the matrix of residuals. For example, this matrix shows that populations 124 and O16 have a remaining variation in relatedness (black square) that is not captured by the bifurcating tree

distant nodes or branches. These events can be interpreted as different migrations events, either ancient or contemporary, that have contributed to generate populations with a mixed ancestry (so-called admixed populations). We can therefore perform simulations for a range of migration events (m).

By adding different migration events, the likelihood of the model (or the total variance explained) is expected to increase (Fig. 15). For example, adding a single migration node substantially increase the proportion of explained variance (+3.1%; *see* Fig. 16 for the corresponding tree topology).

Admixture between populations can be tested using three- and four-population tests. These *f3* and *f4* tests were developed by Reich et al. [67] and Keinan et al. [68], respectively, and are implemented in the TreeMix suite. The tree-population test *f3*(A; B;C) aims at testing if a given population A is admixed between two other populations (B and C). Negative f3 values are indicative of admixture (*see* [67] for methodological details and [6] for empirical tests on oak data).

3.3.5 F_{ST} Fixation Indices

Several bioinformatic solutions were developed to compute measures of differentiation between pools such as F_{ST} (*see* Subheading 3.2.7 for general information about F_{ST}). PoPoolation2 [69] is

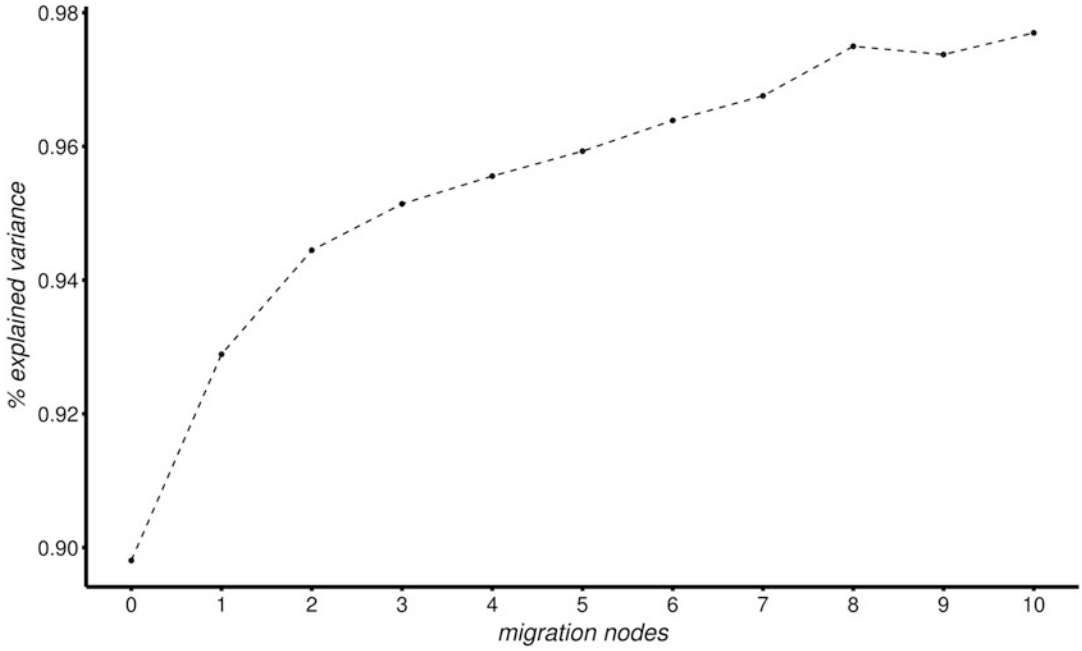


Fig. 15 Proportion of the variance explained for a growing number of migration nodes. Only one simulation was performed per migration node

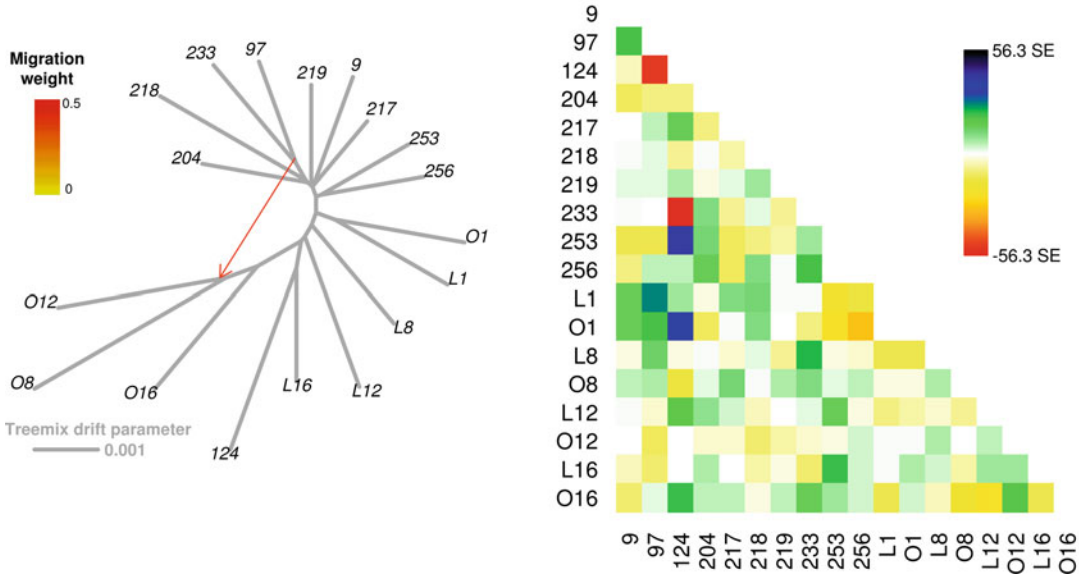


Fig. 16 Population splits inferences under TreeMix assuming a simple bifurcating tree and a migration node. *Left*: Unrooted visualization of the best likelihood tree and the inferred migration node. *Right*: Visualization of the matrix of residuals for this best tree. Unlike in Fig. 14, no strong excess of remaining variation in relatedness between populations 124 and O16 is observed

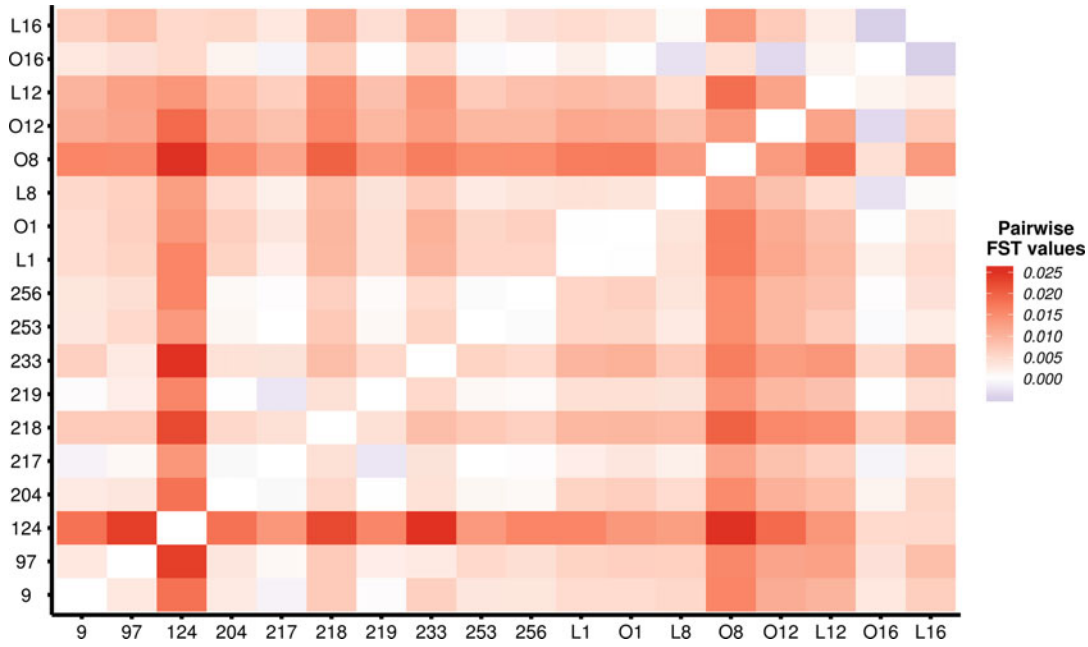


Fig. 17 Pairwise F_{ST} values between the 18 sessile oak pools, as computed by the R package poolfstat. To speed up computations, computations were performed on a random selection of 100,000 SNPs among the whole SNP set

probably the most widely used program for this purpose. In this section, we used the new estimator of F_{ST} recently developed by Hivert et al. [70] because of its higher robustness to different sources of bias associated with pool-seq ([70] for details). In addition, this new F_{ST} estimator is implemented in an R package ("poolfstat" [70]) which also generates input files for BayPass [54], the genome scan method used in Subheading 3.3.6.

Using the R package *poolfstat*, the *computePairwiseFSTmatrix* function can be used to calculate pairwise F_{ST} values over the whole dataset, which can be useful to have a rapid overview of the genetic structure among the different pools (Fig. 17).

F_{ST} values can also be computed for each SNP using the *computeFST* function to detect SNPs that exhibit very high levels of differentiation among all pools (black line, Fig. 18). F_{ST} values can also be estimated for each SNP and each pair of pools using the *computePairwiseFSTmatrix* function with the following argument: "output.snp.values = TRUE" (gray lines, Fig. 18).

3.3.6 Genome Scans of Selection

Unlike the genome scan for selection performed for the African rice (Subheading 3.2.7), we used a Bayesian framework to detect footprints of natural selection. We have chosen the method implemented in BayPass [54], which is equally suited for pool-seq and individual sequencing data. Many other methods are available and of interest too, including Bayenv [71, 72]. Core models of Bayenv and BayPass are indeed very similar. First, the population structure

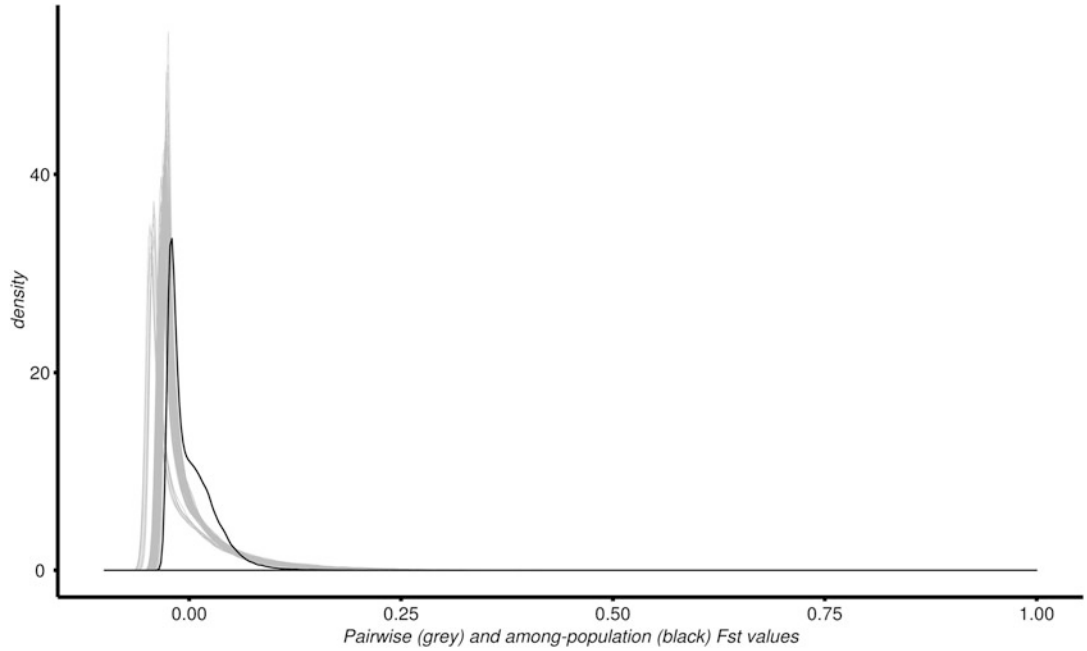


Fig. 18 Distributions of pairwise (gray) and among-population F_{ST} (black) values. Each gray line corresponds to the distribution of F_{ST} for one of the 153 (i.e., $\frac{18*(18-1)}{2}$) possible pairs

is captured by computing a covariance matrix of allele frequencies across all populations. This matrix is particularly convenient since it makes technically possible to perform extensive neutral simulations assuming this inferred covariance matrix in order to calibrate a measure of differentiation (Pseudo-Observed DataSets, PODS) and then identify threshold values based on these neutral simulations (but see also **Note 3**). Under Bayenv or BayPass, the differentiation metric used is the XtX, which can be considered as a SNP-specific F_{ST} explicitly accounting for the population structure. Outlier SNPs are the observed variants (red, Fig. 19) deviating from neutral expectations, i.e., those exhibiting greater XtX values than expected based on the simulations (black, Fig. 19).

XtX outliers are not randomly distributed along the genome but rather cluster in several genomic regions (black dots, Fig. 20). All these regions show an excess of differentiation among populations as compared to the expectations based on the variance-covariance matrix.

3.3.7 Genotype-Environment Association (GEA)

BayPass can also identify association between allele frequency differences and population-specific covariables, such as environmental or phenotype data (e.g., temperature, height, or yield). Assuming that climatic or phenotypic data is available for the set of populations under investigation, it is possible to identify allele frequency variation along these climatic or phenotypic gradients (so-called

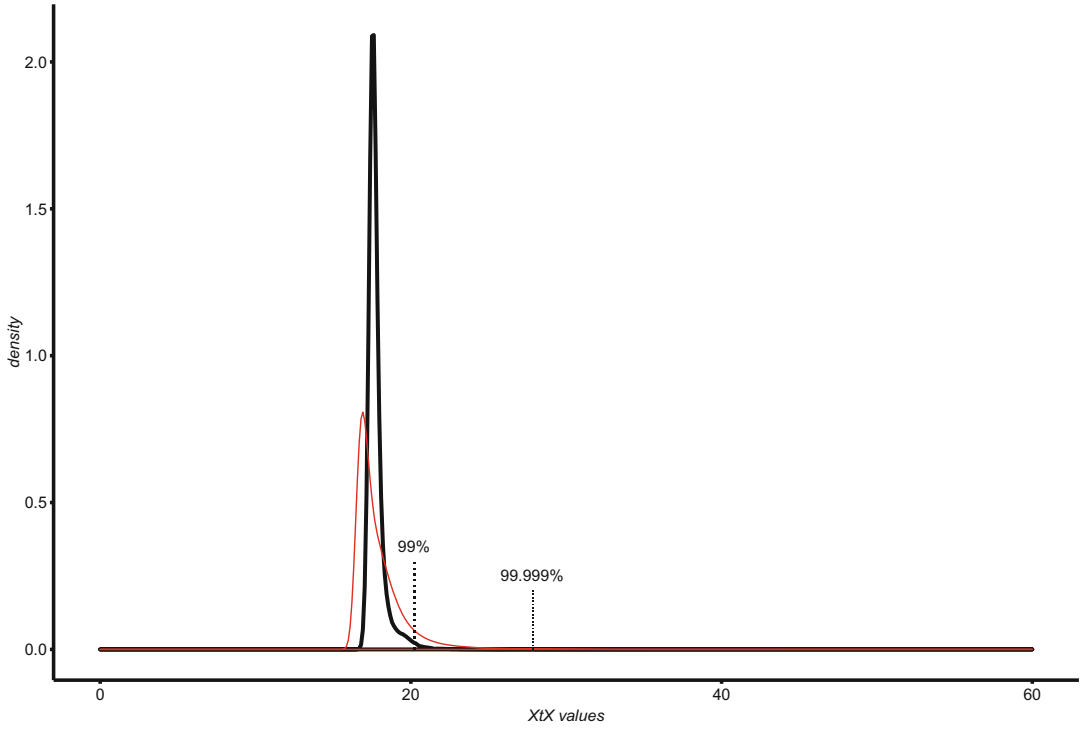


Fig. 19 Distributions of the XtX values for the observed dataset (red) and for the simulations assuming the variance-covariance matrix (black). Thresholds corresponding to the top 1% and 0.001% of the XtX values based on simulations are shown by the dotted lines

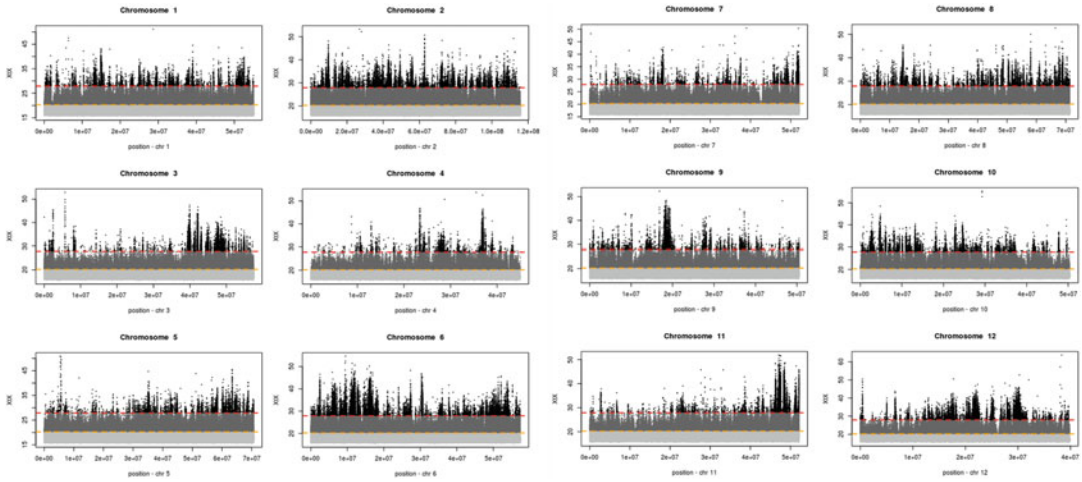


Fig. 20 Manhattan plots showing the chromosome positions of all SNPs and the corresponding XtX value as computed under BayPass. SNPs with empirical XtX values exceeding the 99% and 99.999% thresholds based on Pseudo-Observed DataSets (PODS, orange and red lines, respectively) are shown in dark gray and black, respectively

genetic clines). Associations to environmental covariables are often referred to as genotype-environment associations (GEA), while associations to phenotype are often referred to as genotype-phenotype associations (GPA) or population Genome-Wide Association Study (pGWAS). The strategy is to find correlations between allele frequencies at a given locus for a set of populations and mean values for a given trait for the same populations. In a nutshell, BayPass infers this "environmental effect" through a locus-specific regression coefficient parameter (β). In BayPass, the significance of this parameter can be tested using different decision rules (*see* [54] for details). Here, we used a simple comparison of models with and without association (i.e., a model assuming $\beta \neq 0$ versus $\beta = 0$) and quantify this support using Bayes factors (BF). The most positive BF values correspond to SNPs with the highest support for the model with a significant environmental or phenotypic effect. In general, SNPs of great interest are those simultaneously exhibiting both allele frequency differences among populations (highest XtX values) and associations (highest BF values, Fig. 21).

Manhattan plots showing chromosome positions of the associated SNPs (Fig. 22) reveal clusters of associated SNPs in some genomic regions, particularly on chromosomes 1, 9, 10, and 12. Such investigations can lead to the identification of important genes for local adaptation possible, for example, here adaptations to cold/warm conditions or drought/waterlogging. It is, however,

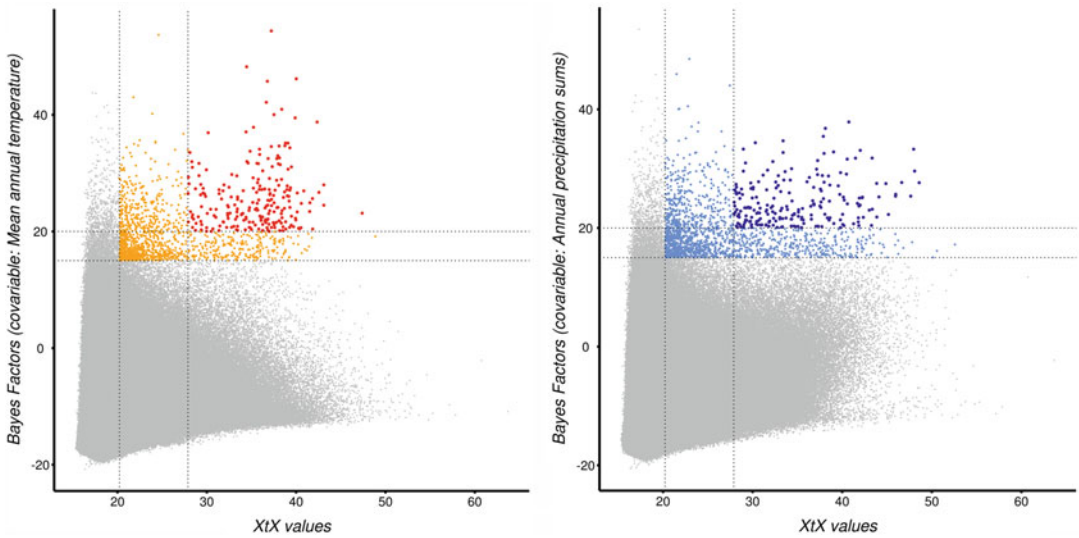


Fig. 21 Whole-genome scan for genetic differentiation (XtX) and association (Bayes factors, BF) with mean annual temperature (*left*) or precipitation sums (*right*) covariables and identification of SNPs of interests (orange or light blue; best candidates, red and dark blue). A simple rule-of-thumb decision was used to identify the most strongly associated SNPs: BF = 15 and BF = 20. As an alternative, it is also possible to use the PODS to calibrate the BF metric, in the same way as for the XtX (*see* Leroy et al. [6])

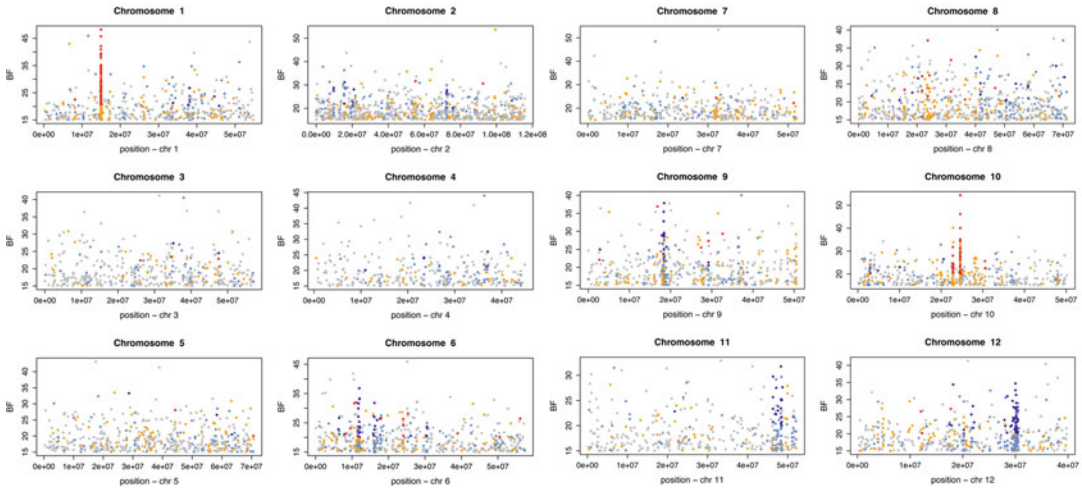


Fig. 22 Manhattan plots showing the chromosome positions of the SNPs exhibiting elevated Bayes factors (BF) as detected using BayPass. Significant SNPs in Fig. 20 are shown in colors. To facilitate readability, only SNPs with XtX $>$ 15 and BF $>$ 15 for either the mean annual temperature covariable or mean annual precipitation sums are shown

crucial to keep in mind the following statement when interpreting the results: correlation does not imply causation. GEA and GPA analyses can provide ecologically meaningful information, but these analyses are also prone to over-interpretation and storytelling (e.g., [73]).

4 Notes

1. For several analyses (e.g., PCA, clustering methods), it is important to note that linkage disequilibrium (LD), the non-random association of alleles within a genome between a given locus and its genomic neighborhood, is an important factor to control. For species with a relatively limited extent of LD across the genome (in general native species with a high genetic diversity), this bias is expected to be limited but can become substantial for some species, particularly domesticated ones. Various SNP pruning methods (e.g., SNPrune [74]) are currently being increasingly used for that purpose. We recommend using these methods. Advices on how to use these methods are available on the github repository.
2. It is also important to note that TreeMix (Subheading 3.3.4) fits single admixture pulses assuming homogeneous gene flow along the genome. This assumption is likely to be violated because migration is expected to be impeded at some genes maintaining genetic differences between hybridizing populations (e.g., [33] for empirical evidence). As a consequence,

TreeMix provides a good way to investigate potential migration events, but the exact direction of gene flow and the intensity of the migration edges should be interpreted with some caution. Some more advanced modeling approaches, albeit computationally intense, can decipher the evolutionary history of the investigated species with more confidence. These methods can explicitly account for heterogeneous migration rates (i.e., presence of barriers to gene flow). These methods provide considerably stronger statistical support for migration between populations, as well as temporal changes in effective population sizes, e.g., Approximate Bayesian Computation (ABC [33, 75]) or dadi [32]. A growing number of empirical studies have used the former (e.g., [75, 76]), the latter (e.g., [77, 78]), or both methods (e.g., [61]).

3. Deciphering the evolutionary history of a given species is an important step because demography can generate a substantial background noise weakening genome scan analyses [79]. To perform robust identification of variants under selection (or variants in close vicinity), one of the ongoing challenges is to better take into account the evolutionary history of the population. Extensive simulations under the inferred most likely evolutionary scenario can provide an accurate distribution of the expected differences in allele frequencies (e.g., F_{ST}, XtX, or similar), thereby allowing the identification of variants under selection among loci deviating from these demographic expectations. Some early attempts to explicitly take into account the inferred demography to scan genome for selection have recently emerged (e.g., [61, 77, 80, 81]). In future, we suspect the emergence of new methods inferring at once the most likely demographic scenario and variants departing from neutrality assuming this scenario.

Acknowledgments

The analyses benefited from the Montpellier Bioinformatics Biodiversity (MBB) platform services, the genotoul bioinformatics platform Toulouse Midi-Pyrenees (Bioinfo Genotoul), the Bird Platform of the University of Nantes and Compute Canada (Graham servers). This work takes its source from a diverse range of research contributions and projects we achieved during the last 5 years. During this period, TL was supported by different postdoctoral fellowships from the French *Agence Nationale de la Recherche* (ANR, Genoak project, PI: Christophe Plomion, 11-BSV6-009-021 and BirdIslandGenomic, PI: Benoit Nabholz, ANR-14-CE02-0002), from the European Research Council (ERC, Treepeace, PI: Antoine Kremer, Grant Agreement

no. 339728), and from the University of Vienna, Austria (PI: Christian Lexer). QR was supported by the government of Canada through Genome Canada, Genome British Columbia and Genome Quebec. QR wants to thank Louis Bernatchez for the opportunity to develop various projects during his postdoctoral research. We want to thank Jean-Marc Aury, Antoine Kremer, and Christophe Plomion for providing access to the oak sequencing data. We also thank Philippe Vigouroux and Philippe Cubry for information concerning the African rice data and Pierre-Alexandre Gagnaire and Nicolas Bierne for discussions on TreeMix. This book chapter is dedicated to Prof. Christian Lexer, who through his career greatly advanced our knowledge of population genomics and evolutionary botany.

