Methods in Molecular Biology 2264

Springer Protocols

Pasquale Tripodi *Editor*

Crop Breeding

Genetic Improvement Methods

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.

Crop Breeding

Genetic Improvement Methods

Edited by

Pasquale Tripodi

Council for Agricultural Research and Economics - Research Centre for Vegetable and Ornamental Crops (CREA-OF), Pontecagnano, SA, Italy

Humana Press

Editor
Pasquale Tripodi
Council for Agricultural Research
and Economics - Research Centre
for Vegetable and Ornamental Crops
(CREA-OF)
Pontecagnano, SA, Italy

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-0716-1200-2 ISBN 978-1-0716-1201-9 (eBook)
https://doi.org/10.1007/978-1-0716-1201-9

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

The challenges of modern agriculture require increasingly innovative investigation methodologies to be applied to the genetic improvement of crops. Cutting-edge technologies for genome sequencing and plant phenotyping along with computational and bioinformatic tools are speeding the throughput and deepness of the investigation of complex traits. Genome editing approaches are increasingly applied for genetic improvement in many plant species. The successful use of these methodologies requires solid lab-based knowledge in order to prepare the experimental study populations, samples for analysis as well to develop an accurate workflow. The complexity of the methods requires deep specialization and greater interdisciplinarity of skills.

The present volume "Crop Breeding: Genetic Improvement Methods" is a result of a collaboration with leading scientists from main international universities and research institutions working in the crop breeding sector.

Aiming at covering all topics, the present volume describes breeding methods for the development of biparental and multiparental mapping populations, lab protocols for high-throughput isolation of nucleic acids and metabolites, different high-performing genotyping approaches, mapping strategies for QTLs and mutation identifications, computational and bioinformatic pipelines, tissue culture-based and transformation methods for androgenesis, ploidy modification, and RNA interference. The book highlights recent developed genome editing protocols including CRISPR and TALEN methods and methodologies for in-field/in-soil plant phenotyping.

"Crop Breeding: Genetic Improvement Methods," therefore, cover all aspects, being addressed to the broadest audience of students, breeders, and scientists applying current protocols or interested in the knowledge of the described methodologies.

Pontecagnano, SA, Italy ***Pasquale Tripodi***

Contents

Contributors

Mouhannad Alsalem • *Department of Crop Sciences, Institute of Agronomy, University of Natural Resources and Life Sciences Vienna, Tulln, Austria*

Saleh Alseekh • *Max-Planck-Institute of Molecular Plant Physiology, Potsdam-Golm, Germany; Center of Plant System Biology and Biotechnology, Plovdiv, Bulgaria*

Joana S. Amaral • *Centro de Investigação de Montanha (CIMO), Instituto Politécnico de Bragança, Bragança, Portugal*

Mariette Andersson • *Department of Plant Breeding, Swedish University of Agricultural Sciences, Alnarp, Sweden*

Gernot Bodner • *Department of Crop Sciences, Institute of Agronomy, University of Natural Resources and Life Sciences Vienna, Tulln, Austria*

Carmen Capel • *Centro de Investigación en Biotecnología Agroalimentaria, Departamento de Biología y Geología, Universidad de Almería, Almería, Spain*

Elisa Cappetta • *Department of Agricultural Sciences, University of Naples Federico II, Portici, Italy*

Cristina Cecchini • *Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria (CREA), Centro di Ricerca Ingegneria e Trasformazioni Agroalimentari, Monterotondo, Rome, Italy*

Tomaš Čermák • *Department of Genetics, Cell Biology and Development & Center for Genome Engineering, University of Minnesota, St. Paul, MN, USA; Inari Agriculture, Cambridge, MA, USA*

Felice Contaldi • *CREA Research Centre for Vegetable and Ornamental Crops, Pontecagnano Faiano, Italy*

Andrea Copetta • *Council for Agricultural Research and Economics Research Centre for Vegetable and Ornamental Crops (CREA-OF), Sanremo, IM, Italy*

Corrado Costa • *Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria (CREA), Centro di Ricerca Ingegneria e Trasformazioni Agroalimentari, Monterotondo, Rome, Italy*

Joana Costa • *REQUIMTE-LAQV, Faculdade de Farmácia, Universidade do Porto, Porto, Portugal*

Nunzio D'Agostino • *Department of Agricultural Sciences, University of Naples Federico II, Portici, Italy*

Isidore Diouf • *INRAE, UR1052, Génétique et Amélioration des Fruits et Légumes, Centre de Recherche PACA, Montfavet, France*

Salvatore Esposito • *CREA Research Centre for Vegetable and Ornamental Crops, Pontecagnano Faiano, Italy*

Ann-Sofie Fält • *Department of Plant Breeding, Swedish University of Agricultural Sciences, Alnarp, Sweden*

Alisdair R. Fernie • *Max-Planck-Institute of Molecular Plant Physiology, Potsdam-Golm, Germany; Center of Plant System Biology and Biotechnology, Plovdiv, Bulgaria*

Giovanna Festa • *CREA Research Centre for Vegetable and Ornamental Crops, Pontecagnano Faiano, Italy*

Simone Figorilli • *Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria (CREA), Centro di Ricerca Ingegneria e Trasformazioni Agroalimentari, Monterotondo, Rome, Italy*
Teresa Rosa Galise • *Department of Agricultural Sciences, University of Naples Federico II, Portici, Italy*
Edgar García-Fortea • *Instituto de Conservación y Mejora de la Agrodiversidad Valenciana, Universitat Politècnica de València, Valencia, Spain*
Ana García-Pérez • *Instituto de Conservación y Mejora de la Agrodiversidad Valenciana, Universitat Politècnica de València, Valencia, Spain*
Karla Gasparini • *Departamento de Ciências Biológicas, Escola Superior de Agricultura "Luiz de Queiroz", Universidade de São Paulo, Piracicaba, SP, Brazil*
Esther Gimeno-Páez • *Instituto de Conservación y Mejora de la Agrodiversidad Valenciana, Universitat Politècnica de València, Valencia, Spain*
Pietro Gramazio • *Instituto de Conservación y Mejora de la Agrodiversidad Valenciana, Universitat Politècnica de València, Valencia, Spain*
Liliana Grazina • *REQUIMTE-LAQV, Faculdade de Farmácia, Universidade do Porto, Porto, Portugal*
Per Hofvander • *Department of Plant Breeding, Swedish University of Agricultural Sciences, Alnarp, Sweden*
José M. Jiménez-Gómez • *Institut Jean-Pierre Bourgin, INRAE, AgroParisTech, Université Paris-Saclay, Versailles, France*
Zoltán Kevei • *Cranfield Soil and Agrifood Institute, Cranfield University, Cranfield, UK*
Pawan L. Kulwal • *State Level Biotechnology Centre, Mahatma Phule Agricultural University, Rahuri, Maharashtra, India*
Marina Laura • *Council for Agricultural Research and Economics Research Centre for Vegetable and Ornamental Crops (CREA-OF), Sanremo, IM, Italy*
Jianjun Lei • *Key Laboratory of Biology and Germplasm Improvement of Horticultural Crops in South China, Ministry of Agriculture, College of Horticulture, South China Agricultural University, Guangzhou, China*
Rafael Lozano • *Centro de Investigación en Biotecnología Agroalimentaria, Departamento de Biología y Geología, Universidad de Almería, Almería, Spain*
Isabel Mafra • *REQUIMTE-LAQV, Faculdade de Farmácia, Universidade do Porto, Porto, Portugal*
Aimee A. Malzahn • *Department of Plant Science and Landscape Architecture, University of Maryland College Park, College Park, MD, USA*
Marina Martínez-López • *Instituto de Conservación y Mejora de la Agrodiversidad Valenciana, Universitat Politècnica de València, Valencia, Spain*
Alireza Nakhforoosh • *Global Institute for Food Security, University of Saskatchewan, Saskatoon, SK, Canada*
Alessandro Nicolia • *CREA Research Centre for Vegetable and Ornamental Crops, Pontecagnano Faiano, Italy*
Federico Pallottino • *Consiglio per la Ricerca in Agricoltura e l'Analisi dell'Economia Agraria (CREA), Centro di Ricerca Ingegneria e Trasformazioni Agroalimentari, Monterotondo, Rome, Italy*
Laura Pascual • *Department of Biotechnology-Plant Biology, School of Agricultural, Food and Biosystems Engineering, Universidad Politécnica de Madrid, Madrid, Spain*
Mariola Plazas • *Instituto de Conservación y Mejora de la Agrodiversidad Valenciana, Universitat Politècnica de València, Valencia, Spain*

Jaime Prohens • *Instituto de Conservación y Mejora de la Agrodiversidad Valenciana, Universitat Politècnica de València, Valencia, Spain*
Yiping Qi • *Department of Plant Science and Landscape Architecture, University of Maryland College Park, College Park, MD, USA; Institute for Bioscience and Biotechnology Research, University of Maryland, Rockville, MD, USA*
Ravinder Singh • *School of Biotechnology, Sher-e-Kashmir University of Agricultural Sciences and Technology of Jammu, Jammu, India*
Binmei Sun • *Key Laboratory of Biology and Germplasm Improvement of Horticultural Crops in South China, Ministry of Agriculture, College of Horticulture, South China Agricultural University, Guangzhou, China*
Pasquale Termolino • *Institute of Biosciences and Bioresources (IBBR), National Research Council of Italy (CNR), Naples, Italy*
Pasquale Tripodi • *Council for Agricultural Research and Economics - Research Centre for Vegetable and Ornamental Crops (CREA-OF), Pontecagnano, SA, Italy*
Santiago Vilanova • *Instituto de Conservación y Mejora de la Agrodiversidad Valenciana, Universitat Politècnica de València, Valencia, Spain*
Fernando J. Yuste-Lisbona • *Centro de Investigación en Biotecnología Agroalimentaria, Departamento de Biología y Geología, Universidad de Almería, Almería, Spain*
Zhangsheng Zhu • *Key Laboratory of Biology and Germplasm Improvement of Horticultural Crops in South China, Ministry of Agriculture, College of Horticulture, South China Agricultural University, Guangzhou, China; Peking University-Southern University of Science and Technology Joint Institute of Plant and Food Sciences, Department of Biology, Southern University of Science and Technology, Shenzhen, China*
Agustin Zsögön • *Departamento de Biologia Vegetal, Universidade Federal de Viçosa, Viçosa, MG, Brazil*

Chapter 1

Methods of Development of Biparental Mapping Populations in Horticultural Crops

Pasquale Tripodi

Abstract

Biparental mapping populations consist of a set of individuals derived from crosses between two parents often belonging to diverse species of a botanical genus and differing in terms of phenotype and traits to share. The development of such recombinant libraries represents a powerful strategy for dissection of the genetic basis of complex traits in crops and these are largely utilized to develop pre-breeding sources to use in crop improvement. This chapter provides an overview of methods and strategies to follow, for the construction of different types of populations, from a plant breeder point of view. Starting from the initial crossing between founder lines toward the further selection steps, here are described the populations commonly established in autogamous species including F_2, double haploids, backcrosses and recombinant inbreds, and introgression lines.

Key words Breeding, Backcrosses, Segregation, Introgression lines, Double haploids, Recombinant inbred lines, Quantitative traits

1 Introduction

Breeding activities rely on the strategies toward the development of improved varieties through the generation of novel highly heritable variation and the selection of desirable traits. The factors determining the success of a breeding strategy are related to the genetic nature of the traits to transfer and to the selection method adopted. The complexity of a breeding program can vary if the target trait to transfer is a single gene with a greater effect on the phenotype or if it is a quantitative trait. Since the origins of agriculture and from the passage of the communities from hunter-gatherer to farmers, several transitions occurred which involved natural and artificial selection. These processes, known as domestication, lead to a major adaptation of crops to cultivation and utilization by humans [1]. The discoveries of Mendel and the advent of modern agriculture and new technologies in the field of molecular biology have led to a boost of traditional and marker-assisted breeding programs in

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_1,

the last two centuries [2]. Over the last 30 years, the main target in the genetic improvement of crops has been the transfer of hundreds of alleles from exotic resources to cultivated germplasm [3, 4]. This has been facilitated through the development of various types of experimental mapping populations (EMP) [5]. Despite the efforts done, there is a tremendous need to establish novel breeding programs; in fact, in the few next decades, the agricultural scenario will face multiple challenges related to population increases and climatic variations [6]. Therefore, it is essential to develop advanced genetic materials with high yield and with resistance/tolerance to various biotic and abiotic stresses. Most of the traits of agricultural interest have a quantitative nature and are known as quantitative trait loci (QTLs). QTLs can be controlled by few or many loci with large or minor effects on the phenotype. In both cases, the transfer of QTLs and the investigation of their genetic basis are possible through appropriate crossbreeding programs leading to the development of EMP. EMP can be obtained through biparental crosses leading to the combination and fixing of alleles from two sources, in advanced generations. These populations can be classified according to their genetic constitution and level of segregation [7]:

1. Permanent populations consist of a set of lines with a fixed genetic background established after several generations of selfing. Each line differs from the other for the represented fragments of the founder lines. These can be evaluated over years and across locations.
2. Segregant (or temporary) populations for which the genetic constitution is not fixed and can change every generation of inbreeding: In general, they are used in single experiments.

Although the literature is full of articles reporting the results of genetic and genomic studies through the use of EMP, in many cases the strategies of development are not well described. This chapter aims to provide, from a breeder point of view, the method of establishment of biparental EMP starting from the crossings between parents until the selection of successive generations.

2 Materials

Hereafter are reported facilities, types of equipment, and materials needed to develop a biparental plant breeding program.

2.1 Facilities

A controlled environment to make crosses and further selection steps are required. Different types of facilities can be counted including greenhouses or glasshouses or tunnels. For a major efficiency, the possibility to control temperatures and humidity is preferable.

2.1.1 Greenhouses

1. Structure: concrete and iron-glass or iron-polycarbonate.
2. Sensors to register internal environmental conditions (temperature, humidity, light).
3. A cooling system using evaporative fans and pads: Pads are mounted on a sidewall and are supplied with water. Air drawn through the pads by fans on an opposite wall drops the indoor temperature by 5–10 °C. Alternatively it is possible to use a fogging system using high-pressure water delivery to emit very fine water particles, and the drops evaporate in the air, reducing the temperature.
4. Shading screen double layer (Ludvig Svensson) able to reduce the sun's intensity, as well as retain heat overnight.
5. Metal benches for in-pot plant growing. Plants can be grown directly in the soil.
6. Control of cold temperature using a tube rail heating system.
7. Automated fertigation system able to supply water and fertilizer jointly.
8. Lamps for lighting supply especially during the winter season (e.g., LED bulbs or energy-saving fluorescent lights).

2.1.2 Tunnels

Alternatively, a tunnel covered by polyethylene and closed by an insect-proof net in order to avoid cross-pollen contamination. Plants can be grown directly in the soil or in pots positioned on the soil or benches. For irrigation it is possible to use both perforated plastic hoses positioned directly on the ground or rainy sprinklers. Tunnels can use meteorological station or thermometer (minimum-maximum) with magnet for moving mercury back to start at the end of each day. Shade cloths can be applied in order to decrease the temperature in the warmer season.

2.2 Equipment

1. Pointed tweezers for floral emasculation.
2. Magnifying glass lens (wearable glass type).
3. Pollination work aprons.
4. Breeding labels with strings.
5. Permanent ink markers, highlighters.
6. Ethanol and toilet paper to sterilize equipment between pollinations.
7. A vibrator (or small brushes) for collecting pollen.
8. Eppendorf for pollen storage.
9. Bag to isolate flowers for self-pollination and/or after crosses.
10. Notebook for taking note of crosses and observed phenotypic traits.
11. Correction fluid/correction tape.

2.3 Plant Materials

The two parental lines used for a biparental breeding program are the recurrent cultivar which in general is an elite variety lacking few traits which are transferred by the donor line. Therefore, the choice of parent lines is the fundamental point behind the success of a cross-breeding program. This depends on the level of polymorphism and how distantly related they are, the purity in terms of homozygosity level, the selected trait(s) to be introgressed, and the final background in which traits need to be transferred. For instance, disease resistance traits are usually transferred from wild species which are used as a donor parent. For qualitative traits, accessions at different biological statuses (e.g., landraces, wilds, cultivars) can be selected as a donor. For mapping quantitative traits it is highly recommended to choose the most phenotypically and genetically distant founders.

3 Methods

3.1 F_2 Populations

Two pure parental lines differing for many traits to be studied are selected as parents, parent 1 (P1) and parent 2 (P2). Contrasting characteristics can be linked to the qualitative and morphological features of the fruits, plant architecture, resistances, metabolic diversity, etc.

P1 and P2 are sown and grown in pots, at flowering stage; for hermaphrodite species (flowers with both staminate, male, pollen-producing, and carpellate, female, ovule-producing, parts; e.g., tomato, pepper, bean) flowers of P1 are emasculated and pollinated using the pollen collected from mature flowers of P2. Emasculation consists of the removal of anthers and stamen using pointed tweezers without damaging the other parts of the flower such as the ovary and the pistil. This process needs to be done before flower maturity in order to avoid any release of pollens from anthers. In the case of species bearing male and female organs on separate flowers (monoecious such as watermelon, corn) the male flower from P1 can be removed and pollen from P2 will be used to pollinate female flowers of P1. The crossed flower can be then covered using the appropriate envelopes. Once fruits are grown, seeds are harvested and sown, and the resulting plants are the F_1 (hybrid) progeny.

For autogamous crops, if P1 and P2 are homozygous at all *loci*, all individuals of the F_1 generation have the same genotype and phenotype, in agreement with the principle of uniformity of Mendel. For allogamous plants, instead, a rate of heterozygosity in P1 and P2 is expected. The resulting hybrid is then self-fertilized to produce the F_2 population segregating for all traits differing between the two parents (Fig. 1).

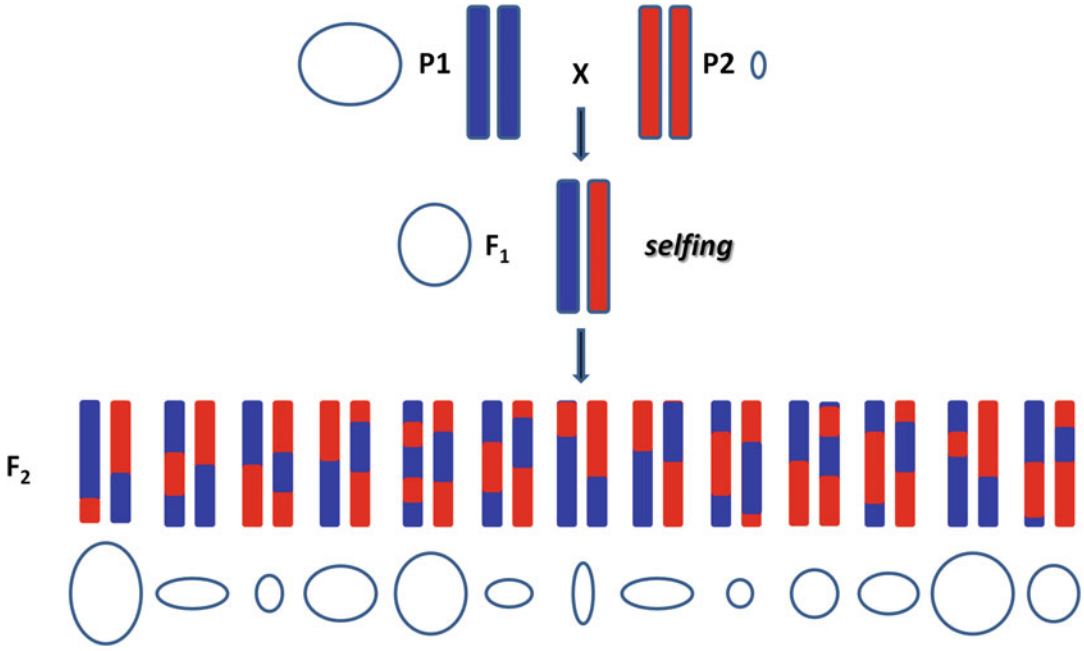


Fig. 1 Method of development of a F_2 progeny. Two parent lines P1 and P2, differing for fruit size (circles), are crossed. The resulting hybrid is self-fertilized for one generation. The segregation of the progeny is shown by the size of the circles

For self-fertilizing, the plant must be grown in controlled environments avoiding the entrance of insects (e.g., bees, butterflies) causing any cross-pollination. Flowers can be covered by bags before opening; alternatively, the whole plant can be covered using an insect-proof net. Once flowers are fertilized and fruits grown, the covers can be removed. Seeds are then harvested from mature fruit and the derived plants are F_2. Once established the F_2 population can be used for trait mapping experiments (*see* **Note 1**). QTL analysis requires the development of a linkage map able to associate the observed phenotypes to the recombinant events occurring in individuals of the population [8]. Although there is not a defined number of individuals required for QTL mapping since it depends on the number of traits to be scored and the inheritance mechanism, more than 100 individuals are advisable in order to allow a good resolution of linked (>1% recombinant frequency ~1 cM) loci at feasible costs [7]. QTL analysis can be done also using the phenotype of F_3 progenies and the genotype of F_2-related parents [9].

3.2 Double Haploids (DH)

Haploids are individuals containing a single complete set of chromosomes. Double haploids (DH) are diploids produced from chromosome doubling of haploids, therefore containing identical set of chromosome pairs in cells [10].

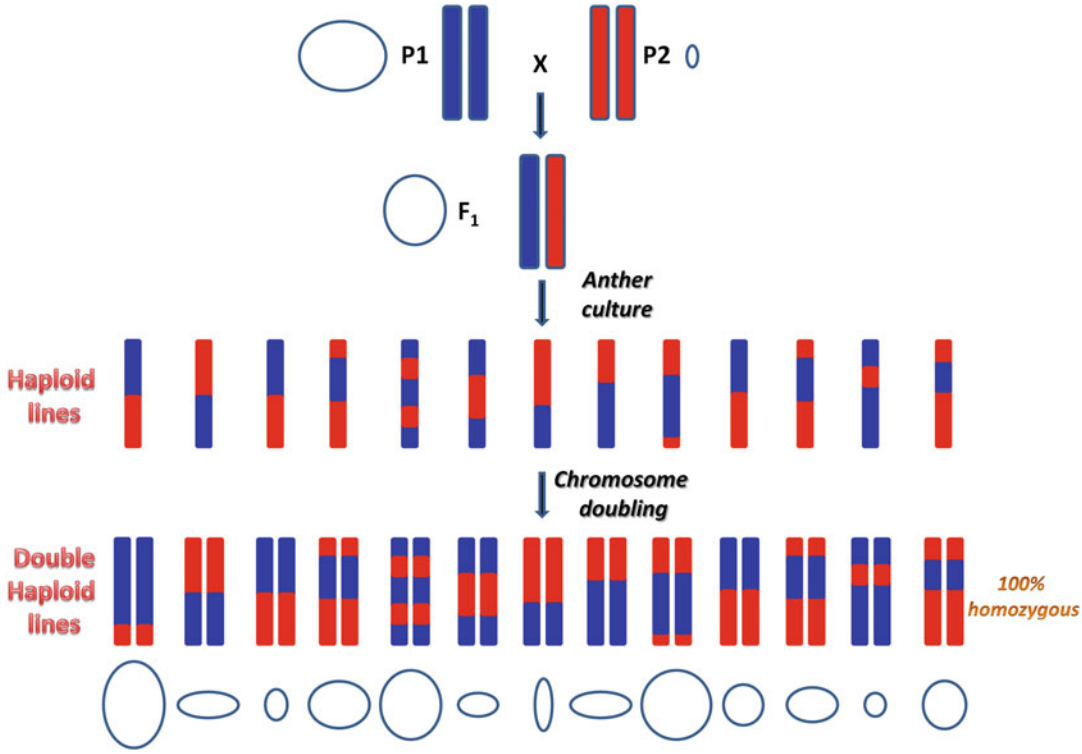


Fig. 2 Method of development of double haploids (DH). Two parent lines P1 and P2, differing for fruit size (circles), are crossed. From selfing individuals of the resulting hybrid, the anthers are taken and cultured on a special media. Haploid plants are then regenerated from haploid cells and double haploids are developed by treatment using colchicines

The development of DH aims to achieve stabilized populations starting from the F_2 avoiding several selfing cycles (*see* **Note 2**). In in vitro methods, the flowers of each segregating F_2 individual are harvested and anthers put on a nutrient medium in order to produce haploid cells (HC) (androgenesis). Thereafter, from each HC of the gametophyte, plants are regenerated. Haploid plants are sterile and cannot produce any progeny; therefore, each line is treated by colchicines [10], a compound that inhibits the cell division at metaphase and the related separation of chromosomes, leading to the formation of di-haploid cells with double chromosome numbers (Fig. 2). Gynogenesis is another in vitro method consisting of the culture of unfertilized isolated ovules and ovaries isolated from flower buds in order to develop embryos from cells of the embryo sac [7]. DH can also be obtained in vivo through parthenogenesis or chromosome elimination after hybridization [10]. Methods of development specific for each crop are extensively described in the literature [7, 10]. Individuals of the resulting population are called double haploids (DH). Although fewer recombinations occur with respect to F_2, for QTL analysis more than 100 individuals are recommended.

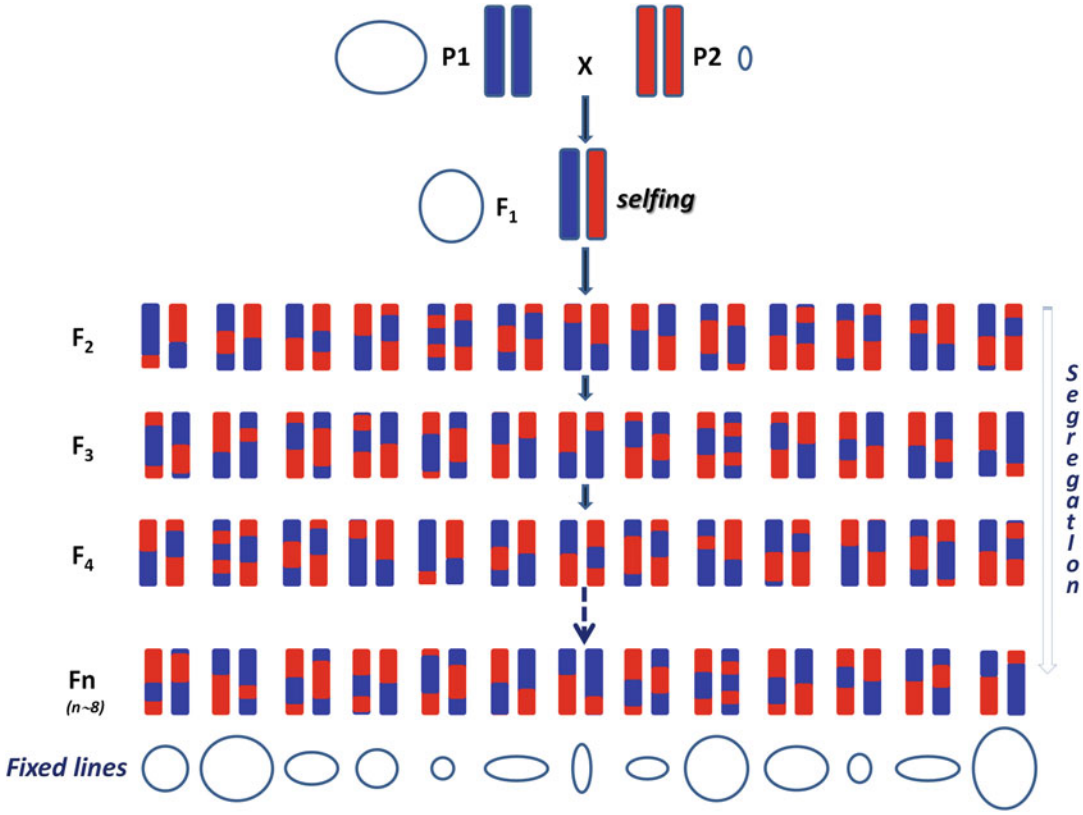


Fig. 3 Method of development of recombinant inbred lines (RILS). Two parent lines P1 and P2, differing for fruit size (circles), are crossed. The F_2 is selfed by single seed descent for *n* generations in order to develop a population of fixed "immortal" lines nearly homozygous at each locus

3.3 Recombinant Inbred Lines (RILs)

Recombinant inbred lines (RILs) are developed by self-fertilizing for *n* generations of the segregating individuals of an F_2 population through single-seed descent. At each generation, the degree of homozygosity at each locus increases, being as general rule 75% at F_3, 87.5% at F_4, 92.25% at F_5, 96.8% at F_6, 98.4% at F_7, and 99.22% at F_8 (Fig. 3). The homozygosity rate is due to the recombination frequencies occurring between linked loci. Low recombination frequency allows the development of a higher proportion of homozygotes and stabilizes the population more rapidly. Procedures for flower selfing are carried out for each individual with the methodology described above (Subheading 3.1). Blocks of alleles are inherited from each parent. Therefore, each individual is the result of a mixture of the genome of founder lines. Closer loci have a higher probability of descending from the same parent, while recombination can mix loci at each selfing generation. At F_6 and further generations, each RIL is the result of a mix of the P1 and P2 genome (mosaic genome). Given the several rounds of meiosis done before homozygosity, the degree of recombination is higher

compared to F_2 and DH populations. Only tightly linked loci do not recombine during the repetitive self-fertilizations. The resulting molecular maps show a higher resolution with respect to those generated in F_2 allowing to determine the position of tightly linked markers (*see* **Note 3**).

3.4 Backcross Inbred Lines (BILs)

Backcross lines are developed by crossing back the F_1 obtained by P1 (recurrent) × P2 (donor) with the recurrent parent. Backcrossing is widely used in crop breeding for the improvement of one or a few major traits which are present in the donor line and lacking in the recurrent one (an elite cultivar). The method allows the transfer of the target trait and the recovery of the recurrent genome through multiple hybridizations cycles. The process continues until the progeny is highly similar to the elite cultivar except that for the target trait transferred from the donor.

Through this crossing scheme it is possible to develop a population of backcross inbred lines (BIL, also called inbred backcross lines) carrying portion of donor fragment in the recurrent genotype. Once the first generation of a backcross is developed, namely BC_1, a large number of individuals can be self-fertilized for *n* generation by means of single-seed descent to establish a BC_1_BIL.

Alternatively, they can be backcrossed to P1 to develop the BC_2 generation. In some cases, the BC_1 to backcross can be selected by molecular markers (e.g., marker close to the gene of interest) or for phenotype characteristics (e.g., high content of ascorbic acid). Individual plants of BC_2 can be self-pollinated for *n* generations in order to develop a BC_2_BIL or backcrossed to P1 to produce a BC_3. Individuals in the BC_3 population are self-pollinated until they reach homozygosity. In general, the number of self-cycles required is five or more generations; however, in order to keep a rate of heterozygosity it is possible to stop at the second or third generation. A BIL population consists of a set of backcross-inbred individuals stabilized; therefore, it can be considered permanent (*see* **Note 4**). More backcrosses allow recovering a larger amount of the recurrent parent (Fig. 4). Contrariwise, the probability of recovering the genes from the donor parent is reduced by half for each generation due to the backcrossing process.

More backcrosses are in general preferred when few genes need to be transferred or few traits are mapped; in fact, they allow reducing the number and size of donor fragments, separating by segregation the unlinked fragments, and minimizing linked fragments due to recombination with the recurrent parent. Molecular markers can help to select individuals with desired traits.

3.5 Introgression Lines (ILs)

Introgression lines (reported also as chromosome segment substitution lines) are germplasm libraries in which each individual harbors a single fragment of an alien genome (donor parent) in the

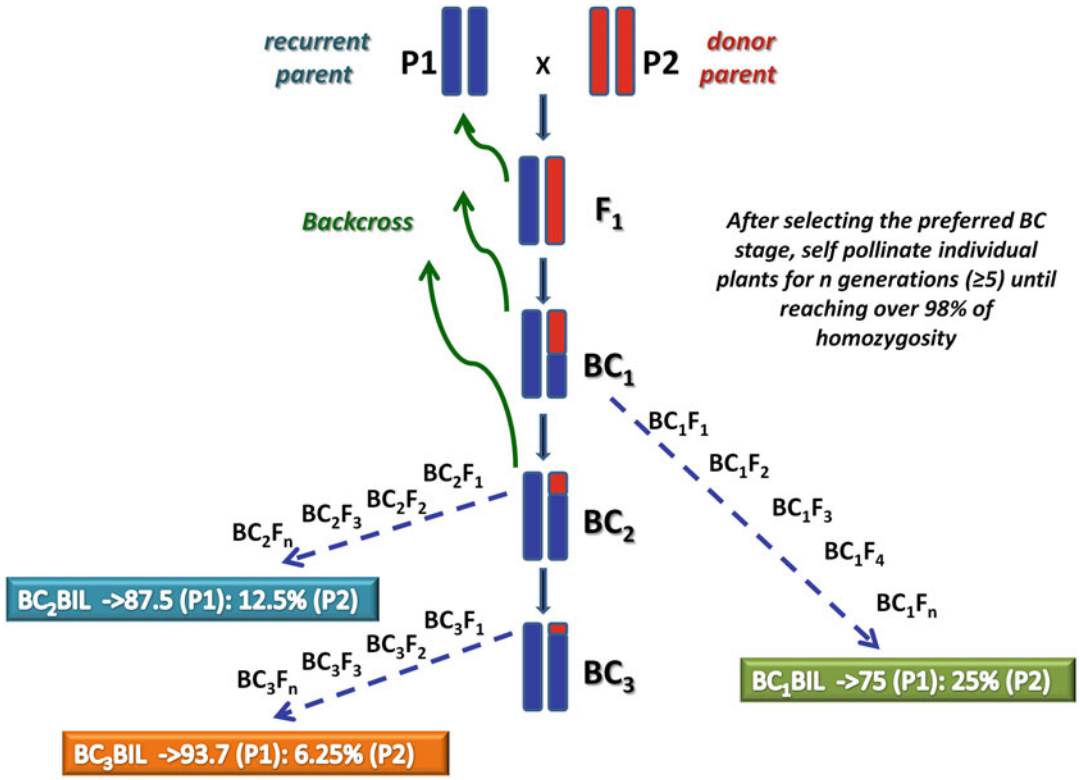


Fig. 4 Method of development of backcrossed inbred lines (BILs). After the cross of the two parent lines P1 and P2, the hybrid is backcrossed to the recurrent parent to produce BC lines which can be selfed by single seed descent for *n* generations or repeatedly backcrossed to the same recurrent parent. With each round of backcrossing, the number and size of genomic fragments of the donor parent are reduced by 50%. The BIL populations can be considered highly homozygous at all loci after about five cycles of selfing

genetic background of an elite variety (recurrent parent). The hybrid developed from the initial cross is backcrossed to the recurrent parent to obtain the BC_1 generation. The backcross scheme continues up to the fifth or sixth generation as described above (Subheading 3.4). At each BC offspring, the genome fraction of recurrent parent is halved while that of recurrent parent increases in a proportion equal to $1 - (1/2)x$ where "x" is the generation of backcrossing (Fig. 5). The linkage often occurring with undesired genes surrounding the target gene slows the process of development of ILs. In this case, genomic marker-assisted selection (GMAS) is performed at each BC step in order to determine the introgressions and the genes of interest contained in each individual plant. From BC_1 to BC_6, GMAS allows selecting lines harboring a single introgression in a specific chromosomal region. Early BC generations in general carry several introgressions in the whole-genome regions. As backcross progresses, the size and the number of introgressions of each individual decrease. The final steps of self-

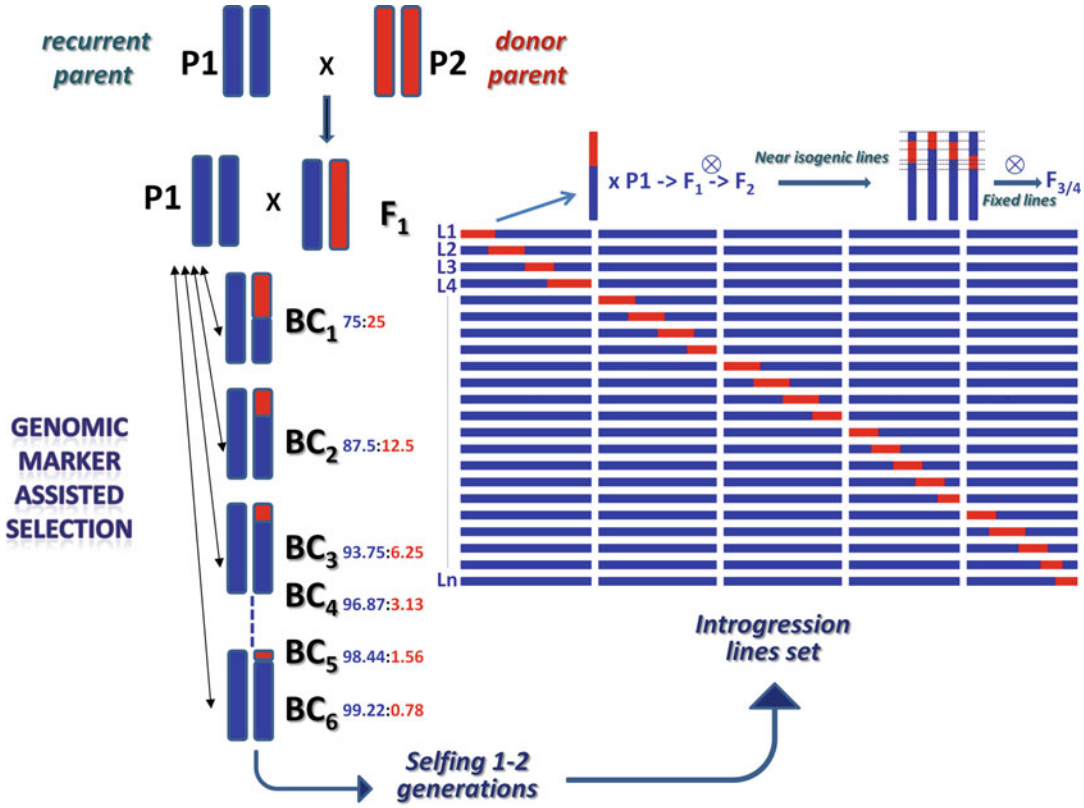


Fig. 5 Introgression line (IL) development scheme. After the cross of the two parent lines P1 and P2, the hybrid is backcrossed to the recurrent parent to produce BC progenies. Further backcrosses and marker-assisted selection allow detecting the lines with the desired introgression. At each backcross step, the proportion of the donor genome is reduced by 50%. Final self-fertilization for 1–2 cycles allows making the selected lines homozygous. The entire IL population is represented by a set of individuals with a single introgression able to cover the whole genome. Near-isogenic lines (NIL) can be developed from a single IL through backcrossing with the recurrent parent and further marker-assisted selection

fertilization allow making the selected lines homozygous. The overall target is to select a set of lines, each carrying a single donor fragment with minimum overlap, able to represent the whole alien genome. In the case of necessity to reduce the size of the introgression in order to estimate the value of a single gene or a cluster of genes, each IL can be fragmented in sub-ILs by backcrossing with the recurrent parent and marker-assisted selection of the F_2 generation [11]. Markers allow detecting the breakpoints of each introgression. The whole set of sub-IL represents the entire original IL. Several are the advantages of ILs with respect to the other mapping populations for QTL identification and for estimation of the genotypic and environmental factors underlying the variation of quantitative traits (*see* **Note 5**).

4 Notes

Below are reported the potentialities and constraints of each of the mapping population described in this chapter:

1. F_2
   (a) Advantage: Rapid and easy to develop.
   (b) Disadvantage: Few recombinations (one round); it is not an immortal population since each individual segregates. F_3 families are still highly heterozygous and cannot be replicated.
2. Double haploids (DH):
   (a) Advantage: DH lines constitute a permanent resource, to be tested in different seasons and locations.
   (b) Disadvantage: Difficult to produce since not all species are amenable; moreover, there is only a single round of recombination.
3. Recombinant inbred lines (RILs):
   (a) Advantage: RILs constitute a permanent resource, to be tested in different seasons and locations. Powerful for the estimation of the genotype × environment interaction. Within RILs it is possible to observe high recombination. Good population for QTL studies.
   (b) Disadvantage: RIL development is not possible in species that are completely self-incompatible; moreover, their development is costly and time consuming due to the several rounds of selfing required. Dominance and epistasis cannot be measured because no heterozygous lines are available.
4. Backcross inbred lines (BILs):
   (a) Advantage: All advantages already listed for RILs. The population is more addressed for breeding purposes due to the higher amount of the recurrent genome in each line. Novel varieties can be directly obtained by a few crosses and minimal step improvement. Optimal for QTL analysis using single-factor analysis. The population is suitable for fine-mapping studies. Pyramiding of gene/traits is possible thanks to the small fragment of the donor parent within each individual.
   (b) Disadvantage: Time and cost consuming for their development. Limited ability to study epistatic interactions due to high homozygosity (unless fewer cycles of selfing are done). Difficulty in studying the interaction of multiple, unlinked genes from the donor parent due to the small representation of the donor genome in each line.

5. Introgression lines (ILs):
   (a) Advantage: Permanent resources powerful and precise for detecting and analyzing qualitative and quantitative traits, to estimate G × E interaction, for functional genomics and evolutionary studies. The population is suitable for fine-mapping studies. Pyramiding of gene/traits is possible thanks to the small fragment of the donor parent within each individual. Each line can be directly used in breeding to develop novel varieties.
   (b) Disadvantages: Costly and time consuming.

References

1. D'Agostino N, Tripodi P (2017) NGS-based genotyping, high-throughput phenotyping and genome-wide association studies laid the foundations for next-generation breeding in horticultural crops. Diversity 9(3):38. https://doi.org/10.3390/d9030038
2. Moose SP, Mumm RH (2008) Molecular plant breeding as the foundation for 21st century crop improvement. Plant Physiol 147:969–977
3. Hajjar R, Hodgkin T (2007) The use of wild relatives in crop improvement: a survey of developments over the last 20 years. Euphytica 156:1–13
4. Nair KP (2019) Gene flow between cultivated plants and their wild relatives. In: Combating global warming. Springer, Cham, pp 49–52. https://doi.org/10.1007/978-3-030-23037-1_10
5. Xu Y, Li P, Yang Z, Xu C (2017) Genetic mapping of quantitative trait loci in crops. Crop J 5:175–184
6. Keating BA, Herrero M, Carberry PS, Gardner J, Cole MB (2014) Food wedges: framing the global food demand and supply challenge towards 2050. Glob Food Secur 3:125–132
7. Xu Y (2010) Molecular plant breeding. CAB International, Wallingford
8. Schnable PS, Hsia A-P, Nikolau BJ (1998) Genetic recombination in plants. Curr Opin Plant Biol 1(2):123–129. https://doi.org/10.1016/S1369-5266(98)80013-7
9. Zhang YM, Xu SZ (2004) Mapping quantitative trait loci in F_2 incorporating phenotypes of F_3 progeny. Genetics 166:1981–1993
10. Maluszynski M, Kasha KJ, Forster BP, Szarejko I (2003) Doubled haploid production in crop plants: a manual. Kluwer Academic, Dordrecht
11. Alseekh S, Ofner I, Pleban T, Tripodi P, Di Dato F, Cammareri M, Mohammad A, Grandillo S, Fernie AR, Zamir D (2013) Resolution by recombination: breaking up *Solanum pennellii* introgressions. Trends Plant Sci 18:536–538

Chapter 2

Multiparental Population in Crops: Methods of Development and Dissection of Genetic Traits

Isidore Diouf and Laura Pascual

Abstract

Multiparental populations are located midway between association mapping that relies on germplasm collections and classic linkage analysis, based upon biparental populations. They provide several key advantages such as the possibility to include a higher number of alleles and increased level of recombination with respect to biparental populations, and more equilibrated allelic frequencies than association mapping panels. Moreover, in these populations new allele's combinations arise from recombination that may reveal transgressive phenotypes and make them a useful pre-breeding material. Here we describe the strategies for working with multiparental populations, focusing on nested association mapping populations (NAM) and multiparent advanced generation intercross populations (MAGIC). We provide details from the selection of founders, population development, and characterization to the statistical methods for genetic mapping and quantitative trait detection.

Key words Genetic mapping, QTL detection, Multiparental populations (MPP), Linkage analysis, Pre-breeding populations

1 Introduction

Effective breeding strategies require knowledge about the genetic control of the target traits to improve, as well as variability to perform the selection. Unluckily, most of the key breeding traits are quantitative and present a polygenic basis complicating the identification of the causal molecular variants [1]. Traditionally, the identification of these quantitative trait loci (QTLs) has been based on linkage analysis in biparental populations such as F2, backcrosses (BC), or recombinant inbred lines (RILs) [2]. However this strategy presents two major drawbacks: it allows only the identification of QTLs segregating among two individuals, and mapping resolution is low, as it is based on small number of generations. Association mapping, based on the analysis of diverse populations, was proposed to overcome these limitations [3]. Genome-wide association studies (GWAs), based on diverse

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_2,

association panels, provide a wider range of diversity and accumulate historical recombination. In this case, the main limitations are linkage disequilibrium, which is variable along the genome, and population structure [4].

During the twenty-first century an integrated strategy has been proposed to combine both approaches, the development of experimental populations derived from the crosses between a diverse set of founders, or multiparental populations (MPP). The mating design of such populations provides a tool to control population structure and to balance allele frequencies, while the number of founders facilitates the inclusion of wider genetic diversity than in biparental populations. When identifying QTLs in such populations mapping resolution is increased, as the recombination generated during population development is coupled with the historical recombination events between the founders. Moreover, in these populations new allele's combinations arise from recombination and this may reveal transgressive phenotypes and make them a useful pre-breeding material. The research interest in MPP is clear. Actually, in the last 10 years there has been an increasing number of scientific publications related with this topic (Fig. 1).

Several designs have been proposed to develop multiparental populations from the establishment of the first inter-mated population in mice [5]. In crops two main types of MPP have been developed, nested association populations (NAM, [6]) and multiparent advanced generation intercrosses (MAGIC, [7, 8]). NAM populations were first proposed in maize (*Zea mays*) by Yu et al. [6], and include RIL families derived from the F_1s produced by

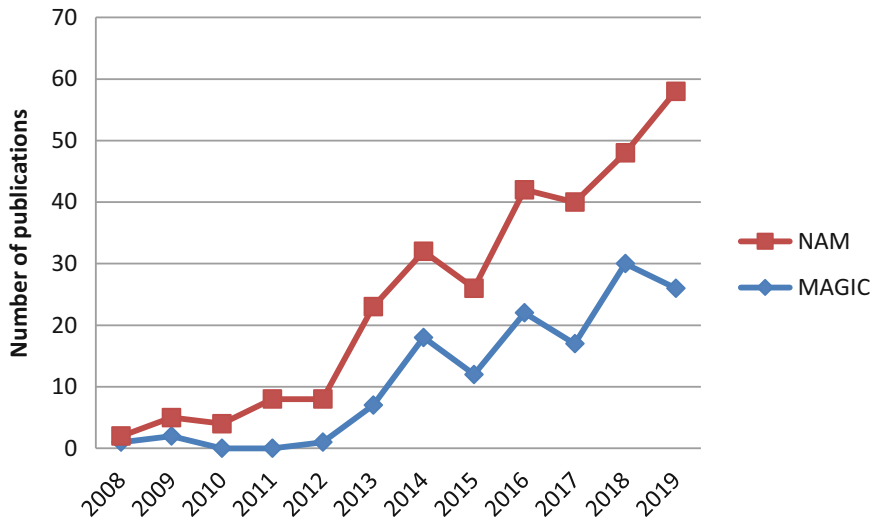


Fig. 1 Number of publications related with multiparental populations since 2008 according to Scopus. Publications contain the searched term in the title, keywords, or abstract. "Nested association mapping" in blue and "Multiparent advance generation intercross" in red

intercrossing a set of diverse founders. The preferred design is based on crossing one recurrent founder with *n* diverse founder lines, thus mirroring breeding schemes for incorporating traits in a specific background. NAM populations can be used as research and breeding tools. MAGIC populations' potential in crops was first highlighted by MacKay and Powell [7]. To construct a MAGIC population a set of founders are inter-mated, to develop derived inbred lines. The diverse population created is composed by a set of lines whose genomes are fine-scale mosaics of contributions from all founders [9], thus providing a pre-breeding population where new phenotypes arise from the combination of different backgrounds.

The potential and complementarity of MPP with respect to traditional linkage analysis and association mapping have been demonstrated in crops like tomato (*Solanum lycopersicum*) [10]. The major drawback of MPP is the greater initial investment in time and effort needed to develop such populations. Thus, it is important to pay extra attention to MPP establishment to guarantee its relevance as a long-term genetic resource. Here we describe the strategies for working with multiparental populations, focusing on NAM and MAGIC. We provide details from the selection of founders, population development, and characterization to the statistical methods for genetic mapping and quantitative trait detection.

2 NAM Population Development

Once developing a NAM population, it must be taken into account that the number and genetic diversity of founders, coupled with the mating design, will directly impact the power to detect genetic associations [11]. Always considering that phenotyping is a key step in the process limited by the total size of the population, typically composed by thousands of lines (Table 1).

2.1 Founder Selection

Regarding the number of founders, most NAM populations developed range between 10 and 50 founder lines (Table 1). The optimal number of founders should provide a balance between the number of different progenies that can be maintained and self-pollinated to develop the NAM RILs and the number of lines that will capture the available genetic diversity. There are two approaches: the first consisting on the maximization of the wide range of diversity through the following steps: (1) genetic characterization of a diverse set of accessions from which the founders will be selected; (2) selection with algorithms like Core Hunter [25] maximizing the expected proportion of heterozygous loci in the offspring (HE parameter) that can be used to increase the number

Table 1
NAM populations developed in crops. The genome or breeding tag in the mating design indicates if the common reference founder was chosen based on its uses for breeding

Species	Pop size	Mating design	RIL size	Founders	References
Maize	5000	REF-genome	200	26	Yu et al. [6]
Maize	2267	2-REF-breeding[a]	94	24	Bauer et al. [12]
Barley	1420	BC1-REF-breeding[b]	55	26	Maurer et al. [13]
Barley	796	BC2-REF-breeding[b]	32	26	Nice et al. [14]
Bread wheat	852	REF-breeding	85	11	Bajgain et al. [15]
Rice	1879	REF-breeding	187	11	Fragoso et al. [16]
Sorghum	2214	REF-breeding	220	11	Bouchet et al. [17]
Rapeseed	2425	REF-genome	161	16	Hu et al. [18]
Soybean	5600	REF-breeding	140	41	Diers et al. [19]
Bread wheat	2100	REF-breeding	75	29	Jordan et al. [20]
Maize	1257	BC1-REF-breeding[b]	210	6	Chen et al. [21]
Barley	6160	REF-breeding	69	89	Hemshrot et al. [22]
Sorghum	771	REF-breeding	257	4	Marla et al. [23]
Durum wheat	6208	REF-breeding	125	50	Kidane et al. [24]

[a]Two different NAM populations derived from crossing two different references, and crosses among the two reference lines to connect the populations
[b]Each founder was crossed with the reference, but the F1 was backcrossed once or twice with the reference before SSD

of QTLs potentially segregating; and (3) testing of different founder sizes in order to determine the point where adding more founders does not increase significantly the diversity included.

In the second approach, if the species to analyze present subspecies or ecotypes that differ greatly on flowering or ripening time, we can discard accessions that flower or ripe outside predefined intervals, in order to avoid phenotyping bias. Moreover, if it is desired to employ the developed NAM for direct breeding purposes, some of the founders may be decided based exclusively on their phenotype (e.g., when the accession presents resistance to a specific pest). In this case, we can use the Core Hunter algorithm inside a subset of accessions, or after fixing some of the accessions that will be included, to select the rest of founders. When this approach is taken the power for QTL detection may be lower.

2.2 Mating Design

The most widely used mating design for developing NAM populations mimics the one firstly proposed by Yu et al. [6], in maize (Table 1), from now on REF (Reference Design). This design offers key advantages in terms of simplicity, possibility to impute RIL genotypes, and a higher homogeneity on flowering and ripening

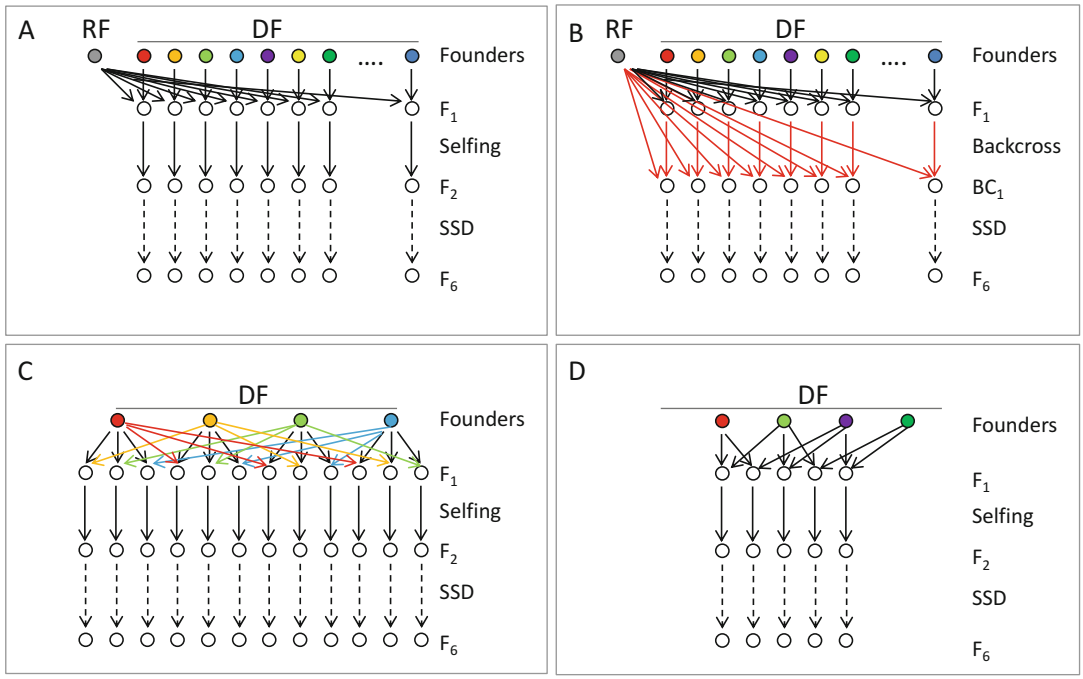


Fig. 2 Mating schemes for NAM population development. (**a**) Single reference design (REF). (**b**) Backcross reference design (BC REF). (**c**) Diallel crossing (DIA). (**d**) Double round robin design (DRR). Reference founder (RF), diverse founders (DF)

times among the different progenies [6]. In this case, it is important to determine the founder line that will be used as reference. A useful approach will be to select the line employed as reference for the species genome, so it will be fully characterized at genomic level, or to select an elite variety widely used for breeding, if the aim is to introduce the generated NAM lines directly on breeding programs.

In the REF design a set of RIL populations will be obtained from the crosses between the reference founder and the rest of diverse founders (Fig. 2a). Thus, the developed NAM population will be formed by, as many RIL families, as diverse founders selected. After the first F_1 crosses, each RIL family will be developed by SSD method (single seed descent, [26]) by self-pollination during several generations until reaching F_6 or F_7. The optimum population size will be between 100 and 200 lines by RIL family. Smaller RIL family's sizes may pose difficulty in the estimation of allele effects, and in those cases, it will not be possible to use standard analysis developed for biparental populations. In order to accelerate the development of the population it is possible to derive DH (double haploids) instead of RILs. However, it must be taken into account that this procedure will decrease the number of recombination events produced during the population development. In this case, we advise to derive the DH from F_2 plants instead of F_1, a strategy that according to Stich [11] simulations provides a good balance between time and power to detect QTLs.

Apart from the REF design, different authors have proposed alternative designs for NAM population development, like the BC-REF design (backcross reference, Fig. 2b) where the F_1s obtained are backcrossed to the reference founder before starting the RIL development [13]; this approach is recommended when the diverse founders are exotic lines. In this case, a smaller portion of the diverse founder genome is present in each line, and thus the effects of agronomically unadapted alleles are reduced, allowing estimates of the value of exotic alleles in the context of cultivated germplasm [14].

Other mating designs are based on different schemes for obtaining the F_1s that according to simulations increase the power to detect QTLs [11, 27–29]. For autogamous species the most promising design is the double round robin design (DRR), that is, crossing the founder 1 × founder 2, founder 1 × founder 3, founder 2 × founder 3, and founder 2 × founder 4, and then deriving a RIL population from each cross (Fig. 2d). For allogamous species where crosses are more easily performed, diallel (DIA) design (Fig. 2c), where all the founders are crossed with each other [30], and then a RIL population derived from each cross, performs better. The power of these designs is derived from more balanced allele frequencies and greater number of QTLs potentially segregating. However, the diverse backgrounds between each RIL may pose difficulty in the phenotypic trials. Considering all the facts, we recommend to use the proposed REF design [6], or alternatively to choose two or three references maximum and cross them with the rest of the diverse founders, to avoid phenotyping bias.

Regarding the development of RIL families once the mating design is chosen, it is important to avoid any phenotypic selection during the rounds of self-pollination. Moreover, due to the size of the population if it is not possible to perform all the lines development on the same greenhouse or field; it is advisable to develop an equal number of each RIL population lines in each of the available environments. This procedure will minimize differences between RIL populations due to inadvertent selection.

2.3 NAM Population Characterization

Once the population is constructed it will be necessary to characterize it at phenotypic and genetic levels. Regarding the genetic characterization it is important to consider that density and quality of the available genotypes will determine the power and precision when performing QTL detection.

There are two possible strategies for genotyping; the optimal one will depend on the crop genomic information available. For species without a reference genome a good option is to genotype the founders and the NAM population with high-throughput markers like RADseq (restriction-site associated DNA sequencing, [31]) or DArTseq (diversity array technology sequence, [32]) that will provide tens of thousands of polymorphic SNPs. If there is an

available genotyping array in the species that can be used, however, if the founders of the population have not been used for developing such an array, a high proportion of the markers may be monomorphic in the population.

For species with available reference genome, it is advisable to re-sequence the founders at high coverage (20–30×), as those data can be used for selecting the set of markers to genotype the population and imputing missing data in NAM RILs. For genotyping the population different approaches can be taken. Yu et al. [6] and McMullen et al. [33] proposed to select a set of SNPs for which the reference founder presents a rare allele distributed along the whole genome. The NAM lines are genotyped with these set of SNPs and those data are used for determining recombination blocks. Finally, high-density genotypic data from the founders are overlaid on the recombination blocks identified for each RIL [34]. In this approach imputation of RIL genotypes may be biased by the selected markers, and there is a risk of missing double-recombination events between the selected markers. Considering the decrease on high-throughput genotyping technologies like GBS (genotyping by sequencing [35]) we recommend to perform GBS or sequencing at low coverage (0.5×) for the full set of RILs and later to impute/overlay any missing data with the available sequence from the founders [36].

Regarding the phenotypic characterization of the NAM RILs, as in biparental QTL mapping or association panel analysis, any essay must be carried out in at least two different environments (years and/or locations). When designing the phenotyping trials the main limiting factor will be the size of the developed population that typically ranges between hundreds and thousands of lines (Table 1). Robust results will be obtained if randomized complete design (RCD) is used. Whenever the population size does not allow to phenotype the full set of lines an augmented design can be carried out; in this case founders can be used as check varieties. The final design, number of replicates, and environments to test should be determined based on the traits characterized. There are no specific guidelines for NAM populations, as experimental designs present the same advantages and disadvantage as for other types of populations where highly homozygous lines are used.

2.4 NAM Populations Developed in Crops

The first crop where a NAM population was developed was maize [6]; since then this approach has been extended to different species mainly in cereals (Table 1). This approach can be extended to a wide range of species, as proved by its success on allogamous species like maize and sorghum (*Sorghum bicolor*) [6, 21, 23] as well as on autogamous ones such as wheat species (*Triticum* spp.), barley (*Hordeum vulgare*), and rice (*Oryza sativa*) [13–16, 20, 22, 24].

NAM populations have been used to dissect a wide range of complex traits, from genes regulating recombination [20] or implicated in crop domestication [21] to key breeding traits like yield

and resistance to pests [15, 19]. Moreover, some populations have been already designed to be used as breeding materials [24]. Thus, NAM populations constitute a basic resource for genetic dissection of complex traits and breeding that can be potentially developed in any species.

3 MAGIC Population Development

The development of MAGIC populations is time consuming and requires considerable effort depending on the species, as it includes hundreds of lines (Table 2). Besides being an important resource for genetic analyses, MAGIC populations are also useful as breeding resources from which elite lines could be readily derived for the release of new varieties.

3.1 Founder Selection

Founder selection is an important process that will determine the suitability of a MAGIC population according to the pursued goals. The number of founder lines selected for the development of MAGIC population in plants varies from 4 [47] to 19 parents [40, 49]. The number of founders to select might be important for the QTL analysis, especially for models based on parental haplotype probabilities. Selection criteria may be based on different prospects in crops, encompassing geographical origins of the lines [47, 64, 65], their genetic diversity [53, 54], agronomic performance [40], disease resistance [66], tolerance to abiotic stresses [39], and crop quality requirements [67]. The geographical origin of founder lines is important for MAGIC population intended for breeding programs in specific regions. Screening genetic and phenotypic diversity of locally adapted accessions could enhance and accelerate the chance to release new varieties presenting different combinations of adapted favorable alleles. Founder selection could be based also on genotypic information, which is more and more accessible for several crops. A high number of crop accessions are conserved in national gene banks and the genetic information of such collections might help for parental line selection. Founder selection based on maximizing genetic diversity is a good strategy for developing a MAGIC population that is representative of specific collections, and can be performed as described for the NAMs. Interspecific MAGIC populations could be developed in the absence of reproductive barriers. This strategy would be beneficial to include genetic/phenotypic diversity that is not present in a single species, but efforts should be deployed to balance the different species avoiding bias due to minimum allele frequencies. Both intra- and interspecific MAGIC populations present interest for genetic analyses, though interspecific populations might be less advantageous for efficient breeding purposes due to linkage drag. Besides, segregation distortion (SD) problems may arise in interspecific crosses.

Table 2
Software and statistical programs documented for QTL mapping in MAGIC populations and case study examples of effective use

		Examples				
Software/model	**Approach**	**Species**	**Parents**	**RIL pop**	**References**	**Trait**
TASSEL [37]	BA	Cotton	11	547 S_6	Islam et al. [38]	Fiber quality
		Rice	8	200 S_4	Bandillo et al. [39]	Biotic/abiotic stress and grain quality
		Sorghum	19	200 S_7	Ongom and Ejeta [40]	Plant height
GAPIT [41]	BA	Cotton	11	547 S_6	Islam et al. [38]	Fiber quality
		Cotton	11	550	Naoumkina et al. [42]	Fiber length
ASReml [43]	BA	Maize	8	951 hybrids	Giraud et al. [44]	Silage performance
	PPA	Maize	8	951 hybrids	Giraud et al. [45]	Biomass production
HAPPY [46]	PPA	Wheat	4	1100 F_6	Huang et al. [47]	Plant height, hectoliter weight
		Arabidopsis	19	700 S_7	Gnan et al. [48]	Yield
		Arabidopsis	19	459 S_6	Kover et al. [49]	Flowering time, development
mpMap [50]	PPA	Barley	8	533 DH	Sannemann et al. [51]	Flowering time
		Wheat	8	394 $F_{6:8}$	Stadlmeier et al. [52]	Powdery mildew resistance
		Tomato	8	397 S_3	Pascual et al. [53]	Fruit weight
		Cowpea	8	305 F_8	Huynh et al. [54]	Flowering, seed size, growth habit, maturity
R/qtl2 [55]	PPA	Arabidopsis	19	374	de Jong et al. [56]	Flowering time, plant growth
MagicQTL [57]	PPA	Arabidopsis	19	426	Wei and Xu [57]	Bolt to flowering, growth rate

(continued)

Table 2
(continued)

		Examples				
Software/model	**Approach**	**Species**	**Parents**	**RIL pop**	**References**	**Trait**
MPWGAIM [58]	PPA	Wheat	4	672	Verbyla et al. [58]	Lodging
MVMPWGAIM [59]	PPA	Wheat	4	1063	Verbyla et al. [59]	Seed size, flowering time
GAPL [60]	PPA	Cowpea	8	305 F_8	Shi et al. [61]	Flowering time
FarmCPU [62]	PPA	Maize	8	332	Butrón et al. [63]	Fusarium resistance

BA biallelic approach, *PPA* parental probability approach

3.2 Mating Design

Once the parental lines have been selected, the following step will notably consist of designing the crossing scheme to deliver MAGIC recombinant inbred lines (RILs). The special feature of MAGIC populations relies on the intermating of all parental lines such that genomes of the final RILs will constitute a mosaic from the contribution of all parental genomes. The mating design will ultimately determine the population structure which ideally should be balanced, each parent roughly contributing equally to the genome of the final RILs. From the classical approach, the first level of the crossing might correspond to a half-diallel design with multiple two-way crosses of the parents where each parent is crossed to its analog. In the case of four and eight parents, a total of 6 and 28 crosses are possible, each generating a different F_1 family. The two-way F_1 are then intercrossed in one cycle to obtain four-way F_1 (Fig. 3a) and two cycles to obtain eight-way F_1. At this stage, one or two cycles of intercrossing may be added before selfing, which will increase the recombination; however it would extend the time for development. After this highly homozygous lines will be developed to establish the MAGIC population. SSD and DH are both methods used in plants to generate highly homozygous lines. Comparatively, SSD may present some advantages over DH as it allows additional recombination events. Selfing progenies with SSD generate however a residual heterozygosity and the final RILs are not fully homozygous at all loci. Seven cycles of SSD selfing resulted in 305 F_8 lines in cowpea (*Vigna unguiculata*) presenting 99% of homozygosity [54]. In tomato, with three generations of SSD, Pascual et al. [53] obtained 12% of residual heterozygosity. Although it might give rise to problems on genotyping, residual heterozygosity may be beneficial for the creation of heterogeneous inbred families (HIFs) and subsequent analyses such as fine mapping regions of interest [68].

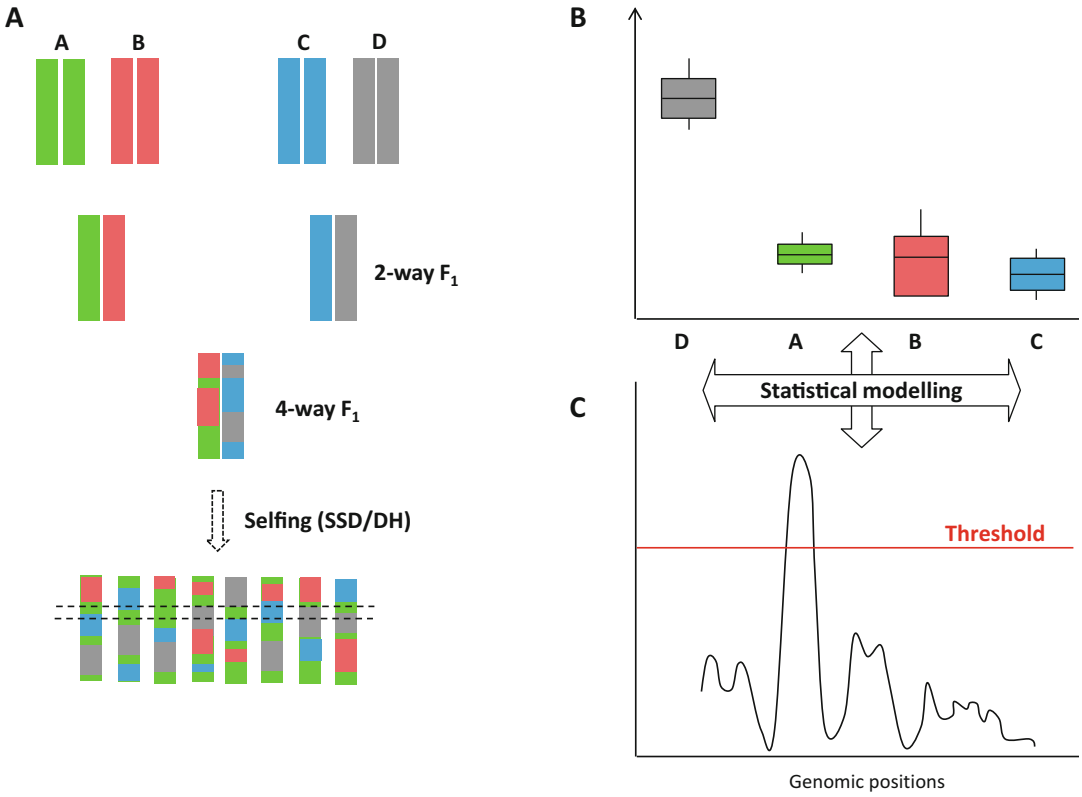


Fig. 3 Principles of the parental probability approach for QTL mapping in MAGIC populations. (**a**) Crossing scheme for the development of a four-way MAGIC population. (**b**) Phenotypic distribution of the MAGIC RILs grouped according to their haplotype status at a given region. (**c**) The LOD score of the QTL derived along the whole-genome scan through simple regression or mixed linear models

3.3 MAGIC Population Characterization

MAGIC populations are valuable resources for marker-trait association analysis and mapping of quantitative trait loci and genes of economically breeding importance. Such analyses though require an adequate characterization of the genetic and phenotypic variation of a whole set of individuals forming the mapping population. Genotyping and phenotyping of MAGIC populations are amenable after sufficient cycles of selfing (five to eight cycles of SSD) to obtain nearly homozygous lines at most loci. Genotyping is usually conducted after selection of a set of relevant molecular markers that are able to discriminate the origin of the recombination blocks in each MAGIC line. The improvement in sequencing technologies allowed successful completion of genome sequence in diverse plant species [69] facilitating the identification of polymorphism variation between individuals. Thus, before selecting the markers it is recommended to conduct whole-genome resequencing of MAGIC parental lines when a reference genome is available. Marker selection should then be based on SNP quality, their frequent distribution along the genome, and the allelic profile over the parents to

avoid redundant information. Besides the advantage of low-cost genotyping due to an efficient selection of the marker set, such a strategy also presents advantages for QTL mapping and candidate gene identification [53, 70]. For species with large genome size or without a reference, genome-wide characterization can be realized directly through GBS to deliver a large set of SNP polymorphisms [31]. Precision in phenotyping is another important aspect to consider for reliable marker-trait association analyses. If conducted at the whole population level, it can lead to high resolution of QTL mapping. However, the large number of RILs generated when creating MAGIC populations may impede phenotyping at the whole population level. Thus, a selection of a subset of the population to an appropriate number for the experimental facilities is required. Using only a subset of the population for phenotypic measurements could reduce the power and precision of QTL detection. It is therefore important to efficiently select the population subset. Lines for phenotyping could be randomly selected or selection could be based on the genetic distance between lines to optimize the representative diversity of a population.

MAGIC populations present the particularity of being immortal populations for which genotype by environment interaction (G × E) could be assessed. Phenotyping can be conducted several times in different locations, years, or cultural conditions. This can be used to identify genetic determinants of G × E and for the selection of superior lines to include in breeding programs. For G × E analyses, attention should be paid on MAGIC line selection to be sure that the same set of lines is tested in different environments.

3.4 MAGIC Populations Developed in Crops

The MAGIC populations enclose large genetic diversity with different combinations of parental alleles generally leading to transgressive segregation. They constitute breeding resources from which superior lines could be selected toward new variety creation. From an S_2:bulk in MAGIC *indica* rice population, 400 lines have been selected on the basis of agronomic traits and then tested in mega-environment trials toward the selection of elite lines according to targeted environments [39]. MAGIC rice lines have been also included in multi-environment trial assays and currently some lines are under selection for tolerance to salinity and nutrient deficiency [71]. In cotton (*Gossypium hirsutum*), marker-assisted selection has been applied for fiber quality. The strategy was based on best allelic combination at different SNP markers that were strongly associated to fiber length and other fiber quality traits [38]. The development of advanced lines combining positive alleles at different genomic regions or for different traits is an appealing strategy for efficient breeding. A promising method called MAGReS (for multiparent advanced generation recurrent selection) has been described [9]. This strategy requires first the

identification of significant QTL for traits of interest and parents carrying positive alleles. The following step will consist of the efficient selection of MAGIC lines with the best alleles and design of crossing scheme between those lines. Valente et al. [72] developed a decision support tool they called OptiMAS that is intended to identify the optimal crossing scheme to combine best alleles in offspring lines and which is effective for multiparental populations. Genomic selection based on single SNP or haplotype is also another approach for breeding in MAGIC populations that could be applied for highly polygenic traits or for multi-trait selection criteria.

4 Dissection of Quantitative Traits: Analysis Methods

4.1 Constructing Genetic Maps on MAGICs

The complex design of MAGIC populations causes several challenges for the construction of genetic linkage maps for marker-trait association analyses. Although MAGIC populations are importantly developed in crops and used for genetic analyses, few software are yet available for the construction of genetic maps. R/mpMap is a commonly used package for the development of marker map in MAGIC populations [50]. This package is implemented in R [73] and is particularly adapted to MAGIC populations derived from four and eight parents. Genetic maps have been developed from MAGIC populations with mpMap package, outperforming other biparental genetic maps in wheat [74] and tomato [53]. Further improvement was brought by a new version of the package R/MpMap2 which is extended to biparental and 16-way MAGIC populations with less computational time for large population size and marker number [75]. Recently, Zheng et al. [76] provided magicMap, another program for the construction of genetic map in different population types which go beyond $2n$ MAGIC populations and also include other multiparental populations. Furthermore, the magicMap program allows missing genotypes in founders and explicit modeling of heterozygotes for genetic map construction.

4.2 QTL Analysis on MAGIC Populations

Although MAGIC populations present multiple advantages, application of suitable statistical models is required for genetic analyses in QTL mapping. A review of different statistical modeling for quantitative analyses of MAGIC populations was conducted by Huang et al. [9]. Two main approaches can be used for QTL mapping in MAGIC populations, biallelic approach (BA) lying in the use of marker scores such as in biparental populations or parental probability approach (PPA), where the probabilities of parental allele affiliation are inferred at each marker position (Fig. 3). Recently, Ogawa et al. [77] proposed another approach based on the use of haplotype blocks that has not yet been largely explored in

MAGIC populations and which could be an intermediate between the BA and PPA. Mott et al. [46] highlighted the limitations of BA strategy due to the fact that not all information of the parental polymorphisms is accounted. The limits with the BA strategy are well illustrated by Verbyla et al. [58]. Methods for reconstruction of the parental probabilities have been proposed [50] and attested efficient in different analyses with real data [47, 53]. The R/mpMap package developed by Huang and George [50] allows mapping QTLs through simple linear regression model where QTL effects are estimated for every parent at each marker position where parental probabilities are computed. On the basis of parental probabilities, Broman et al. [55] proposed new methods adapted for more complex designs of MAGIC populations and implemented them in R/qtl2 package. This package offers possibility to use regression models based on the Haley-Knott regression method [78] or to perform genome scan with a linear mixed model which could account for polygenic residual variance by modeling genetic relatedness between individuals. Previous studies had already proposed the use of linear mixed model in MAGIC populations which modeled also the polygenic variance and allowed for detection of more than one QTL per chromosome or linkage group [57, 59]. Mixed models are highly powerful for MAGIC populations because of the flexibility they offer for modeling any complexity from the mating design. In maize for instance, Giraud et al. [44] applied mixed model for QTL analysis and considered the structure from the parental belonging to different heterotic groups. Other software and linear mixed models efficiently used for QTL mapping analyses in MAGIC populations have been documented and summarized in Table 2.

4.3 QTL Analysis on NAM Populations

Individual NAM families do not differ from classical RIL mapping populations; thus if each family includes a high number of lines theoretically they could be analyzed separately, and then the results are integrated. This approach is not recommended, as it will be missing the link of the common founder (when REF mating is employed), and reducing the population size to family size, losing statistical power to identify QTLs.

In other to avoid these problems, some studies perform methods used for GWAs in association panels with the full set of NAM lines, including a kinship matrix as cofactor [14]. However, in this case recombination information and linkage phase are not used, and thus the method does not provide a specific position for the QTL essential for map-based cloning.

Specific procedures designed for NAM populations will increase the power to detect QTLs [79]. Here we will briefly describe different methods and the software in which they have been implemented. In order to choose the best method, we recommend taking into account free available software, as well as, the required bioinformatics skills to perform the analysis.

A specific NAM population procedure based on MLM (mixed linear model, [80]) is the one implemented on the package NAM R [81]. To account for the particular structure of a NAM population the marker alleles are recoded to work with haplotypes and a genomic relation matrix is built (GRM). Based on this, MLM is applied and then a Bayes approach is used to refit the model, which is evaluated with the likelihood ratio test to call QTLs (described in detail in [81]). This method solves part of the problem and performs better than typical MLM methods as it takes into account the linkage phase. Besides, it is implemented on a free R package, facilitating the analysis. However, it does not take into account the recombination information and does not provide a genetic position for the QTL.

One of the most used procedures is joint inclusive composite interval mapping (JICIM, [79]). This method has been used for the maize NAM population [82] and according to simulations it is able to detect a QTL within 1cM with an 85% chance, when the QTL overlapped with markers [79]. It is based on a two-step approach; first general linear model (GLM) is used, employing stepwise regression to estimate the parameters in the model. Second, based on the parameter estimation a scanning similar to the second step of ICIM (inclusive composite interval mapping, [83]) is run, to determine the position of the QTLs (described in detail in [79]). Thus, the method requires a joint linkage map which can be constructed with software like magicMap [76] or Icimapping [84].

The method can be run as described by Chen et al. [21]. First, running the stepwise linear regression fixed model implemented in the PROC GLMSELECT procedure in SAS software version 9.3 [85] and second, conducting a one-dimensional scanning to determine QTL position and confidence interval (this approach requires a SAS software license). An alternative is the procedure described by Jordan et al. [20] using TASSEL v.5.2.42 for the first step, and the ICIM v 4.1.0.0 program taking into account the nested family effect for the second step. However, we recommend the Icimapping software [84]. Its version 4.2 (http://www.isbreeding.net/) released on 2019-07-25 is freely available under registration. Icimapping runs under Windows 10 and presents a user-friendly graphical interface. The software includes the tools to integrate individual linkage maps for each NAM family into a joint linkage map and performs QTL detection with the JICIM method.

Finally, Bian and Holland [86] proposed a new algorithm TAGGING (thinning and aggregating) based on ensemble learning [87] to solve the problem of collinearity among markers that arises from the high-throughput genotyping technologies used nowadays. This method consists of thinning dense marker maps into a set of smaller maps, predicting QTLs in each of the smaller

maps, and later aggregating the predictions. The potential of this method for joint family QTL detection is demonstrated by applying specific procedures for the QTL prediction part (described in detail in [86]). However, this procedure is not implemented in user-friendly software hampering its application.

5 Conclusions

Both NAM and MAGIC populations have the advantage of accumulating higher diversity compared to classical biparental populations. Their development is time consuming and the founder line selection combined with the different possibilities for crossing schemes offer high flexibility for population creation. Thus all the aspects described in this chapter, from the founder selection to the available analysis tools, should be considered before starting, in order to create a durable and useful genetic resource for research and/or breeding. Actually, the first choice will be between NAM and MAGIC population. As conclusion here we present a summary of the main differences between both resources highlighting their advantages.

1. MAGIC populations require smaller sizes than NAM to achieve comparable resolution, as mixing all the founder backgrounds allows the detection of more recombination breakpoints.
2. MAGIC populations present more allelic combinations; however it is difficult to know how a specific allele will perform in an elite background.
3. NAM populations allow the inclusion of wider genetic diversity (more founders), without increasing the development time. However in MAGIC populations the inclusion of higher number of founders requires extra generations.
4. NAM populations can be easily extended by adding new RIL families.
5. The haplotype assignation is more straightforward in NAM, as the lines present only two possibilities.
6. NAM lines, when developed using as a reference an elite cultivar, already contain 50% of elite genetic background, facilitating the inclusion in breeding programs.

Examples of direct application of multiparental populations for breeding have been described in the sections above. According to breeding goals, the identification of promising lines from the allelic variation at significantly detected QTLs is a strategy that might be applied in both NAM and MAGIC. For polygenic characters and multi-trait selection criteria, genomic selection in multiparental populations is a promising strategy. However, few studies have

investigated the effective application of genomic prediction model in multiparental populations. Lehermeier et al. [88] presented some guidelines regarding the genetic structure and required sample size of data sets for model training that can be used as guidelines to fully exploit multiparental populations.

References

1. Morrell PL, Buckler ES, Ross-Ibarra J (2012) Crop genomics: advances and applications. Nat Rev Genet 13:85–96
2. Price AH (2006) Believe it or not, QTLs are accurate! Trends Plant Sci 11:213–216
3. Korte A, Farlow A (2013) The advantages and limitations of trait analysis with GWAS: a review. Plant Methods 9:29
4. Visscher PM, Brown MA, McCarthy MI, Yang J (2012) Five years of GWAS discovery. Am J Hum Genet 90:7–24
5. Churchill G, Airey DC, Allayee H, Angel JM, Attie AD, Beatty J et al (2004) The Collaborative Cross, a community resource for the genetic analysis of complex traits. Nat Genet 36:1133–1137
6. Yu J, Holland JB, McMullen MD, Buckler ES (2008) Genetic design and statistical power of nested association mapping in maize. Genetics 178:539–551
7. Mackay I, Powell W (2007) Methods for linkage disequilibrium mapping in crops. Trends Plant Sci 12:57–63
8. Cavanagh C, Morell M, Mackay I, Powell W (2008) From mutations to MAGIC: resources for gene discovery, validation and delivery in crop plants. Curr Opin Plant Biol 11:215–221
9. Huang BE, Verbyla KL, Verbyla AP, Raghavan C, Singh VK, Gaur P et al (2015) MAGIC populations in crops: current status and future prospects. Theor Appl Genet 128:999–1017
10. Pascual L, Albert E, Sauvage C, Duangjit J, Bouchet J-P, Bitton F et al (2016) Dissecting quantitative trait variation in the resequencing era: complementarity of bi-parental, multi-parental and association panels. Plant Sci 242:120–130
11. Stich B (2009) Comparison of mating designs for establishing nested association mapping populations in maize and Arabidopsis thaliana. Genetics 183:1525–1534
12. Bauer E, Falque M, Walter H et al (2013) Intraspecific variation of recombination rate in maize. Genome Biol 14:R103
13. Maurer A, Draba V, Jiang Y, Schnaithmann F, Sharma R, Schumann E et al (2015) Modelling the genetic architecture of flowering time control in barley through nested association mapping. BMC Genomics 16(1):290
14. Nice LM, Steffenson BJ, Brown-Guedira GL, Akhunov ED, Liu C, Kono TJY et al (2016) Development and genetic characterization of an advanced backcross-nested association mapping (AB-NAM) population of wild × cultivated barley. Genetics 203:1453
15. Bajgain P, Rouse MN, Tsilo TJ, Macharia GK, Bhavani S, Jin Y et al (2016) Nested association mapping of stem rust resistance in wheat using genotyping by sequencing. PLoS One 11: e0155760
16. Fragoso CA, Moreno M, Wang Z, Heffelfinger C, Arbelaez LJ, Aguirre JA et al (2017) Genetic architecture of a rice nested association mapping population. G3 7:1913–1926
17. Bouchet S, Olatoye MO, Marla SR, Perumal R, Tesso T, Yu J et al (2017) Increased power to dissect adaptive traits in global Sorghum diversity using a nested association mapping population. Genetics 206:573–585
18. Hu J, Guo C, Wang B, Ye J, Liu M, Wu Z et al (2018) Genetic properties of a nested association mapping population constructed with semi-winter and spring oilseed rapes. Front Plant Sci 9:1740
19. Diers BW, Specht J, Rainey KM, Cregan P, Song Q, Ramasubramanian V et al (2018) Genetic architecture of soybean yield and agronomic traits. G3 8:3367–3375
20. Jordan KW, Wang S, He F, Chao S, Lun Y, Paux E et al (2018) The genetic architecture of genome-wide recombination rate variation in allopolyploid wheat revealed by nested association mapping. Plant J 95:1039–1054
21. Chen Q, Yang CJ, York AM, Xue W, Daskalska LL, DeValk CA et al (2019) TeoNAM: a nested association mapping population for domestication and agronomic trait analysis in maize. Genetics 213:1065–1078

22. Hemshrot A, Poets AM, Tyagi P, Lei L, Carter CK, Hirsch CN et al (2019) Development of a multiparent population for genetic mapping and allele discovery in six-row barley. Genetics 213:595–613
23. Marla SR, Burow G, Chopra R, Hayes C, Olatoye MO, Felderhoff T et al (2019) Genetic architecture of chilling tolerance in Sorghum dissected with a nested association mapping population. G3 9:4045–4057
24. Kidane YG, Gesesse CA, Hailemariam BN, Desta EA, Mengistu DK, Fadda C et al (2019) A large nested association mapping population for breeding and quantitative trait locus mapping in Ethiopian durum wheat. Plant Biotechnol J 17:1380–1393
25. Thachuk C, Crossa J, Franco J, Dreisigacker S, Warburton M, Davenport GF (2009) Core Hunter: an algorithm for sampling genetic resources based on multiple genetic measure. BMC Bioinformatics 10:243
26. Knott DR, Kumar J (1975) Comparison of early generation yield testing and a single seed descent procedure in wheat breeding. Crop Sci 15:295–299
27. Guo B, Sleper DA, Beavis WD (2010) Nested association mapping for identification of functional markers. Genetics 186:373–383
28. Klasen JR, Piepho HP, Stich B (2012) QTL detection power of multi-parental RIL populations in Arabidopsis thaliana. Heredity 108:626–632
29. Li J, Bus A, Spamer V, Stich B (2016) Comparison of statistical models for nested association mapping in rapeseed (Brassica napus eL.) through computer simulations. BMC Plant Biol 16:26
30. Griffing B (1956) Concept of general and specific combining ability in relation to diallel crossing systems. Aust J Biol Sci 9:463–493
31. Poland JA, Brown PJ, Sorrells ME, Jannink J-L (2012) Development of high-density genetic maps for barley and wheat using a novel two-enzyme genotyping-by-sequencing approach. PLoS One 7:e32253
32. Sansaloni C, Petroli C, Jaccoud D et al (2011) Diversity Arrays Technology (DArT) and next-generation sequencing combined: genome-wide, high throughput, highly informative genotyping for molecular breeding of Eucalyptus. BMC Proc 5:P54
33. McMullen MD, Kresovich S, Villeda HS, Bradbury P, Li H, Sun Q et al (2009) Genetic properties of the maize nested association mapping population. Science 325:737–740
34. Guo B, Beavis WD (2011) In silico genotyping of the maize nested association mapping population. Mol Breed 27:107–113
35. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e1937
36. Zan Y, Payen T, Lillie M, Honaker CF, Siegel PB, Carlborg O (2019) Genotyping by low-coverage whole-genome sequencing in intercross pedigrees from outbred founders: a cost-efficient approach. Genet Select Evol 51:44
37. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635
38. Islam MS, Thyssen GN, Jenkins JN, Zeng L, Delhom CD, McCarty JC et al (2016) A MAGIC population-based genome-wide association study reveals functional association of GhRBB1_A07 gene with superior fiber quality in cotton. BMC Genomics 17. https://doi.org/10.1186/s12864-016-3249-2
39. Bandillo N, Raghavan C, Muyco PA, Sevilla MAL, Lobina IT, Dilla-Ermita CJ et al (2013) Multi-parent advanced generation inter-cross (MAGIC) populations in rice: progress and potential for genetics research and breeding. Rice 6:11
40. Ongom PO, Ejeta G (2018) Mating design and genetic structure of a multi-parent advanced generation intercross (MAGIC) population of Sorghum (Sorghum bicolor (L.) Moench). G3 8:331–341
41. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ et al (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–2399
42. Naoumkina M, Thyssen GN, Fang DD, Jenkins JN, McCarty JC, Florane CB (2019) Genetic and transcriptomic dissection of the fiber length trait from a cotton (Gossypium hirsutum L) MAGIC population. BMC Genomics 20:112
43. Butler D, Cullis B, Gilmour A, Gogel B (2007) ASRemlR reference manual. State of Queensland Department of Primary Industries and Fisheries
44. Giraud H, Bauland C, Falque M, Madur D, Combes V, Jamin P et al (2017) Linkage analysis and association mapping QTL detection models for hybrids between multiparental

population from two heterotic groups: application to biomass production in maize (Zea mays L.). G3 7:3649–3657

45. Giraud H, Bauland C, Falque M, Madur D, Combes V, Jamin P et al (2017) Reciprocal genetics: identifying QTL for general and specific combining abilities in hybrids between multiparental populations from two maize (Zea mays L.) heterotic groups. Genetics 207:1167–1180
46. Mott R, Talbot CJ, Turri MG, Collins AC, Flint J (2000) A method for fine mapping quantitative trait loci in outbred animal stocks. Proc Natl Acad Sci U S A 97:12649–12654
47. Huang BE, George AW, Forrest KL, Kilian A, Hayden MJ, Morell MK et al (2012) A multiparent advanced generation inter-cross population for genetic analysis in wheat. Plant Biotechnol J 10:826–839
48. Gnan S, Priest A, Kover PX (2014) The genetic basis of natural variation in seed size and seed number and their trade-off using Arabidopsis thaliana MAGIC lines. Genetics 198:1751
49. Kover PX, Valdar W, Trakalo J, Scarcelli N, Ehrenreich IM, Purugganan MD et al (2009) A multiparent advanced generation inter-cross to fine-map quantitative traits in Arabidopsis thaliana. PLoS Genet 5:e1000551
50. Huang BE, George AW (2011) R/mpMap: a computational platform for the genetic analysis of multiparent recombinant inbred lines. Bioinformatics 27:727–729
51. Sannemann W, Huang BE, Mathew B, Leon J (2015) Multi-parent advanced generation inter-cross in barley: high-resolution quantitative trait locus mapping for flowering time as a proof of concept. Mol Breed 35:86
52. Stadlmeier M, Hartl L, Mohler V (2018) Usefulness of a multiparent advanced generation intercross population with a greatly reduced mating design for genetic studies in winter wheat. Front Plant Sci 9:1825
53. Pascual L, Desplat N, Huang BE, Desgroux A, Bruguier L, Bouchet J-P et al (2015) Potential of a tomato MAGIC population to decipher the genetic control of quantitative traits and detect causal variants in the resequencing era. Plant Biotechnol J 13:565–577
54. Huynh B-L, Ehlers JD, Huang BE, Munoz-Amatriain M, Lonardi S, Santos JRP et al (2018) A multi-parent advanced generation inter-cross (MAGIC) population for genetic analysis and improvement of cowpea (Vigna unguiculata L. Walp.). Plant J 93:1129–1142
55. Broman KW, Gatti DM, Simecek P, Furlotte NA, Prins P, Sen S et al (2019) R/qtl2: software for mapping quantitative trait loci with high-dimensional data and multiparent populations. Genetics 211:495–502
56. de Jong M, Tavares H, Pasam RK, Butler R, Ward S, George G et al (2019) Natural variation in Arabidopsis shoot branching plasticity in response to nitrate supply affects fitness. PLoS Genet 15:e100836
57. Wei J, Xu S (2016) A random-model approach to QTL mapping in multiparent advanced generation intercross (MAGIC) populations. Genetics 202:471
58. Verbyla AP, George AW, Cavanagh CR, Verbyla KL (2014) Whole-genome QTL analysis for MAGIC. Theor Appl Genet 127:1753–1770
59. Verbyla AP, Cavanagh CR, Verbyla KL (2014) Whole-genome analysis of multienvironment or multitrait QTL in MAGIC. G3 4:1569–1584
60. Zhang L, Meng L, Wang J (2019) Linkage analysis and integrated software GAPL for pure-line populations derived from four-way and eight-way crosses. Crop J 7:283–293
61. Shi J, Wang J, Zhang L (2019) Genetic mapping with background control for quantitative trait locus (QTL) in 8-parental pure-line populations. J Hered 110:880–891
62. Liu X, Huang M, Fan B, Buckler ES, Zhang Z (2016) Iterative usage of fixed and random effect models for powerful and efficient genome-wide association studies. PLoS Genet 12:e1005767
63. Butron A, Santiago R, Cao A, Samayoa LF, Malvar RA (2019) QTLs for resistance to Fusarium ear rot in a multiparent advanced generation intercross (MAGIC) maize population. Plant Dis 103:897–904
64. Mackay IJ, Bansept-Basler P, Barber T, Bentley AR, Cockram J, Gosman N et al (2014) An eight-parent multiparent advanced generation inter-cross population for winter-sown wheat: creation, properties, and validation. G3 4:1603–1610
65. Sallam A, Martsch R (2015) Association mapping for frost tolerance using multi-parent advanced generation inter-cross (MAGIC) population in faba bean (Vicia faba L.). Genetica 143:501–514
66. Campanelli G, Sestili S, Acciarri N, Montemurro F, Palma D, Leteo F et al (2019) Multi-parental advances generation inter-cross population, to develop organic tomato genotypes by participatory plant breeding. Agronomy 9:119
67. Meng L, Zhao X, Ponce K, Ye G, Leung H (2016) QTL mapping for agronomic traits using multi-parent advanced generation inter-

cross (MAGIC) populations derived from diverse elite indica rice lines. Field Crops Res 189:19–42

68. Tuinstra MR, Ejeta G, Goldsbrough PB (1997) Heterogeneous inbred family (HIF) analysis: a method for developing near-isogenic lines that differ at quantitative trait loci. Theor Appl Genet 95:1005–1011
69. Goodwin S, McPherson JD, McCombie WR (2016) Coming of age: ten years of next-generation sequencing technologies. Nat Rev Genet 17:333–351
70. Diouf IA, Derivot L, Bitton F, Pascual L, Causse M (2018) Water deficit and salinity stress reveal many specific QTL for plant growth and fruit quality traits in tomato. Front Plant Sci 9:279
71. Ponce KS, Ye G, Zhao X (2018) QTL identification for cooking and eating quality in indica rice using multi-parent advanced generation intercross (MAGIC) population. Front Plant Sci 9:868
72. Valente F, Gauthier F, Bardol N, Blanc G, Joets J, Charcosset A et al (2014) OptiMAS: a decision support tool to conduct marker-assisted selection programs. Crop Breed Methods Protoc 1145:97–116
73. R Core Team (2019) R: a language and environment for statistical computing. In: R Foundation for Statistical Computing. Available via DIALOG. https://www.R-project.org/. Accessed 31 Jan 2020
74. Gardner KA, Wittern LM, Mackay IJ (2016) A highly recombined, high-density, eight-founder wheat MAGIC map reveals extensive segregation distortion and genomic locations of introgression segments. Plant Biotechnol J 14:1406–1417
75. Shah R, Huang E (2019) Map construction using multi-parent populations (Version v0.0.6). In: Zenodo. Available via DIALOG. https://doi.org/10.5281/zenodo.2613114. Accessed 31 Jan 2020
76. Zheng C, Boer MP, van Eeuwijk FA (2019) Construction of genetic linkage maps in multi-parental populations. Genetics 212:1031–1044
77. Ogawa D, Yamamoto E, Ohtani T, Kanno N, Tsunematsu H, Nonoue Y et al (2018) Haplotype-based allele mining in the Japan-MAGIC rice population. Sci Rep 8:4379
78. Haley CS, Knott SA (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. Heredity 69:315–324
79. Li H, Bradbury P, Ersoz E, Buckler ES, Wang J (2011) Joint QTL linkage mapping for multiple-cross mating design sharing one common parent. PLoS One 6:e17573
80. Kang HM, Sul JH, Service SK, Zaitlen NA, Kong S-Y, Freimer NB et al (2010) Variance component model to account for sample structure in genome-wide association studies. Nat Genet 42:348–U110
81. Xavier A, Xu S, Muir WM, Rainey KM (2015) NAM: association studies in multiple populations. Bioinformatics 31:3862–3864
82. Buckler ES, Holland JB, Bradbury PJ, Acharya CB, Brown PJ, Browne C et al (2009) The genetic architecture of maize flowering time. Science 325:714–718
83. Li H, Ye G, Wang J (2007) A modified algorithm for the improvement of composite interval mapping. Genetics 175:361–374
84. Meng L, Li H, Zhang L, Wang J (2015) QTL IciMapping: integrated software for genetic linkage map construction and quantitative trait locus mapping in biparental populations. Crop J 3:269–283
85. SAS Institute (2011) SAS/STAT 9.3 user's guide. SAS Institute Inc, Cary, NC, USA. Available via DIALOG. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.226.6407&rep=rep1&type=pdf. Accessed 31 Jan 2020
86. Bian Y, Holland JB (2015) Ensemble learning of QTL models improves prediction of complex traits. G3 5:2073–2084
87. Dietterich TG (2000) Ensemble methods in machine learning. In: International workshop on multiple classifier systems. Springer, Berlin, pp 1–15
88. Lehermeier C, Kramer N, Bauer E et al (2014) Usefulness of multiparental populations of maize (Zea mays L.) for genome-based prediction. Genetics 198:3–16

Chapter 3

Using Metabolomics to Assist Plant Breeding

Saleh Alseekh and Alisdair R. Fernie

Abstract

Recent methodological advances in both gas chromatography-mass spectrometry (GC-MS) and liquid chromatography-mass spectrometry (LC-MS) have provided a deep understanding of metabolic regulation occurring in plant cells. The application of these techniques to agricultural systems is, however, subject to more complex interactions. Here we summarize a step-by-step modern metabolomics methodology that generates metabolome data toward the implementation of metabolomics in crop breeding. We describe a metabolic workflow, and provide guidelines for handling large sample numbers for the specific purpose of metabolic quantitative trait loci approaches.

Key words Metabolomics, Natural genetic variation, QTL mapping, Crop improvement

1 Introduction

The metabolites of the plant kingdom are extremely diverse; a commonly quoted estimate is that plants produce somewhere in the order of 200,000 unique chemical structures [1, 2]. Recently, there has been an increasing use of analytical technologies such as metabolomics for the comprehensive profiling of metabolites in biological samples and its subsequent application in several related research areas such as human nutrition, drug discovery, and plant breeding [3, 4]. Given the diversity of structural classes of metabolites, ranging from primary metabolites such as carbohydrates, amino acids, and organic acids to very complex secondary metabolites such as phenolics, alkaloids, and terpenoids, there is no single methodology that can measure the complete metabolome in one step. It is, therefore, often necessary to combine different techniques to detect (even a significant proportion of) all metabolites within a complex mixture [5]. Both gas chromatography-mass spectrometry (GC-MS) and liquid chromatography-mass spectrometry (LC-MS) have been intensively used to profile a broad natural variance in the form of recombinant inbred lines (RILs), introgression lines (ILs), and, more recently, genome-wide

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264,
https://doi.org/10.1007/978-1-0716-1201-9_3,

association mapping panels in order to boost our understanding of the regulation of plant primary and secondary metabolite levels [6–8].

Harnessing the vast genetic potential that exists in wild exotic species and modern crop elite varieties for plant breeding requires the establishment of rapid, predictive tools and concepts to understand the mechanistic basis for traits and to associate traits with genomic or other diagnostic information [9]. This first step enables subsequent crop improvement by breeding and selection, using the diagnostic information to guide plant breeding to combine key traits in improved varieties. Large-scale germplasm enhancement programs work to develop techniques to associate markers with phenotypes impacting crop quality and phenotyping. In general, phenotyping of large set of germplasm and natural population requires a time- and cost-intensive process. Current technologies such as genetic markers allow marker-assisted selection [10]. However, recently, metabolomics has emerged as a highly promising approach for prediction of a variety of agronomically important phenotypes of crop plants and particularly for discovering of signature metabolites for traits of interest [11]. Therefore, metabolomics has emerged and been proposed to display promising prospects to accelerate the selection of improved breeding materials and screen a wide range of crop varieties [12, 13]. Integration of metabolomics with modern plant genomics tools, such as genotype-based sequencing (GBS), genome-wide genetic variants, and whole-genome sequencing, opens further exciting horizons for crop improvement [14]. Metabolomics is direct, not dependent on genotyping, and addresses the features that are directly relevant to biological function and thus to plant phenotype and agronomic traits [15]. For instance, metabolomics has been used to predict phenotypic performance in model species, such as *Arabidopsis thaliana* [16] and crop species such as tomato (*Solanum lycopersicum*) [4, 17], maize (*Zea mays*) [18, 19], wheat (*Triticum* spp.) [20, 21], barley (*Hordeum vulgare*) [22, 23], and rice (*Oryza sativa*) [24, 25]. This allows us to reduce the gap between phenotype and genotype and leads to precision breeding [26].

Application of metabolomic platforms in plant breeding programs has several challenges. In contrast to genetic markers, metabolomics is dependent on metabolite composition which is known to be highly influenced by environmental and experimental variation; this fact renders experimental design and sample preparation critical. Therefore, it is important to understand and control factors that contribute to sources of variation within the data sets. This includes the variability during sample collection, preparation, and storage [27, 28]. In addition, analytical variation caused by suboptimal performance of the chosen apparatus and instrument drift over time are major issues in large-scale metabolomics studies [29]. While there is no single best way to conduct metabolomic

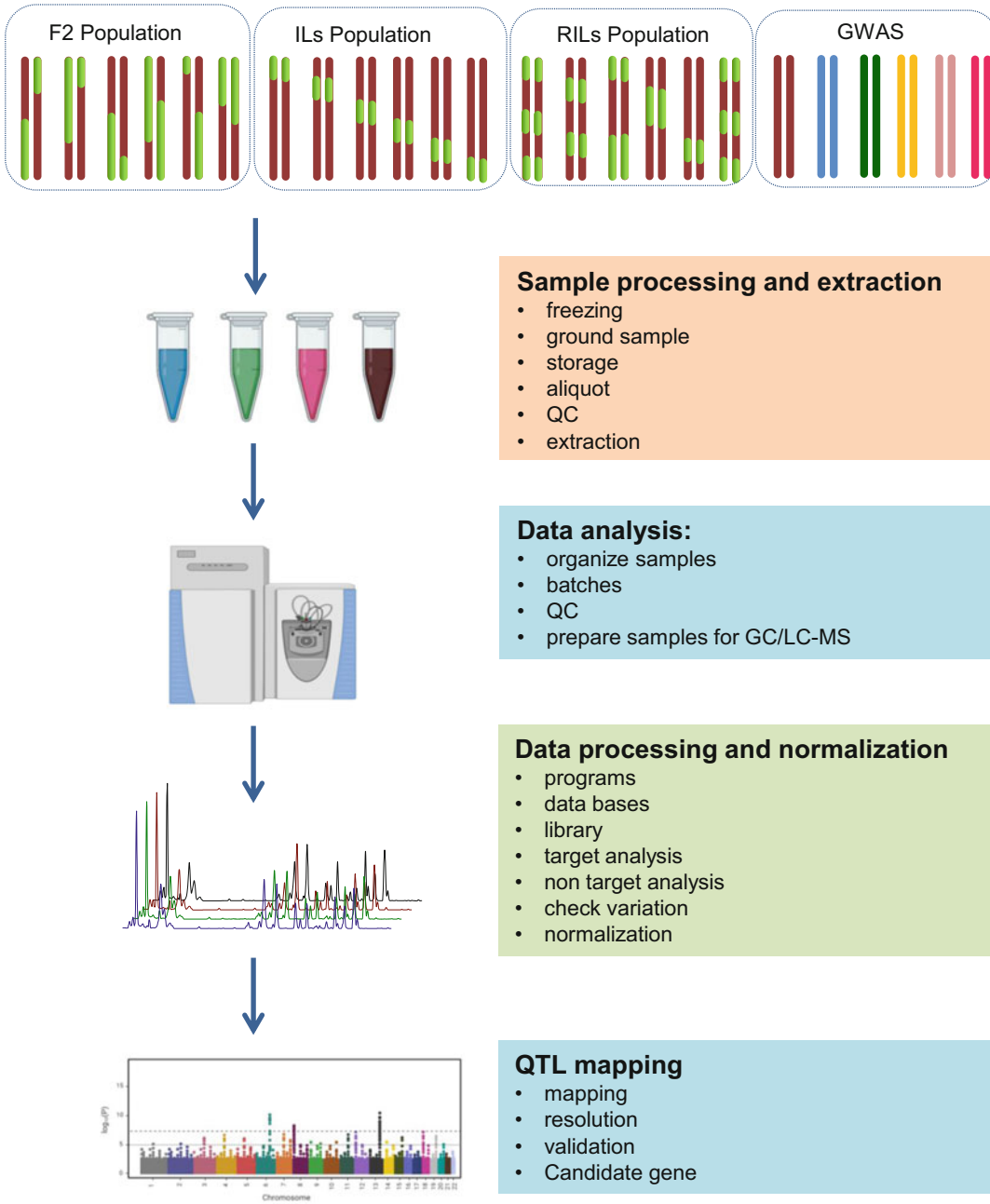


Fig. 1 Flowchart of the metabolomics study in plants. Showing the different steps for experimental design, sample preparation, and process for QTL experimental study. Part of the figure was prepared using biorender.com

studies, there are a number of pitfalls and known problems which need to be carefully avoided [30, 31]. In this chapter, we describe a metabolic workflow (Fig. 1), which provides guidelines for handling large sample numbers for the specific purpose of metabolic quantitative trait loci approaches.

2 The Workflow of Metabolomics Analysis

2.1 Plant Population Growth

1. Suitable plant populations are ILs, DHs, RILs, or GWA panels that are commonly used to investigate the genetic architecture of metabolite accumulation.
2. Suitable field or greenhouse growth conditions that are large enough to accommodate the populations in a manner that facilitates rapid harvest of samples from individual plants.

Precise details concerning the appropriate extraction protocols, machine settings, and running conditions are provided in [32–34] for GC-MS and LC-MS. Here we solely concentrate on aspects pertinent to the large-scale analysis of genetic populations and normalization aspects that need to be adopted to ensure proper cross-sample comparability as well as downstream analysis of the data within the framework of quantitative loci and association mapping analyses.

2.2 Sample Preparation

Sample harvesting and preparation are crucial steps in metabolomics as they greatly affect the reliability and final metabolomics results [30, 31]. In large-scale experiments with vast sample size and genotype numbers (e.g., ILs, RILs, or GWAS) it is conceivable that these may be slightly different in terms of their developmental age adding yet another source of variation. However, experimental design is key to any metabolomics experiment and having a large number of biological replicates is an essential means to minimize metabolite variation during sample preparation. After harvesting, plant organs (e.g., leaves, flowers, or fruits) should be immediately snap-frozen in liquid nitrogen and stored at −80 °C, or immediately ground to a powder and extracted (*see* **Note 1**). Many extraction protocols are available for plant metabolomics and have been discussed in detail before (for example *see* refs. 34, 35 for LC-MS, and ref. [33] for GC-MS). However, there are certain key steps at which these protocols should be adapted when handling a large number of plant samples. For example quality control (QC) (*see* **Note 2**) and pooled control samples (*see* **Note 3**) are necessary for data normalization to reduce the analytical errors and batch-to-batch deviation [32]. While most metabolomics studies are carried out under the highly controlled conditions of the laboratory, most mQTL studies carried out for crop species have been conducted in the field. For this reason and in order to minimize the variation there are several crucial points to take into consideration during harvest (*see* **Note 4**).

2.3 Sample Processing and Extraction

After harvesting, plant organs (e.g., leaves, flowers, or fruits) or dissected tissues should be immediately snap-frozen in liquid nitrogen and stored at −80 °C, or immediately ground to a powder and extracted. Sample grinding is required to optimize solvent

extraction and additionally aids in homogenizing the sample material [36]. Many extraction protocols are available for plant metabolomics and have been discussed in detail before [33, 34]. That said some important points require adaptation within these protocols when handling the large number of plant samples required for QTL analysis. First, quality control (QC) is essential throughout the entire sample preparation process, from the field to the sample storage location and through distribution to chemical analytics and data normalization strategies. In taking great caution to reduce analytical errors the quality of data is improved allowing for the robust detection of relatively small metabolomics changes between genotypes. Prior to extraction QC samples should be prepared by pooling aliquots of individual study samples; the QC samples should then be distributed across all machine batches and aliquots thereof should be extracted, derivatized, and analyzed at the same time as the individual study samples.

2.4 Sample Preparation and Analysis

Following extraction samples must subsequently be prepared for analysis. In the case of LC-MS, once the samples are extracted an aliquot of the extract can directly be introduced to the apparatus (*see* ref. 34 for LC-MS). In GC-MS-based metabolomics, however, additional preparative steps are required either to confer volatility to the metabolites via silylation or to simplify chromatography of sugars via methoxyamination [37–39]. We recommend to divide the samples in batches so that each batch contains 50–80 samples with a large number of QC samples distributed across the sequence run (*see* **Note 5**). Metabolite profiling via GC-MS involves several general steps (for details *see* ref. 40). In the case of LC-MS-based metabolomics approaches, the most frequently used protocols use C18-based reversed-phase columns coupled to soft ionization (*see* **Note 6**).

2.5 Data Processing

Once the samples have been analyzed, automatic data processing tools are required for peak picking and mass peak alignment. In GC-MS several such tools, software, and databases have been established and used for this purpose (for more details *see* refs. 39–41). Data processing for LC-MS is, however, considerably more complex (*see* **Note 7** and [34] for details). For both GC-MS and LC-MS methods, manual checking of the peaks is strongly recommended.

2.6 Data Normalization

Metabolomics depends on its ability to detect and quantify biologically related metabolite changes in complex biological samples. As with any high-throughput technology, systematic biases are often observed in LC-MS and GC-MS metabolomics data [42, 43]. As the number of samples in the data set increases there is a corresponding time-dependent variation in the metabolite data. The variability in samples can arise from multiple sources including

the physiological differences they aim to detect and variability within the analytical method itself. Removing sources of variability such as systematic errors is thus one of the top priorities in metabolomic data preprocessing. However, metabolite diversity leads to different responses to variations at given experimental conditions, rendering normalization a highly demanding task [44]. For the effective elimination of different sources of analytical variation preprocessing steps should follow a specific sequence [32]. Here the quality control (QC) samples are of key importance. They are ideally prepared by pooling equal volumes of material from all of the biological samples to be analyzed. Alternatively, a chemically defined mixture of authenticated reference compounds [45] that mimics the metabolic composition of the investigated biological material can be employed. Both synthetic mixtures and biological QC samples are then subjected to the same sample extraction, instrumental analyses (ideally distributed across the analytical run), and data processing, thus providing quality checks for technical and analytical error, and quantitative calibration to eliminate batch effects for the final processed data. This normalization is a crucial step for minimizing the batch-to-batch data variability across extended periods and has recently been re-suggested as a community standard for metabolomics (Alseekh et al., in review).

3 QTL Analysis in Mapping Populations

The principle of quantitative trait locus (QTL) mapping is based on detecting the association of genetic markers with the phenotype of interest in the resultant offspring [46]. Markers are used to partition the mapping population into different genotypic groups based on the presence or absence of a particular marker locus and to determine whether significant differences exist between groups with respect to the trait being measured [44]. If a QTL is linked to a marker locus, then the individuals with different marker locus genotypes will have different mean values of the quantitative trait under study. In plants, the use of such mapping populations—often referred to as immortal populations—holds great utility since the use of stable populations permits the growth of clonal replicates as well as multiple analyses of genetically identical individuals across multiple harvests. There are several structural populations and methods which have been used to detect the QTL and mapping. Therefore, choosing the proper population for such experiments is a key determinant in the success of any given project. There are several factors influencing the detection of QTL that should be considered in advance of planning such experiments (*see* **Note 8**). In the following sections, we briefly describe the most commonly used structural populations for the QTL mapping.

3.1 RIL Mapping

Immortal mapping populations consisting of homozygous individuals have been much used to map loci for complex traits in plants. Recombinant inbred lines (RILs) can be obtained relatively easily and are produced by successively selfing the progeny of individual F_2 plants (single seed descent method), from which the F_8 generation and onwards are practically homozygous lines that will produce further progeny that is essentially identical to the previous generation [47]. Such a population can also be produced by induced chromosomal doubling of haploids, such as for doubled haploids (DHs; [48–50]). However, RILs are likely advantageous over DHs since they are characterized by a higher frequency of recombination within the population, resulting from multiple meiotic events occurred during repeated selfing [51].

3.2 IL Mapping

Another type of immortal population consists of introgression lines (IL) which are obtained through repeated backcrossing and extensive genotyping. These can also be referred to as near-isogenic lines (NILs; [52]), or backcross inbred lines (BILs; [53, 54])—although the latter are slightly different in nature. These lines contain a single or a small number of genomic introgression fragments from a donor parent into an otherwise homogeneous genetic background.

3.3 Genome-Wide Association Mapping

Although IL and RILs have historically been the most common types of experimental populations used for the analysis of quantitative traits and represent powerful methods to identify regions of the genome that co-segregate with a given trait, they do suffer from some limitations [44]. Namely, only allelic diversity that segregates between the parents of the particular F_2 cross or within the RIL population can be assayed [44], and secondly, the amount of recombination that occurs during the creation of the RIL population places a limit on the mapping resolution [55]. The basic principle of genome-wide association studies (GWAS), which was initially developed for use in medical genetics, is that the incidence of nucleotide polymorphisms is associated with the presence of variance that overcomes the limitations of using the IL and RILs. This approach has several major advantages over conventional QTL mapping. First, a much larger and more representative gene pool can be surveyed. Second, it bypasses the expense and time of mapping studies and enables the mapping of many traits in one set of genotypes. Third, a much finer mapping resolution can be achieved, resulting in small confidence intervals of the detected loci compared to classical mapping, where the identified loci need to be fine-mapped. Finally, it has the potential not only to identify and map QTLs but also to identify the causal polymorphism within a gene that is responsible for the difference in two alternative phenotypes [44]. A major issue with association studies is false positives, and the main sources of such false positives are the linkage between causal and noncausal sites [56, 57].

4 Conclusion

Both gas chromatography-mass spectrometry (GC-MS) and liquid chromatography-mass spectrometry (LC-MS) are widely used analytical tools for profiling highly complex mixtures of primary and secondary metabolites, respectively. Use of these techniques in high throughput is faced with a large number of potential sources of nonbiological variation that can compromise the interpretation of the results. However, by following several recommendations prior to and during the conductance of large-scale genomics and QTL mapping experiments such problems can be circumvented in a relatively facile manner. This will allow us to move improving crop composition from one metabolite at a time to more comprehensive changes. Owing to technical limitations, researchers traditionally focus on a single or at most a handful of metabolic traits that were of greatest importance for either industrial or nutritional value. Prime examples of these targeted approaches include carotenoid content of tomato, protein content of maize, and starch content of potato and rice [12]. The tomato hybrid AB2 harbors a QTL from *S. pennellii* and is currently a leading processing variety. Another interesting example is the recent identification, by association mapping, of lycopene ε cyclase as a key determinant of provitamin A levels in maize. This finding is particularly pertinent given the severe health disorders that result from vitamin A deficiency. These strategies were at least partially reliant on association mapping; however they did not yet embrace the possibilities offered by metabolome profiling. In recent years metabolomics has allowed huge insight into the genetic architecture of hundreds of metabolites (*see* for example refs. 4, 58–66), the metabolic shifts that occurred on domestication [62, 67], and early metabolite markers that are able to predict yield [68–70]. It thus seems likely that we have just begun to exploit the possibilities offered in metabolomics-associated breeding.

5 Notes

1. Given that the levels of metabolites vary through the day, and that some experiments are too large to allow harvest in a single day it is essential to harvest control samples for each temporally separate harvest. Also as plant metabolomics experiments are generally performed at the organ level (developing fruit, whole leaf, root, etc.), it is recommended to have pooled samples to reduce the level of variation within genotype. These issues are especially important when the harvest sessions of a given experiment are numerous or when each session requests several people harvesting to limit its duration. The age, or preferably

the developmental stage, of the plants or their organs needs to be defined relative to standardized growth conditions and/or phenology descriptors, by using dedicated ontologies (Plant Ontology at http://www.plantontology.org/ for phenology or reference articles [71] for Arabidopsis and [72] for tomato) when available.

2. The quality control (QC) samples should qualitatively and quantitatively represent the entire collection of samples included in the study, providing an average of all of the metabolomes analyzed in the study. Sample is prepared by pooling aliquots of individual study samples, either all or a subset representative for the study. The QC sample has (should have) an identical or a very similar (bio) chemical diversity as the study samples. The QC samples are evenly distributed over all the batches and are extracted, derivatized, and analyzed at the same time as the individual study samples as part of the total sequence order. The data from the QC samples is used to monitor drift, separate high- and low-quality data, equilibrate the analytical platform, correct for drift in the signal, and allow the integration of multiple analytical experiments. The data analysis technique such as principal component analysis can be used to quickly assess the reproducibility of the QC samples in an analytical run. The QC samples are used to determine the variance of a metabolite feature.
3. In case the experiment is too large to allow harvesting in a single and relatively short time, it is essential to harvest control samples from each temporally separate harvest. Further, plant metabolomics experiments are generally performed at the organ levels, and recommend to harvest pooled samples (several fruits or leafs) per biological replicate. In addition, the age or developmental stage should be carefully considered according to standardized growth condition and phenology descriptors.
4. All samples for a given experiment should follow exactly the same procedure before grinding, and during extraction, storage, and analyzing. Sample grinding is usually required to optimize solvent extraction and additionally aids in homogenizing the sample material.
5. While online derivatization instrumentation is available that allows each sample to be derivatized for the same time prior to injection—it frequently breaks down and the need for such equipment can easily be circumvented by simple randomization approaches.
6. In LC-MS, by contrast to electron impact ionization, applied in GC-TOF/MS, ionization typically involves soft ionization techniques, such as electrospray ionization or atmospheric

pressure chemical ionization, resulting in protonated (in positive mode) or deprotonated (in negative mode) molecular ions. Modern high-resolution instruments with exact mass detection, such as TOF/MS, ion cyclotron FT-MS, or orbitrap FT-MS, nowadays enable the profiling of hundreds to thousands of compounds in plant extracts, combined with elemental formula calculations of the detected masses [38, 73].

7. Chromatograms from the UPLC–FT/MS runs can be analyzed and processed with REFINER MS® 10.0 (GeneData, http://www.genedata.com). Molecular masses, retention time, and associated peak intensities for each sample are extracted from the .raw files. The chemical noise was subtracted automatically. The chromatogram alignments are performed using a pairwise alignment-based tree using m/z windows of five points and RT windows of five scans within a sliding frame of 200 scans. The further processing of the MS data includes isotope clustering, adduct detection, and library searches. Resulting data matrices with peak ID, retention time, and peak intensities in each sample are generated.
8. Factors influencing QTL mapping: The environmental effects may have a large influence on the expression of quantitative traits. The size of the population used in the mapping study is also highly important; the larger the population, the more accurate the mapping study and the more likely it is to allow detection of QTL with smaller effects. Further, QTL mapping studies should be independently confirmed or validated. Such confirmation studies (referred to as validation or replication studies) can be achieved by repeating the experiment and the QTL mapping at different sites, seasons, or years. The conserved detected QTL throughout several repeated experiment is most likely the QTL that has strong genetic effect (high heritability) and that can be chosen as a region to focus on in further analysis. A second type of validation may involve independent populations constructed from the same parental genotypes or closely related genotypes used in the primary QTL mapping study. In the GWAS, once an association between a particular SNP and variation in a trait of interest has been established, a crucial but yet too often overlooked step is to replicate the association in an independent mapping population. As the number of studies documenting significant associations between SNPs and variation in quantitative traits of interest accumulates, increasing emphasis should be placed on replicating studies to validate the effects of significant associations.

Acknowledgments

SA and ARF acknowledge funding of the PlantaSYST project by the European Union's Horizon 2020 Research and Innovation Programme (SGA-CSA no. 664621 and no. 739582 under FPA no. 664620).

References

1. Alseekh S, Fernie AR (2018) Metabolomics 20 years on: what have we learned and what hurdles remain? Plant J 94(6):933–942. https://doi.org/10.1111/tpj.13950
2. Rai A, Saito K, Yamazaki M (2017) Integrated omics analysis of specialized metabolism in medicinal plants. Plant J 90(4):764–787. https://doi.org/10.1111/tpj.13485
3. Bijlsma S, Bobeldijk L, Verheij ER, Ramaker R, Kochhar S, Macdonald IA, van Ommen B, Smilde AK (2006) Large-scale human metabolomics studies: a strategy for data (pre-) processing and validation. Anal Chem 78 (2):567–574. https://doi.org/10.1021/ac051495j
4. Schauer N, Semel Y, Roessner U, Gur A, Balbo I, Carrari F, Pleban T, Perez-Melis A, Bruedigam C, Kopka J, Willmitzer L, Zamir D, Fernie AR (2006) Comprehensive metabolic profiling and phenotyping of interspecific introgression lines for tomato improvement. Nat Biotechnol 24(4):447–454. https://doi.org/10.1038/nbt1192
5. Fiehn O (2002) Metabolomics - the link between genotypes and phenotypes. Plant Mol Biol 48(1–2):155–171. https://doi.org/10.1023/a:1013713905833
6. Fu J, Keurentjes JJB, Bouwmeester H, America T, Verstappen FWA, Ward JL, Beale MH, de Vos RCH, Dijkstra M, Scheltema RA, Johannes F, Koornneef M, Vreugdenhil D, Breitling R, Jansen RC (2009) System-wide molecular evidence for phenotypic buffering in Arabidopsis. Nat Genet 41(2):166–167. https://doi.org/10.1038/ng.308
7. Rowe HC, Hansen BG, Halkier BA, Kliebenstein DJ (2008) Biochemical networks and epistasis shape the Arabidopsis thaliana metabolome. Plant Cell 20(5):1199–1216. https://doi.org/10.1105/tpc.108.058131
8. Wentzell AM, Rowe HC, Hansen BG, Ticconi C, Halkier BA, Kliebenstein DJ (2007) Linking metabolic QTLs with network and cis-eQTLs controlling biosynthetic pathways. PLoS Genet 3(9):1687–1701. https://doi.org/10.1371/journal.pgen.0030162
9. Zabotina OA (2013) Metabolite-based biomarkers for plant genetics and breeding. In: Lübberstedt T, Varshney RK (eds) Diagnostics in plant breeding. Springer, Dordrecht, Netherlands, pp 281–309. https://doi.org/10.1007/978-94-007-5687-8_14
10. Varshney A, Mohapatra T, Sharma RP (2005) Molecular mapping and marker assisted selection of traits for crop improvement. In: Srivastava PS, Narula A, Srivastava S (eds) Plant biotechnology and molecular markers. Springer, Dordrecht, Netherlands, pp 289–330. https://doi.org/10.1007/1-4020-3213-7_20
11. Sumner LW, Lei Z, Nikolau BJ, Saito K (2015) Modern plant metabolomics: advanced natural product gene discoveries, improved technologies, and future prospects. Nat Prod Rep 32 (2):212–229. https://doi.org/10.1039/c4np00072b
12. Fernie AR, Schauer N (2009) Metabolomics-assisted breeding: a viable option for crop improvement? Trends Genet 25(1):39–48. https://doi.org/10.1016/j.tig.2008.10.010
13. Alseekh S, Bermudez L, de Haro LA, Fernie AR, Carrari F (2018) Crop metabolomics: from diagnostics to assisted breeding. Metabolomics 14(11):148. https://doi.org/10.1007/s11306-018-1446-5
14. Davey JW, Hohenlohe PA, Etter PD, Boone JQ, Catchen JM, Blaxter ML (2011) Genome-wide genetic marker discovery and genotyping using next-generation sequencing. Nat Rev Genet 12(7):499–510. https://doi.org/10.1038/nrg3012
15. Stitt M, Fernie AR (2003) From measurements of metabolites to metabolomics: an 'on the fly' perspective illustrated by recent studies of carbon-nitrogen interactions. Curr Opin Biotechnol 14(2):136–144. https://doi.org/10.1016/s0958-1669(03)00023-5
16. Meyer RC, Steinfath M, Lisec J, Becher M, Witucka-Wall H, Torjek O, Fiehn O, Eckardt A, Willmitzer L, Selbig J, Altmann T (2007) The metabolic signature related to high plant growth rate in Arabidopsis thaliana. Proc Natl Acad Sci U S A 104(11):4759–4764. https://doi.org/10.1073/pnas.0609709104
17. Schauer N, Zamir D, Fernie AR (2005) Metabolic profiling of leaves and fruit of wild species tomato: a survey of the Solanum lycopersicum

complex. J Exp Bot 56(410):297–307. https://doi.org/10.1093/jxb/eri057

18. Wen W, Li K, Alseekh S, Omranian N, Zhao L, Zhou Y, Xiao Y, Jin M, Yang N, Liu H, Florian A, Li W, Pan Q, Nikoloski Z, Yan J, Fernie AR (2015) Genetic determinants of the network of primary metabolism and their relationships to plant performance in a maize recombinant inbred line population. Plant Cell 27(7):1839–1856. https://doi.org/10.1105/tpc.15.00208
19. de Abreu E, Lima F, Li K, Wen W, Yan J, Nikoloski Z, Willmitzer L, Brotman Y (2018) Unraveling the lipid metabolism in maize with time-resolved multi-omics data. Plant J 93 (6):1102–1115
20. Hill CB, Taylor JD, Edwards J, Mather D, Bacic A, Langridge P, Roessner U (2013) Whole-genome mapping of agronomic and metabolic traits to identify novel quantitative trait loci in bread wheat grown in a water-limited environment. Plant Physiol 162(3):1266–1281. https://doi.org/10.1104/pp.113.217851
21. Chen J, Hu X, Shi T, Yin H, Sun D, Yuanfeng H, Xia X, Luo J, Fernie AR, He Z, Chen W (2020) Metabolite-based genome-wide association study enables dissection of the flavonoid decoration pathway of wheat kernels. Plant Biotechnol J. https://doi.org/10.1111/pbi.13335
22. Cao D, Lutz A, Hill CB, Callahan DL, Roessner U (2017) A quantitative profiling method of phytohormones and other metabolites applied to barley roots subjected to salinity stress. Front Plant Sci 7:2070. https://doi.org/10.3389/fpls.2016.02070
23. Tohge T, Ramos MS, Nunes-Nesi A, Mutwil M, Giavalisco P, Steinhauser D, Schellenberg M, Willmitzer L, Persson S, Martinoia E, Fernie AR (2011) Toward the storage metabolome: profiling the barley vacuole. Plant Physiol 157(3):1469–1482. https://doi.org/10.1104/pp.111.185710
24. Peng M, Shahzad R, Gul A, Subthain H, Shen SQ, Lei L, Zheng ZG, Zhou JJ, Lu DD, Wang SC, Nishawy E, Liu XQ, Tohge T, Fernie AR, Luo J (2017) Differentially evolved glucosyltransferases determine natural variation of rice flavone accumulation and UV-tolerance. Nat Commun 8:1975. https://doi.org/10.1038/s41467-017-02168-x
25. Yang ZG, Nakabayashi R, Okazaki Y, Mori T, Takamatsu S, Kitanaka S, Kikuchi J, Saito K (2014) Toward better annotation in plant metabolomics: isolation and structure elucidation of 36 specialized metabolites from Oryza sativa (rice) by using MS/MS and NMR analyses. Metabolomics 10(4):543–555. https://doi.org/10.1007/s11306-013-0619-5
26. Sulpice R (2019) Closing the yield gap: can metabolomics be of help? J Exp Bot 71 (2):461–464. https://doi.org/10.1093/jxb/erz322
27. Biais B, Bernillon S, Deborde C, Cabasson C, Rolin D, Tadmor Y, Burger J, Schaffer AA, Moing A (2012) Precautions for harvest, sampling, storage, and transport of crop plant metabolomics samples. In: Hardy NW, Hall RD (eds) Plant metabolomics: methods and protocols. Humana, Totowa, NJ, pp 51–63. https://doi.org/10.1007/978-1-61779-594-7_4
28. Gibon Y, Rolin D (2012) Aspects of experimental design for plant metabolomics experiments and guidelines for growth of plant material. Methods Mol Biol 860:13–30. https://doi.org/10.1007/978-1-61779-594-7_2
29. Sysi-Aho M, Katajamaa M, Yetukuri L, Oresic M (2007) Normalization method for metabolomics data using optimal selection of multiple internal standards. BMC Bioinformatics 8:93. https://doi.org/10.1186/1471-2105-8-93
30. Fernie AR, Aharoni A, Willmitzer L, Stitt M, Tohge T, Kopka J, Carroll AJ, Saito K, Fraser PD, DeLuca V (2011) Recommendations for reporting metabolite data. Plant Cell 23 (7):2477–2482. https://doi.org/10.1105/tpc.111.086272
31. Lu W, Su X, Klein MS, Lewis IA, Fiehn O, Rabinowitz JD (2017) Metabolite measurement: pitfalls to avoid and practices to follow. Annu Rev Biochem 86(1):277–304. https://doi.org/10.1146/annurev-biochem-061516-044952
32. Alseekh S, Wu S, Brotman Y, Fernie AR (2018) Guidelines for sample normalization to minimize batch variation for large-scale metabolic profiling of plant natural genetic variance. Methods Mol Biol 1778:33–46. https://doi.org/10.1007/978-1-4939-7819-9_3
33. Osorio S, Do PT, Fernie AR (2012) Profiling primary metabolites of tomato fruit with gas chromatography/mass spectrometry. In: Hardy NW, Hall RD (eds) Plant metabolomics: methods and protocols. Humana, Totowa, NJ, pp 101–109. https://doi.org/10.1007/978-1-61779-594-7_7
34. Shimizu T, Watanabe M, Fernie AR, Tohge T (2018) Targeted LC-MS analysis for plant secondary metabolites. Methods Mol Biol 1778:171–181. https://doi.org/10.1007/978-1-4939-7819-9_12
35. Salem MA, Juppner J, Bajdzienko K, Giavalisco P (2016) Protocol: a fast, comprehensive and reproducible one-step extraction method for the rapid preparation of polar and semi-polar metabolites, lipids, proteins, starch and cell wall polymers from a single sample. Plant Methods

12:45. https://doi.org/10.1186/s13007-016-0146-2

36. Markert B (1995) Sample preparation (cleaning, drying, homogenization) for trace element analysis in plant matrices. Sci Total Environ 176(1–3):45–61. https://doi.org/10.1016/0048-9697(95)04829-4
37. Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L (2000) Metabolite profiling for plant functional genomics. Nat Biotechnol 18(11):1157–1161. https://doi.org/10.1038/81137
38. Allwood JW, De Vos RCH, Moing A, Deborde C, Erban A, Kopka J, Goodacre R, Hall RD (2011) Plant metabolomics and its potential for systems biology research: background concepts, technology, and methodology. In: Jameson D, Verma M, Westerhoff HV (eds) Methods in systems biology, Methods in enzymology, vol 500, pp 299–336. https://doi.org/10.1016/b978-0-12-385118-5.00016-5
39. Lisec J, Schauer N, Kopka J, Willmitzer L, Fernie AR (2006) Gas chromatography mass spectrometry-based metabolite profiling in plants. Nat Protoc 1(1):387–396. https://doi.org/10.1038/nprot.2006.59
40. Kopka J, Fernie A, Weckwerth W, Gibon Y, Stitt M (2004) Metabolite profiling in plant biology: platforms and destinations. Genome Biol 5(6):109. https://doi.org/10.1186/gb-2004-5-6-109
41. Kopka J, Schauer N, Krueger S, Birkemeyer C, Usadel B, Bergmuller E, Dormann P, Weckwerth W, Gibon Y, Stitt M, Willmitzer L, Fernie AR, Steinhauser D (2005) GMD@CSB.DB: the Golm Metabolome Database. Bioinformatics 21(8):1635–1638. https://doi.org/10.1093/bioinformatics/bti236
42. Karpievitch YV, Nikolic SB, Wilson R, Sharman JE, Edwards LM (2014) Metabolomics data normalization with EigenMS. PLoS One 9(12):e116221. https://doi.org/10.1371/journal.pone.0116221
43. Fiehn O, Kopka J, Dormann P, Altmann T, Trethewey RN, Willmitzer L (2001) Metabolite profiling for plant functional genomics. Nat Biotechnol 19(2):173
44. Sehgal D, Singh R, Rajpal VR (2016) Quantitative trait loci mapping in plants: concepts and approaches. In: Rajpal VR, Rao SR, Raina SN (eds) Molecular breeding for sustainable crop improvement, Sustainable development and biodiversity, vol 2, pp 31–59. https://doi.org/10.1007/978-3-319-27090-6_2
45. Strehmel N, Hummel J, Erban A, Strassburg K, Kopka J (2008) Retention index thresholds for compound matching in GC-MS metabolite profiling. J Chromatogr B Anal Technol Biomed Life Sci 871(2):182–190. https://doi.org/10.1016/j.jchromb.2008.04.042
46. Broman KW (2001) Review of statistical methods for QTL mapping in experimental crosses. Lab Anim 30(7):44–52
47. Crow JF (2007) Haldane, Bailey, Taylor and recombinant-inbred lines. Genetics 176(2):729–732
48. Han F, Ullrich SE, Kleinhofs A, Jones BL, Hayes PM, Wesenberg DM (1997) Fine structure mapping of the barley chromosome-1 centromere region containing malting-quality QTLs. Theor Appl Genet 95(5):903–910. https://doi.org/10.1007/s001220050641
49. Rae AM, Howell EC, Kearsey MJ (1999) More QTL for flowering time revealed by substitution lines in Brassica oleracea. Heredity 83:586–596. https://doi.org/10.1038/sj.hdy.6886050
50. von Korff M, Wang H, Leon J, Pillen K (2004) Development of candidate introgression lines using an exotic barley accession (Hordeum vulgare ssp. spontaneum) as donor. Theor Appl Genet 109(8):1736–1745. https://doi.org/10.1007/s00122-004-1818-2
51. Jansen RC (2004) Quantitative trait loci in inbred lines. In: Handbook of statistical genetics. Wiley, Chichester, UK. https://doi.org/10.1002/0470022620.bbc16
52. Monforte AJ, Tanksley SD (2000) Development of a set of near isogenic and backcross recombinant inbred lines containing most of the Lycopersicon hirsutum genome in a L-esculentum genetic background: a tool for gene mapping and gene discovery. Genome 43(5):803–813. https://doi.org/10.1139/gen-43-5-803
53. Jeuken MJW, Lindhout P (2004) The development of lettuce backcross inbred lines (BILs) for exploitation of the Lactuca saligna (wild lettuce) germplasm. Theor Appl Genet 109(2):394–401. https://doi.org/10.1007/s00122-004-1643-7
54. Blanco A, Simeone R, Gadaleta A (2006) Detection of QTLs for grain protein content in durum wheat. Theor Appl Genet 113(3):563–565. https://doi.org/10.1007/s00122-006-0313-3
55. Jamann TM, Balint-Kurti PJ, Holland JB (2015) QTL mapping using high-throughput sequencing. In: Alonso JM, Stepanova AN (eds) Plant functional genomics: methods and protocols, Methods in molecular biology, vol 1284, 2nd edn, pp 257–285. https://doi.org/10.1007/978-1-4939-2444-8_13
56. Platt A, Vilhjalmsson BJ, Nordborg M (2010) Conditions under which genome-wide

association studies will be positively misleading. Genetics 186(3):1045–1052. https://doi.org/10.1534/genetics.110.121665

57. Larsson SJ, Lipka AE, Buckler ES (2013) Lessons from Dwarf8 on the strengths and weaknesses of structured association mapping. PLoS Genet 9(2):e1003246. https://doi.org/10.1371/journal.pgen.1003246
58. Luo J (2015) Metabolite-based genome-wide association studies in plants. Curr Opin Plant Biol 24:31–38. https://doi.org/10.1016/j.pbi.2015.01.006
59. Loudet O, Chaillou S, Merigout P, Talbotec J, Daniel-Vedele F (2003) Quantitative trait loci analysis of nitrogen use efficiency in Arabidopsis. Plant Physiol 131(1):345–358. https://doi.org/10.1104/pp.102.010785
60. Lisec J, Steinfath M, Meyer RC, Selbig J, Melchinger AE, Willmitzer L, Altmann T (2009) Identification of heterotic metabolite QTL in Arabidopsis thaliana RIL and IL populations. Plant J 59(5):777–788. https://doi.org/10.1111/j.1365-313X.2009.03910.x
61. Alseekh S, Tohge T, Wendenberg R, Scossa F, Omranian N, Li J, Kleessen S, Giavalisco P, Pleban T, Mueller-Roeber B, Zamir D, Nikoloski Z, Fernie AR (2015) Identification and mode of inheritance of quantitative trait loci for secondary metabolite abundance in tomato. Plant Cell 27(3):485–512. https://doi.org/10.1105/tpc.114.132266
62. Zhu GT, Wang SC, Huang ZJ, Zhang SB, Liao QG, Zhang CZ, Lin T, Qin M, Peng M, Yang CK, Cao X, Han X, Wang XX, van der Knaap E, Zhang ZH, Cui X, Klee H, Fernie AR, Luo J, Huang SW (2018) Rewiring of the fruit metabolome in tomato breeding. Cell 172 (1–2):249. https://doi.org/10.1016/j.cell.2017.12.019
63. Wu S, Alseekh S, Cuadros-Inostroza A, Fusari CM, Mutwil M, Kooke R, Keurentjes JB, Fernie AR, Willmitzer L, Brotman Y (2016) Combined use of genome-wide association data and correlation networks unravels key regulators of primary metabolism in Arabidopsis thaliana. PLoS Genet 12(10):e1006363. https://doi.org/10.1371/journal.pgen.1006363
64. Sauvage C, Segura V, Bauchet G, Stevens R, Do PT, Nikoloski Z, Fernie AR, Causse M (2014) Genome-wide association in tomato reveals 44 candidate loci for fruit metabolic traits. Plant Physiol 165(3):1120–1132. https://doi.org/10.1104/pp.114.241521
65. Nunes-Nesi A, Alseekh S, de Oliveira Silva FM, Omranian N, Lichtenstein G, Mirnezhad M, Gonzalez RRR, Sabio YGJ, Conte M, Leiss KA, Klinkhamer PGL, Nikoloski Z, Carrari F, Fernie AR (2019) Identification and characterization of metabolite quantitative trait loci in tomato leaves and comparison with those reported for fruits and seeds. Metabolomics 15(4):46. https://doi.org/10.1007/s11306-019-1503-8
66. Matsuda F, Nakabayashi R, Yang ZG, Okazaki Y, Yonemaru J, Ebana K, Yano M, Saito K (2015) Metabolome-genome-wide association study dissects genetic architecture for generating natural variation in rice secondary metabolism. Plant J 81(1):13–23. https://doi.org/10.1111/tpj.12681
67. Beleggia R, Rau D, Laido G, Platani C, Nigro F, Fragasso M, De Vita P, Scossa F, Fernie AR, Nikoloski Z, Papa R (2016) Evolutionary metabolomics reveals domestication-associated changes in tetraploid wheat kernels. Mol Biol Evol 33(7):1740–1753. https://doi.org/10.1093/molbev/msw050
68. Brauner PC, Schipprack W, Utz HF, Bauer E, Mayer M, Schon CC, Melchinger AE (2019) Testcross performance of doubled haploid lines from European flint maize landraces is promising for broadening the genetic base of elite germplasm. Theor Appl Genet 132 (6):1897–1908. https://doi.org/10.1007/s00122-019-03325-0
69. Riedelsheimer C, Endelman JB, Stange M, Sorrells ME, Jannink JL, Melchinger AE (2013) Genomic predictability of interconnected biparental maize populations. Genetics 194 (2):493–503. https://doi.org/10.1534/genetics.113.150227
70. Westhues M, Schrag TA, Heuer C, Thaller G, Utz HF, Schipprack W, Thiemann A, Seifert F, Ehret A, Schlereth A, Stitt M, Nikoloski Z, Willmitzer L, Schon CC, Scholten S, Melchinger AE (2017) Omics-based hybrid prediction in maize. Theor Appl Genet 130 (9):1927–1939. https://doi.org/10.1007/s00122-017-2934-0
71. Boyes DC, Zayed AM, Ascenzi R, McCaskill AJ, Hoffman NE, Davis KR, Gorlach J (2001) Growth stage-based phenotypic analysis of Arabidopsis: a model for high throughput functional genomics in plants. Plant Cell 13 (7):1499–1510. https://doi.org/10.1105/tpc.13.7.1499
72. Brukhin V, Hernould M, Gonzalez N, Chevalier C, Mouras A (2003) Flower development schedule in tomato Lycopersicon esculentum cv. sweet cherry. Sex Plant Reprod 15 (6):311–320. https://doi.org/10.1007/s00497-003-0167-7
73. Allwood JW, Clarke A, Goodacre R, Mur LAJ (2010) Dual metabolomics: a novel approach to understanding plant-pathogen interactions. Phytochemistry 71(5–6):590–597. https://doi.org/10.1016/j.phytochem.2010.01.006

Chapter 4

High-Throughput DNA Isolation in Vegetable Crops for Genomics Applications

Pasquale Tripodi and Giovanna Festa

Abstract

Isolating high-quality DNA is essential for several applications in molecular biology and genomics. Performing whole-genome sequencing in crops and development of reduced representation genomic libraries for genotyping require precise standard on DNA in terms of concentration and purity. For screening large populations it is essential to increase the extraction throughput at affordable costs. In this chapter a homemade protocol is provided that is able to isolate in 96-well plates 198 samples of DNA in a single extraction. The method has been validated in tomato and pepper and can be applied in several vegetable species.

Key words DNA extraction, Vegetable crops, High-throughput, Genomic applications

1 Introduction

Isolation of high-quality deoxyribonucleic acid is a key step for the success of many molecular biology applications. In recent years, the rise of cutting-edge technologies in genomics has required standardized parameters to be reached in terms of quality and quantity of DNA prior to processing. Furthermore, the need to analyze large sets of samples for experimental mapping population development, QTL studies, and marker-assisted selection has required the increase of the throughput of extraction at affordable costs. Major constraints occurring during DNA isolation regard the separation of nucleic acid from carbohydrates, proteins, and polyphenols which could interfere with the various steps required in next-generation sequencing methods such as library construction and amplification.

In this chapter, we describe a homemade reagent-based protocol suitable for genomic applications toward the isolation of large number of samples in 96-well plates to be performed in less than 2 h. The method is a modification of a microprep-based protocol for PCR marker routine analysis in tomato [1].

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_4,

2 Materials

2.1 Equipment

- Freeze dryer lyophilizer (Benchtop, Thermo Fisher Scientific).
- TissueLyser II (Qiagen).
- Millipore Milli-Q® IQ 7003/05/10/15 Ultrapure & Pure Water Purification System (Merck).
- 96-Well MegaBlock, 1.2 mL (Sarstedt).
- Tungsten carbide beads, 3 mm (Qiagen).
- Centrifuge with microplate rotor (SL16R, rotor code 75003624, Thermo Scientific).
- Multichannel pipetman, 12 channels (G 12 × 20 μL, G 12 × 300 μL, Gilson).
- Graduated cylinders and beakers with volumes of 250, 500, and 1000 mL (Vetrochimica).
- Polystyrene reagent reservoirs 100 mL (Thermo Fisher Scientific).
- Nanodrop spectrophotometer (ND-1000, Thermo Fisher Scientific).
- Qubit flex fluorometer (Thermo Fisher Scientific).
- Temperature-controlled water bath (Precision™ Circulating Water Baths, Thermo Scientific).
- Horizontal gel electrophoresis system (Sub-Cell Model 192, Biorad).
- Gel Doc XR+ Gel Documentation System (Biorad).

2.2 Reagents

- Trizma base (Sigma-Aldrich).
- Sorbitol (Sigma-Aldrich).
- *N*-lauryl sarcosine sodium salt (Sigma-Aldrich).
- Ethylenediaminetetraacetic acid disodium salt dehydrate (EDTA) (Sigma-Aldrich).
- Sodium chloride (NaCl) (Sigma Aldrich).
- CTAB (Sigma-Aldrich).
- Sodium bisulfite (Sigma-Aldrich).
- Acetic acid glacial (Sigma-Aldrich).
- RNase A 20 mg/mL (Thermo Fisher Scientific).
- Isopropanol (Sigma-Aldrich), store at −20 °C.
- Ethanol (70% v/v) (Sigma-Aldrich), store at −20 °C.
- Agarose, molecular biology grade (Thermo Fisher Scientific).

- SYBR™ Safe DNA Gel Stain (Thermo Fisher Scientific).
- Gel loading buffer (Sigma-Aldrich).
- Lambda DNA/HindIII Marker (Thermo Fisher Scientific).

2.3 Solutions

All solutions must be prepared with molecular biology-grade chemicals, in sterile nuclease-free Milli-Q water (hereafter NFH_2O). Autoclave at 120 °C for 20 min of all plasticware and other accessories.

- 1 M Tris–HCl, pH 7.5: In a baker, dissolve 121.1 g of Trizma base in 600 mL of NFH_2O in a 1 L beaker. Add a stir bar to the beaker and leave it on a stir plate. Add more NFH_2O, adjust the pH to 7.5 with concentrated HCl (~60 mL 12 N) until the solution is completely dissolved, and then adjust the final volume to 1 L with NFH_2O.
- 0.5 M EDTA, pH 8.0: Dissolve 186.01 g of EDTA disodium salt in 600 mL of NFH_2O in a 1 L beaker. Add a stir bar to the beaker and stir vigorously on a magnetic stirrer. Adjust the pH to 8.0 with NaOH (~20 g of NaOH pellets) until the solution is completely dissolved and then adjust the final volume to 1 L with NFH_2O.
- 5 M NaCl: In a graduated beaker dissolve 146.1 g of NaCl with 500 mL NFH_2O. Stir vigorously on a magnetic stirrer until it is fully dissolved.
- DNA extraction buffer (DB): 0.35 M Sorbitol, 0.1 M Tris base, pH 7.5, 0.5 M EDTA, pH 8.0. For 500 mL of solution, take 31.88 g of sorbitol, 6.05 g of Trizma base, and 1 mL of 0.5 M EDTA, pH 8.0. Add 300 mL of NFH_2O in a beaker and dissolve all components with a magnetic stirrer. Add 300 μL of HCl to reach 8.26 pH, and then adjust the final volume to 0.5 L with NFH_2O. The solution can be autoclaved or kept fresh and stored at 4 °C.
- Nuclei lysis buffer (LB): 1 M Tris–HCl, pH 7.5; 0.5 M EDTA, pH 8.0; 5 M NaCl and CTAB. For 500 mL of solution, add 100 mL of Tris–HCl (final concentration 200 mM), 50 mL of EDTA (final concentration 0.1 M), 200 mL of NaCl (final concentration 2 M), and 1.0 g of CTAB. Dissolve all components adding 150 mL of NFH_2O on a magnetic stirrer. The solution can be autoclaved or kept fresh and stored at room temperature.
- 5% Sarkosyl buffer (SB): Dissolve 2.5 g of *N*-lauryl-sarcosine in 50 mL of NFH_2O. Conserve the solution at room temperature.
- Elution TE buffer: 1 M Tris–HCl, pH 7.5, 0.5 M EDTA, pH 8.0. Take 5 mL of 1 M Tris–HCl (final concentration 10 mM) and add 1 mL of EDTA (final concentration 1 mM). Adjust the final volume to 500 mL with NFH_2O.

- Chloroform:isoamyl alcohol 24:1 (Sigma Aldrich); alternatively, under chemical fume hood, add in a glass bottle 240 mL of chloroform to 10 mL of isoamyl alcohol.
- Main extraction buffer (MEB): Add in proportion 2.5 of DB, 2.5 of LB, and 1.0 of SB. For each 100 mL of solution add 0.5 g of sodium bisulfite, e.g., 50 mL BD + 50 mL LB + 20 mL SB + 0.6 g of sodium bisulfite.
- TAE 50× buffer: Trizma base, 0.5 M EDTA, pH 8.0, acetic acid glacial. For 500 mL of solution, weigh out 121 g of Trizma base and dissolve in approximately 300 mL of NFH_2O. Add 28.55 mL of 100% acetic acid glacial and 50 mL of 0.5 M EDTA, pH 8.0. Adjust the solution to a final volume. The solution can be stored at room temperature.

3 Methods

3.1 DNA Isolation

- Collect small leaves (~4 g fresh weight) from young plants (3 weeks older) and put into 96-well plates (Fig. 1). Each well is representative of a plant. Take note in an Excel file of the correspondence between well and sample (e.g., A1 = acc1, A2 = acc2).
- Add two beads to each well.
- Before grinding, lyophilize the plate with samples (without the cover) for 24 h or immerge it in liquid nitrogen for 5–6 s (*see* **Note 1**).
- Grind samples in TissueLyser II at a frequency of 30 Hz for 30 s. Be aware that plates are closed (*see* **Note 2**).
- Ensure that all samples are ground (e.g., looking at the bottom of the plate or in each well), prepare fresh MEB, and dispense 35 mL in the reagent reservoir. Add 300 μL of MEB to each well using the multichannel (*see* **Note 3**).
- Add 4 μL of 20 mg/mL of RNase A to each well using 10 μL multichannel.
- Incubate in water bath at 65 °C for 30 min with periodic mixing by inverting the plate (~5 mixing steps) (*see* **Note 4**).
- After the incubation add 300 μL of mL of chloroform:isoamyl (24:1) to each well using the G 12 × 300 μL multichannel pipette. The step must be done under the chemical fume hood. Tighten the caps and mix gently by inverting for 1 min.
- Centrifuge at 2272 × *g* (4000 rpm) for 5 min.
- After the centrifuge, two layers will be observed: a supernatant (aqueous phase) containing the DNA plus RNA and other solubles, and a bottom phase containing leaf matter with proteins, carbohydrates, and other substances (*see* **Note 5**).

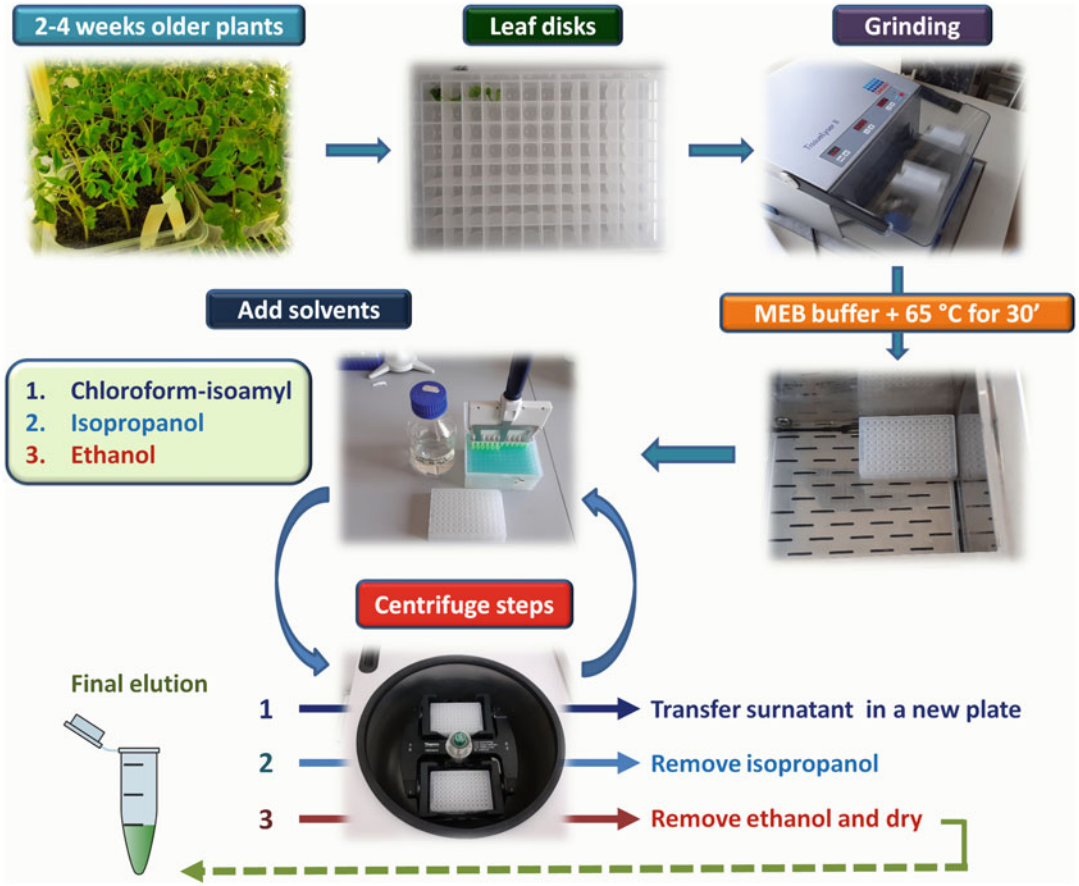


Fig. 1 Principal steps of DNA extraction details in Subheading 3.1

- Under the chemical fume hood, aspirate the supernatant (~250 μL) with the multichannel pipette avoiding touching the bottom layer.
- Transfer then the supernatant into a new 96 deep-well plate.
- Add an equal volume of isopropanol at −20 °C, gently shake the samples by inversion for 30 s, and then leave the plate in ice for 30 min (*see* **Note 6**).
- Centrifuge at 2272 × *g* (4000 rpm) for 5 min.
- Remove isopropanol by using multichannel pipette or by inverting the plate (*see* **Note 7**).
- Add 300 μL of ethanol 70% in order to wash the pellet, leave for 5 min at room temperature, and then centrifuge at maximum speed for 5 min.
- Remove ethanol and leave the pellet drying at room temperature for 15 min, placing the plates upside down.

- Once the pellets are dried, resuspend in 30 mL of TE buffer and incubate at 65 °C for 15 min in a thermostatic water bath.
- Centrifuge for 10 min at 2000 × *g* to obtain the finalized eluted DNA.

3.2 DNA Quantification

- The quantity (ng/μL) and quality (ratio 260/280 and 260/230) of DNA can be determined using Nanodrop or Qubit (*see* **Note 8**).
- In the case of use of Nanodrop it is highly recommended to check concentration and degradation by means of 1% agarose gel in 1× TAE buffer. For 500 mL of 1× TAE, add 10 mL of TAE 50× stock solution in 490 mL of NFH_2O. Prepare gel adding 1 g of agarose for every 100 mL of 1× TAE. Shake gently and pour it into the electrophoresis support plate.
- For each sample to be quantified, add 1 μL of DNA, 2 μL of gel loading buffer 6×, and 9 μL of NFH_2O.
- As reference add Lambda DNA/HindIII marker at a concentration of 50 ng/μL and 100 ng/μL.
- Run gel for 30 min at 90 V constant voltage and visualize at trans/UV using Gel Doc XR+.
- The concentration of high-molecular-weight DNA will be calculated by comparing the band intensity of the Lambda DNA/-HindIII control to the DNA sample.

4 Notes

1. It is possible to store samples at −80 °C after lyophilization. Furthermore, to increase the performance of grinding, the lyophilized samples can be placed for 10 min at −80 °C prior to grinding. For soft materials (e.g., very young leaves) it is possible to collect them fresh and keep for 1 h at −80 °C before grinding.
2. Ensure that plates are well closed, the hooks of TissueLyser II fixed, and the instrument is balanced.
3. Considering that 96 × 300 μL is 28.8 mL, it is recommended to add a larger volume to avoid losses due to pipetting.
4. Tighten well the caps in order to avoid solution spillage. A layer of film can be added.
5. During this step, proteins, polysaccharides, and other debris are separated. Indeed, chloroform breaks the bonds between proteins and DNA. The centrifuge allows separating DNA which remains in the aqueous phase, while the remaining compounds, being heavier, are decanted on the bottom. A clear aqueous phase highlights a good separation.

6. Cold isopropanol allows the DNA to precipitate. In case no precipitation occurs (e.g., low DNA), it is possible to increase the incubation in ice or at −20 °C up to 8 h or overnight. To speed up the process, the plate with isopropanol can be stored at −80 °C for 60–90-min incubation.
7. It is important that the pellet precipitated at the bottom. It is possible to increase the time of centrifuge up to 20 min to facilitate the precipitation.
8. The ratio of absorbance at 260 and 280 nm is used to assess the purity of DNA. For "pure" DNA, a ratio of ~1.8 is generally accepted. Lower values indicate the presence of protein, phenol, or other contaminants that absorb strongly at or near 280 nm. The ratio 260/230 is used as a secondary measure of nucleic acid purity. The expected value for "pure" nucleic acid is commonly in the range of 2.0–2.2. Lower values may indicate the presence of contaminants which absorb at 230 nm.

Reference

1. Fulton TM, Chunwongse J, Tanksley SD (1995) Microprep protocol for extraction of DNA from tomato and other herbaceous plants. Plant Mol Biol Rep 13:207–209

Chapter 5

High-Resolution Melting Analysis as a Tool for Plant Species Authentication

Liliana Grazina, Joana Costa, Joana S. Amaral, and Isabel Mafra

Abstract

High-resolution melting (HRM) analysis is a cost-effective, specific, and rapid tool that allows distinguishing genetically related plants and other organisms based on the detection of small nucleotide variations, which are recognized from melting properties of the double-stranded DNA. It has been widely applied in several areas of research and diagnostics, including botanical authentication of several food commodities and herbal products. Generally, it consists of the main steps: (1) in silico sequence analysis and primer design; (2) DNA extraction from plant material; (3) amplification by real-time PCR with an enhanced fluorescent dye targeting a specific DNA barcode or other regions of taxonomic interest (100–200 bp); (4) melting curve analysis; and (5) statistical data analysis using a specific HRM software. This chapter presents an overview of HRM analysis and application, followed by the detailed description of all the required reagents, instruments, and protocols for the successful and easy implementation of a HRM method to differentiate closely related plant species.