References

1. Charlesworth B (2010) Molecular population genomics: a short history. Genet Res 92:397–411. https://doi.org/10.1017/S0016672310000522
2. Wang W, Mauleon R, Hu Z et al (2018) Genomic variation in 3,010 diverse accessions of Asian cultivated rice. Nature 557:43–49. https://doi.org/10.1038/s41586-018-0063-9
3. 1001 Genomes Consortium. Electronic address: magnus.nordborg@gmi.oeaw.ac.at, 1001 Genomes Consortium (2016) 1,135 genomes reveal the global pattern of polymorphism in *Arabidopsis thaliana*. Cell 166:481–491. https://doi.org/10.1016/j.cell.2016.05.063
4. Hartl DL, Clark AG (1998) Principles of population genetics. Sinauer, Sunderland, MA
5. Cubry P, Tranchant-Dubreuil C, Thuillet A-C et al (2018) The rise and fall of African Rice cultivation revealed by analysis of 246 new genomes. Curr Biol 28:2274–2282.e6. https://doi.org/10.1016/j.cub.2018.05.066
6. Leroy T, Louvet J-M, Lalanne C, et al (2019) Adaptive introgression as a driver of local adaptation to climate in European white oaks bioRxiv 584847. https://doi.org/10.1101/584847
7. Li H (2013) Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv 1303.3997
8. Li H, Durbin R (2009) Fast and accurate short read alignment with burrows–wheeler transform. Bioinformatics 25:1754–1760. https://doi.org/10.1093/bioinformatics/btp324
9. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with bowtie 2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923
10. Makino T, Rubin C-J, Carneiro M et al (2018) Elevated proportions of deleterious genetic variation in domestic animals and plants. Genome Biol Evol 10:276–290. https://doi.org/10.1093/gbe/evy004
11. Meyer RS, Purugganan MD (2013) Evolution of crop species: genetics of domestication and diversification. Nat Rev Genet 14:840
12. Falush D, Stephens M, Pritchard JK (2003) Inference of population structure using multilocus genotype data: linked loci and correlated allele frequencies. Genetics 164:1567
13. Hubisz MJ, Falush D, Stephens M, Pritchard JK (2009) Inferring weak population structure with the assistance of sample group information. Mol Ecol Resour 9:1322–1332. https://doi.org/10.1111/j.1755-0998.2009.02591.x
14. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945
15. Novembre J, Stephens M (2008) Interpreting principal component analyses of spatial population genetic variation. Nat Genet 40:646
16. Baird NA, Etter PD, Atwood TS et al (2008) Rapid SNP discovery and genetic mapping using sequenced RAD markers. PLoS One 3:e3376. https://doi.org/10.1371/journal.pone.0003376
17. Durand E, Jay F, Gaggiotti OE, François O (2009) Spatial inference of admixture proportions and secondary contact zones. Mol Biol Evol 26:1963–1973. https://doi.org/10.1093/molbev/msp106
18. Corander J, Marttinen P (2006) Bayesian identification of admixture events using multilocus molecular markers. Mol Ecol 15:2833–2843. https://doi.org/10.1111/j.1365-294X.2006.02994.x

19. Raj A, Stephens M, Pritchard JK (2014) fastSTRUCTURE: variational inference of population structure in large SNP data sets. Genetics 197:573. https://doi.org/10.1534/genetics.114.164350
20. Frichot E, François O (2015) LEA: an R package for landscape and ecological association studies. Methods Ecol Evol 6:925–929. https://doi.org/10.1111/2041-210X.12382
21. Frichot E, Mathieu F, Trouillon T et al (2014) Fast and efficient estimation of individual ancestry coefficients. Genetics 196:973. https://doi.org/10.1534/genetics.113.160572
22. Caye K, Deist TM, Martins H et al (2016) TESS3: fast inference of spatial population structure and genome scans for selection. Mol Ecol Resour 16:540–548. https://doi.org/10.1111/1755-0998.12471
23. Charlesworth B, Morgan MT, Charlesworth D (1993) The effect of deleterious mutations on neutral molecular variation. Genetics 134:1289
24. Pont C, Leroy T, Seidel M et al (2019) Tracing the ancestry of modern bread wheats. Nat Genet 51:905–911. https://doi.org/10.1038/s41588-019-0393-z
25. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123:585
26. Charlesworth B (2009) Fundamental concepts in genetics: effective population size and patterns of molecular evolution and variation. Nat Rev Genet 10:195–205. https://doi.org/10.1038/nrg2526
27. Sigwart J (2009) Coalescent theory: an introduction. Syst Biol 58:162–165. https://doi.org/10.1093/schbul/syp004
28. Terhorst J, Kamm JA, Song YS (2017) Robust and scalable inference of population history from hundreds of unphased whole genomes. Nat Genet 49:303–309. https://doi.org/10.1038/ng.3748
29. Li H, Durbin R (2011) Inference of human population history from individual whole-genome sequences. Nature 475:493
30. Schiffels S, Durbin R (2014) Inferring human population size and separation history from multiple genome sequences. Nat Genet 46:919
31. Excoffier L, Dupanloup I, Huerta-Sánchez E et al (2013) Robust demographic inference from genomic and SNP data. PLoS Genet 9:e1003905. https://doi.org/10.1371/journal.pgen.1003905
32. Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD (2009) Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genet 5:e1000695. https://doi.org/10.1371/journal.pgen.1000695
33. Roux C, Fraïsse C, Romiguier J et al (2016) Shedding light on the Grey zone of speciation along a continuum of genomic divergence. PLoS Biol 14:e2000234. https://doi.org/10.1371/journal.pbio.2000234
34. Akashi H, Osada N, Ohta T (2012) Weak selection and protein evolution. Genetics 192:15. https://doi.org/10.1534/genetics.112.140178
35. Lu J, Tang T, Tang H et al (2006) The accumulation of deleterious mutations in rice genomes: a hypothesis on the cost of domestication. Trends Genet 22:126–131. https://doi.org/10.1016/j.tig.2006.01.004
36. Yang J, Mezmouk S, Baumgarten A et al (2017) Incomplete dominance of deleterious alleles contributes substantially to trait variation and heterosis in maize. PLoS Genet 13:e1007019. https://doi.org/10.1371/journal.pgen.1007019
37. Liu Q, Zhou Y, Morrell PL, Gaut BS (2017) Deleterious variants in Asian Rice and the potential cost of domestication. Mol Biol Evol 34:908–924. https://doi.org/10.1093/molbev/msw296
38. Ramu P, Esuma W, Kawuki R et al (2017) Cassava haplotype map highlights fixation of deleterious mutations during clonal propagation. Nat Genet 49:959
39. Zhou Y, Massonnet M, Sanjak JS et al (2017) Evolutionary genomics of grape (Vitis vinifera ssp. vinifera) domestication. Proc Natl Acad Sci USA 114:11715. https://doi.org/10.1073/pnas.1709257114
40. Stein JC, Yu Y, Copetti D et al (2018) Genomes of 13 domesticated and wild rice relatives highlight genetic conservation, turnover and innovation across the genus Oryza. Nat Genet 50:285–296. https://doi.org/10.1038/s41588-018-0040-0
41. Marsden CD, Ortega-Del Vecchyo D, O'Brien DP et al (2016) Bottlenecks and selective sweeps during domestication have increased deleterious genetic variation in dogs. Proc Natl Acad Sci U S A 113:152. https://doi.org/10.1073/pnas.1512501113
42. Ng PC, Henikoff S (2001) Predicting deleterious amino acid substitutions. Genome Res 11:863–874
43. Choi Y, Sims GE, Murphy S et al (2012) Predicting the functional effect of amino acid substitutions and indels. PLoS One 7:e46688. https://doi.org/10.1371/journal.pone.0046688
44. Peischl S, Excoffier L (2015) Expansion load: recessive mutations and the role of standing

genetic variation. Mol Ecol 24:2084–2094. https://doi.org/10.1111/mec.13154

45. Henn BM, Botigué LR, Bustamante CD et al (2015) Estimating the mutation load in human genomes. Nat Rev Genet 16:333
46. Henn BM, Botigué LR, Peischl S et al (2016) Distance from sub-Saharan Africa predicts mutational load in diverse human genomes. Proc Natl Acad Sci U S A 113:E440. https://doi.org/10.1073/pnas.1510805112
47. Simons YB, Turchin MC, Pritchard JK, Sella G (2014) The deleterious mutation load is insensitive to recent population history. Nat Genet 46:220–224. https://doi.org/10.1038/ng.2896
48. Lewontin RC, Krakauer J (1973) Distribution of gene frequency as a test of the selective neutrality of polymorphisms. Genetics 74:175
49. Bierne N, Roze D, Welch JJ (2013) Pervasive selection or is it. . .? Why are FST outliers sometimes so frequent? Mol Ecol 22:2061–2064. https://doi.org/10.1111/mec.12241
50. Bierne N, Welch J, Loire E et al (2011) The coupling hypothesis: why genome scans may fail to map local adaptation genes. Mol Ecol 20:2044–2072. https://doi.org/10.1111/j.1365-294X.2011.05080.x
51. Lotterhos KE, Whitlock MC (2015) The relative power of genome scans to detect local adaptation depends on sampling design and statistical method. Mol Ecol 24:1031–1046. https://doi.org/10.1111/mec.13100
52. Nei M, Maruyama T (1975) Lewontin-Krakauer test for neutral genes. Genetics 80:395
53. Robertson A (1975) Remarks on the Lewontin-Krakauer. Genetics 80:396
54. Gautier M (2015) Genome-wide scan for adaptive divergence and association with population-specific covariates. Genetics 201:1555. https://doi.org/10.1534/genetics.115.181453
55. Whitlock MC, Lotterhos KE (2015) Reliable detection of loci responsible for local adaptation: inference of a null model through trimming the distribution of FST. Am Nat 186:S24–S36. https://doi.org/10.1086/682949
56. Luu K, Bazin E, Blum MGB (2017) Pcadapt: an R package to perform genome scans for selection based on principal component analysis. Mol Ecol Resour 17:67–77. https://doi.org/10.1111/1755-0998.12592
57. Abdellaoui A, Hottenga J-J, de Knijff P et al (2013) Population structure, migration, and diversifying selection in the Netherlands. Eur J Hum Genet 21:1277
58. Jackson DA (1993) Stopping rules in principal components analysis: a comparison of Heuristical and statistical approaches. Ecology 74:2204–2214. https://doi.org/10.2307/1939574
59. Schlötterer C, Tobler R, Kofler R, Nolte V (2014) Sequencing pools of individuals—mining genome-wide polymorphism data without big funding. Nat Rev Genet 15:749
60. Gautier M, Foucaud J, Gharbi K et al (2013) Estimation of population allele frequencies from next-generation sequencing data: pool-versus individual-based genotyping. Mol Ecol 22:3766–3779. https://doi.org/10.1111/mec.12360
61. Leroy T, Rougemont Q, Dupouey J-L, et al (2018) Massive postglacial gene flow between European white oaks uncovered genes underlying species barriers. bioRxiv. https://doi.org/10.1101/246637
62. Plomion C, Aury J-M, Amselem J et al (2018) Oak genome reveals facets of long lifespan. Nat Plants 4:440–452. https://doi.org/10.1038/s41477-018-0172-3
63. De Vries SMG, Alan M, Bozzano M, Burianek V, Collin E, Cottrell J, Ivankovic M, Kelleher CT, Koskela J, Rotach P, Vietto L, Yrjänä L (2015) Pan-European strategy for genetic conservation of forest trees and establishment of a core network of dynamic conservation units. XF2017001223. EUFORGEN/BI, Paris. http://www.euforgen.org/fileadmin/templates/euforgen.org/upload/Publications/Thematic_publications/EUFORGEN_FGR_conservation_strategy_web.pdf
64. Lindner MS, Kollock M, Zickmann F, Renard BY (2013) Analyzing genome coverage profiles with applications to quality control in metagenomics. Bioinformatics 29:1260–1267. https://doi.org/10.1093/bioinformatics/btt147
65. Kofler R, Orozco-terWengel P, De Maio N et al (2011) PoPoolation: a toolbox for population genetic analysis of next generation sequencing data from pooled individuals. PLoS One 6:e15925. https://doi.org/10.1371/journal.pone.0015925
66. Pickrell JK, Pritchard JK (2012) Inference of population splits and mixtures from genome-wide allele frequency data. PLoS Genet 8:e1002967. https://doi.org/10.1371/journal.pgen.1002967
67. Reich D, Thangaraj K, Patterson N et al (2009) Reconstructing Indian population history. Nature 461:489
68. Keinan A, Mullikin JC, Patterson N, Reich D (2007) Measurement of the human allele frequency spectrum demonstrates greater genetic

drift in east Asians than in Europeans. Nat Genet 39:1251

69. Kofler R, Pandey RV, Schlötterer C (2011) PoPoolation2: identifying differentiation between populations using sequencing of pooled DNA samples (Pool-Seq). Bioinformatics 27:3435–3436. https://doi.org/10.1093/bioinformatics/btr589
70. Hivert V, Leblois R, Petit EJ et al (2018) Measuring genetic differentiation from Pool-seq data. Genetics 210:315. https://doi.org/10.1534/genetics.118.300900
71. Coop G, Witonsky D, Di Rienzo A, Pritchard JK (2010) Using environmental correlations to identify loci underlying local adaptation. Genetics 185:1411–1423. https://doi.org/10.1534/genetics.110.114819
72. Günther T, Coop G (2013) Robust identification of local adaptation from allele frequencies. Genetics 195:205. https://doi.org/10.1534/genetics.113.152462
73. Pavlidis P, Jensen JD, Stephan W, Stamatakis A (2012) A critical assessment of storytelling: gene ontology categories and the importance of validating genomic scans. Mol Biol Evol 29:3237–3248. https://doi.org/10.1093/molbev/mss136
74. Calus MPL, Vandenplas J (2018) SNPrune: an efficient algorithm to prune large SNP array and sequence datasets based on high linkage disequilibrium. Genet Sel Evol 50:34. https://doi.org/10.1186/s12711-018-0404-z
75. Roux C, Tsagkogeorga G, Bierne N, Galtier N (2013) Crossing the species barrier: genomic hotspots of introgression between two highly divergent *Ciona intestinalis* species. Mol Biol Evol 30:1574–1587
76. Fraïsse C, Roux C, Gagnaire P-A et al (2018) The divergence history of European blue mussel species reconstructed from approximate Bayesian computation: the effects of sequencing techniques and sampling strategies. PeerJ 6:e5198. https://doi.org/10.7717/peerj.5198
77. Rougemont Q, Gagnaire P-A, Perrier C et al (2017) Inferring the demographic history underlying parallel genomic divergence among pairs of parasitic and nonparasitic lamprey ecotypes. Mol Ecol 26:142–162. https://doi.org/10.1111/mec.13664
78. Tine M, Kuhl H, Gagnaire P-A et al (2014) European sea bass genome and its variation provide insights into adaptation to euryhalinity and speciation. Nat Commun 5:5770
79. Hermisson J (2009) Who believes in whole-genome scans for selection? Heredity 103:283–284
80. Fraïsse C, Roux C, Welch JJ, Bierne N (2014) Gene-flow in a mosaic hybrid zone: is local introgression adaptive? Genetics 197:939. https://doi.org/10.1534/genetics.114.161380
81. Le Moan A, Gagnaire P-A, Bonhomme F (2016) Parallel genetic divergence among coastal–marine ecotype pairs of European anchovy explained by differential introgression after secondary contact. Mol Ecol 25:3187–3202. https://doi.org/10.1111/mec.13627

Chapter 17

The Application of Flow Cytometry for Estimating Genome Size, Ploidy Level Endopolyploidy, and Reproductive Modes in Plants

Jaume Pellicer, Robyn F. Powell, and Ilia J. Leitch

Abstract

Over the years, the amount of DNA in a nucleus (genome size) has been estimated using a variety of methods, but increasingly, flow cytometry (FCM) has become the method of choice. The popularity of this technique lies in the ease of sample preparation and in the large number of particles (i.e., nuclei) that can be analyzed in a very short period of time. This chapter presents a step-by-step guide to estimating the nuclear DNA content of plant nuclei using FCM. Attempting to serve as a tool for daily laboratory practice, we list, in detail, the equipment required, specific reagents and buffers needed, as well as the most frequently used protocols to carry out nuclei isolation. In addition, solutions to the most common problems that users may encounter when working with plant material and troubleshooting advice are provided. Finally, information about the correct terminology to use and the importance of obtaining chromosome counts to avoid cytological misinterpretations of the FCM data are discussed.

Key words Chromosome number, DAPI, DNA ploidy level, Endopolyploidy, Genome size, Flow cytometry, Flow histogram, C-value, PI, Plant nuclei isolation, Relative fluorescence

1 Introduction

The total amount of DNA in the nucleus of an organism is generally referred to as the genome size and is measured either in picograms (pg; i.e., 1×10^{-9} g) or megabase pairs (Mbp, with 1 pg = 978 Mbp, [1]). People started to investigate genome size in plants even before the structure of DNA was worked out, with the first plant to have its genome size estimated being *Lilium longiflorum* in 1951 [2]. Since then, the genome sizes of over 12,000 species have been estimated [3, 4] with the data being used not only for practical applications (e.g., How much will it cost to sequence a genome? How many clones are needed for making BAC libraries?) but also for providing valuable insights into many biological fields, including evolution, systematics, ecology, population genetics, and plant breeding (e.g., [5–15]).

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_17,

Over the years, several methods have been used to estimate genome sizes in plants (e.g., Feulgen densitometry, reassociation kinetics). Nevertheless, in recent years, due to a variety of reasons, flow cytometry (FCM) has become the method of choice [16]. Briefly, the method involves three steps: (1) a sample of plant tissue is chopped in a suitable buffer to release the nuclei while maintaining their integrity; (2) the nuclei are stained with a fluorochrome that binds quantitatively to the DNA, so the bigger the genome, the more stain that is bound to the DNA; (3) the nuclei are passed through a flow cytometer which measures the amount of stain bound to each nucleus (Fig. 1a). By preparing a combined sample which includes a plant species with a known DNA amount (reference standard), the relative intensity of fluorescence

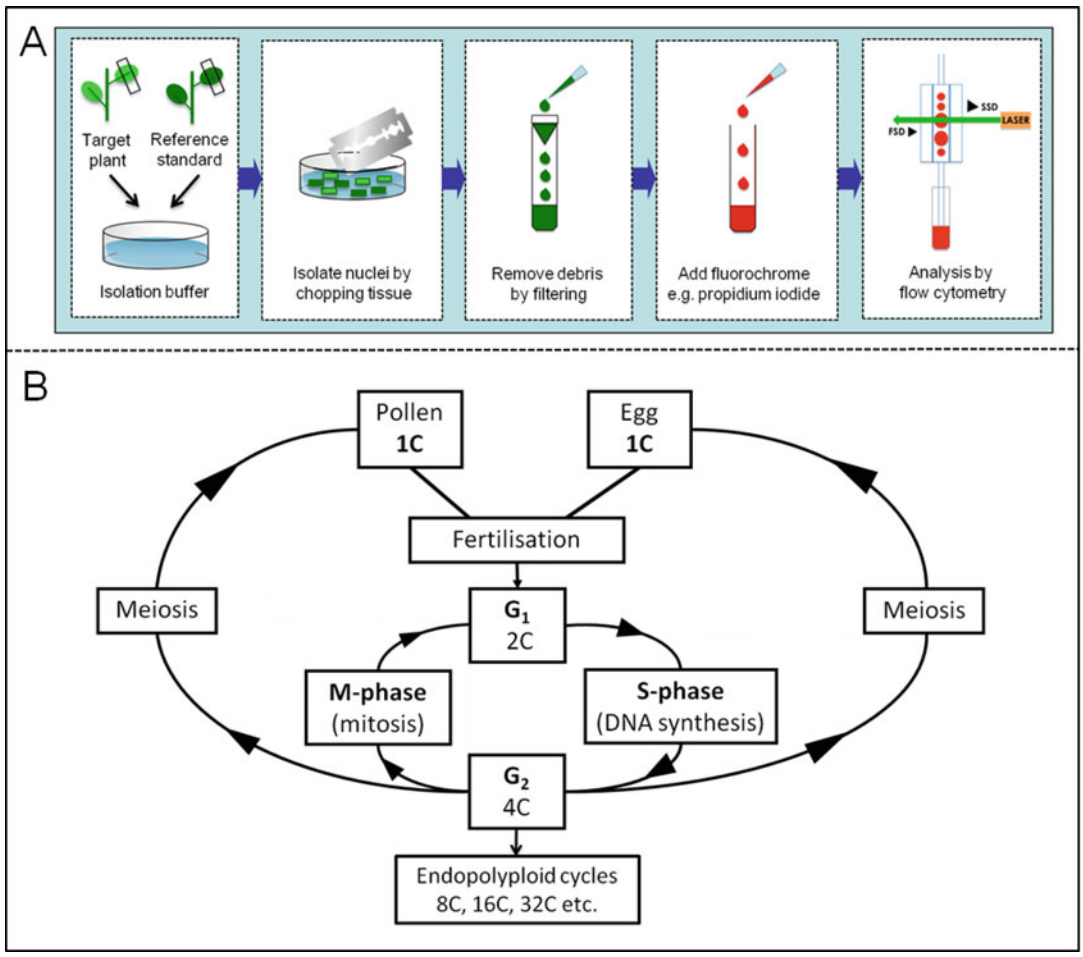


Fig. 1 (**a**) The basic steps involved in the estimation of genome size and ploidy level by flow cytometry. (**b**) Changes in the holoploid C-value at different stages of the cell cycle and following meiosis and endopolyploidy (N.B. cells which undergo endopolyploidy (i.e., DNA synthesis not accompanied by mitosis) will have C-value greater than 4C (i.e., 8C, 16C, 32C, etc., depending on the number of rounds of DNA replication))

from the target plant can be converted into an absolute genome size. It is important to realize that FCM only gives information about the relative or absolute DNA amount of the isolated nuclei, but it does not provide cytological information. Yet without such information, interpretations of the chromosome number and/or ploidy level of the species can be flawed. Subheading 1.2 below highlights the importance of obtaining such cytological data and the pitfalls and errors that can arise without it.

FCM can also be used to estimate the ploidy level of a plant based on comparing the genome size of the target species either with the genome size of a specimen of known ploidy (i.e., determined karyologically) or with an internal standard (in that case, the reference standard must be kept constant throughout the experiment, and the ploidy level of at least one target sample should be karyologically determined). However, in such cases, the ploidy level is referred to as the "DNA ploidy" to distinguish it from studies where ploidy level has been determined karyologically [17]. Such approaches are now being increasingly used to survey the diversity of cytotypes across plant populations and have uncovered a surprising diversity of hitherto unsuspected ploidy variation in some species [18, 19]. Besides ploidy levels, FCM can be also employed to explore and investigate the occurrence of endopolyploidy. Endopolyploidy, where several rounds of DNA synthesis occur without mitosis (i.e., producing cells with 4C, 8C, 16C, etc., Fig. 1b), can occur in certain cell types within a plant and can reach very high levels (e.g., the endosperm haustorium cells of *Mesembryanthemum crystallinum* (Aizoaceae) have undergone 16 endocycles and hence, the nuclei are 65,536C) [20]. Differences in the occurrence and levels of endopolyploidy between tissues can result in different tissues having different DNA amounts [21]. In addition to this variation within plants, the frequency of endopolyploidy also varies between families and lineages of land plants (e.g., it is reported to be absent in lycophytes and liverworts but common in mosses).

Another application of FCM is the inference of reproductive pathways in plants (sexual vs. apomixis). Based on the same principles as for the analysis of DNA ploidy levels, this approach consists of establishing the ratio between the relative DNA contents of the embryo and the endosperm in the seed. This method has been called "flow cytometric seed screening (FCSS)" [22] and has been widely used since then in multiple studies which aim to gain insights into the impact of apomixis in plants and its long-term evolutionary consequences (e.g., [23–25]). In brief, seeds from most flowering plants formed via the sexual pathway are expected to display an embryo-endosperm ratio of ~1.5 (2C:3C) resulting from double fertilization comprising (1) the fusion of one of the haploid sperm cells with the haploid egg cell to make a 2C zygote and (2) the fusion of the other haploid sperm with two haploid polar nuclei to

form a triploid endosperm (3C). Any deviation from this ratio is assumed to have arisen from a variety of apomictic pathways [24].

This chapter outlines the general method used to estimate genome size and/or determine the DNA ploidy level in plants using FCM and its multiple applications in different fields (Subheadings 2–4). However, given the immense diversity of plants in terms of their morphology (e.g., woody, succulent, herbaceous) and biochemistry (e.g., presence of pigments, tannins, phenolics, mucilaginous compounds), several problems may well be encountered. The majority of these arise mainly from the interaction between chemicals present in the cell cytoplasm and the binding of the fluorochrome to the DNA [26–34] leading to erroneous results. Thus, this chapter also outlines some of the more commonly encountered problems and ways in which the poor results might be improved to overcome these issues.

In addition to the information given here, it is worth checking databases such as the Plant DNA C-values Database [3] and the Genome Size in Asteraceae Database (GSAD, ([35])) to see whether a given genus of interest has previously been studied by FCM and if so, whether any particular modifications were made to the buffers used, etc. This will help to overcome specific problems associated with the particular genus being analyzed. In addition, checking such databases can be helpful to get some idea about the range of genome sizes one might expect for a given taxon. The Plant DNA C-values Database [3] contains data for all the major groups of land plants and three algal lineages, while the more focused database containing genome size data for Asteraceae (GSAD—[36]) is ideal for specific studies focused on this family of angiosperms. Such prior information can save a lot of time and frustration!

1.1 Terminology Used for Genome Size Studies

Given that the amount of DNA varies throughout the cell cycle (i.e., G_2 nuclei have twice the DNA amount as G_1 nuclei) (Fig. 1b) and following meiosis and endopolyploidy (=somatic polyploidy), considerable confusion can arise when discussing genome sizes. To overcome such issues, Greilhuber et al. [37] proposed the following terminologies which have now been widely adopted:

1. *Holoploid 1C-value* (abbreviated to 1C-value) refers to the amount of DNA in the unreplicated gametic nucleus (e.g., pollen or egg cell of angiosperms) regardless of the ploidy level of the cell. The 2C-value represents the amount of DNA in a somatic cell at the G_1 stage of the cell cycle, while the 4C-value is the amount in a somatic cell at the G_2 stage, following DNA synthesis (S-phase) (*see* Fig. 1b).
2. *Monoploid 1C-value* (abbreviated to 1Cx-value) refers to the amount of DNA in the unreplicated monoploid (x) chromosome set. For a diploid organism where 2n = 2x,

the 1C- and 1Cx-values are the same; however, for a polyploid organism, the 1Cx is always smaller than the 1C-value (e.g., for a tetraploid where 2n = 4x, then 1Cx = 1/2 1C, whereas for a hexaploid where 2n = 6x, then 1Cx = 1/3 1C, etc.).

1.2 The Importance of Cytological Data for Genome Size Studies

As noted above, FCM only measures the total amount of DNA in the nucleus and gives no specific information about the chromosome number or ploidy level of the plant analyzed (although this can be deduced in certain cases as outlined in Subheading 3.2.3). Despite this, many studies report a ploidy level or chromosome number for the analyzed plant which has either been taken from the literature or based on comparisons of DNA amounts found in related species. This is, to some extent, acceptable in a stable cytological system where there is little variation in chromosome number and size between species. However, in plants, such situations are probably the exception rather than the rule, even between closely related taxa, as many genera show considerable cytological diversity—for example, (1) polyploidy, both within (e.g., [18]) and between species (e.g., [38, 39]), is frequent; (2) large divergences in genome size among closely related species with the same ploidy have been reported (e.g., *see* Fig. 2); and (3) increases in ploidy level or chromosome number are not necessarily accompanied by proportional changes in DNA amount (e.g., *see* [40]).

Examples of problems and misinterpretations of genome size that can arise through assuming the ploidy level and/or chromosome number of a species have been discussed by Suda et al. [17]. Below are a few examples to illustrate the pitfalls that can arise when karyological information is not obtained in parallel with genome size data.

1. In species with a constant chromosome number but a big range in size, an absence of chromosome data could lead to the erroneous suggestion that polyploids may be present to explain the large range of genome sizes encountered. This is illustrated by the genus *Cypripedium* (Orchidaceae) where most species have a chromosome count of 2n = 20 but genome size has been shown to vary over tenfold between species (1C = 4.1–43.1 pg) [41] (*see* Fig. 2a and b). Another example can be found in the genus *Heloniopsis* (Melanthiaceae) where the genome sizes for some species are twice the size of others and yet all species have the same chromosome number of 2n = 34 [42]. A similar situation has also been reported in the genus *Artemisia* (Asteraceae), where the diploid species *Artemisia annua* with 2n = 18 chromosomes has a 1C-value of 1.75 pg [43], while *Artemisia leucodes* with the same chromosome number has a 1C = 7.70 pg [44]. Without doing a chromosome count, one could easily assume that *A. leucodes* was a polyploid given such differences in genome size.