Key words HRM, Species identification, Authenticity, Botanical origin, Food, Herbal products

1 Introduction

Recently, there has been a high number of reports worldwide and a growing evidence regarding the occurrence of fraudulent practices in products of plant origin. These include several foods of high economic importance, such as spices, wine, and olive oil, and different products containing medicinal plants, namely herbal infusions, traditional herbal medicines, and plant food supplements [1–4]. The increasing concern of stakeholders, such as regulatory entities, industries, and consumers, has prompt the development of different methods aiming at the botanical origin authentication. DNA-based methods have undoubtedly proved to be suited for the identification of species, presenting advantages over phenotypic and chemical approaches in terms of specificity and reliability. Advances in molecular biology techniques over the last couple of decades lead to the development of high-resolution melting

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_5,

(HRM) analysis, as a simple, fast, and cost-effective tool for plant species authentication.

HRM analysis is a post-PCR approach based on monitoring the gradual denaturation of double-stranded DNA (dsDNA) of amplified fragments, which allows detecting small nucleotide differences. It has emerged from high-resolution real-time PCR instrumentation and new-generation fluorescent dyes. EvaGreen, LCGreen-PLUS, SYTO®9, or ResoLight are examples of enhanced fluorescent DNA-binding dyes that can be used at higher concentrations than the classical SYBR Green I dye, resulting in enhanced fluorescence signals and increased sensitivity, without causing PCR inhibition [5–8]. High-resolution equipment, capable of small temperature increments, high acquisition rate, and high melting accuracy, as well as appropriate software is also required [6–9]. When dsDNA dissociates into single strands (ssDNA), the dye is released, causing a fluorescence decrease that is plotted against temperature, generating a melting curve (Fig. 1a). The melting curve profile and estimated melting temperature depend on the amplicon length, sequence, and GC content. The temperature at which half of the amplicons are single stranded is called melting temperature (Tm). The Tm can be determined from the conventional melting curve analysis, corresponding to the melt peak obtained by plotting the negative derivative of the fluorescence (F) over temperature (T) ($-\mathrm{d}F/\mathrm{d}T$) versus the temperature [6, 7] (Fig. 1b). Amplicons that substantially differ in length and/or nucleotide composition present distinct melting profiles and, consequently, different Tm, being easily differentiated by simply using standard instrumentation with the SYBR Green dye. However, when amplicons differ in just one or few nucleotides, they may present similar melting curve profiles with small shifts in Tm, disabling their differentiation. In such cases, further data treatment using specific HRM software is required to normalize data. This allows the removal of the fluorescence variance of the pre- and post-melting temperature regions, leaving the curve range between the bars as a new normalized plot that magnifies profile differences (Fig. 1c). To better visualize the differences between individual melting curves, some HRM software applications enable plotting the difference curve data obtained from subtracting the sample melting curves from a pre-defined reference set (Fig. 2).

A key issue in HRM analysis is the selection of the target region. Small-length amplicons (<300 bp) containing sequence variations, such as single-nucleotide polymorphisms (SNP) and small insertions or deletions, among a set of plant species or even cultivars are recommended [7]. DNA barcodes are informative short sequence of nuclear, plastidial, or mitochondrial regions with high potential to serve as taxon identifiers due to their genome low intraspecific and high interspecific variability [10]. The combination of HRM analysis with DNA barcodes, or more precisely mini-barcodes

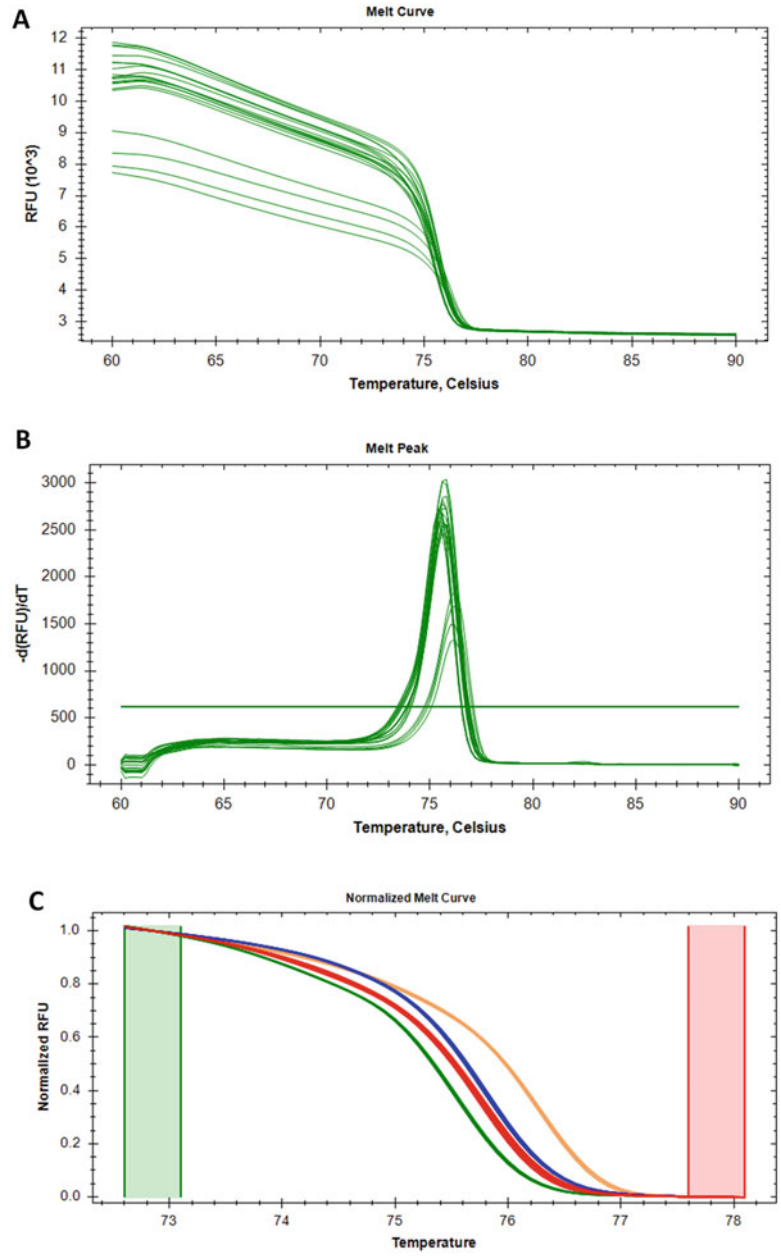


Fig. 1 HRM analysis applied in the differentiation of *Lavandula stoechas*, *L. pedunculata*, *L. viridis*, and wild *Lavandula*. Raw (**a**) and derivative (**b**) and normalized (**c**) melting curves

(<300 bp), has been designated as Bar-HRM and considered as a powerful tool to differentiate among closely related plant species. In opposition to animal species, mitochondrial regions are not recommended for plants because they present low evolutionary rates and low nucleotide substitution. Therefore, nuclear and

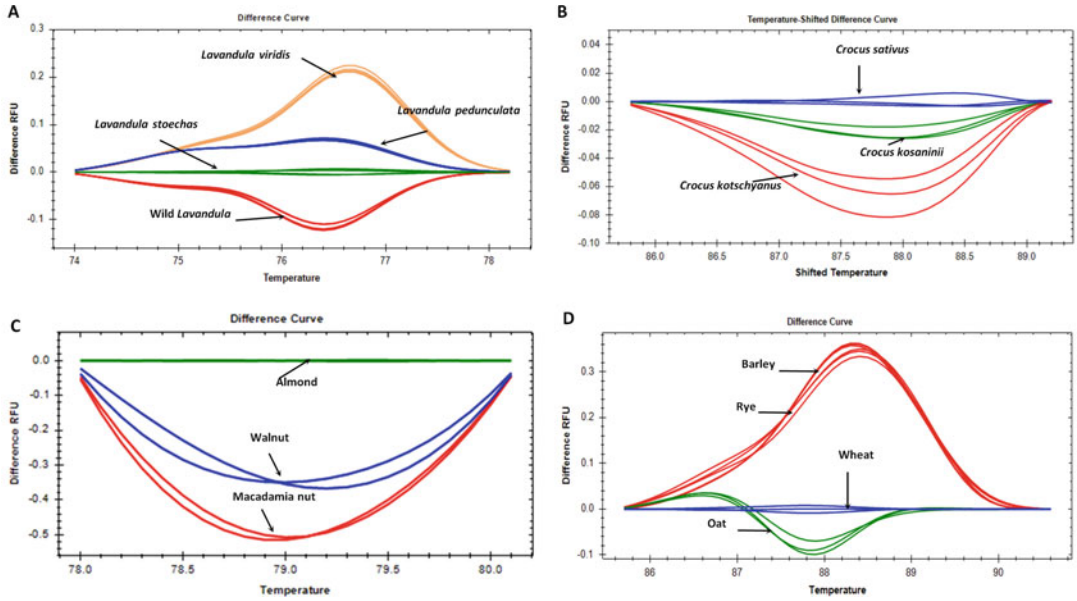


Fig. 2 Application of HRM analysis (difference melt curves) in the discrimination of plant material at species (**a** and **b**) and genus (**c** and **d**) levels

plastidial regions, such as ribulose-1,5-bisphosphate carboxylase oxygenase (*rbcL*), maturase k (*matK*), intergenic spacer regions (*trnH-psbA*), and internal transcribed spacers 1 and 2 (ITS, ITS2), have been proposed as alternative barcodes for plants. The choice of the best region is often an arduous and challenging task since there is no single locus that works as a universal plant barcode [10–12].

The use of DNA-mini-barcodes coupled with HRM analysis has been successfully applied in the discrimination of plant species in various products, including *Lavandula* spp. to determinate the botanical origin of honey [13] (Fig. 2a), different *Crocus* spp. in commercial saffron spices [14] (Fig. 2b), *Tinospora* spp. to authenticate herbal medicines [15], and *Hypericum* spp. to authenticate herbal infusions [3]. HRM analysis targeting an allergen-encoding gene was successfully applied to discriminate *Prunus dulcis* from other tree nuts [16] (Fig. 2c) and to differentiate wheat (*Triticum* spp.) from other gluten-containing cereals (rye, barley, and oat) in gluten-free foods [17] (Fig. 2d). Overall, HRM analysis is considered a fast and reliable tool to discriminate among closely related plant species, being considered also a cost-effective and high-throughput approach since it does not require any post-PCR analysis or sequencing as in several other DNA-based methods.

2 Materials

Prepare all solutions using ultrapure water (deionized water) and analytical grade reagents. Prepare and store all reagents at room temperature (unless indicated otherwise). Use molecular biology-grade consumables (e.g., tips, reaction tubes, PCR tubes, real-time PCR strips and caps) for DNA extraction and PCR analysis (sterile, DNase and RNase free). The rest of the materials and consumables, which can be bought in non-sterile conditions, should be chemically decontaminated (e.g., DNA-ExitusPlus, AppliChem GmbH, Darmstadt, Germany) or in-house autoclaved (121 °C, 15 min). The use of a PCR workstation, especially when manipulating the DNA extracts and PCR reagents, as well as when performing all tasks associated with the preparation of PCR or real-time PCR mixes, is highly recommended. All waste disposal regulations should be followed when disposing waste materials.

2.1 Target Genes and Software

1. Depending on the selected plant species to be differentiated, different coding genes (*rbcL* and *matK*), as well as noncoding spacers (*trnH-psbA*, ITS and ITS2), can be tested to discriminate DNA sequences at species level [18].
2. Table 1 lists software applications that can be used to search for the available DNA sequences, within genes or regions with high sequence homology, but having enough interspecific variability.

2.2 Reagents

1. DNA extraction: Nucleospin Plant II (Macherey-Nagel, Düren, Germany) DNA extraction kit (for alternatives *see* **Note 1**).
2. PCR mix: SuperHot *Taq* DNA polymerase (e.g., Genaxxon Bioscience GmbH, Ulm, Germany), chemically inactivated prior to an activation step (normally at 95 °C for several minutes), including respective 10× buffer and 25 mM of $MgCl_2$; PCR-grade water; 10 mM of dNTP mix and primers (forward and reverse) synthesized outsourced (e.g., Eurofins Genomics, Ebersberg, Germany).
3. Agarose gel electrophoresis of PCR products: 1.5% of agarose in 1× SGTB (Grisp, Porto, Portugal) or 2% agarose in TAE (40 mM Tris-acetate, 1 mM EDTA) buffer with 1× GelRed (Biotium Inc., Hayward, CA, USA); DNA marker (e.g., DNA 100 bp marker, Bioron GmbH, Römerberg, Germany); loading buffer (4% (w/v) sucrose, 0.05% (w/v) bromophenol blue, 0.12 M EDTA).
4. Purification of PCR products: GRS PCR and Gel Band Purification kit (Grisp, Porto, Portugal).

Table 1
Examples of algorithms available online for free use listed according to application

Software	Description	URL
Sequence databases		
NCBI database	National Center for Biotechnology Information provides access to biomedical and genomic information	https://www.ncbi.nlm.nih.gov/
Sequence alignment		
BLASTn	Finds regions of similarity between nucleotide sequences to sequence databases and calculates their statistical significance	https://blast.ncbi.nlm.nih.gov/Blast.cgi?PAGE_TYPE=BlastSearch
ClustalW	Multiple Sequence Alignment	https://www.genome.jp/tools-bin/clustalw
MEGA	Molecular Evolutionary Genetics Analysis	https://www.megasoftware.net/
Clustal Omega	Multiple Sequence Alignment	https://www.ebi.ac.uk/Tools/msa/clustalo/
BIOEDIT vs7.2	Biological Sequence Alignment Editor	https://bioedit.software.informer.com/versions/
Primer design		
Primer-Blast	Design primers specific to PCR template	https://www.ncbi.nlm.nih.gov/tools/primer-blast/
Primer3	Pick primers from a DNA sequence	http://bioinfo.ut.ee/primer3-0.4.0/
GenScript Online PCR Primers Designs Tool	Online tool to design PCR primers	https://www.genscript.com/tools/pcr-primers-designer
Eurofins Genomics PCR Primer Design Tool	PCR primer design tool analyzes the entered DNA sequence and chooses the optimum PCR primer pairs	https://www.eurofinsgenomics.eu/en/ecom/tools/pcr-primer-design/
Primer3Plus	Select primer pairs to detect the given template sequence	http://www.bioinformatics.nl/cgi-bin/primer3plus/primer3plus.cgi
Primer properties		
OligoCalc	Provide information regarding the physical properties of oligonucleotides, self-complementarity, and hairpin loop formation	http://biotools.nubic.northwestern.edu/OligoCalc.html
OligoEvaluator	Provide information about basic physical properties of oligonucleotides, formation of secondary structure, and primer dimer	http://www.oligoevaluator.com/LoginServlet

(continued)

Table 1
(continued)

Software	Description	URL
Sequencing analysis		
FinchTV	Viewing trace data from Sanger DNA Sequencing (scf or ab1 file formats)	https://digitalworldbiology.com/FinchTV
MEGA	Molecular Evolutionary Genetics Analysis	https://www.megasoftware.net/
BIOEDIT vs7.2	Biological Sequence Alignment Editor	https://bioedit.software.informer.com/versions/

5. Mix for real-time PCR with HRM: Use pre-prepared mixes for real-time PCR containing all the reaction components (enzyme, buffer, dNTP, and Mg^{2+}) and the new-generation fluorescent DNA-binding dye (e.g., EvaGreen) (e.g., SsoFast EvaGreen Supermix, Bio-Rad Laboratories, Hercules, CA, USA) (other alternatives, *see* **Note 2**); PCR-grade water; 10 mM of dNTP mix; and primers (forward and reverse) synthesized outsourced (e.g., Eurofins Genomics, Ebersberg, Germany).

2.3 Equipment

1. Refrigerated centrifuge (e.g., Heraeus Fresco 17, Thermo Scientific, Osterode am Harz, Germany).
2. Thermomixer block (e.g., Thermomixer comfort, Eppendorf AG, Hamburg, Germany).
3. Water bath (0–110 °C).
4. Vortex stirrer.
5. Microplate reader UV/Vis spectrophotometer, with microvolume plate accessory for nucleic acid and protein quantification (Synergy HT with Take 3 plate, BioTek, Winooski, VT, USA) (*see* **Note 3**).
6. Electrophoresis apparatus (electrophoresis tank and power supply).
7. PCR workstation with UV-cleaner-recirculator, UV light, and white lamp (e.g., VWR International GmbH, Darmstadt, Germany).
8. UV light photographic system (e.g., UV light tray Gel Doc™ EZ Imager, Bio-Rad Laboratories, Hercules, CA, USA).
9. Thermal cycler (e.g., MJ Mini personal thermal cycler, Bio-Rad Laboratories, Hercules, CA, USA).
10. High-resolution real-time PCR instrumentation (e.g., CFX96 real-time PCR system, Bio-Rad Laboratories, Hercules, CA, USA) capable of reading one fluorophore (FAM or SYBR

Green) and respective software for real-time PCR data treatment (Bio-Rad CFX manager 3.1, Bio-Rad Laboratories, Hercules, CA, USA), combined with the specific HRM software (Precision Melt Analysis version 1.3, Bio-Rad Laboratories, Hercules, CA, USA) (other alternatives, *see* **Note 4**).

3 Methods

3.1 In Silico Analysis

1. Select the gene or DNA region for the potential discrimination of the target species and check if there are consensus sequences available at NCBI database. For this purpose, the BLASTn algorithm (Table 1), also at NCBI database, can be used to search for DNA sequences based on their similarity.
2. Download and align the selected sequences using an alignment algorithm (e.g., BIOEDIT vs7.2) (Table 1) (*see* **Note 5**). Within alignment, search for regions of high homology to design primers, but make sure that the amplicons will have some nucleotide mismatches within the entire sequence to allow interspecific variability.
3. Design primers either manually or using primer designing tools, such as Primer-Blast (Table 1). Verify primers' proprieties (physicochemical parameters, absence of hairpins, 3′ complementary, and self-annealing) using specific algorithms (e.g., OligoCalc) (Table 1). Check the complementary of the designed primers toward the target sequences using the software Primer-Blast (Table 1) (*see* **Notes 6–8**).
4. Order primer synthesis in specialized outsourcing facilities (e.g., Eurofins Genomics, Ebersberg, Germany). This step can take 2 or 3 days, depending on the selected production facility.

3.2 DNA Extraction

1. To extract DNA from plant material, select an appropriate DNA extraction method, such as Nucleospin Plant II (Macherey-Nagel, Düren, Germany) (*see* **Notes 9** and **10**). Follow kit instructions performing minor alterations, if necessary. The example given below follows the protocol with PL1 buffer.
2. Weigh 20–100 mg of grounded (lyophilized or dried) plant material in a 2.0 mL sterile reaction tube. Add 400 μL of PL1 buffer (preheated at 65 °C) to each tube, make strong vortex, and incubate for 1 h at 65 °C in thermomixer (900 rpm). Make frequent vortex to the samples during the lysis.
3. After incubation, leave tubes at room temperature and add 10 μL of RNase A (10 mg/mL) for 5 min (other conditions can be used, *see* **Notes 11** and **12**).

4. Centrifuge the tubes at 4 °C, 17,000 × *g*, for 10 min. Remove the supernatant carefully to a new tube, transfer it to a Nucleospin filter column, and centrifuge for 2 min at 11,000 × *g* at room temperature.
5. Collect the filtrate to a 1.5 mL sterile reaction tube and add 450 μL of PC buffer (DNA-binding buffer). Mix gently by pipetting and transfer the entire volume to the Nucleospin plant II column. Centrifuge for 1 min, at 11,000 × *g* at room temperature, and discard the flow through (be aware that the column has a maximum volume of 700 μL, so it can only be loaded with 700 μL each time; repeat loading until the entire volume of sample has passed the column).
6. Wash the silica membrane (Nucleospin plant II column) with 400 μL of PW1 and centrifuge for 1 min at 11,000 × *g* (room temperature), discarding the flow through (first washing step).
7. Wash the silica membrane (spin column) with 700 μL and 200 μL of PW2, and centrifuge for 1 and 2 min at 11,000 × *g* (room temperature), respectively (second and third washing steps). After each centrifugation, always discard the flow through. Make sure that the column is dry after the final 2-min centrifugation (residues of ethanol will damage DNA extracts).
8. Place column in a new 1.5 mL sterile reaction tube, add 50 μL of elution buffer (PE) preheated at 65 °C, and incubate for 5 min at 65 °C. Elute through 1-min centrifugation at 11,000 × *g* (room temperature). Repeat last step, in order to obtain 100 μL of DNA extract.

3.3 Determination of DNA Yield and Purity

1. Use a microplate UV/Vis spectrophotometer instrument (BioTek Instruments, Inc., Winooski, VT, USA), starting with the calibration of the Take 3 microvolume plate accessory (16 spots) with 4 μL of pure water (e.g., PCR water).
2. Place 4 μL of each DNA extract (in duplicate) on the plate spot and read absorbencies at 260, 280, and 320 nm using the UV/Vis spectrophotometer microplate reader. The yield and purity of each DNA extract will be determined automatically, following the nucleic acid quantification protocol with sample type defined for double-strand DNA in the Gen5 data analysis software version 2.01 (BioTek Instruments, Inc., Winooski, VT, USA).
3. Dilute DNA extracts to a specific concentration (in the case of extracts from plant material, final DNA concentration of 5–10 ng/μL is highly recommended). Store DNA extracts and dilutions at −20 °C until analysis.

3.4 Qualitative PCR

1. Prepare the reaction mix by the addition of all the components for a total volume of 25 μL. Each reaction mix must contain PCR-grade water (volume adjusted to the amount of remaining reagents), buffer 10× (2.5 μL), 10 mM of dNTP (2.0 μL), $MgCl_2$ (final concentration of 1.5 up to 3.0 mM), primers (100–500 nM), and 1.0 U of SuperHot Taq DNA polymerase (0.2 μL).
2. Distribute 23 μL of reaction mix by each well or tube and add 2 μL of DNA template (10–20 ng). Positive (DNA from target species) and negative (no-template DNA) controls should be included. Close wells or tubes and place them on the thermal cycler.
3. Define program of temperatures in the thermal cycler. Example of a program: initial denaturation at 95 °C for 5 min; 40 cycles at 95 °C for 30 s, 65 °C (this temperature must be previously optimized along with the Mg^{2+} concentration for each primer pair) for 30 s, and 72 °C for 30 s; and a final extension at 72 °C for 5 min. The number of cycles must also be optimized for each primer pair.
4. To visualize the obtained amplicons, prepare a 1.5% agarose gel stained with GelRed 1× (Biotium Inc., Hayward, CA, USA). Mix 5–20 μL of PCR product with 1–4 μL loading buffer, apply to gel wells, and run electrophoresis using SGTB 1× (Grisp, Porto, Portugal) for 25–30 min at 200 V. For each gel, use a 100 bp DNA marker. If the DNA marker is not pre-stained, add loading buffer.
5. After the electrophoresis, visualize the agarose gel with a UV light tray Gel Doc EZ Imager using GelRed dye protocol. Record a digital image with Image Lab software version 5.2.1 (Bio-Rad Laboratories, Hercules, CA, USA) and analyze the results.

3.5 Real-Time PCR with HRM Analysis

1. When performing a real-time PCR run (e.g., CFX96 Real-Time PCR system) with HRM analysis, set the program of temperatures and design the plate following the steps defined by the software (e.g., Bio-Rad CFX manager 3.1).
2. Open the wizard setup of Bio-Rad CFX manager 3.1, and define the program of temperatures as the protocol. This program must include the real-time PCR amplification, followed by the melt curve. An example of program is presented in Fig. 3 (*see* **Note 13**).
3. Prepare the plate, by selecting the correct fluorophore (SYBR Green) and the plate type (white for white strips/plate, clear for clear strips/plates). Set the number of samples and replicates (3–4 replicates per sample by run are recommended), by defining their place in the plate (example in Fig. 4) (*see* **Note 14**).

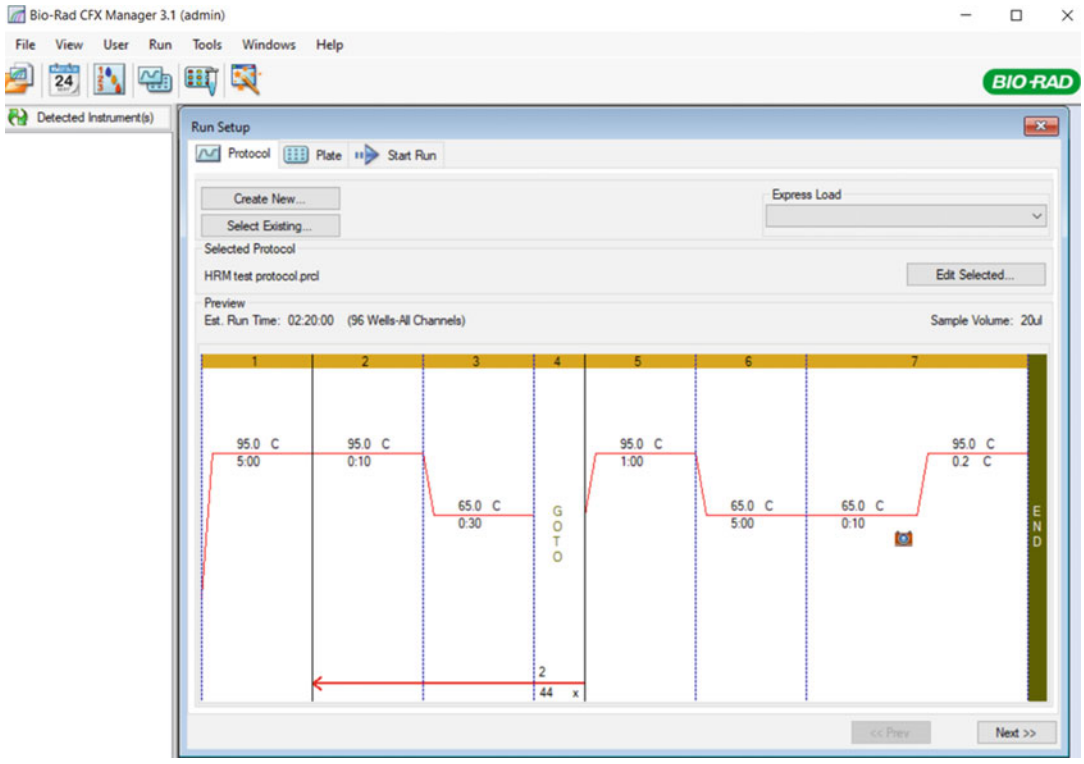


Fig. 3 Example of a real-time PCR program of temperatures with melt curve protocol adjusted for posterior HRM analysis

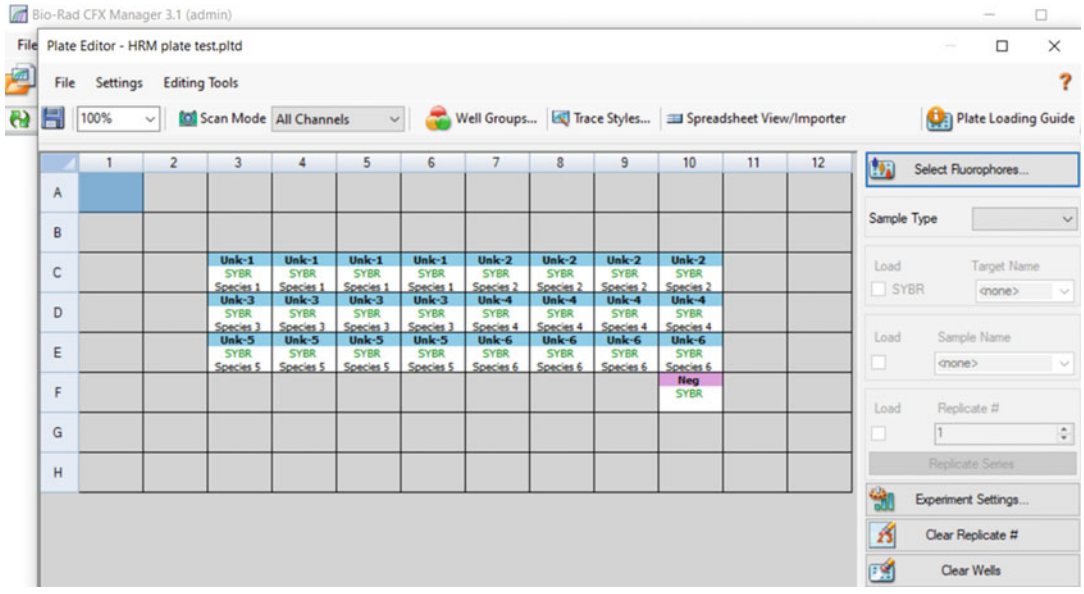


Fig. 4 Example of a real-time PCR plate

4. Prepare the reaction mix by adding all the components needed for all wells, except the DNA template. Each reaction mix must contain PCR-grade water (volume adjusted to the amount of remaining reagents), 1× of SsoFast EvaGreen Supermix (10 μL), and primers (100–500 nM) for a total of 20 μL reaction volume.
5. Place the necessary number of strips (or plate), distribute 18 μL of reaction mix per each well, and add 2 μL of DNA extract. Include a negative control (no-DNA template). Close wells and use a PCR spinner to ensure that all volume is at the bottom of the wells. Place the reaction strips (or plate) on the thermal cycler and start run.
6. After finishing the real-time PCR run, open the Precision Melt Analysis version 1.3 and create a new melt file by choosing the real-time PCR file recently generated by Bio-Rad CFX manager 3.1. Save the newly generated melt file.
7. Open the melt file and analyze the results (example in Fig. 5). The file is generated using predefined automated parameters, which might be adjusted with respect to the type of analysis that is being processed (*see* **Notes 15–18**).

3.6 DNA Sequencing for Method Validation

1. To validate HRM analysis, Sanger sequencing of PCR products of template species (species under study) is recommended (*see* **Notes 19** and **20**).
2. Follow the steps described in this section regarding "Qualitative PCR" to obtain the PCR products from the template species. Use a purification kit (e.g., GRS PCR and Gel Band Purification kit, Grisp, Porto, Portugal) in order to purify the amplified PCR products (removal of components from amplification reaction) according to manufacturers' instructions.
3. Send the purified PCR products for direct sequencing of both strands in opposite directions of each target amplicon in specialized research facilities (Eurofins Genomics, Ebersberg, Germany).
4. Check the quality of electropherograms using FinchTV software (or MEGA software) (Table 1). Only electropherograms with high quality should be analyzed and further aligned (BIOEDIT or MEGA software) (Table 1).
5. Critically analyze HRM results in relation to sequencing data.

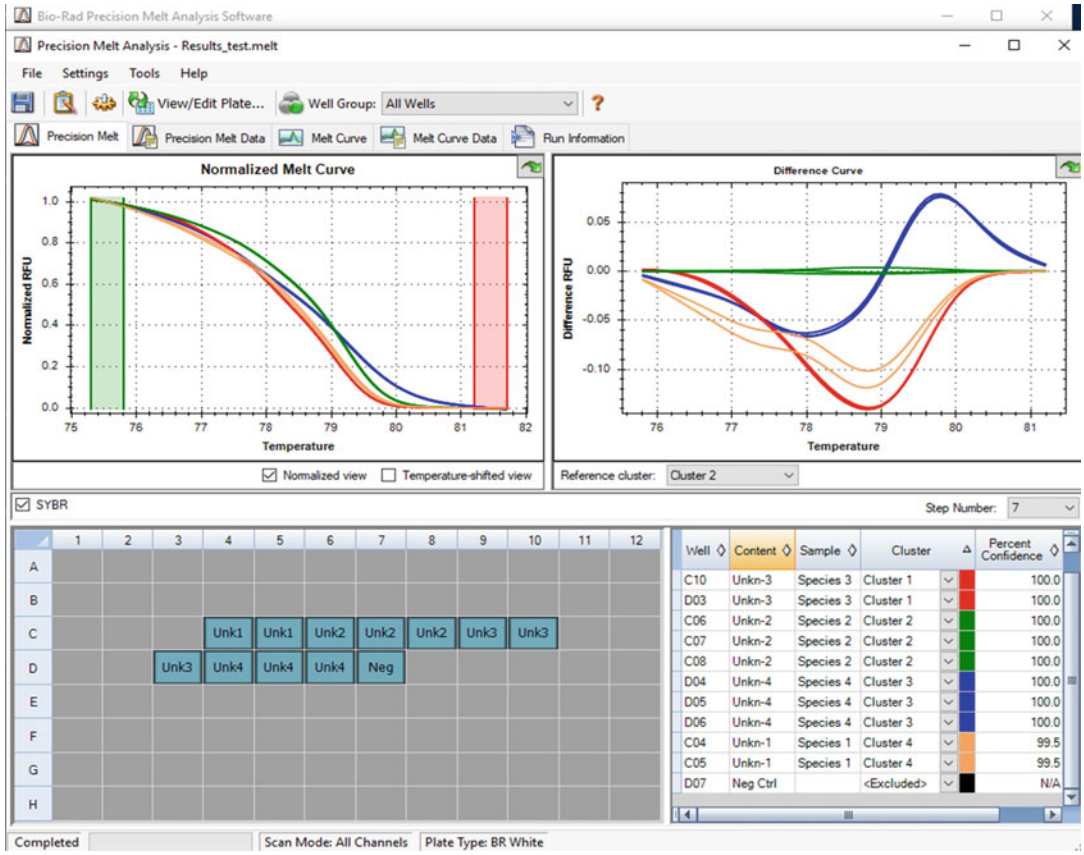


Fig. 5 Example of a melt file generated with the Precision Melt Analysis version 1.3, showing the normalized melt curve, the difference curve, the plate, and the classification of species by cluster (and respective degree of confidence)

4 Notes

1. Several commercial kits specialized in the DNA extraction of plant material can be used as alternative to Nucleospin Plant II (Macherey-Nagel, Düren, Germany) to extract DNA from plant source material, namely DNeasy Plant Mini kit (Qiagen GmbH, Hilden, Germany), E.Z.N.A. plant DNA DS Mini kit (Omega Bio-Tek, Inc., Norcross, GA, USA), and E.Z.N.A. SP Plant DNA Kit (Omega Bio-Tek, Inc., Norcross, GA, USA). Some of the commercial kits use lysis buffers with cationic (cetyltrimethyl ammonium bromide, CTAB) or anionic (sodium dodecyl sulfate, SDS) detergents, which can be combined with spin columns with silica-based membranes for retrieving high-purity DNA extracts. Nonetheless, most of these commercial kits allow extracting DNA with higher yield than most in-house-developed methods like the CTAB method, thus providing high-yield, -quality, and -purity plant

DNA extracts. Based on bead-beating technology, instead of the common detergents such as CTAB, DNeasy Plant Pro kit (Qiagen GmbH, Hilden, Germany) can also be considered a potential choice to extract high-quality cellular DNA from plant cells, tissues, and seeds. Besides commercial kits, the in-house-developed methods like the CTAB and wizard-based method [19, 20] can also be used.

2. One of the most well-known dyes for HRM analysis is the EvaGreen™, which can be used in a pre-prepared mix as in this protocol (SsoFast EvaGreen Supermix, Bio-Rad Laboratories, Hercules, CA, USA) or acquired separately (Biotium, CA, USA) and added to the in-house-prepared mix. EvaGreen is nonfluorescent, but it becomes highly fluorescent upon "release-on-demand" mechanism of binding to dsDNA. Besides, it is non-mutagenic and noncytotoxic, as it does not cross cell membranes. Other dyes, such as based on LC Green (addition of these dyes increases the melting temperature of DNA by 1–3 °C), Syto 9, and Chromofy (both dyes show high enhancement of fluorescence upon binding to double-stranded nucleic acid sequences), can also be used for successful HRM analysis. Examples of commercially available dyes for HRM application: LightCycler 480 ResoLight High Resolution Melting Dye (Roche Molecular Diagnostics Inc., Pleasanton, CA, USA), LCGreen PLUS (Idaho Technology Inc., Salt Lake City, Utah, USA), MeltDoctor™ HRM Master Mix (Applied Biosystems, Thermo Fisher Scientific, Waltham, MA, USA), and Type-it HRM PCR Kit (Qiagen GmbH, Hilden, Germany).
3. The use of a microvolume UV/Vis spectrophotometer is recommended to allow the direct reading of the extract, using as little extract as possible and avoiding extract dilutions and manipulations. Presently, the most well-known equipment is the nanodrop (e.g., NanoDrop™ 2000, Thermo Fisher Scientific, Delaware, USA), which is capable of doing measurements with just 1–2 μL, but it only allows one reading at a time, while the proposed system with the microvolume plate allows 16 reads simultaneously using similar volumes. To ensure more precise measurements, a volume of 4 μL of each extract is highly recommended.
4. There are other choices for high-resolution real-time PCR instruments (including their respective HRM analysis software): LightCycler® 480 Instrument II with LightCycler® 480 Gene Scanning Software (Roche Molecular Diagnostics Inc., Pleasanton, CA, USA), Rotor-Gene Q HRM System with Rotor-Gene ScreenClust HRM Software (Qiagen GmbH, Hilden, Germany), and 7500 Fast Real-time PCR System with High Resolution Melt (HRM) Software v2.0 (Applied Biosystems, Thermo Fisher Scientific, Waltham, MA, USA).

5. Besides BIOEDIT, there are other algorithms that can be used for DNA sequence alignment, either available online or for free download, namely ClustalW (Multiple Sequence Alignment), MEGA (Molecular Evolutionary Genetics Analysis), and Clustal Omega (Multiple Sequence Alignment) (Table 1).
6. In real-time PCR analysis, especially in HRM analysis, one important parameter to consider is the length of the amplicon. Therefore, to ensure a more accurate HRM analysis, the PCR products should range between 90 and 200 bp, although bigger fragments can also be used.
7. When designing primers manually, consider the following criteria/tips [21–23]: (a) the length of the primers should be 18–24 bp, with 40–60% G/C content (if possible, with the 3′ of a primer ending in C or G to promote binding); (b) avoid more than 1 or 2 G/C pairs at the 3′- and 5′-ends; (c) the Tm of the primers should range between 50 and 60 °C, although primers with Tm closer to 60 °C allow better amplifications; (d) the pair of primers should have closely matched melting temperatures to maximize PCR product yield (difference >5 °C between primers can lead to no amplification); (e) avoid runs of four or more of one base or dinucleotide repeats (for example, ACCCC or ATATATAT); and (f) primer pairs should not have complementary regions.
8. Several online platforms allow designing primers in alternative to the ones presented above. Examples: Primer3, GenScript Online PCR Primer Design Tool, Eurofins Genomics PCR Primer Design Tool, and Primer3Plus. Other software to check primers' properties: OligoEvaluator (Table 1).
9. Extracting DNA from plant material is often a challenge, which means that it is frequently recommended to test different kits or in-house-developed methods for the successful extraction of amplifiable DNA from certain plant species (e.g., seeds, spices, fruits, and leaves).
10. Nucleospin Plant II (Macherey-Nagel, Düren, Germany) allows extracting DNA from plant material using two different and independent protocols. One protocol is based on a lysis buffer containing the cationic detergent CTAB (PL1), while the other protocol uses an anionic (SDS) lysis buffer (PL2), which requires the subsequent precipitation of proteins by adding a potassium acetate solution (PL3). Both methods use a silica-based membrane combined with spin columns to ensure DNA extracts with high yield, quality, and purity.
11. In lysis step, the incubation can be performed for 10 min at 65 °C with previous addition of 10 μL of RNase A (10 mg/mL) (as suggested by the manufacturers), although longer incubation periods are recommended in order to increase DNA yields, which is the case of the example provided above.

12. For plant material, the use of RNase is normally recommended. Its use allows obtaining DNA extracts with more stability and with the adequate purity (1.8–2.0), thus enabling a better performance both by qualitative PCR and real-time PCR. However, care should be taken with its use since it also degrades DNA, reducing drastically the final yield.
13. Before starting the melt curve protocol, it is highly recommended to fully denature the PCR products and allow their correct annealing of the DNA complementary strands (DNA duplexes) by adding two steps, which were referred in the program of temperatures (denaturation of PCR products at 95 °C for 1 min, annealing of DNA duplexes at 65 °C for 5 min).
14. The use of white strips or plates is highly recommended for real-time PCR coupled to HRM analysis because they reduce the background noise and enhance fluorescence.
15. HRM analysis software (e.g., Precision Melt Analysis version 1.3, Bio-Rad Laboratories, Hercules, CA, USA) analyzes the fluorescence signal collected upon each increment of temperature during the melt curve. If the melt curve has big increments of temperature during small intervals of time (e.g., 0.5 °C for 10 s), the HRM software may not have enough data for a correct analysis. In such cases, the software accounts the fact that the generated file by the real-time PCR run with the melt curve does not comply with the recommended melt curve parameters. For optimal high-resolution melt data, the recommended increments of temperature during the melt curve should not exceed 0.2 °C between steps and a hold time minimum of 10 s during the melt curve protocol.
16. The melt file generated by the Precision Melt Analysis Software 1.3 (Bio-Rad Laboratories, Hercules, CA, USA) is analyzed by automated settings. The HRM analysis using the automated settings normally gives the best classification of samples into respective clusters, but depending on the type and number of nucleotide differences among the target sequences, some settings need to be adjusted. The software uses the data from the real-time PCR file and generates melting curves as a function of temperature, followed by the normalized melting curves and respective difference curves for easier visualization of the clusters. The melting curve shape sensitivity establishes the stringency used to categorize melting curves into different clusters. A high percentage value for this parameter allows increasing stringency and presents the results in more heterozygote clusters. The parameter of Tm difference threshold determines the lowest amount of Tm difference among samples. Like the melting curve shape sensitivity parameter, the Tm difference threshold defined to higher levels yields more heterozygote clusters.

17. By adjusting the settings, some nucleotide differences can be highlighted, while others can be neglected, which means that the HRM analysis needs to be controlled. This control is made by the level of confidence (expressed as percent of confidence). Each sample is mapped onto each cluster's probability distribution, based on their similarity to the mean melt curve across each sample in the cluster. The confidence value is an indication of the probability of a sample being included in a cluster; therefore the percentage of confidence should be as close as possible to 100%. Levels above 95% are normally considered as evidence of high confidence levels for the cluster classification.
18. The classification of samples into clusters is highly dependent on the type and number of nucleotide differences. Therefore, it is important to consider the samples/species that are intended to be separated by HRM analysis. When testing many genotypes containing several nucleotide differences, the classification into clusters might not distinguish groups with only one or two nucleotide differences. Even when testing few genotypes, but with few and many nucleotide differences among them, the classification into clusters might not be as evident as expected. For example, when testing three species, one with ten nucleotide differences comparing with two species with only one to two nucleotide mismatches between both, the products might be classified as two clusters instead of the expected three clusters. This can be explained by the high similarity of two species in relation to the third one, which might justify their inclusion in only one cluster in relation to a much distant cluster containing the ten-nucleotide difference species.
19. When developing a new real-time PCR method with HRM analysis, it is highly recommended to validate it by sequencing the PCR products of template species.
20. Before purifying PCR products for sequencing, it is highly recommended to perform an electrophoresis in 1.5% agarose gel (using only 2 μL of the amplified product) following the instructions described in Subheading 3 (Qualitative PCR).

Acknowledgment

This work was supported by FCT (Fundação para a Ciência e Tecnologia) under the Partnership Agreements UIDB/50006/2020 and UIDB/00690/2020. L. Grazina is grateful to FCT grant (SFRH/BD/132462/2017) financed by POPH-QREN (subsidized by FSE and MCTES).

References

1. Villa C, Costa J, Oliveira MBPP, Mafra I (2017) Novel quantitative real-time PCR approach to determine safflower (*Carthamus tinctorius*) adulteration in saffron (*Crocus sativus*). Food Chem 229:680–687. https://doi.org/10.1016/j.foodchem.2017.02.136
2. Böhme K, Calo-Mata P, Barros-Velazquez J, Ortea I (2019) Review of recent DNA-based methods for main food-authentication topics. J Food Agric Chem 67:3854–3864. https://doi.org/10.1021/acs.jafc.8b07016
3. Costa J, Campos B, Amaral JS, Nunes ME, Oliveira MBPP, Mafra I (2016) HRM analysis targeting ITS1 and *matK* loci as potential DNA mini-barcodes for the authentication of *Hypericum perforatum* and *Hypericum androsaemum* in herbal infusions. Food Control 61:105–114. https://doi.org/10.1016/j.foodcont.2015.09.035
4. Grazina L, Amaral JS, Mafra I (2020) Botanical origin authentication of dietary supplements by DNA-based approaches. Compr Rev Food Sci Food Saf 19:1080–1109. https://doi.org/10.1111/1541-4337.12551
5. Wittwer CT, Reed GH, Gundry CN, Vandersteen JG, Pryor RJ (2003) High-resolution genotyping by amplicon melting analysis using LCGreen. Clin Chem 49:853–860. https://doi.org/10.1373/49.6.853
6. Druml B, Cichna-Markl M (2014) High resolution melting (HRM) analysis of DNA – its role and potential in food analysis. Food Chem 158:245–254. https://doi.org/10.1016/j.foodchem.2014.02.111
7. Sun W, Li J-J, Xiong C, Zhao B, Chen S-L (2016) The potential power of Bar-HRM technology in herbal medicine identification. Front Plant Sci 7:367. https://doi.org/10.3389/fpls.2016.00367
8. Pereira L, Gomes S, Barrias S, Fernandes JR, Martins-Lopes P (2018) Applying high-resolution melting (HRM) technology to olive oil and wine authenticity. Food Res Int 103:170–181. https://doi.org/10.1016/j.foodres.2017.10.026
9. Montgomery JL, Sanford LN, Wittwer CT (2010) High-resolution DNA melting analysis in clinical research and diagnostics. Expert Rev Mol Diagn 10:219–240. https://doi.org/10.1586/erm.09.84
10. Hollingsworth PM, Forrest LL, Spouge JL, Hajibabaei M, Ratnasingham S, van der Bank M et al (2009) A DNA barcode for land plants. Proc Natl Acad Sci U S A 106:12794–12797. https://doi.org/10.1073/pnas.0905845106
11. 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
12. Techen N, Parveen I, Pan Z, Khan IA (2014) DNA barcoding of medicinal plant material for identification. Curr Opin Biotechnol 25:103–110. https://doi.org/10.1016/j.copbio.2013.09.010
13. Soares S, Grazina L, Costa J, Amaral JS, Oliveira MBPP, Mafra I (2018) Botanical authentication of lavender (*Lavandula* spp.) honey by a novel DNA-barcoding approach coupled to high resolution melting analysis. Food Control 86:367–373. https://doi.org/10.1016/j.foodcont.2017.11.046
14. Villa C, Costa J, Meira L, Oliveira MBPP, Mafra I (2016) Exploiting DNA mini-barcodes as molecular markers to authenticate saffron (*Crocus sativus* L.). Food Control 65:21–31. https://doi.org/10.1016/j.foodcont.2016.01.008
15. Osathanunkul M, Osathanunkul R, Madesis P (2018) Species identification approach for both raw materials and end products of herbal supplements from *Tinospora* species. BMC Complem Altern Med 18:111. https://doi.org/10.1186/s12906-018-2174-0
16. Costa J, Mafra I, Oliveira MBPP (2012) High resolution melting analysis as a new approach to detect almond DNA encoding for Pru du 5 allergen in foods. Food Chem 133:1062–1069. https://doi.org/10.1016/j.foodchem.2012.01.077
17. Martín-Fernández B, Costa J, de-los-Santos-Álvarez N, López-Ruiz B, Oliveira MBPP, Mafra I (2016) High resolution melting analysis as a new approach to discriminate gluten-containing cereals. Food Chem 211:383–391. https://doi.org/10.1016/j.foodchem.2016.05.067
18. Kress WJ (2017) Plant DNA barcodes: applications today and in the future. J Syst Evol 55:291–307. https://doi.org/10.1111/jse.12254
19. Lipp M, Brodmann P, Pietsch K, Pauwels J, Anklam E (1999) IUPAC collaborative trail study of a method to detect genetically modified soy beans and maize in dried powder. J AOAC Int 82:923–928
20. Mafra I, Silva SA, Moreira EJMO, da Silva CSF, Oliveira MBPP (2008) Comparative study of DNA extraction methods for soybean derived food products. Food Control 19:1183–1190.

https://doi.org/10.1016/j.foodcont.2008.01.004

21. Christensen H, Olsen JE (2018) Primer design. In: Christensen H (ed) Introduction to bioinformatics in microbiology. Springer International Publishing, Cham, Switzerland, pp 81–102
22. Primer design tips. https://www.thermofisher.com/pt/en/home/life-science/oligonucleotides-primers-probes-genes/custom-dna-oligos/oligo-design-tools.html. Accessed Jan 2020
23. Primer design for PCR. https://www.addgene.org/protocols/primer-design/. Accessed Jan 2020

Chapter 6

Specific-Locus Amplified Fragment Sequencing (SLAF-Seq) as High-Throughput SNP Genotyping Methods

Zhangsheng Zhu, Binmei Sun, and Jianjun Lei

Abstract

Most plant agronomic traits are quantitatively inherited. Identification of quantitative trait loci (QTL) is a challenging target for most scientists and crop breeders as large-scale genotyping is difficult. Molecular marker technology has continuously evolved from hybridization-based technology to PCR-based technology, and finally, sequencing-based high-throughput single-nucleotide polymorphisms (SNPs). High-throughput sequencing technologies can provide strategies for sequence-based SNP genotyping. Here we describe the SLAF-seq that can be applied as the SNP genotyping approach. The high-throughput SNP genotyping methods will prove useful for the construction of high-density genetic maps and identification of QTLs for their deployment in plant breeding and facilitate genome-wide selection (GWS) and genome-wide association studies (GWAS).

Key words Quantitative trait loci, SLAF-seq, High-throughput, SNP, Genotyping

1 Introduction

The yield, quality, and resistance traits are the most important agronomic traits of crop plants [1–3]. However, most of these traits are controlled by multiple major genes or quantitatively inherited [4, 5]. The use of molecular assisted selection (MAS) can shorten the breeding cycle and accelerate the breeding process of elite varieties [6]. Therefore, the identification of the QTLs or major genes that control these traits in various genetic backgrounds is imperative. The MAS depending on marker quality and density of the genetic map is performed for QTL mapping [7]. The high-quality and high-density genetic map could be used for detecting QTL for important traits, and the narrowed QTL interval or provided promising candidate genes to develop molecular markers for MAS [1, 2]. The quality and density of the genetic map highly depend on the genotyping method applied. During the last decades, extensive studies were dedicated to the development of genotyping methods [8–11]. Molecular marker technology has

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_6,

continuously evolved from hybridization-based RFLPs to PCR-based RAPDs, AFLPs, and SSRs, and next-generation sequencing (NGS)-based high-throughput SNPs (Table 1). Especially, conventional PCR-based molecular markers such as AFLP, RAPD, and SSR have played important roles in genotyping assays during the past two decades [8, 9]. However, with this strategy, the number of markers is too small to meet the requirements of high-throughput genotyping. Subsequently, the hybrid-based high-throughput genotyping method microarray analysis has been developed by Affymetrix company and this method with a fairly high cost-performance ratio. However, with microarray analysis, the distribution of the marker on the target genome is uncontrollable and unable to do de novo marker discovery. NGS technologies can be applied to discover large quantities of SNPs in the whole-genome scale [12]. Genotyping-by-sequencing (GBS) is fairly straightforward for small genomes; target enrichment or reduction of genome complexity must be employed to ensure sufficient overlap in sequence coverage for species with large genomes [10]. Reducing genome complexity with restriction enzymes is easy, quick, specific, and highly reproducible, and may reach important regions of the genome that are inaccessible to sequence capture approaches [7, 13]. Combining NGS with restriction enzyme digestion, several genotyping methods have been developed. Restriction site-associated DNA sequencing (RAD-seq) is a rapid and cost-effective polymorphism identification and genotyping method for high-density SNP discovery and genotyping [14]. Subsequently, a similar technology genotyping-by-sequencing (GBS) was developed [10]. However, compared with RAD-seq, the GBS procedure is substantially less complicated; the generation of restriction fragments with appropriate adapters is more straightforward [10, 15]. The selection of digested DNA fragment sizes is critically important for improving the efficiency of tag utilization. GBS does not select the size of the digested fragment before PCR amplification, and the RAD-seq conducts the size-selection step of the digested fragment before PCR amplification [12, 15]. However, traditional RAD-seq technology has shortcomings in more operation steps and shorter read length [12, 15]. SLAF-seq combines bioinformatics and RAD-seq technology [11]. SLAF-seq is an optimized version of double-digestion RAD-seq, specifically intended for large-scale genotyping experiments [11]. The enzymes and the sizes of the restriction fragments are optimized with training data to ensure even distribution and avoid repeats. The fragments are also selected over a tight range to optimize PCR amplification. This approach can effectively avoid repetitive sequences in the genome, develop SNP markers with uniform distribution in the genome, and improve the efficiency of molecular marker development. SLAF-seq is a fairly efficient genotyping approach for plants and has been widely used in many species

Table 1
Comparison between SLAF-seq and other genotyping technologies

	Marker discovery		Marker controllability				Throughput and cost		
Technology	De novo marker discovery	Genotyping method	Number controllability	Distribution controllability	Reduction of repetitive regions	Frequency normalization	Marker throughput	Sample throughput	Cost
RFLP	Yes	Electrophoresis	Uncontrollable	Uncontrollable	Weak	Unnecessary	Low	Low	Expensive
SSR	No	Electrophoresis	Controllable	Uncontrollable	Weak	Unnecessary	Low	Low	Expensive
AFLP	Yes	Electrophoresis	Uncontrollable	Weak	Weak	Unnecessary	Low	Low	Expensive
RAPD	Yes	Electrophoresis	Uncontrollable	Weak	Weak	Unnecessary	Low	Low	Expensive
Golden Gate	No	Hybridization	Controllable	Uncontrollable	Strong	Unnecessary	Medium	Medium	Medium
Affymetrix	No	Hybridization	Controllable	Uncontrollable	Strong	Unnecessary	High	Medium	Cheap
RAD-seq	Yes	SE sequencing	Uncontrollable	Weak	Medium	Low	High	Medium	Cheap
GBS	Yes	PE sequencing	Controllable	Strong	Strong	Strong	High	High	Cheap
SLAF-seq	Yes	PE sequencing	Controllable	Strong	Strong	Strong	High	High	Cheap

such as rice [16], soybean [16], cotton [17], and pepper [1] genotyping and QTL mapping. The SLAF-seq technique described here, including predesign analysis and experiment, which allows for efficient, low-cost, simple, and high-throughput genotyping.

2 Materials

Prepare all solutions using ultrapure water prepared by purifying deionized water, to attain a sensitivity of 18.2 MΩ cm at 25 °C. The chemical reagents were at analytical grade reagents.

2.1 Enzymes

*Hae*III restriction enzyme, DNA polymerase I, large (Klenow) fragment, T4 DNA ligase, T4 DNA ligase.

2.2 Kits and Reagents

- QIAquick PCR Purification Kit.
- Gel Extraction Kit, Phusion High-Fidelity PCR Master Mix with HF Buffer.
- E.Z.N.A. Cycle Pure Kit.
- Agarose.
- CTAB extraction buffer.
- dATP, 1 mM: Add 1 μL of 100 mM dATP to 99 μL of sterile water. Store at −20 °C for up to 1 year.
- dNTPs, 10 mM: Add 1 mL each of 100 mM dATP, dTTP, dCTP, and dGTP to 6 mL of sterile water. Aliquot into 1.5 mL tubes. Store at −20 °C for up to 1 year.
- 10× TE buffer: 100 mM Tris–HCl, 10 mM EDTA, pH 8.0. Sterilize solutions by autoclaving for 20 min at 121 °C on the liquid cycle. Store the buffer at room temperature until use.
- 3 M NaOAc, pH 5.2: Add 204.15 g sodium acetate NaOAc to 500 mL of doubly distilled water, and add the acetic acid to adjust the solutions with pH 5.2. Sterilize solutions by autoclaving for 20 min at 121 °C. Store the buffer at room temperature until use.
- Absolute ethyl alcohol (Sigma-Aldrich, cat. no. 792780).
- 70% (vol/vol) Ethanol, to 35 mL of absolute ethyl alcohol, add 15 mL of sterile water.
- Annealing buffer stock (10×): 100 mM Tris–HCl, (pH 8), 500 mM NaCl, 10 mM EDTA.

2.3 Equipment

- Microcentrifuge tubes, 1.5 mL.
- Microcentrifuge (Eppendorf, Germany).
- Thermocycler (Bio-Rad, USA).
- Low-binding 96-well PCR plates.

- Tube/plate rotator.
- Gel electrophoresis system.
- Gel visualization system.
- NanoDrop One spectrophotometer.

3 Methods

3.1 Plant Materials and DNA Extraction

1. Extract high-quality DNA from 0.1 g young healthy leaves by the method of CTAB [18].
2. Dissolve DNA dissolved in 1× TE for a total of 30 μL.
3. The DNA sample concentration and purity are quantified with NanoDrop One spectrophotometer and integrity by electrophoresis in a 1% agarose gel with lambda DNA as a standard. Ensure that the DNA concentration is higher than 50 ng/μL, and the amount of DNA is higher than 5 μg for each sample.

3.2 Predesign Experiment

1. Flowchart of SLAF: SLAF-seq is an efficient method of large-scale genotyping, which is based on the reduced representation library (RRL) and high-throughput sequencing [11]. The procedure is shown in Fig. 1. All experimental procedures are carried out at room temperature unless otherwise specified. The results presented in this protocol were used in the *Capsicum* SLAF-seq data. The *Hae*III enzyme and sizes of restriction fragments were evaluated accordingly to the training data pepper (*Capsicum annuum*) reference genome. Three criteria were considered: (a) The number of SLAFs must be suitable for the requirements of the research goal; (b) repeated SLAFs should be avoided; and (c) the SLAFs must be random and evenly distributed throughout the genome sequences to be studied. Taking these into considerations can improve the efficiency of SLAF-seq (*see* **Note 1**). The simulation of SLAF distribution is presented in Fig. 1a, b.
2. To maintain the sequence depth uniformity of different fragments, a tight length range of about 30–50 bp is used (Fig. 1a).
3. A pilot PCR amplification was performed to check the RRL features in this target length range, which would ordinarily include fragments with similar amplification features on the gel (Fig. 1c). When nonspecific amplified bands appear on the gel, we will repeat the predesign step to produce a new scheme. After that the qualified library is subject to deep sequencing according to the Illumina paired-end sequencing protocol (Fig. 1d).

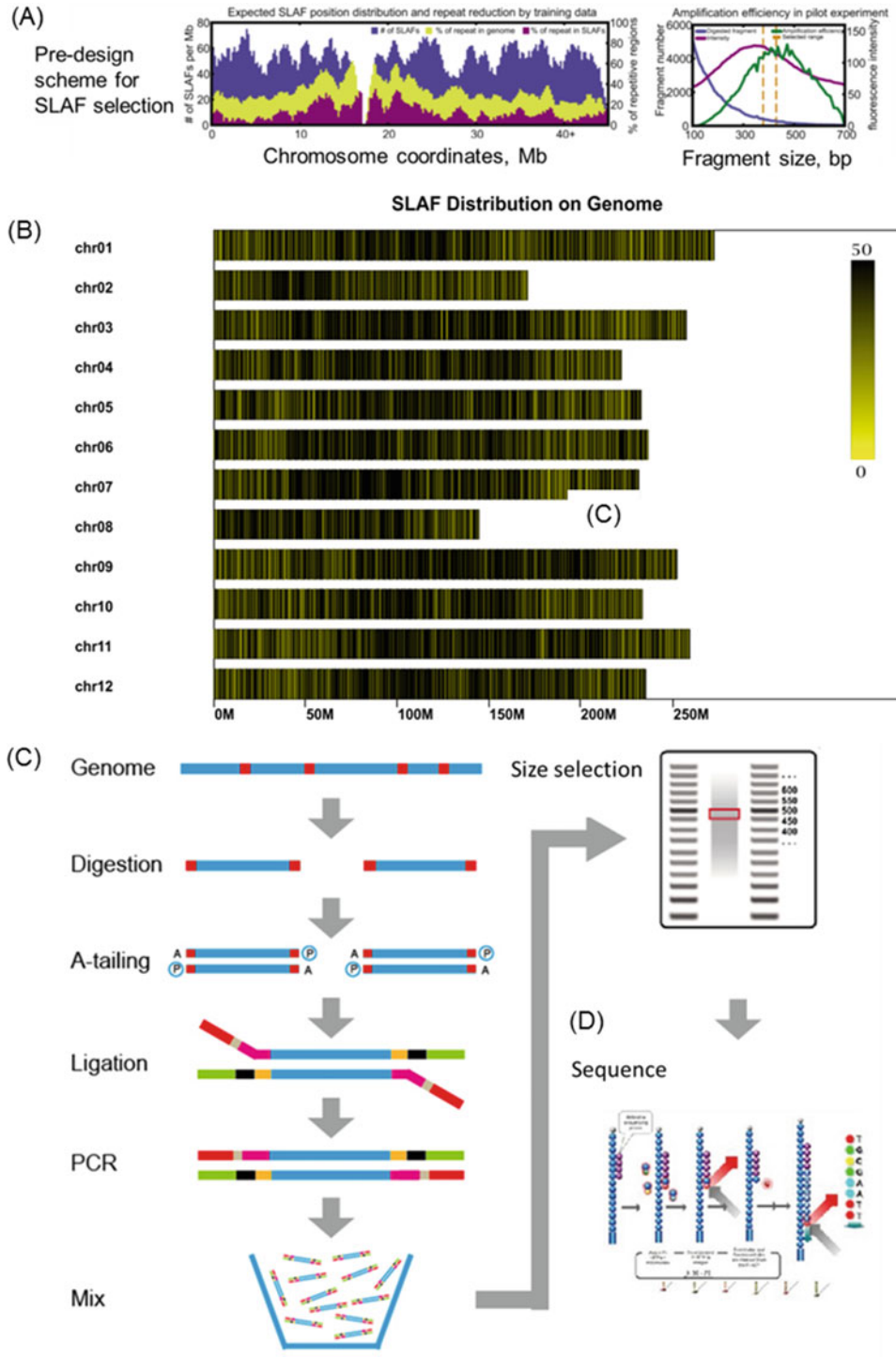


Fig. 1 Flowchart of SLAF. (**a**) Predesign scheme for SLAF selection using training data. The reduced representation design must be decided based on marker efficiency characteristics, which include random distribution throughout the genome, uniqueness in the genome, and consistent amplification efficiency among

3.3 SLAF Library Construction

The SLAF library construction is done in accordance to the predesigned scheme.

1. To create an adapter, combine each oligo with its complementary oligo in a 1:1 ratio in annealing buffer for a final concentration at 1× buffer for the concentration of 40 μM at a total of 100 μL.
2. Incubate at 96 °C for 2.5 min at a thermocycler, and then cool at a rate of not greater than 3 °C/min until the solution reaches a temperature of 25 °C. Store the solution at −20 °C for up to 6 months.
3. Prepare final working strength concentrations of annealed adapters from this annealed stock. For convenience, it is possible to store the adapters at 4 °C while in active use.
4. The reaction components are 0.5–1 μg of genomic DNA, 10× CutSmart Buffer, and 2 μL of *Hae*III, and add ddH_2O to a final amount of 50 μL. To scale up, multiply all reaction amounts proportionately.
5. The reaction was incubated at 37 °C for 3 h. Restriction-ligation reactions were heat-inactivated at 80 °C for 20 min.
6. Sample purification: Add 5 μL of 3 M NaOAc and 100 μL of ice-cold ethanol to the sample. Vortex for 5 s to mix it.
7. Incubate at −20 °C freezer for at least 20 min. Centrifuge at 20,000 × *g* for 10 min at 4 °C to pellet the DNA.
8. Discard the supernatant by decanting. Wash the pellet with 1 mL of 70% ethanol.
9. Centrifuge at 20,000 × *g* for 10 min at 4 °C. Discard the supernatant, quick-spin for 5 s at 3000 × *g* at room temperature, and pipette off any remaining ethanol, being careful not to disturb the DNA pellet.
10. Allow the pellet to dry for 10–15 min at room temperature. Be sure that the pellet is completely dry before resuspending.
11. Resuspend the DNA pellet in 32 μL of nuclease-free water. Place it at 37 °C for 5 min to help dissolve the DNA.

Fig. 1 (continued) selected markers. A pilot experiment is performed to evaluate the amplification efficiency based on the predesigned scheme (refer to [11]). (**b**) All SLAFs (black lines) distributed on 12 chromosomes. *x*-coordinate numbers indicate the length of the chromosome. The yellow bar indicates a chromosome. The black line indicates SLAF, and the chromosomes were displayed in 1 M slide window, and the deeper color indicates more SLAF tags. (**c**) SLAF library construction. The optimal restriction enzymes are selected to digest the native genomic DNA according to the predesign experiment. The enzyme-digested DNA fragments are subject to an A-tailing procedure. The A-tailing DNA fragments are ligated with dual-index sequence adaptors. The diluted restriction ligation samples are amplified by PCR with a specific primer containing a barcode. The PCR productions are purified and pooled. The DNA fragment with indicated seize is selected. (**d**) The qualified library is subject to deep sequencing according to the Illumina paired-end sequencing protocol

12. A-tailing: Add the following reagents to set up the A-tailing reaction: 5 μL of 10× NEBuffer2, 10 μL of 1 mM dATP, 3 μL of Klenow fragment (3′–5′ exo-; 5 U/μL), and 32 μL of digested DNA fragment.
13. Mix gently and quick-spin at room temperature for 5 s at 3000 × *g*. Incubate at 37 °C for 30 min.
14. Sample cleanup: Add 5 μL of 3 M NaOAc and 100 μL of ice-cold absolute ethanol to the sample. Vortex to mix it.
15. Repeat **steps 7–9** to perform ethanol precipitation.
16. Resuspend DNA in 32 μL of 1× TE. Place the sample at 37 °C for 5 min to aid resuspension.
17. Adapter ligation: Add 5 μL T4 DNA ligase, 5 μL 10 mM ATP, and 8 μL of adapter to the 32 μL DNA sample to a final volume of 50 μL. Samples were incubated at 22 °C for 1 h and heated to 65 °C for 30 min to inactivate the T4 ligase.
18. Sample cleanup: Add 5 μL of 3 M NaOAc and 100 μL of ice-cold absolute ethanol to the sample. Vortex to mix it. Repeat **steps 7–9** to perform ethanol precipitation.
19. Resuspend DNA and place the sample at 37 °C for 10 min to aid resuspension. Dilute to an appropriate concentration of 5 ng/μL. Assemble the PCR reaction components: 2.5 μL of 10 μM forward *Hae*III primer, 2.5 μL of 10 μM forward *Hae*III primer, 25 μL of 2× Phusion master mix, and 20 ng of template DNA; add ddH_2O to a final volume of 50 μL.
20. Load the samples into a thermocycler and run the following program: denature at 94 °C for 2 min; denature at 98 °C for 30 s; 98 °C for 15 s and 60 °C for 30 s to anneal; 72 °C for 1 min to extend; repeat the steps for 10–20 times; extend to 72 °C for 5 min.
21. Using E.Z.N.A. Cycle Pure Kit purify and pool the PCR productions. The pooled sample is incubated at 37 °C with *Hae*III primer, T4 DNA ligase, ATP, and Solexa adapter. Using a Quick Spin column purify the sample, and then run out on a 2% agarose gel.
22. Isolate fragments with indicated length (with indexes and adaptors) in size (*see* **Note 2**) using a gel extraction kit. Elute DNA samples in a final volume of 50 μL.
23. Assemble the PCR reaction components: 5 μL of 10 μM Solexa amplification primer mix, 25 μL of 2× Phusion master mix, and 20 ng of template DNA; add ddH_2O to a final volume of 50 μL. PCR amplification of the fragment products with Phusion master mix and Solexa amplification primer mix is done to add barcode 2.

24. Phusion PCR settings are as listed in the Illumina sample preparation guide. Purify the samples with gel, excising DNA in designed length, which is diluted for sequencing.
25. The paired-end sequencing is performed upon the selected SLAFs using an Illumina high-throughput sequencing platform (Illumina, USA).

3.4 SLAF Library Evaluation

1. Sequencing raw data process: Clean the sequencing adapter from the reads, filter out reads with 10% of the uncertain base, and trim 4–5 bp terminal sites.
2. Read quality evaluation: Filter out low-quality reads with quality score <30e (30e indicates that the average of base error probability was 0.001). Presentation of simulation data of *Capsicum* reads can be seen in Fig. 2a.
3. Distribution-type base checks are used to detect AT and GC separation phenomena, which may be caused by sequencing or library construction, and can affect subsequent analysis (Fig. 2b).

3.5 SLAF-Seq Data Grouping and Genotyping

1. Map the clean reads to the reference genome using Burrows-Wheeler Aligner (BWA) software. Reads are defined as the same SLAF if they are mapped on the same position with over 90% identity (Fig. 3a).
2. Use the GATK software and Samtools/bcftools [19] to detect SNPs between the parents.
3. Filter out low-quality SNPs. The process is based on the following criteria: (a) minimum read depth less than 10×; (b) average base quality less than 30e; (c) SNPs in each offspring individual anchored on different positions; and (d) SNPs in offspring with more than 40% missing data (*see* **Note 3**).

4 Genotyping

Use the MAF evaluation to define alleles in each SLAF locus (Fig. 3b). For further genetic analysis, the polymorphic SLAFs are transformed into genotype code with universal genetic two allele rule and the code (Fig. 3c). The simulation read data are retrieved from *C. annuum* and *C. chinense* that are diploid species; one locus could harbor at most four SLAF tags; locus containing more than four tags are filtered out as repetitive SLAFs, and those with two, three, and four tags are identified as polymorphic SLAFs. F_2 population is obtained from a cross of two *Capsicum* inbred lines with the genotype aa or bb; therefore, only the SLAF markers, which had segregation patterns of aa × bb (Fig. 3d), are used in further analysis (*see* **Note 4**).

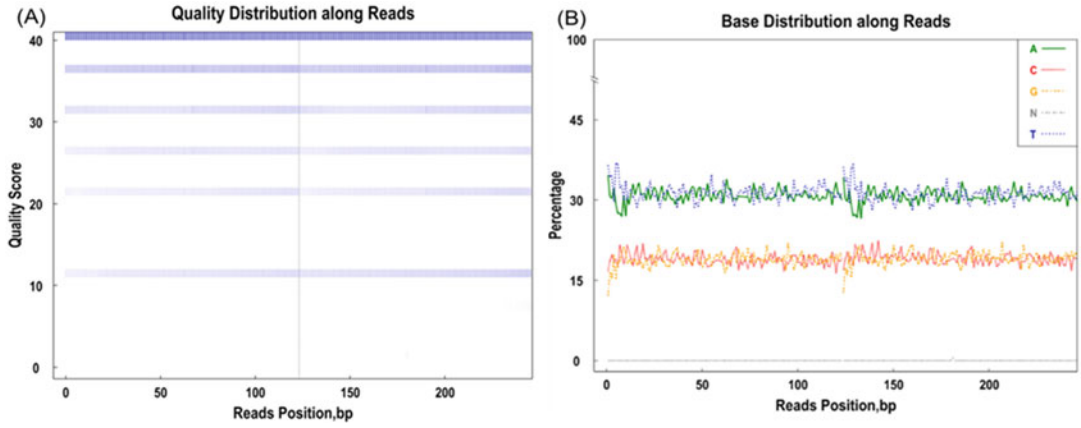
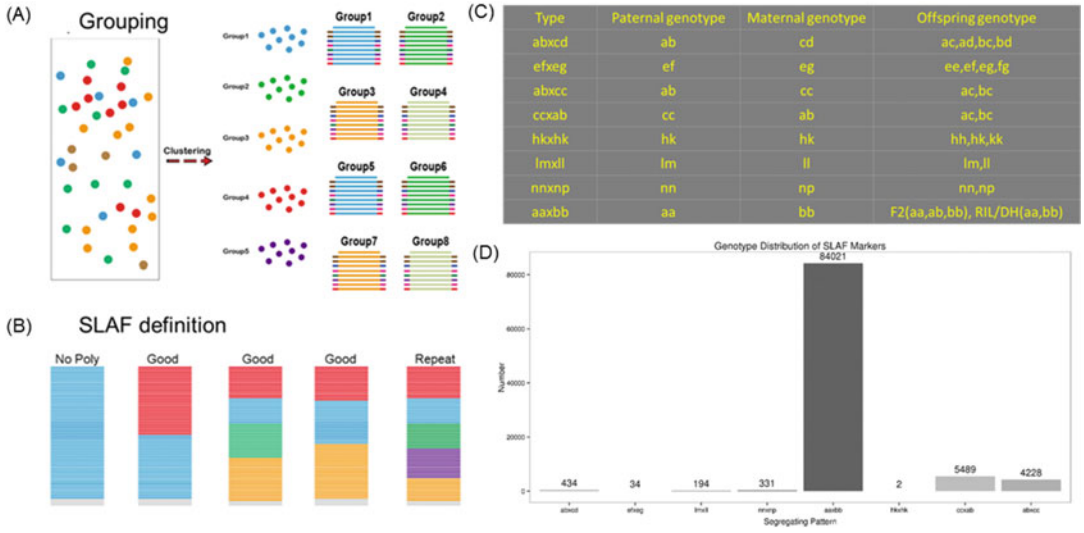


Fig. 2 Sequencing data quality evaluation. (**a**) Sequencing quality value distribution evaluation. The *x*-coordinate is the base position in the reads, and the *y*-coordinate is the single base error rate. The first 100 bp is the error rate distribution of the first end of the two-ended sequencing sequence, and the error rate distribution of the second 100 bp is the sequence at the other end. (**b**) Base distribution evaluation



Type	Paternal genotype	Maternal genotype	Offspring genotype
abxcd	ab	cd	ac,ad,bc,bd
efxeg	ef	eg	ee,ef,eg,fg
abxcc	ab	cc	ac,bc
ccxab	cc	ab	ac,bc
hkxhk	hk	hk	hh,hk,kk
lmxll	lm	ll	lm,ll
nnxnp	nn	np	nn,np
aaxbb	aa	bb	F2(aa,ab,bb), RIL/DH(aa,bb)

Fig. 3 Genotype definition process of SLAF-seq. (**a**) Samples were distinguished by barcodes and data grouping by sequence similarity. (**b**) Minor allele frequency (MAF) filtering and SLAF definition. The diploid species' one locus could harbor at most four SLAF tags; locus containing more than four tags was filtered out as repetitive SLAFs, and those with two, three, and four tags were identified as polymorphic SLAFs. One group of more than five seed tags is recognized as a repeat. (**c**) Eight potential segregating patterns of genotype codes. (**d**) The number of markers for eight segregation patterns. The *x*-axis indicates eight segregation patterns of polymorphic SLAF markers; the *y*-axis indicates the number of markers. An F_2 population is obtained from a cross of two pepper inbred lines with the genotype aa or bb; therefore, only the SLAF markers, which had segregation patterns of aa $\times$ bb, are used in map construction

5 Notes

1. The applied restriction enzyme used for DNA restriction depends on the preliminary simulation experiment. Firstly, for the double digestion, the reaction buffer compatibility among two restriction enzymes should be taken into consideration. In addition, the applied enzyme must keep the following criteria: the number of SLAFs must be suitable for the requirements of the research goal such as the requirements of sequence depth are varied among genetic map construction, QTL-seq, BAS--seq, and GWAS analysis; repeated SLAFs should be avoided and the SLAFs must be random and evenly distributed throughout the genome sequences to be studied.
2. DNA fragment lengths have considerable influence on the PCR amplification efficiency. In previous studies, because the predesign step is ignored, researchers tended to select relatively long length ranges to identify as many enzyme sites as possible leading to the selection of fragments with different copy numbers. This reduces sequence efficiency. Generally, the DNA fragments with 300–500 bp in length have a relative amplification efficiency. In addition, the selection of a tighter length range of about 30–100 bp may help to obtain fragments with similar copy numbers and to ensure similar sequence depths among fragments.
3. The quality of SLAF read is crucial for genotyping. Previous studies showed that the error ratio of genotype calling dropped greatly from 1× to 4× and that further increases in sequencing depth above 4× sequencing depth had relatively little influence on sequencing error rates. To ensure the quality of the SLAF, the minimum read depth should be more than 10×; however, it also reported that reads with higher than 6× in depth are also acceptable for analysis. Although it recommended that the SNPs in offspring with more than 40% missing data should be filtered. But this depends on the number of polymorphic SLAFs that were obtained if the SLAF met the requirement of the project; only 15% SNP missing data in the individual can also be omitted.
4. In the genotyping procedure, the ploidy and population type of plants must be taken into consideration. For diploid species, one SLAF locus can contain no more than four allele tags, so SLAF loci with more than four alleles and SLAFs with two to four alleles were identified as polymorphic markers. All polymorphism SLAF loci were genotyped with consistency in the parental and offspring SNP loci. We must keep in mind that the marker code of the polymorphic SLAFs was analyzed according to the population type. Generally, aa × bb segregation type is

always applied in inbreed population such as F_2, RIL, and DH population genotyping. ab × cd, ef × eg, ab × cc, cc × ab, hk × hk, lm × ll, and nn × np segregation types are always applied in the cross-pollination populations.

Acknowledgments

This work was supported by the National Key Research and Development Program (2018YFD1000800), National Natural Science Foundation of China (31572124).