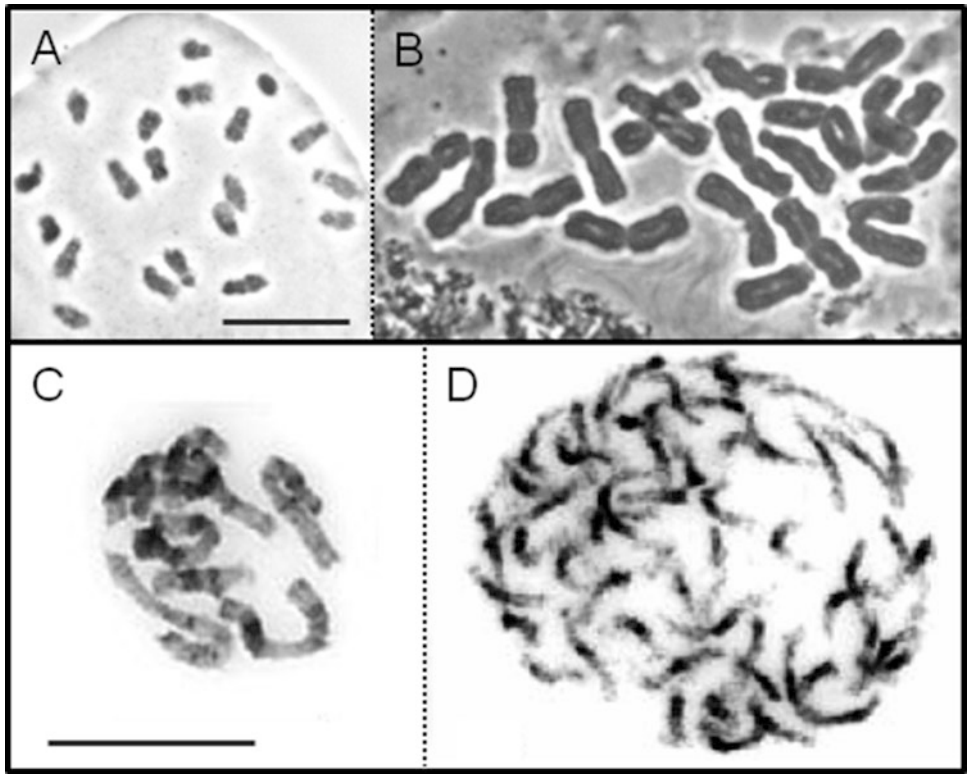


Fig. 2 Examples where chromosome size and number in related species do not correlate with genome size. Chromosomes of (**a**) *Cypripedium molle* (1C = 4.1 pg) and (**b**) *C. calceolus* (1C = 32.4 pg) taken at the same magnification showing an eightfold range in genome size but a constant chromosome number of 2n = 20. Image reproduced with permission from [41] (scale bar = 10 μm). Chromosomes of (**c**) diploid *Physaria bellii* (2n = 2x = 8; 1C = 2.34 pg) compared with those from (**d**) the high polyploid *P. didymocarpa* (2n = 14x = 56; 1C 2.23 pg). Both species have similar genome sizes but very different chromosome numbers and sizes. Image reproduced with permission from [46] (scale bar = 5 μm)

2. Erroneous assumptions of ploidy level in a studied species can arise when increases in chromosome number via polyploidy have not been accompanied by proportional increases in genome size. This is likely to be a common problem since genome downsizing following polyploidy is frequently encountered in angiosperms [45]. In extreme cases, species with higher ploidy levels may have the same or lower genome sizes than related species of lower ploidy. An example of this is provided by the genus *Physaria* (Brassicaceae) where the high polyploid *Physaria didymocarpa* with 2n = 14x = 56 actually has a smaller genome (1C = 2.23 pg) than a related diploid *P. bellii* (2n = 2x = 8) with 1C = 2.34 pg [46] (Fig. 2c and d).
3. Nonproportional changes in DNA content have also been reported in different cytotypes of the same species. Once again, this has the potential to lead to erroneous deductions

of ploidy level based on genome size data alone. Both increases and decreases in the size of monoploid genomes have been reported with increasing ploidy levels. For example, in *Larrea tridentata* (Zygophyllaceae), the hexaploid cytotype was reported to have just 1.25 times more DNA than the tetraploid [47]. Without chromosome data to support this, it is possible that the hexaploid could have been misidentified as a pentaploid based on DNA amount alone.

In addition to these examples, it is also important to note that many chromosomal changes and variations (e.g., aneuploidy, chromosome duplications and deletions, sex and supernumerary chromosomes, supernumerary segments, etc.) can arise which are detectable as changes in DNA amount. Without identifying these through cytological analysis, further misinterpretations of the data may arise.

Overall, these examples serve to illustrate how serious mistakes can be made in the absence of karyological information. Thus, it is strongly recommended that chromosome counts are made of the plant used for genome size estimation. If this is not possible, then the ploidy level should always be referred to as the "DNA ploidy level" as discussed by Suda et al. [17].

2 Materials

Detailed information about plant tissues, reagents, composition of the isolation buffers, as well as the technical equipment needed to carry out genome size and ploidy estimations using FCM are described below.

2.1 Plant Tissue and Reference Standards

Of the potential plant tissues suitable for genome size estimation, leaf tissue is preferred by researchers because it generally gives the best results. Nevertheless, other plant tissues such as petals, flower stalks, young stems (incl. petioles), cambial tissue or decorticate twigs of woody species, roots, pollen grains (incl. pollinia), fruiting capsules, and seeds (dried or fresh) [13, 34, 48–51] can be considered as viable alternatives for genome size estimations. When fresh plant tissues are selected, they should be as fresh as possible and collected from young and actively growing parts of the plant as such material is likely to give the best results. Old and senescent tissues will probably result in higher levels of background signal and may contain high proportions of nuclei at the G_2 phase of mitosis.

In addition, silica-dried leaves and herbarium vouchers may be used to estimate DNA ploidy levels (e.g., [14, 52, 53]). However, given that DNA deterioration is likely to occur in such samples, the material is not considered suitable for high-quality estimations of genome size in absolute units. More recently, improved protocols

for silica-desiccated samples, including long-term storage at −20 °C in the genus *Juniperus*, have shown highly consistent results between fresh and preserved samples [14], therefore opening new opportunities to explore genome size diversity in preserved materials. That said, the suitability and consistency of the method should be rigorously tested in each case prior to the use of such material for genome size studies.

Notwithstanding, as an alternative to using desiccated material, the suitability of glycerol-preserved nuclei for estimating genome size in absolute units for material up to at least a few weeks old has been investigated [54]. This method has been designed for field research, and although it has some limitations (i.e., high-quality results are only obtained when samples are kept in ice-cold buffer), it demonstrates the efforts that researchers in this discipline may go to in order to overcome problems associated with the current limited timescale available to analyze large batches of fresh material without compromising quality of the results.

Concerning reference standards, we recommend that several species, covering a broad range of genome sizes, are kept growing in the laboratory to enable the most appropriate standard to be selected for each particular analysis. Many species have been used, but we summarize some of the most popular ones in Table 1 which work well with FCM.

2.2 Equipment

1. Set of pipettes with disposable tips (100 μL, 1 mL).
2. Razor blades (double-edged) or scalpel with replaceable blades. A razor blade holder or alternative protective device (e.g., cork or silicon bung) is also recommended.
3. Plastic petri dishes (c. 5–6 cm diameter).
4. Disposable nylon mesh filters (30–42 μm pore size; e.g., Sysmex CellTrics, cat. no. 04-0042-2316). Alternatively, regular nylon mesh cut into squares and fitted on disposable tips can be used.
5. Sample tubes suitable for the particular flow cytometer being used (check manufacturer's specifications in each case).
6. 1.5 mL tubes.
7. Sample tube racks.
8. Plastic and/or expanded polystyrene containers to fill with ice.
9. Latex, nitrile, or vinyl gloves. Safety goggles and lab coat.
10. Centrifuge fitted with a rotor suitable for 1.5 mL tubes.
11. Fridge and freezer.
12. Flow cytometer fitted with the light source suitable for excitation of the DNA fluorochrome used in the study (check fluorochrome's excitation and emission spectra to select the suitable excitation sources following the manufacturer's recommendations).

Table 1
Several reference standard species recommended for genome size estimation

Plant species	1C DNA content (pg)	References
Oryza sativa 'IR-36'	0.50	[76]
Raphanus sativus L. 'Saxa'	0.55	[77]
Solanum lycopersicum L. 'Stupiké polní rané'	0.98	[77]
Vigna radiata 'Berken'	1.20	[76]
Glycine max Merr. 'Polanka'	1.25	[78]
Petunia hybrida Vilm. 'PxPc6'	1.42	[79]
Petroselinum crispum 'Champion Moss Curled'	2.22	[80]
Zea mays L. 'CE-777'	2.71	[81]
Pisum sativum L. 'Express Long'	4.18	[79]
Pisum sativum L. 'Ctirad'	4.54	[82]
Pisum sativum L. 'Minerva Maple'	4.86	[76]
Secale cereale L. 'Daňkovské'	8.09	[82]
Vicia faba L. 'Inovec'	13.45	[77]
Allium cepa L. 'Ailsa Craig'	17.44	[83]
Allium cepa L. 'Alice'	17.42	[82]

13. Analytical software for evaluation of flow cytometric data (usually provided by the manufacturer of the flow cytometer).
14. Fume cupboard to carry out nuclei isolation using buffers supplemented with either β-mercaptoethanol or dithiothreitol (DTT) (*see* **Note 1**).
15. Cleaning and decontamination solutions for flow systems. Domestic sodium hypochlorite (bleach) diluted 1:5 in distilled water.
16. Calibration particles: fluorescent beads [e.g., Sysmex, cat. nos. 05-4006_R (green) and 05-4022 (UV)].

2.3 Reagents

2.3.1 Fluorochromes

The following fluorochromes are the most popular dyes used in flow cytometry to estimate genome size and for DNA ploidy analysis:

1. PI (propidium iodide—*see* **Notes 1** and **2**): Intercalating fluorescent dye (IFD). Prepare a 1 mg/mL stock solution and filter through a 0.22 μm filter. Store in 1 mL aliquots at −20 °C (*see* **Note 1**).
2. DAPI (4′,6-diamidino-2-phenylindole—*see* **Notes 1** and **2**): Base-specific dye (BSD). Prepare 0.1 mg/mL stock solution

and filter through a 0.22 μm filter. Store in 1 mL aliquots at −20 °C (*see* **Note 1**).

SYBR Green I has also been used in a few genome size and ploidy studies (e.g., [55]), although much less frequently than PI and DAPI. Its use is not described here although information on how to prepare it for genome size estimation is given in **Note 2**.

2.3.2 Isolation Buffers (See *Notes 3* and *4*)

1. *General Purpose Buffer [56] for the one-step protocol (Subheading 3.1.1)*: 0.5 mM spermine·4HCl, 30 mM sodium citrate, 20 mM MOPS (*see* **Note 1**), 80 mM KCl, 20 mM NaCl, 0.5% (v/v) Triton X-100. Adjust to pH 7.0. Store the buffer either at 4 °C if used regularly or at −20 °C in 10 mL aliquots.
2. *Otto buffer [57] for the two-step protocol (Subheadings 3.1.2 and 3.1.3)*: Otto I: 100 mM citric acid monohydrate, 0.5% (v/v) Tween 20 (*see* **Note 43**). Store at 4 °C. Otto II: 400 mM Na_2HPO_4 (*see* **Notes 39** and **44**). Store at room temperature. The fluorochrome (DAPI or PI; see above) can be added to Otto II before adjusting the final volume of the stock solution. If this is done, the buffer should be stored in the dark at room temperature. Alternatively, the fluorochrome can be added directly to the sample at **step 10** of Subheading 3.1.2 or **step 7** of Subheading 3.1.3.

   Further modifications to the composition of the Otto buffer (including the addition of different amounts of Tween 20, HCl, HNO_3, and acetic acid) which were shown to improve the estimation of genome size in some species are given in Šmarda et al. [34].

3 Methods

3.1 Isolation of Plant Nuclei

Nuclei suspensions can be prepared according to either the one-step protocol (Subheading 3.1.1) or the two-step protocol (Subheading 3.1.2). The one-step protocol using General Purpose Buffer (*see* **Note 3**) works with many plant species. However, for some plant groups, the two-step protocol using the Otto buffers will provide histograms with much higher-quality peaks. A simplified version of the two-step protocol using the Otto buffer is given in Subheading 3.1.3.

We recommend (unless specified otherwise) working under cold conditions (i.e., keep all solutions, buffers, and prepared samples waiting for analysis on ice, and do the chopping step in a petri dish resting on a bed of ice). Together, this helps to inhibit the negative effect of many cytosolic compounds that may be present (e.g., DNase, phenolics, tannins, etc.), and it can be especially helpful when working with recalcitrant samples. Furthermore, in some cases, the quality of the results can be further improved if the

tubes are kept in ice cold water while the sample is being run on the flow cytometer.

3.1.1 Isolation of Plant Nuclei Using the One-Step Protocol

1. Place a small amount of the selected plant tissue (usually about 1 cm^2 or 20 mg) in a 6 cm petri dish (*see* **Note 6**).
2. Add 1 mL of ice-cold General Purpose Buffer to the petri dish. This isolation buffer performs well with a wide range of plant families. However, a selection of alternative buffers is given in **Note 4**.
3. Chop the tissues in the buffer using a new razor blade or sharp scalpel (*see* **Note 7**).
4. Add another 1 mL of the same ice-cold buffer as used in **step 2** (*see* **Note 8**).
5. Mix the crude suspension by gently shaking the petri dish.
6. Filter the homogenate through a 30–42 μm nylon mesh filter into a labelled flow cytometry tube (*see* **Note 9**). The chopping and filtration processes might result in a reduction in the final volume, especially when working with dried samples. To reduce any critical effect, (1) dried samples can be presoaked in buffer for up to 5–10 min before **step 2**, and (2) filters can also be soaked in buffer prior to filtration.
7. Add the appropriate volume of fluorochrome (Subheading 2.3.1) to the nuclei suspension and vortex gently. For a typical sample which is c. 2 mL, the amount of stock PI added is 100 μL (i.e., working concentration: 50 μg/mL), while for DAPI, 80 μL (i.e., working concentration: 4 μg/mL) should be added (*see* **Notes 10** and **11**).
8. Keep samples on ice until ready to analyze (*see* **Note 12**).
9. Proceed to analyze the nuclear DNA content, vortexing the sample before putting it on the flow cytometer (follow instructions in Subheading 3.2).

3.1.2 Isolation of Plant Nuclei Using the Two-Step Protocol

This procedure uses the Otto buffer (*see* Subheading 2.3.2; *see* also **Note 5** for an alternative buffer which can be used here).

1. Place a small amount of the selected plant tissue (usually about 1 cm^2 or 20 mg) in a 6 cm petri dish (*see* **Note 6**).
2. Add 1 mL of ice-cold Otto I buffer.
3. Chop the tissues in the buffer using a new razor blade or sharp scalpel (*see* **Note 7**).
4. Mix the crude suspension by gently shaking the petri dish.
5. Filter the homogenate through a 30–42 μm nylon mesh filter into a labelled 1.5 mL tube.

6. Pellet the nuclei by centrifuging at 150 × *g* for 5 min (*see* **Notes 13** and **14**).
7. Carefully remove the supernatant leaving approximately 100 μL of the buffer (*see* **Note 15**).
8. Resuspend the pellet by gently shaking and add a further 100 μL of the buffer used in **step 2** (*see* **Note 16**).
9. Add 1 mL of room temperature Otto II buffer (*see* **Note 17**).
10. Add the appropriate volume of the fluorochrome to the nuclei suspension (if it is not already in the buffer—*see* Subheading 2.3.2) and vortex gently. For a typical sample which is c. 1.2 mL, the amount of stock PI added is 60 μL, while for DAPI, 50 μL should be added.
11. Incubate the samples at room temperature for few minutes in the dark (*see* **Note 18**).
12. Proceed to analyze the nuclear DNA content, vortexing the sample before putting it on the flow cytometer (follow instructions in Subheading 3.2).

3.1.3 Isolation of Plant Nuclei Using a Simplified Two-Step Protocol

This procedure uses the Otto isolation buffer (*see* Subheading 2.3.2 above; *see* also **Note 19** for alternative buffers which can be used here).

1. Place a small amount of the selected plant tissue (usually about 1 cm^2 or 20 mg) in a 6 cm petri dish (*see* **Note 6**).
2. Add 0.5 mL of ice-cold Otto I.
3. Chop the tissues in the buffer using a new razor blade or sharp scalpel (*see* **Note 7**).
4. Mix the crude suspension by gently shaking the petri dish
5. Add 2 mL of Otto II buffer.
6. Filter the homogenate through a 30–42 μm nylon mesh filter into a labelled flow cytometry tube (*see* **Note 9**).
7. For a typical sample which is c. 2.5 mL, the amount of stock PI added is 100 μL to give a working concentration of 50 μg/mL, while for DAPI, 100 μL of the stock should be added to give a working concentration of 4 μg/mL to the nuclei suspension (if it is not already in the buffer—*see* Subheading 2.3.2) and vortex gently.
8. Incubate at room temperature for few minutes in the dark (*see* **Note 18**).
9. Proceed to analyze the nuclear DNA content, vortexing the sample before putting it on the flow cytometer (follow instructions in Subheading 3.2).

3.2 Analysis of the Nuclear DNA Content and DNA Ploidy Level

The flow cytometer allows the measurement of several optical properties of the isolated particles (i.e., nuclei) that move one by one through the flow capillary tube illuminated by a laser beam or mercury light source. Prior to analyzing any plant sample, check that the instrument is properly aligned using fluorescent calibration beads (*see* Subheading 2.2). Subsequently test the linearity of the flow cytometer by running a plant sample (e.g., reference standard) and comparing the ratio between the 4C/2C peaks, which ideally should be in the range of 1.98–2.02 sensu Doležel et al. [16].

The first step in the analysis of a new target species requires the user to determine its relative nuclear DNA fluorescence. This step is described in Subheading 3.2.1. Based on this information, the user can then proceed either to Subheading 3.2.2 to determine the absolute DNA amount or to Subheading 3.2.3 to determine the DNA ploidy level.

3.2.1 Measurement of the Relative Nuclear DNA Fluorescence of a Sample

1. Load the tube containing the suspension of stained nuclei onto the flow cytometer sample port and run for a few seconds at low speed until the flow has stabilized through the tubing system (*see* **Notes 20** and **21**).
2. Adjust the flow rate to a speed of 15–25 nuclei/s (*see* **Notes 22** and **23**).
3. Once the sample is running through the flow cytometer, a flow histogram with peaks will start to appear. The peak positions can then be adjusted using the instrument gain settings to move the peaks within the histogram (*see* **Notes 24** and **25**). It is also possible to adjust the lower limit threshold so that undesirable low-channel signals (e.g., from cell debris and autofluorescent compounds) are excluded from the histogram. If there is a large amount of cell debris/background fluorescence in the flow histogram, then *see* **Note 26**, while if additional, unexpected peaks appear, then *see* **Note 27**.
4. Measure 5000 particles (*see* **Note 28**).
5. Use the software provided by the flow cytometer manufacturer to assess the quality of histograms by (1) estimating the proportion of background, (2) checking peak symmetry, and (3) evaluating the peak width, expressed as the coefficient of variation, CV% (=SD of peak/mean channel position of the peak × 100) (*see* **Notes 29** and **30**).
6. Save the histogram if appropriate (*see* **Note 31**).

3.2.2 Measurement of the Absolute Nuclear DNA Content of a Sample Using a Reference Standard

Once the target sample has been run on its own to determine what reference standard to use and what gain the machine should be set at (*see* Subheading 3.2.1), a combined sample which includes both the target species and reference standard can then be prepared and

run to determine the absolute nuclear DNA content of the target species.

To ensure the estimate of nuclear DNA content in absolute units is as accurate as possible, FCM researchers have adopted several best practice approaches. These include the following recommendations: (1) three specimen plants are collected per population/species and three independent replicates are processed per sample, or (2) five specimens are collected per population/species and two independent replicates are processed per specimen; (3) only intercalating fluorochromes (e.g., PI) should be used; base-specific fluorochromes such as DAPI are not suitable for estimating nuclear DNA content.

1. Load the sample which contains a suspension of stained nuclei of both the target species and the selected internal reference standard (based on results obtained in Subheading 3.2.1) onto the flow cytometer sample port and run for a few seconds at low speed until the flow has stabilized through the tubing system (*see* **Notes 20** and **21**).
2. Adjust the flow rate to a speed of 15–25 nuclei/s (*see* **Notes 22** and **23**).
3. Once the sample is running through the flow cytometer, a flow histogram with peaks will start to appear. The peak positions can then be adjusted, if necessary, using the instrument gain settings to move the peaks within the histogram. It is also possible to adjust the lower limit threshold so that undesirable low-channel signals (e.g., from cell debris and autofluorescent compounds) are excluded from the histogram.
4. Check to see if there is any evidence of negative effects caused by the presence of cytosolic compounds which can affect the accuracy of the genome size estimation. This is done by comparing the position of the G_1 peak of the reference standard in this combined sample with its position in a sample containing just the reference standard (*see* Subheading 3.2.1). Both samples must be run at the same gain.
5. When this situation arises, alternative isolation methods should be tested (*see* **Note 32**); otherwise, proceed with the next step.
6. Measure 5000 particles (*see* **Note 28**) (in some protocols, 10,000 particles are recommended) and save the data (*see* **Note 31**).
7. Use the software provided by the flow cytometer manufacturer to assess the quality of histograms (*see* **step 5** of Subheading 3.2.1). Assuming the quality of the histogram is suitable (i.e., CVs $< 3\%$) (*see* **Note 29**), also obtain the statistical information for the histogram (i.e., mean peak position).

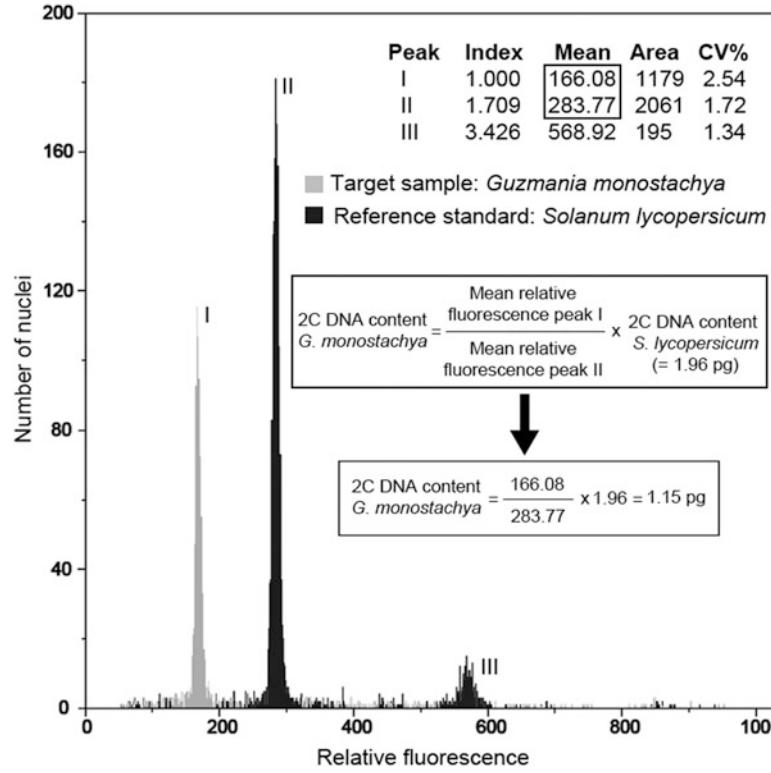


Fig. 3 A typical flow histogram to illustrate how genome size is calculated for the target species *Guzmania monostachya* using *Solanum lycopersicum* as the internal reference standard. Using the output data from the flow cytometer software, the mean relative fluorescence of the G_1 peak of *G. monostachya* (gray peak labelled I, i.e., 166.08) is divided by that of the mean G_1 peak of the standard *S. lycopersicum* (black peak labeled II, i.e., 283.77). This ratio is then multiplied by the 2C DNA content of *S. lycopersicum* to give the 2C value of *G. monostachya*. To convert between pg and Mbp, use the conversion factor 1 pg = 978 Mbp [1] (N.B. peak III is the G_2 peak of the reference standard)

8. Calculate the nuclear DNA amount (2C-value) of the target plant in each replicate as follows (*see* **Notes 33** and **34**):

$$\text{2C DNA content target (pg)} = \frac{\text{target sample mean } G_1 \text{ peak}}{\text{standard sample mean } G_1 \text{ peak}} \times \text{2C DNA content standard(pg)}$$

   For an illustrative sample histogram output and calculation, *see* Fig. 3.

9. Calculate the mean nuclear DNA content and the standard deviation for the species (including all specimens and replicates) (*see* **Note 35**). To convert between picograms (pg) and megabase pairs (Mbp), use: 1 pg = 978 Mbp [1].

3.2.3 Measurement of the Relative Nuclear DNA Content of a Sample Using a Reference Standard to Determine DNA Ploidy Level

Among the multiple uses of FCM, DNA ploidy estimation is becoming highly popular as it allows the rapid screening of multiple samples. The protocol described below is optimized to work at either the species level or within species complexes.

1. Load the sample which contains a suspension of stained nuclei of both the target species of unknown ploidy and either a reference sample comprising a species of known ploidy (i.e., karyologically determined) or another internal standard (in that case, as mentioned above, the ploidy level of at least one target sample must be karyologically determined) (*see* **Note 36**) onto the flow cytometer sample port and run for a few seconds at low speed until the flow has stabilized through the tubing system (*see* **Note 20**).
2. Perform **steps 2** and **3** (Subheading 3.2.2).
3. Measure at least 3000 particles (*see* **Note 37**) and save the data (*see* **Note 31**).
4. Use the software provided by the flow cytometer manufacturer to obtain the statistical information for the histogram (e.g., peak position and ratio, area, CV% (*see* **step 5** of Subheading 3.2.1)).
5. Calculate the relative nuclear DNA amount (DNA ploidy) of the target plant as follows:
   a. If the reference sample used (with known ploidy) is the same species as the target sample, a perfect overlap of G_1 peaks will indicate they both have the same ploidy.
   b. If multiple peaks appear, then calculate the ploidy level using the following formula:

$$\text{Target sample ploidy} = \frac{\text{target sample mean } G_1 \text{ peak}}{\text{standard sample mean } G_1 \text{ peak}} \times \text{reference sample ploidy}$$

   c. If one of the cultivars listed in Table 1 is used as the reference standard, ploidy levels can be inferred by means of the ratio between the G_1 peaks of both the standard and the target samples (keeping in mind that the chromosome number of at least one of the target samples must be known). An example of the FCM analysis of ploidy level in the genus *Sorbus* (Rosaceae) is given in Fig. 4.

3.2.4 Measurement of the Relative Nuclear DNA Content of a Sample to Determine the Extent of Endopolyploidy

The protocol for detecting and measuring the level of endopolyploidy is very similar to that described for the measurement of DNA ploidy (Subheading 3.2.3 above) with just a few small adjustments as outlined below.

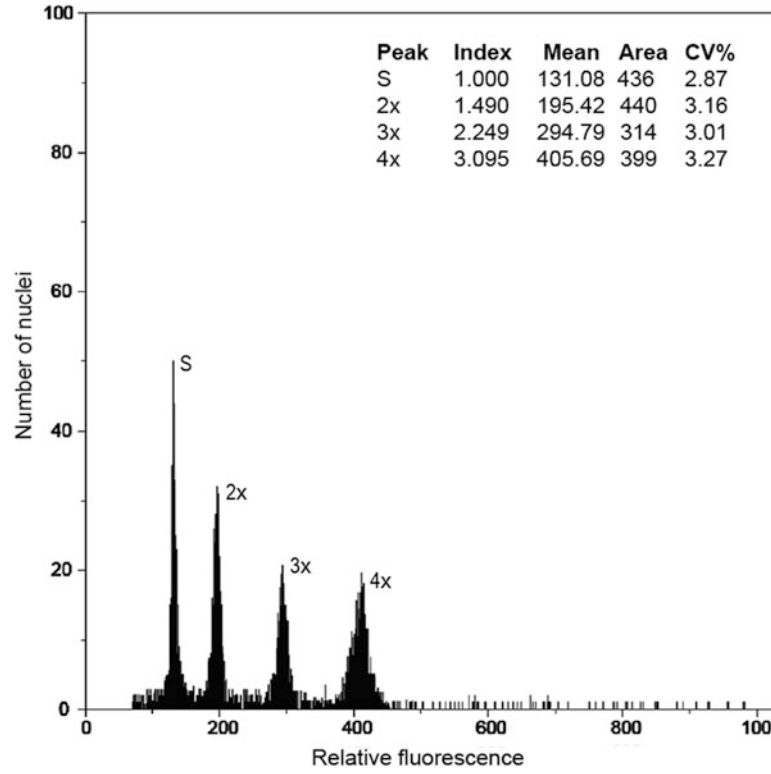


Fig. 4 Flow cytometric ploidy analysis in *Sorbus*. DNA ploidy was assessed in different species of *Sorbus* using the internal reference standard (*Oryza sativa*). Diploid *Sorbus aria* (whose chromosome number has been counted) was used as a reference to uncover higher ploidy levels in related species of unknown ploidy by determining the ratio between the peaks of these *Sorbus* species and the internal reference standard. S = G_1 peak of the internal standard (*Oryza sativa*); 2x = G_1 peak of the chromosomally determined diploid *S. aria*; 3x = G_1 peak of the triploid *S. saxicola*; 4x = G_1 peak of the tetraploid *S. rupicola*

1. Load the sample which contains a suspension of stained nuclei of the target species and reference standard onto the flow cytometer sample port and run for a few seconds at low speed until the flow has stabilized through the tubing system (*see* **Note 20**). Although it is recommended to include a reference standard with the target species when measuring endopolyploidy, in some cases, this may not be possible as the peak (s) of the reference standard may overlap with the peaks of the target species. If a suitable reference standard is not obtained (due to the overlapping peaks), the target species can be run alone.
2. Perform **steps 2** and **3** (Subheading 3.2.2). If a large number of peaks are observed (due to high levels of endopolyploidy), it may be necessary to adjust to a log scale so that all the peaks are displayed within the flow histogram (e.g., Fig. 5).

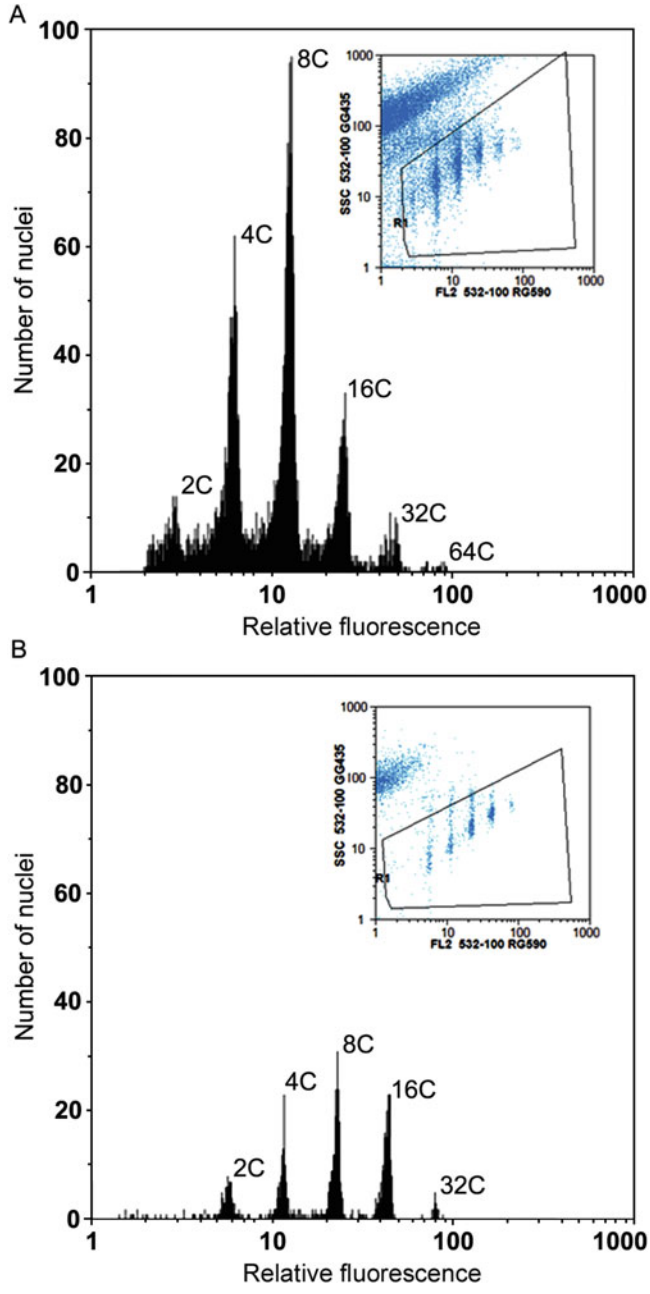


Fig. 5 Analysis of endopolyploidy in leaf material of (**a**) *Oscularia deltoides* (Aizoaceae) and (**b**) *Tylecodon paniculatus* (Crassulaceae) with different peaks corresponding to different levels of endopolyploidy, with some nuclei having undergone six endocycles to reach 64C as observed in *O. deltoides* and five endocycles to reach 32C in *T. paniculatus*

3. Measure at least 10,000 particles and save the data (*see* **Note 31**). It is important to measure this increased number of particles when exploring endopolyploidy to ensure that all peaks are visible on the histogram.
4. Use the software provided by the flow cytometer manufacturer to obtain the statistical information for the histogram (e.g., peak position and ratio, area, CV% (*see* **step 5** of Subheading 3.2.1)).
5. Using the measured area of each peak, the endoreduplication index (Ei) can be measured using the following formula [58]:

$$Ei = \left(\frac{(0\ x\ n_{2c}) + (1\ x\ n_{4c}) + (2\ x\ n_{8c}) + (3\ x\ n_{16c}) \ldots}{(n_{2c} + n_{4c} + n_{8c} + n_{16c} \ldots)} \right.$$

where n is equal to the area (number of nuclei) of each peak.

3.2.5 Using Seeds to Determine Reproductive Pathways Based on the Ratio between the Relative Nuclear DNA Content of the Embryo and the Endosperm

1. Follow the procedure to isolate nuclei describe in Subheading 3.1.1 using seeds as the starting material. Ideally, individual seeds are used; however, if seeds are very small, then several seeds (c. 5–10 seeds) can be processed and analyzed together.
2. Load the tube containing the suspension of stained nuclei onto the flow cytometer sample port and run for a few seconds at low speed until the flow has stabilized through the tubing system (*see* **Notes 20** and **21**).
3. Adjust the flow rate to a speed of 15–25 nuclei/s (*see* **Notes 22** and **23**).
4. Once the sample is running through the flow cytometer, a flow histogram with peaks will start to appear. The peak positions can then be adjusted using the instrument gain settings to move the peaks within the histogram (*see* **Notes 24** and **25**). If there is a large amount of cell debris/background fluorescence in the flow histogram, then *see* **Note 26,** while if additional, unexpected peaks appear, then *see* **Note 27**.
5. Measure between 1000 and 3000 particles or until the embryo and endosperm fluorescence peaks can be clearly identified and analyzed with the flow cytometer software (*see* **Note 28**).
6. Use the software provided by the flow cytometer manufacturer to assess the quality of histograms (*see* **Notes 29** and **30**).
7. Calculate the ratio between the fluorescence peak of the embryo and the endosperm. Interpretation of the results can be found in Matzk et al. [22].

4 Notes

1. Many of the chemicals that are used in FCM are considered hazardous, and so suitable protective equipment (i.e., gloves, lab coat, fume cupboard) should be used to avoid health risks, and manufacturer's safety recommendations should be followed when using them. For example:

   MOPS (3-morpholino-propanesulfonic acid) and DTT (dithiothreitol) may cause irritation to the eyes, respiratory system, and skin.

   β-Mercaptoethanol is very hazardous and can be fatal if inhaled, swallowed, or absorbed through skin contact.

   PI is a potential mutagen and may cause irritation to the eyes, respiratory system, and skin.

   DAPI is a potential carcinogen and may cause irritation to the eyes, respiratory system, and skin.

   DMSO (dimethyl sulfoxide) itself is not considered as a hazardous substance but in contact with other potentially toxic chemicals might enhance their absorption through the skin.

2. Intercalating fluorescent dyes (IFD) bind to double-stranded DNA and RNA with no base preference. These are suitable for genome size estimations in absolute units (the majority of studies in plants use PI). Fluorescent dyes that bind preferentially to base-specific rich DNA (BSD), either AT-rich or GC-rich DNA, are not suitable for estimating genome size but are frequently used for ploidy level studies.

   If users wish to test SYBR Green I as a fluorochrome, then it can be prepared as follows: The stock solution provided by the manufacturer is usually 10,000× concentrate, and manufacturers recommend a working concentration of 10×. The stock should first be diluted 100-fold in DMSO (dimethyl sulfoxide—*see* **Note 1**) to give a diluted solution of 100× (e.g., 50 μL SYBR I in 4.95 mL of DMSO). This 100× solution can be stored in 5 mL aliquots at −20 °C. For use, the appropriate volume of this diluted 100× solution is added to the nuclei isolation buffer to give a final working concentration of 10×.

3. Isolation buffers must be prepared using either single or double distilled water, filtered through a 0.22 μm filter to remove suspended particles, and stored as specified. The isolation buffer might precipitate after a while if it has not been stored at the appropriate temperature or when poor quality water has been used (*see* **Note 38**). The pH of the buffer is adjusted either with 1 M NaOH or 1 N HCl (*see* **Note 39**). Further information about the roles of the different buffer components is given in **Notes 40** and **41**.

4. While the General Purpose Buffer given in Subheading 2.3.2 works for many plant species, the selection of the most appropriate buffer needs to be determined empirically for each plant group. In many cases, the same buffer works well across a family, while in other cases, different buffers are needed for different genera, or even within a genus. Other buffers which have been shown to work across a diversity of plants include Ebihara's, LB01, Galbraith, woody plant buffer, and commercial buffers such as CyStain and OxProtect (Sysmex Ltd.).

   A comprehensive list of alternative buffers and their components is given below. These isolation buffers must be prepared using either single or double distilled water, filtered through a 0.22 μm filter to remove suspended particles, and stored as specified. Most buffers remain stable for up to 3 months if appropriately stored. As indicated below, some buffers can be stored longer by freezing them in aliquots at −20 °C. However, if this is done, then once thawed, the buffer should not be refrozen.

   (a) *LB01 buffer [59]*: 15 mM Tris, 2 mM Na_2EDTA, 0.5 mM spermine·4HCl, 80 mM KCl, 20 mM NaCl, 0.1% (v/v) Triton X-100. Adjust to pH 7.5. Add β-mercaptoethanol to give a final concentration of 15 mM (*see* **Note 1**). Store the buffer either at 4 °C if used regularly or at −20 °C in 10 mL aliquots.

   (b) *Tris $MgCl_2$ buffer [60]*: 200 mM Tris, 4 mM $MgCl_2$, 0.5% (v/v) Triton X-100. Adjust pH to 7.5 and store at 4 °C.

   (c) *Galbraith buffer [61]*: 45 mM $MgCl_2$, 20 mM MOPS (*see* **Note 1**), 30 mM sodium citrate, 0.1% (v/v) Triton X-100. Adjust pH to 7.0. Store the buffer either at 4 °C if used regularly or at −20 °C in 10 mL aliquots.

   (d) *Woody plant buffer [56]*: 200 mM Tris, 4 mM $MgCl_2$, 2 mM Na_2EDTA, 86 mM NaCl, 10 mM sodium metabisulfite, 1% PVP-10 (*see* **Note 41**), 1% (v/v) Triton X-100. Adjust to pH 7.5. Store the buffer either at 4 °C if used regularly or at −20 °C in 10 mL aliquots.

   (e) *$MgSO_4$ buffer [62]*: 9.53 mM $MgSO_4$, 47.67 mM KCl, 4.77 mM HEPES, 6.48 mM DTT (*see* **Note 1**), 0.25% (v/v) Triton X-100. Adjust to pH 8.0. Store the buffer either at 4 °C if used regularly or at −20 °C in 10 mL aliquots.

   (f) *Bino's buffer [63]*: 200 mM mannitol, 10 mM MOPS (*see* **Note 1**), 0.05% (v/v) Triton X-100, 10 mM KCl, 10 mM NaCl, 2.5 mM DTT (*see* **Note 1**), 10 mM spermine·4HCl, 2.5 mM Na_2EDTA, 0.05% (w/v) sodium azide (*see* **Note 1**). Adjust to pH 5.8 and store at 4 °C.

(g) *De Laat's buffer [64]*: 15 mM HEPES, 1 mM Na_2EDTA, 0.2% (v/v) Triton X-100, 80 mM KCl, 20 mM NaCl, 15 mM DTT (*see* **Note 1**), 0.5 mM spermine·4HCl, 300 mM sucrose. Adjust to pH 7.0 and store at 4 °C.

(h) *Ebihara's buffer [65]*: 50 mM Na_2SO_3, 50 mM Tris, 40 mg/mL PVP-40 (*see* **Note 41**), 140 mM β-mercaptoethanol (*see* **Note 1**). Adjust to pH 7.5 and store at 4 °C.

(i) *Seed buffer [66]*: 5 mM $MgCl_2$, 85 mM NaCl, 100 mM Tris, 0.1% Triton X-100. Adjust to pH 7.0 and store at 4 °C (*see* **Note 42**).

(j) *Gif nuclear buffer (GNB) [67]*: 45 mM $MgCl_2$, 30 mM sodium citrate and 60 mM MOPS, pH 7.0, 1% PVP 10,000, 0.1% Triton X-100 and 10 mM sodium metabisulfite ($S_2O_5Na_2$). This buffer can be stored at 4 °C (*see* **Note 42**), but the metabisulfite is added daily.

(k) *Baranyi buffer [68]*: Baranyi solution I: 100 mM citric acid monohydrate, 0.5% (v/v) Triton X-100. Store at 4 °C.

Baranyi solution II: 400 mM Na_2HPO_4, 10 mM sodium citrate, 25 mM sodium sulfate. Store at room temperature.

The fluorochrome (DAPI or PI; *see* Subheading 2.3.1) can be added to Baranyi solution II before adjusting the final volume of the stock solution. If this is done, the buffer should be stored in the dark at room temperature. Alternatively, the fluorochrome can be added directly to the sample at **step 10** of Subheading 3.1.2 or **step 7** of Subheading 3.1.3.

(l) *Mishiba buffer [69]*:Solution A: see recipe for Galbraith buffer, i.e., buffer (c) above.

Solution B: 10 mM Tris, 50 mM sodium citrate, 2 mM $MgCl_2$, 1% PVP-40 (original recipe used PVP K-30—*see* **Note 41**), 0.1% Triton X-100, 18 mM β-mercaptoethanol (*see* **Note 1**). Adjust to pH 7.5. Store at 4 °C.

(m) *Commercially available buffers*: *Sysmex PI Absolute P* (Sysmex cat. no. 05-5022) and *Sysmex CyStain PI OxProtect* (Sysmex cat. no. 05-5027). Follow manufacturer's instructions for sample preparation. Both buffers contain PI-based staining solutions which need to be kept in the dark for long-term storage. Store at between 2 and 8 °C.

5. The Baranyi buffer (comprising solutions I and II—see buffer (l) in **Note 4** [68]) can be tried as an alternative to the Otto buffer in the two-step protocol (Subheading 3.1.2) and simplified two-step protocol (Subheading 3.1.3) if the results from using the Otto buffers give poor histograms.

6. The amount of tissue used needs to be determined empirically, taking into account the number of nuclei released and the proportion of debris produced. For internal standardization, when needed, also add leaf tissue of the appropriate reference standard species (*see* Table 1).
7. It is very important to use very sharp razor blades or scalpels to chop the tissue into a crude suspension, while minimizing damage to the nuclei. It is therefore recommended that each razor blade or scalpel is used only once. The chopping must be vigorous, quick, and short to avoid drying of the sample. We recommend empirical adjustments, especially to the chopping intensity, so that optimal numbers of nuclei are released without generating too much cell debris which can lead to high background signal in the flow histogram and low numbers of nuclei in the G_1 peak.
8. The working volume can be modified, but remember that if this is done, then the volume of the fluorochrome added at **step 7** will need to be adjusted accordingly to maintain the appropriate final concentrations.
9. Check carefully that the sample is free of particles after filtration to minimize the possibility of blockages in the flow cytometer.
10. If the samples have become brown/dark just a few minutes after adding the fluorochrome, this is indicative that the sample is undergoing oxidation due to the presence of secondary metabolites in the cytoplasm. The reaction can be slowed down by keeping everything cool during nuclei isolation and analysis (e.g., using ice-cooled samples, petri dishes, isolation buffers, and sample tubes and placing the prepared sample of isolated nuclei on ice during flow cytometry). Sometimes, this problem can also be avoided by supplementing the isolation buffer with reducing agents such as β-mercaptoethanol and DTT (*see* **Note 30**). Another option that might help is the addition of PVP-10, PVP-40, or higher molecular weights such as PVP-360 (*see* **Note 41**), which will help improve histogram quality and sample stability, especially if tannins are present. If the problem persists, then alternative isolation buffers (*see* **Note 4**) should be tested and the chopping intensity reduced (*see* **Note 7**).
11. Many protocols add RNase (Ribonuclease II-A) at 50 μg/mL at this stage when PI is used as the fluorochrome. This is because PI intercalates into double-stranded (ds) nucleic acids so it can stain dsRNA as well as dsDNA. Nevertheless, since RNase is only active between 15 and 70 °C, with an optimal temperature of 60 °C, it can be left out of any protocol that lacks an incubation step within this temperature range. Since the protocols described here do not include such an

incubation step, RNase has not been included. Nevertheless, if users want to include an RNase incubation step, a stock RNase solution can be prepared by heating 1 mg/mL RNase to 80 °C for 15 min (to inactivate DNases) and filtering through a 0.22 μm filter. The stock can be stored in 1 mL aliquots at −20 °C (note that 100 μL of the stock into a 2 mL final sample volume should be added).

12. The time between staining (**step 7**) and running the sample on a flow cytometer can vary from a few minutes to up to 1 h (and in rare cases up to half a day). While for some plant samples, a short incubation works fine, for others, a longer incubation can give better results. In cases where there is a large amount of debris, keeping the incubation time to just a few minutes can improve the quality of the flow histograms generated. Thus, incubation time needs to be adjusted empirically for each plant species to optimize results.
13. The relative centrifugal speed and time may need to be empirically adjusted.
14. Samples are stable in Otto I (or Baranyi solution I—*see* **Note 5**); hence, it is possible to prepare several samples in advance and simultaneously centrifuge them together.
15. It is important to do this step very gently so as not to remove the pelleted nuclei.
16. As samples are stable in Otto I (or Baranyi solution I—*see* **Note 5**), it is possible to prepare many replicates and store them at either room temperature or 4 °C for up to several hours.
17. The addition of Otto II (or Baranyi solution II —*see* **Note 5**) raises the pH of the sample to c. 7.3 and increases salt concentration. To keep these parameters within a working range, the amount of buffer added at this stage should be about fourfold that of Otto I (or Baranyi solution I—*see* **Note 5**) which now comprises c. 200–250 μL.
18. The optimal incubation time should be adjusted in each case, but short incubation times (e.g., less than 5 min) usually provide the best results because nuclei may not remain stable for a long time after this step.
19. When following the simplified two-step protocol (Subheading 3.1.3), Baranyi or Mishiba buffer (*see* **Note 4** above) can also be tried if the histograms obtained using the Otto buffer are of poor quality.

    If Mishiba buffer is used, follow the protocol in Subheading 3.1.3 with the following modifications:

    Step 2.* Add 0.2 mL of ice-cold Mishiba solution A.

    Step 4.* Incubate for 5 min at room temperature.

Step 5.* Add 1 mL of Mishiba solution B.

Step 7.* For a typical sample using Mishiba buffer, the volume is usually c. 1.2 mL; thus, the amount of stock PI added is 60 μL, while for DAPI, 50 μL should be added.

Step 8.* Incubate at room temperature in the dark for 20 min.

20. The user can clear the acquisition results as many times as needed until the flow rate becomes stabilized. Do not be tempted to start recording data for analysis until the flow rate has stabilized (usually 0.5–1 min after the start of the run) as this can lead to poor histograms and inaccurate results.
21. If no peaks are appearing in the flow histogram, and assuming that the flow cytometer is properly set up, the peaks are probably off the scale due to an inappropriate gain setting for the sample being analyzed. To locate the peaks, adjust the gain setting of the machine. This can sometimes be done more easily by using the log scale setting of the relative fluorescence (*x* axis) scale. Once the position of the peak has been located, adjust back to a linear scale to perform analyses. Remember that the gain of the machine should be kept within the range that is recommended by the manufacturer of the flow cytometer to ensure the machine is operating optimally.
22. If the flow rate is slow, there are several explanations and possible solutions:
    (a) This could be a technical problem with the flow cytometer. Any blockage in the flow chamber or in the tubing system can cause a reduction in the number of nuclei recorded. Check that the pressure in the system is within the recommended range for the machine and clean the flow system using either a decontaminant solution or a diluted bleach solution (*see* Subheading 2.2) to wash out any potential blockage.
    (b) Alternatively, this could be a biological problem caused by the plant material being analyzed. The concentration of nuclei in the suspension can vary significantly between samples depending on tissue type, quantity of material used, etc. Hence, the flow rate will need to be adjusted each time a new sample is loaded onto the machine. If the concentration of nuclei in the sample is low, this will necessitate a high flow rate, and this can result in a broadening of peaks and high CVs (*see* **Notes 29** and **30**). If possible, it is best that this is overcome by preparing a new sample using more material to increase nuclei concentration, rather than running the sample at a high flow rate. When using flow cytometers with a pre-set sample acquisition rate (e.g., slow, medium, high), we recommend

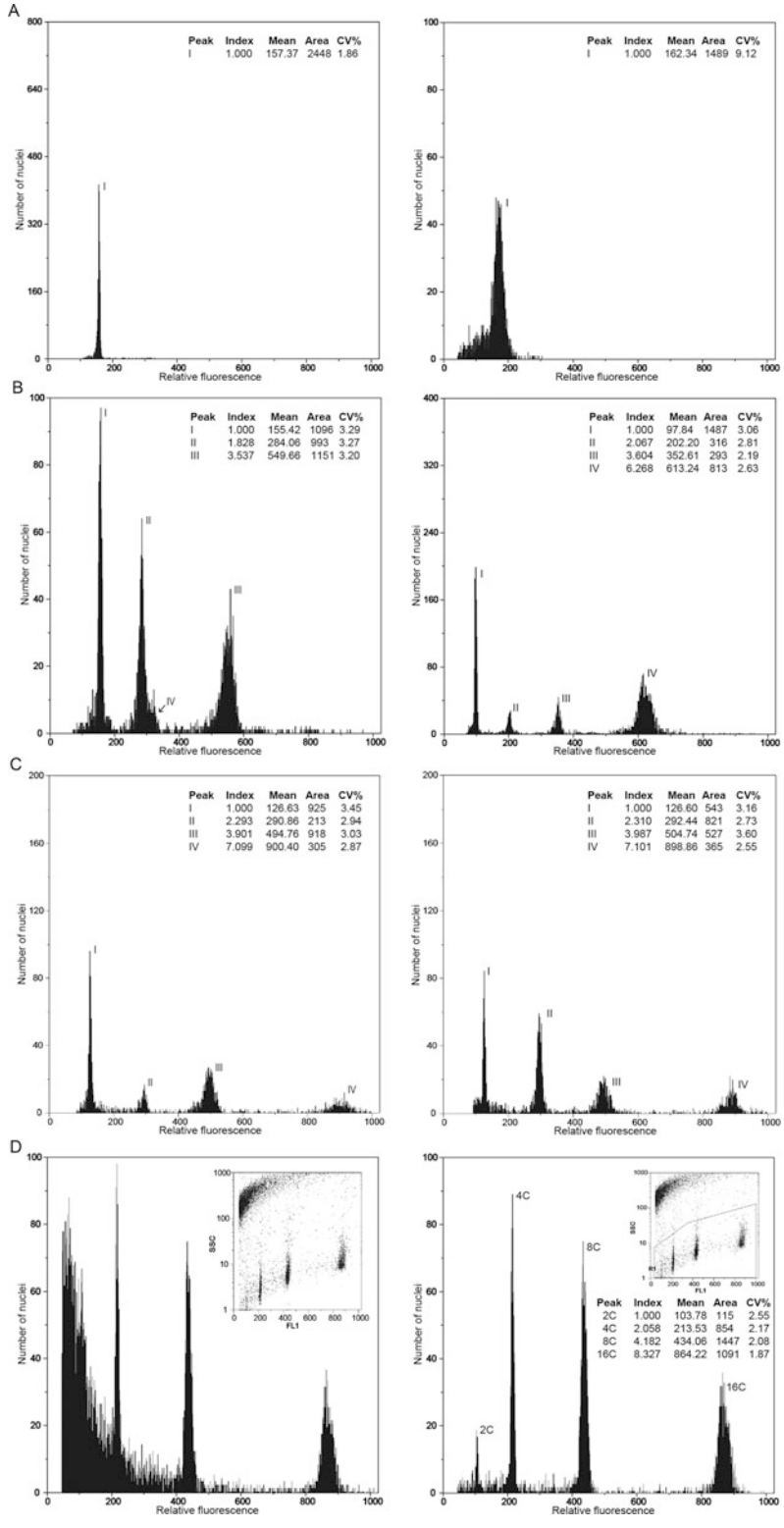


Fig. 6 Troubleshooting problems encountered during flow cytometric analysis of plant material. (**a**) Fluorescence histograms obtained after analysis of isolated nuclei of *Clusia multiflora* (Clusiaceae). Samples in both histograms were prepared using the same leaf and the same isolation buffer (woody plant buffer—[56]) but

using the slow rate and only increase it if absolutely necessary.

Other possible causes of a slow flow rate include inappropriate chopping intensity, the tissue used is not suitable, and/or the isolation buffer selected is not appropriate. Such problems can be overcome by, for example, increasing the amount of tissue used, adjusting the chopping intensity, and testing different types of plant material (*see* Subheading 2.1 and Fig. 6b which illustrates the effect of changing from leaf material to pollinia in the orchid *Dactylorhiza*). Even changing the end of the leaf used for analysis can result in a dramatic change in the proportion of G_1 nuclei released (*see* Fig. 6c). Changing the isolation buffer (*see* **Note 4**) can also have a large effect, especially if the sample is releasing mucilaginous compounds into the chopping buffer. Indeed, many plants contain mucilage in their cytoplasm, and isolated nuclei may bind to this during the chopping process leading to a low number of released nuclei. Increasing the percentage of detergent (e.g., Triton X-100 up to 4%) can help, but keep in mind that a higher concentration of detergent can also result in higher levels of cell debris and hence a lower quality of flow histograms, so compromise may be necessary.

If the sample is releasing large amounts of mucilaginous compounds, improved results may be obtained by passing the filtered sample (after **step 6**, Subheadings 3.1.1 and 3.1.3; **step 5**, Subheading 3.1.2) through cotton wool saturated in buffer as outlined in Lee and Lin

Fig. 6 (continued) supplemented with different types of polyvinylpyrrolidone (PVP) to illustrate the dramatic effect on the quality of the flow histogram. (Left) using PVP-360 and (right) using PVP-40. (**b**) Flow histograms of the relative fluorescence in *Dactylorhiza* sp. (Orchidaceae) illustrating the utility of using alternative tissues to leaf samples to estimate nuclear DNA contents. (Left) genome size estimated using pollinia of *Dactylorhiza* sp. and *Solanum lycopersicum* as internal standard [standard: peak I (G_1) and IV (G_2), pollinia: peak II (1C-G_1) and III (2C-G_2); calculated 1C-value of *Dactylorhiza* sp. = 3.58 pg]. (Right) genome size estimated using leaf tissue of *Dactylorhiza* sp. and *Solanum lycopersicum* as internal standard [standard: peak I (G_1) and II (G_2), *Dactylorhiza* sp. leaf: peak III (2C-G_1) and IV (4C-G_2); calculated 2C = 7.06 pg]. (**c**) Flow histograms of leaf tissue from the orchid *Dracula* sp. (using *Oryza sativa* as internal standard (peak I)), illustrating how different parts of the same leaf can have very different proportions of G_1 and G_2 nuclei. Using a young and actively growing leaf of *Dracula* (c. 1.5 cm long), the apical tip was seen to have a much lower proportion of G_1 nuclei (peak II, left histogram) compared with the basal part of the leaf (peak II, right histogram). (N.B. peaks III and IV correspond to G_2 and partial endopolyploid nuclei, respectively.) (**d**) Flow histograms of relative fluorescence in leaf tissue of *Kalanchoe marnieriana* (Crassulaceae) illustrating how poor histograms with much debris (left, ungated histogram) can be improved by gating the histogram (right) to reveal not only the G_1 nuclei of *K. marneriana* which was hidden in the debris of the left histogram but also the presence of several endopolyploid cycles

[70]. Also, increasing the acidity by adding HCl, HNO_3, or acetic acid to the Otto I buffer can help to dissolve mucilaginous or other substances which are preventing the nuclei release (e.g., *see* [34] for details).