References

1. Zhu Z, Sun B, Wei J, Cai W, Huang Z, Chen C et al (2019) Construction of a high density genetic map of an interspecific cross of *Capsicum chinense* and *Capsicum annuum* and QTL analysis of floral traits. Sci Rep 9(1):1054
2. Collard BC, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos Trans R Soc Lond Ser B Biol Sci 363 (1491):557–572
3. Cuthbert JL, Somers DJ, Brule-Babel AL, Brown PD, Crow GH (2008) Molecular mapping of quantitative trait loci for yield and yield components in spring wheat (*Triticum aestivum* L.). Theor Appl Genet 117 (4):595–608
4. Miao L, Yang S, Zhang K, He J, Wu C, Ren Y et al (2020) Natural variation and selection in *GmSWEET39* affect soybean seed oil content. New Phytol 225(4):1651–1666
5. Liang Y, Liu Q, Wang X, Huang C, Xu G, Hey S et al (2019) ZmMADS69 functions as a flowering activator through the ZmRap2.7-ZCN8 regulatory module and contributes to maize flowering time adaptation. New Phytol 221 (4):2335–2347
6. Mohan M, Nair S, Bhagwat A, Krishna TG, Yano M, Bhatia CR et al (1997) Genome mapping, molecular markers and marker-assisted selection in crop plants. Mol Breed 3 (2):87–103
7. Jena KK, Mackill DJ (2008) Molecular markers and their use in marker-assisted selection in rice. Crop Sci 48(4):1266–1276
8. Williams JGK, Kubelik AR, Livak KJ, Rafalski JA, Tingey SV (1990) DNA polymorphisms amplified by arbitrary primers are useful as genetic markers. Nucleic Acids Res 18 (22):6531–6535
9. Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Hornes M et al (1995) AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res 23(21):4407–4414
10. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES et al (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6(5):e19379
11. Sun X, Liu D, Zhang X, Li W, Liu H, Hong W et al (2013) SLAF-seq: an efficient method of large-scale *de novo* SNP discovery and genotyping using high-throughput sequencing. PLoS One 8(3):e58700
12. Peterson BK, Weber JN, Kay EH, Fisher HS, Hoekstra HE (2012) Double digest RADseq: an inexpensive method for *de novo* SNP discovery and genotyping in model and non-model species. PLoS One 7(5):e37135
13. Gore MA, Chia JM, Elshire RJ, Sun Q, Ersoz ES, Hurwitz BL et al (2009) A first-generation haplotype map of *Maize*. Science 326 (5956):1115–1117
14. Miller MR, Dunham JP, Amores A, Cresko WA, Johnson EA (2007) Rapid and cost-effective polymorphism identification and genotyping using restriction site associated DNA (RAD) markers. Genome Res 17(2):240–248
15. Gore M, Bradbury P, Hogers R, Kirst M, Verstege E, van Oeveren J et al (2007) Evaluation of target preparation methods for single-feature polymorphism detection in large complex plant genomes. Crop Sci 47:S135–S148
16. Yang X, Xia X, Zeng Y, Nong B, Zhang Z, Wu Y et al (2018) Identification of candidate genes for gelatinization temperature, gel consistency and pericarp color by GWAS in rice based on SLAF-sequencing. PLoS One 13(5):e0196690

17. Zhang Z, Shang H, Shi Y, Huang L, Li J, Ge Q et al (2016) Construction of a high-density genetic map by specific locus amplified fragment sequencing (SLAF-seq) and its application to Quantitative Trait Loci (QTL) analysis for boll weight in upland cotton (*Gossypium hirsutum*). BMC Plant Biol 16(1):79
18. Porebski S, Bailey LG, Baum BR (1997) Modification of a CTAB DNA extraction protocol for plants containing high polysaccharide and polyphenol components. Plant Mol Biol Reporter 15(1):8–15
19. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al (2009) The sequence alignment/map format and SAMtools. Bioinformatics 25(16):2078–2079

Chapter 7

Effective Mapping by Sequencing to Isolate Causal Mutations in the Tomato Genome

Fernando J. Yuste-Lisbona, José M. Jiménez-Gómez, Carmen Capel, and Rafael Lozano

Abstract

Forward genetic analysis remains as one of the most powerful tools for assessing gene functions, although the identification of the causal mutation responsible for a given phenotype has been a tedious and time-consuming task until recently. Advances in deep sequencing technologies have provided new approaches for the exploitation of natural and artificially induced genetic diversity, thus accelerating the discovery of novel allelic variants. In this chapter, a mapping-by-sequencing forward genetics approach is described to identify causal mutations in tomato (*Solanum lycopersicum* L.), a major crop species that is also a model species for plant biology and breeding.

Key words Mapping-by-sequencing, Mutations, Gene discovery, Tomato, *Solanum lycopersicum*

1 Introduction

Tomato (*Solanum lycopersicum* L.) is a commercially important crop throughout the world because of its high nutritive value for both fresh market and processing industries, but it is also regarded as a model species for studying developmental processes, especially for fleshy fruit biology [1, 2]. As a research model, tomato presents many interesting agronomic and genetic features, such as short life cycle, high multiplication rate, self-pollination, ease of mechanical crossing, suitability for genetic manipulation, and availability of a high-quality full genome sequence [3, 4]. In addition, whole-genome sequencing data from hundreds of tomato cultivars and most wild tomato species are available [5–11], as well as numerous induced mutant collections for forward genetic screens [12–16]. In this regard, a large number of these mutants have already proven their utility as powerful tools for assessing gene functions.

Fernando J. Yuste-Lisbona and José M. Jiménez-Gómez contributed equally with all other contributors.

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_7,

Nevertheless, the identification of the causal mutation underlying a particular phenotype had until recently been a laborious and time-consuming process. The availability of reference genome sequences and the decreasing costs of high-throughput sequencing have accelerated the cloning process by combining genetic mapping with whole-genome sequencing, an approach known as mapping-by-sequencing that has been successfully carried out in different model species such as *Saccharomyces cerevisiae* [17], *Caenorhabditis elegans* [18], *Drosophila melanogaster* [19], and *Arabidopsis thaliana* [20].

Forward genetics approach to isolate a genetic variant that is responsible for a particular mutant phenotype is a multistep process that usually involves genetic mapping to localize the chromosomal region harboring the mutation of interest, followed by searching for candidate mutations in genes within this genomic region, and subsequent validation through functional approaches [21]. In the genetic mapping step, a mutant displaying a phenotype of interest is used to generate a mapping population segregating for both wild-type and mutant individuals. Attempts to reduce the scale, costs, and complexity of mutation mapping were introduced several decades ago through bulked-segregant analysis [22]. With this method, DNA from the mapping population is pooled based on the phenotypes of individuals and then compared using molecular markers [22] or, more recently, whole-genome sequencing [21]. High-throughput sequencing reveals nucleotide changes by aligning the sequencing reads to the reference genome and allows for the quantification of allele frequency ratios in contrasting pooled samples in order to detect biases uncovering linkage to the mutant phenotype. At the unlinked genomic regions, approximately 50% of the reads should come from each parental genome, whereas reads closely linked to the causal mutation should only derive from the mutant parent. Finally, the identified candidate genomic region is filtered for unique homozygous genetic variants, among which will be the causal mutation responsible for the mutant phenotype of interest. Thereby, mapping-by-sequencing has become a powerful and efficient method for gene discovery that overcomes the inherent difficulties of other gene cloning procedures.

Mapping-by-sequencing has been successfully employed to identify causal mutations in a tomato EMS-induced mutant population [23]. In this strategy, the mapping population was created by backcrossing a homozygous mutant plant to its non-mutagenized parent, and then self-crossing the resulting BC_1F_1 hybrid to generate a BC_1F_2 progeny that segregated for the trait of interest. However, the efficient identification of the causal mutation depends on both the density of EMS mutations and the extent of linkage disequilibrium in the region of interest, as in this case the polymorphisms used to calculate allele frequency ratios are only

those caused by EMS-induced mutagenesis. A small number of polymorphisms between wild-type and mutant parents (lowly mutagenized populations or spontaneous mutants) may represent a difficulty for a precise gene mapping. Another limitation of using BC_1F_2 populations for mapping-by-sequencing is that similar allelic frequency ratios are expected for tightly linked EMS mutation, which hinders the direct identification of the causal mutation, thereby making determination of the causal polymorphism a daunting task. In this chapter, an alternative protocol for mapping-by-sequencing is described, in which the wild tomato species *S. pimpinellifolium*, genetically distant from cultivated tomato, is outcrossed with the mutant of interest to generate a F_2 interspecific mapping population. The use of another species provides a huge amount of polymorphisms for linkage analysis, simplifying the bioinformatic workflow (Fig. 1), ensuring a high mapping accuracy and thus facilitating the identification of the causal mutations.

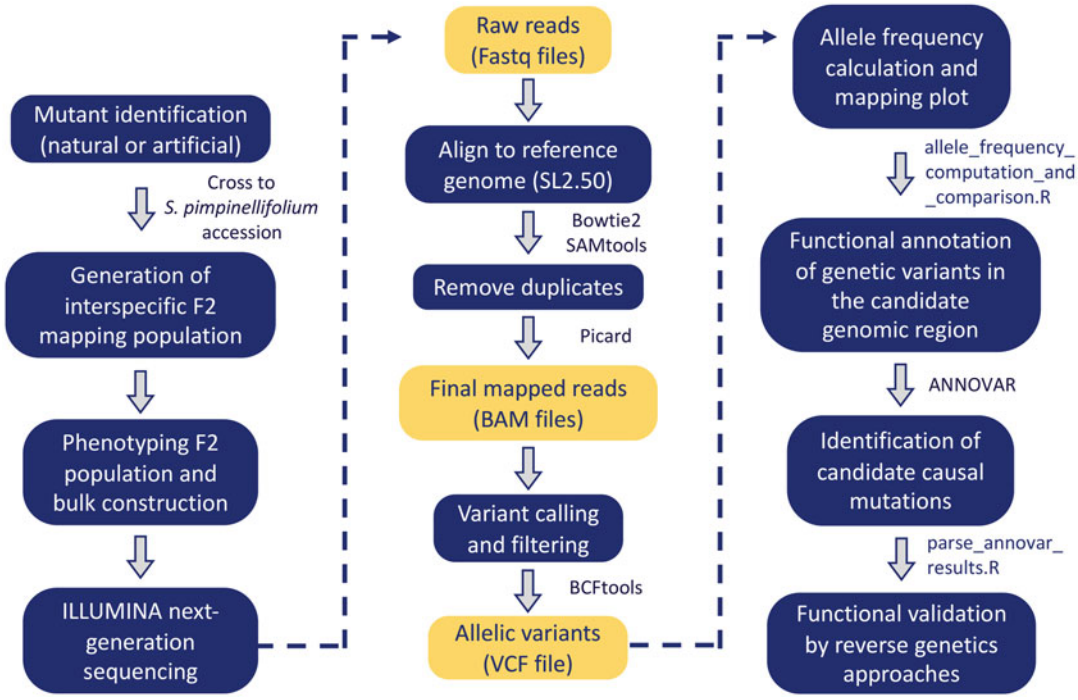


Fig. 1 Flowchart of causal mutation identification by mapping-by-sequencing approach

2 Materials

2.1 Plant Material

An interspecific F_2 mapping population generated by crossing the mutant of interest to wild tomato *Solanum pimpinellifolium* accession LA1589 and self-fertilizing the F_1 plants. LA1589 can be substituted by any other *S. pimpinellifolium* accession for which re-sequencing data is available.

2.2 DNA Extraction and Illumina Sequencing

Materials and reagents for DNA isolation will differ depending on the lab. The CTAB DNA extraction method [24] is recommended, but any other method can be used to obtain pure DNA. After validation of DNA quality, wild-type and mutant pools will be formed using an equal amount of DNA from wild-type and mutant F_2 plants, respectively. The genomic DNA is from the tomato parental line where the mutant of interest was recognized and the pooled samples will be paired-end sequenced using Illumina NGS platform. A minimum coverage of 10× is recommended for each sample. Libraries will be generated according to Illumina standard library preparation protocols. Illumina sequencing data will be delivered as compressed FASTQ files.

2.3 System Requirements

System should have a Unix-based operating environment. Computer requirements vary depending on the data to be analyzed but a minimum of 16 GB RAM and at least 1 TB of free disk space are recommended.

2.4 Software Requirements

All software detailed below are commonly used for the analysis of high-throughput sequencing data, are free for research use, and have detailed manuals for installation and usage. The most common problems arising from the utilization of these programs can be solved by searching in bioinformatic forums such as Biostar (https://www.biostars.org/) or SEQanswers (http://seqanswers.com/).

- Bowtie 2, a software package for aligning sequencing reads to long reference sequences (http://bowtie-bio.sourceforge.net/bowtie2/index.shtml) [25].
- SAMtools, a suite of programs for interacting with high-throughput sequencing data (http://www.htslib.org/) [26].
- Picard, a set of command-line tools for manipulating high-throughput sequencing data and formats such as Sequence Alignment Map (SAM), Binary Alignment Map (BAM), and variant call format (VCF) (https://broadinstitute.github.io/picard/).
- GATK, a set of bioinformatic tools for analyzing high-throughput sequencing and VCF data (https://gatk.broadinstitute.org/hc/en-us/) [27].

- R, a free software environment for statistical computing and graphics (http://www.r-project.org/) [28].
- ANNOVAR, a tool to functionally annotate genetic variants detected from diverse genomes (https://doc-openbio.readthedocs.io/projects/annovar/) [29].

3 Methods

3.1 Generation of the Segregating F_2 Mapping Population

First, cross the wild tomato *S. pimpinellifolium* LA1589 accession to the mutant plant. It is advised to use the tomato mutant plant as female parent as anther emasculation will be easier and fruits will have more seeds. In the case that mutation of interest causes alterations in carpels or female gametophyte development, the wild tomato *S. pimpinellifolium* will be used as female parent. To perform the crossing, remove immature anthers from flowers at pre-anthesis stage in the female parent, without damaging the pistils, to prevent self-pollination. Dissect the stamens of 4–6 flowers to get the pollen from the male parent. Cross-pollinate the female parent depositing the pollen from the male parent over the stigma of the anther-removed flower. F_1 seeds will be collected from red ripe fruits.

Second, grow F_1 plants and allow selfing to produce F_2 seeds. Evaluate 12–16 F_1 plants to determine the genetic inheritance of the mutant phenotype in this new interspecific genetic background (*see* **Note 1**). Plants of the LA1589 accession and the mutant line will be grown, together with F_1 plants, as controls.

3.2 Phenotypic Evaluation of the F_2 Population and Bulk Construction

Grow at least 300 plants of the F_2 offspring together with the parental lines, which will be used as controls, and phenotype them for the trait of interest. For a recessive mutation, a 3:1 wild-type: mutant phenotype ratio is expected, and the mutant phenotype will be present in 25% of the F_2 population (around 75 plants). Collect young leaf tissue from 100 ($\pm$10) plants clearly showing the wild-type phenotype, as well as 50 ($\pm$10) plants showing a clear mutant phenotype. Please note that the 100 wild-type plants will include individuals that are heterozygous and homozygous for the wild-type allele (expected ratio of 2:1 heterozygous:homozygous). If needed, store the samples at -80 °C until use. Extract the genomic DNA of each sample according to the chosen protocol. Once quality and quantity of DNA extractions will be assessed, constitute the wild-type and mutant bulks by pooling equal amount of DNA from individual samples of each phenotype. For dominant and incomplete dominant mutations *see* **Note 2**.

3.3 Library Construction and Next-Generation Sequencing

A step-by-step method to generate sequencing libraries or the resulting sequencing reads is not detailed here, as these protocols are usually performed in the sequencing facilities. In this study's case, sequencing libraries are generated according to Illumina standard library preparation protocols, and sequenced using paired-end technology on an Illumina platform. Each bulk is sequenced to a depth of 20–40× coverage of the tomato genome (minimum of 10×). This sequencing method generates two FASTQ files (R1 and R2) containing paired-end reads for each sample. Commonly, sequencing facilities perform the quality assessment of all short-read data, as well as the preprocessing of raw reads included trimming of adapter sequences and low-quality bases (*see* **Note 3**). After quality validation, sequencing reads are ready for mapping to the reference genome.

3.4 Sequencing Read Alignment to the Tomato Genome Reference

Paired-end Illumina sequencing will produce two FASTQ files (corresponding to the paired-end R1 and R2 files) for each sample (tomato parental line, wild-type bulk, and mutant bulk):

```
tomato_parent_R1.fastq.gz
tomato_parent_R2.fastq.gz
wild_type_bulk_R1.fastq.gz
wild_type_bulk_R2.fastq.gz
mutant_bulk_R1.fastq.gz
mutant_bulk_R2.fastq.gz
```

In order to map reads to the tomato genome, first it is needed to build an index of the genome sequence for fast read alignment.

1. Download the tomato genome reference sequence. In this tutorial, the SL2.50 version is used, but later versions can also be used.

```
wget ftp://ftp.solgenomics.net/tomato_genome/assembly/
build_2.50/S_lycopersicum_chromosomes.2.50.fa
```

2. Create an index for the tomato reference genome with Bowtie 2 [25], create a new directory, change the fasta file into that directory, and run:

```
bowtie2-build S_lycopersicum_chromosomes.2.50.fa tomato_index
```

This will generate a set of six files with suffixes .1.bt2, .2.bt2, .3.bt2, .4.bt2, .rev.1.bt2, and .rev.2.bt2.

3. Map the reads from each sample onto the tomato reference genome. This step generates an output file in SAM format, an international standard (https://samtools.github.io/hts-specs/). It is necessary to specify the name of the sample with --rg and --

rg-id for later identifying genetic variants in each sample. It is advised, if possible, to use as many as eight processors to perform this task, which can be set with the parameter -p 8:

```
bowtie2 -x tomato_index -p8 -1 tomato_parent_R1.fastq.gz -2
tomato_parent_R2.fastq.gz --rg SM:tomato_parent --rg-id toma-
to_parent -S tomato_parent.sam
bowtie2 -x tomato_index -p8 -1 wild_type_bulk_R1.fastq.gz -2
wild_type_bulk_R2.fastq.gz --rg SM:wild_type_bulk --rg-id
wild_type_bulk -S wild_type_bulk.sam
bowtie2 -x tomato_index -p8 -1 mutant_bulk_R1.fastq.gz -2
mutant_bulk_R2.fastq.gz --rg SM:mutant_bulk --rg-id mutant_-
bulk -S mutant_bulk.sam
```

4. Convert SAM file to the BAM format using SAMtools [26]. Only aligned reads mapped with a good quality score (higher than 10) will be used (option -q 10):

```
samtools view -bhS -q 10 tomato_parent.sam > tomato_parent.bam
samtools view -bhS -q 10 wild_type_bulk.sam > wild_type_bulk.bam
samtools view -bhS -q 10 mutant_bulk.sam > mutant_bulk.bam
```

5. Sort and index the BAM file to accelerate downstream analysis:

```
samtools sort tomato_parent.bam -o tomato_parent.sorted.bam
samtools sort wild_type_bulk.bam -o wild_type bulk.sorted.bam
samtools sort mutant_bulk.bam -o mutant_bulk.sorted.bam
```

6. Use Picard (https://broadinstitute.github.io/picard/) to remove duplicated reads and then index the output BAM file. In this step, only the procedure for the mutant bulk sample is shown. Repeat the same process for the tomato parental line and wild-type bulk samples:

```
PicardCommandLine MarkDuplicates INPUT=mutant_bulk.sorted.bam
OUTPUT=mutant_bulk.sorted.Duplrm.bam METRICS_FILE=mutant_-
bulk.sorted.Duplrm.metrics.txt REMOVE_DUPLICATES=true VALIDA
TION_STRINGENCY=SILENT
PicardCommandLine BuildBamIndex INPUT=mutant_bulk.sorted.
Duplrm.bam OUTPUT=mutant_bulk.sorted.Duplrm.bam.bai VALIDA
TION_STRINGENCY=SILENT
```

3.5 Variant Calling and Filtering

The three BAM files obtained above will be used for variant calling analysis in order to identify the polymorphisms (SNPs and indels) between each sample and with respect to the reference genome *S. lycopersicum* v2.50 (Heinz 1706 cultivar). This will produce an output file including three types of polymorphisms: (1) natural polymorphisms existing in the genetic background of the tomato

mutant; (2) natural polymorphisms existing in the *S. pimpinellifolium* LA1589 accession; and (3) polymorphisms induced by the mutagenic agent. Variant calling analysis will be carried out using the HaplotypeCaller tool from GATK [27], and should be performed on the three alignment files (tomato_parent, wild_type_bulk, and mutant_bulk) simultaneously. The HaplotypeCaller algorithm takes an assembly-based novel approach that determines genotype likelihoods independently in each sample and then jointly considers data from all samples in the cohort to increase the confidence of each variant call, filtering low-quality sites. The output file will be in the standard format VCF (https://www.internationalgenome.org/wiki/Analysis/vcf4.0/):

```
gatk --java-options "-Xmx8G" HaplotypeCaller -R S_lycopersicum_chromosomes_2_50.fa -I tomato_parent.sorted.Duplrm.Realn.bam -I wild_type_bulk.sorted.Duplrm.Realn.bam -I mutant_bulk.sorted.Duplrm.Realn.bam -O all_variants.vcf
```

Next, compare genome-wide allele frequencies between the two pooled samples. For this, it is advised to use only biallelic SNPs that are called with high confidence (minimum of ten reads in each sample). A table with these type of variants will be obtained in a convenient format for the calculation of frequencies and graphical representation using BCFtools, which is included in the SAMtools package [26]:

```
bcftools view --samples mutant_bulk,wild_type_bulk --types snps --min-alleles 2 --max-alleles 2 --exclude 'GT[*]=="mis" | FMT/DP[*]<1' -Ou all_variants.vcf | bcftools query --print-header --format '%CHROM\t%POS\t[%AD{0}\t%AD{1}\t]\n' -o filtered_variants.wild_type_and_mutant_bulks.tsv
```

3.6 Genome-Wide Allele Frequency Analysis

The genome-wide distribution of allele frequencies in wild-type and mutant bulks will be determined to associate the mutant phenotype to a tomato genomic region. The necessary calculations and graphics will be performed in the statistics software R (http://www.r-project.org/) [28]. After installing R, open the program and select the folder containing the TSV file created above as "Working Directory." Then, open and run in R the script file "*allele_frequency_computation_and_comparison.R*" (https://github.com/AGR-176/Mapping-by-sequencing-in-tomato/), which computes the allele frequency for each genetic variant in each sample, then calculates the average allele frequency in 1 Mb sliding windows, and graphically represents the results in the output file "*allele_frequency_comparison.pdf*." The pdf should contain a graph that will allow for the identification of the chromosomal region where the causal mutation is located. As shown in Fig. 2, for a recessive monogenic

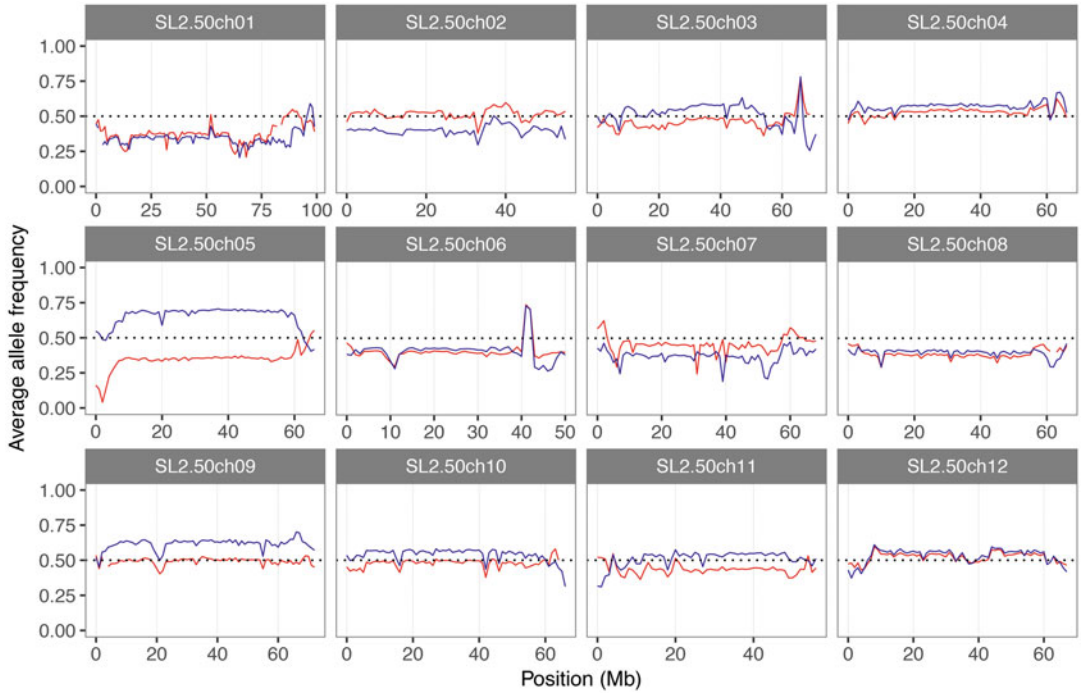


Fig. 2 Mapping the causal mutation responsible for a specific trait that follows a monogenic pattern of inheritance. The mutation of interest is located at the beginning of the chromosome 5 since in this genomic region the average allele frequency ratios of the mutant bulk (red) drop to 0, while those of the wild-type bulk (blue) remain close to 0.5

mutation (*see* **Notes 4** and **5**), it is expected that allele frequencies from the mutant bulk will drop to 0 in the genomic region harboring the mutation responsible for the phenotype of interest, while those from the wild-type bulk will tend to be 0.67 (*see* **Note 6**).

3.7 Identification of the Candidate Causal Mutations

Once the chromosomal region carrying the causal mutation has been identified, extract the genomic region of interest from the VCF file (in the example shown in Fig. 2: chromosome 05, positions from 0 to 3,000,000 bp) using BCFtools with the aim to examine only informative genetic variants and simplify the following analyses. In the same command, variants that appear in all three samples (not interesting for the analysis) are removed:

```
bcftools view -e 'COUNT(GT="AA")=N_SAMPLES' --targets
SL2.50ch05:0-3000000 -o all_variants.ch05_0_3000000.vcf all_-
variants.vcf
```

Then, ANNOVAR [29] is used to predict the functional impact of all variants in the mutant bulk sample. The variant positions will be annotated according to the tomato reference genome annotation ITAG2.4, which can be downloaded as a GFF3 file:

```
wget ftp://ftp.solgenomics.net/tomato_genome/annotation/ITA-
G2.4_release/ITAG2.4_gene_models.gff3
```

The GFF3 file downloaded needs to be converted to be used with ANNOVAR; use the *gff3ToGenePred* tool (http://hgdownload.cse.ucsc.edu/admin/exe/linux.x86_64/gff3ToGenePred/) from UCSC Genomics Institute to convert the GFF3 file to GenePred file:

```
gff3ToGenePred ITAG2.4_gene_models.gff3 Tomato_refGene.txt
```

Generate a transcript FASTA file using the "*Tomato_refGene.txt*" GenePred file with the Perl script "*retrieve_seq_from_fasta.pl*":

```
perl retrieve_seq_from_fasta.pl --format refGene --seqfile
S_lycopersicum_chromosomes.2.50.fa Tomato_refGene.txt --out
Tomato_refGene.fa
```

Then, use the Perl script "*convert2annovar.pl*" to convert the filtered VCF file into ANNOVAR format, which generates the "all_variants.ch05_0_3000000.mutant_bulk.avinput" file:

```
perl convert2annovar.pl -allsample -includeinfo -outfile all_-
variants.ch05_0_3000000 -format vcf4 all_variants.
ch05_0_3000000.vcf
```

Finally, functionally annotate the detected genetic variants in the mutant bulk with the Perl script "*annotate_variation.pl*":

```
perl annotate_variation.pl -geneanno all_variants.
ch05_0_3000000.mutant_bulk.avinput -dbtype generic -buildver
Tomato_refGene -outfile mutant_bulk_functional_annotation
```

ANNOVAR will produce two tab-separated-values files: one finished in ".variant_function" that contains all mutations annotated with their respective positions on the tomato genome, and another one finished in ".exonic_variant_function" that only contains mutations located within exons and their functional consequences.

Then, open the script "*parse_annovar_results.R*" (https://github.com/AGR-176/Mapping-by-sequencing-in-tomato/) which will be used to generate a table with all variants in the region of interest ranked by their likelihood of causality and annotated with the functional description of the gene and the genotype in the mutant bulk, the wild-type bulk, and the tomato parent (*see* **Note 7**). The script produces as output the "*candidate_mutations.tsv*" file that can be opened with spreadsheet applications such as Apache OpenOffice Calc, Google Spreadsheet, or Microsoft Excel. In this

file, loss-of-function mutations that are absent in the tomato parental line, homozygous in the mutant bulk, and heterozygous in the wild-type bulk are ranked in the first place (*see* **Note 8**).

3.8 Cosegregation Analysis and Functional Validation

Once the functional impact of the candidate causal mutations is determined, the main candidates to be responsible for the mutant phenotype will be those whose functional effects cause changes in protein-coding sequence. To support that these selected variants are responsible for the observed mutant phenotype, use Sanger sequencing to validate the candidate variants detected by Illumina's base calling algorithm and perform a cosegregation analysis of the selected candidate variants and the phenotype of interest in the F_2 mapping population. To do this, selected variants will be converted to PCR-based markers and subsequently analyzed in the F_2 progeny to confirm that markers are completely linked to the mutant phenotype.

Finally, involvement of the locus/gene underlying the mutant phenotype needs to be confirmed. Different reverse genetic approaches could be used for the functional validation of causal mutations, among them (1) gene knockdown or gene knockout of the target gene to copy the mutant phenotype in a wild-type background; (2) gene complementation or gene overexpression of the candidate gene in the mutant background in order to restore the wild-type phenotype; and (3) identification of new allelic variants of the candidate gene in mutagenized populations by TILLING methodologies.

4 Notes

1. The mutant phenotype will only be observed in F_1 plants for dominant or incomplete dominant mutations; the proportion of wild-type and mutant plants should be conforming to the 0:1 or 1:1 wild-type:mutant expected ratio, depending on whether the mutant male parent was homozygous or heterozygous for the mutation of interest, respectively. In the case of a recessive mutation, the phenotype of interest will not be observed until the evaluation of the F_2 offspring.
2. For dominant mutations, a 1:3 wild-type:mutant phenotype ratio is expected. Thus, the mutant phenotype is composed of a 2:1 ratio of individuals that are heterozygous and homozygous for the mutant allele, respectively. In this case, store young leaf tissue from all F_2 plants at −80 °C until use and collect F_3 seeds from the F_2 plants displaying a mutant phenotype. Grow and phenotype 12–16 plants from each F_3 family, which will allow for the identification of those F_3 families derived from F_2 homozygous for the mutant allele (i.e., F_3 families where all

individuals show a mutant phenotype). Combine equal amount of genomic DNA from wild-type F_2 plants to generate the wild-type bulk, as well as from F_2 plants homozygous for the mutant allele to construct the mutant bulk. The production of the F_3 population will require at least 6 months for a complete analysis. In the case of incomplete dominant mutations, heterozygote plants will display a third intermediate phenotype resulting from a combination of the phenotypes of both alleles. Thus, a 1:2:1 wild-type:intermediate-mutant:severe-mutant phenotype ratio is expected. Hence, to generate the mutant bulk, select the F_2 individuals showing the most severe mutant phenotype, which represent individuals homozygous for the mutant allele.

3. In the case that sequencing facilities do not implement the quality assessment of raw reads, a plethora of open software tools exists to perform these tasks in all existing computer platforms. For example, FastQC (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) can be used for quality control of sequencing data.
4. In the case of monogenic dominant and incomplete dominant mutations, the expected allele frequency ratios in the genomic region harboring the causal mutation will be 0 in the mutant bulk, while those from the wild-type bulk will be 1 as it is only formed with the DNA of homozygous plants for the wild-type allele.
5. Although this method is mainly used for mapping mutations responsible for traits that follow a monogenic pattern of inheritance, it could also be used to map mutant traits controlled by either digenic or oligogenic inheritance. An example of a mutant phenotype caused by two independent recessive mutations is shown in Fig. 3 and described by Yuste-Lisbona et al. [30]. Application of this approach allowed for the identification of *EXCESSIVE NUMBER OF FLORAL ORGANS* (*ENO*), an AP2/ERF transcription factor which interacts synergistically with *LOCULE NUMBER* (*LC*) to regulate floral meristem activity and fruit size [30]. In these cases, the expected segregation ratio in the F_2 population will be 15:1 (wild-type: mutant phenotype) and the mutant bulk will be conformed with the DNA of plants homozygous for both mutant alleles. Thus, it is expected that the allele frequency ratios in the mutant bulk will drop to 0 in two genomic regions, which harbor the mutations under study.
6. In the case of a recessive monogenic mutation, the expected allele frequency ratios in the genomic region harboring the causal mutation will tend to be 0.67 in the wild-type bulk due to the fact that the ratio between the number of individuals that

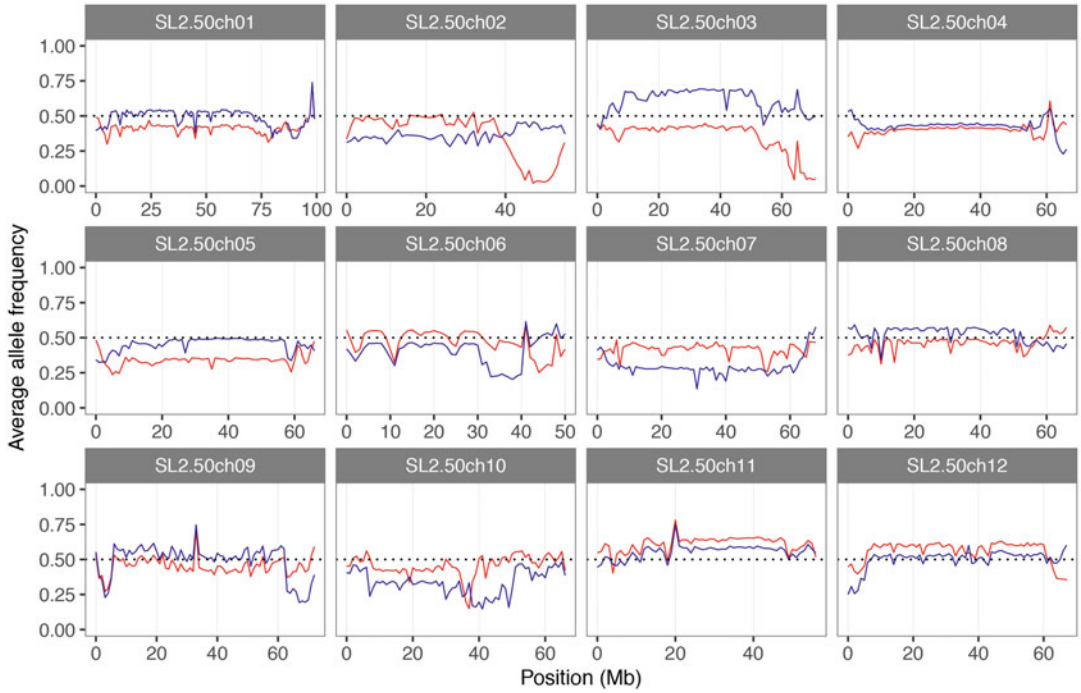


Fig. 3 Mapping the causal mutations responsible for a particular trait with a digenic recessive inheritance pattern. The graphical representation of the average allele frequency ratios shows that these drop to 0 at the end of chromosomes 2 and 3 in the mutant bulk (red), while the average allele frequency ratios remain close to 0.5 in the wild-type bulk (blue), which indicates that the mutations under study are located in these genomic regions (*see* ref. 30)

are heterozygous at the mutated locus and the number of those that are homozygous for the wild-type allele is expected to be 2:1. However, these allele frequency ratios could vary depending on the genomic region where the causal mutation is located since the wild-type bulk is formed with F_2 plants from an interspecific cross, which could give rise to bias toward one parent allele.

7. As the mutation under study has been originated in a cultivated tomato genetic background, the mutant bulk should not have natural polymorphisms from the wild *S. pimpinellifolium* LA1589 accession in the chromosomal region carrying the causal mutation. Thus, the LA1589 variants are only used for mapping purpose, allowing for the calculation of the allele frequency ratios along tomato chromosomes.
8. In the case of using mapping-by-sequence approach to isolate several causal mutations from an induced mutant collection generated in the same tomato genetic background, an alternative strategy could be performed that avoids sequencing the tomato parental line and the wild-type bulks. This procedure is

based on the parallel comparison of the allele frequencies among all mutant bulks. In this case, each mutant bulk is expected to be heterozygous for tomato and *S. pimpinellifolium* alleles except in the genomic region where causal mutations are located, which will be homozygous for cultivated tomato alleles. Hence, the chromosomal location of the causal mutations can be mapped by comparing in parallel the average allele frequency ratios of each mutant bulk. Then, each candidate genomic region will be screened for unique mutations, since for a particular mutant, the variant responsible for a mutant phenotype must be present in homozygous state in its corresponding mutant bulk and absent in the remaining ones.

Acknowledgments

This work was supported by the Spanish Ministry of Economy and Competitiveness (grant AGL2015-64991-C3-1-R) and the BRESOV (breeding for resilient, efficient, and sustainable organic vegetable production) project. BRESOV project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 774244. J.M.J.-G received funding from ANR projects (ANR-17-ERC2-0013-01, ANR-18-CE92-0039-01, and ANR-17-CE20-0024-02). IJPB benefits from the support of Saclay Plant Sciences—SPS (ANR-17-EUR-0007).

References

1. Meissner R, Jacobson Y, Melamed S, Levyatuv S, Shalev G, Ashri A et al (1997) A new model system for tomato genetics. Plant J 12:1465–1472
2. Lozano R, Giménez E, Cara B, Capel J, Angosto T (2009) Genetic analysis of reproductive development in tomato. Int J Dev Biol 53:1635–1648
3. Ranjan A, Ichihashi Y, Sinha NR (2012) The tomato genome: implications for plant breeding, genomics and evolution. Genome Biol 13:167
4. Tomato Genome Consortium (2012) The tomato genome sequence provides insights into fleshy fruit evolution. Nature 485:635–641
5. Causse M, Desplat N, Pascual L, Le Paslier MC, Sauvage C, Bauchet G et al (2013) Whole genome resequencing in tomato reveals variation associated with introgression and breeding events. BMC Genomics 14:791
6. Tomato Genome Sequencing Consortium (2014) Exploring genetic variation in the tomato (*Solanum* section *Lycopersicon*) clade by whole-genome sequencing. Plant J 80:136–148
7. Lin T, Zhu G, Zhang J, Xu X, Yu Q, Zheng Z et al (2014) Genomic analyses provide insights into the history of tomato breeding. Nat Genet 46:1220–1226
8. Tieman D, Zhu G, Resende MF Jr, Lin T, Nguyen C, Bies D et al (2017) A chemical genetic roadmap to improved tomato flavor. Science 355:391–394
9. Zhu G, Wang S, Huang Z, Zhang S, Liao Q, Zhang C et al (2018) Rewiring of the fruit metabolome in tomato breeding. Cell 172:249–261.e12
10. Gao L, Gonda I, Sun H, Ma Q, Bao K, Tieman DM et al (2019) The tomato pan-genome uncovers new genes and a rare allele regulating fruit flavor. Nat Genet 51:1044–1051

11. Razifard H, Ramos A, Della Valle AL, Bodary C, Goetz E, Manser EJ et al (2020) Genomic evidence for complex domestication history of the cultivated tomato in Latin America. Mol Biol Evol 37(4):1118–1132. https://doi.org/10.1093/molbev/msz297
12. Meissner R, Chague V, Zhu Q, Emmanuel E, Elkind Y, Levy AA (2000) A high throughput system for transposon tagging and promoter trapping in tomato. Plant J 38:861–872
13. Menda N, Semel Y, Peled D, Eshed Y, Zamir D (2004) In silico screening of a saturated mutation library of tomato. Plant J 38:861–872
14. Carvalho RF, Campos ML, Pino LE, Crestana SL, Zsogon A, Lima JE (2011) Convergence of developmental mutants into a single tomato model system: 'Micro-Tom' as an effective toolkit for plant development research. Plant Methods 7:18
15. Just D, Garcia V, Fernandez L, Bres C, Mauxion J, Petit J et al (2013) Micro-Tom mutants for functional analysis of target genes and discovery of new alleles in tomato. Plant Biotechnol 30:225–231
16. Pérez-Martín F, Yuste-Lisbona FJ, Pineda B, Angarita-Díaz MP, García-Sogo B, Antón T et al (2017) A collection of enhancer trap insertional mutants for functional genomics in tomato. Plant Biotechnol J 15:1439–1452
17. Birkeland SR, Jin N, Ozdemir AC, Lyons RH Jr, Weisman LS, Wilson TE (2010) Discovery of mutations in *Saccharomyces cerevisiae* by pooled linkage analysis and whole-genome sequencing. Genetics 186:1127–1137
18. Sarin S, Prabhu S, O'Meara MM, Pe'er I, Hobert O (2008) *Caenorhabditis elegans* mutant allele identification by whole-genome sequencing. Nat Methods 5:865–867
19. Blumenstiel JP, Noll AC, Griffiths JA, Perera AG, Walton KN, Gilliland WD et al (2009) Identification of EMS-induced mutations in *Drosophila melanogaster* by whole-genome sequencing. Genetics 182:25–32
20. Cuperus JT, Montgomery TA, Fahlgren N, Burke RT, Townsend T, Sullivan CM et al (2010) Identification of MIR390a precursor processing-defective mutants in Arabidopsis by direct genome sequencing. Proc Natl Acad Sci U S A 107:466–471
21. Schneeberger K (2014) Using next-generation sequencing to isolate mutant genes from forward genetic screens. Nat Rev Genet 15:662–676
22. Michelmore RW, Paran I, Kesseli RV (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
23. Garcia V, Bres C, Just D, Fernandez L, Tai FW, Mauxion JP et al (2014) Rapid identification of causal mutations in tomato EMS populations via mapping-by-sequencing. Nat Protoc 11:2401–2418
24. Dellaporta SL, Wood J, Hicks JB (1983) A plant DNA minipreparation: Version II. Plant Mol Biol Rep 1:19–21
25. Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. Nat Methods 9:357–359
26. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N et al (2009) 1000 genome project data processing subgroup, the sequence alignment/map (SAM) format and SAMtools. Bioinformatics 25:2078–2079
27. DePristo MA, Banks E, Poplin R, Garimella KV, Maguire JR, Hartl C et al (2011) A framework for variation discovery and genotyping using next-generation DNA sequencing data. Nat Genet 43:491–498
28. R Development Core Team (2011) R: a language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria
29. Wang K, Li M, Hakonarson H (2010) ANNOVAR: functional annotation of genetic variants from next-generation sequencing data. Nucleic Acids Res 38:e164
30. Yuste-Lisbona FJ, Fernández-Lozano A, Pineda B, Bretones S, Ortíz-Atienza A et al (2020) *ENO* regulates tomato fruit size through the floral meristem development network. Proc Natl Acad Sci U S A 117:8187–8195

Chapter 8

Association Mapping in Plants

Pawan L. Kulwal and Ravinder Singh

Abstract

Quantitative trait loci mapping has become a common practice in crop plants and can be accomplished using either biparental populations following interval mapping or natural populations following the approach of association mapping. Because of its ability to use the natural diversity and to search for functional variants in a broader germplasm, association mapping is becoming popular among researchers. An overview of the different steps involved in association mapping in plants is provided in this chapter.

Key words Association mapping, GWAS, Linkage disequilibrium, Population structure, Multi-parental populations, False discovery rate

1 Introduction

The advances in the area of molecular marker development and computational analysis have facilitated quantitative trait locus (QTL) mapping in crop plants. Consequently, QTL mapping has become a common practice among plant breeders and large numbers of marker-trait associations (MTAs) have been identified in different crops for a variety of traits [1]. Based on the type of population used for analysis (biparental/multi-parental/natural population) and the objective of the program, QTL mapping can be performed using the principle of either interval mapping (IM) or linkage disequilibrium (LD)-based association mapping (AM). Since the QTLs identified in a biparental population using the IM approach are relevant only in those breeding programs involving parents differing for specific QTLs, they are of little use in the other breeding programs [2]. Association mapping, also known as LD mapping or genome-wide association study (GWAS), is a method of finding MTAs based on historic recombinations in natural populations possessing wider trait variation [3] to identify candidate genes and regions for a trait. The basic principle involved in AM is co-inheritance of alleles of SNPs in adjoining regions in a

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_8,

population due to LD [4] and a strong correlation between allele variants and trait in a natural population.

AM is becoming popular over the commonly used biparental QTL mapping because of its ability to exploit the natural diversity and to search for functional variants in broader germplasm offering advantages over the latter resulting in increased QTL resolution [1, 2, 5–7]. The resolution achieved by AM in the identification of genes is much higher due to a number of factors including its ability to densely genotype nucleotide-level variations and low LD among natural populations. This technique does outperform more widely used QTL mapping based on biparental mapping populations by identifying strong correlations between alleles and trait variations encompassed in natural populations. Moreover, the associations can be identified in a much shorter period of time as compared to the time taken for QTL identification in biparental mapping populations. Thus, LD-based AM provides an alternative approach for the identification of MTAs using a set of genotypes of known/unknown ancestry carrying maximum genetic variability for the trait of interest [2].

Based on the objective of a specific investigation, AM can be classified as either a genome-wide association study (GWAS) or candidate gene (CG)-based analysis. When the objective of the study is to find all the genomic regions involved in controlling the trait variation, GWAS is utilized. The results can give an overview of the genetic basis of the trait and the associations which are most promising can be identified and shortlisted for further study using this approach. However, if information about the CG for the target trait is available, one would like to confirm the genes that control the trait of interest following CG-based AM. In this chapter, an attempt has been made to describe the basic steps involved in association analysis in plants along with different issues that need to be considered during analysis of the data. While doing so, the complex statistical issues are not discussed. However, for more details, readers can refer to the published literature [1, 2, 5–8].

2 Materials

For undertaking any AM study in plants, three essential requirements are (1) association mapping population evaluated for the trait of interest, (2) marker genotypic data on the AM population, and (3) computer software to perform statistical (including population structure, marker imputation) and genetic analysis for identification of MTAs. All these issues are described individually in the following sections.

2.1 Different Populations Used for Association Mapping in Plants

The successful outcome of any AM program depends on the population which is used for analysis. Although AM study in plants can be carried out using a variety of populations, one can choose either natural population, breeding material, germplasm, or multi-parental populations for this purpose. Based on the composition of the population used, AM studies have also been classified either as broad-based or narrow-based [9]. Broad-based studies use germplasm, landraces, cultivars, and natural populations while narrow-based studies use multi-parental populations specifically developed for this purpose. Ideally, a population comprising diverse genotypes (which can capture the diversity for the trait of interest) is best suited for such studies. The majority of the AM studies carried out in plants have utilized a set of diverse genotypes in their analysis. However, care needs to be taken to account for the effect of population structure in analysis to rule out the possibility of identifying false-positive QTL (*see* Subheading 3.2 on population structure later).

The advantages of using breeding populations rely on the phenotypic data routinely generated in breeding programs that can be used for analysis with a minor additional effort for genotyping. Moreover, the results obtained in terms of phenotypic data and materials developed are of direct relevance to the breeders. However, in recent years, multi-parental populations like multiparent advanced generation intercross (MAGIC) and nested association mapping (NAM) populations are also being used in many crops for AM. Depending on the nature of the trait being studied, one can decide the composition of AM population. For example, population comprising exotic material, core collection, or gene bank collection are more suited for traits which are less influenced by adaptation [10]. On the contrary, elite material is more suited for traits that are difficult to phenotype (*see* **Notes 1** and **2**).

The data scored for various traits in a population could be controlled by a few or large numbers of genes. A trait controlled by a large number of genes is likely to have higher variance as compared to traits controlled by a lesser number of genes. The association analysis tries to identify strong relationships between the covariance of the genetic polymorphisms and trait variations [7]. It is important to note that the accuracy of association analysis is directly proportional to the accuracy with which the trait data has been scored (*see* **Note 3**).

2.2 Molecular Markers for Association Mapping in Plants

The variation in traits is governed partly by the polymorphism at the contributing loci and its interaction with the environment. A number of marker systems have been used to genotype genetic polymorphisms. An ideal marker system should be able to quantify multi-allelic polymorphisms at individual loci, yet the marker system should be able to scan whole genome or candidate gene regions with enough markers to dissect traits at a finer scale. As

mentioned above, the number of markers and LD would determine the accuracy of the association study. So, the marker system for association analysis should have the following characteristics: (1) the markers should be easy to develop and genotype, (2) a marker system amenable to multiplexing or high-throughput assays is more preferred, (3) multi-allelic marker system is needed to associate the trait variance with different alleles, and (4) the markers need to be present at high frequency and uniformly dispersed in genome. Not all marker system fulfills above requirements. While simple sequence repeats (SSRs) were one of the first marker systems to be utilized for GWAS in crop plants, single-nucleotide polymorphism (SNPs) and other nucleotide-level sequence variations have proved their utility for gene identification through GWAS analysis for various traits (*see* **Note 4**).

Single-nucleotide polymorphisms as genetic markers for AM have been used in a number of crops for mapping of genes for a variety of traits including agronomic, quality, biotic, and abiotic stress resistance. The major advantage of using SNPs is that these are present in abundance uniformly across the genome, even in species with a narrow genetic base.

In recent years, a shift toward the use of high-throughput multiplexed SNP genotyping technologies has also been observed. Genotyping using genome-wide SNP markers would allow the detection of a complete landscape of the genes/QTL for a particular trait. Most technologies do not allow the processing of a large number of samples for genome-wide SNPs. For this, arrays (both standard panels and customized) are available from Affymetrix (www.thermofiisher.com) and Illumina Inc. (www.illumina.com). These panels can genotype thousands of samples for millions of SNPs faster than any other technology. Illumina's Infinium iSelect microarray has the capability to target from 3072 to 700K custom SNPs. The Affymetrix has developed the Axiom genotyping solution able to accommodate up to 650K altogether on a single microarray. The genotyping-by-sequencing (GBS) discussed below gives another dimension to SNP development and genotyping.

With advances in the sequencing techniques, the cost of marker development has reduced drastically and the total usable sequence data has also increased manifold. The GBS in totality represents various methods that are used to (re)sequence genome partially or in whole for the development of genome-wide sequence tags followed by identification of SNPs [11, 12]. These genome-wide sequence tags developed across a group of diverse individuals can lead to simultaneous development and genotyping of nucleotide variations in the form of SNPs as genetic markers. Earlier SNP development was restricted to major crops including cereals, pulses, and oilseeds. However, recent advances in sequencing technologies have changed this dynamics and high-throughput sequencing

technologies can be applied to even lesser known crops for the development of genome-wide SNP markers for association analysis. Subjecting large collections of natural populations to GBS has also allowed capturing rare and minor alleles for association analysis.

2.3 Software for Performing Association Analysis

With the advances in high-throughput phenotyping and genotyping techniques, the volume of data generated in any AM experiment is huge, and the data is processed for a number of parameters (including marker imputation, PCA, kinship, population structure) before being put to use for AM. Different software packages are required to carry out all these activities and perform the analysis with sufficient power to detect significant MTAs. In recent years several software packages have been developed to perform association analysis in an efficient manner (Table 1). Depending on the volume of data and the expertise of the person handling it, one can use any software for analysis. Ideally software having the ability to perform multiple tasks and which is user friendly is preferred. Although several software packages are available for this purpose, TASSEL (Trait cc by Association, Evolution, and Linkage) is the software that was designed while working with plant systems [13] and has been widely used by the plant breeding community. For a beginner, this software can be very useful for the analysis (*see* **Notes 5** and **6**). Another suitable software for plants is GAPIT, which is R based [14], and performs association analysis and genomic prediction. Besides these, "R" is also commonly used for this purpose and scripts for performing different types of analysis are freely available (*see* **Note 7**).

3 Methods

With the availability of trait data and the marker genotypic data generated on the population, one can start analysis of data for identifying MTAs using appropriate software. However, before the marker data is put to use for association analysis, it is necessary to perform imputation of the missing marker data and study the population structure/stratification in the population used for AM and remove markers with minor allele frequency less than 5% (*see* **Note 8**). These issues and the steps involved in association analysis are described below.

3.1 Marker Imputation

Although the advances in high-throughput marker development techniques (next-generation sequencing and GBS) generate huge amount of data in a short span of time, it often creates the problem of missing marker data which leads to identification of false/spurious MTAs in AM studies. In order to increase the utility of the missing information, marker imputation is carried out. Marker imputation basically means replacing the missing values for a

Table 1
List of commonly used software packages for association mapping in plants. Reproduced with permission from [2]

Program	Features	Web address (verified 16 October 2020)
TASSEL	LD statistics, GLM, MLM, CMLM, P3D, genomic selection; graphical interphase, PCA, and kinship, free	http://www.maizegenetics.net/tassel/
GAPIT	R based, CMLM, fast computation, free	http://www.maizegenetics.net/gapit
R	Generic, commonly used for programming, free	http://www.r-project.org/
PLINK	Handles virtually unlimited numbers of SNPs; MDS to visualize substructure, free	http://pngu.mgh.harvard.edu/~purcell/plink/
EMMA	Mixed model, corrects for the confounding from population structure and genetic relatedness, free	http://mouse.cs.ucla.edu/emma/
EMMAX	Large-scale association mapping, corrects for the confounding from population structure and genetic relatedness, increased computational speed, free	http://genetics.cs.ucla.edu/emmax/
EIGENSOFT	Uses principal component analysis to explicitly model ancestry differences between cases and controls, free	http://genepath.med.harvard.edu/~reich/Software
GenAMap	Performs automatic structured association mapping (SAM) using different algorithms; good graphical presentation, free	https://github.com/blengerich/GenAMap
Matapax	GWAS is performed in R environment with EMMA and GAPIT libraries; performs all essential steps for basic GWAS, population structure, fast computation, free	http://matapax.mpimp-golm.mpg.de
Merlin	Includes an integrated genotype inference feature for improved analysis when some genotypes are missing, does not control for population stratification of its own, free	http://www.sph.umich.edu/csg/abecasis/merlin/tour/assoc.html
ASReml	Handles large data set, calculates population structure and pedigree-based kinship, commercial	http://www.vsni.co.uk/software/asreml
SAS	Generic program commonly used in data analysis, commercial	http://www.sas.com
JMP Genomics	Calculates population structure and marker-based kinship, commercial	http://www.jmp.com/software/genomics/
SVS	Comprehensive package with better visualization of the results; offers different options, commercial	http://www.goldenhelix.com/SNP_Variation/
GenStat	Performs GLM and MLM, takes care of population structure, commercial	http://www.vsni.co.uk/software/genstat
FaST-LMM	For analysis of large data sets (up to 120,000 individuals), free	http://fastlmm.codeplex.com/

particular genotype with those of the predicated marker alleles [2]. Although different programs or approaches are available for this purpose, the most convenient way is to use information from the neighboring markers to impute the values. This is often accomplished based on the observed genotypes at neighboring markers. Marker imputation can also be performed using the data from the reference genome. Imputation of missing marker data reduces cost and time in genotyping again for the missing values and increases the utility of the marker data. Marker imputation is very important when one wants to perform meta-analysis. Since the value of missing marker data set is increased upon imputation, it is recommended that one should impute the missing marker data before it is used for association analysis (*see* **Note 9**).

3.2 Study of Population Structure and Family Relatedness

Population structure is the presence of subpopulations in the population used for AM. If not taken into account, population structure can influence the results of analysis by identifying spurious or false-positive associations. For finding population structure, one can use a variety of software solutions including the widely used S TRUCTURE [15]. Either all the marker data or a set of unlinked markers from the available set of markers can be used to find out the number of subpopulations present using this software. The analysis can be done with or without the admixture model and considering the variable number of subpopulations (K) (*see* **Note 10**). Alternatively, it is possible to perform principal component analysis (PCA) using the available marker data and use the first few components (explaining a major portion of the total variation; or equivalent to the number of subpopulations as obtained using STRUCTURE analysis) as covariates to control for population structure. In TASSEL [13], one can perform PCA and the PCs can be used as covariates (Q) to account for population structure (*see* **Note 11**). In addition, this software can also work out kinship using the marker data which is very effective in handling the issue of family (genetic) relatedness (relationships among individuals) and population structure.

3.3 Performing Association Analysis

When all the data sets are available, one can start the analysis of data using appropriate software. Depending on whether to use the independent variables as a fixed effect variable or random variables, two commonly used approaches for association analysis are the general linear model (GLM) and the mixed linear model (MLM). In GLM, only the fixed effect model is used to identify the association between genetic markers and phenotype. Additionally, population structure (Q) can be used as a covariate. On the other hand, MLM approach uses genotypic data and population stratification as a fixed effect, and variance-covariance matrix calculated as kinship coefficients among individuals as a random effect. These models are used to calculate significant associations between genetic markers

and trait of interest on the basis of *p*-value, and variance components for the trait are calculated using ANOVA to work out allelic effects. Although, as compared to GLM, the time required for analysis using the MLM approach is more, particularly when the data set is large, it has higher statistical power than GLM and can detect more true associations.

While analyzing data using GLM approach, one can select files containing phenotypic data and genotypic data and can perform analysis either without taking into account the population structure (naive model) or including the population structure (*Q*) (*see* Fig. 1). In case of MLM, one has to derive a kinship matrix (*K*) from the marker data and include it into analysis as random effects along with the phenotypic and genotypic data. Additionally, population structure (*Q*) can also be included in the analysis. When a kinship matrix (*K*) is used along with population structure (*Q*), the "*Q* + *K*" approach improves statistical power compared to "*Q*" only [16]. All these analyses will provide *p*-values for the markers based on which the significant associations can be identified.

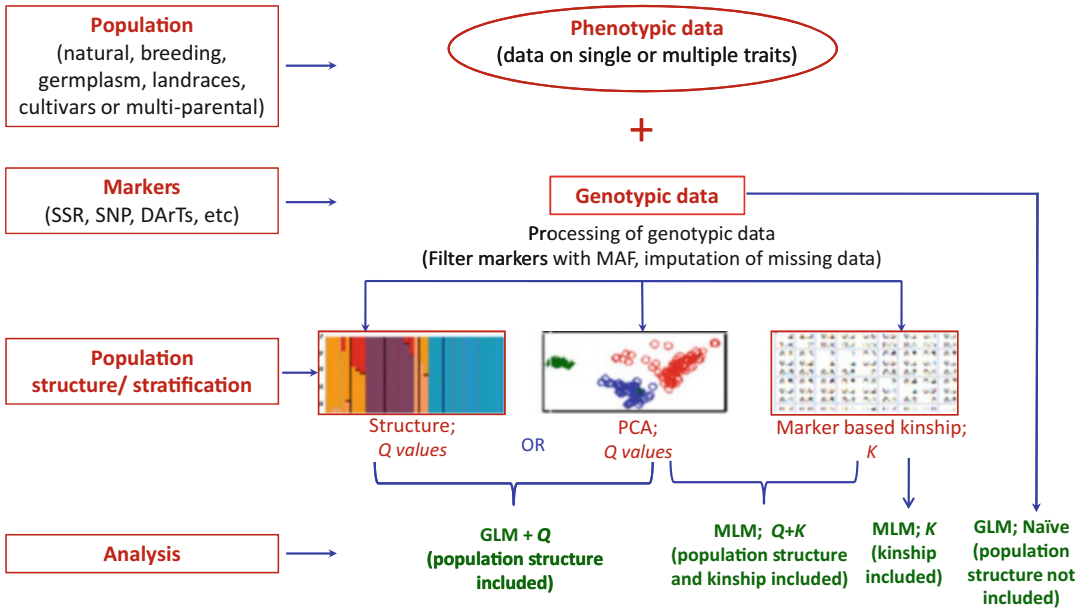


Fig. 1 Schematic representation of steps involved in association mapping. On the left are shown the resources required and steps involved in association mapping including different populations used, different markers to be used for genotyping, study of population structure, and different ways to performing analysis. On the right it is shown that association analysis can be conducted following general linear model (GLM) by including population structure (*Q* values obtained using STRUCTURE software or using principal components) (GLM + *Q*) or including marker-based kinship (*K*) following mixed linear model (MLM) approach or including *Q* along with *K* (*Q* + *K*) in MLM or without including the effect of population structure in GLM ("naive" approach; only phenotypic and genotypic data are used without taking care of population structure)

3.4 Calculating False Discovery Rate and False-Positive QTLs

The AM studies involve a large number of markers to be tested for association with the population of varying sizes. Each marker is independently tested for association leading to a total number of associations being tested equal to the total number of markers. With such large marker data sets, the chances of identifying false-positive/spurious associations are very high. Therefore, in order to determine the statistical significance threshold, different statistical procedures accounting for multiple testing have been proposed [2]. Two major adjustments/corrections for *p*-value used in GWAS analysis are false discovery rate (FDR; [17]) and Bonferroni correction [18]. The FDR represents a value below which all *p*-values are taken as a significant association. For calculation of FDR at $\alpha = 0.05$, the *p*-values are arranged in an increasing order followed by identification of *p*-value equal or less than $r/m \times 0.05$, where "r" represents the rank of the *p*-value and "m" represents the total number of genetic markers. For Bonferroni correction, the significant *p*-value is calculated using α/n, where α is the significance threshold (usually 0.05) and "n" represents the number of tests being tested independently (equal to number of markers) (*see* **Note 12**). However, the basic assumption that all tests be independent is sometimes violated in Bonferroni correction due to the presence of LD when a large number of genetic markers are used, leading to an overcorrection for *p*-value. Besides these two tests, permutation test (generally 1000 times) can also give corrected *p*-values and can minimize the chances of identifying false-positive associations. Software TASSEL has a provision of permutation analysis while using GLM approach.

3.5 Preparing QQ Plots and Manhattan Plots

A marker is said to be linked with the trait of interest if its *p*-value is less than the significance threshold assigned. The most significant associations can be identified by arranging the marker *p*-values in ascending order. At the same time the results of the analysis can also be shown graphically using quantile-quantile (QQ) plots and the Manhattan plots.

QQ plots: The number of genetic markers (mostly SNPs) used for GWAS tends to run into multiples of 100K, and each genetic marker tested for MTA generates a *p*-value. The *p*-values obtained form a basis to understand association analysis results, and visualizing all the *p*-values in a single representation tends to help interpret and identify true associations. The QQ plots represent observed and expected (theoretical) *p*-values of associations between genetic marker and trait. The observed and expected *p*-values should fall on or around a straight line with minimum deviations, under the null hypothesis that "none of the genetic markers are associated with trait" (*see* **Note 13**). The observed *p*-values deviating away from the straight line (at the top) tend to indicate that corresponding SNPs are significantly associated with the trait.

Manhattan plots: These plots are used to visualize GWAS results by plotting negative logarithm (log) of *p*-values on *Y*-axis and chromosome-wise SNP genomic coordinates on *X*-axis. Each point on the plot represents corresponding negative log value and genomic coordinates on *X*- and *Y*-axes. FDR/Bonferroni-corrected (threshold) *p*-values are represented by a horizontal line on Manhattan plots to identify SNPs significantly associated with trait. The SNPs above threshold *p*-value are suggested to be significantly associated with traits, as compared to other SNPs. The negative log *p*-values are color-coded on the basis of their corresponding SNPs on chromosomes.

Both these types of plots can be plotted using different software. While TASSEL has a provision to draw these plots after analysis, several "R" scripts are also freely available for this purpose (*see* **Note 14**).

4 Notes

1. While preparing the AM panel (AM population), make sure that heterozygote individuals (segregating lines) are not included along with the inbred lines, as segregating lines may give inconsistent phenotype across the years and location, apart from creating trouble for proper scoring of the marker alleles.
2. If the trait of interest is influenced by major traits like days to flowering and maturity, then care must be taken to remove extreme genotypes from the panel, so that proper scoring of trait data is possible. This will enable the identification of robust MTAs. Do not include any wild species/relative in the AM panel just for the sake of creating diversity.
3. The phenotyping of the trait must be done with high accuracy. Any wrong scoring of trait data can influence the results and increase the chances of false-positive associations. Such associations may be statistically significant, but may not have any biological relevance. Upon phenotyping of the data rather than creating separate files for different traits, the phenotypic data of multiple traits can be saved in one single file which can be imported for analysis. This will enable analysis of all the traits at a time.
4. The number of markers used for association analysis depends on the extent of LD, which is a measure of correlation (*r*-square) between two SNPs across a set of individuals. A number of factors can influence LD estimates including the relatedness of individuals in a population, type of genetic markers, and regions of genome being genotyped. Less number of

markers are required for regions with high LD and vice versa. The use of a large number of genetic markers (covering the entire genome) increases the probability to capture all possible causal variants for polygenic traits.

5. It is always better to read the user manual of any software completely before the data is used for analysis. The important thing is to prepare the data files according to the requirement of the software being used. Most of the time, people are stuck at this point only. Software like TASSEL have relatively simple format. It is always better to start with the tutorial data provided with the software.
6. Make sure that there are no missing data in the files. If so, it must be substituted with the appropriate symbol so as to distinguish it from the remaining data points. Although there are different options available in any software, for the beginners, it is always convenient to use the default settings of the software.
7. If the analysis is being done using R-based packages, different scripts for performing AM are also available free online. One can search them online using appropriate key words and use these scripts as such.
8. Before using the marker data for analysis, remove the markers with minor allele frequencies (MAF) less than 5% from the analysis. Most of the software can do this or one can identify such markers using MS Excel also.
9. The genotypes for which the marker genotypic data is missing, it must first be imputed using appropriate software. Use the imputed values for analysis. It will increase the value of the marker data and avoid re-genotyping for the missing values. One can also perform analysis with and without imputed marker data and see the difference in the results obtained.
10. The best way of understanding the effect of population structure and family relatedness is to perform the analysis without accounting for it (naive) followed by including them in the analysis. The comparison of the results obtained using both these analyses can tell the effect of population structure on the overall analysis.
11. If you are using software TASSEL, make sure that the missing marker data is imputed before performing PCA.
12. If the total numbers of markers used in the analysis are 1100, and the strongest associated SNP/marker is having a *p*-value of 1.5025×10^{-4}, then the threshold value will be 9.0909×10^{-6} at a significant level of 1% after Bonferroni correction (0.01/1100), making it nonsignificant.

13. In QQ plots, the significant deviation of the observed values from that of the expected values suggests that there are too many false-positive associations (if the graph of observed values is above that of expected values). This generally happens when population structure is not taken into account. However, if the graph of the observed values is below that of the expected values, it indicates that too many parameters have been included in the analysis for correction of population structure and family relatedness. Generally, this happens when $Q + K$ is used. However, it also depends on the trait being analyzed.
14. Several data sets (containing phenotypic and genotypic data) as well as "R" scripts are available with the published literature. One can use these data sets and perform association analysis using the parameters given in the research articles to gain the hands-on training.

Acknowledgment

During the course of writing this chapter, P.L.K. received financial assistance from the Department of Agriculture, Cooperation and Farmers Welfare, Ministry of Agriculture and Farmers Welfare, Government of India, for a research project.

References

1. Gupta PK, Kulwal PL, Jaiswal V (2019) Association mapping in plants in the post-GWAS genomics era. Adv Genet 104:75–154. https://doi.org/10.1016/bs.adgen.2018.12.001
2. Gupta PK, Kulwal PL, Jaiswal V (2014) Association mapping in crop plants: opportunities and challenges. Adv Genet 85:109–147
3. Bergelson J, Roux F (2010) Towards identifying genes underlying ecologically relevant traits in Arabidopsis thaliana. Nat Rev Genet 11 (12):867–879
4. Bush WS, Moore JH (2012) Genome-wide association studies. PLoS Computational Biol 8:12
5. Zhu C, Gore M, Buckler ES, Yu J (2008) Status and prospects of association mapping in plants. Plant Genome 1:5–20
6. Gupta PK, Kulwal PL, Mir RR (2013) QTL mapping: methodology and applications in cereal breeding. In: Gupta PK, Varshney RK (eds) Cereal genomics II. Springer, Netherlands, pp 275–318
7. Gupta PK, Rustgi S, Kulwal PL (2005) Linkage disequilibrium and association studies in higher plants: present status and future prospects. Plant Mol Biol 57:461–485
8. Kulwal PL (2018) Trait mapping approaches through linkage mapping in plants. Adv Biochem Eng Biotechnol 164:53–82. https://doi.org/10.1007/10_2017_49
9. Islam MS, Thyssen GN, Jenkins JN, Zeng L, Delhom CD, McCarty JC, Deng DD, Hinchliffe DJ, Jones DC, Fang DD (2016) A MAGIC population-based genome-wide association study reveals functional association of *GhRBB1_A07* gene with superior fiber quality in cotton. BMC Genomics 17(1):903
10. Kulwal P, Ishikawa G, Benscher D, Feng Z, Yu LX, Jadhav A, Mehetre S, Sorrells ME (2012) Association mapping for pre-harvest sprouting resistance in white winter wheat. Theor Appl Genet 125(4):793–805
11. Hyten DL, Cannon SB, Song Q, Weeks N, Fickus EW, Shoemaker RC et al (2010) High-throughput SNP discovery through deep resequencing of a reduced representation library to anchor and orient scaffolds in the soybean whole genome sequence. BMC Genomics 11:38. https://doi.org/10.1186/1471-2164-11-38

12. Elshire RJ, Glaubitz JC, Sun Q, Poland JA, Kawamoto K, Buckler ES, Mitchell SE (2011) A robust, simple genotyping-by-sequencing (GBS) approach for high diversity species. PLoS One 6:e19379. https://doi.org/10.1371/journal.pone.0019379
13. Bradbury PJ, Zhang Z, Kroon DE, Casstevens TM, Ramdoss Y, Buckler ES (2007) TASSEL: software for association mapping of complex traits in diverse samples. Bioinformatics 23:2633–2635
14. Lipka AE, Tian F, Wang Q, Peiffer J, Li M, Bradbury PJ et al (2012) GAPIT: genome association and prediction integrated tool. Bioinformatics 28:2397–2399
15. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155:945–959
16. Yu J, Pressoir G, Briggs WH, Bi IV, Yamasaki M, Doebley JF, McMullen MD, Gaut BS, Nielsen DM, Holland JB, Kresovich S (2006) A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. Nat Genet 38 (2):203–208
17. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. J R Stat Soc 57:289–300
18. Bonferroni CE (1936) Teoria statistica delle classi e calcolo delle probabilità. Pubblicazioni del R Istituto Superiore di Scienze Economiche e Commerciali di Firenze 8:3–62

Chapter 9

Practical Workflow from High-Throughput Genotyping to Genomic Estimated Breeding Values (GEBVs)

Felice Contaldi, Elisa Cappetta, and Salvatore Esposito

Abstract

The global climate is changing, resulting in significant economic losses worldwide. It is thus necessary to speed up the plant selection process, especially for complex traits such as biotic and abiotic stresses. Nowadays, genomic selection (GS) is paving new ways to boost plant breeding, facilitating the rapid selection of superior genotypes based on the genomic estimated breeding value (GEBV). GEBVs consider all markers positioned throughout the genome, including those with minor effects. Indeed, although the effect of each marker may be very small, a large number of genome-wide markers retrieved by high-throughput genotyping (HTG) systems (mainly genotyping-by-sequencing, GBS) have the potential to explain all the genetic variance for a particular trait under selection. Although several workflows for GBS and GS data have been described, it is still hard for researchers without a bioinformatics background to carry out these analyses. This chapter has outlined some of the recently available bioinformatics resources that enable researchers to establish GBS applications for GS analysis in laboratories. Moreover, we provide useful scripts that could be used for this purpose and a description of key factors that need to be considered in these approaches.

Key words Next-generation breeding, Machine learning, Single-nucleotide polymorphisms (SNPs), Genomic estimated breeding values (GEBVs), Stacks, rrBLUP

1 Introduction

In the era of next-generation breeding, genomic selection (GS) is paving new opportunities to increase plant performance, especially for traits with polygenic inheritance [1–3]. In contrast to the traditional breeding such as marker-assisted selection (MAS), which has been successfully used for simple traits with a few major-effect genes [4–6], GS estimates the genetic potential of individual genotypes by using all genome-wide marker data (mainly SNPs), accelerating the breeding for traits regulated by a large number of small-effect genes (many "minor" gene effects) [1, 7]. Indeed, although each marker may have different effects (even minor), they are all useful to explain all the genetic variance within an experiment [8]. In this context, the selection process is

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_9,

achieved through statistical methods, which are able to accurately predict marker effects, increasing the rate of genetic gain per unit of time. Firstly, a GS prediction model is built using genotypic and phenotypic data of a training population and its accuracy is then determined by using a testing population. Finally, the trained model is employed to predict the GEBVs of non-phenotyped individuals, thus selecting superior genotypes. In this way, genomic screening of breeding populations accelerates the genetic gain obtained in each cycle, especially when selection is performed for traits with not a high heritability [9]. The potential breeding value of an individual is estimated through high-throughput genotyping (HTG) systems, which now allow to identify several thousand of variants in the entire genomes. Among HTGs, genotyping-by-sequencing (GBS) methods allow to rapidly detect nucleotide variation in many individuals simultaneously, and it is the most used method in plant genetic studies [10]. Genotypic data are then combined with phenotypic data taken through the next-generation phenomic systems, which combine high-throughput agri-systems and high-performance computing technologies for big data generation [10]. Finally, the GS models will be employed to predict breeding values for superior individuals through bioinformatic pipelines, which are needed to analyze and interpret the obtained results. However, it is still not easy for researchers without a bioinformatic background to carry out these analyses. In light of this reason, in this chapter, we focus on the entire protocol to process GBS data in the presence of a complete or draft genome (Fig. 1). We also provide the main scripts to help researchers in performing these analyses.

2 Materials

In this section are listed the main bioinformatics software and tools used to carry out a variant calling analysis from GBS and/or digest restriction site-associated DNA (RADseq) as well as genomic selection data. In detail, we show how researchers can install the needed packages in Linux.

1. **FastQC** (http://www.bioinformatics.babraham.ac.uk/projects/fastqc/) aims to provide a QC report where problems generated either in the sequencer or in the starting library material can be identified [11]. Since FastQC is a java application, we strongly suggest uploading the latest version of Java Runtime Environment (JRE) before starting with the installation process. Users can easily download FastQC, unzip the file, and launch it as follows:

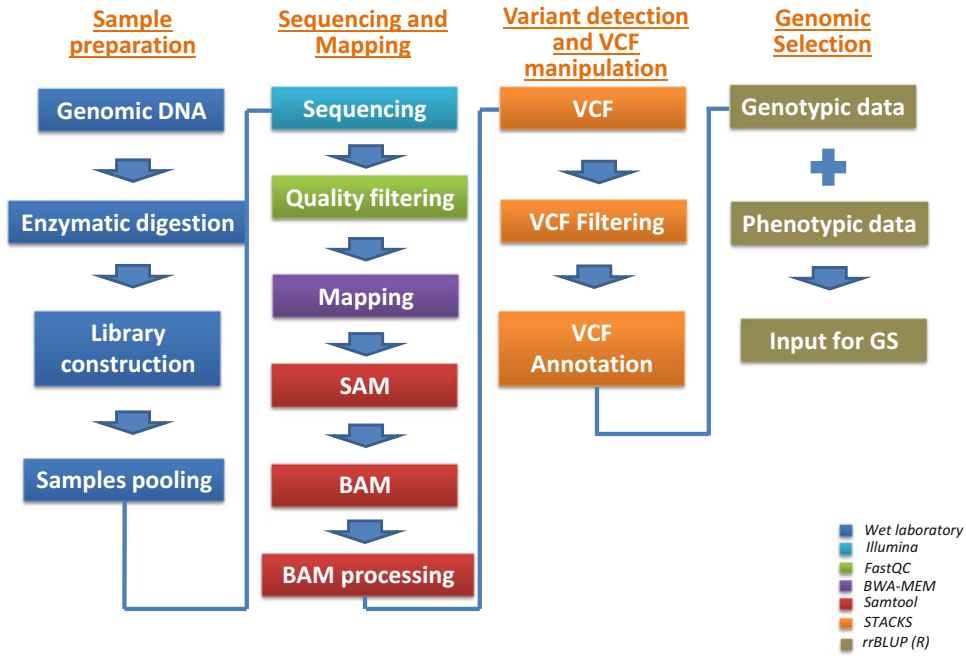


Fig. 1 Workflow from sample preparation to bioinformatic computation and pipelines toward genomic selection

```
$ wget http://www.bioinformatics.babraham.ac.uk/projects/
fastqc/fastqc_x.xx.x.zip
$ unzip fastqc_vx.xx.x.zip
$ cd FastQC
$ chmod +x fastqc
$ sudo mv FastQC/ /usr/local/
$ sudo ln -s /usr/local/FastQC/fastqc /usr/local/bin/fastqc
```

2. **Cutadapt** (https://cutadapt.readthedocs.io/en/stable/) is a stand-alone tool that can easily trim low-quality regions from Illumina sequencing reads [12]. Cutadapt is mainly written in Python, although the alignment algorithm is implemented in Python module:

```
$ wget http://cutadapt.googlecode.com/files/cutadapt-x.x.tar.gz
$ tar xzvf cutadapt-x.x.tar.gz
$ cd cutadapt-x.x
$ sudo apt install cutadapt
```

3. **BWA** (http://bio-bwa.sourceforge.net/) is a software package for mapping low-divergent sequences against a large reference genome [13]. It consists of three algorithms: BWA-backtrack, BWA-SW, and BWA-MEM. The latter is specifically designed for Illumina sequence reads up to 1 Mb and is generally recommended for high-quality queries as it is faster and more accurate. BWA can be installed in the following way:

```
$ wget http://downloads.sourceforge.net/project/bio-bwa/bwa-
x-x-x.tar.bz2
$ bunzip2 bwa-x.x.x.tar.bz2
$ tar xvf bwa-x.x.x.tar
$ cd bwa-x.x.x
$ make
$ export PATH=$PATH:/path/to/bwa-x.x.x
#Add bwa to your PATH by editing ~/.bashrc
```

Then execute the following commands to run it and test if the installation was done successfully:

```
$ source ~/.bashrc
$ bwa
```

4. **SAMtools** (http://samtools.sourceforge.net/) is a package of utilities designed for manipulating alignments in the SAM (sequence alignment/map) or BAM (binary alignment/map) format, including sorting, merging, indexing, and generating alignments in a per-position format [14]. SAMtools and BCFtools are distributed as individual packages. The code uses HTSlib internally (a C library for reading/writing high-throughput sequencing data), and both can be built independently as follows:

```
$ wget https://github.com/samtools/samtools/releases/down-
load/x.x.x/samtools-x.x.x.tar.bz2 -O samtools.tar.bz2
$ tar -xjvf samtools.tar.bz2
$ cd samtools-{version}
$ make
$ sudo make prefix=/usr/local/bin install
```

In order to use the BCFtools plug-ins, the environment variable must be set and linked to the correct location:

```
$ export BCFTOOLS_PLUGINS=/path/to/bcftools/plugins
```

Type "make install" to install the bcftools executable and associated scripts and a manual page to /usr/local. This can be changed by using the configure script's --prefix option:

```
$ ./configure --prefix=/path/to/install/dir
```

5. **Picard** (https://github.com/broadinstitute/picard) is a set of command-line tools (in Java) for manipulating high-throughput sequencing (HTS) data and formats such as SAM/BAM/CRAM and VCF [15]:

```
$ wget https://github.com/broadinstitute/picard/releases/download/x.xxx/picard-tools-x.xxx.zip -O picard-tools-x.xxx.zip
$ unzip picard-tools-x.xxx.zip
$ sudo mv picard-tools-x.xxx /usr/local
```

6. **Stacks** (http://creskolab.uoregon.edu/stacks/) allows de novo assembly of short read GBS data and identification of genetic variation in the presence and absence of a reference genome [16]:

```
$ wget http://creskolab.uoregon.edu/stacks
$ tar xfvz stacks-2.xx.tar.gz
$ cd stacks-2.xx
$ ./configure
$ sudo make install
```

7. **Plink** (https://www.cog-genomics.org/plink/1.9/) is a free, open-source whole-genome association analysis toolset designed to perform a range of basic, large-scale analyses in a computationally efficient manner [17]. The software is designed flexibly to perform a wide range of basic, large-scale genetic analyses:

```
$ cd /programinstallers/
$ wget -N https://www.cog-genomics.org/static/bin/plink$ver/plink_linux_x86_64.zip
$ mv plink_linux_x86_64.zip plink_${ver}_linux_x86_64.zip
$ mkdir /usr/local/bin/plink_$ver
$ cd /usr/local/bin
$ unzip /programinstallers/plink_${ver}_linux_x86_64.zip
```

Make a version-independent symlink:

```
$ ln -s plink_$ver plink
```

Add to default PATH for all users:

```
$ sudo nano /etc/profile
 PATH="$PATH:/usr/local/bin/plink"
```

8. **rrBLUP** (https://cran.r-project.org/web/packages/rrBLUP/index.html) is an R library (R version must be 2.14 or advanced) [18]. To install rrBLUP 4.4 click on "Install Package" in the drop-down menu available in the R graphical interface and select rrBLUP. Alternatively, you can install rrBLUP package from local zip (Menu Bar: Packages-> Install package from local zip files). The user manual (pdf version) can be downloaded from https://cran.rproject.org/web/packages/rrBLUP/rrBLUP.pdf.