23. If the flow rate is unstable just after starting acquisition and large numbers of particles are being recorded, even when the flow cytometer is running at a slow speed, this may be due to unstable pressure in the flow cytometer. It can be caused by several factors including the presence of suspended particles (e.g., algae) in the sheath fluid and sheath fluid tubes/filters. Check that the pressure is correct and replace sheath fluid, tubes, and filters. If algae become a recurrent problem, 0.02% sodium azide can be added to the water in the sheath fluid bottle; however, it should be noted that sodium azide is toxic and should be handled appropriately. Alternatively, the sheath fluid bottles can be thoroughly rinsed with domestic bleach (*see* Subheading 2.2) every 2 months, or even more frequently when they are not changed on a daily basis. In addition, many manufacturers recommend that the sheath fluid tubing and filters are replaced every 3 months.

24. Given that most of the measurements will require the use of a reference standard (*see* Subheadings 3.2.2 and 3.2.3), it is strongly recommended that the user knows, in advance, the peak position of a set of reference standards, ranging from small to big genomes (check Table 1 for recommended reference standards). This can be done by adjusting the gain settings so that the G_1 peak of the standard always falls, for example, around channel number 200 (N.B. some flow cytometry machines do not include an option to adjust the gain, in such cases ensure the gate is set correctly to include the peaks of interest). Then, when the target sample is run alone for the first time, the user will be able to determine the best reference standard by testing the peak position of the target plant at the different gains selected for the standards. It is noted that while the G_1 peak is usually the dominant peak, in many cases, G_2 peaks are present which might interfere with the target sample. Care should therefore be taken to note where the peak positions of the target sample fall in relation to both the G_1 and G_2 peaks of the reference.

    The positions of the G_1 peaks in the flow histogram for the target plant and the internal reference standard should be different enough to avoid overlapping peaks. However, ideally, the ratio between the G_1 peaks for the standard and the target plant should not exceed threefold to reduce risk of errors arising due to loss of linearity in the flow cytometer.

25. If the position of the peak appears to be unstable (i.e., the peak in the histogram builds at a different position each time the acquisition data are cleared), it may suggest the incubation time following the addition of the fluorochrome is insufficient (i.e., **step 8**, Subheadings 3.1.1 and 3.1.3; **step 11**, Subheading 3.1.2). Check different incubation times to test staining stability. If the problem persists, test alternative isolation buffers. However, when using the Otto buffers (Subheadings 3.1.2 and 3.1.3) (or Baranyi buffers—*see* **Note 5**), the nuclei may be unstable once Otto II (or Baranyi solution II) has been added (*see* **step 11**, Subheading 3.1.2, or **step 8**, Subheading 3.1.3). For these buffers, increasing the incubation time is only likely to lead to deterioration in the flow histogram quality and unstable peaks.
26. Large amounts of cell debris/background signal in the flow histogram are a commonly encountered problem (e.g., see histogram on the left of Fig. 6d). There are several explanations and solutions:
    (a) The isolation buffer selected is not appropriate for the sample. Test an alternative isolation buffer (*see* **Note 4**).
    (b) The tissue selected is not in good condition or optimal for FCM. Test other plant tissues (*see* Subheading 2.1).
    (c) The length of time that samples are kept on ice before being analyzed (**step 8**, Subheading 3.1.1) or incubated at room temperature (**step 11**, Subheading 3.1.2; **step 8**, Subheading 3.1.3) can influence the quality of the flow histogram, so try adjusting the incubation time.
    (d) Over-chopping of the sample (*see* **step 3** in Subheadings 3.1.1–3.1.3) can, in some cases, lead to large amounts of background debris in the flow histogram. Reduce chopping intensity and use a new sharp razor blade or scalpel for each sample to avoid cell damage. Reduced chopping has been shown to significantly improve the quality of the flow histograms when working with highly succulent species such as those belonging to Aizoaceae, Asphodelaceae, and Crassulaceae [71]. Over-chopping may also yield poor results when working with particularly tough leaves such as those found in some palm (Arecaceae) or gymnosperm species. In such cases, we also suggest first incubating the leaves in the isolation buffer for 5 min on ice so they are easier to chop.
    (e) If none of the above solutions improve the quality of the flow histogram, then gating can be tried if the flow cytometer is fitted with a side scatter detector. For this, the region of interest is selected in the side scatter vs. forward light histogram so that the flow histogram of relative

fluorescence excludes the signals coming from the side scatter. An example of how effective this can be is shown in Fig. 6d.

27. If additional and perhaps unexpected peaks which do not follow an endopolyploid series are present in the flow histogram, this suggests the presence of contaminants such as insects, insect eggs, and fungi in the plant sample. To avoid this problem, always check the plant material carefully before chopping (using a stereomicroscope if necessary) to ensure there are no contaminating organisms. If endoparasites are suspected, then alternative plant parts will have to be tested.
28. The number of particles that need to be recorded will vary depending on the type of analysis being carried out. Usually, it is recommended that 5000 particles are recorded for estimations of genome size, although for some materials, it may not be possible to obtain so many nuclei (e.g., for recalcitrant material or for species where only limited amounts of material are available). For ploidy level estimations, then typically data from 3000 particles are recorded.
29. The CV of a peak is a measure of peak quality and must be kept as low as possible (ideally less than 3%) and always below 5%. Higher CVs are not acceptable for publication unless it has been demonstrated that higher quality cannot be achieved after extensive tests with different buffers, incubation times, types of material, etc. (e.g., samples rich in polyphenols, old silica dried samples, and herbarium vouchers).
30. Broad peaks with high and unacceptable CVs are, unfortunately, commonly encountered in the analysis of plant material. There are several possible explanations and solutions. These can broadly be divided into technical and biological sources:

    Technical

    (a) A loss of pressure in the flow cytometer system might result in a reduction of the peak quality. Check that the pressure is correct.

    (b) The instrument might be out of alignment. Align the instrument light source by using calibration beads (*see* Subheading 2.2).

    (c) Broad peaks are produced when the flow rate is too high. Run the samples at a flow rate that is no greater than c. 20 particles/s.

    (d) Air bubbles in the flow system can cause peaks with high CVs. Clean the flow chamber as recommended by the manufacturer and take extra care to remove any air bubbles from the filter after the sheath fluid bottle has been

refilled. Also make sure that the lid of the sheath fluid bottle is tightly screwed on to seal the system.

(e) As reported by Doležel et al. [16], an obsolete arc lamp used for UV excitation might be the cause of this problem. Replace the lamp and align the instrument.

(f) Weak fluorescence and peaks with large CVs can arise when a sample of DAPI-stained nuclei is analyzed following a sample of PI-stained nuclei. Doležel et al. [16] noted that this situation can arise as a result of fluorochrome interference if the flow cytometer has not been completely cleaned between samples. To avoid this problem, ensure that the machine is thoroughly washed through by running a tube containing a weak solution (1:5 dilution in distilled water) of domestic bleach (do not leave bleach sitting in the system for more than a few minutes) and then washing the system thoroughly with distilled water.

Biological

(a) In some cases, the isolation protocol and/or the buffer used are unsuitable for the material being analyzed, and the result can be a poor quality flow histogram with large CVs. Test alternative isolation buffers (*see* **Note 4**) and protocols (Subheading 3.1).

(b) Secondary metabolites in the cytoplasm may interfere with the fluorochrome staining of the DNA and lead to an increase in CVs. Sometimes, this can be overcome by supplementing the isolation buffer with reducing agents such as β-mercaptoethanol and dithiothreitol (DTT) (250 μL per 200 mL of buffer). Tannins are also frequent in plants, so the addition of PVP-10/PVP-40 is common to help minimize their effects. PVP-360 has also been shown to be effective and, in certain cases, may work when other PVP types have failed (e.g., *see* Fig. 6a and **Note 41**). The effect of secondary metabolites can also be minimized by reducing the chopping intensity (*see* **Note 26 (d)**) and carrying out the nuclei isolation steps on ice and with ice-cold solutions (as recommended in Subheading 3.1).

(c) Some tissues of some plants are just recalcitrant and produce poor results. Test alternative tissues (*see* Subheading 2.1), including pollinia (e.g., *see* Fig. 6b) or different parts of the leaf (e.g., *see* Fig. 6c), or try putting the plant in the dark for a few days prior to analysis.

(d) Doležel et al. [16] reported that excluding RNase from the isolation buffer when PI is used to stain DNA can result in increased CVs, especially in tissues with active protein synthesis, such as root tips. Nevertheless, if this is

the case, then the protocol should include an incubation step at 37 °C for at least 30 min to ensure the RNase has sufficient time to work (*see* **Note 11** for how to prepare RNase).

(e) The wrong concentration of the DNA fluorochrome can also reduce the quality of the flow histogram, so it is important to check that the fluorochrome solution has been prepared correctly.

31. It is recommended that when a run is saved, the file name should include information on the species analyzed, replicate number, buffer used, and internal reference standard (if applicable). If possible, it is also helpful to get the software to list the instrument settings (e.g., gain and lower limit settings) used for each run. This enables histograms to be compared, if appropriate.

32. If a shift in the position of the G_1 peak of the reference standard is detected, then it is necessary to change the sample preparation. Often, the problem can be solved by changing to another isolation buffer (*see* **Note 4**). Alternatively, the addition of various compounds can sometimes eliminate the problem, e.g., the addition of 3% PVP (*see* **Note 41** and Fig. 6a) to bind to polyphenolics or addition of dithiothreitol (DTT) or β-mercaptoethanol which is a good reducing agent (*see* **Note 30**). In addition, the problem can sometimes be overcome by using different plant materials such as roots, stems, bracts, and seeds (e.g., *see* Subheading 2.1 and Fig. 6b and c).

33. For accurate nuclear DNA amount estimations, it is recommended that the number of particles in both the target and the reference standard G_1 peaks should be similar.

34. Some plant breeding material, pollen, and the gametophyte stage of bryophyte groups (i.e., mosses, liverworts, and hornworts) are haploid. In such cases, the first peak of the target sample in a flow histogram (G_1) will correspond to the 1C rather than the 2C-value.

35. Technical factors should not account for more than 2–3% of the variation between different estimates for the same species, although for some materials (e.g., recalcitrant tissues), this type of variation may be greater. Higher levels of variability in C-value estimates for a species may reflect intraspecific variation due to the presence of chromosomal instabilities (e.g., B chromosomes, supernumerary segments or aneuploidy) or taxonomic heterogeneity in the samples analyzed (*see* [8, 72] for further discussion on intraspecific variation).

36. (1) Wherever possible, it is recommended that the "reference" sample of known ploidy is at the lowest ploidy level known for a given species/complex (i.e., diploid). (2) If investigating

ploidy levels within a species, the reference sample can be a sample of the species whose ploidy has been karyologically determined (e.g., a diploid sample). (3) Following the recommendations of Doležel et al. [16], the nuclear DNA may be stained with PI or DAPI, although the latter option may result in higher-quality histograms (i.e., the peaks have smaller CVs). The use of DAPI is also recommended to detect aneuploid specimens.

37. For ploidy level estimations, it is not necessary to measure as many nuclei as needed to estimate the absolute genome size of a sample (*see* Subheading 3.2.2); thus, data for a lower number of particles can be collected.

38. If the isolation buffer becomes cloudy, changes color, or contains suspended particles, it suggests the buffer has been stored incorrectly or that the storage time has been exceeded. In either case, this can result in fungi or bacteria growing in the buffer. If this has happened, then new isolation buffer needs to be prepared and stored as indicated (*see* Subheading 2.3.2 and **Note 3**). Unused buffer should be discarded after 3 months. It is also strongly recommended to prepare small volumes (e.g., 200 mL) so that the stocks are as fresh as possible.

39. The pH of the isolation buffers must be above 4 for PI to stain the DNA; most are around a neutral pH. For protocols using either Otto buffer (*see* Subheadings 3.1.2 or 3.1.3) or Baranyi buffer (*see* **Note 5**), the nuclei are isolated in a citric acid solution which is acidic (i.e., Otto I or Baranyi solution I). The pH is then raised to neutral by the addition of a basic solution containing Na_2HPO_4 to ensure optimum staining of the DNA when the fluorochrome is added.

40. Isolation buffers contain several different components which ensure that enough nuclei are released from the cells and that the DNA is protected from degradation and binds the fluorochrome quantitatively. Typically, isolation buffers include the following components: (1) organic buffers (e.g., Tris, MOPS, HEPES) which stabilize the pH between 7.0 and 8.0 (depending on the buffer) to hence enable DNA staining by the fluorochrome; (2) non-ionic detergents (e.g., Triton X-100 and Tween 20) to facilitate the release of nuclei and prevent their aggregation; (3) chromatin stabilizers (e.g., spermine, $MgCl_2$, $MgSO_4$) to maintain the integrity of DNA; (4) chelating agents (e.g., Na_2EDTA (ethylenediaminetetraacetic acid disodium salt), sodium citrate) to bind divalent cations such as Mg^{2+} and Mn^{2+} and hence block DNase activity; and (5) inorganic salts (e.g., KCl, NaCl) to ensure the correct ionic strength of the buffer. Some buffers also include β-mercaptoethanol, DTT, ascorbic acid, or sulfite which acts as a reducing agent to

prevent protein oxidation, and PVP (*see* **Note 7** below). For a discussion of the effect of different buffer components in a range of plant species, *see* Loureiro et al. [73] and Greilhuber et al. [74].

41. The polymer PVP (polyvinylpyrrolidone) is used to reduce the effect of polyphenols and other secondary metabolites such as tannins that are often present in plant tissues and which can inhibit the quantitative staining of DNA by the fluorochrome. Such secondary metabolites may also increase cell debris leading to a significant reduction in the quality of the peaks in the flow histogram (*see* **Notes 31** and **35**). Generally, PVP-10 and PVP-40 are used although in certain cases, only PVP-360 was shown to result in decent flow histograms (*see* Fig. 6a).
42. A modified version of this buffer was reported by Hörandl et al. [75] who also added 6.1 mM sodium citrate to the buffer.
43. It is *essential* that the cell culture tested grade of Tween 20 from Sigma-Aldrich (cat. no. P2287) is used. Tween 20 for molecular biology (Sigma, cat. no. P9416) is not suitable for FCM.
44. Dissolving $Na_2HPO_4{\cdot}12H_2O$ can be speeded up by heating the solution gently.

Acknowledgments

J. P. benefited from a Ramón y Cajal Fellowship (RYC-2017-2274) funded by the Ministerio de Ciencia, Innovación y Universidades.

References

1. Doležel J, Bartoš J, Voglmayr H, Greilhuber J (2003) Nuclear DNA content and genome size of trout and human. Cytometry A 51A:127–128
2. Ogur M, Erickson RO, Rosen GU, Sax KB, Holden C (1951) Nucleic acids in relation to cell division in *Lilium longiflorum*. Exp Cell Res 2:73–89
3. Leitch IJ, Johnston E, Pellicer J, Hidalgo O, Bennett MD. (2019) Plant DNA C-values database (release 7.1, April 2019). https://cvalues.science.kew.org/
4. Pellicer J, Leitch IJ (2019) The plant DNA C-values database (release 7.1): an updated online repository of plant genome size data for comparative studies. New Phytologist 226(2): 301–305. https://doi.org/10.1111/nph.16261
5. Loureiro J, Travnicek P, Rauchova J, Urfus T, Vit P, Stech M, Castro S, Suda J (2010) The use of flow cytometry in the biosystematics, ecology and population biology of homoploid plants. Preslia 82:3–21
6. Suda J, Kron P, Husband BC, Trávnícek P (2007) Flow cytometry and ploidy: applications in plant systematics, ecology and evolutionary biology. In: Doležel J, Greilhuber J, Suda J (eds) Flow cytometry with plants cells. Wiley-VCH, Weinheim, pp 103–130
7. Dodsworth S, Leitch AR, Leitch IJ (2015) Genome size diversity in angiosperms and its influence on gene space. Curr Opin Genet Dev 35:73–78
8. Greilhuber J, Leitch IJ (2013) Genome size and the phenotype. In: Leitch IJ, Greilhuber J, Doležel J, Wendel JF (eds) Plant genome diversity, vol 2, physical structure, behaviour and evolution of plant genomes. Springer-Verlag, Wien, pp 323–344

9. Guignard MS, Crawley MJ, Kovalenko D, Nichols RA, Trimmer M, Leitch AR, Leitch IJ (2019) Interactions between plant genome size, nutrients and herbivory by rabbits, molluscs and insects on a temperate grassland. Proc R Soc B Biol Sci 286:20182619
10. Pellicer J, Hidalgo O, Dodsworth S, Leitch IJ (2018) Genome size diversity and its impact on the evolution of land plants. Genes 9:88
11. Sliwinska E (2018) Flow cytometry – a modern method for exploring genome size and nuclear DNA synthesis in horticultural and medicinal plant species. Folia Hort 30:103
12. Kreiner JM, Kron P, Husband BC (2017) Evolutionary dynamics of unreduced gametes. Trends Genet 33:583–593
13. Kron P, Husband BC (2012) Using flow cytometry to estimate pollen DNA content: improved methodology and applications. Ann Bot 110:1067–1078
14. Farhat P, Hidalgo O, Robert T, Siljak-Yakovlev S, Leitch IJ, Adams RP, Bou D-KM (2019) Polyploidy in the conifer genus Juniperus: an unexpectedly high rate. Front Plant Sci 10:676
15. Guignard MS, Nichols RA, Knell RJ, Macdonald A, Romila C-A, Trimmer M, Leitch IJ, Leitch AR (2016) Genome size and ploidy influence angiosperm species' biomass under nitrogen and phosphorus limitation. New Phytol 210:1195–1206
16. Doležel J, Greilhuber J, Suda J (2007) Estimation of nuclear DNA content in plants using flow cytometry. Nat Protoc 2:2233–2244
17. Suda J, Krahulcova A, Travnicek P, Krahulec F (2006) Ploidy level versus DNA ploidy level: an appeal for consistent terminology. Taxon 55:447–450
18. Kolář F, Čertner M, Suda J, Schönswetter P, Husband BC (2017) Mixed-ploidy species: Progress and opportunities in polyploid research. Trends Plant Sci 22:1041–1055
19. Husband BC, Baldwin SJ, Suda J (2013) The incidence of polyploidy in natural plant populations: major patterns and evolutionary processes. In: Leitch IJ, Greilhuber J, Doležel J, Wendel JF (eds) Plant genome diversity, vol 2, physical structure, behaviour and evolution of plant genomes. Springer-Verlag, Wien, pp 255–276
20. Barkla BJ, Rhodes T, Tran K-NT, Wijesinghege C, Larkin JC, Dassanayake M (2018) Making epidermal bladder cells bigger: developmental- and salinity-induced endopolyploidy in a model halophyte. Plant Physiol 177:615–632
21. Leitch IJ, Dodsworth S (2017) Endopolyploidy in plants. eLS. https://doi.org/10.1002/9780470015902.a0020097.pub2
22. Matzk F, Meister A, Schubert I (2000) An efficient screen for reproductive pathways using mature seeds of monocots and dicots. Plant J 21:97–108
23. Dobeš C, Lückl A, Hülber K, Paule J (2013) Prospects and limits of the flow cytometric seed screen--insights from Potentilla sensu lato (Potentilleae, Rosaceae). New Phytol 198:605–616
24. Hojsgaard D, Hörandl E (2019) The rise of apomixis in natural plant populations. Front Plant Sci 10:358
25. Schinkel CCF, Kirchheimer B, Dellinger AS, Klatt S, Winkler M, Dullinger S, Hörandl E (2016) Correlations of polyploidy and apomixis with elevation and associated environmental gradients in an alpine plant. AoB PLANTS 8:plw064
26. Noirot M, Barre P, Louarn J, Duperray C, Hamon S (2002) Consequences of stoichiometric error on nuclear DNA content evaluation in Coffea liberica var. dewevrei using DAPI and propidium iodide. Ann Bot 89:385–389
27. Noirot M, Barre P, Louarn J, Duperray C, Hamon S (2000) Nucleus-cytosol interactions - a source of stoichiometric error in flow cytometric estimation of nuclear DNA content in plants. Ann Bot 86:309–316
28. Noirot M, Barre P, Duperray C, Louarn J, Hamon S (2003) Effects of caffeine and chlorogenic acid on propidium iodide accessibility to DNA: consequences on genome size evaluation in coffee tree. Ann Bot 92:259–264
29. Noirot M, Barre P, Duperray C, Hamon S, De Kochko A (2005) Investigation on the causes of stoichiometric error in genome size estimation using heat experiments. Consequences on data interpretation. Ann Bot 95:111–118
30. Price HJ, Hodnett G, Johnston JS (2000) Sunflower (*Helianthus annuus*) leaves contain compounds that reduce nuclear propidium iodide fluorescence. Ann Bot 86:929–934
31. Bennett MD, Price HJ, Johnston JS (2008) Anthocyanin inhibits propidium iodide DNA fluorescence in *Euphorbia pulcherrima*: implications for genome size variation and flow cytometry. Ann Bot 101:777–790
32. Loureiro J, Rodriguez E, Doležel J, Santos C (2006) Flow cytometric and microscopic analysis of the effect of tannic acid on plant nuclei and estimation of DNA content. Ann Bot 98:515–527

33. Cires E, Cuesta C, Fernández MA, Nava HS, Vázquez VM, Fernández JA (2011) Isolation of plant nuclei suitable for flow cytometry from species with extremely mucilaginous compounds: an example in the genus Viola L. (Violaceae). An Jard Bot Madr 68:139–154
34. Šmarda P, Knápek O, Březinová A, Horová L, Grulich V, Danihelka J, Veselý P, Šmerda J, Rotreklová O, Bures P (2019) Genome sizes and genomic guanine+cytosine (GC) contents of the Czech vascular flora with new estimates for 1700 species. Preslia 91:117–142
35. Fernandez P, Gálvez F, Garcia S, Gras A, Hidalgo O, Pellicer J, Siljak-Yakovlev S, Vitales D, Vallès J. (2018) GSAD genome size in Asteraceae database (release 3.0, July 2019). http://asteraceaegenomesize.com/
36. Garnatje T, Canela MÁ, Garcia S, Hidalgo O, Pellicer J, Sánchez-Jiménez I, Siljak-Yakovlev S, Vitales D, Vallès J (2011) GSAD: a genome size in the Asteraceae database. Cytometry A 79A:401–404
37. Greilhuber J, Doležel J, Lysak MA, Bennett MD (2005) The origin, evolution and proposed stabilization of the terms 'Genome size' and 'C-value' to describe nuclear DNA contents. Ann Bot 95:255–260
38. Wood TE, Takebayashi N, Barker MS, Mayrose I, Greenspoon PB, Rieseberg LH (2009) The frequency of polyploid speciation in vascular plants. Proc Natl Acad Sci U S A 106:13875–13879
39. Barker MS, Arrigo N, Baniaga AE, Li Z, Levin DA (2016) On the relative abundance of autopolyploids and allopolyploids. New Phytol 210:391–398
40. Trávníček P, Ponert J, Urfus T, Jersáková J, Vrána J, Hřibová E, Doležel J, Suda J (2015) Challenges of flow-cytometric estimation of nuclear genome size in orchids, a plant group with both whole-genome and progressively partial endoreplication. Cytometry A 87:958–966
41. Leitch IJ, Kahandawala I, Suda J, Hanson L, Ingrouille MJ, Chase MW, Fay MF (2009) Genome size diversity in orchids - consequences and evolution. Ann Bot 104:469–481
42. Pellicer J, Kelly LJ, Leitch IJ, Zomlefer WB, Fay MF (2014) A universe of dwarfs and giants: genome size and chromosome evolution in the monocot family Melanthiaceae. New Phytol 201:1484–1497
43. Torrell M, Vallès J (2001) Genome size in 21 Artemisia L. species (Asteraceae, anthemideae): systematic, evolutionary, and ecological implications. Genome 44:231–238
44. Garcia S, Sanz M, Garnatje T, Kreitschitz A, McArthur ED, Vallès J (2004) Variation of DNA amount in 47 populations of the subtribe Artemisiinae and related taxa (Asteraceae, anthemideae): karyological, ecological, and systematic implications. Genome 47:1004–1014
45. Leitch IJ, Bennett MD (2004) Genome downsizing in polyploid plants. Biol J Linn Soc 82:651–663
46. Lysák MA, Lexer C (2006) Towards the era of comparative evolutionary genomics in Brassicaceae. Plant Syst Evol 259:175–198
47. Poggio L, Burghardt AD, Hunziker JH (1989) Nuclear DNA variation in diploid and polyploid taxa of Larrea (Zygophyllaceae). Heredity 63:321–328
48. Sliwinska E, Pisarczyk I, Pawlik A, Galbraith DW (2009) Measuring genome size of desert plants using dry seeds. Botany-Botanique 87:127–135
49. Sliwinska E, Zielinska E, Jedrzejczyk I (2005) Are seeds suitable for flow cytometric estimation of plant genome size? Cytometry A 64A:72–79
50. Wang N, McAllister HA, Bartlett PR, Buggs RJA (2016) Molecular phylogeny and genome size evolution of the genus Betula (Betulaceae). Ann Bot 117:1023–1035
51. Anamthawat-Jónsson K, Thórsson ÆT, Temsch EM, Greilhuber J (2010) Icelandic birch polyploids — the case of a perfect fit in genome size. J Bot 2010:347254, 9 pages
52. Suda J, Travnicek P (2006) Reliable DNA ploidy determination in dehydrated tissues of vascular plants by DAPI flow cytometry - new prospects for plant research. Cytometry A 69A:273–280
53. Viruel J, Conejero M, Hidalgo O, Pokorny L, Powell RF, Forest F, Kantar MB, Soto Gomez M, Graham SW, Gravendeel B, Wilkin P, Leitch IJ (2019) A target capture-based method to estimate ploidy from herbarium specimens. Front Plant Sci 10:937
54. Kolář F, Lučanová M, Těšitel J, Loureiro J, Suda J (2012) Glycerol-treated nuclear suspensions - an efficient preservation method for flow cytometric analysis of plant samples. Chromosom Res 20:303–315
55. Clarindo WR, Carvalho RC (2011) Flow cytometric analysis using SYBR green I for genome size estimation in coffee. Acta Histochem 113:221–225
56. Loureiro J, Rodriguez E, Doležel J, Santos C (2007) Two new nuclear isolation buffers for plant DNA flow cytometry: a test with 37 species. Ann Bot 100:875–888
57. Otto F (1992) Preparation and staining of cells for high-resolution DNA analysis. In:

Radbruch A (ed) Flow cytometry and cell sorting. Springer-Verlag, Berlin, pp 101–104

58. Barow M, Meister A (2003) Endopolyploidy in seed plants is differently correlated to systematics, organ, life strategy and genome size. Plant Cell Environ 26:571–584
59. Doležel J, Binarova P, Lucretti S (1989) Analysis of nuclear DNA content in plant cells by flow cytometry. Biol Plant 31:113–120
60. Pfosser M, Amon A, Lelley T, Heberlebors E (1995) Evaluation of sensitivity of flow-cytometry in detecting aneuploidy in wheat using disomic and ditelosomic wheat-rye addition lines. Cytometry 21:387–393
61. Galbraith DW, Harkins KR, Maddox JM, Ayres NM, Sharma DP, Firoozabady E (1983) Rapid flow cytometric analysis of the cell cycle in intact plant tissues. Science 220:1049–1051
62. Arumuganathan K, Earle ED (1991) Nuclear DNA content of some important plant species. Plant Mol Biol Report 9:208–218
63. Bino RJ, Lanteri S, Verhoeven HA, Kraak HL (1993) Flow cytometric determination of nuclear replication stages in seed tissues. Ann Bot 72:181–187
64. de Laat AMM, Blaas J (1984) Flow cytometric characterization and sorting of plant chromosomes. Theor Appl Genet 67:463–467
65. Ebihara A, Ishikawa H, Matsumoto S, Lin S-J, Iwatsuki K, Takamiya M, Watano Y, Ito M (2005) Nuclear DNA, chloroplast DNA, and ploidy analysis clarified biological complexity of the *Vandenboschia radicans* complex (Hymenophyllaceae) in Japan and adjacent areas. Am J Bot 92:1535–1547
66. Matzk F, Meister A, Brutovská R, Schubert I (2001) Reconstruction of reproductive diversity in *Hypericum perforatum* L. opens novel strategies to manage apomixis. Plant J 26:275–282
67. Bourge M, Brown SC, Siljak-Yakovlev S (2018) Flow cytometry as tool in plant sciences, with emphasis on genome size and ploidy level assessment. Genet Appl 2:1–12
68. Baranyi M, Greilhuber J (1995) Flow cytometric analysis of genome size variation in cultivated and wild *Pisum sativum* (Fabaceae). Plant Syst Evol 194:231–239
69. Mishiba KI, Ando T, Mii M, Watanabe H, Kokubun H, Hashimoto G, Marchesi E (2000) Nuclear DNA content as an index character discriminating taxa in the genus Petunia sensu Jussieu (Solanaceae). Ann Bot 85:665–673
70. Lee H-C, Lin T-Y (2005) Isolation of plant nuclei suitable for flow cytometry from recalcitrant tissue by use of a filtration column. Plant Mol Biol Report 23:53–58
71. Powell RF, Pulido Suarez L, Magee AR, Boatwright JS, Kapralov MV, Young AJ Genome size variation and endopolyploidy in the diverse succulent plant family Aizoaceae. Bot J Linn Soc. (in press)
72. Šmarda P, Bureš P (2010) Understanding intraspecific variation in genome size in plants. Preslia 82:41–61
73. Loureiro J, Rodriguez E, Doležel J, Santos C (2006) Comparison of four nuclear isolation buffers for plant DNA flow cytometry. Ann Bot 98:679–689
74. Greilhuber J, Temsch EM, Loureiro J (2007) Nuclear DNA content measurement. In: Doležel J, Greilhuber J, Suda J (eds) Flow cytometry with plant cells. Wiley-VCH, Weinheim, pp 67–102
75. Hörandl E, Dobes C, Suda J, Vit P, Urfus T, Temsch EM, Cosendai AC, Wagner J, Ladinig U (2011) Apomixis is not prevalent in subnival to nival plants of the European Alps. Ann Bot 108:381–390
76. Bennett MD, Smith JB (1991) Nuclear DNA amounts in angiosperms. Philos Trans R Soc Lond B Biol Sci 334:309–345
77. Doležel J, Sgorbati S, Lucretti S (1992) Comparison of three DNA fluorochromes for flow cytometric estimation of nuclear DNA content in plants. Physiol Plant 85:625–631
78. Doležel J, Dolezelova M, Novak FJ (1994) Flow cytometric estimation of nuclear DNA amount in diploid bananas (Musa acuminata and *M. balbisiana*). Biol Plant 36:351–357
79. Marie D, Brown SC (1993) A cytometric exercise in plant DNA histograms, with 2C values for 70 species. Biol Cell 78:41–51
80. Obermayer R, Leitch IJ, Hanson L, Bennett MD (2002) Nuclear DNA C-values in 30 species double the familial representation in pteridophytes. Ann Bot 90:209–217
81. Lysák MA, Doležel J (1998) Estimation of nuclear DNA content in Sesleria (Poaceae). Caryologia 52:123–132
82. Doležel J, Greilhuber J, Lucretti S, Meister A, Lysák MA, Nardi L, Obermayer R (1998) Plant genome size estimation by flow cytometry: inter-laboratory comparison. Ann bot 82(Suppl. A):17–26
83. Clark J, Hidalgo O, Pellicer J, Liu H, Marquardt J, Robert Y, Christenhusz M, Zhang S, Gibby M, Leitch IJ, Schneiderrs H (2016) Genome evolution of ferns: evidence for relative stasis of genome size across the fern phylogeny. New Phytol 210:1072–1082

Chapter 18

Molecular Cytogenetics (Fluorescence In Situ Hybridization - FISH and Fluorochrome Banding): Resolving Species Relationships and Genome Organization

Sonja Siljak-Yakovlev, Fatima Pustahija, Vedrana Vičić-Bočkor, and Odile Robin

Abstract

Fluorochrome banding (chromomycin, Hoechst, and DAPI) and fluorescence in situ hybridization (FISH) are excellent molecular cytogenetic tools providing various possibilities in the study of chromosomal evolution and genome organization. The constitutive heterochromatin and rRNA genes are the most widely used FISH markers. The rDNA is organized into two distinct gene families (18S–5.8S–26S and 5S) whose number and location vary within the complex of closely related species. Therefore, they are widely used as chromosomal landmarks to provide valuable evidence concerning genome evolution at chromosomal levels.

Key words Chromomycin, *Crepis*, DAPI, Fluorescence in situ hybridization (FISH), Fluorochrome banding, Hoechst, *Pinus*, rRNA genes

1 Introduction

Molecular cytogenetics provide new possibilities in the study of chromosomal evolution and genome organization which also contribute to a better characterization of the karyotype. Fluorochrome banding and fluorescence in situ hybridization (FISH) are excellent tools for chromosome identification in studies of chromosome evolution and genome organization and also to reveal the relationships between different taxa. These molecular cytogenetic approaches have been widely used for karyotyping in many wild and cultivated plants, e.g., in *Arabidopsis thaliana* [1], *Medicago truncatula* [2], *Picea abies* and *P. omorika* [3], *Agropyron* [4], *Hordeum*, and *Triticum* [5], and for studying evolutionary relationships within many genera, e.g., *Hypochaeris* [6, 7], *Quercus* [8], *Lilium* [9], *Nicotiana* [10], *Pinus* [11, 12], *Juniperus* [13], *Reichardia* [14], and *Cheirolophus* [15].

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_18,

Before the development of fluorochrome banding, Giemsa C-banding has been used to reveal constitutive heterochromatin (highly repetitive DNA sequences which remain condensed during the whole cell cycle). This heterochromatin can be GC or AT rich, or "neutral." The most widely used base-specific fluorochrome chromomycin A3 is a fluorescent stain that binds strongly to GC-rich regions in DNA. DAPI (4′,6′-diamidino-2-phenylindole) or Hoechst (bisbenzimide H33258), on the other hand, is specific for AT-rich DNA. Comparative patterns of fluorochrome banding may be useful not only in identifying homologous chromosomes but also in revealing phylogenetic relationships among species [9, 16, 17].

In the case of two closely related species of the genus *Crepis* [*Crepis praemorsa* (L.) Tausch and *Crepis incarnata* Tauch.] with the same chromosome number and almost identical karyotypes, banding techniques revealed a high intrachromosomal differentiation between two species (Fig. 1). All constitutive heterochromatin in these two species, revealed after Giemsa C-banding, represents AT-rich DNA regions [18]. However, in *C. praemorsa*, heterochromatic regions are limited only to centromeres and nucleolar organizer region (NOR). In *C. incarnata*, this type of heterochromatin is abundant forming the large telomeric and intercalary bands. The AT-rich DNA regions are consequently GC poor and present low fluorescent intensity with appropriate fluorochrome (see negative bands on chromomycin stained chromosomes, Fig. 1d). Before these results, obtained by chromosome banding, in *Flora Europaea*, the *C. incarnata* has been considered only as subspecies [*C. praemorsa* subsp. *dinarica* (Beck) Hayek, synonym = *C. incarnata*] [19]. This and numerous other studies demonstrate the usefulness of fluorochrome banding in resolving systematic and phylogenetic relationships between closely related taxonomic entities and point out the high implication of heterochromatin during differentiation of *C. incarnata* (endemic mountain species from Alps) from *C. praemorsa* (ancestral species from Euro-Asiatic plains with a large geographical repartition). In addition to this study, the reproductive isolation has been also detected which confirmed the specific level of these two taxa [20, 21].

Fluorescence in situ hybridization is a 35-year-old molecular cytogenetic tool that has developed continuously. Schwarzacher and Heslop-Harrison [22] provided the most accurately documented data and protocols concerning FISH techniques in plants.

In eukaryotes, rRNA genes present the most widely used FISH markers. They are organized into two distinct gene families. The first family of rRNA genes, encoding for 18S, 5.8S, and 26S ribosomal RNA (35S rDNA), occurs as tandem arrays at one or several specific regions on chromosomes. The 35S rDNA loci consist of tandemly repeated units of the 18S, 5.8S and 26S rDNA, internal transcribed (ITS1 and ITS2) sequences, and intergenic spacers

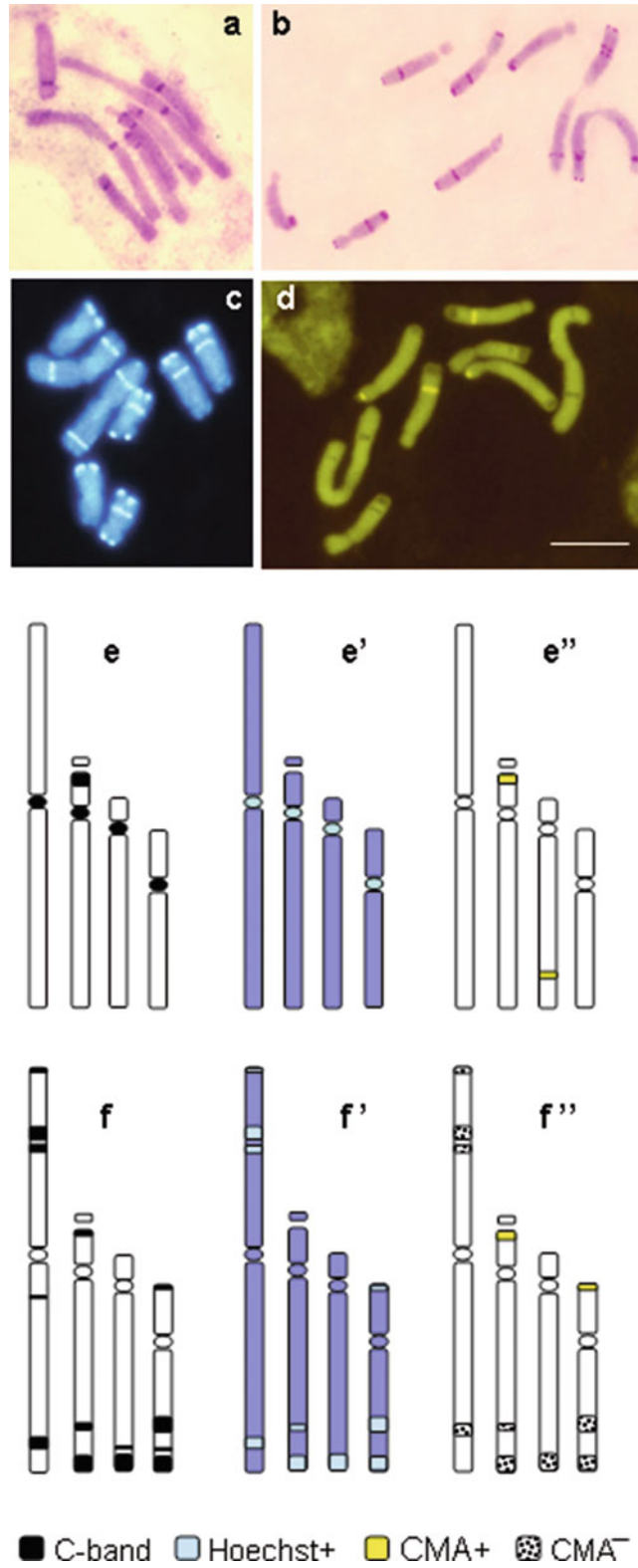


Fig. 1 *Crepis praemorsa*: Giemsa C-banding (**a**), idiogram showing C-bands (**e**), Hoechst (**e′**), and CMA bandings (**e″**). *C. incarnata*: C-banding (**b** and **f**), Hoechst (**c** and **f′**), and CMA bandings (**d** and **f″**). Bar = 10 μm

(IGS). These genes are highly conserved, and the chromosomal segment harboring them is known as a nucleolar organizer region (NOR). The second family is presented by 5S rRNA genes, also highly conserved and widely used as molecular cytogenetic markers.

Due to their high copy number, both families of rRNA genes are easily and reproducibly detectable on chromosomes and constitute suitable landmarks for chromosome identification.

The number and location of rDNA vary within the complex of closely related species; therefore, it can be used as a chromosomal landmark to provide valuable evidence concerning genome evolution at chromosomal levels. The rDNAs can change rapidly both in copy number and chromosome distribution, and rDNA transposition or dispersion in plant genomes is frequently observed [23–27]. These rearrangements are generally in correlation with species differentiation and speciation, and FISH analysis of rDNA is a good tool to detect chromosome variations.

Since 2012 [28, 29], a database of plant rDNA is available (www.plantrdnadatabase.com). Some authors have recently published on the cytogenetic characteristics of rDNA in plants and on this database [30] which we recommend to all researchers working in the field of plant cytogenetics.

Recently, Waminal et al. [31] have developed an alternative approach for efficient karyotyping and genome evolutionary studies using the PLOP-FISH protocol, a simplified and rapid multiplex FISH analysis using pre-labeled oligonucleotide probes (PLOPs) for simultaneous visualization of different target loci with reducing the cost and time for FISH hybridization. The authors analyzed only a few species using probes based on highly conserved regions in plants, animals, and fungi. In our future researches, we will test the proposed protocol and compare the data obtained with previously analyzed plant species with conventional FISH methods.

The following example demonstrates the use of fluorochrome banding and FISH to detect small structural chromosomal differences even at the level of intraspecific taxonomic categories.

The genus *Pinus*, and Pinaceae family in general, is characterized by the same chromosome number ($2n = 24$) and conserved karyotypes with all metacentric chromosome pairs, except one submetacentric. In such cases, when karyotyping based on morphometric analysis is difficult, the comparative patterns of fluorochrome banding and FISH experiment may be useful not only in identifying homologous chromosomes but also in revealing phylogenetic relationships among taxa.

Thus, in the study of *Pinus nigra* J.F. Arnold subspecies [*Pinus nigra* subsp. *laricio* Maire and *Pinus nigra* subsp. *dalmatica* (Vis.) Franco], molecular cytogenetic tools revealed an unsuspected difference in heterochromatin and rDNA organization [11]. DAPI staining after FISH displayed a high number of signals (Figs. 2 and

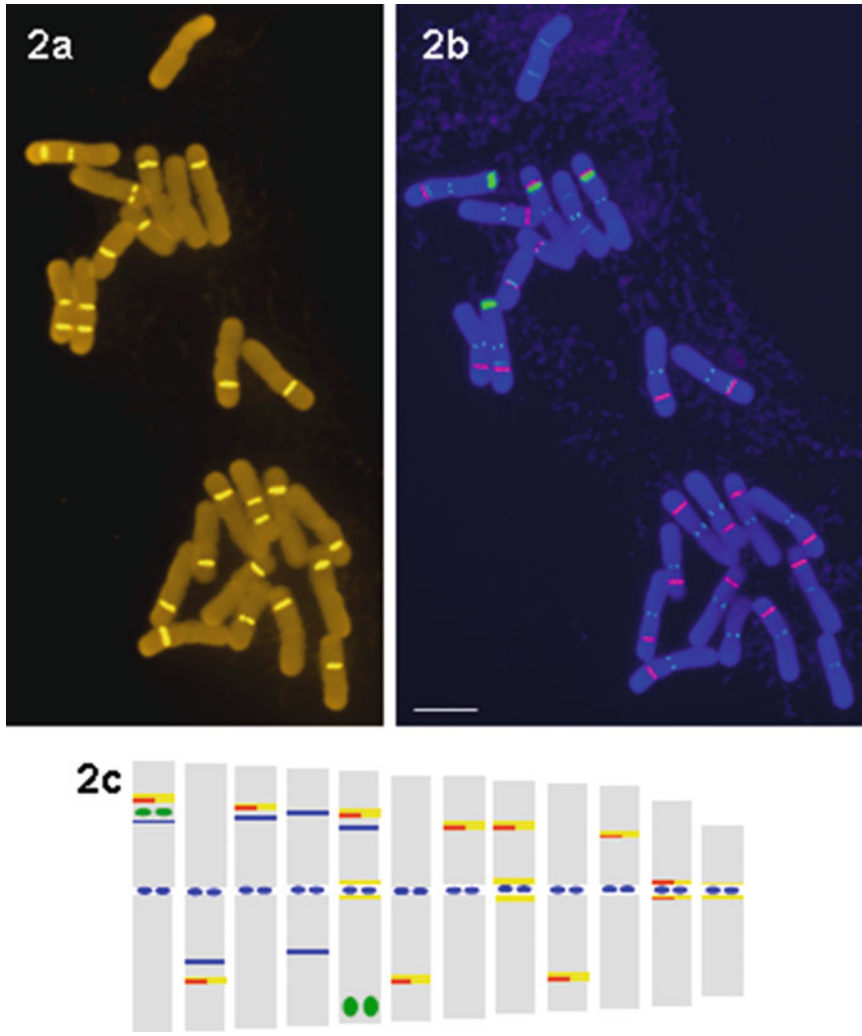


Fig. 2 *Pinus nigra* subsp. *laricio*: CMA-chromomycin banding (**a**); FISH (**b**); corresponding haploid idiogram showing 10 35S and 2 5S rDNA loci, 13 CMA, and 18 DAPI bands (**c**). Bar = 10 μm

3). The number of CMA bands was 26 in ssp. *laricio* (Fig. 2a) and 24 in ssp. *dalmatica* (Fig. 3a) with slightly different positions. Since all the centromeres were DAPI positive, the differences were reflected by the number of intercalary DAPI bands. They were distributed either on one or both chromosome arms. Two DAPI patterns were evident: the first with a lower number of signals (36 in ssp. *laricio*) and the second with a higher number of bands (64 in ssp. *dalmatica)* (Fig. 2b and 3b, respectively). The number and position of 5S rRNA genes were the same, but the number of 18S–26S rDNA loci was 10 for ssp. *laricio* and 8 for ssp. *dalmatica* (Figs. 2b, c and 3b, c, respectively).

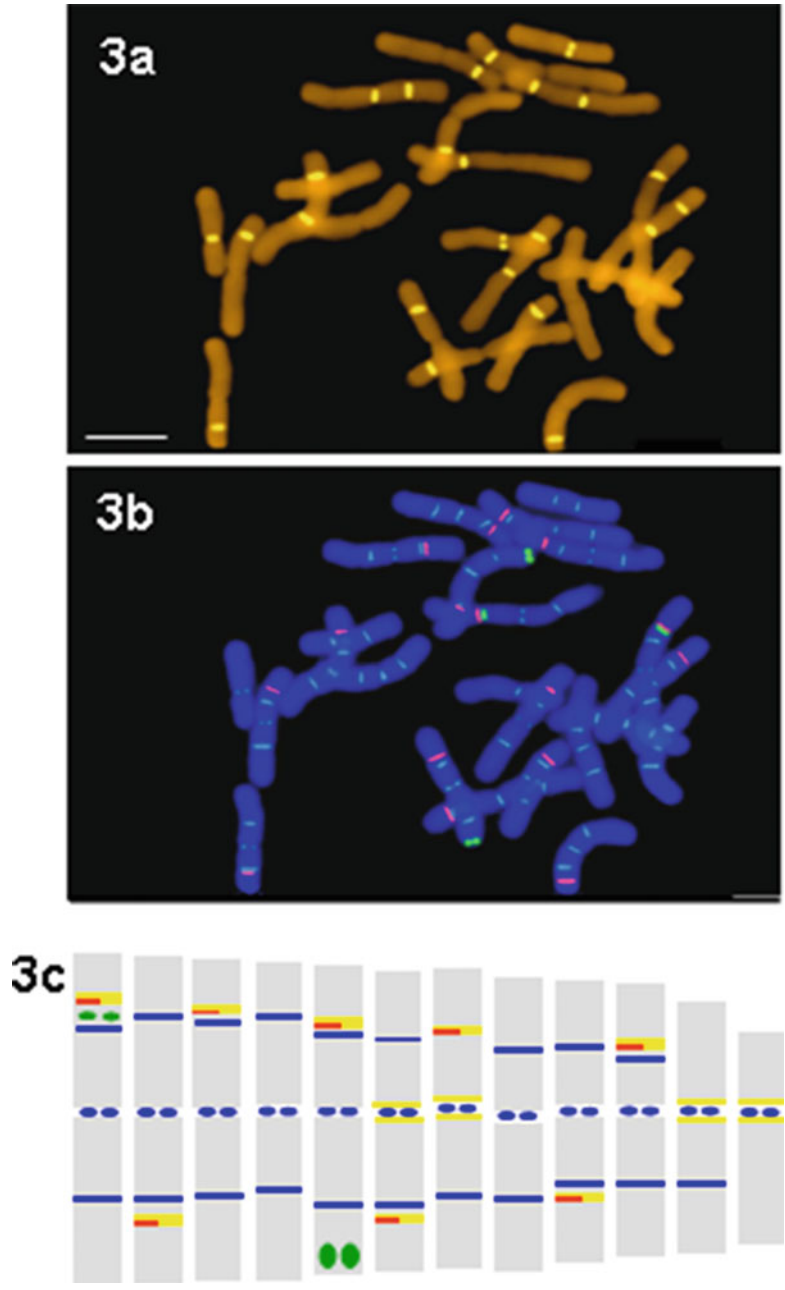


Fig. 3 *Pinus nigra* subsp. *dalmatica*: CMA (**a**); FISH (**b**); haploid idiogram showing 8 35S and 2 5S rDNA loci, 12 CMA, and 32 DAPI bands (**c**). Bar = 10 μm

Therefore, the molecular cytogenetic analysis can unequivocally reveal subtle chromosomal changes even between low taxonomic categories, and by combining it with phytogeography and ecology of representatives of a complex of related species, it

becomes possible to determine processes of species differentiation and evolution and the phylogenetic relationships between taxa.

2 Materials

Use sterile ultrapure water and analytical grade reagents for preparing solutions. Prepare and store all reagents at room temperature, unless indicated otherwise. For long-term storage, the stock solutions can be aliquoted and stored at ≤ -20 °C. For short-term storage, the solutions can be kept at 2–6 °C, protected from light if necessary. Care should be taken in handling and disposal of dyes and all waste materials, according to applicable local regulations.

2.1 Pretreatments and Root Tip Fixations

1. 0.05% (m/v) aqueous colchicine solution: dissolve 0.05 g colchicine in 100 mL water.
2. 0.002 M 8-hydroxyquinoline: dissolve 0.029 g 8-hydroxyquinoline in 100 mL water (*see* **Note 1**).
3. Carnoy I: freshly prepared 3:1 (v/v) ethanol:glacial acetic acid (*see* **Note 2**).
4. Carnoy II: freshly prepared 6:3:1 (v/v) ethanol:chloroform: glacial acetic acid (*see* **Note 3**).

2.2 Buffers

1. 0.01 M Citrate buffer, pH 4.6: solution A: 0.1 M citric acid. Solution B: 0.1 M trisodium citrate, pH 4.6. Mix 25.5 mL solution A and 24.5 mL solution B; adjust volume to 100 mL with ddH_2O. Store at −20 °C.
2. 0.05 M Citrate buffer, pH 4.6: solution A: 0.5 M citric acid. Solution B: 0.5 M trisodium citrate, pH 4.6. Mix 25.5 mL solution A and 24.5 mL solution B; adjust volume to 100 mL with ddH_2O. Store at −20 °C.
3. McIlvaine buffer, pH 5.5: solution A: 0.l M citric acid. Solution B: 0.2 M dibasic sodium phosphate. Mix 21.6 mL of A and 28.4 mL of B; adjust volume to 200 mL with ddH_2O. Store at −20 °C.
4. McIlvaine buffer, pH 7.0: solution A: 0.l M citric acid. Solution B: 0.2 M dibasic sodium phosphate. Mix 18 mL of A and 82 mL of B, diluted to a total of 200 mL with ddH_2O. Store at −20 °C.
5. McIlvaine buffer, pH 7.0 + 5 mM Mg^{2+}: dilute 0.123 g of $MgSO_4 \cdot 7H_2O$ in 100 mL of McIlvaine buffer, pH 7.0 (*see* **Note 4**). Store at −20 °C.

2.3 Enzyme Mixture

1. Hydrolytic enzyme mixture: 4% cellulase "Onozuka" RS (Yakult Pharmaceutical Co.), 1% pectolyase Y-23 (Seishin Pharmaceutical Co.), and 4% hemicellulase (Sigma-Aldrich) in 0.05 M citrate buffer. Store mixture at −20 °C (*see* **Note 5**).

2.4 Fluorochrome Banding

1. Chromomycin A3 (CMA), working solution: dissolve 0.02 g of CMA in 100 mL of McIlvaine buffer, pH 7.0 + Mg^{2+}. Store at −20 °C, protected from light.
2. 0.05% (m/v) methyl green dissolved in pH 5.5 McIlvaine buffer. Store at 4 °C.
3. Hoechst 33258 [Ho; bisbenzimide H33258; 2-[2-(4-Hydroxyphenyl)-6-benzimidazolyl]-6-(1-methyl-4-piperazyl)-benzimidazoletrihydrochloride]: dissolve 1 mg Ho in 100 mL ddH_2O for stock solution and store at −20 °C, protected from light. Work solution: dilute 1 mL of Ho stock solution with 4 mL of McIlvaine buffer, pH 5.5.
4. DAPI (4′,6 diamidino-2-phenylindole): stock solution of 2 μg/mL in ddH_2O. Working solution at 0.1 μg/mL, aliquoted and stored at −20 °C, protected from light.
5. Antifade solution: Citifluor AF_2 (Agar Scientific Oxford Instruments, Stansted, UK) or manually prepared glycerol antifade solution (McIlvaine buffer, pH 7.0 + Mg^{2+}:glycerol = 1:1, v/v).

2.5 FISH (Fluorescence In Situ Hybridization)

1. SSC (20×) (saline-sodium citrate buffer): 3 M sodium chloride, 0.3 M sodium citrate tribasic dihydrate, adjusted to pH 7.0 with 1 M HCl, autoclaved, and stored at room temperature. For use in the hybridization mixture, store at −20 °C.
2. SSC (2×): dilute 100 mL of 20× SSC with 900 mL of ddH_2O.
3. SSC (0.1×): dilute 13 mL of 2× SSC with 237 mL of ddH_2O.
4. RNase A stock solution: dissolve 10 mg of RNase in 1 mL of 10 mM Tris–HCl, pH 8.0. Boil for 15 min and allow to cool. Store at −20 °C in aliquots. Prior to use, dilute 100× in 2× SSC.
5. 0.01 M HCl.
6. Pepsin stock solution: 0.1 mg/mL solution in 0.01 M HCl. Aliquot and store at −20 °C.
7. Proteinase K: 1 mg/mL stock solution in ddH_2O. Store at −20 °C. Prior to use, dilute 100× in 2× SSC.
8. Formamide, deionized.
9. Tween 20.
10. SSCT (4×): dilute 100 mL of 20× SSC with 400 mL of ddH_2O. Add 1 mL of Tween 20.
11. Dextran sulfate (DS): dissolve 50 g of DS in 100 mL of sterile ddH_2O. Store in aliquots at −20 °C.
12. Sodium dodecyl sulfate (SDS): dissolve 1 g of SDS in sterile 10 mL ddH_2O. Store in aliquots at −20 °C.

13. Salmon sperm DNA solution (SS): concentration 10.5 ± 0.5 mg/mL. Store in aliquots at −20 °C.
14. Hybridization buffer: prepare 50 μL for one slide: 50% formamide, 10% dextran sulfate, 0.6% sodium dodecyl sulfate, 1.5 μL salmon sperm, and 5 μL 20× SSC. Calculate the required amount of ddH_2O to the final volume depending on the amounts of 18S–26S and 5S DNA probes added.
15. Modified hybridization buffer: prepare 50 μL for one slide: 50% formamide, 10% dextran sulfate, 5 μL 20× SSC, and 50 mM NaH_2PO_4, pH = 7.0 (*see* **Note 6**). Calculate the required amount of ddH_2O to the final volume depending on the amounts of 18S–26S and 5S DNA probes added.
16. PCR labeling with digoxigenin-11-dUTP: the PCR mixture consists of 1× PCR buffer, 2 mM $MgCl_2$, 0.2 mM dNTP, 1 mM digoxigenin-11-dUTP, 0.2 mM M13 forward primer (universal), 0.2 mM M13 reverse primer (universal), 1 U of Taq polymerase (Promega), and 100 ng of plasmid pTa794, containing a 410 bp fragment of 5S rRNA gene and a spacer region of wheat as a template [32]. After the initial denaturation step at 95 °C for 5 min, 35 cycles of denaturation for 30 s at 94 °C, annealing for 30 s at 55 °C, and elongation for 30 s at 72 °C were done, followed by a final elongation step of 5 min at 72 °C. The obtained PCR products are checked by electrophoresis on 1% agarose gel to verify product length and digoxigenin incorporation. Store in aliquots at −20 °C.
17. Nick translation: 18S–26S rDNA probes are labeled with Cy3 by nick translation using Nick Translation Mix (Roche) according to manufacturer instructions. A plasmid containing a 2.4 kb fragment of 18S rRNA gene from *Cucurbita pepo* is used as a template [33]. Store in aliquots at −20 °C; protect from light.
18. Blocking buffer: dissolve 0.1 g of BSA (bovine serum albumin) in 2 mL 4× SSCT (*see* **Note 7**). Store at −20 °C.
19. Antibody buffer: dilute antibody stock solution (200 μg/mL) 1:75 with blocking buffer. For one slide, mix 49.3 μL of blocking buffer with 0.7 μL of anti-digoxigenin-fluorescein, Fab fragments (ADF) mixture (*see* **Note 8**).
20. Antifade mounting medium: use Vectashield mounting medium with DAPI (Vector Laboratories, Peterborough, UK).
21. DAPI (4′,6 diamidino-2-phenylindole): stock solution of 0.5 μg/μL in ddH_2O. Working solution at 0.1 mg/mL, aliquoted and stored at −20 °C, protected from light.