3 Methods

3.1 The GBS Protocol

The workflow for variant discovery invocates three main steps: (1) raw data processing, (2) read alignment to a reference genome or de novo assembly, and (3) variant discovery, filtering, and annotation (Fig. 1). In the following sections, we deepen the focus on these steps to provide background information for researchers who want to use the available bioinformatics tools to perform various tasks. For each tool, we also provide the main options used (Table 1).

3.1.1 Raw Data Processing

Demultiplexing is the first key step of processing raw sequencing data, which separates reads into their corresponding samples based on barcode matching. Demultiplexing of Illumina reads can be carried out using the program "process_radtags" implemented in the stacks workflow [16]. The program examines raw reads from an Illumina sequencing and checks for intact barcode and RAD cut-site. Then demultiplexing is performed. Process_radtags can be used with both single-end and paired-end Illumina reads, and a list of barcodes is needed to better separate the samples. The program can be launched using the following command line:

```
$ process_radtags -p in_dir [--paired [--interleaved]] [-b
barcode_file] -o out_dir -e enz [-c] [-q] [-r] [-t len]
```

Table 1
List of the main options used in a genotype-by-sequencing (GBS) analysis

Program/tool	Option used	Description
process_radtags	-p	Path to a directory of files
	-b	Path to a file containing barcodes
	-o	Path to output the processed files
	-e	Provide the restriction enzyme used
	-c	Clean data, remove any read with an uncalled base
	-q	Discard reads with low-quality scores
	-r	Rescue barcodes and RAD-Tags
cutadapt	-a	SEQUENCE of an adapter ligated to the 3′ end
	-o	Write trimmed reads to FILE
	-e	Maximum allowed error rate as value between 0 and 1
	-q	Quality cutoff
	-m	Discard reads shorter than LEN. Default: 0
bwa	-p	Prefix of the output database [same as db filename]
	-M	Mark shorter split hits as secondary (for Picard compatibility)
	-t	Number of threads
samtools	view	Command to convert SAM files in BAM
	sort	Command to sort BAM files
	index	Command to index BAM files
Picard	-I	Input file in BAM format
	-O	Output file name
ref_map.pl (Stacks)	--samples_dir	Path to the directory containing the samples BAM
	--popmap	Path to a population map file (format is "TAB," one sample per line)
	--O	Path to an output directory
Populations (Stacks)	-P	Path to the directory containing the Stacks files
	--popmap	Path to a population map (format is "SAMPLE1POP1\n...")
	-p	Minimum number of populations a locus must be present in to process a locus
	-r	Minimum percentage of individuals in a population required to process a locus for that population
	-t	Number of threads to run in parallel sections of code

It is important to note that users need to know where to find the barcodes in the sequencing data. For instance, if your data are single-end or paired-end and the barcodes are localized at the beginning of the reads, users need to specify the --inline_null flag. By contrast, if barcodes are at the end of the first line of fastq file, researchers can use the --index_null flag. Examples of inline and index barcodes are listed in the following link: http://catchenlab.life.illinois.edu/stacks/manual/. Once samples are demultiplexed, a quality assessment and correction of reads by filtering or trimming are necessary to remove the various type of errors and artifacts, such as base calling errors, low-quality bases, adaptor contamination, and duplicate reads. Numerous publicly available software for preprocessing of sequencing reads are available in the literature, including Trimmomatic (http://www.usadellab.org/cms/?page=trimmomatic) [19], FASTX-Toolkit (tool integrated into the Galaxy platform—http://hannonlab.cshl.edu/fastx_toolkit/) [20], and cutadapt (http://code.google.com/p/cutadapt/) [12]. In our pipeline, we used cutadapt [12], which is a fast, easy, and multicore software that removes adapter sequences and low-quality regions (below a user-defined quality threshold). The program can be launched in a terminal window, with the below code, in order to filter fastq files by quality (-q), minimum length (-m), and maximum error rate (-e). The sequence of the adapter is given with the "-a" option:

```
$ cutadapt -a adapters_file.txt -o output_directory -e 0.2 -q
30 -m 50
```

3.1.2 Mapping to Reference Genome

After preprocessing and cleanup, the next step is the mapping of short reads versus a reference genome, although such approaches are also available when reference genome is lacking (*see* **Note 1**). In the last years, several short-read alignment programs, such as MAQ [21], STAMPY [22], Bowtie2 [23], BWA [13], and SOAP2 [24] have been developed, although the BWT-based aligners are preferred since they use up only a limited amount of memory. In our workflow (Fig. 1), we adopted BWA-MEM which is the latest and it is generally recommended for high-quality queries as it is faster and more accurate [13]. BWA-MEM also has better performance than BWA and SOAP2 for 70–100 bp Illumina reads. Firstly, users need to index their reference genome as follows before performing the mapping step:

```
$ bwa index -p genome_folder/genome.index genome.fasta
```

The option -a [bwtsw/is] can also be implemented, where "bwtsw" is used for long and large genomes such as the whole human genome, whereas "is" for short ones. When the indexed

reference genome is ready, the mapping step can be run using the following code:

```
$ bwa mem -M -t 10 reference.fasta sample_r1.fastq sample_r2.
fastq > aln.sam
```

The output files are in SAM format, which contains alignment data in human-readable tab-delimited text. However, SAM files generally tend to be very large, whereas BAM format (a compressed binary version of SAM) is preferred for the downstream variant detection analyses due to its relatively smaller size. Here, we use the "view" command of SAMtools [14] to convert mapped reads from SAM to BAM format:

```
$ samtools view -Sb input.sam > output.bam
```

For downstream analysis, the BAM files must be sorted and indexed according to the chromosomal positions. To achieve this, we use the sort and index utilities of SAMtools:

```
$ samtools sort --threads "$NCPU" -o output.sorted.bam input.
bam
$ samtools index output.sorted.bam
```

One key point in many GBS pipelines is to remove PCR duplicate results originated from the original DNA templates and that have been sequenced many times. In addition, researchers may consider the possibility to discard multi-mapped reads (*see* **Note 2**). Duplicated reads may have a detrimental effect on the quality of the variant calls, especially when the coverage is low, leading to false-positive variant calls. Computational methods for the detection and removal of PCR duplicates have become available and generally rely on the observation of identical alignment positions of reads to the reference genome. Among them, Picard (MarkDuplicates) [14] and SAMTools (rmdup) [15] are the two main software used for PCR duplicate removal. The MarkDuplicates tool implemented in Picard works by comparing sequences in the 5′-positions of both reads in a SAM/BAM file. The main output is a new SAM or BAM file, in which duplicates have been identified in the SAM flag field for each read:

```
$ Java -jar picard.jar MarkDuplicates I=input.bam O=marked_du-
plicates.bam M= marked_dup_metrics.txt
```

As shown in the above example, MarkDuplicates also produces a metric file indicating the numbers of duplicates for both single- and paired-end reads. Finally, duplicates can be removed using the "remove_duplicate" and "remove_sequencing_duplicates" flags.

3.1.3 Variant Calling and Identification of Genomic Variants

The final step is to call sequence variants (mainly SNPs and InDels) from the processed BAM file. Several software tools for variant calling, including SAMtools:mpileup/BCFtools [14], GATK [25], SOAP [26], SNVer [27], and GNUMAP [28], are suitable for this purpose. In our variant calling workflow, we have implemented the most commonly used SNP caller: STACKS [16]. In particular, the "ref_map.pl" script was used for this purpose; it executes the Stacks pipeline by running each of its components individually:

```
$ ref_map.pl --samples dir --popmap path [-s spacer] [--
paired] --o dir [-X prog:"opts" ...]
```

The script uses a population map file in "TAB" format (one sample per line) as follows:

```
Sample1.bam   pop1
Sample2.bam   pop1
Sample3.bam   pop1
Sample4.bam   pop2
Sample5.bam   pop2
```

This is the simplest way to run Stacks and it handles many of the details, such as sample numbering. Once the "ref_map.pl" is executed, the population program (implemented in Stacks) can be launched to finish the analysis [16]. The program will analyze a population of individual samples computing a number of genetics statistics such as expected/observed heterozygosity, π, and F_{IS} at each nucleotide position. In addition, the program creates a variety of standard output formats, including the VCF (variant call format) format containing all the identified variants, which is the emerging standard for storing variant data:

```
$ Populations -P ./stacks/ --popmap ./samples/popmap --smooth
-p 10 -r 0.75 -t8
```

3.1.4 SNP Filtering

Filtering raw SNP candidates is an essential step in the genotyping workflow as it helps in reducing false-positive calls made from biases in the sequencing data and removes those calls that do not fulfill specific thresholds for SNP and genotype properties. Although

most of the currently available variant calling pipelines such as SAMtools [14], GATK [25], and STACKS [16] include SNP filtering of false-positive calls based on read depth and quality threshold, in our pipeline we perform additional filtering based on missing genotyping calls and minor allele frequency (MAF). For this purpose, a Perl script named "filter_vcf.pl" (https://github.com/aquaskyline/16GT/blob/master/filterVCF.pl) can be used to perform filtering based on missing genotype and ignoring SNPs with a MAF less than 5%.

3.2 Looking for Optimal Parameters to Set Up a GS Experiment

Although in literature several GS protocols are reported, little is known regarding the establishment of optimal parameters. Selection response depends on TRN size, marker density, marker linkage disequilibrium, knowledge of the pedigree structure, precision of the phenotyping, and statistical methods used to predict the GEBVs (*see* also **Notes 3** and **4**). It has been shown that the highest accuracy was reached with large TRNs, although the optimal size seems to be highly influenced by the relationship between TRS and TST as revealed by [29, 30]. In particular, if TRS is unrelated or distantly related to the TST, the accuracy tends to be low (0.2 and 0.4), whereas it increases in the case of TRS fully related to TST (0.4 and 0.7) [31]. Thus, developing ad hoc TRN could improve the prediction of accuracy fixing advantageous alleles in the second and third cycles of recombination. In addition to TRN size and its composition, the quality and the density of markers have a strong impact on GS models. Since the goal of GS model is to capture the genetic variation as much as possible, a higher marker density may be suitable to improve prediction accuracy [32]. However, it is important to filter raw SNPs by the percentage of missing values (PMV) and minor allele frequency (MAF) in order to reduce false positive. To date, several statistical methods have been developed to estimate the marker effects in the TRN using filtered SNPs [33]. Briefly, the current GS methods are classified into two groups based on the different assumptions regarding the marker effect distribution and variances. The first group, which includes ridge regression best linear unbiased prediction (rr-BLUP) and genomic best linear unbiased prediction (G)BLUP [34], assumes that all marker effects are normally distributed and that the variance of each marker is the same [35]. By contrast, the second one, mainly Bayesian methods (BayesA, BayesB, Bayesian LASSO, and BayesR), assumes that marker effects have different statistical distributions and variances [35]. In the case of trait affected by many small-effect genes, the GEBV values are more effectively predicted by methods like (G)BLUP, whereas Bayesian methods are suitable when considering traits controlled by larger QTL or when considering prediction of unrelated individuals [36].

3.2.1 Tutorial for GS Using rrBLUP Package in R

In this section we briefly describe how to carry out a GS protocol using R. The original codes have been provided in the rrBLUP manual, although some of them have been modified for specific purposes:

 # Load rrBLUP in R

Open a new R session (version 3.2.3) and run the following codes to create your own folder. Place both input files (genotypes and phenotypes) in the same folder:

```
$ library(rrBLUP)
$ setwd("C:/Users/Desktop/folder_of_choise")
```

 # Import the Genotype/Marker and Phenotype Data

Using the following codes, users can load the markers and phenotype data. SNP data need to be converted in −1,0,1 matrix (1 = homozygous for parent 1, 0 = heterozygous, and −1 homozygous for parent 2). The new matrix will have row-wise for plant IDs and column-wise for markers, whereas phenotype data will contain row-wise plant ID and column-wise phenotypes. Remember to use "Header = F" for SNP data, since the file does not have a header with marker names but "Header = T" for phenotypes:

```
$ Markers <- as.matrix(read.table(file="snp.txt"), header=F)
$ Pheno <-as.matrix(read.table(file ="traits.txt", header=-
TRUE)
```

 # Impute NA markers using A.mat option

```
impute=A.mat(Markers,max.missing=0.5,impute.method="mean",re-
turn.imputed=T)
Markers_impute=impute$imputed
```

 # Remove markers with more than 50% missing data

The function for() loop identifies the NA in any cell deleting that row. Usually, it takes a while to run:

```
$ for(i in 1: nrow(Markers_impute)){for(j in 1: ncol(Marker-
s_impute)){myImputedData<- Markers_impute [which(rowSums(Mar-
kers_impute[,], na.rm =
FALSE, dims=1) != "NA"),]}}
```

 # Define the training and test populations

Here, the size of training and test dataset is defined. Out of 96 total samples, 38 are randomly chosen for training using the function sample(), whereas the function setdiff() determines the numbers that are not in the training population and will be part of the validation population. Finally, "Pheno_train" and "m_train" are the phenotype and marker matrices belonging to the training population, whereas "Pheno_valid" and "m_valid" refer to samples and markers in the validation populations:

```
$ train= as.matrix(sample(1:96, 38))
$ test<-setdiff(1:96,train)
$ Pheno_train=Pheno[train,]
$ m_train=Markers_impute2[train,]
$ Pheno_valid=Pheno[test,]
$ m_valid=Markers_impute2[test,]
```

 # Run mixed.solve[ ] on the trait of interest

The mixed.solve() function calculates maximum likelihood (ML) or restricted ML (REML) solutions for mixed model like $y = X\beta + Zu + e$ # where y is the $n*1$ size vector of observations. If there is any NA or missing value, it will delete the corresponding rows of X and Z. # Z is the $n*m$ sized design matrix for the random effects. By default, it is assumed to be the identity matrix. # K is the $m*m$ sized, positive semi-definite covariance matrix of random effects:

```
$ yield=(Pheno_train[,1])
$ yield_answer <- mixed.solve(yield, Z=m_train, K=NULL, SE=
FALSE, return.Hinv =FALSE)
```

Yield_answer$u is the output of the marker effects. To see the results regarding the first five markers, users can use the function "as.matrix" as follows:

```
$ YLD = yield_answer$u
$ e = as.matrix(YLD)
$ pred_yield_valid = m_valid %*% e
$ pred_yield=(pred_yield_valid[,1])+yield_answer$beta
$ pred_yield
```

 # Determine the model accuracy

In this step the correlation between the predicted and observed values is calculated. Note that the accuracy will

change slightly each time, mainly due to the different number of individuals sampled for the training and validation populations:

```
$ yield_valid = Pheno_valid[,1]
$ YLD_accuracy <-cor(pred_yield_valid, yield_valid, use="com-
plete")
$ YLD_accuracy
```

 # Cross validation for many cycles for yield only

```
traits=1
cycles=500
accuracy = matrix(nrow=cycles, ncol=traits)
for(r in 1:cycles)
{train= as.matrix(sample(1:96, 29))
test<-setdiff(1:96,train)
Pheno_train=Pheno[train,]
m_train=Markers_impute2[train,]
Pheno_valid=Pheno[test,]
m_valid=Markers_impute2[test,]
yield=(Pheno_train[,1])
yield_answer<-mixed.solve(yield, Z=m_train, K=NULL, SE =
FALSE, return.Hinv=FALSE)
YLD = yield_answer$u
e = as.matrix(YLD)
pred_yield_valid = m_valid %*% e
pred_yield=(pred_yield_valid[,1])+yield_answer$beta
pred_yield
yield_valid = Pheno_valid[,1]
accuracy[r,1] <-cor(pred_yield_valid, yield_valid, use="com-
plete" )
mean(accuracy)
```

4 Notes

1. In the absence of a reference genome, paired-end sequencing data generated by RAD-seq or GBS approaches can be assembled de novo using software packages such as STACKS, UNEAK, or RApiD to produce mini-contigs that can be used as a reference for read mapping and genotyping.
2. Multi-mapped reads are those that align to multiple locations within the reference genome sequence. Most eukaryotic organisms, including polyploid plants, harbor a high number of orthologous and paralogous gene families, which contain

multiple isoforms nearly identical or similar sequences. For this reason, shorter reads are less specific, tending to have more multi-mapping events. Although the proportion of multi-mapped reads ranges from 20% to 60%, discarding a high proportion of them will result in a significant loss of valuable information. Therefore, it is a good practice to take into account multi-mapped reads and use some Perl utility scripts such as bowtie2_extract_best_global_hit.pl or bowtie2_extract_best_local_hit.pl to go through the SAM files and identify the best hit from multi-mapped reads.

3. Values will be different every time it is run since different lines will be included in the training or validation sets.
4. Accuracy is affected by training size, validation size, number of markers, and heritability.

Acknowledgment

The authors thank the BRESOV (Breeding for resilient, efficient and sustainable organic vegetable production) and TomGEM (A holistic multi-actor approach toward the design of new tomato varieties and management practices to improve yield and quality in the face of climate change) projects founded by the European Union Horizon 2020 research and innovation program under grant agreement No. 774244 and No. 679796, respectively. We also thank D'Acunzo D.M. for editing the manuscript.

References

1. Heffner EL, Sorrells ME, Jannink J (2009) Genomic selection for crop improvement. Crop Sci 49(1):1–12. https://doi.org/10.2135/cropsci2008.08.0512
2. Crossa J, De Los Campos G, Pérez P et al (2010) Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186:713–724. https://doi.org/10.1534/genetics.110.118521
3. Lorenz AJ, Chao S, Asoro FG et al (2011) Genomic selection in plant breeding. Knowledge and prospects. Adv Agron 110:77–123. https://doi.org/10.1016/B978-0-12-3855312.00002-5
4. Villano C et al (2018) High-throughput genotyping in onion reveals structure of genetic diversity and informative SNPs useful for molecular breeding. Mol Breed 39(1). https://doi.org/10.1007/s11032-018-0912-0
5. Collard BC, Mackill DJ (2008) Marker-assisted selection: an approach for precision plant breeding in the twenty-first century. Philos Trans R Soc Lond Ser B Biol Sci 363:557–572. https://doi.org/10.1098/rstb.2007.2170
6. Crossa J, Pérez-Rodríguez P, Cuevas J et al (2017) Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci 22(11):961–975. https://doi.org/10.1016/j.tplants.2017.08.011
7. Dekkers JCM, Hospital F (2002) The use of molecular genetics in the improvement of agricultural populations. Nat Rev Genet 3:22–32. https://doi.org/10.1038/nrg701
8. Wang X, Xu Y, Hu Z, Xu C (2018) Genomic selection methods for crop improvement: current status and prospects. Crop J 6:330–340. https://doi.org/10.1016/j.cj.2018.03.001
9. Heffner EL, Sorrells ME, Jannink JL (2009) Genomic selection for crop improvement, vol 49, pp 1–12. https://doi.org/10.2135/cropsci2008.08.0512

10. Esposito S, Carputo D, Cardi T, Tripodi P (2019) Applications and trends of machine learning in genomics and phenomics for next-generation breeding. Plants 9(1). https://doi.org/10.3390/plants9010034
11. Van der Auwera GA et al (2013) From FastQ data to high confidence variant calls: the genome analysis toolkit best practices pipeline. Curr Protoc Bioinformatics 43:11.10.1–11.10.33. https://doi.org/10.1002/0471250953.bi1110s43
12. Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J 17:10–12. https://doi.org/10.14806/ej.17.1.200
13. 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
14. Li H, Handsaker B, Wysoker A et al (2009) Genome project data processing subgroup The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352
15. Picard Toolkit (2018) Broad Institute, GitHub Repository. http://broadinstitute.github.io/picard/
16. Catchen J, Hohenlohe P, Bassham S et al (2013) Stacks: an analysis tool set for population genomics. Mol Ecol. https://doi.org/10.1111/mec.12354
17. Purcell S, Neale B, Todd-Brown K et al (2007) PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 81:559–575. https://doi.org/10.1086/519795
18. Endelman JB (2011) Ridge regression and other kernels for genomic selection with R package rrBLUP. Plant Genome 4:250–255. https://doi.org/10.3835/plantgenome2011.08.0024
19. Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btu170
20. Blankenberg D et al (2010) Manipulation of FASTQ data with Galaxy. Bioinformatics 26:1783–1785. https://doi.org/10.1093/bioinformatics/btq281
21. Li H, Ruan J, Durbin R (2008) Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome Res 18:1851–1858. https://doi.org/10.1101/gr.078212.108
22. Lunter G, Goodson M (2011) Stampy: a statistical algorithm for sensitive and fast mapping of Illumina sequence reads. Genome Res 21:936–939. https://doi.org/10.1101/gr.111120.110
23. Langmead B, Trapnell C, Pop M, Salzberg SL (2009) Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. https://doi.org/10.1186/gb-2009-10-3-r25
24. Li R, Yu C, Li Y et al (2009) SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics 25:1966–1967. https://doi.org/10.1093/bioinformatics/btp336
25. McKenna A, Hanna M, Banks E et al (2010) The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res 20:1297–1303. https://doi.org/10.1101/gr.107524.110
26. Li R, Li Y, Kristiansen K, Wang J (2008) SOAP: short oligonucleotide alignment program. Bioinformatics 24:713–714. https://doi.org/10.1093/bioinformatics/btn025
27. Wei Z, Wang W, Hu P, Lyon GJ, Hakonarson H (2011) SNVer: a statistical tool for variant calling in analysis of pooled or individual next-generation sequencing data. Nucleic Acids Res 39:e132. https://doi.org/10.1093/nar/gkr599
28. Clement NL, Snell Q, Clement MJ et al (2010) The GNUMAP algorithm: unbiased probabilistic mapping of oligonucleotides from next-generation sequencing. Bioinformatics 26:38–45
29. Schopp P, Müller D, Technow F, Melchinger AE (2017) Accuracy of genomic prediction in synthetic populations depending on the number of parents, relatedness, and ancestral linkage disequilibrium. Genetics 205:441–454. https://doi.org/10.1534/genetics.116.193243
30. Edwards SM, Buntjer JB, Jackson R et al (2019) The effects of training population design on genomic prediction accuracy in wheat. Theor Appl Genet 132:1943–1952. https://doi.org/10.1101/443267
31. Bassi FM, Bentley AR, Charmet G, Ortiz R, Crossa J (2016) Breeding schemes for the implementation of genomic selection in wheat (*Triticum Spp.*). Plant Sci 242:23–36. https://doi.org/10.1016/j.plantsci.2015.08.021

32. Zhang H, Yin L, Wang M et al (2019) Factors affecting the accuracy of genomic selection for agricultural economic traits in maize, cattle, and pig populations. Front Genet 10:189. https://doi.org/10.3389/fgene.2019.00189
33. Robertsen CD, Hjotrtshøj RL, Janss LL (2019) Genomic selection in cereal breeding. Agronomy 9:1–16. https://doi.org/10.3390/agronomy9020095
34. Whittaker JC, Thompson R, Denham MC (2000) Marker-assisted selection using ridge regression. Genet Res 75:249–252. https://doi.org/10.1017/S0016672399004462
35. Meuwissen THE, Hayes BJ, Goddard ME (2001) Prediction of total genetic value using genome-wide dense marker maps. Genetics 157:1819–1829
36. De los Campos G, Hickey JM, Pong-Wong R et al (2013) Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193:327–345. https://doi.org/10.1534/genetics.112.143313

Chapter 10

Guidelines for Setting Up a mRNA Sequencing Experiment and Best Practices for Bioinformatic Data Analysis

Teresa Rosa Galise, Salvatore Esposito, and Nunzio D'Agostino

Abstract

RNA-sequencing, commonly referred to as RNA-seq, is the most recently developed method for the analysis of transcriptomes. It uses high-throughput next-generation sequencing technologies and has revolutionized our understanding of the complexity and dynamics of whole transcriptomes.

In this chapter, we recall the key developments in transcriptome analysis and dissect the different steps of the general workflow that can be run by users to design and perform a mRNA-seq experiment as well as to process mRNA-seq data obtained by the Illumina technology. The chapter proposes guidelines for completing a mRNA-seq study properly and makes available recommendations for best practices based on recent literature and on the latest developments in technology and algorithms. We also remark the large number of choices available (especially for bioinformatic data analysis) in front of which the scientist may be in trouble.

In the last part of the chapter we discuss the new frontiers of single-cell RNA-seq and isoform sequencing by long read technology.

Key words Transcriptome, RNA-sequencing, Bioinformatics, Experimental design, Biological replicates, Assembly, Summarization, Normalization, Differentially expressed genes, scRNA-seq, Iso-seq

1 Introduction

1.1 The Evolution of Transcriptomics

The transcriptome is the set of all RNAs in one cell or in a population of cells and reflects all genes that are being actively expressed at any given time in cells in response to genetic factors and environmental stimuli. Indeed, not all genes are transcriptionally active in a given cell; in other words, each cell or population of cells show a unique pattern of gene expression.

The term "transcriptome" was proposed by Charles Auffray in 1996 [1] and appeared in a scientific paper in 1997 for the first time [2]. The transcriptome can be analyzed by analogical and digital methods. The former are based on fluorescence intensity detection to measure gene expression, whereas the latter are based on the generation of sequence tags [3].

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_10,

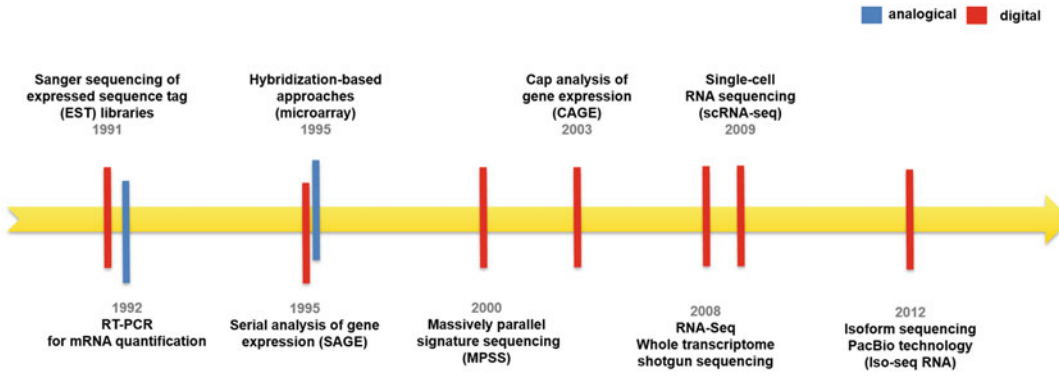


Fig. 1 Timeline of key developments in transcriptome analysis. Analogical and digital approaches are in blue and red, respectively

In Fig. 1 we describe the timeline of key developments in transcriptome analysis and distinguish between analogical and digital approaches.

In 1991 Adams and colleagues started a pilot project to evaluate the use of partial cDNA sequences (i.e., *e*xpressed *s*equence *t*ags; ESTs) for the large-scale investigation of gene expression in humans [4]. For over a decade, EST collections of different species have dramatically increased in size, as they have been the most attractive route for sampling and studying transcriptomes due to their versatility and wide range of applications [5]. Basically, EST projects were developed to complement existing genome sequencing projects [6]; however, they were also established as low-cost alternatives to explore the "gene space" of neglected species [7].

All those projects led to the development of primary [8] and secondary [9, 10] sequence repositories equipped with user-friendly Web interfaces that allow users to investigate in detail EST information content.

In 1992 the quantitative real-time polymerase chain reaction (qPCR) was developed [11] to detect, characterize, and quantify in "real time" nucleic acids. Commonly, in RT-qPCR (*r*everse *t*ranscription *q*uantitative *p*olymerase *c*hain *r*eaction), mRNAs are first reverse transcribed into cDNA and then qPCR is carried out. DNA is amplified by repeating the following three steps: denaturation, annealing, and elongation; fluorescent labeling (i.e., dyes or probes) enables the amount of PCR product to be monitored as PCR progresses. In dye-based qPCR, fluorescent labeling allows amplified DNA molecules to be quantified by using a double-stranded DNA (dsDNA)-binding dye. In this way, only one target at a time can be amplified as the dye will bind to any dsDNA in the sample. Contrariwise, probe-based qPCR requires the design of target-specific probe(s) associated with fluorophore and quencher and can be used to simultaneously target many mRNAs thanks to the specificity of the probes.

Sequence-based approaches complement this method, which however remains widely used above all when investigating the gene expression of a few genes. Over time, more sensitive instrumentation and highly efficient detection chemistries have been developed, thus making this technique more reliable.

The high-throughput quantification of a transcriptome started to become a real possibility with the development of gene expression microarrays [12]. Specific sequences (i.e., probes) are immobilized ("ink-jet printed") to or synthesized in situ in defined positions of a solid surface (i.e., chip array). Then, labeled DNA fragments from a sample are hybridized to the chip array. Messenger RNA abundance can be measured using either a "one-color" or a "two-color" design. While in "one-color" design each sample (be it the test or the control) is labeled and hybridized to a separate microarray, in "two-color" design two biological samples (test sample and control sample) are labeled with different fluorescent dyes, usually cyanine 3 and cyanine 5, and then simultaneously hybridized onto the same chip array (i.e., competitive hybridization). In both cases the measured fluorescence corresponds to the abundance of each mRNA in one sample.

In 1995 a novel digital experimental technique, referred to as *s*erial *a*nalysis of *g*ene *e*xpression (SAGE), was designed to better gain a quantitative measure of gene expression in a particular type of cell or tissue [13]. The SAGE method was based on the isolation of unique sequence tags (9–10 bp in length) from mRNAs and on their concatenation into long molecules to be subjected to Sanger sequencing. The limited size of SAGE tags soon turned out to be not always sufficient to unambiguously detect the gene from which the tag is derived. Therefore, different versions of the original SAGE protocol were developed (i.e., LongSAGE and SuperSAGE) in order to produce longer transcript tags [14, 15].

*M*assively *p*arallel *s*ignature *s*equencing (MPSS) is a method similar to SAGE, introduced in 2000 to acquire in a single operation hundreds of thousands of sequence tags and perform in-depth gene expression profiling [16]. MPSS generates short (17–20 bp) tag sequences adjacent to the 3′ end of mRNAs. Each tag sequence is cloned (roughly 100,000 amplified copies) onto an individual microbead. All the different microbeads (each corresponding to a single mRNA) are then arrayed in a flow cell for Sanger sequencing.

As SAGE and MPSS allow short tags at the 3′ ends of mRNAs to be obtained, a method, never previously described, called *c*ap *a*nalysis *g*ene *e*xpression (CAGE) was introduced in 2003 [17]. CAGE is based on the sequencing of concatemers of DNA tags from the 5′ end of mRNA at the cap sites and allows gene expression profiling and identification of transcriptional start points.

RNA-sequencing, commonly referred to as RNA-seq, is the most recently developed method for the analysis of transcriptomes [18].

1.2 The RNA-Sequencing Revolution

RNA-seq uses high-throughput next-generation sequencing (NGS) technologies and brought qualitative and quantitative improvement to transcriptome analysis. It has clear advantages over existing approaches [3] and has revolutionized our understanding of the complexity and dynamics of whole transcriptomes [19].

RNA-seq combines into a single high-throughput assay the discovery and quantification of transcripts. It is particularly attractive for non-model organisms as it does not require any a priori knowledge of the genome of the target species. Compared with all the methods developed so far for transcriptome analysis (*see* Subheading 1.1), RNA-seq provides better resolution (very high dynamic range) and representativeness, facilitates the discovery of novel genes and isoforms, and has a wide range of applications [20].

A typical RNA-seq experiment consists of a few steps, including RNA isolation, complementary DNA (cDNA) conversion, library preparation, next-generation sequencing, and data analysis using bioinformatic tools. In the next sections, we review all the major steps of a RNA-seq experiment, from experimental design to data analysis.

2 Dissecting the RNA-Seq Workflow

Figure 2 shows the different steps of the general workflow that can be run by users to design and perform a mRNA-seq experiment and to process mRNA-seq data.

Below, we review the individual steps and highlight the challenges associated with each step. The main goal of this chapter is to make available guidance and recommendations for best practices based on recent literature and on the latest developments in technology and algorithms. Providing an exhaustive list of the standard bioinformatic resources/tools for the analysis of mRNA-seq data is not the goal of this chapter, which instead proposes guidelines for completing a mRNA-seq study properly. The latter clearly depends on the species under investigation as well as on scientists' research goals.

2.1 The Importance of Having a Robust Experimental Design

A good experimental design is critical for a successful RNA-seq experiment. Sampling, randomization, blocking, and replication are all basic elements of any well-planned RNA-seq design. The manuscript by [21] provides a detailed overview of all possible statistical designs that allow to distinguish biological variations

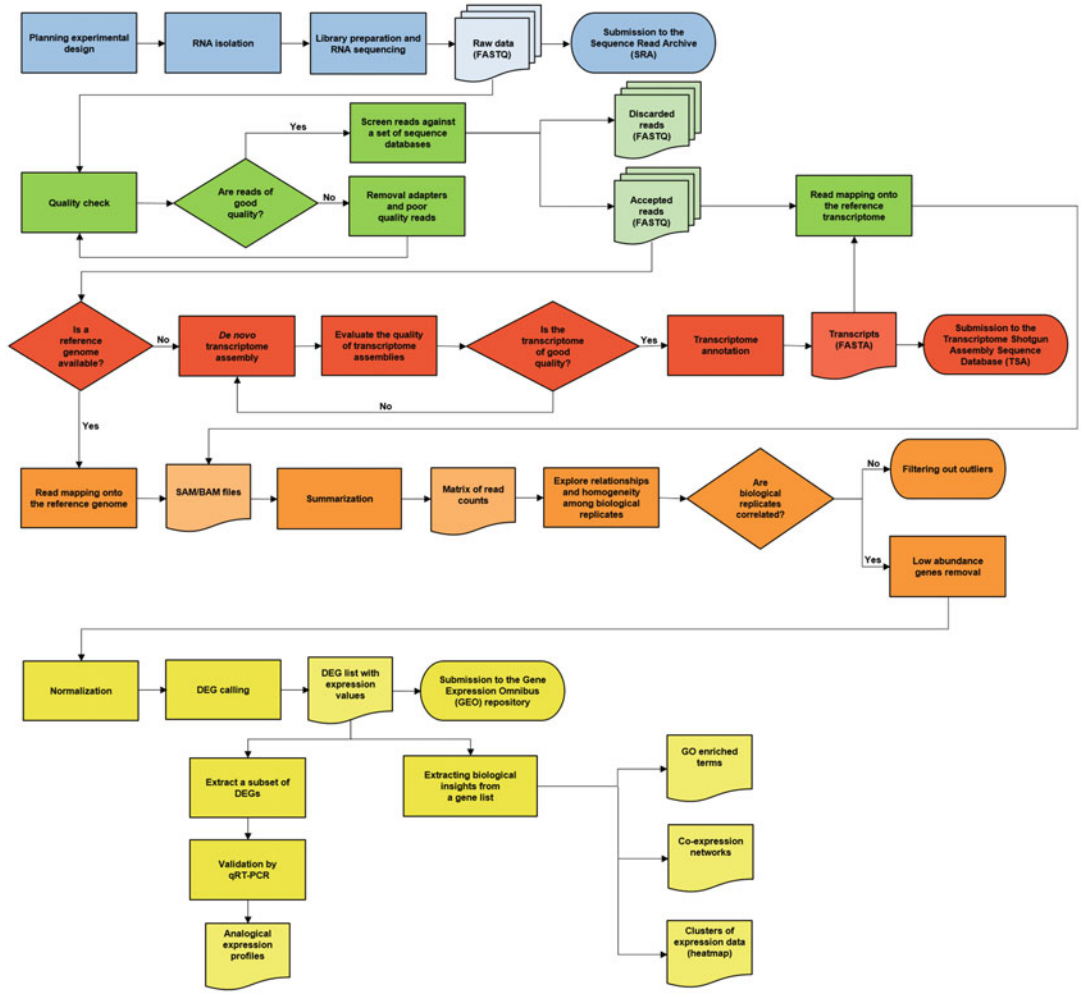


Fig. 2 Decision tree flowchart for mRNA-seq data analysis. (Figure has been deposited at the Figshare repository, https:/doi.org/10.6084/m9.figshare.11877417.v1)

from technical ones. Herein, we are going to summarize the main concepts of this milestone work.

Three levels of sampling can be defined in a RNA-seq experiment. First of all, it is necessary to identify organisms or individuals from a larger population to which results of the study may be generalized (i.e., subject sampling). RNA sampling and fragment sampling follow. They occur during library preparation when RNA molecules are isolated from tissues/organs/cell(s) and only a subset of fragmented RNAs are retained for amplification and sequencing. Clearly, incorrect sampling can contribute to negatively affect the results of a RNA-seq study. Indeed, non-randomly selected samples are subjected to a selection bias, which is the tendency to under- or over-represent a part of the population.

If biological effects (control vs. treatment) are not separable from confounding factors there is no way of knowing whether the observed difference in RNA abundance between treatment groups is due to biological or technical issues.

Two sources of variation may contribute to confounding effects in RNA-seq data: batch effects and lane effects. Batch effects are any errors that occur after RNA fragmentation until it is uploaded into the flow cell (e.g., PCR amplification and reverse transcription artifacts). Lane effects are any errors that occur during the sequencing reaction until base calling.

Typically, independent RNA samples are loaded into different lanes of the flow cell; in this way the sequencing reaction takes place independently for each sample (unblocked design). Multiplexing (i.e., bar-coded samples) eliminates additional sources of variation caused by lane or batch effects because the batch effects are the same for all samples and because the sequencing reaction occurs in a single lane for all samples. This experimental design is referred to as a balanced block design (BBD).

In case the number of treatments exceeds the number of unique bar codes in one lane, BBD is not possible and a balanced incomplete block design can be used [21].

In the absence of biological replicates, the within-group variability cannot be estimated and the generalization of results gathered from unreplicated data can lead to unrealistic conclusions. Generally, the overall costs of a RNA-seq project affect the number of biological replicates. A minimum of three biological replicates *per* condition is generally required. Clearly, the greater the number of biological replicates, the greater the statistical power and the generalizability of the results [22].

2.2 RNA Isolation

Molecular scientists and biotech companies have developed several strategies to isolate RNA from different plant tissues/organs. These approaches are based on the use of enzymes and chemical products to break the cell wall (i.e., cell lysis). A non-negligible issue to be considered is that RNA is fragile compared with DNA; it therefore degrades very easily, mainly due to its sensitivity to RNases (i.e., ribonucleases responsible for the degradation of RNA molecules). For this reason, the most common strategy for RNA isolation is based on free RNase cell lysis buffer [23], which is usually made of stabilization solution to protect cellular RNA, also minimizing the need to immediately process samples. Among the stabilization solution, RNAlater™ (Thermo Scientific, Wilmington, DE, USA, and Qiagen, Valencia, CA, USA) and RNAstable™ (Sigma-Aldrich, St Louis, MO, Canada) are the most widely used for RNA protection. Generally, the RNA extraction can be performed using commonly available reagents (inexpensive alternative) or commercial kits. Some in-house protocols combine the lysis power of TRIzol with spin columns, whereas several commercial kits can be found in

the market to allow scientists to choose the one that best fits their needs. Nowadays, automated RNA extractions are also tempting as an alternative to manual methods [24]. However, although the automated extraction procedure allows standardization of sample processing and promises to reduce or cancel contamination, the manual extraction methods still guarantee higher quality and quantity of nucleic acids and remain less expensive than automatic extraction methods especially for laboratories with medium/low processing capacity. All extraction protocols (both in-house and commercial kits) include three steps: (1) solubilization of sample using detergent and chaotropic agents, (2) tissue/cell disruption, and (3) recovery of RNA from the lysate with organic or solid-phase extraction. Since RNAs come in a wide range of sizes, it is necessary to define which is the target population to be investigated. Poly(A +) RNA enrichment procedure is commonly used to isolate mRNA and long noncoding RNA (lncRNA) even if this type of selection may result in 3′ end bias. By contrast, the removal of ribosomal RNAs (rRNA) from total RNA by negative selection (i.e., rRNA depletion using rRNA-specific probes or exonucleases) is more complex and expensive, but it provides a "near-complete" transcriptome [25]. Once RNA molecules have been extracted, their quantity and quality can be determined through two metrics: *steady-state* RNA and RNA integrity number (RIN), respectively. The former refers to RNA concentration which is linearly dependent on RNA synthesis and degradation [26], although differences in mRNA levels are usually inferred to arise from changes in synthesis. It is calculated as follows:

$$\frac{\mathrm{d}R}{\mathrm{d}t} = tx_{\mathrm{j}}[\mathrm{DNA}] - d_{\mathrm{j}}[\mathrm{RNA_j}]$$

where ($\mathrm{RNA_j}$) is the RNA concentration for gene j, (DNA) is the constant ((DNA) = 1), tx_{j} is the transcription rate, and d_{j} is the degradation rate of gene j.

The Quant-iT™ RiboGreen® RNA assay kit (Thermo Scientific, Wilmington, DE, USA) is commonly used for the quantitation of RNA in solution.

Conversely, RIN evaluates the degree of degradation of RNA molecules and it is usually recorded with the Bioanalyzer (Agilent Technologies, CA, USA), a micro-capillary-based electrophoretic cell that allows separation of RNA samples according to their molecular weight and the subsequent detection via laser-induced fluorescence [27]. The amount of measured fluorescence correlates with the amount of RNA of a given size. The main advantage of this system over traditional gel electrophoresis is the tiny amounts of RNA samples required as input (~1 μl).

Once RNA molecules have been extracted and isolated, it is also important to verify the absence of genomic DNA (gDNA). Indeed, some protocols can carry over gDNA into RNA samples

leading to a counting bias (for example false-positive signals) in the downstream analysis [28]. For this purpose, scientists can use lithium chloride (LiCl), which acts selectively on RNA, leaving DNA in the solution. The protocol is extremely simple, and it is based on adding 1 volume of 5 M LiCl solution to the resuspended RNA. Following chilling at −20 °C for 30–60 min and spinning at 16,000 × *g* for 30 min at 4 °C, the supernatant (DNA) can be discarded, whereas the pellet (purified RNA) can be washed in 70% EtOH and resuspended at user's convenience.

2.3 Library Preparation and RNA-Sequencing

After RNA isolation and purification, the next step is the selection of the sequencing platform [29] and the production of RNA-seq libraries. At present, Illumina is the technology of choice for RNA-seq experiments. RNA-seq library preparation may vary depending on the target RNA population, on the NGS platform, and finally on a series of users' preferences. As a general rule, RNA with a concentration ranging from 100 ng to 4 mg and a RIN value [27] of at least 8 is required for RNA-seq library preparation. Then, RNAs are converted into double-strand cDNAs, which are subjected to shearing (e.g., nebulization or sonication) followed by adapter ligation and PCR amplification.

One of the first aspects to be considered is sequencing depth (i.e., number of reads *per* sample), which depends on the objective of the RNA-seq study. If the final goal of the RNA-seq experiment is to get a snapshot of highly expressed genes, 5–25 million reads are enough. If, instead, the aim is to have a global view of gene expression, 30–60 million reads *per* sample are required. The latter option is generally the most used. A higher number of reads (100–200 million) is necessary for an in-depth view of the transcriptome. Indeed, the greater the sequencing depth, the greater the chance of capturing mRNAs with low expression levels.

The interpretation of the concept of "coverage" for RNA-seq data is puzzling: the size of the transcriptome under investigation is unknown as the transcriptome is dynamic, complex, and tissue/cell dependent. To calculate average coverage, users should divide the total number of reads by the total size of the transcriptome. As the latter is unknown, it could be estimated by clustering the reads, but this will have huge opportunities for errors.

Users can choose between two options on the basis of the desired sequencing depth: (1) pooling multiple samples into the same flow cell lane and (2) sequencing samples across several lanes of the flow cell. The first option is generally tagged as multiplexing or barcoding and ensures a reduction in sequencing costs. Barcodes are short DNA stretches that are attached to mRNA fragments and are used to discriminate each sample from each other in a lane.

A key parameter to be determined is the length of reads (i.e., the number of nucleotides sequenced), as it may affect the mapping procedure onto a reference genome/transcriptome. The shorter

the read, the more likely is that it can be mapped back in multiple positions along with the reference. This must be clearly avoided if errors in the transcript/gene expression level estimates are to be minimized. As for Illumina platforms, read length is fixed and depends on the instrument and chemistry used.

A further option is the choice between non-strand-specific and strand-specific protocols [30]. By applying non-strand-specific (i.e., standard) protocols it is not possible to discriminate between reads originated from the sense or the antisense strand. Ideally, users would like to distinguish reads coming from RNAs of the sense strand from those coming from RNAs of the antisense strand, as this allows ambiguous data (e.g., chimeras due to overlapping transcripts) to be resolved and the differential expression profiling to be more accurate and reliable.

Users have also two options to set up a sequencing run: single-end or paired-end. Single-read sequencing involves sequencing of only one end of a cDNA fragment; paired-end sequencing involves sequencing of both ends of a cDNA fragment tagged as R1 and R2. The sequence between R1 and R2 remains unknown, but its size (i.e., mate inner distance or gap size) is known because it was decided during the preparation phase of the sequencing library (Fig. 3). The insert size is the length of the stretch of sequence between the paired-end adapters (Fig. 3). The choice of the right insert size is an important part of planning the sequencing experiment as it affects fragment size that is given by the insert size plus the length of both adapters (Fig. 3). Fragment size selection is typically done after fragmentation of the input DNA and adapter ligation, using gel electrophoresis or beads.

If the final goal of the RNA-seq experiment is just gene expression profiling in species with gold standard (i.e., well-annotated) genomes, single-end sequencing could be the right choice. Indeed, the use of paired-end sequencing will simply double the costs of the experiment. On the other hand, the use of paired-end sequencing

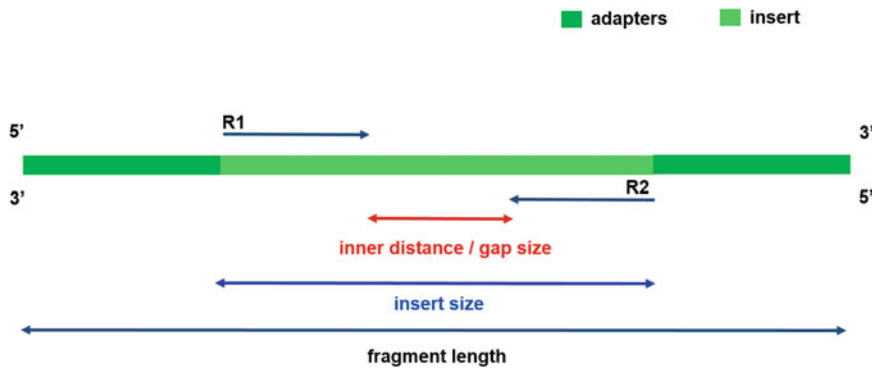


Fig. 3 Schematic representation of a paired-end sequencing. Gap size, insert size, and fragment length are indicated

improves the accuracy of expression estimates in genes with multiple isoforms (i.e., transcripts) as it allows to solve highly complex regions made up of repetitive sequences and it facilitates the read mapping step. Last but not least, a lot of methods and algorithms are being developed with paired-end reads in mind.

2.4 Read Preprocessing

Results of NGS sequencing runs are normally delivered as FASTQ-formatted files [31]. They are text files that include the sequences of the reads together with *per* base quality (*Q*) scores encoded by ASCII characters [31]. *Q* score *per* base is translated in the probability (i.e., score in Phred log scale) of a base being incorrectly called; it ranges between −5 and 41 and depends on the sequencing technology and the base caller used [31].

Once the user gets the FASTQ files, the first thing to do is to perform quality control (QC) checks on raw sequence data. QC is essential to (1) verify whether library construction and sequencing were correctly carried out, (2) exclude any possible biases, and (3) make sure that the sequences are suitable for downstream analyses.

Several tools have been developed for quality assessment of raw reads [32–34], with the most popular being FastQC (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/). FASTQScreen [35] can be optionally run to check if sequencing includes reads of the expected origin or possibly contaminating sequences. It implies the alignment of sequenced reads onto one or more reference genomes (nuclear and/or cytoplasmic), thus allowing to determine from where reads originate and to filter out only those reads that match the genome of interest.

The following step is critical as it can impair downstream analysis if it is not performed properly. It implies removing technical sequences (i.e., adapter sequences and Illumina primers), discarding low-quality reads (i.e., filter reads with high *Q* scores), trimming off poor-quality bases at the 5′ and/or 3′ of reads, and rejecting reads below a specified length. A very important information the user must be aware of is which technical sequences have been used for sample preparation, as Illumina has used different sets of technical sequences throughout the years. In case the user is unable to trace this information, it can always be derived by looking at overrepresented sequences (usually returned by QC tools). An additional aspect that needs to be stressed is that there are no general rules for setting the filtering thresholds, which are fixed based on user's experience and overall quality of the sequencing run.

Several tools have been developed, each with its own pros and cons, and are actively used to perform this task (http://hannonlab.cshl.edu/fastx_toolkit/download.html) [36].

As a general rule, QC is repeated even after having preprocessed the reads in order to verify the successful increase in the overall quality of the dataset.

2.5 Transcriptome Assembly

Typically, the next step after read preprocessing is the alignment of high-quality reads to a reference genome (*see* Subheading 2.8). In case a reference genome is missing, it becomes necessary to generate a robust reference transcriptome assembly. The goal of transcriptome assembly is to reconstruct transcription units from which read tags could have originated. Transcriptome reconstruction is challenging both for the presence of repetitive regions and, especially in eukaryotes, because most gene loci generate alternative mRNA isoforms due to alternative transcription and alternative splicing events [37, 38]. Thus, it becomes demanding to determine which isoform produced each read.

Two different strategies can be used for transcriptome reconstruction [39], namely genome-guided and genome-independent (i.e., de novo). Genome-guided methods rely on the mapping of all reads onto a reference genome, before proceeding to assemble overlapping reads at each locus into transcripts.

Genome-guided approaches are computationally feasible, have a high sensitivity, and are preferable if the final goal is to contribute to the annotation of a genome. Indeed, they can easily capture known information but also novel variations, thus helping to expand the catalog of expressed mRNAs.

Genome-independent methods directly assemble reads into potential transcripts without using a reference genome. Several algorithms have been developed in order to reconstruct de novo a transcriptome from read tags, each of which uses a different approach and has its own limitations and advantages. A partial compendium of de novo assembly software used to build transcriptome can be accessed at [38].

Most of them are based on building a de Bruijn graph from the RNA-seq reads [40]. One of the key parameters in de Bruijn graph assemblers is the *K*-mer length [40]. The term *K*-mer usually refers to a stretch of nucleic acid sequences of length *K*.

Some algorithms use a single *K*-mer (SK) value (e.g., Trinity [41]); others run a set of *K*-mer values and then merge all the assemblies generated from multiple *K*-mer (MK) to obtain a final nonredundant transcriptome (e.g., Velvet/Oases and Trans-AbySS) [42, 43].

It was demonstrated that the MK approach achieves better assembly results [38, 40]. Therefore, in case user's preference falls onto a SK assembler, it would be advisable to run different *K*-mer values and then combine the results.

All de novo assemblers output thousands of transcripts; as a consequence, a commonly accepted practice is to cluster highly similar transcripts and retain a single representative sequence *per*

cluster. This is usually performed by running CD-HIT [44]. A different approach to cut down the number of assembled transcripts is to make use of abundance estimates in a genome-free manner [45, 46]. This means fixing a minimum expression threshold and excluding transcripts that have little read support.

2.6 Evaluating the Accuracy and Completeness of Transcriptome Assemblies

Once the user has obtained a transcriptome assembly, he/she would like to know how accurate and complete it is. Below is a list of the different strategies that can be used, alone or preferably in combination, to characterize the overall quality of a transcriptome assembly.

1. Align back all the RNA-seq reads to the assembly. Generally, at least ~80% of input RNA-seq reads should map onto the reconstructed transcriptome so that it can be considered of good quality.
2. Compare the assembled transcripts against a database of protein sequences (e.g., UniProtKB [47]) to check the number of transcripts that appear to be full length or nearly full length.
3. Run BUSCO, a tool developed to quantitatively assess assembly completeness in terms of gene content [48]. It is based on conserved ortholog datasets for six major phylogenetic clades and provides intuitive metrics to describe transcriptome completeness.
4. Run DETONATE to compute scores that measure the overall quality of transcriptome assemblies [49]. DETONATE includes two packages, namely RSEM-EVAL and REF-EVAL. The former is a reference-free evaluation method based on a probabilistic model that relies only on the RNA-seq reads used for the construction of the assembly and on the assembly itself. The latter additionally requires a reference transcript set and provides a number of reference-based measures.
5. Execute TransRate, a reference-free tool for the evaluation of the quality of de novo transcriptome assemblies that uses only sequenced reads and the assembly as input [50]. TransRate assessment is based on two different statistics: the TransRate contig score, which provides a quantitative measure of the accuracy of assembly for each transcript, and the TransRate assembly score, which provides a measure of the completeness of the assembly.
6. Run rnaQUAST, a tool that computes various metrics for evaluating transcriptome assembly completeness and correctness using reference genome and gene database [51].

2.7 Transcriptome Annotation

Once the transcriptome has been assembled, it is generally annotated using similarity-based searches against different databases (referred to as filtering databases). For plant species, BLASTx searches against the UniProtKB/SwissProt database [52] and the

Arabidopsis thaliana protein complement [53] are commonly carried out. Often, BLASTn searches against Rfam [54] and/or available transcriptomes of related species complete this analysis step [55]. Annotation can be refined using several tools, some publicly available and others proprietary, that allows transcript functions to be described in a standard and controlled vocabulary. Gene Ontology (GO) terms [56], Enzyme Commission (EC) numbers [57], InterPro protein signatures [58], and KEGG pathway assignments can be derived using the proprietary Blast2GO suite [59]. Information about patterns of domains/motifs within sequences, as well as GO and KEGG pathway annotations, can be retrieved also by running InterProScan [60]. To make the InterProScan run faster, it is good practice to extract protein-coding regions (i.e., open reading frame prediction) from the reference assembly using TransDecoder [61] or ESTScan [62] and use them as query sequences.

Finally, the KEGG Automatic Annotation Server (KAAS) can be used to derive KO (KEGG Orthology) assignments that are automatically associated with KEGG pathways [63].

2.8 Read Mapping and Summarization

Typically, the most time-consuming step in the analysis of RNA-seq data is the alignment of high-quality reads for all replicates onto the reference genome or a known transcriptome [64].

Evidently, the more accurate the mapping onto a reference sequence, the more reliable the results of downstream steps will be (e.g., quantification of gene expression).

A big array of tools (i.e., read aligners) have been developed for the alignment of short reads to a reference sequence with different levels of accuracy and speed [64, 65]. Two types of read aligners can be distinguished: unspliced and spliced [66]. The most popular unspliced aligners are BWA [67] and Bowtie2 [68], while the most commonly used spliced aligners are TopHat2 [69, 70], STAR [71], and HISAT2 [72].

Indeed, the algorithms developed for the mapping of reads from RNA-seq must take into account that genes in eukaryotic genomes include introns, while reads from mature mRNAs do not. In addition, introns are of variable length; therefore algorithms must be able to handle spliced alignment with gaps ranging in size from a few dozen bases to thousands of bases.

A key issue in the quantification of gene expression is the handling of reads that map equally well to multiple locations (i.e., "multimaps") of the reference sequence. Aligners handle "multimaps" differently [66]: some opt for a conservative approach, thus discarding reads that mapped to multiple locations; others allocate "multimaps" randomly or on the basis of an estimate of local coverage; still others allocate each multi-mapped read to all of the positions it maps to (e.g., a read mapping to five different locations will count as 20% of a read at each position).

Users must keep in mind that when a known transcriptome is used as reference, the number of multi-mapped reads increases considerably because of reads that derive from exons shared by various isoforms of the same gene. Using longer reads or paired-end sequencing allows mitigating the ambiguity of multimaps to some extent.

Alignment information of short reads mapped against reference sequences is generally saved into the Sequence Alignment/Map (SAM) file format [73]. SAMtools is the software package that provides utilities for processing read alignments in the SAM format [73].

Once locations for as many reads as possible have been obtained, the next step is to summarize and aggregate reads over some biologically meaningful units, such as exons, transcripts, or genes. This process is referred to as summarization [74].

The most common approach is to exploit previously annotated features and it implies the availability of an annotated reference genome. The simplest strategy is to count reads that map along the whole length of the gene, introns included. In this way reads from unannotated exons flow into the count matrix. This strategy is also referred to as the "exon union model" and involves counting all reads that touch any exon (from all mRNA isoforms) within a gene. This provides a global summary of the expression of a particular gene although it does not allow to estimate the abundance of each isoform. The exon intersection model, instead, uses only exons common to all isoforms of a particular gene. This measure is more stable when alternative transcription characterizes a particular gene but does not aggregate all possible reads, thus reducing the power for differential expression analysis.

Several tools have been developed to aggregate reads over biological units and generate a count matrix, with the most popular being featureCounts [75] and HTSeq-count [76]. Some software packages (e.g., RSEM [77] and Salmon [46]) do not rely on the availability of a reference genome and are particularly useful when a de novo transcriptome assembly is used as a reference for quantification. While the combination of spliced aligners with featureCounts/HTSeq-count is very common, tools developed for the estimation of transcript-level abundance (e.g., Salmon) are increasingly used as they generally outperform the former [78]. However, at present, complex transcriptomes continue to be studied at the gene, rather than transcript level.

It is clear that the final matrix of read counts will depend on the summarization strategy used, as different sets of reads can be included or excluded in the table of counts depending on the biological unit chosen to aggregate reads and on the approach used (e.g., exon union or intersection model).

2.9 Normalization and Detection of Differentially Expressed Genes

To determine if gene X is differentially expressed, we would like to know whether the number of reads aligning to gene X tends to be different between experimental conditions. To do this, read counts in the count matrix must first be normalized and then used as input to perform some statistical tests between samples of interest.

Even before, however, it is necessary to explore relationships among sample replicates. Principal component analysis (PCA) is generally carried out to evaluate the homogeneity of the samples. A viable alternative is to calculate Pearson's correlation coefficient (possibly producing a replicate Pearson's correlation heatmap) between biological replicates ($r = 0.9$ is a widely accepted cutoff). In case of replicates that are clear outliers, users might consider removing them from the study as they could be a source of confounding effects.

Genes with a very low level of expression (i.e., low-abundance genes) across all libraries must be filtered out as they are considered not reliable for statistical inference [79]. The identification and filtering of these low-abundance genes may improve detection sensitivity of differentially expressed genes (DEGs) and undoubtedly facilitate the computational work without major loss of information. One of the most commonly used methods (implemented in the edgeR package [80]) is to filter genes with a counts-per-million (CPM) value less than a fixed threshold (e.g., 0.5 or 1) in at least Z samples, where Z represents a subset of all samples, including biological replicates. A data-driven technique, which is not based on selecting an arbitrary threshold value, has been proposed and successfully used to filter low-abundance genes [81].

Read counts need to be properly normalized to extract reliable expression estimates and the choice of normalization has a great influence on the statistical analysis for the call of differentially expressed genes [79].

Normalization is an important issue in mRNA-seq data analysis as it allows removing sources of variability in the data and enables more accurate comparison of expression levels within and between samples.

Normalization methods allow for either inter-sample or intra-sample comparison. Within-sample (i.e., intra-sample) normalization allows quantification of expression levels of each gene relative to all others in the sample. Between-sample (i.e., inter-sample) normalization is crucial for comparing each other read counts from different libraries.

RPKM (*r*eads *per k*ilobase *per m*illion mapped reads) [18] and FPKM (*f*ragments *per k*ilobase of exon model *per m*illion mapped fragments) [82] are the most commonly used metrics (or expression units) that attempt to normalize for sequencing depth and gene length. RPKM was established for single-end RNA-seq, while FPKM was conceived for paired-end RNA-seq. Indeed, paired-end sequencing produces two reads *per* mRNA

fragment, but both reads are not always mappable. Therefore, it was decided to count fragments instead of reads in order to derive expression values. A closely related alternative metric is TPM (*t*ranscripts *per m*illion) [83], introduced to correct inconsistencies while comparing independent samples.

The two most commonly used methods for inter-sample normalization are trimmed mean of *M*-values (TMM) [84] and relative log expression (RLE) [85]. Both showed good performance when compared with other normalization methods (e.g., total count, upper quartile, median, quantile [86]). Noteworthy, they do not correct read counts for gene length, which is irrelevant for inter-sample comparisons [87]. We agree with this interpretation and suggest using one of these methods for normalization.

Typically, box plots of the distribution of read counts before and after normalization are formulated.

Several methods, based on different statistical models, have been developed to find genes that are differentially expressed between conditions [88]. Historically we have gone from statistical models based on Poisson distribution, through negative binomial distribution models, to generalized linear models [89].

At present, edgeR [80] and DESeq2 [85] are the most used tools, both based on generalized linear models. NOISeq is instead particularly useful when high variability is observed across biological replicates [90]. Finally, several tools for RNA-seq time course data have been developed ad hoc and their performances compared in [91].

We suggest a conservative approach to DEG calling that relies on the use of at least two methods (e.g., DESeq2 and edgeR). The list of DEGs independently called by each method is, then, filtered based on fold change and FDR (false discovery rate) and finally a single gene list is obtained from the intersection of the previous ones so that only DEGs called by all methods will be used downstream [92].

2.10 Extracting Biological Insights from a Gene List

Generating lists of differentially expressed genes is not the final step of the analysis.

Enrichment analysis of GO terms, included in the "molecular function" and "biological process" domains, is normally performed to extract biological insight from the gene list. Several command-line and Web-based tools that perform GO enrichment are currently available [93–96]. Usually, these tools require a target set and a reference set of genes as input and seek enrichment by comparing the target set with the reference set. As an alternative, investigators can use the DAVID bioinformatic resource to extract biological meaning from large gene lists [97].

The user-driven tool MapMan [98] could be used to map the identified DEGs and their expression values onto MapMan bins (i.e., diagrams of metabolic pathways or other processes) for data visualization and pathway analysis [92].

A very useful technique in analyzing gene expression data is clustering analysis followed by heatmap generation. Both hierarchical and K-means clustering (i.e., the two most popular types of clustering) allow the identification of patterns in the data. In other words, the clustering of gene expression data supports identifying groups of genes that behave similarly both because they have similar functions (i.e., co-functional genes) and because they are under the same transcriptional control (i.e., co-regulated genes). A plethora of clustering tools for gene expression data are available in R [99]. Recently, a very user-friendly cloud platform, named WebMeV, for analyzing and visualizing expression data has been developed [100].

A further method to extract information from a list of DEGs is to perform gene co-expression network analysis in order to cluster sets of coordinately expressed genes into different modules. Each module or group includes genes that are likely to be functionally associated as their expression levels correlate strongly. To the best of our knowledge, weighted gene co-expression network analysis (WGCNA) is the most widely used method [101].

Finally, genome browsers allow investigators to interactively explore large RNA-seq datasets. Indeed, coverage tracks from RNA-seq data, quantitative data (i.e., gene expression values), and raw read tracks can be viewed using genome browsers [102], among all of which the stand-alone *i*ntegrative *g*enomics *v*iewer (IGV) is arguably the most widely used [103].

2.11 RNA-Seq Validation by Quantitative RT-PCR

Typically, qRT-PCR experiments on several key transcripts are performed to confirm and corroborate digital gene expression profiles derived from RNA-seq data [104]. As qRT-PCR is an accurate and sensitive but low-throughput method, a varying number of genes (generally a dozen) are randomly extracted from the list of DEGs (up- or down-regulated) to quantify their expression levels. Alternatively, target genes can be the ten top-ranked genes in the list or can be selected based on the user's biological knowledge.

The selection of the best reference gene, characterized by having a stable expression (i.e., unchanged expression pattern across tissues, developmental stages), is crucial for this type of analysis, as it is used as internal reaction control for the normalization of mRNA levels between samples [105]. However, at present, normalization with multiple reference genes is becoming the standard because it generates more reliable results [106]. Quantitative RT-PCR results (ΔCt values) are compared with normalized digital gene expression profiles (e.g., FPKM) and qRT-PCR and RNA-seq expression values of each target gene are correlated after log transformation [104].

2.12 Submission to Public Repositories

The final step is to make raw and processed data available to the research community. This is not negligible as data sharing facilitates repeatability and novel discoveries. Submission to the Sequence Read Archive (SRA)/European Nucleotide Archive (ENA)/DDBJ Sequence Read Archive (DRA) of raw reads is highly recommended. Likewise, in silico-assembled transcripts from primary data should be submitted to the Transcriptome Shotgun Assembly Sequence Database (TSA). It is useful to remark that reads used in the assembly procedure must have been experimentally determined by the same submitter. Finally, processed data, such as raw counts of sequencing reads for the features of interest and/or normalized abundance measurements (e.g., from edgeR), can be submitted to the Gene Expression Omnibus (GEO) repository.

3 New Frontiers in Transcriptome Analysis

3.1 Single-Cell RNA-Seq

In general, RNA-seq of bulk tissues derives gene expression changes that are a signal of the average expression of multiple cell types.

At present, single-cell biology is a hot topic and single-cell RNA sequencing (scRNA-seq) is increasingly spreading to explore gene expression dynamics at single-cell resolution [107]. Since its introduction in 2009 (Fig. 1) [108], this technique is generating new knowledge on the mechanisms underlying cell development, cell heterogeneity, and cell response to stimuli [109, 110] and it is disclosing cell-to-cell gene expression variability. A second revolution in transcriptomics [111] is happening and it is accompanied by new challenges in data analysis and management.

Different scRNA-seq protocols have been developed in the past few years, which differ in some critical aspects of single-cell isolation and RNA sequencing [112]. Single-cell isolation is the first step for obtaining transcriptome information from an individual cell and different approaches have been proposed [109, 113], with high-throughput microfluidics-based methods (e.g., Drop-Seq) being those developed more recently [114, 115].

The general workflow for the generation of scRNA-seq libraries and the analysis of scRNA-seq data is basically identical to the one we have just described for bulk mRNA-seq data; however, some steps require the use of tools developed on purpose.

Indeed, methods and tools developed for preprocessing, transcriptome assembly and annotation, read mapping, and summarization can be directly applied to scRNA-seq data [116, 117]. Compared to bulk mRNA-seq, scRNA-seq generates nosier data; therefore, data normalization represents the critical step of the entire workflow. To this end, several normalization methods have been developed ad hoc for scRNA-seq data [118, 119].

However, because the aims of bulk RNA-seq and scRNA-seq are different and since clustering [120] and trajectory analysis [121] are distinguishing features and critical steps for the analysis of scRNA-seq data, the scientific community needed to develop a dedicated set of tools. At the following link https://github.com/seandavi/awesome-single-cell the reader can find a community-curated list of software packages covering every single step of the analysis process and a partial compendium of bioinformatic tools and methods for scRNA-seq has been published by [122]. But at the moment, there is still no consensus on which methods/tools work best for each step of the analysis.

3.2 Isoform Sequencing Using Pacific Biosciences Technology (Iso-Seq)

It is well documented in the literature that in plant species the majority of genes are alternatively spliced and produce multiple transcript isoforms [123].

In the previous paragraph, we have seen how challenging is the assembly of the different isoforms of a gene from short reads [124].

The Iso-Seq method was introduced in 2012 (Fig. 1) and allows obtaining full-length transcripts using single-molecule real-time (SMRT) sequencing [125].

This method is very powerful as it enables (1) to reconstruct a transcriptome without assembling reads and (2) to resolve all possible isoforms of a gene, thus generating a really accurate snapshot of the transcriptome. It has been widely used to investigate transcriptomes across a variety of important crops [126].

The PacBio SMRT analysis module is a suite of applications developed to handle PacBio long-read sequencing data. It is generally used also to process raw Iso-Seq long reads until generating high-quality full-length isoforms [127]. With the spread of the technology, the scientific community began to develop additional tools such as those developed for error corrections [128] and for the alignment of long reads onto reference sequences (https://github.com/Magdoll/cDNA_Cupcake/wiki/Cupcake:-supporting-scripts-for-Iso-Seq-after-clustering-step). Indeed, it would be advisable to always combine Iso-Seq long reads with short reads (e.g., from Illumina platforms), as the latter are used to correct error-prone long reads [129]. A growing list of tools is accessible at the Iso-Seq™ wiki Web page (https://github.com/PacificBiosciences/IsoSeq_SA3nUP/wiki) to support the growth of the Iso-Seq scientific community.

Given the promising results obtained by Iso-Seq, it is very likely to become the chosen approach for the reconstruction and investigation of the full-length transcriptome atlas of a given species.

4 Conclusion

From its introduction onward, RNA-seq contributed importantly to our understanding of transcriptomes. However, contrary to what was initially thought, RNA-seq proved more challenging than expected, as it is characterized by technical artifacts and biases and has not addressed most of the critical issues associated with the statistical analysis of gene expression data. The last years have seen consensus emerge on the best practices for designing and completing a RNA-seq study correctly [130, 131]. The content of this chapter fits exactly in this context and remarks the large number of choices available (especially for bioinformatic data analysis) in front of which the scientist is found.