3 Methods

3.1 Pretreatment and Fixation of Root Tips

Carry out all procedures at room temperature unless otherwise specified.

1. Immerse root tips in colchicine solution for 3–6 h at room temperature (large chromosomes) or 8-hydroxyquinoline solution for 2–4 h at 16 °C (small chromosomes).
2. Fix root tips in Carnoy I or Carnoy II solution for 15–30 min at room temperature and leave in fixative for 24–48 h at 4 °C (*see* **Note 9**).

3.2 Preparation of Protoplasts

Following the technique of Geber and Schweizer [34] with minor modifications.

1. Thaw enzyme mixture at 37 °C and transfer it either to a watch glass in a Petri dish or to a 1.5 mL centrifuge tube (*see* **Note 10**).
2. Wash fixed root tips in 0.05 M citrate buffer for 10 min and then digest in the enzymatic mixture at 37 ° C for 10–60 min (depending on root size; *see* **Note 11**).
3. Transfer root tip meristems by pipette to a drop of 45% acetic acid on a clean slide. Place cover slip and apply gentle pressure to spread the chromosomes. Tapping with needle tweezers on top of the cover slip may improve chromosome spreading.

3.3 Cover Slip Removal

Following the technique of Conger and Fairchild [35] with minor modifications.

1. Rapidly freeze preparation below −70 °C using liquid nitrogen or CO_2 or by placing slide on dry ice or on a metal plate in a –80 °C freezer (*see* **Note 12**).
2. Remove cover slip quickly while frozen, using a razor blade, and rinse briefly in absolute ethanol.
3. Air-dry and store at room temperature for a couple of days until it is time to proceed to the next step (Subheadings 3.4, 3.5, 3.7, or 3.8).

3.4 Chromomycin Banding

Following the modified techniques of Schweiser [36] and Kondo and Hizume [37] and the technique of Siljak-Yakovlev et al. [3].

1. Prepare air-dried slide with cover slip removed as described in Subheading 3.3.
2. Thaw previously prepared and frozen McIlvaine buffers (pH 5.5; pH 7.0; pH 7.0 + Mg^{2+}) and CMA working solution.
3. Add a few drops of McIlvaine buffer, pH 7.0 + Mg^{2+} to the slide and incubate for 15 min. Gently shake off slide.

4. Apply 80 μL of CMA working solution onto the slide and gently cover with a plastic cover slip (cut from autoclavable waste bags) avoiding formation of air bubbles. Incubate for 60–90 min in the dark.
5. Carefully remove the plastic cover slip with tweezers and wash briefly with McIlvaine buffer, pH 7.0.
6. Counterstain with methyl green for 7 min in the dark.
7. Wash slide briefly with McIlvaine buffer, pH 5.5.
8. Mount preparation in glycerol antifade solution.
9. Store slide in the dark. For long-term conservation, store at 4 °C.
10. Observe under an epifluorescence microscope with appropriate filters.

3.5 Hoechst Banding

Following the techniques of Martin and Hesemann [38].

1. Prepare air-dried slide with cover slip removed as described in Subheading 3.3.
2. Thaw McIlvaine buffer, pH 5.5, and Ho work solution.
3. Rehydrate slide by incubating successively in 70, 50, and 30% ethanol series and in ddH_2O for 5 min.
4. Add a few drops of McIlvaine buffer, pH 5.5, to the slide and incubate for 10 min.
5. Gently shake off slide and apply 80–100 μL of Ho working solution to the slide for 2 min. Cover with a plastic cover slip (cut from autoclavable waste bags), avoiding air bubbles, and protect from light.
6. Carefully remove the plastic cover slip with tweezers and wash briefly with McIlvaine buffer, pH 5.5.
7. Apply McIlvaine buffer, pH 5.5, to the whole slide and incubate for 15 min.
8. Gently shake off slide. Add a few drops of ddH_2O to the slide and incubate for 15 min.
9. Shake the water off, dry the slide, and mount it in glycerol antifade solution.
10. Store slide in the dark. For long-term conservation, store at 4 °C.
11. Observe under an epifluorescence microscope with appropriate filters.

3.6 Destaining Slides After Fluorochrome Bandings

1. Immerse slide in Carnoy I in staining dish until cover slip floats off.
2. Successively dehydrate slide in ice-cold ethanol series (70, 90, and 100%) for 5 min each (*see* **Note 13**).
3. Dry slide for a couple of days in a vertical position in a closed plastic box to prevent accumulation of dust.

3.7 FISH

Following the technique of Heslop-Harrison et al. [39] with minor modifications by Siljak-Yakovlev et al. [3].

3.7.1 Day One

1. Prepare a humid chamber using a plastic box with moistened paper tissues in the bottom. Warm up to 37 °C.
2. Add 200 μL of RNase A working solution to each slide, cover with a plastic cover slip, and incubate in a humid chamber at 37 °C for 1 h.
3. Carefully remove plastic cover slip with tweezers and wash slide in a Coplin jar in 2× SSC twice for 5 min.
4. Briefly rinse slide in a 0.01 M HCl solution.
5. Incubate slide with 80–100 μL of pepsin working solution for 10–15 min at 37 °C (*see* **Note 14**).
6. Rinse slide in deionized H_2O for 2 min.
7. Wash slide in 2× SSC two times for 5 min.
8. Facultative step: denaturation in 50 or 70% (for gymnosperms) formamide, 2 min at 70 °C (*see* **Note 15**). Rinse slide in 2× SSC for 5 min.
9. Dehydrate slide in an ethanol series: 70, 90, and 100% (−20 °C); 3 min each.
10. Air-dry slide for 1–2 h.
11. Add 0.5–2 μL of 18S–26S DNA probe (40 ng/μL) and/or 0.5–2 μL of 5S DNA probe (50 ng/μL) to hybridization buffer to obtain a hybridization mixture, 50 μL/slide (*see* **Note 16**).
12. Denature the probe by incubating the hybridization mixture (in an Eppendorf tube) in a water bath (or a heat block) at 72 °C for 10 min, and transfer immediately on ice for a minimum of 5 min (*see* **Note 17**).
13. Add 50 μL of hybridization mixture to slide and cover with a plastic cover slip. Place slide in a plastic box and incubate in a water bath at 72 °C for 10 min (*see* **Note 18**).
14. Transfer the box to another water bath set at 55 °C for 5 min.
15. Place slide in a humid chamber and incubate overnight at 37 °C (*see* **Notes 19** and **20**).

3.7.2 Day Two

1. Preheat the buffers (0.1× SSC, 2× SSC, 4× SSCT) in a water bath at 42 °C.
2. Carefully remove the plastic cover slip with tweezers and immerse slide in a Coplin jar with 2× SSC buffer for 3 min at room temperature.
3. Wash slide twice in 2× SSC for 5 min at 42 °C.
4. Facultative step to reduce background: wash slide in 20% formamide at 42 °C, two times for 5 min.
5. Wash slide in 0.1× SSC for 5 min at 42 °C.
6. Wash slide in 2× SSC for 5 min at 42 °C.
7. Wash slide in 4× SSCT for 5 min at 42 °C (*see* **Note 21**).
8. Blocking: apply 100 µL of blocking buffer on slide, cover with plastic cover slip, and incubate for 5 min at room temperature, protected from light. Carefully remove plastic cover slip.
9. Antibody detection: apply 50 µL of antibody buffer on slide; cover with plastic cover slip, and incubate at 37 °C for 1 h in a preheated plastic humid chamber.
10. Carefully remove plastic cover slip and immerse slide in 4x SSCT buffer three times for 5 min.
11. Shake the buffer off, dry the slide, and counterstain with final antifade mounting medium with DAPI. Leave to stand for 5–10 min and remove surplus medium using paper tissue.
12. Store slide in the dark at 4 °C.
13. Observe under an epifluorescence microscope with appropriate filters.

3.8 Modified FISH Protocol

In this section, we present a modified and much shorter version of our standard FISH protocol, which we already used and verified for some genera (e.g., *Crepis*, *Iris*, *Narcissus*, *Quercus*, and *Triticum*).

3.8.1 Day One

1. Prepare a humid chamber by placing moistened paper tissues on the bottom of a plastic box and preheat to 37 °C.
2. Add 200 µL of RNase A working solution on each slide, cover with a plastic cover slip, and incubate in a humid chamber at 37 °C for 1 h.
3. Immerse slide in a Coplin jar with 2× SSC buffer and wash twice for 5 min. The plastic cover slip will float off the slide during the first wash. Remove it carefully.
4. Add 50 µL of proteinase K working solution to the slide and incubate for 15 min at 37 °C (*see* **Note 14**).
5. Wash slide in 2× SSC for 5 min.
6. Dehydrate slide in an ethanol series: 70, 90, and 100% (−20 °C); 3 min each.

7. Air dry slide for 1–2 h.
8. Add 50 μL of modified hybridization mixture per slide and cover with a cover slip. Place the slide in a plastic box and incubate in a water bath set at 85 °C for 6 min (*see* **Note 18**).
9. Transfer the slide to a humid chamber and incubate overnight (16–20 h) at 37 °C.

3.8.2 Day Two

1. Preheat the buffers (0.1× SSC, 2× SSC, 4× SSCT) in a bath at 42 °C.
2. Immerse slide in a Coplin jar with 2× SSC buffer and wash for 5 min at room temperature.
3. Wash slide twice in 2× SSC for 5 min at 42 °C.
4. Wash slide in 0.1× SSC for 5 min at 42 °C.
5. Wash slide in 2× SSC for 5 min at 42 °C.
6. Wash slide in 4× SSCT for 5 min at 42 °C.
7. Wash slide 5 min in 4× SSCT at room temperature.
8. Blocking: apply 100 μL of blocking buffer on slide, cover with plastic cover slip, and incubate in a humid chamber for 30 min at room temperature, protected from light. Carefully remove plastic cover slip.
9. Antibody detection: apply 25 μL of antibody buffer per slide. Cover with a plastic cover slip and incubate in a humid chamber for 1 h at 37 °C.
10. Immerse slide in a Coplin jar with 4× SSCT buffer and wash twice for 5 min at room temperature. Carefully remove plastic cover slip.
11. Gently shake off excess buffer and counterstain with antifade mounting medium with DAPI. Remove surplus of medium using paper tissue, after 5–10 min. Alternatively, counterstain slide with 0.2 μg/mL DAPI in ddH_2O for 8 min. After a brief wash in 2× SSC, apply the antifade solution and cover with a cover glass. Remove excess medium using a paper tissue.
12. Place slide in a dark place, at 4 °C.
13. Observe under an epifluorescence microscope with appropriate filters.

3.9 Destaining Slides After FISH

1. Immerse slide in 2× SSC in a staining dish until cover slip floats off of the slide.
2. Dehydrate slide in ice-cold ethanol series (70, 90, and 100%) for 5 min (*see* **Note 13**).
3. Dry the slide in vertical position for a couple of days in a closed plastic box to avoid dust.
4. Restart new FISH experiment on the same slide from step 9 (standard protocol) or 6 (modified protocol) on Day One.

4 Notes

1. Store at 4 °C in a dark glass bottle, not longer than 2 months.
2. It is necessary to use fresh solutions to minimize ester formation, stop mitosis, and preserve chromosome structure integrity.
3. This solution is recommended for oily and waxy tissues to increase the penetration ability of the fixative.
4. It is possible to use $MgCl_2$ instead of $MgSO_4$: add 0.1017 g of $MgCl_2 \cdot 6H_2O$.
5. Proposed enzyme composition and concentrations may require modification for different plant species.
6. Hybridization buffer without probes and water can be prepared in excess volume and stored at −20 °C.
7. Put powder in the buffer and keep at 37 °C for a couple of minutes (without shaking) for easier and faster dissolution.
8. Detection step is not needed if FISH is done with directly labelled probes.
9. For long-term preservation, keep material in the Carnoy fixative (4 °C) for a few days and then transfer it to 70% ethanol fixative and store at 4 °C or –20 °C.
10. In case of large chromosomes and low number of available root tips, avoid centrifugation protocol. Enzyme mixtures can be reused several times in which case digestion time might need to be slightly increased after each round of use.
11. Exposure time of meristems to enzyme mixture depends on tissue thickness. It is necessary to verify the homogeneity and successive decomposition of meristems of analyzed species: material should be soft and break up easily for optimal time.
12. When using a freezer, preparations need to stay at −80 °C at least 24 h to avoid the loss of chromosome during cover slip removal.
13. Store alcohol solutions at −20 °C.
14. Incubation in pepsin and proteinase K should be prolonged in case of larger amounts of cytoplasm on the slide.
15. This step is recommended to achieve better denaturation and reduce background.
16. The probes should be added last, and the hybridization mixture should be homogenized by vortexing.
17. Rapid cooling of the hybridization mixture prevents reannealing of the probe.

18. The exact temperature and duration of treatment vary between species and should be experimentally determined if not already published.
19. It is important to prevent moisture loss by evaporation. However, too much moisture can lead to condensation on the slide, which can result in poorly hybridized slide.
20. Duration of hybridization should be prolonged for gymnosperms to up to 48 h.
21. During this step, thaw blocking buffer.

References

1. Murata M, Heslop-Harrison JS, Motoyoushi F (1997) Physical mapping of the 5S ribosomal RNA genes in *Arabidopsis thaliana* by multicolor fluorescence *in situ* hybridization with cosmid clones. Plant J 12(1):31–37
2. Cerbah M, Kevei Z, Siljak-Yakovlev S et al (1999) FISH chromosome mapping allowing karyotype analysis in *Medicago truncatula* lines Jemalong J5 and R 108-1. Mol Plant Microbe 12:947–950
3. Siljak-Yakovlev S, Cerbah M, Couland J et al (2002) Nuclear DNA content, base composition, heterochromatin and rDNA in *Picea omorika* and *Picea abies*. Theor Appl Genet 104:505–512
4. Zhao Y, Xie J, Dou Q et al (2017) Diversification of the P genome among *Agropyron* Gaertn. (Poaceae) species detected by FISH. Comp Cytogenet 11(3):495–509
5. Rey MD, Moore G, Martín C (2018) Identification and comparison of individual chromosomes of three accessions of *Hordeum chilense*, *Hordeum vulgare*, and *Triticum aestivum* by FISH. Genome 61(6):387–396
6. Cerbah M, Coulaud J, Siljak-Yakovlev S (1998) rDNA organization and evolutionary relationships in the genus *Hypochaeris* (Asteraceae). J Hered 89:312–318
7. Weiss-Schneeweiss H, Tremetsberger K, Schneeweiss GM et al (2008) Karyotype diversification and evolution in diploid and polyploid south American *Hypochaeris* (Asteraceae) inferred from rDNA localization and genetic fingerprint data. Ann Bot 101:909–918
8. Zoldos V, Papes D, Cerbah M et al (1999) Molecular-cytogenetic studies of ribosomal genes and heterochromatin reveal conserved genome organization among eleven *Quercus* species. Theor Appl Genet 99:969–977
9. Muratovic E, Robin O, Bogunic F et al (2010) Speciation of European lilies from *Liriotypus* section based on karyotype evolution. Taxon 59:165–175
10. Lim KY, Matyásek R, Lichtenstein CP et al (2000) Molecular cytogenetic analyses and phylogenetic studies in the *Nicotiana* section Tomentosae. Chromosoma 109:245–258
11. Bogunic F, Siljak-Yakovlev S, Muratovic E et al (2011) Different karyotype patterns among allopatric *Pinus nigra* (Pinaceae) populations revealed by molecular cytogenetics. Plant Biol 13:194–200
12. Bogunic F, Siljak-Yakovlev S, Muratovic E et al (2011) Molecular cytogenetics and flow cytometry reveal conserved genome organization in *Pinus mugo* and *P. uncinata*. Ann For Sci 68(1):179–187
13. Vallès J, Garnatje T, Robin O et al (2015) Molecular cytogenetic studies in western Mediterranean *Juniperus* (Cupressaceae): a constant model of GC-rich chromosomal regions and rDNA loci with evidences for paleopolyploidy. Tree Genet Genomes 11(3):43. https://doi.org/10.1007/s11295-015-0860-3
14. Siljak-Yakovlev S, Godelle B, Zoldos V et al (2017) Evolutionary implications of heterochromatin and rDNA in chromosome number and genome size changes during dysploidy: a case study in *Reichardia* genus. PLoS One 12(8):e0182318
15. Hidalgo O, Vitales D, Vallès J et al (2017) Cytogenetic insights into an oceanic island radiation: the dramatic evolution of pre-existing traits in *Cheirolophus* (Asteraceae: Cardueae: Centaureinae). Taxon 66(1): 146–157
16. Hizume M, Aria M, Tanaka A (1990) Chromosome banding in the genus *Pinus*. III. Fluorescent banding pattern of *P. luchuensis* and its relationships among the Japanese diploxylon pines. Bot Mag Tokyo 103:103–111
17. Bogunic F, Muratovic E, Siljak-Yakovlev S (2006) Chromosomal differentiation of *Pinus*

heldreichii and *Pinus nigra*. Ann For Sci 63:267–274

18. Godelle B, Cartier D, Marie D et al (1993) Heterochromatin study demonstrating the non-linearity of fluorometry useful for calculating genomic base composition. Cytometry 14:618–626
19. Sell PD (1976) Crepis. In: Tutin TG et al (eds) Flora Europaea 4. Cambridge University Press, Cambridge, pp 344–357
20. Siljak-Yakovlev S, Cartier D (1986) Heterochromatin patterns in some taxa of *Crepis praemorsa* complex. Caryologia 39:27–32
21. Cartier D, Siljak-Yakovlev S (1992) Cytogenetics study of the F1 hybrids between *Crepis dinarica* and *Crepis froelichiana*. Plant Syst Evol 182:29–34
22. Schwarzacher T, Heslop-Harrison P (2000) Practical *in situ* hybridization, 2nd edn. BIOS, Oxford, UK
23. Schubert I, Wobus U (1985) *In situ* hybridization confirms jumping nucleolus organizing regions in *Allium*. Chromosoma 92:143–148
24. Raina SN, Mukai Y (1999) Detection of a variable number of 18S-5.8S-26S and 5S ribosomal DNA loci by fluorescent *in situ* hybridization in diploid and tetraploid *Arachis* species. Genome 42:52–59
25. Raskina O, Belyayev A, Nevo E (2004) Quantum speciation in *Aegilops*: molecular cytogenetic evidence from rDNA cluster variability in natural populations. Proc Natl Acad Sci U S A 101:14818–14823
26. Datson PM, Murray BG (2006) Ribosomal DNA locus evolution in *Nemesia*: transposition rather than structural rearrangement as the key mechanism? Chromosom Res 14:845–857
27. Rosato M, Moreno-Saiz CJ, Galián AJ et al (2015) Evolutionary site-number changes of ribosomal DNA loci during speciation: complex scenarios of ancestral and more recent polyploid events. AoB Plants 7:plv135. https://doi.org/10.1093/aobpla/plv135
28. Garcia S, Garnatje T, Kovařík A (2012) Plant rDNA database: ribosomal DNA loci information goes online. Chromosoma 121(4):389–394. https://doi.org/10.1007/s00412-012-0368-7
29. Garcia S, Gálvez F, Gras A et al (2014) Plant rDNA database: update and new features. Database (Oxford) 2014:bau063. https://doi.org/10.1093/database/bau063
30. Garcia S, Kovařík A, Leitch AR et al (2017) Cytogenetic features of rRNA genes across land plants: analysis of the plant rDNA database. Plant J 89(5):1020–1030
31. Waminal NE, Pellerin RJ, Kim N-S et al (2018) Rapid and efficient FISH using pre-labeled oligomer probes. Sci Rep 8:8224. https://doi.org/10.1038/s41598-018-26667-z
32. Gerlach WI, Dyer TA (1980) Sequence organization of the repeating units in the nucleus of wheat which contain 5S rRNA genes. Nucleic Acids Res 8:4851–4865
33. Torres-Ruiz RA, Hemleben V (1994) Pattern and degree of methylation in ribosomal genes of *Cucurbita pepo* L. Plant Mol Biol 26:1167–1179
34. Geber G, Schweizer D (1988) Cytochemical heterochromatin differentiation in *Sinapis alba* (Cruciferae) using a simple air-drying technique for producing chromosome spreads. Plant Syst Evol 158:97–106
35. Conger AD, Fairchild LM (1953) A quick freeze method for making smear slide. Stain Technol 28:281–283
36. Schweiser D (1976) Reverse fluorescent chromosome banding with chromomycin and DAPI. Chromosoma 8:307–324
37. Kondo T, Hizume M (1982) Banding for the chromosomes of *Cryptomeria japonica* D. Don. J. Jpn For Soc 4:356–358
38. Martin J, Hesemann CU (1988) Evaluation of improved Giemsa C- and fluorochrome banding techniques in rye chromosome. Heredity 6:459–467
39. Heslop-Harrison LS, Schwarzacher T, Anamthawat-Jonsson K et al (1991) *In situ* hybridization with automated chromosome denaturation. Technique 3:109–116

Chapter 19

GISH: Resolving Interspecific and Intergeneric Hybrids

Nathalie Piperidis

Abstract

Genomic in situ hybridization (GISH) is an invaluable cytogenetic technique which enables the visualization of whole genomes in hybrids and polyploidy taxa. Total genomic DNA from one or two different species/genomes is used as a probe, labeled with a fluorochrome, and directly detected on mitotic chromosomes from root tip meristems. In sugarcane and sugarcane hybrids, we were able to characterize interspecific hybrids of two closely related species as well as intergeneric hybrids of two closely related genera.

Key words GISH, Fluorochrome, Interspecific, Intergeneric, Genome

1 Introduction

1.1 Genomic In Situ Hybridization (GISH)

GISH was derived from fluorescence in situ hybridization (FISH) techniques (Chapter 18) developed in the early 1980s by biomedical researchers, and it was eventually applied to plant chromosomes in the late 1980s. GISH was first demonstrated in synthetic *Hordeum chilense* x *Secale africanum* hybrids [1] and also used to track artificial introgression of chromosomes in wide crosses [2]. The challenges faced by plant chromosome researchers are mainly based on the fact that plants have cell walls, cytoplasmic debris, and more condensed chromosomes status that could affect the probe/DNA accessibility than in the mammalian cells. GISH is a powerful tool and can be used, for example, to distinguish the genome of one parent from the other by preferential labeling of the genome of either parent. It can also be used to detect alien chromosome(s) in addition lines or alien species in recipient parent, for example. GISH is extremely useful to identify parental chromosomes in interspecific or intergeneric hybrids, to test the origin of natural amphiploids, to track down the introgression of alien chromosomes, or to test the occurrence of exchange between the genomes involved [3, 4]. Multicolor GISH allows simultaneous discrimination of multiple genomes and identification of diploid progenitors

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2_19,

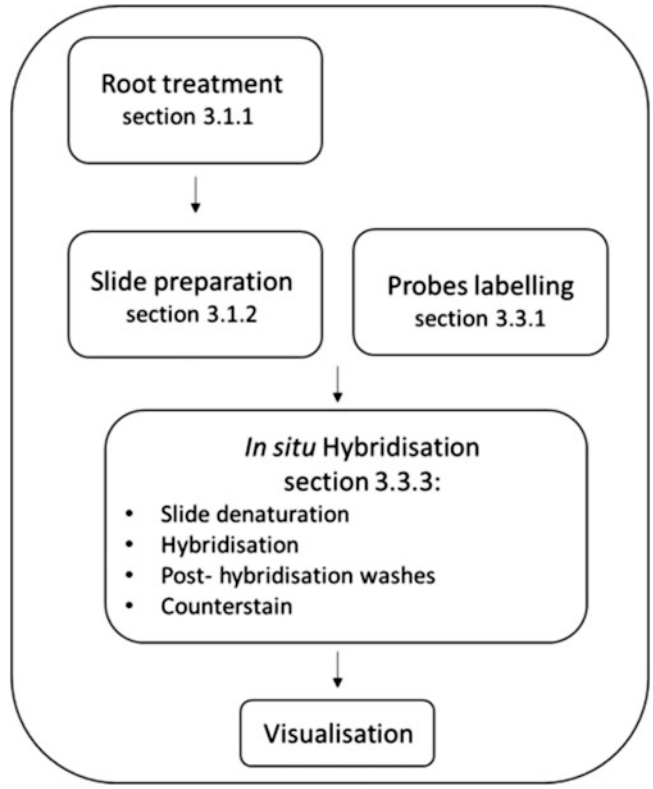


Fig. 1 Overview of the GISH procedure

in allopolyploids. GISH requires labelling of genomic DNA directly with a fluorochrome or with a hapten capable of indirect association with fluorochromes. The nucleic acid fluoro-probe(s) will then provide an assay through complementary pairing with nucleotides of the target DNA on a slide. Fluorochromes provide the ability to visualize in situ homologous regions to the probe within the cellular structure using a fluorescence microscope. Digital camera coupled to the microscope allows to capture permanent images of the fluorescent patterns on the chromosomes. Figure 1 represents the outline of the procedure.

1.2 Example of Application in Sugarcane and Sugarcane Hybrids

Although classical cytological studies in sugarcane [5] allowed a better understanding of the sugarcane genome, molecular cytogenetic methods not only lead to important breakthroughs revealing the level of the complexity of modern sugarcane cultivars but also unraveled the taxonomy of the *Saccharum* genus. Modern sugarcane cultivars are one of the most difficult species to work with on a genetic and molecular level. Sugarcane species are considered to have one of the most complicated genomes studied. Chromosome numbers were determined, uncovering highly polyploid and, frequently, aneuploid members in this genus [6]. The

genome of modern cultivars results in the hybridization of two species of *Saccharum*, the noble cane *Saccharum officinarum* and the wild species *Saccharum spontaneum* which was also revealed by GISH studies. In the past 15 years, molecular cytogenetic techniques have proven to be a very efficient tool to better understand this complex genome and revealed outcomes that classical molecular markers alone could not. These techniques proved to be particularly relevant to refine our understanding of the genome structure of sugarcane and its taxonomy [7, 8]. In our laboratory, we used GISH to characterize interspecific hybrids to taxonomic reclassification of atypical *S. officinarum* as well as intergeneric hybrids involving two different genomes and three different species: *Saccharum officinarum*, *Saccharum spontaneum* and *Erianthus arundinaceus*.

1.2.1 Interspecific Hybrid Between S. officinarum and S. spontaneum

Since the original classification of *Saccharum* species, taxonomy within the genus *Saccharum* has been controversial. *S. officinarum* is known to have 2n = 80 chromosomes; therefore, clones with more than 80 chromosomes should be classified as hybrids. However, Irvine [9] has debated this and suggested that clones that fit the botanical description for *S. officinarum* with more than 80 chromosomes should remain in this classification. GISH studies have contributed to understanding the taxonomic status and relationships of species and clones within the *Saccharum* genus. We used GISH to verify the taxonomic reclassification of atypical *S. officinarum* with more than 80 chromosomes revealed by flow cytometry [7]. GISH results of atypical *S. officinarum* clone Muntok Java are presented in Fig. 2a. Genomic DNA from *S. officinarum* was labeled in "red" with Alexa Fluor 594-5-dUTP, and genomic DNA from *S. spontaneum* was labeled in "green" with Alexa Fluor 488-5-dUTP. Both species are relatively closely related; therefore, *S. officinarum* chromosomes appear "orange," while *S. spontaneum* chromosomes appear "yellow-green" due to some level of cross-hybridization between the two genomes; recombined chromosomes from both species can also be visualized.

1.2.2 Intergeneric Hybrid Between S. officinarum and E. arundinaceus

For our intergeneric GISH characterization, *Erianthus arundinaceus* was labeled in "red" with Alexa Fluor 594-5-dUTP or Rhodamine-5-dUTP, while *S. officinarum* was labeled in "green" with Alexa Fluor 488-5-dUTP or Fluorescein-12-dUTP [8]. GISH of an F1 and backcross 1 (BC1) between the two genomes are presented in Fig. 2b, c. In these intergeneric hybrids, the *E. arundinaceus* chromosome are red and the *S. officinarum* chromosomes are green, as the two species are not as closely related than in the interspecific hybrids. The fluorochrome colors do not overlap as the genome has minimal cross-hybridization.

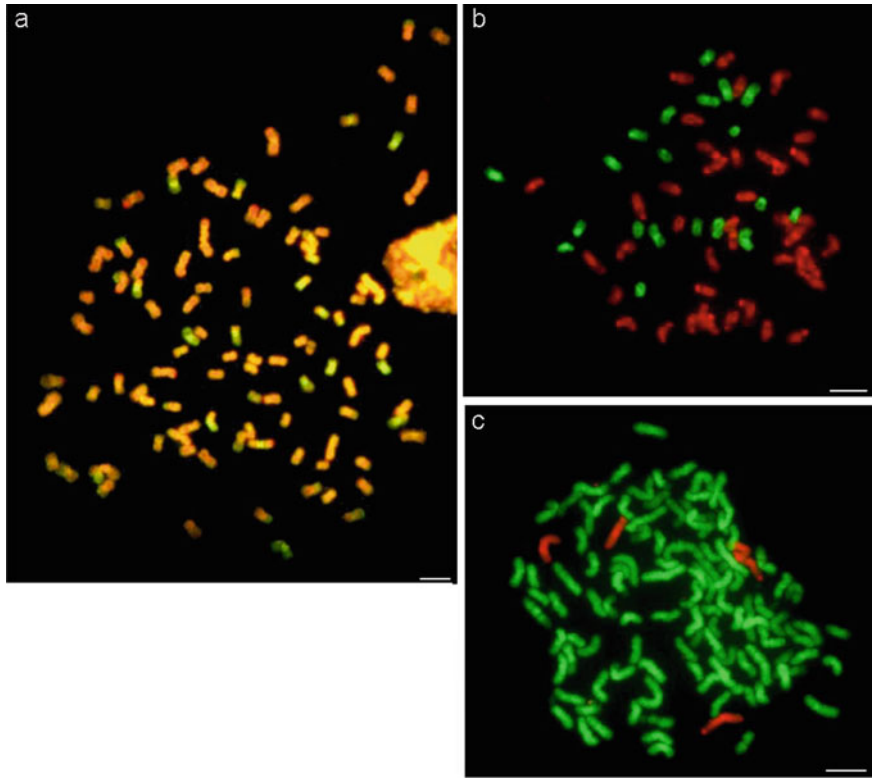


Fig. 2 (**a**) Interspecific chromosome composition of an atypical *S. officinarum* revealed by GISH using total genomic DNA for *S. officinarum* (in orange) and total genomic DNA from *S. spontaneum* (in green), recombined chromosomes appeared in both color. Intergeneric chromosome composition of an F1 (**b**) and a BC3 (**c**) revealed by GISH using total genomic DNA from *S. officinarum* (in green) and *E. arundinaceus* (in red). Scale bar: 5 μm

1.2.3 Revealing the Interspecific and Intergeneric Status of Intergeneric Hybrid Between Saccharum *and* E. arundinaceus

We also characterized intergeneric hybrids using DAPI as a third identification color tool to be able to distinguish the three different species involved, as we wanted to investigate the recombination event between the *Saccharum* and the *Erianthus* genome. The question was whether recombination was preferentially happening between *Saccharum officinarum* and *Erianthus* or *Saccharum spontaneum* and *Erianthus*. To resolve the question, *S. officinarum* was labeled in "red" with Alexa Fluor 594-5-dUTP or Rhodamine-5-dUTP, and *S. spontaneum* was labeled in "green" with Alexa Fluor 488-5-dUTP or Fluorescein-12-dUTP resulting in the *Erianthus arundinaceus* species to be visualize in "blue" from the DAPI stain; therefore, Vectashield without dye was apply on the slide. An *Erianthus* BC3 hybrid is presented in Fig. 3. *S. officinarum* appears pink in this image (resulting from the overlap of colors from the three images (blue, red, and green) necessary to reveal the three different genomes), while *S. spontaneum* appears green, and the *Erianthus* genome appears blue.

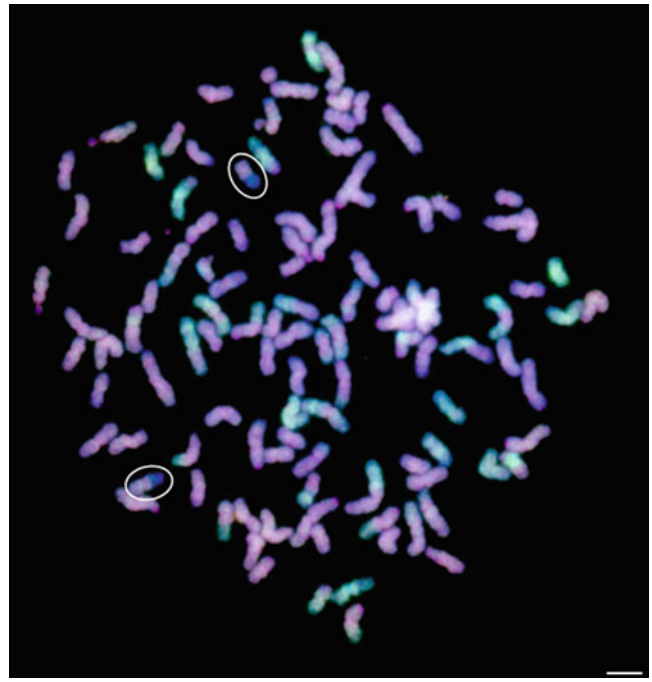


Fig. 3 Intergeneric/interspecific chromosome composition of an *Erianthus* BC3 hybrids revealed by GISH using total genomic DNA for *S. officinarum* (appeared in pink), total genomic DNA from *S. spontaneum* (in green), interspecific recombined chromosomes in pink & green. Two intergeneric recombined chromosomes between *S. officinarum* and *Erianthus arundinaceus* (in pink & blue) are circled in white. The blue (DAPI stain) correspond to the unlabelled *Erianthus* chromosome. Scale bar: 5 μm

2 Materials

Prepare all stock solutions using deionized distilled water (ddH_2O) and chemicals with the highest grade available. For most steps in DNA handling, it is essential that ddH_2O is autoclaved for at least 20 min at 130 °C in order to destroy any DNase activity and ensure sterility. All stock solutions have to be stored at room temperature (RT) unless stated otherwise.

2.1 Equipment

Besides the laboratory standard equipment, few specialized items are needed.

1. CCD (charge-coupled device) camera with image capture and processing software.
2. Coplin jars.
3. Epifluorescence/light microscope.
4. Heating plate with magnetic stirrer.

5. Hot plate with digital temperature control for slide warming.
6. Refrigerated centrifuge (swinging bucket is recommended).

2.2 Stock Solutions Stored at Room Temperature or on Ice for Immediate Use

1. 2× SSC, pH 7.0: Dilute 100 mL of 20× SSC (Saline Sodium Citrate buffer), pH 7.0, in 900 mL for a final volume of 1000 mL.
2. Fixative solution (3:1): Dilute 3 volumes of 100% ethanol to 1 volume of glacial acetic acid.
3. RNase A solution is freshly made upon treatment. Dilute (1/100) the thawed RNase A aliquot on ice: 8 μL of 1% RNase A + 80 μL of 20 × SSC + 712 μL ddH_2O.
4. Hybridization buffer (HB) 50 mL per slide: 25 μL FA, 10 μL DS, 5 μL 20× SSC, 1.5 μL SS DNA, and 80–100 ng of each DNA probe; make up to a final volume of 50 μL with ddH_2O (*see* **Notes 3** and **15**).

2.3 Stock Solutions Stored at 4 °C

1. Antifade for mounting slide: Vectashield Mounting Media with DAPI or Vectashield Mounting Media without DAPI (*see* **Note 1**).
2. 0.25 N HCl: Always work under fume hood; measure 195.56 mL of ddH_2O and then add 4.44 mL of pure HCl (*see* **Note 1**).
3. 0.04% 8-Hydroxyquinoline: Add 40 mg of 8-Hydroxyquinoline to 100 mL of ddH_2O. Place on a stirrer at RT for several hours. Store at 4 °C up to 1 year (*see* **Note 2**).
4. 3 M NaOAc, pH 5.2: Dissolve 40.81 g of sodium acetate trihydrate ($CH3COONa{\cdot}3H_2O$) in 30 mL ddH_2O, titrate pH to 5.2 with glacial acetic acid, and dilute with ddH_2O to a final volume of 100 mL.
5. TE buffer, pH 8.0: 10 mM Tris–HCl, pH 8.0, 1 mM Na_2EDTA. Add 20 μL of 1 M Tris–HCl, pH 8.0, 4 μL of 500 mM Na_2EDTA, pH 8.0, and 1976 μL of ddH_2O.

2.4 Stock Solutions Stored at −20 °C

1. 50% Dextran sulfate (DS): Dissolve 5 g of DS to a final volume of 10 mL of ddH_2O. Stir slowly until dissolved; it could take up to 24 h for the DS to be completely dissolved.
2. BioPrime DNA Labeling System for random priming labeling.
3. 1 μg/μL carrier DNA, Sheared Salmon Sperm DNA (SS DNA): Mix 10 mg of DNA with 10 mL of TE, pH 8.0. Shear in autoclave for 5 min, denature for 10 min in boiling water, and then place on ice. Aliquot and store.
4. Deionized formamide (FA) (*see* **Note 1**):Work under the fume hood. Add 5 g of ion exchange resin for each 100 mL formamide, cover with aluminum, and stir for 30–60 min. Filter twice with Whatman No. 1. Aliquot in 1 mL tubes as well as in 20 mL

tubes and store. Deionize all formamide when a new bottle is opened. Do not keep FA after opening.

5. Digestion citrate buffer: Add 1.47 g of trisodium citrate dihydrate ($Na_3C_6H_5O_7{\cdot}2H_2O$), 1.05 g of citric acid monohydrate ($C_6H_8O_7{\cdot}H_2O$), 2.8 g of KCl, and ddH_2O up to 500 mL. Adjust pH to 4.5, aliquot, and store.
6. Digestion enzyme solution: Add 0.25 g (5% final concentration) of cellulase Onozuka R-10 and 0.05 g (1% final concentration) of pectolyase Y-23 in 5 mL of digestion citrate buffer. Place on stirrer at RT for 1 h. Aliquot into microtubes and store.
7. Ethanol series: Prepare three solutions at 70%, 95%, and 100% ethanol in three Coplin jars and store.
8. 70% FA/2× SSC: Add 35 mL FA and 15 mL of 2× SSC (*see* **Note 1**).
9. Fluorochromes: 1 mM F-x-dUTP (*see* **Note 3**): ChromaTide Alexa Fluor 594-5-dUTP; ChromaTide Alexa Fluor 488-5-dUTP; Fluorescein-12-dUTP; Rhodamine-5-dUTP.
10. dNTP for random priming: dATP, dCTP, dGTP, dTTP (100 mM). Dilute each of the individual dNTP at 10 mM final concentration (10 μL of dNTP + 90 μL of ddH_2O).
11. dNTP Fluorochrome (10 mM) mix (10×): On ice, add 5 μL of each dATP, dGTP, and dCTP and 2.5 μL of dTTP together with 25 μL dUTP-Alexa and 7.5 μL of ddH_2O. Keep at −20 °C for up to 6 months.
12. dNTP Fluorescein and/or Rhodamine mix (10×): On ice, add 5 μL of each dATP, dGTP, and dCTP and 3.25 μL of dTTP together with 17.5 μL dUTP-Alexa and 14.25 μL of ddH_2O.-Keep at −20 °C for up to 6 months.
13. 1% RNase A in 10 mM Tris–HCl, pH 7.5, 15 mM NaCl (DNase-free): Dissolve 10 mg of RNase A in 987 μL ddH_2O with 10 μL of 1 M Tris–HCl, pH 7.5, and 3 μL of 5 M NaCl; incubate in boiling water bath for 15 min; cool slowly; and store in aliquots.

3 Methods

3.1 Root Pretreatment and Slide Preparation

Root tip meristems are the most commonly used plant tissues in cytogenetic methods for preparing mitotic chromosomes as they contain cells in active division. Plants are grown in a glass house in 20 L pots with a mixture of 50/50 vermiculate (coarse grade)/perlite (grade 3) with regular and sufficient application of water and nutrients (*see* **Note 4**). Root tip collection includes a pretreatment in order to arrest as many cells as possible in metaphase and a

fixative treatment, and then the roots can be stored in a 70% ethanol solution at 4 °C. For species with low mitotic index, it could be important to estimate the time of the day where best mitotic index slides are obtained. It is usually recommended to set up an assay where quality/mitotic index of slides is recorded in function of the collection time. The harvesting should be conducted by 1/2 h periods over an 8-h day. For example, in sugarcane, we harvest roots between 10 h 30 min and 11 h 00 min during the optimal growth period days of October to December [7].

3.1.1 Root Treatment

1. Approximately 0.5 cm of roots are harvested with fine forceps and placed directly in 5 mL bottles containing 0.04% 8-Hydroxyquinoline for 4 h at RT to arrest cells in metaphase (*see* **Note 1**).
2. Fix in freshly made 3:1 fixative solution for 72 h at RT.
3. Store roots in 70% ethanol at 4 °C until roots are spread.

3.1.2 Slide Preparation (See **Note 5**)

1. Rinse roots twice in ddH_2O for 10 min at RT.
2. Hydrolyze roots in 0.25 N HCl for 10 min at RT.
3. Rinse roots in ddH_2O for 10 min at RT.
4. Place roots in digestion citrate buffer for 10 min at RT.
5. Cut the distal 1–1.5 mm of the root tip with a fine scalpel; blot away excess moisture with filter paper.
6. Digest root tips in digestion enzyme solution for 90–180 min in a tube place in a water bath at 37 °C. Make sure the root tips are completely covered by the digestive solution. The length of time will vary with species and/or size of the root tips.
7. Carefully remove root tips from tubes and place in ddH_2O in a watch glass for at least 20 min at room temperature. Time must be optimized and the root cap must be removed to avoid high background.
8. Use a Pasteur pipette to carefully remove one root tip and place it on a pre-cleaned slide (*see* **Note 6**).
9. Add one or two drops of freshly prepared 3:1 fixative solution, immediately break apart the tip, and spread it with a pair of fine forceps (*see* **Note 7**).
10. Air-dry and store overnight in a desiccator (37 °C).

3.2 RNase A Treatment

Prior to any GISH experiment, slides are screened to select the ones with the best mitotic chromosome cells. Therefore, to avoid disappointment and reduce the cost of GISH if you work with species prominent to low mitotic index (as in sugarcane for example), we recommend to only hybridize slides with good mitotic preparations, i.e., with at least ten "complete" 2n cells. We also recommend prescreening slides under a 20× objective and recording

coordinates of the good mitotic cells for tracking purposes; this allows to go back straight to the recorded metaphases and avoid photobleaching when capturing images. We are also delimiting the hybridization area (with a diamond pen on the back of the slide) for a targeted and more efficient use of the hybridization buffer.

1. Add 50–100 μL of the freshly made RNase A solution on the slides, cover with a plastic cover slip (*see* **Notes 8** and **9**), and incubate in a humidified incubation chamber (*see* **Note 10**) for 45 min at 37 °C.
2. Rinse slides in a Coplin jar in 2× SSC for 10 min at RT.

3.3 GISH Experiment

The method described here for GISH experiment involves a random priming labeling method with direct fluorochrome. This method is the preferred method in our laboratory as it is very simple and reliable in order to acquire relatively quick and efficient results. There are alternative options to perform GISH in plants. Different methods such as Nick translation (NT) labeling with different types of haptens are extensively described in Zhang and Friebe [10]. One of the most common methods for GISH is NT labeling with biotin and/or digoxigenin, but these haptens will have to be detected and amplified in order to visualize the fluorescent signal. In sugarcane, NT can also be performed with Fluorescein-12-dUTP and/or Rhodamine-5-dUTP with excellent results.

3.3.1 Probe Labeling by Random Priming

Random priming achieves best result with good quality DNA. A mixture of different combinations of hexamers, octamers, or nonamers is annealed randomly to denatured DNA. The annealed small oligonucleotides will then act as primers and allow the synthesis of the complementary DNA strand by the Pol1 fragment of the Klenow enzyme (Pol1 has a DNA polymerase activity as well as exonuclease activity $3' \rightarrow 5'$). Labeled DNA will consist of a mixture of double- and single-stranded fragments. We use the kit BioPrime DNA Labeling System with the "green" and "red" Alexa fluorochromes (F-x-dUTP) or with Fluorescein-12-dUTP and Rhodamine-5-dUTP (*see* **Notes 11** and **12**).

1. On ice, firstly dilute 1 μg of genomic DNA in a volume of 19 μL ddH_2O and add 20 μL of Random Primers (from the kit) in a 1.5 mL tube. Denature the 39 μL in boiling water for 6 min and stand on ice for 15 min.
2. Finally, add 10 μL of 10× dNTP mix and 1 μL of Klenow enzyme. Mix gently, centrifuge briefly, and incubate in a water bath from 5 h to overnight at 37 °C. Longer incubation times usually increase product yield.
3. Add 5 μL of stop buffer.

4. Removal of unincorporated nucleotides and primers is not essential but is recommended to avoid background noise and can be performed by adding 1/10 volume of 3 M NaOAc, pH 5.2, and 2.5 volumes of 100% ethanol and centrifuging at 15,000 × *g* for 30 min at 4 °C. Discard supernatant and add 250 mL of 70% ethanol; centrifuge for 15 min. Discard supernatant carefully. Air-dry tubes for 5 min and resuspend in 20 μL of TE at 37 °C for 5 min. The concentration of the probe should be around 40–50 ng/μL (*see* **Note 12**). Another method to remove unincorporated nucleotides is to use a purification kit where the labeled DNA is purified through columns. This method seems to have a higher ratio of recovery of the labeled probe.
5. The fluorescence of fluorochrome-labeled probes can be estimated by a spot test as follows. Spot 1 μL of fluorochrome labeled probe onto a small piece of nylon membrane, air-dry for approximately 10 min, and then examine the fluorescence intensity under a fluorescence microscope with a suitable filter.
6. Probes can be stored at −20 °C.

3.3.2 Slide Denaturation

Chromosomes are denatured by placing slides on a hot plate at 80 °C in order to be ready for in situ hybridization.

1. Set a hot plate at 80 °C for at least 30 min prior to the denaturation process. We use a digital hot plate for better temperature accuracy (*see* **Note 13**).
2. Apply 200 μL of 70% FA/2× SSC solution, apply cover slip, and place on the hot plate for 3 min at 80 °C (*see* **Note 1**). Denaturation time has to be optimized according to the species and the age of the slides.
3. Remove the cover slip (*see* **Note 8**) and rinse slides in a Coplin jar standing in ice with 2× SSC (at −20 °C) for 3 min.
4. Dehydrate slides 5 min through an ethanol series of 70%, 95%, and finally 100% on ice. Solution of ethanol at 70% and 95% as well as 100% ethanol is kept at −20 °C.
5. Air-dry vertically (*see* **Note 14**).

3.3.3 In Situ Hybridization

1. Denature the freshly made HB for 10 min in boiling water and then place on ice for at least 15 min.
2. Deposit 50 μL of the HB on the dried slide; cover with a plastic cover slip. Avoid bubbles.
3. Place slide in a humidified incubation chamber (*see* **Note 10**) overnight at 37 °C.
4. Prepare three Coplin jars with 2× SSC, 0.5× SSC, and 0.1× SSC in a 42 °C water bath for stringency washes.

5. Remove the cover slip with a squirt of 2× SSC, wash slide in the 2× SSC for 10 min and then in the 0.5× SSC for 10 min, and finally wash with agitation in the last Coplin jar (0.1× SSC) for another 10 min at RT (*see* **Note 16**).
6. Drain slightly one slide at a time without letting it dry. Counterstain the slide with a drop of antifade Vectashield mounting media with DAPI (*see* **Notes 15** and **17**). Cover with a glass cover slip. Seal the cover slip with transparent nail polish. Dry slides horizontally in a slide holder protected from light. Observe under fluorescence microscope with appropriate filter. Store slides horizontally in the dark at 4 °C.

4 Notes

1. Some chemicals, especially HCl, FA, DAPI, and glacial acetic acid, are hazardous/toxic and should be handled with extreme caution. Some products such as FA are more toxic when heated, so always follow good laboratory practice and use the fume hood when required.
2. 8-Hydroxyquinoline is sensitive to light. It is therefore important to store the solution in the dark in a bottle covered with aluminum foil. It is best to place the bottle on a stirrer at least ½ h before using the solution. Finally, just before root collection, fill up 5 mL bottles and keep bottles in a box away from the light to ensure a good efficiency of the active product.
3. Fluorochromes will photo-bleached if exposed to light for long periods of time. During probe labeling preparation, it is recommended to work with a bench lamp directed away from the fluorochromes.
4. Ensure that at least for 4 h prior to harvesting, the roots are not being watered; they will be more accessible if the pots are not soaked. Good size roots are collected approximately every 3 weeks; if roots are not growing properly, it is recommended to use specialized root growth fertilizer.
5. Root treatment for the slide preparation can be performed in the bottle, and the storage solution is removed completely with plastic pipettes. If there is more than one clone/species to be treated, we use a microplate with 24 wells. We treat two to 20 roots from six different species per plate. Each line has four wells in use containing ddH_2O (×2), HCl, ddH_2O, and digestion buffer, respectively. Roots are handled carefully with tweezers in each bath for the 10 required minutes. Root tips/samples are then cut and grouped by size before being set up in digestion enzyme solution. After at least 90 min, the first lots of thinner tips are placed back in the washed microplate containing ddH_2O. The remaining tips are left in the water bath at 37 °C until ready to be spread.

6. Slides are placed in Coplin jars with 100% ethanol and dried just before use with kimwipes. Excess water is removed with a home-made micro Pasteur pipette firstly, and then we use the folded kimwipe to pre-clean the slide. The kimwipe with residual ethanol will suck the remaining water around the root.
7. If chromosomes on the slide have too much cytoplasm/too much cell wall debris, make sure that the root cap was removed before spreading as this increases the quality of the slide preparation. The root cap does not normally detach itself from the tip, and tweezers are most of the time required to remove the cap at this stage without damaging the tip itself. The digested cap-free tip has to be spread evenly on a 32 mm × 40 mm surface of the slide to concentrate the metaphasic chromosomes to a small area. Avoid spreading twice in the same localization.
8. We use pre-cuts of autoclave bags for plastic cover slip as they handle high temperature well and also as it seems that they do not trap too many bubbles.
9. To remove the cover slip, a squirt bottle of 2× SSC is recommended.
10. Our humidified incubation chamber consists of a large petri dish lined with paper at the bottom and soaked with water. The slides are set on plexiglass stick or bended Pasteur pipette so they are not directly in contact with the water.
11. To ensure better result during ethanol precipitation, we are using 100% ethanol at −20 °C, and after adding acetate sodium and ethanol, we leave the tube at −20 °C for 2 h or at −80 °C for 15 min. We also use a refrigerated centrifuge with a swinging bucket as the pellet of DNA would be precipitated at the bottom of the tube. We also preferably use screw cap tubes. After ethanol precipitation, DNA pellets labeled with a red fluorochrome are usually readily seen by the eyes, whereas those labeled with a green fluorochrome are usually of a pale shade of yellow and could not be easily seen. Before resuspending the probe, make sure that all the ethanol has been removed from the tube. Centrifuging tubes for another min at 10,000 × *g* can get rid of the excess ethanol as residual ethanol in probe and slides could result in higher background signal.
12. If your slides present no, weak, or patchy hybridization, it is often the result of labeling problems. Check the quality of the DNA on an agarose gel before labeling as good quality DNA will give a better probe and the length of the probe is also essential.

    Also check the expiry date of the enzymes and dUTPs being used.

13. If you encounter a poor signal from the probe as well as from the counterstain and if chromosome morphology appears abnormal, try denaturing the slide a little less than the recommended 3 min and ensure that the temperature of the hot plate is less than or equal to 80 °C. Poor signal/DAPI stain is very common if chromosomes have a ghostly look just after being spread.
14. The slides can be put in an oven at 37 °C to reduce drying time. Residual ethanol in slides can cause higher background signal.
15. It is recommended to avoid as much as possible exposure to light during the entire procedure (labeling, hybridization, post-hybridization wash, image capture). The laboratory should be entirely dark except from the light coming from a benchtop lamp. When capturing images, be as quick as possible because each exposure to fluorescent light will remove energy from the fluorochrome and therefore decrease its intensity.

    If the hybridization signal is poor, the concentration of the probe used during hybridization might be too low. Try different concentrations of the probe, but ensure that the concentration of the probe after precipitation has not been overestimated. Also make sure that the hybridization solution was mixed thoroughly as the DS solution is very viscous. It is possible to use a special piston pipette or a normal pipette with a cutoff pipette tip to slowly mix the solution up and down.

    Finally, ensure that no bubbles remain between the slide and cover slip after adding the hybridization solution. If bubbles appear, use fine tweezers to lift up and down the cover slip to carefully remove them.
16. Post-hybridization washes are very important to remove unattached probe and therefore reduce the background signal. Ensure that washes are performed according to the procedure.
17. After applying a drop of mounting media (Vectashield with or without DAPI depending on the experiment), we apply gentle pressure on the glass cover slip in order to remove excess media. We use a layer of kimwipe directly on the cover slip and three layers of Whatman paper. Always apply pressure with the thumb when slides are placed on a flat surface in order to prevent breaking the slide and/or cover slip.

Acknowledgments

This work was financially supported by Sugar Research Australia Ltd. (previously BSES Ltd) and also previously by the Australian Centre for International Agricultural Research and the Cooperative Research Centre for Sugar Industry Innovation through Biotechnology.

References

1. Le HT, Armstrong KC, Miki B (1989) Detection of rye DNA in wheat-rye hybrids and wheat translocation stocks using total genomic DNA as a probe. Plant Mol Bio Rep 7:150–158
2. Schwarzacher T, Leitch AR, Heslop-Harrison JS (1989) In situ localization of parental genomes in a wide hybrids. Ann Bot 64:315–324
3. Jiang J, Gill BS (1994) Different species-specific chromosome translocations in *Triticum timopheevii* and *T. turgidum* support the diphyletic origin of polyploid wheats. Chromosom Res 2(1):59–64
4. Jiang J, Gill BS (2006) Current status and the future of fluorescence in situ hybridization (FISH) in plant genome research. Genome 49:1057–1068
5. Sreenivasan TV, Ahloowalia BS, Heinz DJ (1987) Cytogenetics. In: Heinz DJ (ed) Sugarcane improvement through breeding. Elsevier, New York, pp 211–253
6. D'Hont A, Ison D, Alix K, Roux C, Glaszmann J-C (1998) Determination of basic chromosome number in the genus *Saccharum* by physical mapping of ribosomal RNA genes. Genome 41:221–225
7. Piperidis N, Chen JW, Deng HH, Wang LP, Jackson P, Piperidis G (2010) GISH characterization of *Erianthus arundinaceus* chromosomes in three generations of sugarcane intergeneric hybrids. Genome 53:331–336
8. Piperidis G, Piperidis N, D'Hont A (2010b) Molecular cytogenetic investigation of chromosome composition and transmission in sugarcane. Mol Gen Genomics 284: 65–73. https://doi.org/10.1007/s00438-010-0546-3
9. Irvine JE (1999) *Saccharum* species as horticultural classes. Theor Appl Genet 98:186–194
10. Zhang P, Friebe B (2009) FISH on plant chromosomes. In: Liehr T (ed) Fluorescence in situ hybridization (FISH)—application guide. Springer Protocol VI, Berlin, pp 365–394. https://doi.org/10.1007/978-3-540-70581-9_32

INDEX

Pascale Besse (ed.), *Molecular Plant Taxonomy: Methods and Protocols*, Methods in Molecular Biology, vol. 2222, https://doi.org/10.1007/978-1-0716-0997-2, © Springer Science+Business Media, LLC, part of Springer Nature 2021

E

F

G

H

I

K

L

M

N

O

P

Q

R

S

T

U

V