References

1. Piétu G, Mariage-Samson R, Fayein N-A et al (1999) The genexpress IMAGE knowledge base of the human brain transcriptome: a prototype integrated resource for functional and computational genomics. Genome Res 9:195–209. https://doi.org/10.1101/gr.9.2.195
2. Velculescu VE, Zhang L, Zhou W et al (1997) Characterization of the yeast transcriptome. Cell 88:243–251. https://doi.org/10.1016/S0092-8674(00)81845-0
3. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. Nat Rev Genet 10:57–63. https://doi.org/10.1038/nrg2484
4. Adams MD, Kelley JM, Gocayne JD et al (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252:1651–1656. https://doi.org/10.1126/science.2047873
5. D'Agostino N, Traini A, Frusciante L, Chiusano ML (2009) SolEST database: a "one-stop shop" approach to the study of Solanaceae transcriptomes. BMC Plant Biol 9:142. https://doi.org/10.1186/1471-2229-9-142
6. D'Agostino N, Aversano M, Frusciante L, Chiusano ML (2006) TomatEST database: in silico exploitation of EST data to explore expression patterns in tomato species. Nucleic Acids Res 35:D901–D905. https://doi.org/10.1093/nar/gkl921
7. Parkinson J, Blaxter M (2009) Expressed sequence tags: an overview. Methods Mol Biol 533:1–12. https://doi.org/10.1007/978-1-60327-136-3_1
8. Boguski MS, Lowe TMJ, Tolstoshev CM (1993) dbEST—database for "expressed sequence tags". Nat Genet 4:332–333. https://doi.org/10.1038/ng0893-332
9. Lee Y, Tsai J, Sunkara S et al (2005) The TIGR Gene Indices: clustering and assembling EST and know genes and integration with eukaryotic genomes. Nucleic Acids Res 33:D71–D74. https://doi.org/10.1093/nar/gki064
10. Duvick J, Fu A, Muppirala U et al (2008) PlantGDB: a resource for comparative plant genomics. Nucleic Acids Res 36:959–965. https://doi.org/10.1093/nar/gkm1041
11. Higuchi R, Dollinger G, Walsh PS, Griffith R (1992) Simultaneous amplification and detection of specific DNA sequences. Biotechnology 10:413–417. https://doi.org/10.1038/nbt0492-413
12. Schena M, Shalon D, Davis RW, Brown PO (1995) Quantitative monitoring of gene expression patterns with a complementary DNA microarray. Science 270:467–470. https://doi.org/10.1126/science.270.5235.467
13. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW (1995) Serial analysis of gene expression. Science 270:484–487. https://doi.org/10.1126/science.270.5235.484
14. Matsumura H, Reich S, Ito A et al (2003) Gene expression analysis of plant host-pathogen interactions by SuperSAGE. Proc Natl Acad Sci U S A 100:15718–15723. https://doi.org/10.1073/pnas.2536670100

15. Wei CL, Ng P, Chiu KP et al (2004) 5′ Long serial analysis of gene expression (LongSAGE) and 3′ LongSAGE for transcriptome characterization and genome annotation. Proc Natl Acad Sci U S A 101:11701–11706. https://doi.org/10.1073/pnas.0403514101
16. Brenner S, Johnson M, Bridgham J et al (2000) Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays. Nat Biotechnol 18:630–634. https://doi.org/10.1038/76469
17. Shiraki T, Kondo S, Katayama S et al (2003) Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. Proc Natl Acad Sci U S A 100:15776–15781. https://doi.org/10.1073/pnas.2136655100
18. Mortazavi A, Williams BA, McCue K et al (2008) Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods 5:621–628. https://doi.org/10.1038/nmeth.1226
19. Stark R, Grzelak M, Hadfield J (2019) RNA sequencing: the teenage years. Nat Rev Genet 20:631–656. https://doi.org/10.1038/s41576-019-0150-2
20. Han Y, Gao S, Muegge K et al (2015) Advanced applications of RNA sequencing and challenges. Bioinform Biol Insights 9:29–46. https://doi.org/10.4137/BBI.S28991
21. Auer PL, Doerge RW (2010) Statistical design and analysis of RNA sequencing data. Genetics 185:405–416. https://doi.org/10.1534/genetics.110.114983
22. Hansen KD, Wu Z, Irizarry RA, Leek JT (2011) Sequencing technology does not eliminate biological variability. Nat Biotechnol 29:572–573. https://doi.org/10.1038/nbt.1910
23. Farrell RE (1993) Chapter 5 - RNA isolation strategies. In: Farrell RE (ed) RNA methodologies. Academic, Boston, pp 46–92. https://doi.org/10.1016/B978-0-12-374727-3.00005-X
24. Knepp JH, Geahr MA, Forman MS, Valsamakis A (2003) Comparison of automated and manual nucleic acid extraction methods for detection of enterovirus RNA. J Clin Microbiol 41:3532–3536. https://doi.org/10.1128/jcm.41.8.3532-3536.2003
25. Hrdlicková R, Toloue M, Tian B (2016) RNA-Seq methods for transcriptome analysis: RNA-Seq. Wiley Interdiscip Rev RNA 8. https://doi.org/10.1002/wrna.1364
26. Tippmann SC, Ivanek R, Gaidatzis D et al (2012) Chromatin measurements reveal contributions of synthesis and decay to steady-state mRNA levels. Mol Syst Biol 8:593. https://doi.org/10.1038/msb.2012.23
27. Schroeder A, Mueller O, Stocker S et al (2006) The RIN: an RNA integrity number for assigning integrity values to RNA measurements. BMC Mol Biol 7:3. https://doi.org/10.1186/1471-2199-7-3
28. Pereira MA (2017) RNA-seq: applications and best practices, Ch. 1. In: Imada EL (ed) Applications of RNA-Seq and omics strategies - from microorganisms to human health. IntechOpen, Rijeka
29. Chu Y, Corey DR (2012) RNA sequencing: platform selection, experimental design, and data interpretation. Nucleic Acid Ther 22:271–274. https://doi.org/10.1089/nat.2012.0367
30. Levin JZ, Yassour M, Adiconis X et al (2010) Comprehensive comparative analysis of strand-specific RNA sequencing methods. Nat Methods 7:709–715. https://doi.org/10.1038/nmeth.1491
31. Cock PJA, Fields CJ, Goto N et al (2009) The Sanger FASTQ file format for sequences with quality scores, and the Solexa/Illumina FASTQ variants. Nucleic Acids Res 38:1767–1771. https://doi.org/10.1093/nar/gkp1137
32. Cox MP, Peterson DA, Biggs PJ (2010) SolexaQA: at-a-glance quality assessment of Illumina second-generation sequencing data. BMC Bioinformatics 11. https://doi.org/10.1186/1471-2105-11-485
33. Wang L, Wang S, Li W (2012) RSeQC: quality control of RNA-seq experiments. Bioinformatics 28:2184–2185. https://doi.org/10.1093/bioinformatics/bts356
34. Pérez-Rubio P, Lottaz C, Engelmann JC (2019) FastqPuri: high-performance preprocessing of RNA-seq data. BMC Bioinformatics 20:1–11. https://doi.org/10.1186/s12859-019-2799-0
35. Wingett SW, Andrews S (2018) FastQ Screen: a tool for multi-genome mapping and quality control. F1000Research 7:1338. https://doi.org/10.12688/f1000research.15931.2
36. Ballenghien M, Faivre N, Galtier N (2017) Patterns of cross-contamination in a multispecies population genomic project: detection, quantification, impact, and solutions. BMC Biol 15:1–16. https://doi.org/10.1186/s12915-017-0366-6
37. Garber M, Grabherr MG, Guttman M, Trapnell C (2011) Computational methods for

transcriptome annotation and quantification using RNA-seq. Nat Methods 8:469–477. https://doi.org/10.1038/nmeth.1613

38. Hölzer M, Marz M (2019) De novo transcriptome assembly: a comprehensive cross-species comparison of short-read RNA-Seq assemblers. Gigascience 8:1–16. https://doi.org/10.1093/gigascience/giz039
39. Haas BJ, Zody MC (2010) Advancing RNA-Seq analysis. Nat Biotechnol 28:421–423. https://doi.org/10.1038/nbt0510-421
40. Durai DA, Schulz MH (2016) Informed kmer selection for de novo transcriptome assembly. Bioinformatics 32:1670–1677. https://doi.org/10.1093/bioinformatics/btw217
41. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson DA, Amit I, Adiconis X, Fan L, Raychowdhury R, Zeng Q, Chen Z, Mauceli E, Hacohen N, Gnirke A, Rhind N, di Palma F, Birren BW, Nusbaum C, Lindblad-Toh K, Friedman AR (2013) Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. Nat Biotechnol 29:644–652. https://doi.org/10.1038/nbt.1883.Trinity
42. Robertson G, Schein J, Chiu R et al (2010) De novo assembly and analysis of RNA-seq data. Nat Methods 7:909–912. https://doi.org/10.1038/nmeth.1517
43. Schulz MH, Zerbino DR, Vingron M, Birney E (2012) Oases: robust de novo RNA-seq assembly across the dynamic range of expression levels. Bioinformatics 28:1086–1092. https://doi.org/10.1093/bioinformatics/bts094
44. Fu L, Niu B, Zhu Z et al (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics 28:3150–3152. https://doi.org/10.1093/bioinformatics/bts565
45. Roberts A, Pachter L (2013) Streaming fragment assignment for real-time analysis of sequencing experiments. Nat Methods 176:139–148. https://doi.org/10.1016/j.physbeh.2017.03.040
46. Patro R, Duggal G, Love MI et al (2017) Salmon: fast and bias-aware quantification of transcript expression using dual-phase inference. Nat Methods 14:417–419. https://doi.org/10.1038/nmeth.4197.Salmon
47. The UniProt Consortium (2018) Erratum: UniProt: the universal protein knowledgebase (Nucleic acids research (2017) 45 D1 (D158–D169)). Nucleic Acids Res 46:2699. https://doi.org/10.1093/nar/gky092
48. Simão FA, Waterhouse RM, Ioannidis P et al (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics 31:3210–3212. https://doi.org/10.1093/bioinformatics/btv351
49. Li B, Fillmore N, Bai Y et al (2014) Evaluation of de novo transcriptome assemblies from RNA-Seq data. Genome Biol 15:1–21. https://doi.org/10.1186/s13059-014-0553-5
50. Smith-Unna R, Boursnell C, Patro R et al (2016) TransRate: reference-free quality assessment of de novo transcriptome assemblies. Genome Res 26:1134–1144. https://doi.org/10.1101/gr.196469.115
51. Bushmanova E, Antipov D, Lapidus A et al (2016) RnaQUAST: a quality assessment tool for de novo transcriptome assemblies. Bioinformatics 32:2210–2212. https://doi.org/10.1093/bioinformatics/btw218
52. The UniProt Consortium (2019) UniProt: a worldwide hub of protein knowledge. Nucleic Acids Res 47:D506–D515. https://doi.org/10.1093/nar/gky1049
53. Berardini TZ, Reiser L, Li D et al (2015) The Arabidopsis information resource: making and mining the "gold standard" annotated reference plant genome. Genesis 53:474–485. https://doi.org/10.1002/dvg.22877
54. Kalvari I, Argasinska J, Quinones-Olvera N et al (2018) Rfam 13.0: shifting to a genome-centric resource for non-coding RNA families. Nucleic Acids Res 46:D335–D342. https://doi.org/10.1093/nar/gkx1038
55. Vitiello A, Rao R, Corrado G et al (2018) De novo transcriptome assembly of cucurbita pepo l Leaf tissue infested by Aphis gossypii. Data 3:36. https://doi.org/10.3390/data3030036
56. The Gene Ontology Consortium (2019) The Gene Ontology Resource: 20 years and still GOing strong. Nucleic Acids Res 47:D330–D338. https://doi.org/10.1093/nar/gky1055
57. Bairoch A (2000) The ENZYME database in 2000. Nucleic Acids Res 28:304–305. https://doi.org/10.1093/nar/28.1.304
58. Mitchell AL, Attwood TK, Babbitt PC et al (2019) InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res 47:D351–D360. https://doi.org/10.1093/nar/gky1100

59. Götz S, García-Gómez JM, Terol J et al (2008) High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res 36:3420–3435. https://doi.org/10.1093/nar/gkn176
60. Jones P, Binns D, Chang HY et al (2014) InterProScan 5: genome-scale protein function classification. Bioinformatics 30:1236–1240. https://doi.org/10.1093/bioinformatics/btu031
61. Haas BJ, Papanicolaou A, Yassour M et al (2013) De novo transcript sequence reconstruction from RNA-Seq: reference generation and analysis with Trinity. Nat Protoc. https://doi.org/10.1038/nprot.2013.084
62. Iseli C, Jongeneel CV, Bucher P (1999) ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol 1999:138–148
63. Moriya Y, Itoh M, Okuda S et al (2007) KAAS: an automatic genome annotation and pathway reconstruction server. Nucleic Acids Res 35:182–185. https://doi.org/10.1093/nar/gkm321
64. Fonseca NA, Rung J, Brazma A, Marioni JC (2012) Tools for mapping high-throughput sequencing data. Bioinformatics 28:3169–3177. https://doi.org/10.1093/bioinformatics/bts605
65. Lindner R, Friedel CC (2012) A comprehensive evaluation of alignment algorithms in the context of RNA-Seq. PLoS One 7:1–10. https://doi.org/10.1371/journal.pone.0052403
66. Benjamin AM, Nichols M, Burke TW et al (2014) Comparing reference-based RNA-Seq mapping methods for non-human primate data. BMC Genomics 15:1–14. https://doi.org/10.1186/1471-2164-15-570
67. 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
68. Langmead B, Salzberg SL (2013) Bowtie2. Nat Methods 9:357–359. https://doi.org/10.1038/nmeth.1923.Fast
69. Kim D, Pertea G, Trapnell C, Harold Pimentel RK, Salzberg SL (2006) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. ACS Div Environ Chem Prepr Ext Abstr 46:957–961
70. Kim D, Pertea G, Trapnell C et al (2013) TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol 14:R36. https://doi.org/10.1186/gb-2013-14-4-r36
71. Dobin A, Davis CA, Schlesinger F et al (2013) STAR: ultrafast universal RNA-seq aligner. Bioinformatics 29:15–21. https://doi.org/10.1093/bioinformatics/bts635
72. Kim D, Langmead B, Salzberg SL (2015) HISAT: a fast spliced aligner with low memory requirements. Nat Methods 12:357–360. https://doi.org/10.1038/nmeth.3317
73. Li H, Handsaker B, Wysoker A et al (2009) The Sequence Alignment/Map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352
74. Oshlack A (2010) From RNA-seq reads to differential expression results. Genome Biol 11:220
75. Liao Y, Smyth GK, Shi W (2014) FeatureCounts: an efficient general purpose program for assigning sequence reads to genomic features. Bioinformatics 30:923–930. https://doi.org/10.1093/bioinformatics/btt656
76. Anders S, Pyl PT, Huber W (2015) HTSeq-A Python framework to work with high-throughput sequencing data. Bioinformatics 31:166–169. https://doi.org/10.1093/bioinformatics/btu638
77. Li B, Dewey CN (2011) RSEM: accurate transcript quantification from RNA-Seq data with or without a reference genome. BMC Bioinformatics 12:323. https://doi.org/10.1186/1471-2105-12-323
78. Germain PL, Vitriolo A, Adamo A et al (2016) RNAontheBENCH: computational and empirical resources for benchmarking RNAseq quantification and differential expression methods. Nucleic Acids Res 44:5054–5067. https://doi.org/10.1093/nar/gkw448
79. Bullard JH, Purdom E, Hansen KD, Dudoit S (2010) Evaluation of statistical methods for normalization and differential expression in mRNA-Seq experiments. BMC Bioinformatics 11:94. https://doi.org/10.1186/1471-2105-11-94
80. Robinson MD, McCarthy DJ, Smyth GK (2009) edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. Bioinformatics 26:139–140. https://doi.org/10.1093/bioinformatics/btp616
81. Rau A, Gallopin M, Celeux G, Jaffrézic F (2013) Data-based filtering for replicated high-throughput transcriptome sequencing experiments. Bioinformatics 29:2146–2152.

https://doi.org/10.1093/bioinformatics/btt350

82. Trapnell C, Williams BA, Pertea G et al (2011) Transcript assembly and abundance estimation from RNA-Seq reveals thousands of new transcripts and switching among isoforms. Nat Biotechnol 28:511–515. https://doi.org/10.1038/nbt.1621.Transcript
83. Wagner GP, Kin K, Lynch VJ (2012) Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. Theory Biosci 131:281–285. https://doi.org/10.1007/s12064-012-0162-3
84. Robinson MD, Oshlack A (2010) A scaling normalization method for differential expression analysis of RNA-seq data. Genome Biol 11:R25. https://doi.org/10.1186/gb-2010-11-3-r25
85. Love MI, Huber W, Anders S (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. Genome Biol 15:1–21. https://doi.org/10.1186/s13059-014-0550-8
86. Abbas-Aghababazadeh F, Li Q, Fridley BL (2018) Comparison of normalization approaches for gene expression studies completed with high-throughput sequencing. PLoS One 13:1–21. https://doi.org/10.1371/journal.pone.0206312
87. Dillies MA, Rau A, Aubert J et al (2013) A comprehensive evaluation of normalization methods for Illumina high-throughput RNA sequencing data analysis. Brief Bioinform 14:671–683. https://doi.org/10.1093/bib/bbs046
88. Rapaport F, Khanin R, Liang Y et al (2013) Comprehensive evaluation of differential gene expression analysis methods for RNA-seq data. Genome Biol 14:R95. https://doi.org/10.1186/gb-2013-14-9-r95
89. Soneson C, Delorenzi M (2013) A comparison of methods for differential expression analysis of RNA-seq data. BMC Bioinformatics 14:91. https://doi.org/10.1186/1471-2105-14-91
90. Tarazona S, Furió-Tarí P, Turrà D et al (2015) Data quality aware analysis of differential expression in RNA-seq with NOISeq R/Bioc package. Nucleic Acids Res 43:e140. https://doi.org/10.1093/nar/gkv711
91. Spies D, Renz PF, Beyer TA, Ciaudo C (2017) Comparative analysis of differential gene expression tools for RNA sequencing time course data. Brief Bioinform 20:1–11. https://doi.org/10.1093/bib/bbx115
92. Scotti R, D'Agostino N, Zaccardelli M (2019) Gene expression profiling of tomato roots interacting with Pseudomonas fluorescens unravels the molecular reprogramming that occurs during the early phases of colonization. Symbiosis 78:177–192. https://doi.org/10.1007/s13199-019-00611-9
93. Young MD, Wakefield MJ, Smyth GK, Oshlack A (2010) Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol 11:R14. https://doi.org/10.1186/gb-2010-11-2-r14
94. Tian T, Liu Y, Yan H et al (2017) AgriGO v2.0: a GO analysis toolkit for the agricultural community, 2017 update. Nucleic Acids Res 45:W122–W129. https://doi.org/10.1093/nar/gkx382
95. Alexa A, Jorg R (2019) Gene set enrichment analysis with topGO. Encycl Syst Biol, p 806. https://doi.org/10.1007/978-1-4419-9863-7_100552
96. Eden E, Navon R, Steinfeld I et al (2009) GOrilla: a tool for discovery and visualization of enriched GO terms in ranked gene lists. BMC Bioinformatics 10:1–7. https://doi.org/10.1186/1471-2105-10-48
97. Huang DW, Sherman BT, Lempicki RA (2009) Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc 4:44–57. https://doi.org/10.1038/nprot.2008.211
98. Thimm O, Bläsing O, Gibon Y et al (2004) MAPMAN: a user-driven tool to display genomics data sets onto diagrams of metabolic pathways and other biological processes. Plant J 37:914–939. https://doi.org/10.1111/j.1365-313X.2004.02016.x
99. Langfelder P, Horvath S (2012) Fast R functions for robust correlations and hierarchical clustering. J Stat Softw 46:1–17. https://doi.org/10.18637/jss.v046.i11
100. Wang YE, Kuznetsov L, Partensky A et al (2017) WebMeV: a cloud platform for analyzing and visualizing cancer genomic data. Center for Cancer Computational Biology, Dana-Farber Cancer Institute, Boston, MA, pp 1–7
101. Langfelder P, Horvath S (2008) WGCNA: an R package for weighted correlation network analysis. BMC Bioinformatics 9:559. https://doi.org/10.1186/1471-2105-9-559
102. Buels R, Yao E, Diesh CM et al (2016) JBrowse: a dynamic web platform for genome visualization and analysis. Genome Biol 17:1–12. https://doi.org/10.1186/s13059-016-0924-1

103. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. Brief Bioinform 14:178–192. https://doi.org/10.1093/bib/bbs017
104. Everaert C, Luypaert M, Maag JLV et al (2017) Benchmarking of RNA-sequencing analysis workflows using whole-transcriptome RT-qPCR expression data. Sci Rep 7:1–11. https://doi.org/10.1038/s41598-017-01617-3
105. Thellin O, ElMoualij B, Heinen E, Zorzi W (2009) A decade of improvements in quantification of gene expression and internal standard selection. Biotechnol Adv 27:323–333. https://doi.org/10.1016/j.biotechadv.2009.01.010
106. Bustin SA (2002) Quantification of mRNA using real-time reverse transcription PCR (RT-PCR): trends and problems. J Mol Endocrinol 29:23–39. https://doi.org/10.1677/jme.0.0290023
107. Shahan R (2019) The future is now: gene expression dynamics at single cell resolution. Plant Cell 31:933–934. https://doi.org/10.1105/tpc.19.00247
108. Tang W, Tang AY (2019) Biological significance of RNA-seq and single-cell genomic research in woody plants. J For Res 30:1555–1568. https://doi.org/10.1007/s11676-019-00933-w
109. Efroni I, Birnbaum KD (2016) The potential of single-cell profiling in plants. Genome Biol 17:1–8. https://doi.org/10.1186/s13059-016-0931-2
110. Jean-Baptiste K, McFaline-Figueroa JL, Alexandre CM et al (2019) Dynamics of gene expression in single root cells of A. thaliana. Plant Cell. https://doi.org/10.1105/tpc.18.00785
111. Shapiro E, Biezuner T, Linnarsson S (2013) Single-cell sequencing-based technologies will revolutionize whole-organism science. Nat Rev Genet 14:618–630. https://doi.org/10.1038/nrg3542
112. Chen G, Ning B, Shi T (2019) Single-cell RNA-seq technologies and related computational data analysis. Front Genet 10:1–13. https://doi.org/10.3389/fgene.2019.00317
113. Rich-Griffin C, Stechemesser A, Finch J et al (2020) Single-cell transcriptomics: a high-resolution avenue for plant functional genomics. Trends Plant Sci 25:186–197. https://doi.org/10.1016/j.tplants.2019.10.008
114. Macosko EZ et al (2015) Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. Cell 171:1437–1452. https://doi.org/10.4172/2157-7633.1000305.Improved
115. Klein AM, Mazutis L, Akartuna I et al (2015) Droplet barcoding for single-cell transcriptomics applied to embryonic stem cells. Cell 161:1187–1201. https://doi.org/10.1016/j.cell.2015.04.044
116. Hwang B, Lee JH, Bang D (2018) Single-cell RNA sequencing technologies and bioinformatics pipelines. Exp Mol Med 50:96. https://doi.org/10.1038/s12276-018-0071-8
117. Lafzi A, Moutinho C, Picelli S, Heyn H (2018) Tutorial: guidelines for the experimental design of single-cell RNA sequencing studies. Nat Protoc 13:2742–2757. https://doi.org/10.1038/s41596-018-0073-y
118. Tian L, Dong X, Freytag S et al (2019) Benchmarking single cell RNA-sequencing analysis pipelines using mixture control experiments. Nat Methods 16:479–487. https://doi.org/10.1038/s41592-019-0425-8
119. Lytal N, Ran D, An L (2020) Normalization methods on single-cell RNA-seq data: an empirical survey. Front Genet 11:1–14. https://doi.org/10.3389/fgene.2020.00041
120. Kiselev VY, Andrews TS, Hemberg M (2019) Challenges in unsupervised clustering of single-cell RNA-seq data. Nat Rev Genet 20:273–282. https://doi.org/10.1038/s41576-018-0088-9
121. Saelens W, Cannoodt R, Todorov H, Saeys Y (2019) A comparison of single-cell trajectory inference methods. Nat Biotechnol 37:547–554. https://doi.org/10.1038/s41587-019-0071-9
122. Poirion OB, Zhu X, Ching T, Garmire L (2016) Single-cell transcriptomics bioinformatics and computational challenges. Front Genet 7:1–11. https://doi.org/10.3389/fgene.2016.00163
123. Barbazuk WB, Fu Y, McGinnis KM (2008) Genome-wide analyses of alternative splicing in plants: opportunities and challenges. Genome Res 18:1382–1391. https://doi.org/10.1101/gr.053678.106
124. Steijger T, Abril JF, Engström PG et al (2013) Assessment of transcript reconstruction methods for RNA-seq. Nat Methods 10:1177–1184. https://doi.org/10.1038/nmeth.2714

125. Larsen PA, Smith TPL (2012) Application of circular consensus sequencing and network analysis to characterize the bovine IgG repertoire. BMC Immunol 13:1–12. https://doi.org/10.1186/1471-2172-13-52

126. Wang B, Tseng E, Regulski M et al (2016) Unveiling the complexity of the maize transcriptome by single-molecule long-read sequencing. Nat Commun 7:11708. https://doi.org/10.1038/ncomms11708

127. Gordon SP, Tseng E, Salamov A et al (2015) Widespread polycistronic transcripts in fungi revealed by single-molecule mRNA sequencing. PLoS One 10:1–15. https://doi.org/10.1371/journal.pone.0132628

128. Hackl T, Hedrich R, Schultz J, Förster F (2014) Proovread: large-scale high-accuracy PacBio correction through iterative short read consensus. Bioinformatics 30:3004–3011. https://doi.org/10.1093/bioinformatics/btu392

129. Rhoads A, Au KF (2015) PacBio sequencing and its applications. Genomics Proteomics Bioinformatics 13:278–289

130. Martin LBB, Fei Z, Giovannoni JJ, Rose JKC (2013) Catalyzing plant science research with RNA-seq. Front Plant Sci 4:1–10. https://doi.org/10.3389/fpls.2013.00066

131. Van Verk MC, Hickman R, Pieterse CMJ, Van Wees SCM (2013) RNA-Seq: revelation of the messengers. Trends Plant Sci 18:175–179. https://doi.org/10.1016/j.tplants.2013.02.001

Chapter 11

RNA Interference (RNAi) in Tomato Crop Research

Pasquale Termolino

Abstract

RNA interference (RNAi) is a posttranscriptional gene silencing phenomenon induced by double-stranded RNA. It has been widely used as a knockdown technology to analyze gene function in many organisms. In tomato, RNAi technology has widely been used as a reverse genetic tool for functional genomics study. Generally, RNAi is often achieved through transgenes producing hairpin RNA molecules. RNAi lines have the advantage with respect to more modern CRISPR/Cas9 mutants of different levels of downregulation of target gene, and allow the characterization of life-essential genes that cannot be knocked out without killing the organism. Also, RNAi allows to suppress gene expression in multigene families in a regulated manner. In this chapter, an efficient approach to create RNAi stable knockdown-transformed tomato lines is reported. In order, it describes the choice of the target silencing fragment, a highly efficient cloning strategy for the hairpin RNA construct production, a relatively easy procedure to transform and regenerate tomato plants using *Agrobacterium tumefaciens* and a methodology to test the goodness of the transformation procedure.

Key words RNA interference, Tomato, *Agrobacterium*-mediated transformation, Gateway cloning, Silencing

1 Introduction

RNAi was discovered more than 20 years ago; it is a conserved mechanism among many eukaryotes that downregulates gene expression. Sequence-specific dsRNA gene regulation drives inhibition of transcription or translation.

In plants, dsRNA activating RNAi originates from viral sources, transcription of inverted repeats, stress-induced overlapping antisense transcripts, and RNA-directed RNA polymerase (RDR) transcription of aberrant transcripts [1–4].

In brief, dsRNA is recognized inside the cell, processed by Dicer-like (DCL) endoribonucleases from RNase III family [5] into 21–24 nt short interfering RNAs (siRNAs) [6].

The siRNAs are then incorporated in an RNA-induced silencing complex (RISC), a multi-component ribonucleoprotein that recognizes mRNA molecules with homology to the siRNA inside

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_11,

the complex and cuts the corresponding gene(s) inducing degradation of the messenger RNA.

RNAi phenomenon in plants is not limited to the transformed cell, siRNAs are able to move through to neighboring cells via symplast [7], while RNA molecules of a yet unknown nature can move through the apoplast to distant parts of the plant, generating systemic silencing [8, 9].

RNAi has been conventionally based on the use of transgenic plants expressing double-stranded RNAs (dsRNAs) targeting a specific exon (or exons); in crop plants, and in tomato, in particular, a very efficient way to generate dsRNAs is to insert into the plant genome a construct encoding for a hairpin RNA (hpRNA) formed by an inverted repeat sequence of a small fragment homologous to a region of the target gene separated by a spacer [10]. The inserted construct is dominant and therefore phenotypes can be screened in T0 or T1 plants without the need to produce homozygous lines.

For tailoring of constructs for gene silencing it is really important to choose the appropriate vector, promoter, marker, and transformation method; all these elements contribute to the efficiency and modulation of RNAi.

In this work, expression vector was designed using p HELLS GATE 12 (pHG12) plasmid as backbone. pHG12 is engineered with gateway cloning sites flanking the hairpin construct allowing directional recombination in a single step, with no use of restriction enzymes. This system has been successfully used to generate RNAi lines in the model plant *Arabidopsis thaliana* and in tomato (*Solanum lycopersicum*) [11, 12].

2 Materials

Before starting, be sure to have all the facilities required for safe and sterile manipulation of the bacterial and in vitro plant cultures. Some equipment and facilities used in the procedure are listed:

1. Autoclave (model De Lama 70).
2. Static incubation oven (Beckman Coulter).
3. Orbital incubator (VWR).
4. Thermostatic heat block.
5. Platform to visualize agarose electrophoresis gels (ChemiDoc XSR+ from Bio-Rad).
6. Electroporator with 0.1 cm gap sterile electroporation cuvettes (GenePulser from Bio-Rad).
7. Laminar flow hoods, microbiological and in vitro.
8. Climatic room with controlled environment.

It is recommended to prepare all solutions using ultrapure water or (when indicated) with analytical grade solutions.

Autoclaving is always standard cycle 121 °C for 20 min.

2.1 Creation of Expression Clone for RNAi Induction

1. pDonr/zeo entry vector (*see* **Note 1**).
2. pHellsGate12 (or pAgrikola) binary vector for plant stable transformation (*see* **Note 2**).
3. Selected tomato variety genomic DNA or cDNA (*see* **Note 3**).
4. Plasmid DNA miniprep kit.
5. Reagents for both standard and high-fidelity PCR (*see* **Note 4**).
6. Reagents for performing agarose gel electrophoresis.
7. Gateway cloning system (Thermo Fisher Scientific) including Gateway BP Clonase II Enzyme mix and Reagents, Gateway LR Clonase II Enzyme mix and Reagents.
8. Standard M13 sequencing primers.
9. Proteinase K solution.
10. One Shot™ TOP10 Chemically Competent *E. coli* (*see* **Note 5**).
11. Low-salt LB liquid medium (LBL): 10 g/L Tryptone, 5 g/L NaCl, 5 g/L yeast extract. Dissolve ingredients in water. Adjust the pH to 7.5 and bring to volume. Autoclave.
12. Low-salt LB agar plates (LBS): Add 1.5% bacterial agar to nonautoclaved LB medium, mix, and autoclave.
13. Zeocin®, 100 mg/mL stock solution.
14. Spectinomycin, 100 mg/mL stock solution.
15. Miniprep kit for plasmid extraction (any brand).

2.2 Plant Transformation

2.2.1 General Supplies

1. Sterile deionized water.
2. 70% ETOH.
3. NaOCl (4.9% active chloride) + 0.1% SDS solution.
4. Sterile Whatman filter paper.
5. Pipetman complete set and sterile tips.
6. Sterile paper towels.
7. 100 × 20 mm Sterile Petri plates.
8. Parafilm.
9. 50 mL Falcon disposable centrifuge tube, or similar.
10. 150 mL Sterile urine container adapted for in vitro usage (*see* **Note 6**).
11. Magenta™ GA-7 boxes (referred to as Magenta).

2.2.2 Plant Material and Agrobacterium *Strain*

1. 400 Seeds of tomato genotype of interest.
2. Choose a *S. lycopersicum* genotype of interest and be sure to have at least 400 seeds available (*see* **Note 7**).
3. *Agrobacterium tumefaciens* electrocompetent cells (suggested strain: LBA4404, *see* **Note 8**).

2.2.3 Media Components, Stock Solution, and Culture Media

1. YEP liquid medium (YL): 10 g/L Bacto peptone, 10 g/L yeast extract, 5 g/L NaCl. Dissolve ingredients in water. Autoclave.
2. YEP agar plates (YS): Add 1.5% bacterial agar to nonautoclaved YEP medium, mix, and autoclave (*see* **Note 9**).
3. *A. tumefaciens* selection liquid medium: YL + 100 μg/mL kanamycin and 50 μg/mL rifampicin.
4. *A. tumefaciens* selection plates: YS with 100 μg/mL kanamycin and 50 μg/mL rifampicin.
5. Thiamine HCl, 1 mg/mL (*see* **Note 10**).
6. Modified Nitsch vitamins, 1000×: 0.1 g of glycine, 0.5 g of nicotinic acid, 0.025 g of pyridoxine HCl, 0.025 g of thiamine HCl, 0.025 g of folic acid, and 0.002 g of *d*-biotin in 50 mL deionized H_2O. Adjust the pH to 7.00 (*see* **Note 11**).
7. Trans-zeatin, 1 mg/mL (*see* **Note 12**).
8. Kanamycin, 100 mg/mL.
9. Carbenicillin 50 mg/mL.
10. Rifampicin 12 mg/mL (*see* **Note 13**).
11. *SIM*: 20 mM Sodium citrate, 2% sucrose (pH 5.5).
12. *MSO*: 4.30 g/L MS salts including vitamins, 0.4 mg/L thiamine, 10 mg/L myoinositol, 30 g/L sucrose (pH 5.8).
13. *RB1*: 4.30 g/L MS salts including vitamins, 1 mL/L vitamin B5, 30 g/L sucrose, 8 g/L microagar, 1.0 mg/L zeatin riboside (pH 5.8).
14. *RD1*: MS salts including vitamins 4.30 g/L, vitamin B5 1 mL/L, 20 g/L sucrose, 8 g/L microagar, 0.5 mg/L zeatin riboside, 50 mg/L myoinositol, 0.1 mg/L indoleacetic acid (pH 5.8).
15. *RM1*: 4.30 g/L MS salts including vitamins, 1 mL/L vitamin B5, 20 g/L sucrose, 8 g/L microagar (pH 5.8).
16. Sterile electroporation cuvettes with 0.1 cm gap.

Adding appropriate chemical to media: after autoclaving cool the medium to 60 °C, and add hormones and antibiotics before pouring.

Dispense 24 mL of medium per Petri plate. Dispense 30 mL of medium in urine sterile container.

3 Methods

3.1 RNAi Target Design

With RNAi it is possible to downregulate one specific gene (avoiding any possible off-target), or gene families (*see* **Note 14**). It is recommended to use the VIGS designer tool provided by the Solanaceae Genomics Network (http://solgenomics.net/tools/vigs). The tool is very well described by Fernandez-Pozo and collaborators [13]. It is also possible to generate RNAi constructs manually as described by Fantini and Giuliano [14]. Some guidelines are summarized below.

It is recommended to design a fragment on 3′UTR region or exon possibly closer to 3′ end. Size of the target fragment should range from 300 to 550 bp. Once the silencing fragment is selected, design amplification primers using 20–25 nucleotides of the 5′ for the forward primer and 20–25 nucleotides of the 3′ (reverse complement) for the reverse primer. Add the attB cloning sites to the 5′ of the forward and reverse primers, according to the Gateway Instruction Manual: forward attB1 primer: 5′- GGGGA CAAGTTTGTACAAAAAAGCAGGCT -forward amplification primer-3′ reverse attB2 primer: 5′- GGGGACCACTTTGTACAA GAAAGCTGGGT-reverse amplification primer-3′. Use amplification primers without the attB cloning sites for PCR and/or sequencing screening.

3.2 Cloning Procedure

1. Use the designed attB primers in order to amplify the silencing fragment from DNA or cDNA with a high-fidelity DNA polymerase, using the recommended cycling parameters (*see* **Notes 3** and **4**), in a final volume of 20–50 μL.
2. attB-flanked PCR product purification: Evaluate PCR product size on a 1% agarose electrophoresis gel with appropriate DNA stain with an adequate amount of PCR product. Verify that you have a single band of expected size and then proceed further. If multiple bands are present, extract the target band and re-amplify from **step 1** of cloning procedure. It is necessary to obtain a final single PCR product (*see* **Note 15**).
3. Proceed with the BP recombination reaction according to the Gateway Instruction Manual. This reaction involves the attB-flanked PCR product and the donor vector in order to produce the entry clone harboring the silencing fragment flanked by attL recombination sites:
   (a) In a 1.5 mL tube mix 50 fmol of attB-PCR product (*see* **Note 16**) and 50 fmol of donor vector (150 ng/μL, 1 μL) and add TE buffer to reach a total volume of 8 μL.
   (b) Add to the mix 2 μL of BP Clonase II enzyme mix and vortex. Spin briefly the sample and incubate the reaction at 25 °C for 1 or 2 h.

(c) Stop the reaction adding 1 μL of the Proteinase K solution to the sample and vortex briefly. Incubate the sample at 37 °C for 10 min.

4. Transform *E. coli* competent cells with the BP reaction: Add 1 μL of BP reaction into 50 μL of One Shot TOP10 chemically competent *E. coli*. Incubate on ice for 30 min. Heat-shock the cells at 42 °C for 45 s. Incubate on ice for 2 min, and then add 500 μL of LB medium. Incubate in agitation at 37 °C for 1.15 h. Select 100 μL on low-salt LB plates plus 50 μg/mL zeocin.
5. Make a miniprep of plasmid DNA from single colonies and confirm the presence of the silencing fragment either by:
   (a) PCR amplification with universal forward primer and silencing fragment reverse primer without the attB cloning site or vice versa, with universal reverse primer and silencing fragment forward primer without the attB cloning site
   (b) Sequencing with universal M13 forward or reverse primers (*see* **Note 17**)
6. Proceed with LR recombination reaction according to the Gateway Instruction Manual. This reaction involves the entry clone harboring the silencing fragment and the pHellsGate12 destination vector in order to produce the hairpin RNA expression clone:
   (a) In a 1.5 mL tube mix 100 fmol of entry clone harboring the silencing fragment, 50 fmol of destination vector (*see* **Note 18**), and TE buffer to reach a total volume of 8 μL.
   (b) Add to the mix 2 μL of LR Clonase II enzyme mix and mix by vortexing. Spin briefly the sample and incubate the reaction at 25 °C for 1 h to O.N.
   (c) To stop the reaction, add 1 μL of Proteinase K solution to the sample and vortex briefly. Incubate the sample at 37 °C for 10 min.
7. Transform *E. coli*-competent cells with the LR reaction: Add 5 μL of LR reaction into 50 μL of One Shot TOP10 Chemically Competent *E. coli*. Incubate on ice for 30 min. Heat-shock the cells at 42 °C for 45 s. Incubate on ice for 2 min, and then add 700 μL of LB medium. Incubate in agitation at 37 °C for 1.15 h. Select 100–400 μL on LB plates plus 100 μg/mL spectinomycin.
8. Make a miniprep of plasmid DNA from single colonies and confirm the presence of the silencing fragment either by:
   (a) PCR amplification with AGRI51 forward primer and AGRI54 reverse primer for site 1 verification and with AGRI64 forward primer and AGRI69 reverse primer for site 2 (Fig. 1) (*see* **Notes 17** and **18**)

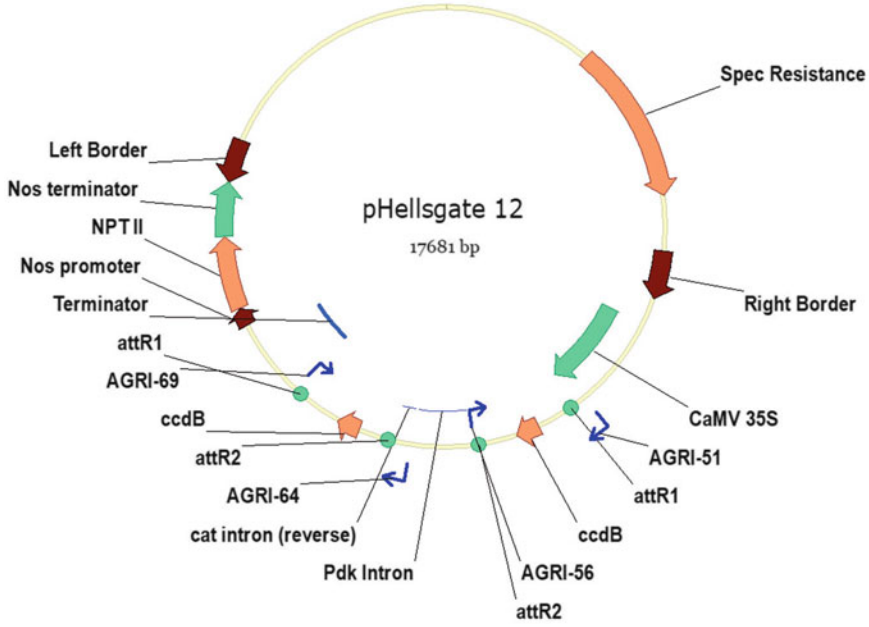


Fig. 1 pHellsGate12 plasmid map; blue arrows indicate primer-binding sites. Bacterial selection is spectinomycin; resistance transferred to plant is kanamycin

Table 1
List of primers needed for vector and insert verification

Primer name	Primer sequence
AGRI51	CAACCACGTCTTCAAAGCAA
AGRI54	CTGGGGTACCGAATTCCTC
AGRI64	CTTGCGCTGCAGTTATCATC
AGRI69	AGGCGTCTCGCATATCTCAT

(b) Sequencing with AGRI51, AGRI54, AGRI64, and AGRI69 primers (Table 1)

3.3 Plant Transformation Procedure

3.3.1 Preparation of A. tumefaciens *Strain*

Transform *A. tumefaciens* LBA4404 electrocompetent with obtained expression clone in an ice-cold 0.1 cm gap sterile electroporation cuvette and add 150 ng of plasmid to 100 μL of *Agrobacterium* LBA4404 competent cells. Perform the electroporation according to the electroporator settings. Immediately resuspend the cells in 900 μL of YL medium and incubate at 28 °C in agitation for 1 h and 30 min. Plate 150 μL of resuspended cells on YS selection plate with appropriate antibiotics. Incubate for 48 h at 28 °C, when colonies reach sufficient size for picking. Check by PCR ten colonies from each transformation (use AGRI primers

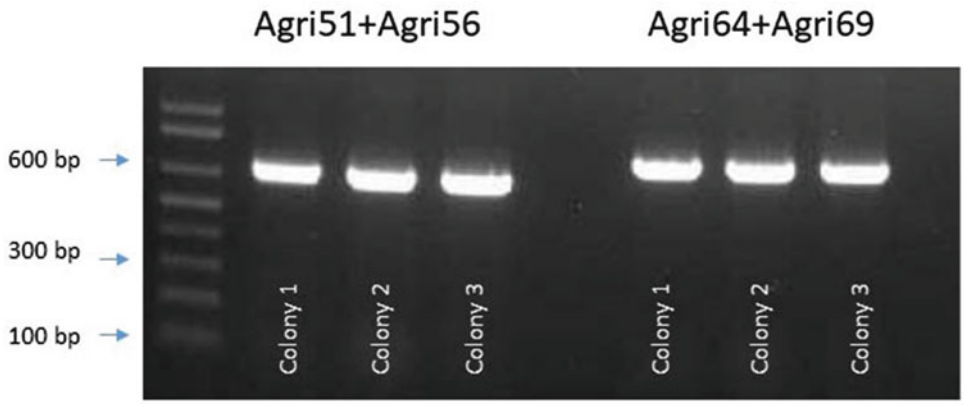


Fig. 2 Example of PCR colony of transformed *A. tumefaciens* strain performed with AGRI primers. All colonies result to be positive for inserted plasmid

from Table 1) (Fig. 2). Pick at least two positive colonies and inoculate into YL at 28° O.N. Plate the same two colonies on fresh YS plate with appropriate selection.

The resulting inoculum can be used for either generating a frozen stock or proceeding with plant transformation (*see* **Note 19**).

3.3.2 Preparation of Plant Material and Transformation

1. Sterilize seeds by submerging in 70% ETOH for 5 min, change solution with fresh 70% ETOH, rinse for 1 min, then wash in 10% of commercial NaOCl (4.9% active chloride) + 0.1% SDS two times for 10 min, and finally rinse five times in sterile H_2O.
2. Sow seeds (20) in two sterile urine containers (10 seeds each) containing RB1 (no zeatin) and incubate in a 25 °C growth room with a photoperiod of 16-h light/8-h dark for 9–10 days.
3. Prepare *Agrobacterium tumefaciens* from Subheading 3.3.1 2 days before transformation as follows: streak a single colony in 5–10 mL YEP with appropriate antibiotics at 28 °C. Grow the bacteria at 1 OD_{600} nm concentration. Centrifuge agro culture at 6000 × *g* for 15 min. Resuspend the pellet in SIM with acetosyringone (0.075 mg/mL) in double volume as starting culture (final OD_{600} should be 0.5). Incubate at 25 °C in agitation in the dark. Centrifuge bacteria at 6000 × *g* for 15 min. Resuspend the pellet in the same volume as starting of MSO with acetosyringone (0.075 mg/mL) (*see* **Note 20**).
4. Cut cotyledons from 6- to 8-day-old seedlings at proximal and distal end; each section should be about 1–0.8 cm. Transfer explants into the *Agrobacterium* solution from the previous step and incubate for 15 min with occasional shaking. Blot explants on sterile filter paper and place with abaxial side down on RB1 + acetosyringone (0.075 mg/mL). Incubate in a 25 °C growth room in the dark for 48 h. After 2 days of

Fig. 3 Wild-type tomato plant (variety M82) and a transformed T01 line after 6 weeks of regeneration on RD1 medium

incubation transfer explants to plates with RD1 + carbenicillin$_{500}$ (500 μg/mL) + kanamycin$_{50}$ (50 μg/mL) and incubate in a 25 °C growth room with 16-h light/8-h dark photoperiod. After 2 weeks, transfer explants to a new fresh RD1 medium + antibiotics and keep on substrate till the regenerated explants reach adequate size. Excise calli from explants, cut in pieces, and transfer again on RD1 + kanamycin$_{50}$ in sterile urine containers. After 5–6 weeks, shoots should appear (Fig. 3); excise and transfer them on RM1 selective medium in sterile Magenta containers to give them more space. Roots should appear in 10 days (*see* **Note 21**) (Fig. 4).

3.3.3 Verification of Transformed Tomato RNAi Lines

RNAi lines have to be verified after transformation. Detection can be made with a simple PCR experiment with AGRI primers (*see* Subheading 3.2) using DNA extracted from the mother plant as a template; these regions do not exist in wild-type tomato plants; thus positive plants to amplification integrate the hpRNA.

Transformation can lead to different levels of expression of the hpRNA. Multiple insertion can activate downregulation of the hpRNA expression itself: it is recommended to verify the copy number of the inserted DNA (T-DNA). In this work to verify the copy number of T-DNA insert, SAQPCR (standard addition qPCR) was performed. Assemble at least one reference plasmid and follow the procedure as described by Huang and collaborators [15]. Copy number check was performed relative to PHYTOENE DESATURASE (single-copy gene ID:Solyc03g123760 in https://solgenomics.net tomato genome database) cloned in a specific entry vector (pDonr/zeo) (*see* **Note 22**).

Fig. 4 Rooted transformed tomato plant into a magenta box

4 Notes

1. It is possible to use different entry vectors; we also used Gateway™ pDONR™221 Vector with kanamycin selection marker, but it is important to use a different antibiotic selection than the destination vector.
2. In this work we used pHellsGate12 vector from CSIRO (www.csiro.au) but other vectors can be used instead. For example, in *Arabidopsis* using the pAgrikola gateway vector [16], Czarnecki and collaborators [17] successfully knocked down six nonfamily genes in *Arabidopsis thaliana*.
3. Be very specific in the choice of starting template DNA or cDNA; do not use different variety or mutant lines as target fragment specificity can decrease. cDNA is usually best to perform amplification since it has no introns.
4. The use of high-fidelity DNA polymerase is preferred in order to avoid mutations in the silencing fragment.
5. For this protocol, an in-house-prepared stock of TOP10 was used. Any *E. coli* commercial competent strain will be suitable. Avoid only ccdB-type strain that has limited selection for the plasmid gateway vectors.
6. It is possible to use any sterile transparent plastic container that can be hermetically closed; in this protocol urine sterile containers are used because they are compact and resistant and hold hermetically gas and liquid very well. The volume of substrate usually optimal is around 25–30 mL.

7. Various tomato genotypes have been transformed with this protocol including M82, Ailsa Craig, MicroTom, and Moneymaker; the best performance has been achieved with Ailsa Craig variety. Use of MicroTom variety can speed up the procedure due to its shorter growth period but avoid it if you are choosing to downregulate any hormone or growth-related gene. MicroTom has many mutations in those pathways that can interfere with selected target silencing [18].
8. In this protocol we used *Agrobacterium tumefaciens* strains LBA4404; this strain does not have a great virulence compared to other *A. tumefaciens* strains but has a better chance to avoid overgrowth on explants. The flocculation level during liquid growth is very high but it is normal. An alternative used with this protocol is *A. tumefaciens* strain GV3101 which has increased virulence and reduced flocculation in a liquid growth medium. Change strain only if you find a low transformation rate.
9. If needed it is possible to make stock preparation of YS in bottles without antibiotic. When needed, melt the medium in a microwave with adequate power for necessary time (depending on the microwave type) and then add the appropriate antibiotic when melted medium reaches 65 °C or less.
10. Wrap in foil and store at 4 °C for max 2 months.
11. Store in 1 mL aliquots at −20 °C.
12. To prepare the stock solution, dissolve 50 mg of trans-zeatin in a few drops of 0.5 M HCl. Add deionized H_2O to a total volume of 50 mL. Filter sterilize and store in 1 mL aliquots at −20 °C. Zeatin must be added after autoclaving since it is heat sensitive.
13. For stock solution, dissolve 120 mg of rifampicin in 10 mL of methanol. Wrap in foil and store at −20 °C.
14. Silencing gene families or single targets depends on the choice of the target fragment region. If you need to be very specific it is better to use the VIGS designer that automatically searches the tomato database to find the more specific target very efficiently; if you need to cluster silence families the best option is to choose a common coding region and design the target amplicon consequently. If trying to design the common fragment you go below 300 bp of fragment size and then do not go below 200 bp; otherwise it is highly possible to have a nonfunctional hairpin RNA.
15. PCR with attB primers can be tricky; using so long primers can lead to multiple PCR targets and using the correct melting temperature (Tm) is not trivial. If you use Vector NTI software use the Tm recommended by Vector output; if you do not use

it then subtract 10 °C from the calculated Tm. In any case you should obtain one single band; if not try to increase Tm first; if it does not work then extract the target band and re-amplify it as in **step 1** of Subheading 3.2.

16. attB-PCR product and donor vector have to be in an equimolar ratio. It is possible to convert femtomole to nanogram, using the following formula: ng = (fmol)(*N*)(660 fg/fmol)(1 ng/10^6 fg) where *N* is the size of the DNA in bp. For attB-PCR products of 300 bp, the amount of attB-PCR product required for the reaction is (50 fmol)(300 bp)(660 fg/fmol)(1 ng/10^6 fg) = 9.9 ng. 50 fmol of donor vector (pDONR/Zeo) is approximately 150 ng.
17. It is convenient to make colony PCR and inoculate at the same time with the same picking tool. Choose adequate number of single colonies, and pick on a new selection plate; the same picking tool can be used to dissolve sample into 20–30 μL of water. It is recommended to boil the colonies for 10 min and then cool on ice for 2 min. Use 5–10 μL of the boiled colony as PCR template.
18. Primer pairs are as follows: first couple: AGRI51 + AGRI56; second couple: AGRI64 + AGRI69. Both PCR reactions can be cycled as follows: 95 °C × 5 min (1 cycle); 95 °C × 30 s + 52 °C × 30 s + 72 °C × 1 min (35 cycles); and end with 1 cycle at 72 °C per 10 min.
19. Prepare the *Agrobacterium* culture when you are ready to make the co-culture; verify bacterial concentration by spectrophotometer; optimal OD concentration should be around 1. If too much concentrated dilute it to 0.8 OD and grow for another 2 h.
20. Use freshly prepared acetosyringone; it tends to precipitate in long storage conservation.
21. Carbenicillin is used to exterminate residual *Agrobacterium* from any explants but it can cause regeneration slowdown; it is possible to substitute with cefotaxime$_{100}$. If experiencing procedure slowdown, halve the concentration of these antibiotics. If rooting does not happen in more than 10 days, stimulate rooting treating with a solution of 1-naphthaleneacetic acid (NAA) or indole-3-acetic acid (IAA). Just dip the shoot into a NAA solution (0.1 mg/L) and place them back into RM1 with no antibiotics.
22. RNAi effect tends to diminish with generations of self-fertilization; it is recommended to propagate T0 or T1 lines clonally.

Acknowledgment

This work was made possible thanks to Dr. Silvana Grandillo from CNR-IBBR; she granted funding and facilities to set up the procedure.

References

1. Hamilton AJ, Baulcombe DC (1999) A species of small antisense RNA in posttranscriptional gene silencing in plants. Science 286:950–952
2. Mette MF, van der Winden J, Matzke MA, Matzke AJ (1999) Production of aberrant promoter transcripts contributes to methylation and silencing of unlinked homologous promoters in trans. EMBO J 18:241–248
3. Borsani O, Zhu J, Verslues PE, Sunkar R, Zhu JK (2005) Endogenous siRNAs derived from a pair of natural cis-antisense transcripts regulate salt tolerance in *Arabidopsis*. Cell 123:1279–1291
4. Luo Z, Chen Z (2007) Improperly terminated, unpolyadenylated mRNA of sense transgenes is targeted by RDR6-mediated RNA silencing in *Arabidopsis*. Plant Cell 19:943–958
5. Vazquez F, Legrand S, Windels D (2010) The biosynthetic pathways and biological scopes of plant small RNAs. Trends Plant Sci 15:337–345
6. Liu Q, Feng Y, Zhu Z (2009) Dicer-like (DCL) proteins in plants. Funct Integr Genomics 9:277–286
7. Dalakouras A, Wassenegger M, Dadami E, Ganopoulos I, Pappas ML, Papadopoulou K (2020) Genetically modified organism-free RNA interference: exogenous application of RNA molecules in plants. Plant Physiol 182:38–50
8. Voinnet O, Baulcombe DC (1997) Systemic signalling in gene silencing. Nature 389:553
9. Palauqui JC, Vaucheret H (1998) Transgenes are dispensable for the RNA degradation step of cosuppression. Proc Natl Acad Sci U S A 95:9675–9680
10. Halliwell C, Waterhouse P (2003) Constructs and methods for high throughput gene silencing in plants. Methods 30:289–295
11. Wielopolska A, Townley H, Moore I, Waterhouse P, Helliwell C (2005) A high-throughput inducible RNAi vector for plants. Plant Biotechnol J 3:583–590
12. Dutta TK, Papolu PK, Banakar P, Choudhary D, Sirohi A, Rao U (2015) Tomato transgenic plants expressing hairpin construct of a nematode protease gene conferred enhanced resistance to root-knot nematodes. Front Microbiol 6:1–14
13. Fernandez-Pozo N, Rosli HG, Martin GB, Mueller LA (2015) The SGN VIGS tool: user-friendly software to design virus-induced gene silencing (VIGS) constructs for functional genomics. Mol Plant 8:486–488
14. Fantini E, Giuliano G (2016) Virus-induced gene silencing as a tool to study tomato fruit biochemistry. Methods Mol Biol 1363:65–78
15. Huang Y, Yin X, Zhu C, Wang W, Grierson D, Xu C, Chenet K (2013) Standard addition quantitative real-time PCR (SAQPCR): a novel approach for determination of transgene copy number avoiding PCR efficiency estimation. PLoS One 8:1–8
16. Hilson P, Allemeersch J, Altmann T, Aubourg S, Avon A, Beynon J, Bhalerao RP, Bitton F, Caboche M, Cannoot B, Chardakov V, Cognet-Holliger C, Colot V, Crowe M, Darimont C, Durinck S, Eickhoff H, de Longevialle AF, Farmer EE, Grant M, Kuiper MT, Lehrach H, Léon C, Leyva A, Lundeberg J, Lurin C, Moreau Y, Nietfeld W, Paz-Ares J, Reymond P, Rouzé P, Sandberg G, Segura MD, Serizet C, Tabrett A, Taconnat L, Thareau V, Van Hummelen P, Vercruysse S, Vuylsteke M, Weingartner M, Weisbeek PJ, Wirta V, Wittink FR, Zabeau M, Small I (2004) Versatile gene-specific sequence tags for Arabidopsis functional genomics: transcript profiling and reverse genetics applications. Genome Res 14(10B):2176–2189
17. Czarnecki O, Bryan AC, Jawdy SS, Yang X, Cheng ZM, Chen JG, Tuskan GA (2016) Simultaneous knockdown of six non-family genes using a single synthetic RNAi fragment in *Arabidopsis thaliana*. Plant Methods 12:16
18. Shikata M, Ezura H (2016) Micro-Tom tomato as an alternative plant model system: mutant collection and efficient transformation, Methods in molecular biology, vol 1363. Humana, New York, NY

Chapter 12

Protoplast-Based Method for Genome Editing in Tetraploid Potato

Alessandro Nicolia, Ann-Sofie Fält, Per Hofvander, and Mariette Andersson

Abstract

The cultivated potato is tetraploid with four probably equivalent loci for each gene. A potato variety is furthermore commonly genetically heterogeneous and selected based on a beneficial genetic context which is maintained by clonal propagation. When introducing genetic changes by genome editing it is then desirable to achieve edits in all four loci for a certain gene target. This is in order to avoid crosses to achieve homozygosity for edited gene loci and at the same time reduce risk of inbreeding depression. In such a context transient transfection of protoplasts for the introduction of mutations, avoiding stable insertion of foreign DNA, would be very attractive. The protocol of this chapter has been shown to be applicable for the introduction of mutations by DNA vectors containing expression cassettes of TALEN, Cas9, and Cas9 deaminase fusions together with sgRNA expression cassettes on either single or separate vectors. Furthermore, the protoplast-based system has been shown to work very efficiently for mutations introduced by in vitro-produced and transfected RNP (ribonucleoprotein) complexes.

Key words Potato, Genome editing, Targeted mutagenesis, TALEN, CRISPR-Cas9, DNA vector, RNP

1 Introduction

Potato, *Solanum tuberosum*, is one of the most important crops worldwide and is expanding in cultivation. The genetics of potato are complex from a breeding point of view, being an autotetraploid. In addition most successful potato varieties are highly heterozygous resulting from beneficial genetic combinations yielding a heterosis effect that are then maintained by clonal propagation.

Lately, genome editing has emerged as an attractive technology to introduce genetic variation and specific traits in crop plants, including potato. Several means of genome editing have successfully been explored in potato using DNA-based TALEN and CRISPR/Cas9 methods, as well as a DNA-free method through RNP (ribonucleoprotein) complexes of Cas9 and sgRNA [1–

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_12,

5]. The two major means of application of genome editing tools to potato cells have been either via *Agrobacterium tumefaciens* transformation of leaf tissue with stable integration of expression cassettes or via transfection and transient application in protoplasts (for a recent review *see* ref. 6).

Although the use of protoplasts could be considered as carrying a higher risk of somaclonal variation than using other tissues and conventional *Agrobacterium* transformation, protoplast transfection and regeneration have the great advantage of minimizing or even avoiding stable integration of recombinant DNA. Then there is no need for crossing out introduced DNA which would disrupt the beneficial genetic context of a particular potato variety. The use of RNP with synthetic sgRNA furthermore eliminates any risk of unintended integration of DNA. In most cases, genotypes with increased genetic variation but with no introduction of recombinant DNA, intended or not, will carry a much lower regulatory burden for taking newly developed genotypes to field trials or even to the market carrying novel valuable traits.

In this chapter we first describe the isolation of protoplasts from potato leaf tissue and their use for application of genome editing tools. This can be used to monitor the efficiency of sgRNAs in the evaluation of different targets to find the best combination to pursue the generation of specific mutations or increase the genetic variation at certain loci. Secondly, we describe the subsequent methods for cell proliferation into calli and regeneration of shoots from dividing calli. The protocol of this chapter has been shown to be applicable for the introduction of mutations by various means such as DNA vectors containing expression cassettes for TALEN and different Cas9 variants as well as to work very efficiently for mutations introduced by in vitro-produced and transfected RNP complexes.

2 Materials

Solutions are diluted in H_2O, unless otherwise stated. All solutions are sterilized through a 0.20 μm filter, except for the phyto agar and Gelrite solutions that are autoclaved before mixed with the filter-sterilized medium.

2.1 Stock Solutions

100× Macro: 74 g KNO_3, 49.2 g $MgSO_4{\cdot}7H_2O$, 3.4 g KH_2PO_4 in 1 L. Store at +4 °C.

100× FE/EDTA: 1.4 g Na_2EDTA, 1.9 g $FeSO_4{\cdot}7H_2O$ in 1 L. Store at +4 °C.

2M $CaCl_2$: 294 g $CaCl_2{\cdot}2H_2O$ in 1 L. Store at +4 °C.

1000× Micro: 1.5 g H_3BO_3, 5.0 g $MnSO_4{\cdot}H_2O$, 1.0 g $ZnSO_4{\cdot}7H_2O$, 0.12 g $Na_2MoO_4{\cdot}2H_2O$, 0.012 g

$CuSO_4 \cdot 5H_2O$, 0.012 g $CoCl_2 \cdot 6H_2O$, 0.38 g KI in 1 L. Store at +4 °C.

1000× Nitsch and Nitsch vitamins: 2 g Glycine, 100 g myoinositol, 0.5 g thiamin-HCl, 0.5 g pyridoxine-HCl, 5 g nicotinic acid, 0.5 g folic acid, 0.05 g biotin in 1 L. Dissolve all except glycine in a small volume of KOH. Store at −20 °C.

100× Vitamin 1: 0.5 g Pantothenic acid, 0.5 g choline chloride, 1 g ascorbic acid, 0.01 g *p*-aminobenzoic acid, 0.5 g nicotinic acid, 0.5 g pyridoxine, 5 g thiamine in 1 L. Mix nicotinic acid, pyridoxine, and thiamine in a small volume of KOH before adding to the solution. Store at −20 °C.

100× Vitamin 2: 0.2 g Folic acid, 0.005 g D(+)biotin, 0.01 g vitamin B_{12} in 1 L. Dissolve folic acid and D(+)biotine in a small volume of KOH. Store at −20 °C.

200× Vitamin 3: 0.01 g Vitamin D_3 in 1 L. Dissolve in a small volume of 70% ethanol. Store at −20 °C.

50× Sugars: 6.25 g Sorbitol, 6.25 g sucrose, 6.25 g D(−)fructose, 6.25 g D(−)ribose, 6.25 g D(+)xylose, 6.25 g D(+)mannose, 6.25 g L(−)rhamnose monohydrate, 6.25 g D(+)cellobiose, 2.5 mg myoinositol in 1 L. Store at +4 °C.

100× Organics: 1 g Pyruvic acid, 2 g fumaric acid, 2 g citric acid monohydrate, 2 g DL-malic acid. Store at +4 °C.

Silver thiosulfate solution (STS): 20.38 mg Silver nitrate ($AgNO_3$) in 10 mL. 0.238 g Sodium thiosulfate pentahydrate ($Na_2S_2O_3 \cdot 5H_2O$) in 10 mL. Mix the two solutions. Store at −20 °C.

BAP: 1 g 6-Benzylaminopurine in 1 L. Store at −20 °C.

NAA: 2 g 1-Naphthaleneacetic acid in 1 L. Dissolve NAA in a small volume of 1 M NaOH. Store at −20 °C.

GA_3: 1 g Gibberellic acid 3 in 1 L. Dissolve in a small volume of 1 M NaOH. Store at −20 °C.

2.2 Potato Leaf Multiplication and Treatment

***Medium A** (Multiplication Medium)*: 4.4 g Murashige and Skoog (MS) medium including vitamins and 30 g sucrose in 1 L. Set pH to 5.8 with KOH. 6 g Phyto agar in 1 L. Make both solutions double concentrated and mix in equal volumes once the phyto agar has cooled down somewhat. Optional: Add 1.33 mL STS. Pour in sterile boxes (e.g., plant container) and store at +4 °C.

***Medium B** (Conditioning Medium)*: 2.7 g MS medium NH_4NO_3 free (e.g., MS mod. No. 4 from Duchefa), 0.1 mL 1000× Nitsch and Nitsch vitamins, 100 mg casein hydrolysate, 2 mg NAA, 0.5 mg BAP. Set to pH 5.8 in 1 L. Prepare fresh the same day and store at +4 °C until use.

Plasmolysis Solution: 91.1 g D-Sorbitol in 1 L. Store at room temperature.

Medium C *(Enzyme Solution)*: 10 mL 100× Macro, 1 mL 1000× micro, 10 mL 100× Fe/EDTA, 5 mL of 100× vitamin 1, 5 mL of 100× vitamin 2, 2.5 mL of 200× vitamin 3, 5 mL 1000× Nitsch and Nitsch vitamins, 20 mL 50× sugars, 10 mL 100× organics, 500 mg casein hydrolysate, 40.63 g D-glucose monohydrate, 37.35 g mannitol, 20 g polyvinylpyrrolidone-10 (PVP-10), 1 mg NAA, 0.4 mg BAP, 10 g cellulase R10, 2 g Macerozyme R10, 3 mL 2M $CaCl_2$ in 1 L.

Add everything except the enzymes and $CaCl_2$. Mix and set volume. Then add Macerozyme followed by cellulase R10. Adjust pH to 5.6. Incubate at 55 °C for 10 min and let it cool down. Add $CaCl_2$. Prepare fresh the same day.

2.3 Protoplast Purification

Wash Solution: 10 mL 100× Macro, 3 mL 2M $CaCl_2$, 1 mL 1000× micro, 14.03 g NaCl, 2 mg NAA, 0.5 mg BAP, 10 mL 100× FE/EDTA in 1 L; set pH to 5.6 with HCl. Prepare fresh the same day. Store at room temperature until use.

Sucrose Solution: 147.19 g Sucrose in 1 L. Store at room temperature.

2.4 Protoplast PEG-Mediated Transformation

Transformation Buffer 1: 34.6 g Mannitol, 14.7 g $CaCl_2 \cdot 2H_2O$, 5 g MES in 1 L. Set pH to 5.6 with KOH. Store at +4 °C.

Transformation Buffer 2: 91.1 g Mannitol, 3.05 g $MgCl_2 \cdot 6H_2O$, 1 g MES in 1 L. Set to pH 5.6 with KOH. Store at room temperature.

PEG Solution: 250 g PEG 4000 (25%), 73 g mannitol, 24 g Ca $(NO_3)_2 \cdot 4H_2O$ in 1 L. Prepare fresh the same day.

2.5 Protoplast Culture

Medium E *(Culture Medium)*: 10 mL 100× Macro, 1 mL 1000× micro, 10 mL 100× Fe/EDTA, 1.25 mL 2M $CaCl_2$, 5 mL of 100× vitamin 1, 5 mL of 100× vitamin 2, 2.5 mL of 200× vitamin 3, 20 mL 50× sugars, 10 mL 100× organics, 500 mg casein hydrolysate, 33.69 g D-glucose monohydrate, 30.98 g mannitol, 2 g bovine serum albumin (BSA), 1 mg NAA, 0.4 mg BAP in 1 L. Set to pH 5.6 with KOH. Store at +4 °C.

Alginate Solution: 28 g Alginic acid sodium salt, 72.88 g sorbitol in 1 L. Dissolve on a magnetic stirrer with medium heating followed by autoclaving in an oversized bottle. Store at +4 °C.

Floating Solution: 72.88 g Sorbitol, 7.35 g $CaCl_2 \cdot 2H_2O$ in 1 L. Store at room temperature.

Setting agar: 72.88 g Sorbitol, 7.35 g $CaCl_2 \cdot 2H_2O$ in 1 L. 8 g Phyto agar in 1 L. Make both solutions double concentrated and mix in equal volumes once the phyto agar has cooled down somewhat. Pour in petri dishes and store at +4 °C.

Medium F *(Callus Growth Medium)*: 2.7 g MS medium NH_4NO_3 free (e.g., MS mod. No. 4 from Duchefa), 107 mg NH_4Cl, 1.0 mL 1000× Nitsch and Nitsch vitamins, 40 mg adenine sulfate, 100 mg casein hydrolysate, 2.5 g sucrose, 54.7 g mannitol, 0.1 mg NAA, 0.5 mg BAP in 1 L. Set to pH 5.8 with KOH.

Release medium: 5.88 g Sodium citrate dihydrate, 91.1 g sorbitol in 1 L. Store at +4 °C.

Medium G *(Shoot Induction Medium 1)*: 2.7 g MS medium NH_4NO_3 free (e.g., MS mod. No. 4 from Duchefa), 267.5 mg NH_4Cl, 1.0 mL 1000× Nitsch and Nitsch vitamins, 80 mg adenine sulfate, 100 mg casein hydrolysate, 2.5 g sucrose, 36.4 g mannitol, 0.1 mg indole-3-acetic acid (IAA), 2.5 mg zeatin riboside trans-isomer (ZEA) in 1 L. Set to pH 5.8 with KOH.

Medium H *(Shoot Induction Medium 2)*: 4.4 g MS medium including vitamins, 10 g sucrose, 2 mg ZEA, 0.01 mg NAA, 0.1 mg GA_3 in 1 L. Set to pH 5.8. 2.5 g Gelrite in 1 L. Make both solutions double concentrated and mix in equal volumes once the Gelrite has cooled down somewhat. Pour in petri dishes and store at +4 °C.

Medium I *(Shoot Induction Medium 3)*: 4.4 g MS medium including vitamins, 20 g sucrose, 0.1 mg GA_3 in 1 L. Set to pH 5.8. 2.5 g Gelrite in 1 L. Make both solutions double concentrated and mix in equal volumes once the Gelrite has cooled down somewhat. Pour in petri dishes and store at +4 °C.

2.6 Equipment

- Sterile plant container.
- Razor blades, carbon steel single edge (TED Pella, Inc).
- Plastic culture dish 100 × 20 mm.
- Aluminum foils.
- Parafilm.
- Temperature-controlled incubator.
- Orbital shaker.
- Sterile 0.20 μm filters.
- Sterile cell strainer filters of 100 and 70 μm.
- Sterile serological pipettes of 10 and 25 mL.
- Automatic pipette controller.
- Sterile plastic conical centrifuge tubes of 50 and 15 mL.

- Bench centrifuge with swing-out rotor.
- Laboratory micropipette and tips.
- Optical microscope with a hemocytometer or a cell counter.

3 Methods

3.1 Protoplast Isolation, Transfection, and Regeneration

1. Internodes of potato containing a single auxiliary bud (*see* **Note 1**) are propagated in vitro using plastic boxes (e.g., plant container) containing 50 mL of Medium A. Growing conditions are 24 °C/20 °C for 16-h light/8-h dark.
2. About 20–30 leaves (1 g) from 4- to 6-week-old plants are excised in sterile condition and placed with the abaxial side down in plastic petri dishes containing 20 mL of Medium B, then sealed with parafilm, covered with aluminum foil, and incubated at 4 °C for 24 h (*see* **Note 2**).
3. Leaves are subsequently cut in slices (1–2 mm) using a sterile blade in a glass petri dish containing a few mL of Medium B and collected in a fresh plastic culture dish containing 10 mL of Medium B (*see* **Note 3**).
4. Medium B is removed, slices are washed with 5–10 ml of plasmolysis solution and then incubated in 20 mL of plasmolysis solution for 30 min, and culture dish is kept covered with an aluminum foil and at room temperature (RT).
5. Plasmolysis solution is removed and substituted with 25 mL of Medium C, and petri dish is sealed with parafilm, wrapped in aluminum foil, and incubated overnight (ON) at 25 °C without shaking. The incubation should not exceed 14 h.
6. The next day the petri dish is incubated for 30 min at RT with very gentle shaking; the solution will turn green due to released protoplasts.
7. Two sterile filters of 100 and 70 μm are mounted together on a sterile 50 mL centrifuge tube and pre-wetted with 5 mL of wash solution. The solution containing released protoplasts is gently aspirated with a pipette and sieved through the filters; remaining protoplasts are washed from the filters using 10 mL of wash solution.
8. The sieved protoplast suspension is transferred to sterile 15 mL centrifuge tubes (8 mL per tube), and the tubes are topped up to 15 mL with additional wash solution. The suspension is centrifuged at 50 × *g* (minimum acceleration and deceleration) for 5 min. Supernatant is subsequently discarded and protoplasts are gently resuspended in 2 mL of wash solution.

9. Fresh sterile 15 mL centrifuge tubes containing 6 mL of sucrose solution are prepared and a maximum of 6 mL of resuspended protoplasts slowly layered on top with a sterile Pasteur pipette or a micropipette with a cut tip, taking care that the interface is not disrupted. The tubes are subsequently centrifuged at 50 × *g* for 15 min (minimum acceleration and deceleration); a thick dark band of protoplasts should appear at the interface of the two solutions (*see* **Note 4**).
10. A fresh sterile 15 mL centrifuge tube containing approximately 3 mL (1–6 mL can be used based on the darkness and thickness of the band) of transformation buffer 1 is prepared. Using a micropipette with a cut tip, the floating protoplasts are gently transferred from the previous step in the tube. A small amount (10–20 μL) of protoplasts is used to quantify density (protoplasts/mL) using a hemocytometer or a cell counter; protoplasts in transformation buffer 1 are stored at 4 °C in the dark during counting.
11. Protoplasts are centrifuged at 50 × *g* for 10 min (minimum acceleration and deceleration), supernatant is subsequently discarded, and protoplasts are gently resuspended in transformation buffer 2 at a concentration of 1.0×10^6 protoplasts/mL.
12. Fresh sterile 15 mL centrifuge tubes are prepared for each transfection or control (i.e., PEG+; PEG−). From 10 μL up to 20 μL of mutation reagents (DNA vector or RNP complex) are pipetted in each tube followed by 100 μL of protoplasts in transformation buffer 2 (approximately 100,000 protoplasts) (*see* **Note 5**).
13. A volume ranging from 110 to 120 μL of PEG solution, accordingly to the volume of protoplast and mutation reagent used, is gently added to each tube which is gently flicked before and after adding the PEG solution. Samples are incubated at RT for 3 min (*see* **Note 6**).
14. Transfection reactions are stopped by carefully adding 5 mL of wash solution to each tube and subsequently centrifuged at 50 × *g* for 5 min (minimum acceleration and deceleration).
15. Supernatant is discarded and transfected protoplasts or controls are gently resuspended in 1 mL of Medium E. The same volume of alginate solution is added to give a final density of 5×10^4 protoplasts/mL (*see* **Note 7**). The two solutions are gently mixed by inverting the tubes and the solution is transferred in aliquots (usually four big drops) to the surface of solid setting agar. The drops are left at RT for a maximum of 2 h to allow solidification of alginate.
16. The alginate lens are subsequently released from the surface of setting agar with the help of 1 mL of floating solution and moved to fresh petri dishes containing 10 mL of Medium

E. Petri dishes are sealed with parafilm, covered with aluminum foil, and incubated at 25 °C for 5 days.

17. After 5 days, constant light is gradually increased by replacing aluminum foil with a white paper sheet and subsequently a mesh filter cloth (app. 100 μm). Once protoplast mini calli are visible to the naked eye (usually after 3 weeks) Medium E is replaced with 10 mL of Medium F and calli will be exposed to full light (approximately 30 μmol/m^2/s) by this stage. Fresh Medium F is provided every week.
18. After 4–6 weeks in Medium F, calli are released from alginate drops adding 5 mL of releasing solution and incubating for a maximum of 10 min; a forceps or a tip can be gently used to help releasing. The releasing solution is carefully aspirated and calli are washed with 10 mL of Medium F; released calli are then incubated in 10 mL of Medium G for another 4–6 weeks. Fresh Medium G is provided every week.
19. Large green calli are then briefly dried on a sterile filter paper, moved individually on petri dishes containing solid Medium H, and incubated in the same conditions used for potato propagation.
20. Calli are moved to fresh Medium H every 10–15 days; shoots usually emerge after 3 months of culture (*see* **Note 8**).
21. Mature shoots are moved to solid Medium I for rooting and plantlets moved to Medium A.

3.2 Analysis

For a fast screening of target specificity and preliminary mutation frequency, analysis can be made on protoplast or callus stage. After an additional 3–12 months, regenerated and elongated shoots with at least four leaves developed can be subjected to screening and characterization. A pool of protoplasts, one to a number of pooled calli, or a small leaf is sufficient for genomic DNA isolation using a method or kit of personal choice, either by extracting manually or by a high-throughput approach using, for example, a DNA extraction robot. The DNA can then be used for next-generation sequencing (NGS) or as a template for PCR amplification-based methods with primers spanning the target sites, preferably using a proofreading or high-fidelity DNA polymerase. The PCR-based methods that have been published are numerous, like probe-based digital PCR, high-resolution fragment analysis (HRFA) [1], loss of restriction enzyme site analysis, high-resolution melt analysis (HRM) [7], T7 endonuclease I, surveyor mismatch assay [8], targeted deep sequencing, Sanger sequencing followed by tracking of insertion and deletions and recombination events (TIDER) [9], or inference of CRISPR edits (ICE) [10] (for more examples and a comparison of different techniques, *see* ref. 11). The HRFA method is only useful for analyses where indels are the consequence of induced mutations, while the other methods can also be used for analyses of lines where base editing is the mode of mutations.

4 Notes

1. Potato in vitro propagation through auxiliary bud is in general successful on MS-based media for a wide range of cultivars, but adjustment of growing condition and media composition may be necessary.
2. This step is important to precondition leaf material and reduce starch granules that are reported to destabilize protoplasts during the extraction process.
3. It is important to use sharp and thin blades and glass petri dishes, in order to avoid tissue crashing and mashing that is detrimental to the transfection and regeneration procedure. Usually a blade should not be used for more than ten leaves; do not reuse blades.
4. Layering of protoplast on top of the sucrose solution without disturbing the interface is critical and may require some training before it can be successfully accomplished. For the centrifugation a swing-bucket rotor is required. The centrifuge breaks should be set to the lowest possible value (possibly zero), as breaking can disturb the layer of protoplast floating at the interface. Only healthy protoplasts will float at the interface, thus purifying them from debris and damaged cells.
5. The following mutation reagents have been shown to work with good efficiency: (a) up to 10 μg of highly pure plasmid DNA and (b) synthetically produced or in vitro-transcribed sgRNA preassembled with 5 μg Cas 9 per target according to the suppliers' instructions.
6. PEG final concentrations ranging from 12.5% up to 40% can be used. Higher PEG concentrations usually allow a higher transfection efficiency, but may have negative impact on regeneration. In our hands a genotype dependence has been noted. Therefore, it is recommended that a range of PEG concentrations are initially tested, e.g., 12.5%, 25%, and 40%. 40% PEG is difficult to dissolve as well as to sterilize. Very mild heating together with stirring can be applied if needed. Also, time for incubation can be varied, and an increase up to 30 min for incubation can be tested.
7. Protoplast density is crucial to initiate the first cell division. Higher density can be tested, if no cell division is observed at 5×10^4 protoplasts/mL.
8. Before shoot differentiation the calli could turn brown; this is cultivar dependent, but it might not affect regeneration. Shoots can continue to be regenerated up to or even beyond 1 year.

Acknowledgments

The authors thank the Mistra Biotech program supported by the Swedish Foundation for Strategic Environmental Research and Lyckeby Research foundation for financial support.

References

1. Andersson M, Turesson H, Nicolia A et al (2017) Efficient targeted multiallelic mutagenesis in tetraploid potato (*Solanum tuberosum*) by transient CRISPR-Cas9 expression in protoplasts. Plant Cell Rep 36:117–128
2. Nicolia A, Proux-Wéra E, Åhman I et al (2015) Targeted gene mutation in tetraploid potato through transient TALEN expression in protoplasts. J Biotechnol 204:17–24
3. Andersson M, Turesson H, Olsson N et al (2018) Genome editing in potato via CRISPR-Cas9 ribonucleoprotein delivery. Physiol Plant 164:378–384
4. Sawai S, Ohyama K, Yasumoto S et al (2014) Sterol side chain reductase 2 is a key enzyme in the biosynthesis of cholesterol, the common precursor of toxic steroidal glycoalkaloids in potato. Plant Cell 26:3763–3774
5. Clasen BM, Stoddard TJ, Luo S et al (2015) Improving cold storage and processing traits in potato through targeted gene knockout. Plant Biotechnol J 14:169–176
6. Nadakuduti SS, Buell CR, Voytas DF et al (2018) Genome editing for crop improvement – applications in clonally propagated polyploids with a focus on potato (Solanum tuberosum L.). Front Plant Sci 9:1607
7. Wittwer CT, Reed GH, Gundry CN et al (2003) High-resolution genotyping by amplicon melting analysis using LCGreen. Clin Chem 49:853–860
8. Vouillot L, Thélie A, Pollet N (2015) Comparison of T7E1 and surveyor mismatch cleavage assays to detect mutations triggered by engineered nucleases. G3 5:407–415
9. Brinkman EK, Kousholt AN, Harmsen T et al (2018) Easy quantification of template-directed CRISPR/Cas9 editing. Nucleic Acids Res 46:e58
10. Hsiau T, Conant D, Rossi N et al (2019) Inference of CRISPR edits from Sanger trace data. BioRxiv. https://doi.org/10.1101/251082
11. Germini D, Tsfasman T, Zakharova VV et al (2018) A comparison of techniques to evaluate the effectiveness of genome editing. Trends Biotechnol 36:147–159

Chapter 13

The Double-Layer Method to the Genesis of Androgenic Plants in *Anemone coronaria*

Andrea Copetta and Marina Laura

Abstract

Homozygous lines occur for plant breeding programs and for studies about gene expression and genetic mapping and they can be derived from anther culture. In this chapter, the method to obtain androgenic plants from an ornamental cut flower, *Anemone coronaria* belonging to the Ranunculaceae family, is described. In this species, androgenic plants were obtained culturing anthers with responsive microspores in Petri dishes containing a double layer of substrate with specific composition. Moreover, thermic treatment has been applied to induce the switch from pollen development program to embryo development program. The method allows to produce both double-haploid plants from diploid mothers (*2n*) and di-haploid plants from tetraploid mothers (*4n*).

Key words Anemone, Anther culture, Heat shock, Breeding, Anomalous microspores

1 Introduction

In recent years, the production of homozygous lines has been occurring for plant breeding programs and for studies about gene expression and genetic mapping. In plant breeding, shortening the length of time required for line development, increases the rate of genetic gain and effective ways to develop new varieties that are adapted to current climates to minimize the effects of climate change [1]. Doubled haploid (DH) populations are produced by regenerating plants by the induction of chromosome doubling from pollen grains, which greatly shortens the line fixation stage because completely homozygous lines are produced immediately [2].

In vitro anther culture is a biotechnological method to obtain homozygous lines defined in androgenic plants which are widely applied for new hybrid achievement in horticultural, cereal, and fruit species [3]. The method involves (1) the application of a thermal shock that causes the switch from pollen development program to embryo development program, and (2) the creation

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_13,

of an environment suitable for the formation and development of embryos and young plantlets. Androgenic plants develop from responsive haploid microspores at a specific stage of development inside anthers stressed with low or high temperature that induce in microspores the switch from pollen development program to embryo development program [4]. Haploid microspores subjected to thermal pretreatment can generate callus or embryos and subsequently young plantlets that can be haploids or diploids (double haploids) for spontaneous DNA duplication [3]. Cultivars of *A. coronaria*, a herbaceous plant with a perennial underground organ, are cultivated for ornamental use, and for garden or cut flower production; are highly heterozygous; and show different levels of ploidy [5].

These allogamous plants show inbreeding depression symptoms [6] and then the traditional methods to produce homozygous lines are unsuccessful. The method described in this chapter was applied for the first time in anemones in 1977 [7] and subsequently implemented [8, 9]. Thus, we have applied anther culture to obtain haploid, double-haploid (from diploid mother), and di-haploid (from tetraploid mother) plants.

2 Materials

2.1 Plant Materials

1. Flower buds of 1–3 cm in length were harvested from the potted field-grown *Anemone coronaria* plants, during the period from January to April.

2.2 Materials for Evaluation of Microspore Developmental Stage

1. 10 g/L of Acetocarmine acid stock: Under chemical hood, measure in a graduated cylinder 45 mL glacial acetic acid and make up the volume to 100 mL with deionized water. Dissolve 1 g of carmine powder in 100 mL glacial acetic acid 45% placing the beaker on a magnetic stirrer. Transfer the solution into a glass bottle. Store at room temperature.
2. Slides, coverslips, scalpel, and forceps (*see* **Note 1**).
3. Microscope (Leica Microsystems).

2.3 Materials for the Surface Sterilization of the Anemone Buds

1. Washing solution: 500 mL of water with 2 mL of liquid soap.
2. Sterilization solution: Aqueous sodium hypochlorite solution NaClO, prepared by mixing one part of commercial bleach (5% free chlorine) and four parts of deionized water (1% free chlorine) plus two drops of Tween 20 as surfactant. Prepare this solution always fresh.
3. Sterile deionized water in glass vessels with caps.
4. Sterile scalpel and forceps.
5. Sterile Petri dishes.

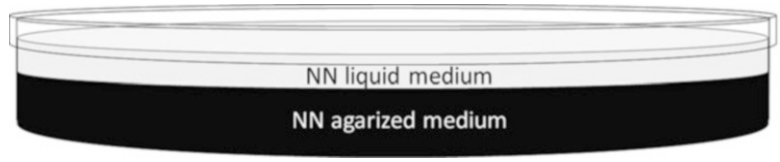


Fig. 1 The Petri dish (Ø 6 cm) contained a double-layer medium: in black, solid lower layer medium (with agar), and in grey, liquid upper layer medium

2.4 Media Preparation for Anther Culture

1. Anther culture medium 1: Semisolid substrate for lower layer (LL medium—*see* Fig. 1): 1.326 g Nitsch and Nitsch (NN) salts and vitamins [10] (*see* **Note 2**), 3% sucrose (*see* **Note 3**), 1% activated charcoal (AC), and 0.8% technical agar (pH 5.7). Dissolve NN salt and vitamin powder and sucrose in 500 mL of deionized water; while stirring the water, add the powder and sucrose and stir until complete dissolution; bring the medium to final volume (600 mL) filling deionized water and adjust the pH of solution using NaOH 1 M or HCl 1 M. Add 0.6 g of AC and 4.8 g of technical agar.
2. Anther culture medium 2: Liquid substrate for upper layer (UL medium—*see* Fig. 1): 0.884 g Nitsch and Nitsch (NN) salts and vitamins [10], 3% sucrose (pH 5.7). Dissolve NN salt and vitamin powder and sucrose in 300 mL of deionized water; while stirring the water, add the powder and sucrose and stir until complete dissolution; bring the medium to final volume (400 mL) filling deionized water and adjust the pH of solution using NaOH 1 M or HCl 1 M (*see* **Note 4**).
3. Sterilize the two media (LL and UL medium) at 121 °C, 1 atm, for 20 min.
4. Dispense LL medium into plastic Petri dishes (Ø 6 cm): Using a pipettor, pour 6 mL of LL medium per each plastic Petri dish. Let LL medium cool and solidify (*see* **Note 5**).
5. Incubators at 33 and 23 °C.

2.5 Medium Preparation for Embryo and Plantlet Culture

1. Glass jars with cap.
2. Develop medium: Half-strength MS salts [11] (*see* **Note 6**), MS vitamins, 3% sucrose, 1% activated charcoal (AC), and 0.8% agar (pH 5.7). To prepare medium, use 500 mL of MS salt stock solution, add 1 mL of vitamin stock solution and 30 g/L of sucrose, and bring the medium to the final volume (1 L) of filling water. Heat and stir the medium until the agar has completely dissolved; dispense about 62.5 mL medium for each glass culture vessel (preparing 16 culture vessels of 500 mL capacity for each liter of medium).
3. Sterilize the medium at 121 °C, 1 atm, for 20 min and allow the medium to cool and solidify prior to plant inoculation.
4. Incubator at 18 °C.

3 Methods

3.1 Evaluation of Microspore Development Stage

1. Select flower buds of 1–3 cm in length (*see* **Note 7**) with a hooked flower stalk with the tepals held within the floral bracts (Fig. 2). Reassess these morphological indicators for each cultivar and plant growth. Excise the bud from potted field-grown plants, keeping part of the stem (*see* **Note 8**).
2. Remove the bracts, move the tepals (Fig. 3), and detach at least five anthers from different areas of the buds (*see* **Note 9**) and place them on a microscope slide.
3. Add two or three drops of acetocarmine acid solution on the slide, close with a coverslip, and crush the anthers by pressing the coverslip with a bottom of the pencil. Wait for a few minutes for the dye to stain the microspores and observe the slide under a microscope (Fig. 4). Anthers containing microspores at the uninucleate stage are the best source of androgenic embryos (*see* **Note 10**).

3.2 Flower Bud Sterilization

1. Rinse the flower buds in the washing solution, with stirring for 10 min.
2. Surface-sterilize the flower buds, with their bracts removed, immersing them in the sterilization solution; stir for 20 min; and rinse the buds twice with sterile distilled water for 10 min. The previous procedure and the following steps should be performed in sterile conditions in a laminar flow hood.

Fig. 2 *A. coronaria* flower bud with a hooked flower stalk, with the tepals held within the floral bracts

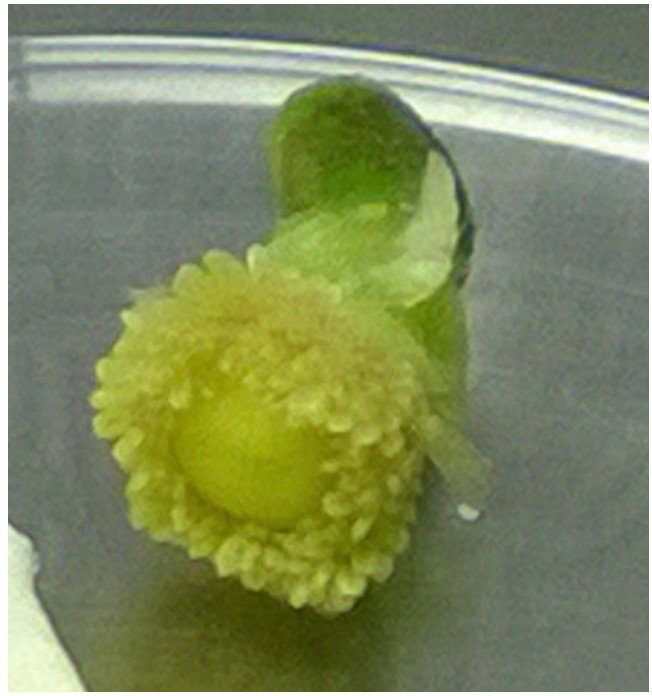


Fig. 3 *A. coronaria* flower bud with removed bracts and tepals shows a crown of pale immature anthers

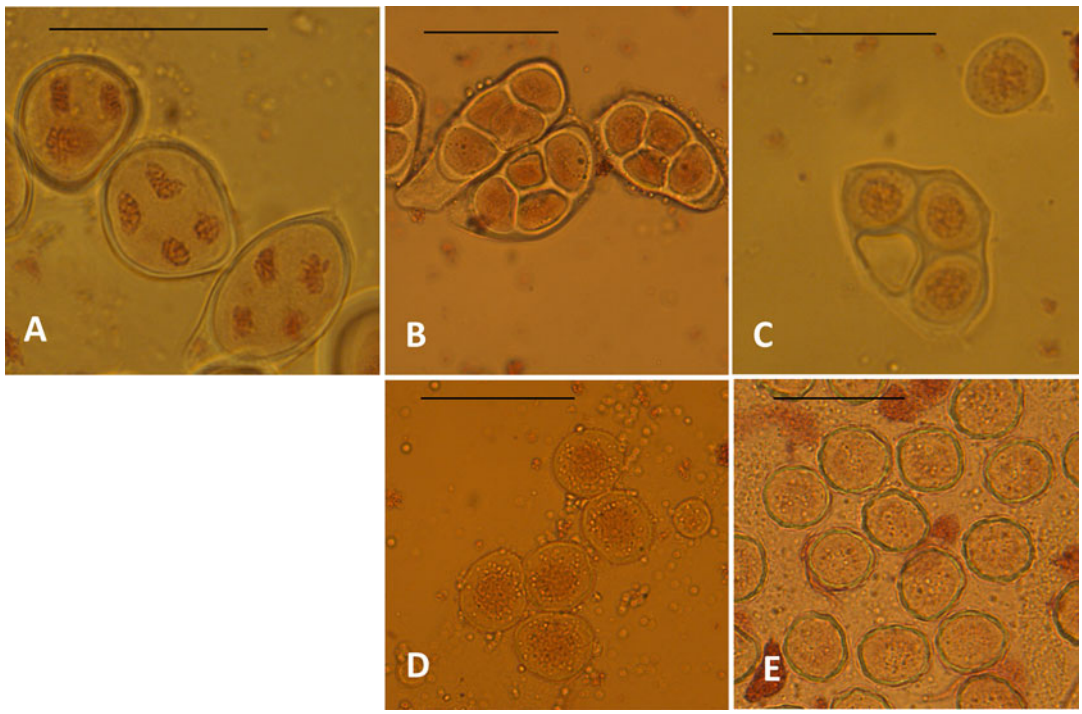


Fig. 4 Different stages of development of microspores stained with acetocarmine solution. (**a**) Immature tetrads in which four microspores are forming. (**b**) Well-formed tetrads. (**c**) Mature tetrad releasing a mature microspore. (**d**) Mature microspores with thin wall. (**e**) Ripening pollens with translucent and ornamented walls. Bars = 50 μm

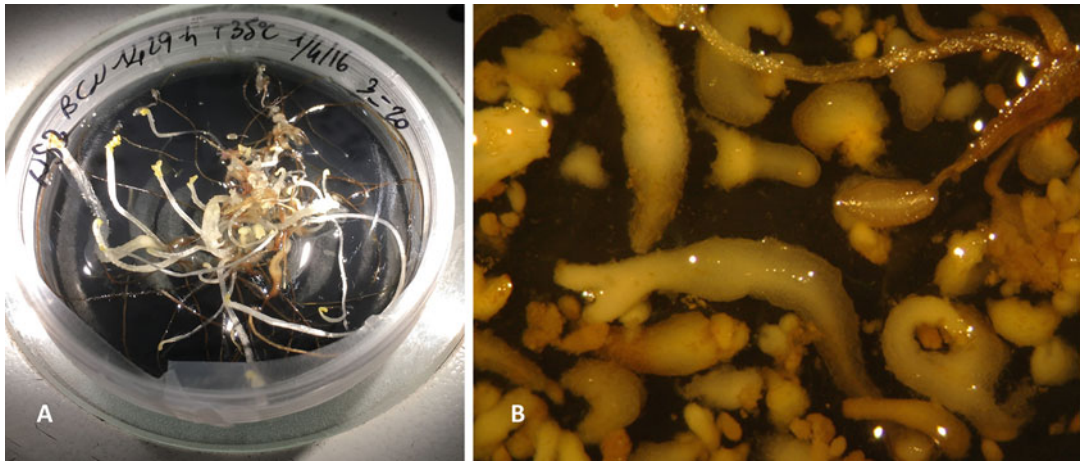


Fig. 5 Plantlets and embryos derived from anther culture. (**a**) Etiolated plantlets derived from 2-month-old anther culture. (**b**) Embryo-like structures and embryos (with cotyledon primordia and a well-developed shoot and root) at different stages of development

3.3 Anther Culture

1. In a sterile Petri dish, remove the tepals from the flower bud, excise the exposed anthers by cutting the filament with forceps or with tip scalpel, and distribute them in Petri dishes (Ø 6 cm) with LL medium (*see* **Note 11**).
2. Add 4 mL UL liquid medium in each Petri dish (Ø 6 cm) on LL medium to allow the floating of the anthers and seal with Parafilm® carefully (*see* **Note 12**).
3. Incubate the Petri dishes with the anthers at 33 ± 1 °C (pre-treatment stress) in complete darkness for 5 days (*see* **Note 13**).
4. Transfer the Petri dishes with the anthers at 23 ± 1 °C in the dark.
5. After about 2 months, the embryo-like structures and plantlets emerged through the wall of the anther pollen sac (Fig. 5) (*see* **Note 14**).

3.4 Embryo and Plantlet Culture

1. Transfer embryos and plantlets derived from the anther culture in flasks containing developed medium (Fig. 6) and incubate at 18 °C with a 16-h/8-h light/dark cycle at 3500 lux for growing.
2. Subculture on the fresh medium at every 30-day interval to grow and develop the plants that will form leaves, bulb, and roots (*see* **Note 15**).

3.5 Acclimatization

1. The acclimatization of the plants can be done in autumn of the same year in which they were formed.
2. Transfer the rooted plants with enlarged tuber (Ø > 1 cm) in a peat-agriperlite substrate (1:1) under unheated greenhouse

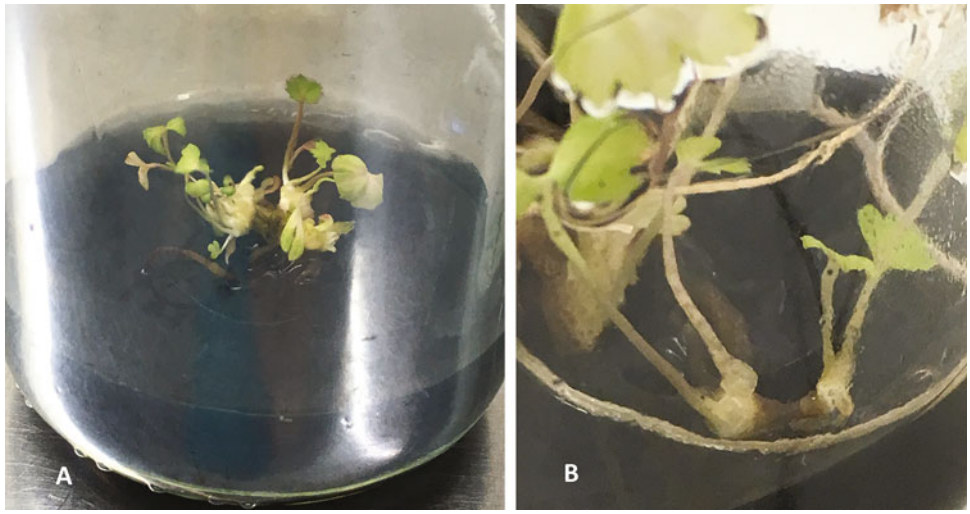


Fig. 6 In vitro culture of androgenic plant of *A. coronaria*. (**a**) Well-formed plant with leaves, bulb, and roots. (**b**) Plant with secondary embryogenesis, new secondary bulbs, and shoot developed from root

(Fig. 7a) under fog or mist system (10 s every 30 min) for the first 7 days, then decrease the humidity gradually, and then provide water properly. The lighting can be supplied by natural light (*see* **Note 16**). When the anemone plants begin to grow, transplant them in pots containing the previously mentioned mixture (Fig. 7b).

4 Notes

1. One forceps with very thin tips is necessary to take stamens with anthers.
2. The culture media should be prepared using commercial salt plus vitamin combination in order to standardize the anther culture.
3. In tests carried out without sucrose or by replacing sucrose with maltose, no androgenic plants were obtained.
4. The quickest way to prepare the two media (LL and UL) is the following: dissolve 2.21 g of NN salt and vitamin powder and 30 g of sucrose in 700 mL of deionized water; while stirring the water, add the powder and sucrose and stir until complete dissolution; bring the medium to final volume (1 L) filling deionized water and adjust the pH. Split the solution in two 1 L glass bottles: one with 400 mL (UL medium) and the other with 600 mL of solution. In the bottle with 600 mL of solution add AC and agar to obtain LL medium.

Fig. 7 Acclimatized androgenic plant of *A. coronaria*. (**a**) In alveolus container. (**b**) Pot culture 2 months after acclimatization

5. Once prepared, the Petri dishes containing the LL medium and the liquid medium UL in the bottle can be stored in the refrigerator and must be moved to room temperature at least 1 h before using them.
6. Dissolve MS powder containing micro- and macroelement complex (43.0 g) in 10 L of deionized water to prepare a stock solution. Dissolve commercial powder containing 25.80 g mixed vitamins to prepare 250 mL 1000× vitamin stock solution in deionized water and stir until completely dissolved. Use 1 mL vitamin stock solution for each liter of culture medium. The basal medium used to rescue anther culture-derived embryos and regenerants was composed of half-strength MS salts, MS vitamins, 10 g/L AC and 30 g/L sucrose in 8 g/L agar.
7. The buds must be treated as soon as they are picked; if it is not possible, they can be stored at 4 °C. Several studies and our experience indicate that the size and the hook morphology of the stalk are good indicators for the presence of immature pale-color anthers with microspores at the correct stage of development. In the middle of spring and with increasing temperatures, the buds tend to anticipate the maturation of the anthers; therefore, uninucleate microspores are more present in buds with a diameter of less than 1 cm.

8. Keeping part of the stem facilitates the collection of the anthers: with forceps the bud is held still and with forceps with thin tips the anthers are removed.
9. This operation allows to verify the presence of microspores at the correct stage of maturation. Each anemone bud can provide from 250 to 400 anthers, whose maturation is asynchronous and centripetal, and therefore, to obtain anthers all at the same stage of maturation from an only bud is impossible. However, the microspores in an anther developed homogenously are at the same stage of development [12].
10. In anemones, only anthers with pollen at uninucleate stage produced androgenic plants. Uninucleate stage is present in mature tetrads and in microspores with thin wall, not translucent and without ornamentations.
11. Put 15–20 anthers from different zones of the same bud in each Petri dish. If during this operation the anthers are damaged it is not a problem because damage to the anther wall increases and facilitates the contact between the microspores and the growth substrate.
12. Seal the Petri dishes with at least three turns of Parafilm; during this operation always keep the Petri dishes horizontal to prevent the anthers from sticking to the lid. After thermal stress, the anthers sticking to the lid darken and are no longer viable.
13. Heat stress ($T > 30$ °C) is the most effective for the formation of androgenic plantlets in Ranunculaceae, but for some varieties, cold stress (first at 7 °C for 7 days or 5 °C for 4 days, and then at 23 ± 1 °C) induces the formation of a greater number of androgenic plantlets.
14. In a single Petri dish, it is possible to observe embryo-like structures and plantlets at different stages of development: free or arising from anther embryos (globular, heart, or torpedo stage); and plantlets with root, shoot, cotyledon primordia, microtuber, and sometimes leaves (Fig. 5).
15. Some embryos and plantlets manifested a propensity for secondary embryogenesis (Fig. 6b): the growing roots form secondary bulbs from which small seedlings develop.
16. For ploidy-level analysis, leaves by flow cytometry or root tips from in vitro- or in vivo-grown plants for the chromosome number count can be used.

Acknowledgment

The authors wish to thank Biancheri Creazioni s.r.l. for partially funding the research.

References

1. Atlin GN, Cairns JE, Das B (2017) Rapid breeding and varietal replacement are critical to adaptation of cropping systems in the developing world to climate change. Glob Food Secur 12:31–37. https://doi.org/10.1016/j.gfs.2017.01.008
2. Mishra R, Rao GJN (2016) In-vitro androgenesis in rice: advantages, constraints and future prospects. Rice Sci 23:57–68. https://doi.org/10.1016/j.rsci.2016.02.001
3. Germanà MA (2011) Anther culture for haploid and double haploid production. Plant Cell Tissue Organ Cult 104:283–300. https://doi.org/10.1007/s11240-010-9852-z
4. Li J, Huang Q, Sun M, Zhang T, Li H, Chen B, Xu K, Gao G, Li F, Yan G, Qiao J, Cai Y, Wu X (2016) Global DNA methylation variations after short-term heat shock treatment in cultured microspores of *Brassica napus* cv. Topas. Sci Rep 6:38401–38412. https://doi.org/10.1038/srep38401
5. Laura M, Allavena A (2007) *Anemone coronaria* breeding: current status and perspectives. Eur J Hortic Sci 72(6):241–247
6. Horovitz A, Galil J, Zohary D (1975) Biological flora of Israel. VI. *Anemone coronaria* L. Isr J Bot 24:26–41
7. Johansson L, Eriksson T (1977) Induced embryo information in anther cultures of several Anemone species. Physiol Plant 40:172–174
8. Laura M, Safaverdi G, Allavena A (2006) Androgenetic plants of *Anemone coronaria* derived through anther culture. Plant Breed 125:629–634
9. Copetta A, Dei F, Marchioni I, Cassetti A, Ruffoni B (2018) Effect of thermal shock in the development of androgenic plants of *Anemone coronaria* L.: influence of genotype and flower parameters. Plant Cell Tissue Organ Cult 134:55–64. https://doi.org/10.1007/s11240-018-1399-4
10. Nitsch JP, Nitsch C (1969) Haploid plants from pollen grains. Science 163:85–87
11. Murashige T, Skoog F (1962) A revised medium for rapid growth and bioassays with tobacco tissue culture. Physiol Plant 15:478–487
12. Dhooghe E, Grunewald W, Reheuln D, Goetghebeur P, Van Labele MC (2012) Floral characteristics and gametophyte development of *Anemone coronaria* L. and *Ranunculus asiaticus* L. (Ranunculaceae). Sci Hortic 138:73–80. https://doi.org/10.1016/j.scienta.2011.10.004

Chapter 14

Ploidy Modification for Plant Breeding Using In Vitro Organogenesis: A Case in Eggplant

Edgar García-Fortea, Ana García-Pérez, Esther Gimeno-Páez, Marina Martínez-López, Santiago Vilanova, Pietro Gramazio, Jaime Prohens, and Mariola Plazas

Abstract

The use of antimitotic agents such as colchicine has been common to obtain polyploid organisms. However, this approach entails certain problems, from its toxicity to the operators for being carcinogenic compounds to the instability of the individuals obtained, and the consequent reversion to its original ploidy because the individuals obtained in most cases are chimeric. In vitro culture allows taking advantage of the full potential offered by the cellular totipotence of plant organisms. Based on this, we present a new in vitro culture protocol to obtain polyploid organisms using zeatin riboside (ZR) and eggplant as a model organism. Flow cytometry is used to identify tetraploid regenerants. The regeneration of whole plants from the appropriate tissues using ZR allowed developing polyploid individuals in eggplant, a crop that tends to be recalcitrant to in vitro organogenesis. Thanks to the use of the polysomatic pattern of the explants, we have been able to develop a methodology that allows to obtain stable non-chimeric polyploid individuals from organogenic processes.

Key words Plant tissue culture, Polysomatic pattern, Polyploid, Flow cytometry, Zeatin riboside, *Solanum melongena*

1 Introduction

Obtaining polyploids is a strategic objective for many seed and breeding companies. They also have great importance in other sectors, such as ornamental plants, since polyploids tend to have larger and more striking organs [1] and triploid individuals, which are sterile, have more durable flowers [2]. Another sector where polyploids could raise interest is biomedicine and pharmacology, since they may have higher levels of biosynthesis and accumulation of bioactive compounds [3]. The crossing between a tetraploid plant and a diploid allows obtaining triploid offspring, which may be completely or partially sterile. This type of organisms presents agronomic traits of great value, such as the complete or partial

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_14,

absence of seeds, adding a great value as seedless fruits are highly appreciated by the consumer.

One of the most widely used methods to develop polyploids is through the application of antimitotic agents, such as colchicine, to induce genome duplication in a variable proportion of cells from embryos, young plants, or adult plant tissues. In addition to posing a risk to the operators, since most of the antimitotic agents are carcinogenic, the use of these chemicals for polyploid development is largely inefficient [4]. Frequently, the results obtained are mixoploids or chimeric that frequently revert to the original diploid status [5].

In order to improve the efficiency of polyploid production, we have developed an eggplant protocol to obtain polyploids without using antimitotic agents [4]. For this, we rely on the polysomatic pattern presented by the different tissues of the plant, which can be detected by means of flow cytometry. In this way, in tissues such as hypocotyl or cotyledons there are different cell populations with naturally diverse ploidy levels [6, 7]. This is a mechanism used by plants during the earliest stages of their development to achieve faster and more efficient cell expansion in terms of energy cost, allowing a fast growth and elongation of the seedling in a very short period.

Therefore, if organogenic processes can be induced in these polysomatic tissues, with a high percentage of probability, it will be possible to obtain polyploid plants. Contrarily to polyploidy plants obtained by antimitotic drugs, these polyploidy plants are generally stable and non-chimeric and thus do not revert to the diploid state. For this, it will be necessary to cultivate hypocotyls or cotyledons, induce the formation of shoots in these tissues, acclimatize them, and evaluate their ploidy level [4].

2 Materials

Prepare all solutions and culture media using ultrapure water (prepared by purifying deionized water, to attain a sensitivity of 18 MΩ cm or lower at 25 °C), or sterile distilled water (autoclaved for 20 min at 121 °C) (*see* **Note 1**). Prepare all reagents and culture media at room temperature and store them at 4 °C. In the case of hormone stocks freeze at −20 °C.

2.1 Solutions

Nuclei extraction buffer: Tris–HCl (15 mM), Na_2EDTA (2 mM), spermine (0.5 mM), KCl (80 mM), and NaCl (20 mM). Add approximately 175 mL distilled H_2O. Leave on the magnetic stirrer until all components dissolve completely. Adjust the pH of the mixture to 7.5 with 1 M HCl. Now add the 2-mercaptoethanol (15 mM) and the Triton X-100 (0.1%) in the gas extraction hood. Leave it shaking in the hood for at least

30 min to fully homogenize the Triton. Once a homogeneous mixture has been obtained, make up to 200 mL with distilled H_2O and store in the refrigerator at 4 °C.

70% Ethanol solution: Prepare a volume of 729 mL of 96% ethanol and bring it up to 1000 mL with sterile distilled water using a test tube (*see* **Note 2**).

20% Bleach solution: Prepare a volume of 200 mL of commercial bleach (37 g/L $HClO_3$) and bring up to 1000 mL with sterile distilled water using a test tube. Add two drops of Tween20 (*see* **Note 3**).

Zeatin riboside (ZR) stock (1 g/L): 20 mg of hormone is dissolved in 2 mL of 1 M NaOH (*see* **Note 4**). Once dissolved, sterile distilled water is added until a total volume of 20 mL is reached. In a laminar flow cabinet, the hormonal stock solution is filtered with a 0.22 μm filter using a plunger syringe and it is distributed in a 2 mL sterile Eppendorf tube (*see* **Note 5**). Store the stock at −20 °C.

Indole butyric acid (IBA) stock (1 g/L): 20 mg of hormone is dissolved in 2 mL of 1 M NaOH (*see* **Note 4**). Once dissolved, sterile distilled water is added until a total volume of 20 mL is reached. In a laminar flow cabinet, the hormonal stock solution is filtered with a 0.22 μm filter using a plunger syringe and it is distributed in a 2 mL sterile Eppendorf tube (*see* **Note 5**). Store the stock at −20 °C.

2.2 Culture Media

E0 medium (germination): To prepare 1 L of culture medium, weigh using a precision balance 2.2 g of MS vitamin salts and 15 g of sucrose, place them in a beaker with 1 L of distilled water, and use a magnetic stirrer to homogenize the mixture. Once this is done adjust the pH to a value of 5.8 (*see* **Note 6**) with the help of a pH meter. Once the pH is adjusted, add 7 g of Gelrite™ (*see* **Note 7**), mix everything well in an autoclavable bottle, close the cap (*see* **Note 8**), and autoclave it for 20 min at 121 °C. Once autoclaved, cool to a temperature between 40 and 50 °C, pour into petri dishes inside a laminar flow cabinet, and let it dry (*see* **Note 9**).

E6 medium (organogenesis induction): To prepare this medium the same steps indicated in the previous section (1 L) are followed with the differences explained below. Remove the medium from the autoclave and once it has tempered, add 2 mL of ZR (1 g/L of ZR stock) in the laminar flow cabinet. The mixture is then shaken vigorously. After this the mixture is poured into petri dishes and left to solidify.

R2 medium (root induction): To prepare this medium the same steps indicated for medium E0 (or E2) are followed with the differences explained below. When the medium is removed

from the autoclave and tempered, 1 mL of IBA [1 g/L of indole butyric acid (IBA) stock] is added to it in the laminar flow cabinet and the solution is shaken vigorously. After this, the solution is poured into petri dishes and left to solidify.

2.3 Plant Material

Seeds of good quality (high germination and pathogen free) need to be sterilized for use in the in vitro culture steps. The detailed procedure is indicated in Subheading 3.2 of this document. In our case, we used seeds of one accession of eggplant (MEL3) kindly provided by the germplasm bank of Universitat Politècnica de València (Valencia, Spain; FAO germplasm bank code: ESP026).

3 Methods

The methodology consists of five main stages (Fig. 1). First, the polysomatic pattern of the genotypes must be evaluated by flow cytometry in leaves, cotyledon, and hypocotyl to verify that there are enough polyploid cells to start the process. The next step is to cultivate the mixoploid explants to induce shoot formation that, after the rooting and acclimatization process, could give rise to polyploid plants. Finally, the ploidy of the regenerated plants must be checked again using flow cytometry.

3.1 Polysomatic Pattern Study

1. Using DAPI staining, the polysomatic pattern of tissues used in in vitro culture is evaluated. For this, explants of hypocotyl (1 cm), cotyledon (1 cm^2), and leaf (as a control; 1 cm^2) are processed. Each sample is cut with a scalpel blade in a petri dish together with 500 µL of nuclei extraction buffer. Several parallel longitudinal cuts are made and then the sample is rotated 90° to repeat the process until the extraction buffer becomes green.
2. Subsequently, the maximum possible volume of the resulting liquid is pipetted and filtered using CellTrics filters in a Sarstedt tube. Finally, 500 µL of the DAPI stain is added to the solution previously filtered.
3. The samples are processed in a cytometer (in our case we used a Cyflow® ploidy analyzer; Partec, Münster, Germany) with a gain of 417 (*see* **Note 10**) at a rate of ~2 µL/s for about 30 s.
4. The diploid peak of the control (true leaf) is set at an arbitrary fluorescence value of 50 units on the *X*-axis; consequently, the peak corresponding to the G2 phase is located at the arbitrary value of 100 units of the same axis.
5. After the analysis, the tissues (generally hypocotyl, cotyledon, or both) with higher proportion of tetraploid cells that could potentially undergo the process of organogenic induction and

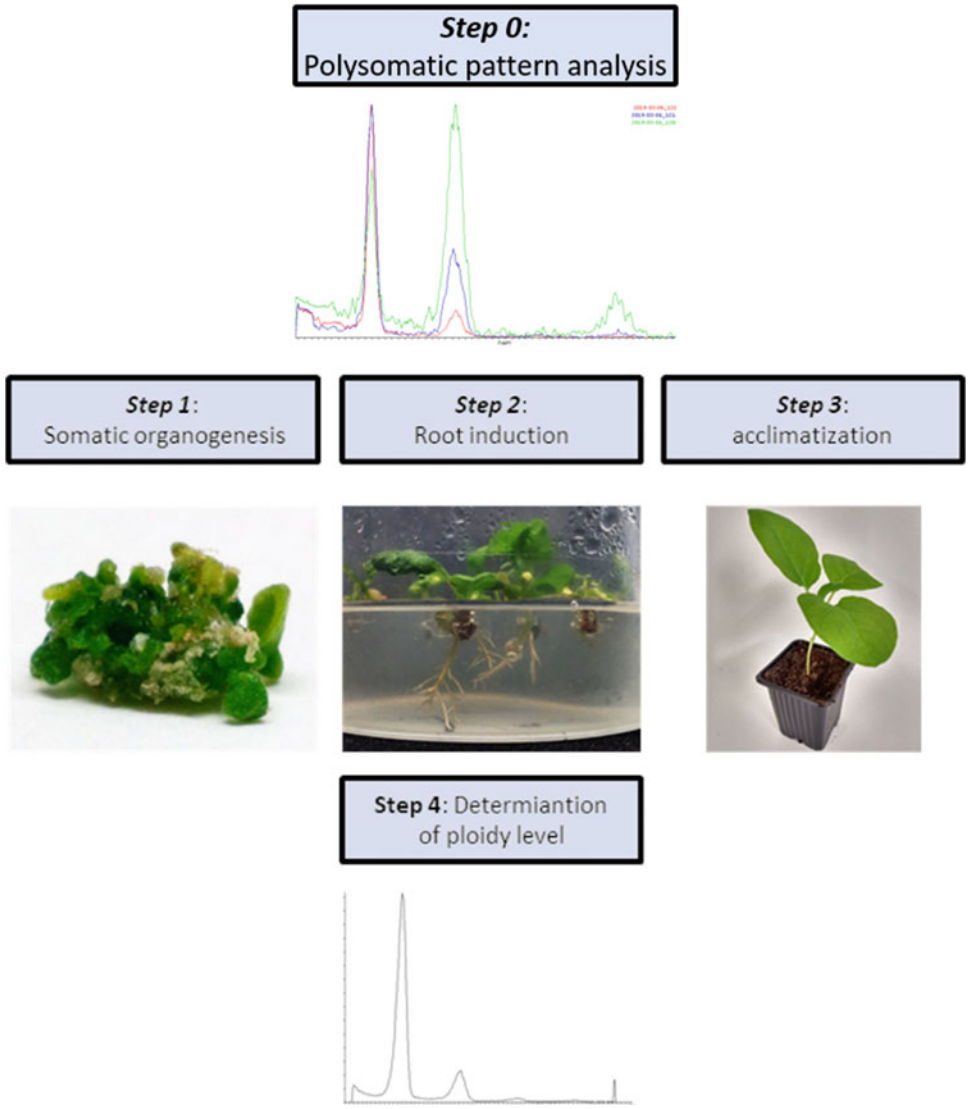


Fig. 1 Workflow of the production and identification of polyploid through in vitro culture. First the polysomatic pattern is studied. After that, organogenic processes are induced in polysomatic tissues, then root formation is induced, and finally the plants are acclimated. Finally, their level of ploidy is evaluated by flow cytometry

give rise to adult tetraploid plants are chosen. Usually, both cotyledon and hypocotyl tissue are good options. As can be seen in the example of Fig. 2, the cotyledon had between three- and fivefold more cells than the true leaf at the G2-phase peak. In the case of hypocotyl, the number of cells of the G2 peak was between seven and nine times greater than the peak of the true leaf. Both hypocotyl and cotyledon tissue showed a peak in the arbitrary fluorescence value of 200, a nonexistent peak in the analysis of the true leaf, which indicated that this excess of cells

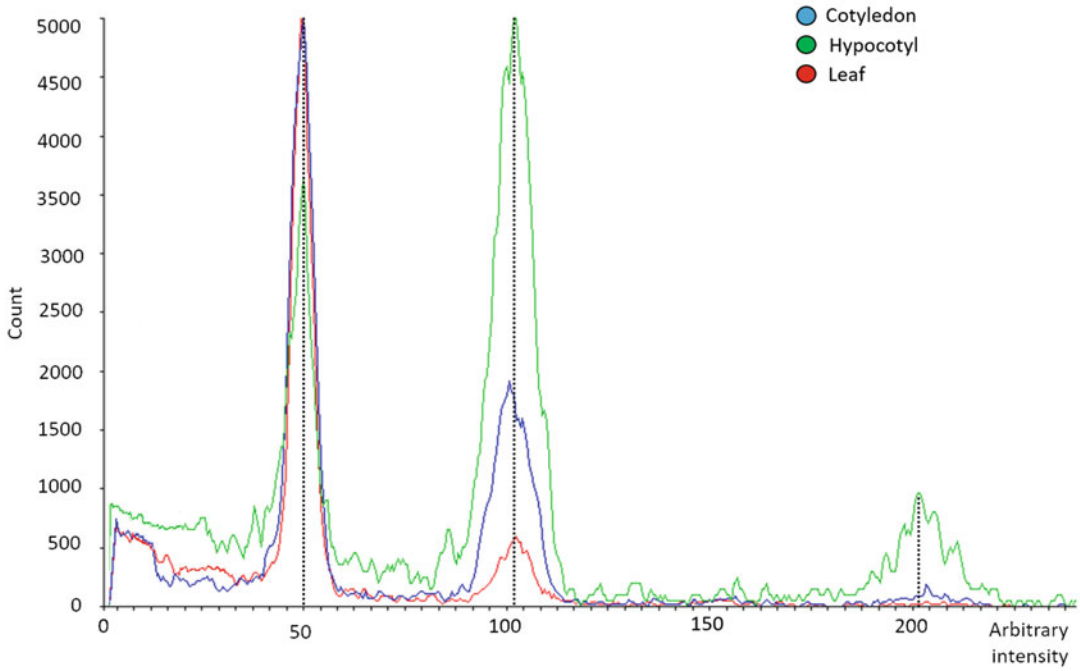


Fig. 2 Flow cytometry histogram in which the relative content of nuclear DNA in different tissues of eggplant is represented. The *X*-axis represents the fluorescence, which is proportional to the amount of DNA. The peak located at the value of 50 corresponds to the diploid nuclei in G1 phase. The *Y*-axis represents the number of nuclei analyzed

in the G2-phase peak of cotyledons and hypocotyls are tetraploid cells fully functional that were dividing. This results in a fluorescence peak in the value of 200.

3.2 In Vitro Polyploid Production

Once the tissue that is more likely to convert to polyploid is determined (*see* **Note 11**), seeds need to be germinated in sterile conditions. Seeds are sterilized using tea filters or muslin sachets.

1. In a laminar flow cabinet, a preliminary 70% ethanol wash is performed for 30 s.
2. Secondly, seeds are soaked in a 20% commercial bleach (with two drops of Tween20) for 10 min.
3. Finally, three washes are performed with sterile distilled water for 1 min each of them. Shake vigorously to wash bleach residues that would have remained in the seeds.
4. Subsequently, the seeds are germinated in petri dishes with E0 medium under sterile conditions, and then the dishes are incubated in the dark. Around 1 month is generally needed until the seedlings showed long hypocotyls and curled cotyledons (*see* **Note 12**).

5. E6 medium is prepared 3–4 days before the experimental session to rule out potential contamination events.
6. Under sterile conditions, five 1 cm long hypocotyl fragments are cultured in each plate with E6 (*see* **Note 13**) medium. Five 1 cm^2 cotyledon explants are also cultivated with two cuts, one in the distal part and another in the proximal part in each of the E6 medium petri dishes. The plates are then kept in a culture chamber at a temperature of 25 °C and photoperiod conditions of 16-h light and 8-h dark.
7. A minimum of 20 plates with five explants each is recommended for each tissue in order to obtain good yields in terms of regenerated polyploid plants.
8. When the buds that formed (*see* **Note 14**) in the surface of the explants have a size of around 0.5–1 cm or they have the size that allows to make an incision and individualize the plantlet (*see* **Note 15**), they are separated from the hypocotyl or cotyledon tissue and subcultured in R2 medium (*see* **Note 16**).
9. When the rooted plant has at least two leaves and a root system with at least two main roots and five secondary roots in each of them, the plants can be transferred to the growing substrate (generally any commercial growing substrate for vegetables is appropriate).
10. After transplant and acclimatization, a sample of 1 cm^2 of leaf is taken for the ploidy analysis as explained in the next section.
11. It is highly advisable to place an inverted plastic glass covering the plant (as a kind of mini greenhouse) for 2 weeks after transplant to prevent it from suffering or dying from dehydration. It is also advisable to humidify it daily with a nebulizer.

3.3 Ploidy Check with Flow Cytometry

This analysis is carried out using the same method as indicated in Subheading 3.1.

After evaluating each of the plants, a cytogram such as the one represented in Fig. 3a or b is observed. Figure 3a corresponds to a diploid plant, as it has the same profile as a conventional diploid eggplant used as a control. Depending on the equipment, we will adjust the gain with which the control sample is interrogated (in our case it has a value of 417), so that we place the peak in phase G1 at an arbitrary value equal to 50.

All those samples, which display a main peak at the value of 100 and a secondary peak at the value of 200, are tetraploid (Fig. 3b). Tetraploid plant peaks appear displaced in the cytogram because they contain twice as much DNA inside their cell nuclei. At this point in vitro tetraploid plants that most likely come from polyploid cells present in the starting tissue have been generated.

From this point the tetraploid plants will grow, and in the case of eggplant, they are generally fertile and able to produce seeds by selfing, which after germination give tetraploid plants. Therefore,

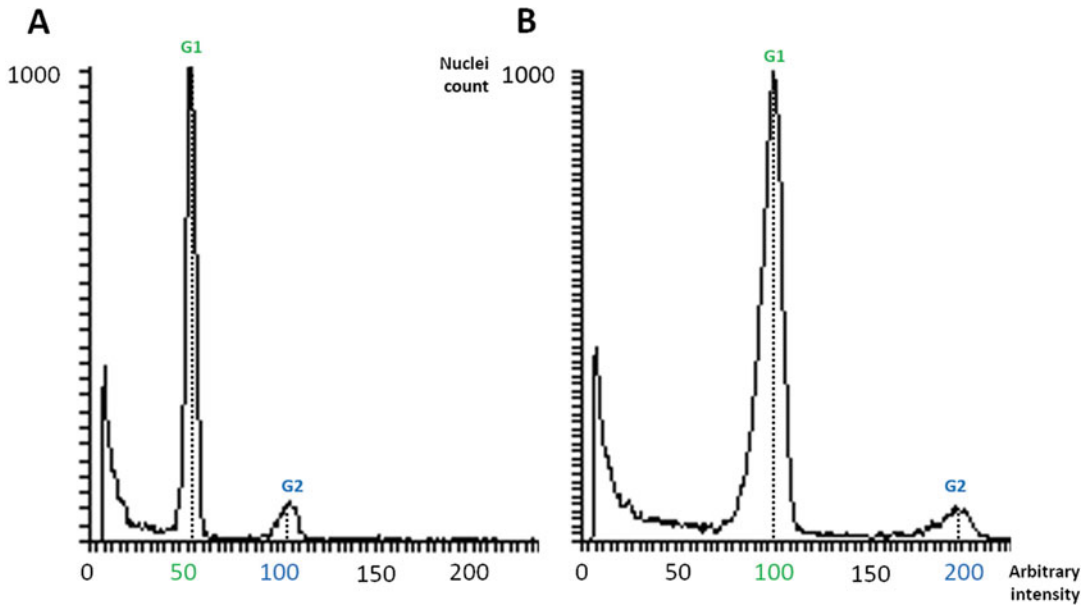


Fig. 3 Cytogram of a diploid eggplant plant (**a**). Cytogram of a tetraploid eggplant plant (**b**). The corresponding peak with the G1 cell phase is indicated in green; the G2 phase is indicated in blue. The *Y*-axis indicates the number of nuclei analyzed

with this method it is possible to not only generate stable non-chimeric tetraploid plants but also propagate them sexually by selfing and immortalize them as tetraploid materials. One of the added values of this methodology is that antimitotic agents that are very harmful to human health are not used.

4 Notes

1. The use of ultrapure water is not decisive, although it is recommended, but with sterile distilled water it is possible to perform the protocol.
2. You can work as if it were absolute ethanol, thus measuring a volume of 700 mL to prepare 70% ethanol solution.
3. Tween20 is a detergent that helps to penetrate the bleach into the tissue and perform a more efficient disinfection. Constant agitation also helps more effective disinfection.
4. In this case, ZR and IBA dissolve very well with 1 M NaOH, but this change depends on the growth regulator. It is something that must be checked if other growth regulators are used.
5. It is highly recommended not to fill the Eppendorf to the limit. When freezing, the liquid inside expands and the lid opens, losing sterility and making the stock useless. It is also highly recommended to seal it with parafilm.

6. It is better to adjust the pH to 5.9, as during the autoclaving process the sugars slightly acidify the culture medium; in this way we correct this effect.
7. The gelling agent should always be added after measuring the pH to prevent the lumps that form from damaging the pH meter electrode.
8. We should always leave the cap partially unscrewed to prevent the bottle from exploding because of pressure during the autoclaving process.
9. About 20 min should be enough for the Gelrite™ to solidify. A good practice is to wait for the drops of condensation on the sides of the petri dishes to disappear to avoid the accumulation of water.
10. This gain value is only valid for eggplant and must be calibrated for each species that is analyzed as it depends on the size of its genome.
11. The yield on the number of polyploid regenerated plants depends largely on the organogenic competence of its polyploid cells. This means that we can find species in which this yield is much higher than in eggplant and others in which we may not obtain any polyploid regenerant.
12. It is advisable to leave the petri dishes with the germinated seeds in photoperiod conditions (16-h light/8-h dark) for a day before using the tissues in **step 6** of subheading 3.2 (In vitro polyploids production). This way the cotyledons will have just roll out, and it will be much more comfortable to work with them.
13. This medium is indicated to induce the organogenic processes in eggplant. For other species it will be necessary to use the corresponding medium that does not necessarily have the same composition than E6.
14. In eggplant, this will start happening after a minimum of 30 days after placing the explants in the E6 medium.
15. In case the explants have not elongated enough after a period of 40 days they can be subcultured in E0 medium for a week and then re-subcultured in E0 + gibberellic acid (GA_3) 1 mg/L until the elongation of the shoots is achieved.
16. It is advisable to prepare this medium plastic pots with a membrane filter in the lid to allow gaseous exchange (Microbox containers O118/120 + OD118/120 #10 (G), SAC02, Nevele, Belgium), so that the plant can reach a good size that will allow us to take a sample for the subsequent cytometric analysis, as well as facilitate the acclimatization of the plant.

Acknowledgment

This research was funded by the Spanish Ministerio de Ciencia, Innovación y Universidades, Agencia Estatal de Investigación and Fondo Europeo de Desarrollo Regional (grant RTI-2018-094592-B-100 from MCIU/AEI/FEDER, UE) and by Universitat Politècnica de València. The Spanish Ministerio de Educación, Cultura y Deporte, funded a predoctoral fellowship granted to Edgar García-Fortea (FPU17/02389). The Generalitat Valenciana and Fondo Social Europeo funded a postdoctoral fellowship granted to Mariola Plazas (APOSTD/2018/014).

References

1. Razdan Tiku A, Razdan MK, Raina SN (2014) Production of triploid plants from endosperm cultures of *Phlox drummondii*. Biol Plant 58:153–158. https://doi.org/10.1007/s10535-013-0372-7
2. Wang X, Cheng Z-M, Zhi S (2015) Breeding triploid plants: a review. Czech J Genet Plant Breed 52:41–54. https://doi.org/10.17221/151/2015-CJGPB
3. Gao SL, Zhu DN, Cai ZH, Xu DR (1996) Autotetraploid plants from colchicine-treated bud culture of *Salvia miltiorrhiza* Bge. Plant Cell Tissue Organ Cult 47:73–77. https://doi.org/10.1007/BF023189684
4. García-Fortea E, Lluch-Ruiz A, Pineda-Chaza BJ, García-Pérez A, Bracho-Gil JP, Plazas M, Gramazio P, Vilanova S, Moreno V, Prohens J (2020) A highly efficient organogenesis protocol based on zeatin riboside for in vitro regeneration of eggplant. BMC Plant Biol 20:6. https://doi.org/10.1186/s12870-019-2215-y
5. Lehrer JM, Brand MH, Lubell JD (2008) Induction of tetraploidy in meristematically active seeds of Japanese barberry (*Berberis thunbergii var. atropurpurea*) through exposure to colchicine and oryzalin. Sci Hort 119:67–71. https://doi.org/10.1016/j.scienta.2008.07.003
6. Gilissen LJW, van Staveren MJ, Creemers-Molenaar J, Verhoeven HA (1993) Development of polysomaty in seedlings and plants of *Cucumis sativus* L. Plant Sci 91:171–179. https://doi.org/10.1016/0168-9452(93)90140-U
7. Smulders MJM, Rus-Kortekaas W, Gilissen LJW (1994) Development of polysomaty during differentiation in diploid and tetraploid tomato (*Lycopersicon esculentum*) plants. Plant Sci 97:53–60. https://doi.org/10.1016/0168-9452(94)90107-4

Chapter 15

Assembly of TALEN and mTALE-Act for Plant Genome Engineering

Aimee A. Malzahn and Yiping Qi

Abstract

Transcription activator-like effector (TALE) is a DNA-binding domain that can be paired with a nuclease to create DNA double-strand breaks, or with an effector protein to alter gene transcription. The ability to precisely alter plant genomes and transcriptomes has provided many insights into gene function and has recently been utilized for crop improvement. Easy design and construction of TALE make the tool more accessible to a variety of researchers. Here, we describe two TALE-based systems: transcription activator-like effector nucleases (TALEN), for creating targeted mutations in a gene of interest, and multiplex TALE activation (mTALE-Act), for activating one or a few genes of interest at the transcription level. Assembly of these tools is based on Golden Gate cloning and Gateway recombination, which are cost-effective and streamlined cloning methods.

Key words Transcription activator-like effector, TALEN, mTALE-Act, Genome editing, Golden gate, CRISPR

1 Introduction

Genome engineering utilizes sequence-specific nucleases (SSNs) to create genetic mutations or applies synthetic transcriptional activators or repressors to alter transcription within the genome. It relies on a series of tools for genome editing and transcriptional regulation. The genome editing technology started with meganucleases, zinc-finger nucleases (ZFN), and then transcription activator-like effector nucleases (TALEN), and now CRISPR-Cas systems are the most popular choices due to their ease of use and high efficiency [1, 2]. Harnessing DNA-binding features of ZFN, TALEN, and CRISPR-Cas systems, transcriptional regulation tools can be further developed. In this chapter, we describe methodology to construct two genome engineering tools: TALEN and mTALE-Act (multiplex TALE-Activator).

Transcription activator-like effectors (TALEs) are produced by *Xanthomonas*, a bacterial plant pathogen which excretes TALEs

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_15,

during infection to alter host gene transcription [3]. The discovery that TALEs recognize DNA targets, with one TAL repeat domain for one DNA nucleotide, was an important one in biology [4, 5]. TALEs have a central DNA-binding domain which is composed of many repeats. These repeats are almost identical, except in the 12th and 13th amino acid positions which is referred to as a repeat variable di-residue (RVD). Each RVD binds to a specific nucleotide: HD = C, NG = T, NI = A, and NN = G and A. Decoding the secrets of DNA binding by TALEs immediately shed lights to de novo engineering of DNA-binding domains for any DNA sequences of interest, which led to the development of TALEN [6]. TALEN is based on two TALE-FokI monomers. Directing a pair of such monomers to proximal sequences on both strands of DNA will create a DNA double-stranded break due to FokI dimerization (Fig. 1a). TALEN was successfully used in genome editing in plants such as tobacco (*Nicotiana tabacum*) [7], *Arabidopsis* [8], and rice (*Oryza sativa*) [9]. TALEs can also be fused to a transcriptional activator such as VP64 for engineering synthetic transcriptional activators (Fig. 1b) [10]. Binding upstream of the gene of interest by such synthetic transcriptional activators can result in high gene expression, which is a useful tool for studying gene

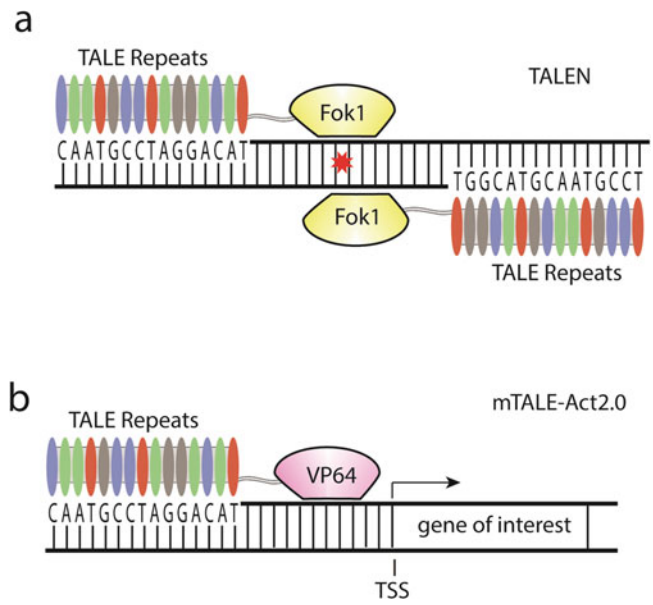


Fig. 1 Diagram of TALEN and mTALE-Act2.0 systems bound to target DNA. Individual RVDs bind to nucleotides and make up the TALE repeats. TALE repeats are fused to an effector protein. (**a**) Fusion of TALE repeats to Fok1 nuclease creates TALEN. Two TALENs are required to create a dimer of Fok1 and induce a DNA double-strand break. (**b**) Fusion of TALE repeats to VP64 activator creates mTALE-Act2.0. Binding upstream of the transcription start site (TSS) activates transcription of the gene of interest

function and regulation. We previously described mTALE-Act system which allows for simultaneous expression of three synthetic TALE-VP64 transcriptional activators for activating up to three genes at once [10].

TALE-based genome engineering is less widespread than CRISPR because it requires more time-consuming construction procedures. However, TALE-based systems are very specific and have unique properties in certain situations. For example, we found that our mTALE-Act system resulted in higher gene expression than CRISPR-Act2.0, which is an improved transcriptional activation system based on CRISPR-Cas9 [10]. Additionally, TALEN is better suited to genome engineering applications that require protein-only nucleases as is the case with mitochondria editing [11]. Here we describe a two-step Golden Gate cloning method for assembling TALE repeats which are sub-cloned into different expression vectors for final assembly of T-DNA vectors based on Gateway recombination for making a TALEN [12] or mTALE-Act system in plants [10].

2 Materials

1. DNA editing computer software such as ApE, Snapgene, and DNA Star and access to TAIR or Genbank.
2. Golden Gate TAL Effector Kit 2.0 from Voytas Lab on Addgene (http://www.addgene.org/kits/voytas-taleffector-goldengatev2/#kit-details). pFUS_A8, pYPQ121, pYPQ127B, and pYPQ202. Plasmids can be found from the Qi Lab on Addgene (http://www.addgene.org/Yiping_Qi/). pZHY013 is also available at Addgene (https://www.addgene.org/36185/).
3. Restriction enzymes *Bsa*I or *Bsa*I-HFv2, *Esp*31/*Bsm*BI, *Eco*RI, *Xba*I, *Bam*HI, *Nhe*I, and *Bgl*II.
4. T4 DNA ligase and 10× T4 DNA ligase buffer.
5. 100 mM Dithiothreitol (DTT).
6. Plasmid-Safe nuclease (Epicentre Biotechnologies, Madison, WI, USA).
7. 25 mM ATP.
8. Taq DNA polymerase, buffer, and dNTPS.
9. Gel electrophoresis equipment.
10. DH5α chemically competent cells.
11. SOC medium: 5 g/L Yeast extract, 20 g/L tryptone, 20 mM dextrose, 10 mM sodium chloride, 2.5 mM potassium chloride, 10 mM magnesium chloride.

12. 50 mg/mL Spectinomycin, carbenicillin/ampicillin, and kanamycin stock.
13. LB plates and liquid media.
14. 40 mg/mL X-gal dissolved in dimethyl sulfoxide or *N*′,*N*-dimethylformamide.
15. 100 mM IPTG (isopropyl β-D-1-thiogalactopyranoside) dissolved in water.
16. Miniprep Kit (QIAprep Spin Miniprep Kit, Qiagen).
17. Primers:
    pCR8_F1: 5′ TTGATGCCTGGCAGTTCCCT 3′
    pCR8_R1: 5′ CGAACCGAACAGGCTTATGT 3′
    TAL_F1: 5′ TGGCGTCGGCAAACAGTGG 3′
    TAL_R2: 5′ GGCGACGAGGTGGTCGTTGG 3′
18. Optional restriction enzymes: *Afl*II, *Bsp*EI.

3 Methods

Two rounds of Golden Gate reactions will be used to assemble 14–16 TALE repeats for recognition of 14–16 bp DNA target sequence. The full 14–16 TALE repeats are broken apart into two segments with each of 7–8 TALE repeats, to be assembled into pFUS_A and pFUS_B vector series, respectively, in the first round of Golden Gate reaction. These two segments will be assembled in the second round of Golden Gate reaction to generate the full-length TALE repeats. Then, fully assembled TALE repeats will be cloned into pYPQ121 and pYPQ127B for multiplex transactional activation by mTALE-Act (Fig. 1b) or cloned into pZHY013 for genome editing by TALEN (Fig. 1a). The final T-DNA vectors are generated by LR Gateway reactions with appropriate attR1-attR2 destination vectors.

3.1 Golden Gate Assembly of TALE Repeats Step 1

1. Download the sequence for *Arabidopsis Cleavage stimulating factor 64 (CSTF64, At1g71800)*, *Glabrous 1 (GL1, At3g27920)*, *RNA-binding protein-defense-related 1 (RBP-DR1, At4g03110)*, and *Alcohol Dehydrogenase 1 (ADH1, At1g77120)* from TAIR (https://www.arabidopsis.org/) or Genbank (https://www.ncbi.nlm.nih.gov/genbank/). The sequence should include ~1000 bp upstream of the transcriptional start site for *CSTF64*, *GL1*, *and RBP-DR1*, which are targeted for transcriptional activation by mTALE-Act. *ADH1* will be targeted for mutagenesis by TALEN. Use a DNA editing software such as ApE, Snapgene, and DNA Star.

2. If designing mTALE-Act for transcriptional activation, select a 14–16 bp target site about ~150 to 350 bp upstream of the gene of interest using TALEN Effector Targeter (https://tale-nt.cac.cornell.edu/node/add/single-tale) of TAL Plasmids Sequence Assembly Tool (http://bao.rice.edu/Research/BioinformaticTools/assembleTALSequences.html). If designing for TALEN for genome editing, select two proximal target sites within an exon, preferably within the first half of the gene to increase the likelihood of creating a knockout. Upload the sequence for the gene of interest and select 14 bp for the minimum and 16 bp for the maximum (*see* **Notes 1** and **2**). We selected the following targets for mTALE-Act: *CSTF64* (5′-ttccttttaaccaaaat-3′), *GL1* (5′-acgtattgatgtgagt-3′), and *RBP-DR1* (5′-ttaatttctcccaact-3′) [10], and the following targets for TALEN: ADH1-left (5′ CCGGATGCTCCTCTT 3′) and ADH1-right (5′ AGTTGTGGTTTGTCT 3′) [8].
3. Select plasmids containing RVD domains for targeting selected sequence. Each RVD corresponds to one nucleotide. The plasmid names correspond to the RVD and the location in the TALE repeat assembly (*see* **Note 3**). Our design uses 8 TALE repeats per vector to increase the efficiency of vector assembly. The first 8 are placed into the pFUS_A8 and then another 6–8 TALE repeats are assembled into pFUS_B(N) using a *Bsa*I-based Golden Gate reaction. The "N" in pFUS_A(N) responds to the number of repeats assembled into the vector. Because pFUS_B houses the final TALE repeat, which is added in a later step, the vector number will be the number of repeats minus 1. For example, one would use pFUS_B7 to house 7 out of 8 TALE repeats, and the last repeat will be added later. The table in **step 4** contains the RVDs used to target our three sequences. The last repeat vector is labeled "LR." pFUS_B8 is used for the repeats of CSTF64, pFUS_B7 for GL1, pFUS_B7 for AtRBP-DR1, pFUS_B7 for ADH1-left, and pFUS_B6 for ADH1-right (*see* **Note 4**).
4. Select the modular plasmids corresponding to the target sequence. The table below describes the vectors used to assemble TALEs for our three target sites.

TALE	RVDs in pFUS_A8 vector							
CSTF64	pNG1	pNG2	pHD3	pHD4	pNG5	pNG6	pNG7	pNG8
GL1	pNI1	pHD2	pNN3	pNG4	pNI5	pNG6	pNG7	pNN8
AtRBP-DR1	pNG1	pNG2	pNI3	pNI4	pNG5	pNG6	pNG7	pHG8
ADH1-left	pHD1	pHD2	pNN3	pNN4	pNI5	pNG6	pNN7	pHD8
ADH1-right	pNI1	pNN2	pNI3	pHD4	pNI5	pNI6	pNI7	pHD8

TALE	RVDs in pFUS_B(N) vector								
CSTF64	pNI1	pNI2	pHD3	pHD4	pNI5	pNI6	pNI7	pNI8	pLR-NG
GL1	pNI1	pNG2	pNN3	pNG4	pNN5	pNI6	pNN7	pLR-NG	
AtRBP-DR1	pNG1	pHD2	pHD3	pHD4	pNI5	pNI6	pHD7	pLR-NG	
ADH1-left	pNG1	pHD2	pHD3	pNG4	pHD5	pNG6	pNG7	pLR-NG	
ADH1-right	pHD1	pNI2	pHD3	pNI4	pNI5	pHD6	pLR-NG		

5. Assemble the modular RVD vectors simultaneously into the pFUS_A8 and pFUS_B(N) vectors using the following Golden Gate reaction (*see* **Notes 5**–**7**):

Golden Gate recipe		Golden Gate program		
Each modular RVD vector pFUS_A8 or pFUS_B(N)	150 ng 75 ng	37 °C 16 °C	5 min 10 min	10×
*Bsa*I	1 μL	50 °C	5 min	
T4 DNA ligase	1 μL	80 °C	5 min	
10× T4 DNA ligase buffer Water	2 μL Up to 20 μL			

6. Cool the reactions on ice and add the following:
   (a) 1 μL 25 mM ATP
   (b) 1 μL Plasmid-Safe nuclease

   Incubate at 37 °C for 1 h (*see* **Note 8**). This step removes all incomplete ligations.

7. Transform 5 μL of reaction into 50 μL of *E. coli* strain DH5α using heat shock or electroporation and rescue with SOC media. Grow at 37 °C for 1 h, and spin and remove supernatant before plating and growing at 37 °C overnight. Use blue/white screening to select correct colonies by using spectinomycin (50 mg/L) LB plates with X-gal and IPTG. If few or no white colonies are observed, then consider replacing the T4 ligase buffer or supplementing with DTT (*see* **Note 6**). Screen both pFUS-A8 and pFUS_B(N) vectors through colony PCR with primers pCR8-F1 and pCR8-R1 using the following recipe and program:

PCR recipe		PCR program		
10× Standard Taq Reaction Buffer	2.5 μL	95 °C	60 s	
10 mM dNTPS	0.5 μL	95 °C	20 s	
pCR8-F1 (10 μM)	0.5 μL	55 °C	30 s	30×

(continued)

PCR recipe		PCR program	
pCR8-R1 (10 μM)	0.5 μL	68 °C	60 s
Taq DNA polymerase	0.1 μL	68 °C	5 min
Water	20.9 μL	10 °C	Hold

The products from the colony PCR will be the full size of the repeat and form a ladder of bands, which is expected from repetitive TALEs (*see* **Note 9**). The full repeat array for 8 RVDs is ~900. The ladder should start around 200 bp and occur every 100 bp. Select colonies and grow overnight in liquid LB with spectinomycin.

8. Purify plasmids from the LB cultures. Correct clones can be confirmed using digestion with *Afl*II and *Xba*I, although this step is optional. RVD sequences can also be confirmed using *Bsp*EI digestion or sequencing with pCR8-F1 and pCR8-R1 primers (*see* **Note 10**).

3.2 Golden Gate Assembly of TALE Repeats Step 2

1. Use a Golden Gate reaction to fuse pFUS_A8, pFUS_B(N), and pLR-NG together into expression vector pZHY500 (*see* **Notes 5**–**7**). This reaction will fuse together the TALE repeats A and B and the last RVD into a backbone that has appropriate restriction enzyme sites for the next cloning step. This reaction will be carried out for each target gene. In our example, the Golden Gate reaction for mTALE-Act targeting *RBP-DR1* will yield pZHY500-1, *CSTF64* vector is pZHY500-2, and *GL1* is pZHY500-3. For TALEN targeting *ADH1*, ADH1-left will yield pZHY500-4 and ADH1-right will yield pZHY500-5.

Golden Gate recipe		Golden Gate program		
pFUS_A8 pFUS_B(N)	150 ng 150 ng	37 °C 16 °C	5 min 10 min	10×
pLR-NG	150 ng	50 °C	5 min	
pZHY500 *Esp*3I/*Bsm*BI T4 DNA ligase 10× T4 DNA ligase buffer Water	75 ng 1 μL 1 μL 2 μL Up to 20 μL	80 °C	5 min	

Plasmid-safe nuclease treatment is not necessary as the expression vector does not have homology with the repeats.

2. Transform into *E. coli* and incubate overnight at 37 °C on LB plates with 50 mg/L carbenicillin or ampicillin. Supplement with X-gal and IPTG for blue and white screening.
3. Select white colonies for inoculation into liquid LB media with carbenicillin/ampicillin. Confirm using colony PCR with primers TAL_F1 and TAL_R2 (*see* **Note 9**). The brightest band

should be the full TALE, which in this example is about 1700 bp. Grow colonies overnight and harvest for miniprep. Further confirmation with digestion and sequencing should be done for the final vectors.

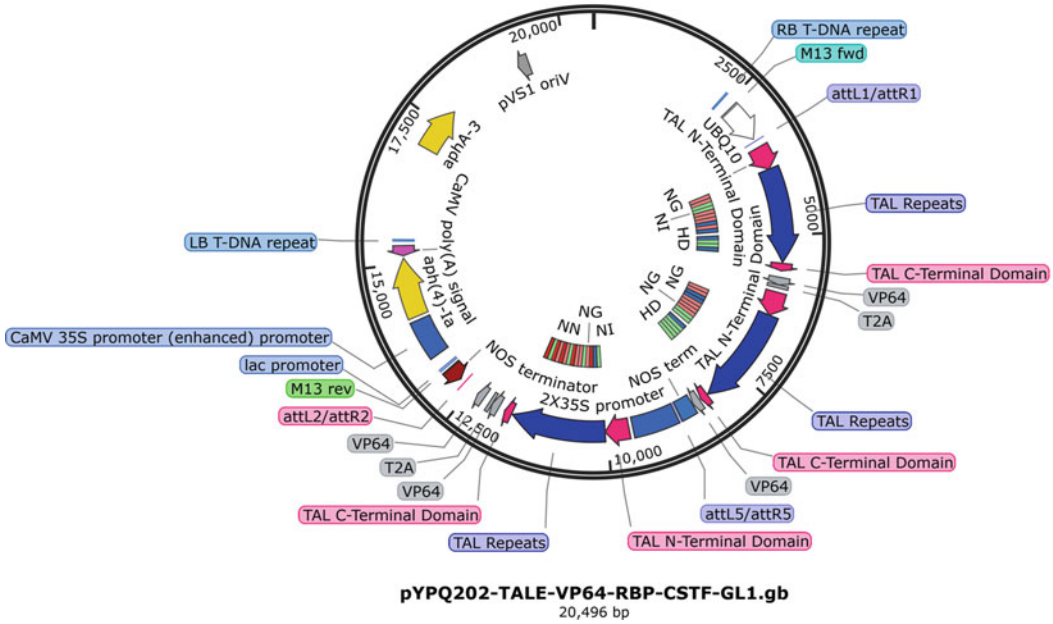
3.3 Assembly of Final T-DNA Vectors for Transcriptional Activation with mTALE-Act

1. Excise the full repeats from pZHY500-1 using restriction enzymes *Xba*I and *Bam*HI and ligate into pYPQ121 using restriction enzyme cloning. This step generates vector pYPQ121-1.
2. Excise the full repeats from pZHY500-2 using restriction enzymes *Nhe*I and *Bgl*II and ligate into pYPQ121-1. This step generates vector pYPQ121-2. Multiplexed mTALE-Acts are separated by a T2A ribosomal skipping motif which allows translation of multiple proteins (the left and right TALE-VP64) from a single transcript [8].
3. Excise the full repeats from pZHY500-3 using restriction enzymes *Xba*I and *Bam*HI and ligate into pYPQ127B (which contains an AtUBQ10 promoter) to generate pYPQ127B-1.
4. The final T-DNA vector is generated through Multisite Gateway recombination with pYPQ121-2, pYPQ127B-1, and destination vector pYPQ202 (*see* **Note 11**) (Fig. 2). This assembles three TALE-VP64 into one vector (*see* **Note 12**). If using one or two TALE-VP64s, use pYPQ140 filler plasmid instead of pYPQ127B.

   Use the following Multisite Gateway recombination recipe (*see* **Note 13**):

Multisite Gateway recombination	
pYPQ121-2	80 ng
pYPQ127B-1	80 ng
pYPQ202	100 ng
LR Clonase II	1 μL
Total volume	7 μL

   Incubate at room temperature overnight.
5. Transform into competent *E. coli* via heat shock or electroporation and plate onto LB media with kanamycin.
6. Select colonies and culture overnight in liquid LB media with kanamycin at 37 °C. Miniprep and verify correct clones by digesting with restriction enzyme *Eco*RI. A correct vector (Fig. 2) will yield bands of 7934, 5423, 3911, 2878, and 350 bp.



pYPQ202-TALE-VP64-RBP-CSTF-GL1.gb
20,496 bp

Fig. 2 Vector map of T-DNA vector housing multiplexed mTALE-Act2.0 system targeting *RPB-DR1*, *GL1*, and *CSTF64* for simultaneous transcriptional activation of three genes in *Arabidopsis*

3.4 Assembly of Final T-DNA Vectors for Genome Editing with TALEN

1. Excise the full repeats from pZHY500-4 using restriction enzymes *Xba*I and *Bam*HI and ligate into pZHY013 using restriction enzyme cloning. This step generates vector pZHY013-1.
2. Excise the full repeats from pZHY500-5 using restriction enzymes *Nhe*I and *Bgl*II and ligate into pZHY013-1 at compatible *Xba*I and *Bam*HI sites. This step generates vector pZHY013-2. The T2A ribosomal skipping motif allows translation of the left and right TALEN from a single transcript [8].
3. The final T-DNA vector is generated through Multisite Gateway recombination with pZHY013-2 and pYPQ202 (*see* **Note 11**) (Fig. 3). This assembles two TALEN into one vector driven by an AtUBQ10 promoter (*see* **Note 12**).

   Use the following Multisite Gateway recombination recipe (*see* **Note 13**):

Multisite Gateway recombination	
pZHY013-2	80 ng
pYPQ202	100 ng
LR Clonase II	1 μL
Total volume	7 μL

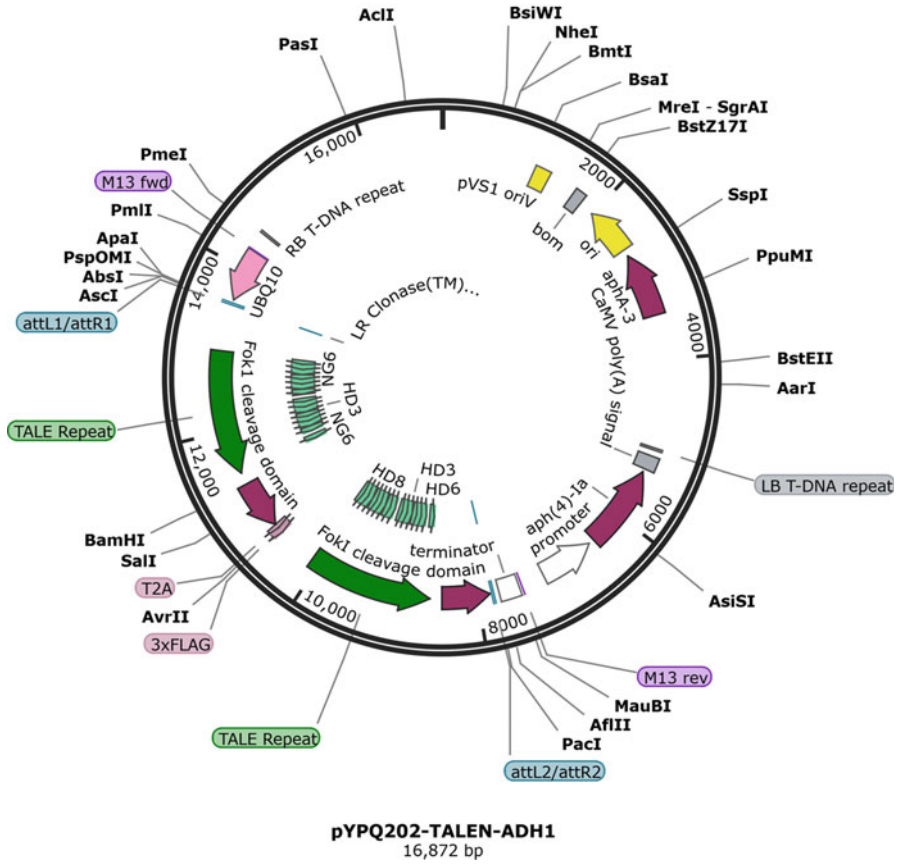


Fig. 3 Vector map of T-DNA housing two TALENs targeting *ADH1* in *Arabidopsis*

Incubate at room temperature overnight.

4. Transform into competent *E. coli* via heat shock or electroporation and plate onto LB media with kanamycin.
5. Select colonies and culture overnight in liquid LB media with kanamycin at 37 °C. Miniprep and verify correct clones by digesting with restriction enzyme *Eco*RI. A correct vector (Fig. 3) will yield bands of 7934, 5486, 2878, 485, and 350 bp.

4 Notes

1. We recommend using T as the first base of a TALE target site or at the −1 position [13]. This is sometimes referred to as the "upstream base." We also recommend using T as the last base of a TALE target site.
2. Streubel et al. showed that NH binds the nucleotide G more specifically than NN, although TALEN Effector Targeter has

the option to select either RVD for the design. We recommend using NN to bind to G since higher binding affinity has been observed with NN than NH [10, 14].

3. *E. coli* housing vector pHD2 is slow growing, and it is unclear why. Additionally, plasmid yields from *E. coli* strain DH5α are higher than from strain DH10B. Plasmids from Addgene are in one of these two strains. It is unclear what mechanism is behind this difference.
4. The vectors in pFUS_B are numbered starting with 1, even though they would follow the RVDs in the pFUS_A(N) vector. In this example, this would be the ninth position in the TALE assembly.
5. Use *Bsa*I which works better in the Golden Gate reaction than *Bsa*I-HF New England Biolabs has released, *Bsa*I-HFv2 (NEB #R3733), which is designed to work well in Golden Gate reactions. This enzyme was not used in this study and we cannot comment directly on the efficiency of the enzyme.
6. Use fresh T4 DNA ligase buffer as it can lose potency and is often the cause of an inefficient Golden Gate reaction. We recommend aliquoting ligase buffer into ~25 μL volumes. A simple way to determine if T4 DNA ligase buffer has gone bad is the absence of the sulfuric smell of dithiothreitol (DTT) in the buffer. If there is no smell, DTT can be added to 1 mM concentration in the final reaction.
7. Golden Gate reactions can be performed in 10 μL reactions to save reagents, but difficult reactions may require 20 μL reactions.
8. The plasmid-safe protocol recommends inactivating the enzyme by heating the reaction to 70 °C for 30 min, but this is an optional step.
9. Select between 3 and 10 colonies to screen with colony PCR. When screening via colony PCR, correct colonies can appear as a smear on the agarose gel. However, there should be a slightly brighter band at the lengths of the repeats within the smear.
10. Sequencing TALEs can be difficult due to the repeats but it is important to confirm correct sequences. Rarely, mutations can occur that affect activity.
11. Multisite Gateway recombination is sensitive to concentration; attention should be paid to the concentration of the vectors, which should be diluted if necessary.
12. It is easy to use a different promoter to drive the expression of mTALE-Act or TALEN. Instead of using pYPQ202, other destination vectors containing different promoters with attR1-attR2 Gateway recombination sites may be used.

13. We have found that Invitrogen Gateway LR Clonase II enzyme mix works well and it is not necessary to use Multisite-Gateway Pro kit which is more expensive.

Acknowledgments

This work is supported by NSF Plant Genome ECA-PGR Award (IOS-1758745) to the University of Maryland, College Park.

References

1. Paul JW, Qi Y (2016) CRISPR/Cas9 for plant genome editing: accomplishments, problems and prospects. Plant Cell Rep 35:1417–1427. https://doi.org/10.1007/s00299-016-1985-z
2. Malzahn A, Lowder L, Qi Y (2017) Plant genome editing with TALEN and CRISPR. Cell Biosci 7. https://doi.org/10.1186/s13578-017-0148-4
3. Römer P, Hahn S, Jordan T, Strauß T, Bonas U, Lahaye T (2007) Plant pathogen recognition mediated by promoter activation of the pepper Bs3 resistance gene. Science 318:645–648. https://doi.org/10.1126/science.1144958
4. Boch J, Scholze H, Schornack S, Landgraf A, Hahn S, Kay S, Lahaye T, Nickstadt A, Bonas U (2009) Breaking the code of DNA binding specificity of TAL-type III effectors. Science 326:1509–1512. https://doi.org/10.1126/science.1178811
5. Moscou MJ, Bogdanove AJ (2009) A simple cipher governs DNA recognition by TAL effectors. Science 326:1501. https://doi.org/10.1126/science.1178817
6. Christian M, Cermak T, Doyle EL, Schmidt C, Zhang F, Hummel A, Bogdanove AJ, Voytas DF (2010) Targeting DNA double-strand breaks with TAL effector nucleases. Genetics 186:757–761. https://doi.org/10.1534/genetics.110.120717
7. Zhang Y, Zhang F, Li X, Baller JA, Qi Y, Starker CG, Bogdanove AJ, Voytas DF (2013) Transcription activator-like effector nucleases enable efficient plant genome engineering. Plant Physiol 161:20–27. https://doi.org/10.1104/pp.112.205179
8. Christian M, Qi Y, Zhang Y, Voytas DF (2013) Targeted mutagenesis of Arabidopsis thaliana using engineered TAL effector nucleases. G3 3:1697–1705. https://doi.org/10.1534/g3.113.007104
9. Li T, Huang S, Jiang WZ, Wright D, Spalding MH, Weeks DP, Yang B (2011) TAL nucleases (TALNs): hybrid proteins composed of TAL effectors and FokI DNA-cleavage domain. Nucleic Acids Res 39:359–372. https://doi.org/10.1093/nar/gkq704
10. Lowder LG, Zhou J, Zhang Y, Malzahn A, Zhong Z, Hsieh T-F, Voytas DF, Zhang Y, Qi Y (2018) Robust transcriptional activation in plants using multiplexed CRISPR-Act2.0 and mTALE-act systems. Mol Plant 11:245–256. https://doi.org/10.1016/j.molp.2017.11.010
11. Gammage PA, Moraes CT, Minczuk M (2018) Mitochondrial genome engineering: the revolution may not be CRISPR-Ized. Trends Genet 34:101–110. https://doi.org/10.1016/j.tig.2017.11.001
12. Cermak T, Doyle EL, Christian M, Wang L, Zhang Y, Schmidt C, Baller JA, Somia NV, Bogdanove AJ, Voytas DF (2011) Efficient design and assembly of custom TALEN and other TAL effector-based constructs for DNA targeting. Nucleic Acids Res 39:e82. https://doi.org/10.1093/nar/gkr218
13. Doyle EL, Booher NJ, Standage DS, Voytas DF, Brendel VP, VanDyk JK, Bogdanove AJ (2012) TAL Effector-Nucleotide Targeter (TALE-NT) 2.0: tools for TAL effector design and target prediction. Nucleic Acids Res 40:W117–W122. https://doi.org/10.1093/nar/gks608
14. Streubel J, Blücher C, Landgraf A, Boch J (2012) TAL effector RVD specificities and efficiencies. Nat Biotechnol 30:593–595. https://doi.org/10.1038/nbt.2304

Chapter 16

Genome Editing to Achieve the Crop Ideotype in Tomato

Tomaš Čermák, Karla Gasparini, Zoltán Kevei, and Agustin Zsögön

Abstract

For centuries, combining useful traits into a single tomato plant has been done by selective crossbreeding that resulted in hundreds of extant modern cultivars. However, crossbreeding is a labor-intensive process that requires between 5 and 7 years to develop a new variety. More recently, genome editing with the clustered regularly interspaced short palindromic repeats (CRISPR)/Cas9 system has been established as an efficient method to accelerate the breeding process by introducing targeted modifications to plant genomes via generation of targeted double-strand breaks (DSBs). CRISPR/Cas9 has been used to generate a variety of specific changes ranging from gene knockouts to gene replacements, and can also be easily multiplexed to modify several targets simultaneously. Given that (1) generating knockout mutations only requires a DSB that is frequently repaired by the error-prone nonhomologous end joining (NHEJ) pathway resulting in gene function inactivation, and (2) the genetic basis of many useful agronomic traits consists of loss of gene function, multiple traits can be created in a plant in one generation by simultaneously introducing DSBs into multiple genes of interest. On the other hand, more precise modifications, such as allele replacement, can be achieved by gene targeting—a less efficient process in which an external template is used to repair the DSB by homologous recombination (HR). These technical breakthroughs allow the design and customization of plant traits to achieve the ideal plant type ("ideotype"). Here, we describe protocols to assemble CRISPR/Cas9 constructs for both single and multiplex gene knockouts as well as gene targeting and to generate and identify genome-edited tomato plants via *Agrobacterium*-mediated transformation in tissue culture.

Key words CRISPR/Cas9, Gene-editing, Gene replacement, *Solanum lycopersicum*, Tomato

1 Introduction

The ideotype is a theoretical model of an ideal crop plant with a combination of highly desirable agronomic traits in a given environment [1]. The main conceptual difference between "ideotype breeding" and conventional breeding is that instead of generating random variation and selecting useful traits, desirable traits are first modeled and then purposefully bred into crops. Originally proposed for wheat, this breeding approach was limited by the dependence on preexisting phenotypic variation and the difficulty of combining multiple useful traits into a single plant [2, 3]. Progress

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264,
https://doi.org/10.1007/978-1-0716-1201-9_16,

in understanding the genetic basis of valuable agronomic traits in crops, coupled with the advent of genome engineering, now makes it possible to accelerate breeding using the ideotype modeling approach [4].

The clustered regularly interspaced short palindromic repeat/CRISPR-associated system 9 (CRISPR/Cas9) is used by bacteria to prevent bacteriophage infection [5, 6] but has been repurposed as a biotechnological tool to generate targeted DNA mutations [7, 8]. CRISPR/Cas9 is a site-specific nuclease system that consists of the Cas9 endonuclease and an engineered single chimeric guide RNA (sgRNA). The sgRNA contains 20 nucleotides of spacer sequence that guide Cas9 to the complementary sequence in the genome (the "protospacer"), next to a protospacer adjacent motif (PAM). For the *Streptococcus pyogenes* Cas9 (SpCas9) used in our protocols, the PAM sequence is 5′-NGG-3′. Upon binding of sgRNA to the genomic target next to a PAM, Cas9 cleaves both DNA strands to create a double-strand break (DSB). The DSB is then repaired by the nonhomologous end joining (NHEJ) pathway leading to either short insertion/deletion (Indel) mutations that result in gene knockout, or under specific conditions, by homologous recombination (HR) leading to precise repair. The repair by HR is much less frequent in plant cells compared to NHEJ, making gene replacement applications rather challenging. However, a few methods have been developed to improve the efficiency of HR-mediated precise editing in plants [9, 10]. Of those, geminivirus replicons (GVRs) that increase the copy number of the DNA template for HR by replication have been successfully implemented in tomato [11–13].

Many useful agronomic traits are the result of mutations inactivating gene function. Re-creating these mutations in cash crops using CRISPR/Cas9 offers the possibility of bypassing the laborious process of crossing and selection [14]. Furthermore, creating multiple gene knockouts in a single transformation event is possible with CRISPR/Cas9 vectors expressing multiple sgRNAs targeting different genes at the same plant line [15]. We have used this approach in a wild relative of tomato to produce an ideotypic phenotype with agronomic potential [16]. Our de novo domestication approach using genome engineering could be extended to other crop species, provided that knowledge of the genetic basis of the target traits is available [17]. Furthermore, gene targeting can be used to create traits that are not based on loss-of-function mutations, and a combination of both approaches paves the way for realizing the tomato ideotype.

To facilitate the rapid construction of CRISPR/Cas9 vectors for plants, we have recently developed a toolkit that uses Golden Gate cloning to assemble different gene editing components into a single construct [18] (http://crispr-multiplex.cbs.umn.edu/). The individual components are pre-cloned in three intermediate

plasmids—modules A, B, and C, which are assembled into a T-DNA transformation backbone. Alternatively, simple knockout vectors can also be created in one step by inserting sgRNAs directly into one of the DIRECT expression vectors. Individual vectors can be obtained from plasmid repositories Addgene (https://www.addgene.org/) and ABRC (https://abrc.osu.edu/). In this chapter, we describe detailed procedures to generate different types of gene editing constructs for tomato using this toolkit, as well as their application in *Agrobacterium*-mediated tomato transformation.

2 Materials

2.1 Equipment

1. Thermocycler.
2. Autoclave.
3. Microwave oven.
4. Laminar flow.
5. Agarose gel electrophoresis equipment.
6. Shaking and static incubator.
7. Water bath.
8. pH meter.
9. Spectrophotometer for nucleic acid quantification.
10. Centrifuge (for microtubes and Falcon tubes).
11. Microtubes (1.5 mL and PCR tubes).
12. Petri dishes.
13. Filter paper.
14. Angled tweezers.
15. Scalpel.
16. Sterile blades.
17. Syringe filters.
18. Sterile syringe.
19. Sterile Pasteur pipettes.
20. Pipettes.
21. Pipette tips.
22. Falcon tubes.
23. Glass bottles.
24. Electroporator plus accessories for *Agrobacterium* transformation.

2.2 Reagents

1. Vectors pDIRECT_22A (Addgene #91133; ABRC #CD3-2667), pDIRECT_22C (Addgene #91135; ABRC #CD3-2668), pMOD_A0101 (Addgene #90998; ABRC #CD3-

2547), pMOD_A0501 (Addgene #91011, ABRC #CD3-2560), pMOD_B2515 (Addgene #91072; ABRC #CD3-2613), pMOD_B2103 (Addgene #91061, ABRC #CD3-2602), pMOD_C0000 (Addgene #91081; ABRC #CD3-2620), pMOD_C3006 (Addgene #91096, ABRC #CD3-2635), pTRANS_220d (Addgene #91114; ABRC #CD3-2651), pTRANS_221 (Addgene #91115, ABRC #CD3-2652).

2. Oligonucleotides and primers:
   (a) Complementary oligonucleotides containing sgRNA spacer sequence or primers to amplify sgRNA array fragments (see protocol).
   (b) M13F: 5′-GTAAAACGACGGCCAGT-3′.
   (c) M13R(-48): 5′-CGGATAACAATTTCACACAG-3′.
   (d) NB424: 5′-CAGCGAGTCAGTGAGCGAGGAAGC-3′.
   (e) NB442: 5′-GCAATCCTGACGAAGACTGGATGT-3′.
   (f) NB457: 5′-CAAGAATTGGGACAACTCCAG-3′.
   (g) NB463: 5′-CGAACGGATAAACCTTTTCACG-3′.
   (h) TC089R: 5′-GGAACCCTAATTCCCTTATCTGG-3′.
   (i) TC214F: 5′-GAAGAGAAGCAGGCCCATTTAT-3′.
   (j) TC320: 5′-CTAGAAGTAGTCAAGGCGGC-3′.
   (k) TC370: 5′-AATGTGTCAAATCGTGGCCT-3′.
   (l) TC430: 5′-GTTGGATCTCTTCTGCAGCA-3′.
   (m) ZY015F primer: 5′- GGAATAAGGGCGACACG GAAATG-3′.
   (n) Forward primer for left donor homology arm amplification: 5′- CGCGTAGTCCTCGG TANNNNNNNNNNNNNNNNNN -3′ (see protocol).
   (o) Reverse primer for right donor homology arm amplification: 5′- TGACTTGAAGTA CACTCNNNNNNNNNNNNNNNNNN -3′ (see protocol).
3. Sterile deionized water (ddH_2O).
4. Sodium hypochlorite.
5. Tween 20.
6. High-fidelity DNA polymerase (*see* **Note 1**).
7. Taq DNA polymerase.
8. T4 DNA ligase or T7 DNA ligase.
9. Deoxynucleotide mix (dNTP).
10. T4 polynucleotide kinase.

11. Restriction enzymes: *Aar*I, *Ban*I, *Esp*3I, *Sap*I.
12. Gibson Assembly Master Mix.
13. ccdB-sensitive chemically competent *E. coli* cells (such as DH5α).
14. Top 10 competent cells.
15. Electro-competent *Agrobacterium* (strains LBA4404 or GV3101) cells.
16. Carbenicillin.
17. Kanamycin.
18. Ampicillin.
19. Rifampicin.
20. Streptomycin.
21. Spectinomycin.
22. Bacto-tryptone.
23. Timentin.
24. Acetosyringone.
25. 1-Naphthaleneacetic acid (NAA).
26. Indole-3-acetic acid (IAA).
27. 6-Benzylaminopurine (BAP).
28. Zeatin (ZEA).
29. Isopropyl-beta-D-thiogalactopyranoside (IPTG).
30. 5-Bromo-4-chloro-3-indolyl-β-*d*-galactopyranoside (X-Gal).
31. Agarose.
32. DNA ladders.
33. Agar.
34. Phytagel.
35. Murashige and Skoog (MS) basal medium.
36. Yeast extract.
37. Sucrose.
38. Glucose.
39. Sodium chloride (NaCl).
40. Potassium chloride (KCl).
41. Magnesium chloride ($MgCl_2$).
42. Magnesium sulfate ($MgSO_4$).
43. Hydrochloric acid (HCl).
44. Potassium hydroxide (KOH).
45. Sodium hydroxide (NaOH).
46. Dimethyl sulfoxide (DMSO).

47. Dimethylformamide.
48. Plasmid extraction and purification kit.
49. Gel and PCR cleanup kit.
50. Reagents for genomic DNA extraction.
51. DNA cloning kit (*see* **Note 2**).

2.3 Solutions

1. Luria-Bertani (LB) liquid medium: Dissolve 10 g bacto-tryptone, 5 g yeast extract, and 10 g NaCl with a little ddH_2O.-Mix well, bring the volume up to 1 L, and adjust the pH to 7.5 with 4 M NaOH. Autoclave for 20 min at 121 °C.
2. Luria-Bertani (LB) solid medium: Dissolve 10 g bacto-tryptone, 5 g yeast extract, and 10 g NaCl with a little ddH_2O.-Mix well, bring the volume up to 1 L, and adjust the pH to 7.5 with 4 M NaOH. Add 15 g agar. Autoclave for 20 min at 121 °C. After cooling the medium to 50 °C, add appropriate antibiotics. For blue-white selection, add 40 μL of X-Gal (20 mg/mL) and 10 μL of IPTG (100 mM). X-Gal and IPTG are spread on the surface of the solidified LB media and allowed to dry in a sterile hood.
3. General propagation medium: Dissolve 4.43 g Murashige and Skoog (MS) basal medium with vitamins and 30 g sucrose with a little ddH_2O. When fully dissolved, bring the volume up to 1 L and adjust pH to 5.7 with 1 M KOH. Then add 7 g agar (*see* **Note 3**). Autoclave for 20 min at 121 °C.
4. Germination medium: Half-strength MS medium. Dissolve 2.2 g Murashige and Skoog (MS) basal medium with vitamins and 15 g of sucrose with a little ddH_2O. When fully dissolved, bring the volume up to 1 L and adjust pH to 5.7 with 1 M KOH. Then add 2.3 g phytagel and take it to the microwave to heat it. Dispense 40 mL into 500 mL flasks. Autoclave for 20 min at 121 °C.
5. Liquid MS medium: Dissolve 4.43 g Murashige and Skoog (MS) basal medium with vitamins and 30 g sucrose with a little ddH_2O. When fully dissolved, bring the volume up to 1 L and adjust pH to 5.7 with 1 M KOH. Autoclave for 20 min at 121 °C.
6. Root induction medium (RIM): Dissolve 4.43 g Murashige and Skoog (MS) basal medium with vitamins and 30 g sucrose with a little ddH_2O. When fully dissolved, bring the volume up to 1 L and adjust pH to 5.7 with 1 M KOH. Then add 7 g agar. Autoclave for 20 min at 121 °C. After cooling the medium to 50 °C, add 0.4 μM 1-naphthaleneacetic acid (NAA) and 100 μM acetosyringone.
7. Shoot induction medium (SIM): Dissolve 4.43 g Murashige and Skoog (MS) basal medium with vitamins and 30 g sucrose

with a little ddH_2O. After it is dissolved, bring the volume up to 1 L and adjust pH to 5.7 with 1 M KOH. Then add 7 g agar. Autoclave for 20 min at 121 °C. After cooling the medium to 50 °C, add 5 μM 6-benzylaminopurine, 300 mg/L timentin, and selection antibiotic (*see* **Note 4**).

8. SOC liquid medium: Dissolve 20 g tryptone, 5 g yeast extract, and 500 mg NaCl in 950 mL of ddH_2O. Add 10 mL of 250 mM KCl solution and adjust pH to 7.0 with 1 M NaOH. Autoclave for 20 min at 121 °C. After cooling the medium (below 60 °C), add 5 mL 2 M $MgCl_2$ solution, 5 mL 2 M $MgSO_4$ solution, and 20 mL 1 M glucose solution. Top to 1 L with sterile ddH_2O.
9. Acetosyringone solution (200 mM): Dissolve 392.4 mg acetosyringone in 10 mL of dimethyl sulfoxide (DMSO). Dispense into 1 mL aliquots in microtubes and store at 20 °C.
10. Indole-3-acetic acid (IAA) (0.1 mM): Dissolve 1.75 mg IAA in few drops of potassium hydroxide (KOH), bring the volume up to 100 mL with ddH_2O, and filter sterilize. Dispense into 1 mL aliquots in microtubes and store at 4 °C.
11. 1-Naphthaleneacetic acid (NAA) (0.4 mM): Dissolve 3.7 mg NAA in few drops of potassium hydroxide (KOH), bring the volume up to 50 mL with ddH_2O, and filter sterilize. Dispense into 1 mL aliquots in microtubes and store at 4 °C.
12. 6-Benzylaminopurine (BAP) (5 mM): Dissolve 56.3 mg BAP in a few drops of 1 M hydrochloric acid (HCl), bring the volume up to 50 mL with ddH_2O, and filter sterilize. Dispense into 1 mL aliquots in microtubes and store at 4 °C.
13. Kanamycin (50 mg/mL): Dissolve 500 mg with a little ddH_2O and bring the volume up to 10 mL with ddH_2O. Sterilize by filtration. Dispense into 1 mL aliquots in microtubes and store at −20 °C.
14. Zeatin (ZEA) (5 mM): Dissolve 54.8 mg ZEA in few drops of 1 M hydrochloric acid (HCl), bring the volume up to 50 mL with ddH_2O, and filter sterilize. Dispense into 1 mL aliquots in microtubes and store at −20 °C.
15. Rifampicin (50 mg/mL): Dissolve 500 mg rifampicin in 10 mL of DMSO. Dispense into 1 mL aliquots in microtubes and store at −20 °C.
16. Timentin (300 mg/mL): Dissolve 3 g timentin in 10 mL of ddH_2O and filter sterilize. Dispense into 1 mL aliquots in microtubes and store at −20 °C.
17. Ampicillin (100 mg/mL): Dissolve 1 g ampicillin in 10 mL of ddH_2O and filter sterilize. Dispense into 1 mL aliquots in microtubes and store at −20 °C.

18. Streptomycin (100 mg/mL): Dissolve 1 g streptomycin in 10 mL of ddH_2O and filter sterilize. Dispense into 1 mL aliquots in microtubes and store at −20 °C.
19. Spectinomycin (100 mg/mL): Dissolve 1 g spectinomycin in 10 mL of ddH_2O and filter sterilize. Dispense into 1 mL aliquots in microtubes and store at −20 °C.
20. X-Gal (20 mg/mL): Dissolve 200 mg 5-bromo-4-chloro-3-indolyl β-D-galactopyranoside in 10 mL of dimethylformamide. Dispense into 500 μL aliquots in microtubes covered with aluminum paper and store at −20 °C.

3 Methods

3.1 Construction of CRISPR/Cas9 Vectors

This section is a detailed description of the procedure to design and build CRISPR/Cas9 expression vectors for *Agrobacterium*-mediated transformation of tomato (*Solanum lycopersicum*). The protocol starts with the selection of target sites for sgRNA(s) in the gene of interest and proceeds with Golden Gate assembly of the gene editing constructs. We will give an example of vector construction for targeted knockout of one or up to six genes with or without a GFP marker, as well as for homologous recombination/template-mediated precise editing of an endogenous site using geminivirus replicons (GVRs). The sgRNA for a single-gene knockout is cloned in one step into the pDIRECT_22A T-DNA backbone. Similarly, vectors expressing six sgRNAs can be constructed in a single step by cloning the sgRNAs into the pDIRECT_22C T-DNA backbone. Alternatively, if a green fluorescent protein (GFP) marker is required in the final construct, one- and six-gene knockout constructs can be created in two steps by assembling the modular vectors pMOD_B2515 or pMOD_B2103 (with pre-cloned sgRNAs) and pMOD_C3006 into the pTRANS_220d T-DNA backbone. Finally, GVR gene targeting constructs require three cloning steps, the first two of which can be performed in parallel: cloning the sgRNA(s) into the modular vector pMOD_B2515 or pMOD_B2103, cloning of the donor template into the modular vector pMOD_C0000, and final assembly of the modular vectors into the T-DNA expression backbone pTRANS_221.

3.1.1 Selection of Target Sites for Cas9/sgRNA Binding and Cleavage

Target site in the gene of interest can be selected either manually or using one of the freely available online tools, such as CRISPR RGEN tools (http://www.rgenome.net/) or CRISPR-P 2.0 (http://crispr.hzau.edu.cn/cgi-bin/CRISPR2/CRISPR). Sites closer to the 5′ end of the coding sequence (CDS) or to a conserved functional domain of the gene are most effective for gene knockouts, while sites closer to the position of the desired custom

modification are necessary for efficient replacement/modification by gene targeting. Generally, any 5′-N_{20}-NGG-3′ site (where N_{20} represents the protospacer sequence and NGG the PAM) is a potential target for Cas9/sgRNA binding and cleavage, although there are some minimal constraints on the sgRNA sequence imposed by the cloning approach and requirements for efficient transcription from certain promoters (*see* **Notes 5** and **6**). The optional requirement to avoid unintended mutations at off-target sites may represent another constraint. Some of the tools for sgRNA design (such as the ones mentioned above) also offer the option to predict and/or avoid off-target mutations specifically in the tomato genome.

3.1.2 Assembling CRISPR/Cas9 Expression Vectors and Agrobacterium*-Mediated Transformation (*See ***Note 7****)*

Simple knockout vectors expressing Cas9 with one or six sgRNAs, an antibiotic selection gene, and no additional components can be built rapidly in a single step using pDIRECT vectors. Alternatively, adding additional components such as the GFP marker or the DNA donor template for gene targeting is facilitated by the modular vector assembly system. In this approach, the sgRNAs and the donor template are first cloned into separate intermediate module vectors and subsequently assembled along with Cas9 and GFP expression cassettes from pre-made module vectors into the pTRANS backbones.

3.1.3 Design and Construction of Single-Gene Knockout Vectors

1. Vectors targeting a single site using one sgRNA can be created using synthetic oligonucleotides that are annealed and directly ligated with the vector backbone. Design and synthesize two complementary oligonucleotides containing the 20 nt spacer sequence (20 nt of the genomic protospacer sequence preceding the PAM), as shown below. The Ns in the oligonucleotide with the 5′-GATT overhang represent the sequence of the PAM-containing strand of the genomic target, while the Ns in the oligonucleotide with the 5′-AAAC overhang represent the sequence of the opposite strand. Note that the G nucleotide immediately adjacent to the 5′-GATT overhang (and its C counterpart) *is fixed* and is a part of the 20 nt protospacer sequence (see below). This is necessary for efficient transcription from the *At*U6 promoter (*see* **Note 6**).

```
5′-GATTGNNNNNNNNNNNNNNNNNNN
          ||||||||||||||||||||
          CNNNNNNNNNNNNNNNNNNNCAAA-5′
```

2. Phosphorylate the oligonucleotides using the following reaction:
   (a) 3 μL 10× T4 DNA ligase buffer (contains ATP).
   (b) 3 μL 100 μM Forward oligonucleotide.

(c) 3 μL 100 μM Reverse oligonucleotide.

(d) 2 μL T4 polynucleotide kinase.

(e) 19 μL H_2O.

Incubate the reaction for 1 h at 37 °C. Alternatively, oligonucleotides may be purchased with 5′ phosphates.

3. Anneal the oligonucleotides in a thermocycler using the following program: 95 °C/5 min + ramping down to 85 °C at −2 °C/s + ramping down to 25 °C at −0.1 °C/s + 4 °C hold. Alternatively, boil the reaction for 2 min and let it gradually cool down to room temperature.
4. Dilute the reaction 25 times with water (1 μL of the reaction + 24 μL H_2O).
5. Clone the annealed oligonucleotides into the pDIRECT_22A vector via the following Golden Gate reaction:
   (a) 2 μL 10× T4 DNA ligase buffer.
   (b) 50 ng pDIRECT_22A plasmid.
   (c) 1 μL 25× diluted phosphorylated and annealed sgRNA oligonucleotides.
   (d) 0.5 μL *Aar*I restriction enzyme.
   (e) 0.4 μL *Aar*I oligonucleotide (comes with the *Aar*I enzyme).
   (f) 1 μL T4 DNA ligase (400 U).
   (g) H_2O up to 20 μL.
6. Place the Golden Gate reaction in a thermocycler and run the following program: 37 °C for 5 min + 16 °C for 10 min + 37 °C for 15 min + 80 °C for 5 min (*see* **Note 8**).
7. Transform competent *ccdB-sensitive E. coli* with 5 μL of the Golden Gate reaction and plate on LB + 50 mg/L kanamycin supplemented with X-Gal and IPTG for blue-white colony screening. Incubate at 37 °C overnight.
8. PCR screen 2–4 white colonies using Taq DNA polymerase, the forward sgRNA oligonucleotide as forward primer and NB463: 5′-CGAACGGATAAACCTTTTCACG-3′ binding to pDIRECT_22A vector backbone as reverse primer.
9. Inoculate 1–2 positive clones (yielding a 456 bp PCR product) into 5 mL LB + 50 mg/L kanamycin and incubate overnight in a 37 °C shaking incubator.
10. Isolate plasmid DNA via miniprep and confirm by Sanger sequencing with TC214_F primer: 5′- GAAGAGAAG CAGGCCCATTTAT-3′.

3.1.4 Design and Construction of Six-Gene Knockout Vectors (See **Note 9**)

1. To assemble vectors expressing an array of six sgRNAs to target six sites simultaneously, PCR is necessary to amplify seven unique fragments containing the promoter and pieces of the sgRNA units that are then seamlessly assembled into the final array in a Golden Gate reaction. Use the "Primer Design and Map Construction" tool at http://crispr-multiplex.cbs.umn.edu/assembly.php to design a reverse and a forward primer for each sgRNA. The example below (Fig. 1) illustrates the design of the primers for the first two sgRNAs. The primers for the remaining sgRNAs are designed in the same way as for the second sgRNA.

   To use the online tool, first generate a list of DNA protospacer sequences (*not* including the PAM sequence) for all six targets in Fasta format. This can be done using a text editor such as Windows Notepad as in this example:

   >target1
   TCCACTTCAAAATATTGTCA
   >target2
   AGCGCACCTCAAACAAGCCT
   >target3
   GATCCTAAGATGTCTAGGCG
   >target4
   TGGTTATGTGTCGGTAAAAT
   >target5
   AAGTTGAAGTGGTATAGATG
   >target6
   CCTGATTGATTGGTTGATAC

   Save as a ".txt" file and use the "Browse. . ." button to open the file in the Primer Design and Map Construction tool. Select pDIRECT_22C as the target vector, CmYLCV promoter, Esp3I restriction enzyme, and Csy4 splicing system and click "Submit" (see Note 10). Synthesize the primers from the list generated by the program and download the vector map for your reference.

2. Generate the template for the first PCR reaction to amplify the promoter fragment of the sgRNA array by digestion of pDIRECT_22C (*see* **Note 11**) with *BanI*:
   (a) 3–5 μg pDIRECT_22C plasmid DNA.
   (b) 2 μL enzyme buffer.
   (c) 1 μL *Ban*I restriction enzyme.
   (d) H_2O up to 20 μL.

   Incubate the reaction at 37 °C for 1 h and run on an agarose gel.

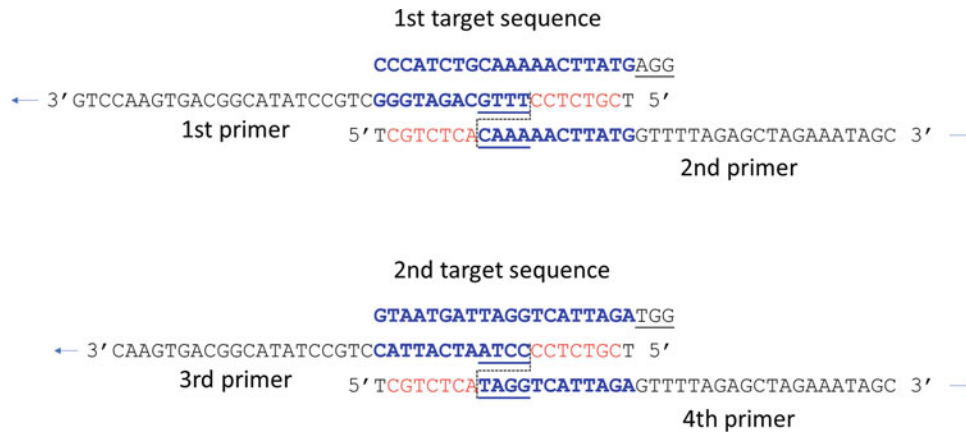


Fig. 1 Illustration of primer design for the first two sgRNAs in six sgRNA array. The reverse primer is extended with the first 12 nt of the target sequence in reverse complement (blue) and an *Esp*3I site (red). The forward primer is extended with the 12 bases in positions 9–20 of the protospacer sequence in 5′–3′ orientation (blue) and an *Esp*3I site (red). The 4 bp of overlapping sequence (underlined) in each of the primers will produce an overhang upon digestion of the PCR product with *Esp*3I, which will serve as the sticky ends (dashed lines) for ligation

3. Excise and purify the largest (4641 bp) fragment using a DNA gel extraction kit. Measure DNA concentration using spectrophotometer (or similar instrument).
4. PCR amplify seven unique overlapping fragments for sgRNA array assembly. Set up seven PCR reactions using primers specified for each reaction by the Primer Design and Map Construction tool. Use the *Ban*I fragment of pDIRECT_22C isolated in the previous step as template ONLY in the first PCR reaction. Use undigested pDIRECT_22C as template in all other reactions. Each reaction should contain the following in a total volume of 50 μL:
   (a) 10 μL 5× HF polymerase buffer.
   (b) 1 μL 10 mM dNTPs.
   (c) 2.5 μL 10 mM Forward primer.
   (d) 2.5 μL 10 mM Reverse primer.
   (e) 5–20 ng Template DNA.
   (f) 0.5 μL High-fidelity DNA polymerase.
   (g) H_2O up to 50 μL.
5. Run the following PCR program: 98 °C for 1 min + 30× (98 °C for 10 s + 60 °C for 20 s + 72 °C for 15 s) + 72 °C for 2 min + 4 °C hold.
6. Run 5 μL of each PCR product on an agarose gel to confirm successful amplification. The product of the first PCR reaction contains the promoter and is longer than the products of the other reactions (Fig. 2).

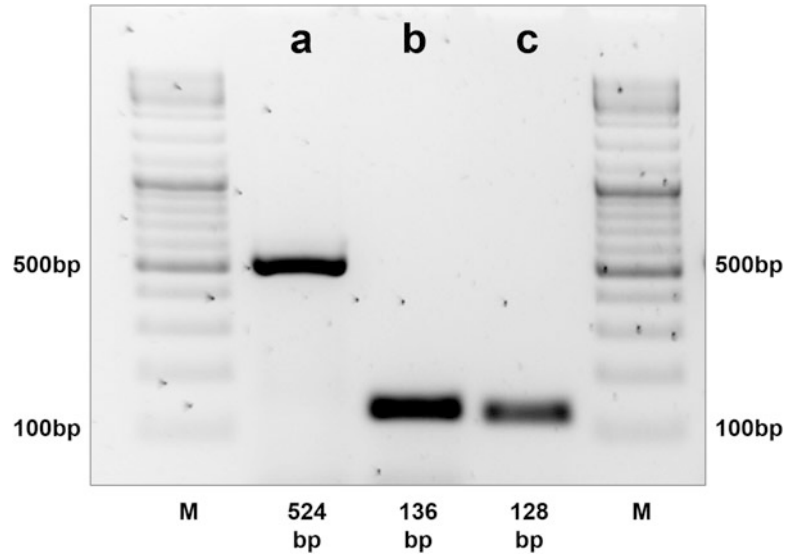


Fig. 2 Gel image of PCR products from **step 6** of Subheading 3.1.4. Products of the first (a), second (b), and last (c) PCR reaction are shown. Products of the third to sixth reaction (not shown) are identical in length to the product of the second PCR reaction. The length of each amplicon is indicated below the picture. 1.5% agarose gel was used. M, GeneRuler DNA Ladder Mix (Thermo Scientific)

7. Dilute each PCR reaction ten times with water (1 μL of the reaction + 9 μL H_2O).
8. Assemble the final sgRNA array in pDIRECT_22C via the following Golden Gate reaction:
   (a) 10 μL 2× T7 DNA ligase buffer (*see* **Note 12**).
   (b) 50 ng pDIRECT_22C plasmid DNA.
   (c) 0.5 μL of each 10× diluted PCR product (3.5 μL total).
   (d) 0.5 μL *Sap*I restriction enzyme (*see* **Note 13**).
   (e) 0.5 μL *Esp*3I restriction enzyme.
   (f) 1 μL T7 DNA ligase (*see* **Note 12**).
   (g) H_2O up to 20 μL.
9. Place the Golden Gate reaction in a thermocycler and run the following program: 10× (37 °C for 5 min + 25 °C for 10 min) + 4 °C hold (*see* **Notes 14** and **15**).
10. Transform competent *E. coli* with 5 μL of the Golden Gate reaction and plate on LB + 50 mg/L kanamycin. Incubate at 37 °C overnight.
11. PCR screen 8–16 colonies using Taq DNA polymerase, the forward primer TC320: 5′-CTAGAAGTAGTCAAGGCGGC-3′ and the reverse primer TC089R: 5′- GGAACCC TAATTCCCTTATCTGG-3′. The correctly assembled clones should yield an 881 bp product which usually also contains a

ladder of smaller products caused by misalignment of DNA strands during PCR amplification due to the repetitive nature of the amplified array.

12. Inoculate 2–3 positive clones into 5 mL LB + 50 mg/L kanamycin and incubate overnight in a 37 °C shaking incubator.
13. Isolate plasmid DNA via miniprep and confirm by Sanger sequencing with TC320: 5′-CTAGAAGTAGTCAAGGCGGC-3′ and M13F: 5′-GTAAAACGACGGCCAGT-3′ primers. Use the vector map created by the Primer Design and Map Construction tool as a reference.

3.1.5 Design and Construction of Single- or Six-Gene Knockout Vectors Expressing GFP

1. Create module B vector with one sgRNA for single-gene knockout vectors or six sgRNAs for six-gene knockout vectors. To insert one sgRNA into module B, follow **steps 1–10** of Subheading 3.1.3, with the following modifications:
   (a) In **step 5**, **use pMOD_B2515** and the following Golden Gate reaction:
   - 2 μL 10× T4 DNA ligase buffer.
   - 50 ng pMOD_B2515 plasmid.
   - 1 μL 25× Diluted phosphorylated and annealed sgRNA oligonucleotides.
   - 0.5 μL *Esp*3I restriction enzyme.
   - 1 μL T4 DNA ligase (400 U).
   - H_2O up to 20 μL.

   (b) In **steps 7** and **9**, use carbenicillin/ampicillin selection instead of kanamycin. X-Gal and IPTG are not necessary.
   (c) In **step 8**, use ZY015F: 5′-GGAATAAGGGCGACACG GAAATG-3′ as reverse primer (306 bp product) instead of NB463.

   To insert up to six sgRNAs into module B, follow **steps 1–13** of Subheading 3.1.4, with the following modifications:
   (a) Substitute pMOD_B2103 for pDIRECT_22C (**steps 1, 2, 4,** and **8**).
   (b) In **steps 10** and **12**, use carbenicillin/ampicillin selection instead of kanamycin.
   (c) In **step 13**, substitute TC089R: 5′- GGAACCC TAATTCCCTTATCTGG-3′ for M13F.
2. Assemble the final T-DNA expression vector via the following Golden Gate reaction:
   (a) 75 ng pTRANS_220d plasmid (transformation backbone).
   (b) 150 ng pMOD_A0101 plasmid (module A plasmid with Cas9).

(c) 150 ng Sequence confirmed module B plasmid with one or six sgRNAs (previous step).

(d) 150 ng pMOD_C3006 (module C plasmid with GFP).

(e) 0.5 μL *Aar*I.

(f) 0.4 μL *Aar*I oligonucleotide (comes with the *Aar*I enzyme).

(g) 1 μL T4 DNA ligase (400 U).

(h) 2 μL 10× T4 DNA ligase buffer.

(i) H_2O up to 20 μL.

3. Place the Golden Gate reaction in a thermocycler and run the following program: 10× (37 °C for 5 min + 16 °C for 10 min) + 37 °C for 15 min + 80 °C for 5 min + 4 °C hold (*see* **Note 15**).
4. Transform competent *E. coli* with 5 μL of the Golden Gate reaction and plate on LB + 50 mg/L kanamycin. Incubate at 37 °C overnight.
5. Inoculate 1–2 colonies into 5 mL LB + 50 mg/L kanamycin and incubate overnight in a 37 °C shaking incubator (*see* **Note 16**).
6. Isolate plasmid DNA via miniprep and confirm by Sanger sequencing with primers spanning each Golden Gate cloning junction (*see* **Note 17**): M13R(-48): 5′-CGGATAACAATTTCACACAG-3′ (spanning T-DNA to module A), TC430: 5′-GTTGGATCTCTTCTGCAGCA-3′ (spanning module A to module B), NB457: 5′-CAAGAATTGGGACAACTCCAG-3′ (spanning module B to module C), and NB463: 5′-CGAACGGATAAACCTTTTCACG-3′ (spanning T-DNA to module C).

3.1.6 Design and Construction of GVR Vectors for Targeted Sequence Insertion/Replacement

Gene targeting can generally be used to make two types of modifications: targeted insertion or targeted replacement. In targeted insertion, the sequence of interest is inserted into a specific site in the genome and no genomic sequence is lost/removed. In targeted replacement, a genomic region ranging in length from single bases to whole genes is replaced with a predesigned sequence containing custom modifications. Although a single targeted double-strand break (DSB) is sufficient to induce both targeted insertion and replacement, the latter is more effectively achieved through excision of the target sequence region spanning all the intended modifications via two DSBs, especially when two or more modifications that are distant from each other are being introduced.

During gene targeting, the DSB(s) are repaired using an external DNA donor template. The donor sequence is designed to contain the desired modifications, flanked on each side with a homology arm: 500 bp to 1 kb of sequence matching the genomic region immediately adjacent to the targeted DSB(s) to allow for

annealing and repair of the broken DNA strand. In addition to the desired modifications, the donor template should also contain synonymous mutations to disrupt the PAM sequence(s) of the sites targeted by sgRNA(s) to prevent cleavage of the donor sequence. Such donor templates can be created by PCR amplification of the target loci in two or more overlapping fragments from genomic DNA of the target species, followed by Gibson assembly. The sequence modifications are introduced in the overlaps included in synthetic primer extensions as shown in Fig. 3. Alternatively, the full-length donor template can be commercially synthesized. Many companies provide DNA fragment synthesis services such as Gene Fragments from Twist Bioscience (https://www.twistbioscience.com) or gBlocks from Integrated DNA Technologies (https://www.idtdna.com).

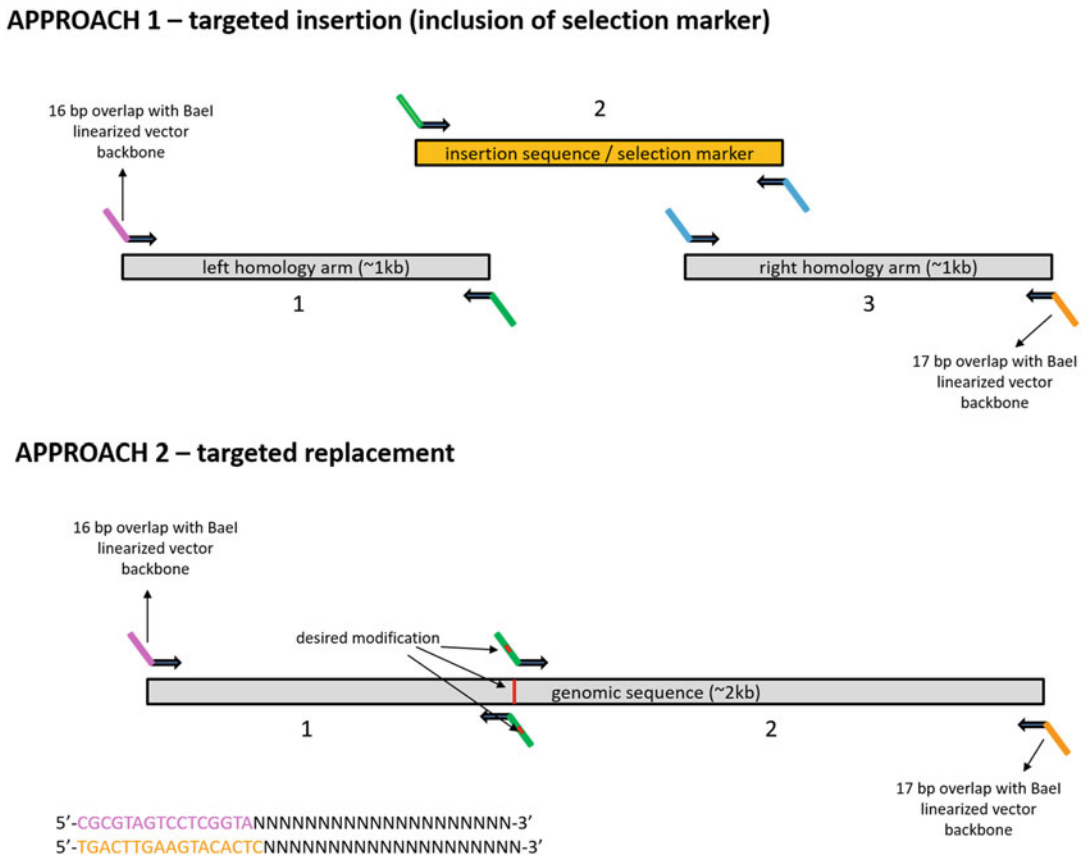


Fig. 3 Primer design for donor template assembly. Three (targeted insertion) or two (targeted replacement) parts are amplified using extended primers. Primer extensions labeled in the same color represent overlapping sequences. Overlaps with the vector backbone are labeled in pink and the sequence is specified on the bottom of the image. The Ns are replaced with the sequence complementary to the 5′ end of the left and 3′ end of the right homology arm. (Figure modified from [18], Supplemental Methods—Protocols)

To incorporate a donor template into a GVR gene targeting construct, it is first cloned into the module C vector pMOD_C0000 by Gibson assembly and then Golden Gate assembled with Cas9 in a module A vector and one sgRNA (for targeted insertion) or two sgRNAs (for targeted replacement) in a module B vector into the GVR T-DNA backbone pTRANS_221.

1. Create module B vector with one or two sgRNAs for targeted insertion or replacement, respectively. Follow **step 1** of Subheading 3.1.5, and then proceed to **step 2** below. Note that the protocol for the construction of six sgRNA vectors can be scaled down to two sgRNAs simply by starting with sequences for two targets instead of six. This procedure can be done in parallel with **steps 2–10** below.
2. Design and synthesize primers, and then PCR amplify parts of the donor template. Use Fig. 3 as a guideline (*see* **Notes 18** and **19**).
3. Gel purify the fragments of the correct size and measure the concentrations using NanoDrop spectrophotometer. **Steps 2** and **3** can be done in parallel with **steps 4** and **5**.
4. To linearize the module C vector backbone for Gibson assembly, set up the following digest:
   (a) 2–3 μg pMOD_C0000 plasmid DNA.
   (b) 2 μL of restriction enzyme buffer.
   (c) 1 μL *Bae*I restriction enzyme.
   (d) 1.25 μL 320 μM SAM (*S*-adenosylmethionine).
   (e) H_2O up to 20 μL.

   Incubate the reaction at 25 °C for 1 h and run on an agarose gel.
5. Excise and purify the linearized backbone (2091 bp) using DNA gel purification kit. Measure DNA concentration using spectrophotometer.
6. Insert the donor template into the linearized pMOD_C0000 via the following Gibson assembly reaction:
   (a) 10 μL of 2× Gibson Assembly Master Mix.
   (b) 50 ng of the *Bae*I linearized pMOD_C0000 from the previous step.
   (c) Equimolar amounts of each purified PCR fragment from **step 3** in threefold molar excess over the module C vector backbone.
   (d) H_2O up to 20 μL.

   Incubate the reaction at 50 °C for 1 h.

7. Transform competent *E. coli* with 5 μL of the Gibson assembly reaction and plate on LB + 50 mg/L carbenicillin/ampicillin. Incubate at 37 °C overnight.
8. PCR screen 3–5 colonies using Taq DNA polymerase, a forward primer specific to the donor template sequence, and the reverse primer ZY015F: 5′- GGAATAAGGGCGACACG GAAATG-3′.
9. Inoculate 2–3 positive colonies into 5 mL LB + 50 mg/L carbenicillin/ampicillin and incubate overnight in a 37 °C shaking incubator.
10. Isolate plasmid DNA via miniprep and confirm by Sanger sequencing with primers NB424: 5′- CAGCGAGTCAGT GAGCGAGGAAGC -3′, ZY015F: 5′- GGAATAAGGGCGA CACGGAAATG-3′, and/or donor template-specific primers.
11. Assemble the final GVR T-DNA gene targeting vector via the following Golden Gate reaction:
    (a) 75 ng pTRANS_221 plasmid (transformation backbone).
    (b) 150 ng pMOD_A0101 (for vectors with a single sgRNA) or pMOD_A0501 (for vectors with two sgRNAs)—module A plasmid with Cas9.
    (c) 150 ng Sequence confirmed module B plasmid with one or two sgRNAs (**step 1**).
    (d) 150 ng Sequence confirmed module C plasmid with the donor template from the previous step.
    (e) 0.5 μL *Aar*I.
    (f) 0.4 μL *Aar*I oligonucleotide (comes with the *Aar*I enzyme).
    (g) 1 μL T4 DNA ligase (400 U).
    (h) 2 μL 10× T4 DNA ligase buffer.
    (i) H_2O up to 20 μL.
12. Place the Golden Gate reaction in a thermocycler and run the following program: 10× (37 °C for 5 min + 16 °C for 10 min) + 37 °C for 15 min + 80 °C for 5 min + 4 °C hold.
13. Transform competent *E. coli* with 5 μL of the Golden Gate reaction and plate on LB + 50 mg/L kanamycin. Incubate at 37 °C overnight.
14. Inoculate 1–2 colonies into 5 mL LB + 50 mg/L kanamycin and incubate overnight in a 37 °C shaking incubator.
15. Isolate plasmid DNA via miniprep and confirm by Sanger sequencing with primers spanning each Golden Gate cloning junction (*see* **Note 11**): M13R(-48): 5′-CGGATAACAATTT CACACAG-3′ (spanning T-DNA to module A), TC430: 5′-GTTGGATCTCTTCTGCAGCA-3′ (spanning module A to

module B), TC214F: 5′-GAAGAGAAGCAGGCCCATTTAT-3′—one sgRNA **OR** TC320: 5′- CTAGAAGTAGT CAAGGCGGC -3′—two sgRNAs (spanning module B to module C), and NB442: 5′- GCAATCCTGACGAAGACTG GATGT-3′ (spanning T-DNA to module C).

3.2 Tomato Transformation, Regeneration, and Molecular Characterization of Gene-Edited Events

The following protocol has been successfully used to transform some of the most common commercial cultivars of tomato, including M82, Moneymaker, and Ailsa Craig [19].

3.2.1 Introducing the CRISPR/Cas9 Vectors into Agrobacterium

1. Thaw on ice 40 μL aliquots of *Agrobacterium* electrocompetent cells stored at −80 °C.
2. Add 0.5–1.0 μg plasmid and mix by gentle tapping.
3. Transfer mix to electroporation cuvette and incubate on ice for 3–5 min.
4. Pulse in electroporator set to 1.8 kV voltage and 400 Ω resistance (for a 0.1 cm cuvette gap).
5. Immediately after pulse add 1 mL of SOC (or LB) medium and transfer to a 1.5 mL microcentrifuge tube. Incubate horizontally with gentle shaking at 28 °C, 140–180 rpm, for 4 h (*see* **Note 20**).
6. Plate 50 and 100 μL of electroporated cells on plates with LB containing the appropriate antibiotic. Seal plate with wrap and incubate for 48–72 h at 28 °C. Plates can be then stored for 1–2 weeks at 4 °C (*see* **Note 21**).
7. Confirm correctly transformed colonies by PCR with gene-specific primers.

3.2.2 In Vitro Propagation and Tomato Explant Preparation

1. Sterilize tomato seed in sodium hypochlorite solution (2.5% v/v) plus two drops of Tween 20 for 15 min on a magnetic stirrer.
2. In laminar flow, wash the seeds with ddH_2O. Use a sieve to aid the process.
3. Place 30 seeds in a 500 mL flask containing 40 mL half-strength MS medium. Seal flask with plastic film. Store in the dark at room temperature for 4 days, and then transfer to growth room under 50 μmol/m^2/s irradiance, 16-h photoperiod, and 25 ± 2 °C.
4. After 8 days of seed inoculation, cut cotyledon explants with a scalpel (*see* **Note 22**) and place 20 explants (abaxial side down) on a Petri dish containing 20 mL of root inducer medium (RIM).

3.2.3 Agrobacterium Inoculum Preparation and Tomato Transformation

The inoculum is prepared from transformed and isolated colonies previously grown in solid LB medium with appropriate antibiotics. The inoculum preparation must be started 2 days before starting the transformation.

1. Using a single colony from the previously prepared or freshly streaked transformed *Agrobacterium* plate, prepare a suspension in 5 mL liquid LB medium with appropriate antibiotics and incubate at 28 °C in an orbital shaker at 140–180 rpm for 16–24 h (*see* **Note 20**).
2. From this initial culture, collect 100 μL and add to 50 mL fresh liquid LB medium with appropriate antibiotics, prepared in a sterile flask (*see* **Note 23**).
3. Incubate the *Agrobacterium* culture overnight at 28 °C in orbital shaker at 140–180 rpm (*see* **Note 20**).
4. Take 1 mL aliquots from the suspension and check OD_{600} on a spectrophotometer (*see* **Note 24**). Centrifuge the cultures at 4000 rpm (112 × *g*) for 15 min. Resuspend in an appropriate volume of MS medium so that OD_{600} is between 0.3 and 0.4. Make sure that the volume is sufficient for the transformation of all explants. Add acetosyringone (final concentration of 100 μM) and wait for 10 min before starting the explant inoculation. Invert gently over 10 min.
5. Dispense two drops of *Agrobacterium* suspension per explant (previously distributed on plates with RIM) using a micropipette (*see* **Note 25**). Incubate for 10 min at room temperature. Slightly tilt the plate with the explants and remove the excess of *Agrobacterium* suspension with a pipette. Dry with sterile paper filter. Seal all plates with parafilm.
6. After 2 days of co-cultivation in the dark at room temperature, transfer the explants to plates with 20 mL of shoot inducing medium (SIM). Cultivate under 20–30 μmol/m^2/s irradiance, 16-h photoperiod, at 25 ± 1 °C for 3 weeks. If necessary, carry out subculture every 2 weeks after that period.
7. After the emergence of shoots (2–5 mm), gently cut the explant and transfer shoots to flasks containing 30 mL hormone-free general propagation medium supplemented with appropriate antibiotics. Seal the flasks with plastic film and identify the transformation events (*see* **Note 26**). Incubate the flasks for under 20–30 μmol/m^2/s irradiance, 16-h photoperiod, at 25 ± 1 °C for 3 weeks.
8. Transfer the well-rooted seedlings to pots with autoclaved soil (no fertilizer); keep covered with a plastic bag or with the top of the PET bottle cut in half for a few days. Cut the end of the plastic bag or remove the cap from the PET bottle after approximately 4–5 days. Pay attention to the development and

acclimatization of the plants before completely removing the plastic bag or PET bottle. After this period, the plants can be grown in a greenhouse, under shade for protection against radiation in the first days.

3.2.4 Identification and Isolation of Transgenic Events

Positive transformation events can be confirmed by kanamycin spray on T_0 acclimatized plants followed by PCR.

Kanamycin Screening

Kanamycin solution can be used for a preliminary visual selection and identification of candidate transformed plants.

1. Prepare a kanamycin solution (400 mg/L) and place in a garden sprayer.
2. Spray the solution on acclimatized plants, taking care to completely wet all the leaves. Repeat this step on 3–5 consecutive days, preferably spraying at the end of the light period.
3. Plants that do not show yellowing in the leaves 1 week after spraying with kanamycin solution can be selected as potentially transgenic candidates. Yellowing will be most evident in young leaflets closer to the apical meristem region.

PCR Screening

1. Collect young leaflets for DNA extraction (commercial DNA extraction kits or a standard laboratory protocol can be used).
2. Use the genomic DNA as template for PCR with Cas9- or kanamycin-specific primers.
3. Run PCR products on an agarose gel to confirm the correct size of the product.

3.2.5 Molecular Characterization of Transgenic Events (See **Note 27**)

1. PCR amplify 400–800 bp of the target locus that harbors the potential mutation using genomic DNA isolated from the transgenic lines. For amplification and sequencing of the target region, it is recommended to use a high-fidelity polymerase (see **Note 28**).
2. Verify the integrity and size of the product by agarose gel electrophoresis.
3. Purify the PCR product with an appropriate kit and quantify the purification product.
4. Calculate the appropriate amount of PCR product (insert) to include in the ligation reaction, following the equation

$$\frac{\text{ng Vector} \times \text{kb size insert}}{\text{kb size of vector}} \times \text{insert : Vector molar ratio} = \text{ng of insert}$$

Table 1
Components of a standard ligation reaction of PCR products into a cloning vector for Sanger sequencing

Reaction component	Standard reaction	Positive control	Background control
2× Rapid ligation buffer	5 μL	5 μL	5 μL
pGEM®-T Easy vector	1 μL	1 μL	1 μL
PCR product	*X* μL[a]	–	–
Control insert DNA	–	2 μL	–
T4 DNA ligase	1 μL	1 μL	1 μL
Nuclease-free water to a final volume of	10 μL	10 μL	10 μL

[a]Amount of insert from **step 3** of this section

5. Ligate the purified PCR product to the cloning vector (the steps below refer to cloning in pGEM®-T Easy vector—Promega). Set up the following reactions (*see* **Note 29**) (Table 1).
6. Transform *E. coli* Top 10 competent cells with 5 μL of the ligation reaction and plate on LB + 50 mg/L carbenicillin/ampicillin, 40 μL of X-Gal (20 mg/mL), and 10 μL of IPTG (100 mM). X-Gal and IPTG are spread on the surface of the solidified LB media and allowed to dry. Incubate at 37 °C overnight.
7. PCR screen 5–10 white colonies using Taq DNA polymerase, and the primers specific to the vector backbone that come with the kit.
8. Inoculate 3–5 positive colonies into 5 mL LB + 50 mg/L carbenicillin/ampicillin and incubate overnight in a 37 °C shaking incubator.
9. Isolate plasmid DNA via miniprep and analyze by Sanger sequencing with the primers specific to the vector backbone that come with the kit.

4 Notes

1. Either Phusion® (Thermo Fisher) or Q5® (New England Biolabs) high-fidelity DNA polymerases are equally suitable. Here, we have described the protocol for the former as per the manufacturer's instructions.
2. For instance, pGEM®-T Easy (Promega).
3. It is preferable to add the medium to a fresh bottle containing agar to avoid uneven distribution and sedimentation onto the bottom.

Table 2
Some commonly used antibiotics and their recommended concentration for *Agrobacterium* selection

Antibiotic	Stock solution (mg/mL)	For 1 mL of MS medium (μL)	Final concentration in MS medium (mg/L)
Kanamycin	50	2	100
Ampicillin	100	0.5	50
Spectinomycin	100	1	100
Streptomycin	100	3	300
Rifampicin	50	1	50
Timentin	300	1	300

4. Examples of antibiotics and their concentration used for *Agrobacterium* selection (Table 2).
5. The following restriction enzymes are used in the cloning process: *AarI* (pDIRECT_22A, pMOD_B2515, pMOD_B2103), and *Esp3I* and *SapI* (pDIRECT_22C, pMOD_B2515, pMOD_B2103). For this reason, the sgRNA target sequence must not contain these restriction sites or start with 5′-GTG-3′ (a sequence that creates an *AarI* site after cloning into the pMOD_B2103 or pDIRECT_22C vectors).
6. Transcription from the *Arabidopsis* U6 RNA polymerase III (*At*U6) promoter used in the pDIRECT_22A and pMOD_B2515 vectors for cloning of one sgRNA starts preferentially at a guanine (G) nucleotide. Target sites starting with a G (5′-GN_{19}-NGG-3′) should be selected when using these vectors. This does not apply to the pDIRECT_22C or pMOD_B2103 vectors for the cloning of six sgRNAs. These vectors use the RNA polymerase II promoter CmYLCV and can start transcription from any nucleotide.
7. For this book, we only describe the protocols for assembly of T-DNA vectors, but a large selection of other pDIRECT and modular vectors is available to create other types of constructs suitable for biolistic or protoplast transformation, or for use with other selectable or visible markers. For details, see http://crispr-multiplex.cbs.umn.edu/.
8. This program is usually sufficient to recover colonies with the correctly assembled vector. If correct clones cannot be obtained, the efficiency of the cloning reaction can be increased by running 5–10 cycles of the restriction-ligation steps (37 °C for 5 min + 16 °C for 10 min).

9. Note that this protocol can be used for assembly of vectors expressing any number of sgRNAs between 2 and 6 by simply starting with the desired number of target sequences. Protocols for assembly of vectors expressing more than six sgRNAs are also available http://crispr-multiplex.cbs.umn.edu/ but are out of the scope of this book chapter.
10. In our experience, this is the most efficient combination of parameters for the assembly of effective multiplex gene editing vectors. However, it is possible to modify the protocol if desired, including options to use different expression backbones, promoters, restriction enzymes, and splicing systems.
11. If available, pMOD_B2103 can also be used (as the PCR template). It is a smaller plasmid and produces a smaller number of smaller fragments upon digestion with *BanI*, which might simplify the purification. The size of the largest fragment (to be purified) is 1591 bp.
12. T4 DNA ligase and the respective buffer can be substituted for T7 DNA ligase and buffer, although this may result in lower efficiency of Golden Gate assembly.
13. *SapI* tends to settle down in the tube. Mix the enzyme solution by pipetting up and down several times before adding to the reaction. In our experience, *SapI* easily loses activity over time. In the case of Golden Gate reaction failure, we recommend making sure that *SapI* is active by running a test digest.
14. This reaction must NOT be heat-inactivated. The PEG in the T7 reaction buffer may have a negative impact on the viability of *E. coli* after heating.
15. The efficiency of the Golden Gate reaction may be further increased by increasing the number of cycles up to 20.
16. Colony PCR is usually not necessary due to the high efficiency of the Golden Gate reaction. Optionally, the junction spanning module C to T-DNA backbone can be verified using primers TC370: 5′-AATGTGTCAAATCGTGGCCT-3′ and NB463: 5′-CGAACGGATAAACCTTTTCACG-3′.
17. In addition to Sanger sequencing, we recommend verifying the integrity of the construct by an analytical digest.
18. The length of each Gibson overlap/primer extension should be at least 16 nt with a melting temperature (Tm) of at least 48 °C. Tm can be calculated by adding 2 °C for each AT pair and 4 °C for each GC pair. Start at 16 nt and extend the sequence until it reaches the Tm of 48 °C or higher.
19. Sequences containing *Aar*I sites should be avoided when designing donor templates. Alternatively, *Aar*I sites can be mutated in the donor sequence by introducing synonymous mutations.

20. Shaker speed and growth time can be adjusted based on *Agrobacterium* growth, which can be checked periodically by determining OD_{600}.
21. *Agrobacterium* colonies on plates will have reduced viability and growth capacity if stored for long periods. It is ideal to always prepare permanent stocks by inoculating a single colony from the plate into liquid LB medium containing the appropriate antibiotic, incubating for 16 h, 140–180 rpm, at 28 °C. In microtubes containing 50% filtered glycerol, add 500 μL of the solution containing *Agrobacterium*. Microtubes are then snap-frozen in liquid nitrogen and stored at −80 °C.
22. Explants must be cut in Petri dishes containing sterilized filter paper and hydrated with ddH_2O. The base and tip of cotyledons must be removed, and then the cotyledons must be cut transversally in two or three pieces.
23. This step is to ensure that the bacterial cultures are in the active log phase of growth; grow at least two flasks of 50 mL culture, one with 50 μL and the other with 100 μL of the initial culture medium with *Agrobacterium*.
24. The measurement of OD_{600} before centrifugation is recommended to calculate the amount of MS medium required for dilution of the pellet. We recommend the use of initial OD_{600} between 6 and 1.0. Values greater than 1.0 correspond to the stationary phase or bacterial death and may compromise the transformation.
25. Make sure that the explants are in contact with the suspension, especially the cut sides. If necessary, add a few more drops around the explants.
26. The same explant can generate more than one regeneration event, which may or may not differ with respect to the mutation. In that case, it is necessary to separate them for further analysis.
27. This protocol describes characterization by Sanger sequencing. Alternatively, plants can be analyzed by amplicon sequencing or T7EI assay.
28. If analyzing precise edits, make sure that the primers bind to the genome region outside of the donor homology arms used in the GVR to prevent amplification of the donor template.
29. If cloning blunt-ended PCR products generated by proofreading polymerases, follow the A-tailing procedure described in the pGEM®-T Easy manufacturer's protocol. Alternatively, CloneJET PCR Cloning Kit (Thermo Fisher) can be used to directly clone blunt PCR products without further modifications.

References

1. Donald CM (1968) The breeding of crop ideotypes. Euphytica 17:385–403
2. Martre P, Quilot-Turion B, Luquet D et al (2015) Model-assisted phenotyping and ideotype design. In: Crop physiology. Academic, Cambridge, pp 349–373
3. Peng S, Khush GS, Virk P et al (2008) Progress in ideotype breeding to increase rice yield potential. Field Crops Res 108:32–38
4. Zsögön A, Cermak T, Voytas D et al (2017) Genome editing as a tool to achieve the crop ideotype and de novo domestication of wild relatives: case study in tomato. Plant Sci 256:120–130
5. Horvath P, Barrangou R (2010) CRISPR/Cas, the immune system of bacteria and archaea. Science 327:167–170
6. Barrangou R, Fremaux C, Deveau H et al (2007) CRISPR provides acquired resistance against viruses in prokaryotes. Science 315:1709–1712
7. Doudna JA, Charpentier E (2014) The new frontier of genome engineering with CRISPR-Cas9. Science 346:1258096
8. Jinek M, Chylinski K, Fonfara I et al (2012) A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science 337:816–821
9. Fauser F, Roth N, Pacher M et al (2012) In planta gene targeting. Proc Natl Acad Sci U S A 109:7535–7540
10. Baltes NJ, Gil-Humanes J, Cermak T et al (2014) DNA replicons for plant genome engineering. Plant Cell 26:151–163
11. Čermák T, Baltes NJ, Čegan R et al (2015) High-frequency, precise modification of the tomato genome. Genome Biol 16:232
12. Vu TV, Sivankalyani V, Kim E-J et al (2020) Highly efficient homology-directed repair using transient CRISPR/Cpf1-geminiviral replicon in tomato. Plant Biotechnol J. https://doi.org/10.1111/pbi.13373
13. Dahan-Meir T, Filler-Hayut S, Melamed-Bessudo C et al (2018) Efficient in planta gene targeting in tomato using geminiviral replicons and the CRISPR/Cas9 system. Plant J 95:5–16
14. Olsen KM, Wendel JF (2013) A bountiful harvest: genomic insights into crop domestication phenotypes. Annu Rev Plant Biol 64:47–70
15. Wolter F, Schindele P, Puchta H (2019) Plant breeding at the speed of light: the power of CRISPR/Cas to generate directed genetic diversity at multiple sites. BMC Plant Biol 19:1–8
16. Zsögön A, Čermák T, Naves ER et al (2018) De novo domestication of wild tomato using genome editing. Nat Biotechnol 36:1211–1216
17. Fernie AR, Yan J (2019) De novo domestication: an alternative route toward new crops for the future. Mol Plant 12:615–631
18. Čermák T, Curtin SJ, Gil-Humanes J et al (2017) A multipurpose toolkit to enable advanced genome engineering in plants. Plant Cell 29:1196–1217
19. Van Eck J (2018) Genome editing and plant transformation of solanaceous food crops. Curr Opin Biotechnol 49:35–41

Chapter 17

Root System Phenotying of Soil-Grown Plants via RGB and Hyperspectral Imaging

Gernot Bodner, Mouhannad Alsalem, and Alireza Nakhforoosh

Abstract

Phenotyping root systems provide essential information for plant breeding, particularly aiming for better abiotic stress resistance. Rhizobox systems provide a field-near growth environment for in situ imaging of root systems in soil. A protocol for RGB and hyperspectral imaging of rhizobox-grown plants is presented that enables gathering of root structural (morphology, architecture) as well as functional (water content, decomposition) information. The protocol exemplifies the setup of a root phenotyping platform combining low-cost RGB with advanced short-wave infrared hyperspectral imaging. For both types of imaging approach, the essential steps of an image analysis pipeline are provided to retrieve biological information on breeding-relevant traits from the imaging datasets.

Key words Root phenotyping, Hyperspectral imaging, Root architecture analysis, Image processing, Rhizobox platform

1 Introduction

Root systems are increasingly recognized as essential targets for crop improvement. Cultivars with superior root systems are expected to enhance yield stability, resource-use efficiency, and resistance against environmental stresses (e.g., [1, 2]). Awareness of the role of root systems as breeding target is rather recent. In addition, breeding for better root systems can potentially make use of high diversity among plant genetic resources [3]. Therefore, it can be expected that root systems constitute an underutilized organ with a still large margin to improve its efficiency when explicitly included as breeding targets for novel cultivars. However, root systems are complex and multivariate compound organs with high spatiotemporal variability and strong interaction with the environment. Defining target traits for improved root systems is therefore challenging and largely dependent on an accurate definition of the pedo-climatic conditions of the target environment [4–6].

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_17,

The major bottleneck for the integration of root sciences and breeding relies upon the shortcoming of methodologies to efficiently study roots at the throughput requirements of breeding. Numbers in plant breeding strongly contrast the throughput of traditional and most of the current (field) root sampling and measurement methods. Thus, overcoming the root measurement bottleneck is one of the key motivations for recent efforts in the development of new imaging technologies and platforms for phenotyping root systems [7, 8]. The need for detailed insights into traits constituting root architecture and functioning as well as the demand for field-near phenotyping conditions has also motivated the development of a soil-filled rhizobox phenotyping system for RGB and hyperspectral root imaging at the University of Natural Resources and Life Sciences, Vienna.

Rhizobox imaging and analysis of root architecture are most commonly based on RGB images in the visible (VIS) range. Currently, however, also other spectral domains beyond VIS are explored for functional root phenotyping. Here we detail the two types of root imaging established at the BOKU root phenotyping platform: (1) a simple RGB imaging setup with a common digital camera and a custom-made dark chamber, and (2) a more complex shortwave infrared (SWIR) hyperspectral imaging setup using a scanner system.

The key advantage of RGB imaging is (1) low cost of hardware, (2) high resolution of camera systems for capturing even small root structures, (3) low data load of captured images, and (4) moderate requirements for data handling and analysis and availability of several image analysis tools.

Hyperspectral imaging allows targeting chemical properties of soil-grown root systems via the specific absorption properties of biochemical substances. In the case of rhizobox-grown plants, some of the key questions addressed are, for example, spatially resolved quantitative mapping of water in soil and roots, organic molecules composing plant tissues [9], and soil organic matter [10]. The advantage of hyperspectral systems to image functional root traits beyond RGB-based determination of root morphology and architecture, however, implies (1) high costs of the imaging setup, (2) longer image acquisition time (about 16 min per rhizobox), (3) higher requirement of data storage (13.7 GB per rhizobox image), (4) lower resolution of the hyperspectral camera (320×256 Pixel), and (5) more complex image processing and analysis.

The objective of this protocol is to provide general guidelines for root imaging and image analysis based on the gathered experiences with different experimental questions obtained so far in the BOKU root phenotyping platform.

2 Materials

Phenotyping root system architecture of soil-grown plants requires adequate experimental systems for in situ observation. These systems, commonly named rhizoboxes or rhizotrons, allow imaging all plant root axes growing along a transparent front side, with the average proportion of surface-visible roots between 20% and 90% of total root length depending on plant species [11]. Various types of rhizoboxes and imaging methods are in use in different phenotyping platforms. The following section describes the rhizobox type and imaging systems of the BOKU root phenotyping platform.

2.1 Rhizobox Construction and Filling

1. Rhizobox material: Rhizoboxes (Fig. 1) are built from two mineral glass plates (8 mm) and side frames made of gray PVC with 15 mm strength. The two glass sides allow imaging from both sides, if necessary/desirable. The glass backside is permanently fixed to the PVC frame by metal rails screwed into the sidewalls, while the front side is removable for opening and (horizontal) filling. At the bottom, the PVC frame contains three holes for drainage of excess water. Alternatively, also rhizoboxes are in use with the backside constructed of gray PVC plates.

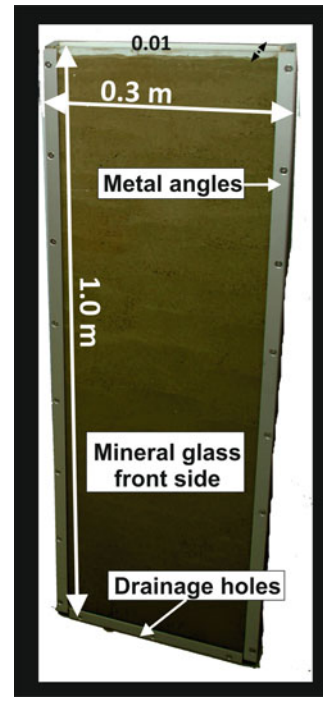


Fig. 1 Rhizoboxes used for plant growth and root imaging at the BOKU root phenotyping platform

2. Rhizobox size: The size of rhizoboxes is 30 × 100 cm with 1 cm inner diameter (weight: 12.4 kg). Generally, the dimension of rhizoboxes should be adapted to the prevailing experimental questions studied in the phenotyping platform. Main questions for fixing rhizobox size are (a) the scheduled duration of an experiment (with larger size for experiments aiming to capture later phenological stages), (b) the expected maximum horizontal/vertical extensions of the root system, and (c) the number of plants in a rhizobox (i.e., single plants or [multiple] neighboring plants). Rhizoboxes are a 2D phenotyping system (i.e., imaging across the transparent front window captures a 2D image section of the surface-visible root architecture). Increasing the inner diameter >1 cm to allow spatial extension into a 3D volume would only be a reasonable choice in case of tap-rooted plants with higher space constraint due to secondary thickening. Larger size of rhizobox systems increases their weight and thereby the ease of handling. For example, the weight of a rhizobox of 3000 cm^3 filled with soil at a bulk density of 1.3 g/cm^3 and water content at field capacity is about 17.3 kg.
3. Hydraulic considerations: Rhizoboxes are suitable systems for phenotyping plants growing in field soil without necessity to supplement with sand and/or organic material (e.g., peat). This, however, implies that rhizobox height must be sufficient to ensure aeration of the plant root zone: height is equivalent to the gravitational potential distribution of water in the rhizobox and, together with the soil water retention curve of the substrate, determines the volume of air- vs. water-filled pore space. For example, in a rhizobox of 100 cm height, at equilibrium (i.e., when no water drains out of the rhizoboxes) the gravitational potential at the top of the rhizobox is 100 cm (=100 hPa) and corresponds to an equivalent (negative) matrix potential of −100 hPa. In case of a water retention curve as shown in Fig. 2, the water content at the top of the rhizobox is 33 Vol.% vs. 12 Vol% for soil and sand, respectively, while the air-filled pore space is 12 Vol.% vs. 35 Vol.%. This size consideration is particularly relevant in case of fine-textured soil where fast-draining macropores are a minor proportion of the whole pore size distribution and thus reduced height can risk oxygen stress. Also for drought-stress settings, it is evident that sand is not an adequate substrate for a rhizobox due to the low equilibrium water content. For details on hydraulic considerations in plant experiments *see* ref. 12.
4. Substrate: The rhizoboxes are filled with field soil which is sieved to 2 mm particle size. Other substrates with different grain size and composition can be used too, according to the prevailing experimental question.

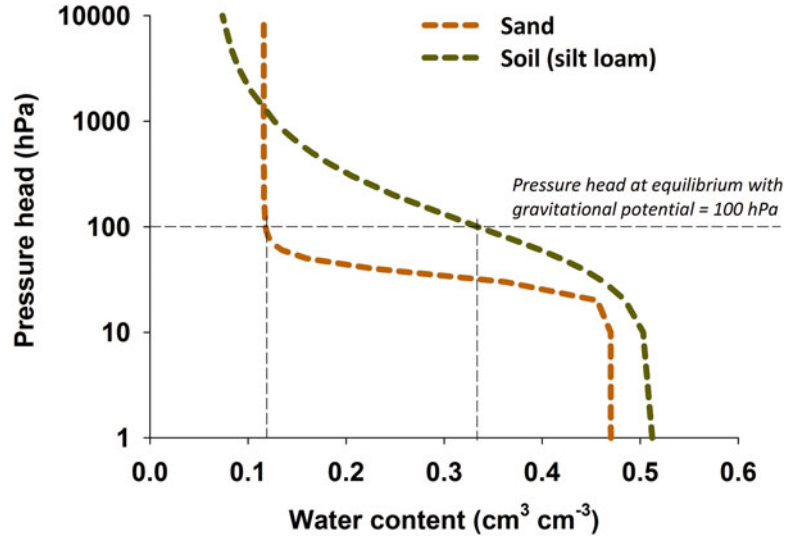


Fig. 2 Soil water retention curve for field soil and sand used in rhizobox experiments

5. Filling of rhizoboxes: Filling is done in horizontal position into the opened rhizoboxes using prewetted substrate. Horizontal filling avoids vertical layering and segregation between fine and coarse particles compared to filling the rhizoboxes in vertical position by pouring the substrate through the upper opening. Still, even for rhizoboxes of small inner diameter and with prewetting of the substrate, a certain degree of unmixing cannot be avoided. Therefore, prewetted substrate is recommended for filling to minimize the unmixing of fine particles. Depending on the type of substrate (particularly silt and clay content), premixing the substrate with water, however, is only possible up to a limited water content to prevent from smearing and structure degradation. The difference in water between the premixing and the target moisture content is added using a spray bottle with fine nozzle (i.e., providing small-sized water drops to prevent aggregate disruption) after filling the rhizoboxes. An example calculation protocol for rhizobox filling with a defined target bulk density and water content is given in Table 1.

   Finally, we point to the fact that the homogeneous water content at filling the horizontally positioned rhizoboxes will change/redistribute when setting the boxes in their final position following the resulting potential gradient. This is a physical process occurring in all plant growth systems, similar to field environments, with water content and pressure head distribution following the geometry (height) of the system as well as the hydraulic properties of the substrate.

Table 1
Setting of water regime in rhizoboxes and filling protocol with respective amounts of soil and water and the resulting final weight for monitoring irrigation/transpiration

Hydraulic properties and settings	
Target water level control (% PAW)	80
Target water level drought (% PAW)	40
$FC_{v.v.}$ (cm^3/cm^3)	0.350
$PWP_{v.v.}$ (cm^3/cm^3)	0.120
80% $PAW_{v.v.}$ (cm^3/cm^3)	0.304
40% $PAW_{v.v.}$ (cm^3/cm^3)	0.212
Bulk density (g/cm^3)	1.30
$FC_{w.w.}$ (g/g)	0.269
$PWP_{w.w.}$ (g/g)	0.092
80% $PAW_{w.w.}$ (g/g)	0.234
40% $PAW_{w.w.}$ (g/g)	0.163
Filling data	
Filling weight dry soil dry (g per rhizobox)[a]	3705.0
Prewetting water amount (g)	400.0
$WC_{w.w.}$ after prewetting (g/g)	0.108
Additional water for control level (g)	466.4
Additional water for drought level (g)	204.2
Weight for irrigation and transpiration monitoring	
Empty weight of rhizobox (g)	12,400
Weight of rhizobox + soil at target WC control treatment	16,971.4
Weight of rhizobox + soil at target WC drought treatment	16,409.2

[a]A margin of 5 cm is left at the top of the rhizoboxes to facilitate watering of the inclined boxes; therefore, the rhizobox soil volume is 2850 cm^3

6. Rhizobox positioning and root visibility: Rhizoboxes are put in a metal framework, holding them at an inclination between 35° and 45°. While higher angles maximize gravitropism-driven root visibility, they also increase space requirements for the experimental setup (e.g., in climate rooms) and reduce the gravitational (and corresponding matrix) potential in the rhizoboxes with the resulting hydraulic behavior more distant to field conditions. The transparent glass sides of rhizoboxes have to be covered, e.g., using wooden plates or black foil, to keep the root zone in the dark and avoid algae growth due to light coming to the glass surface.

We frequently observed that besides gravitropism, root visibility is strongly influenced by aggregation/air spaces: Roots tend to make use of pathways of lower mechanical resistance, thus growing preferentially into larger pores, including space between the glass observation window and soil. Consequently, we found that root visibility was higher in more aggregated/coarser soil compared to finer particle sizes. Thus an enhanced root-soil contact with fine material at the observation window could be at the expense of root visibility.

2.2 Plant Establishment and Experimental Settings

1. Sowing: Rhizoboxes are generally sown with one seed per box in order to optimize the identification of root architecture. However, depending on the experimental question (e.g., root competition), also two or more plants per box can be grown. Root overlap, however, makes architectural quantification more difficult in case of multiple plants. In addition, space constraints can restrict the lateral expansion of root axes and thereby bias an accurate architectural description of root architecture.

   For plant establishment the seeds are either pre-germinated or directly sown. Pre-germination avoids the sowing of nonviable plants and improves homogeneity of plants in the rhizoboxes. Sowing of pre-germinated seeds, however, requires much care for not harming the radicle. Therefore, in case of seeds with high germination rate and vigor (e.g., modern cultivars), we often used direct sowing without pre-germination to avoid the risk of damaging the radicle and minimize risks of fungal infections during pre-germination time.

2. Management of rhizoboxes: Management (irrigation, fertilization) depends on the specific experimental question. Accurate irrigation is done by regularly weighing the rhizoboxes and adding water to a preset level. Fertilizer is provided in liquid form together with irrigation water, except in case of experimental treatments involving nutrient deficiencies and/or specific fertilizer type/application.

2.3 RGB Imaging Setup

1. Black box: A black box (Fig. 3) is used to shield from ambient light during image acquisition, homogeneously illuminate the rhizobox surface, and fix the camera in a defined position for image capture.

   The custom-made black box consists of a metal frame (1.5 m width × 1 m height for the rhizoboxes described above) with pressboard plates fixed to the sides. At the front, the camera is mounted at two positions with a distance of 80 cm from the rhizobox. Depending on the geometry of the

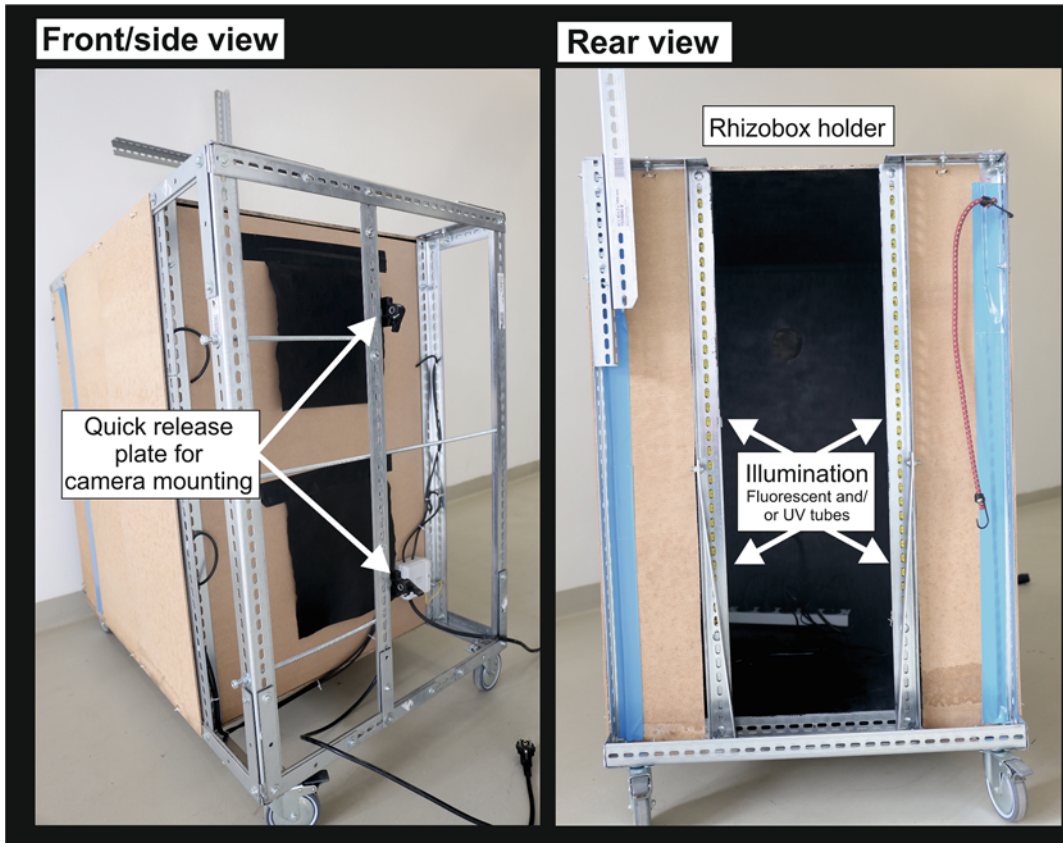


Fig. 3 RGB imaging setup (black box) of BOKU root phenotyping platform with front side (left) where camera is mounted and rear side (right) where rhizobox is fixed

imaging system (rhizobox size, camera distance) the entire rhizobox surface is imaged at once, or two images from top and bottom are taken and stitched together.

The rhizobox is illuminated via four 24 W fluorescent tubes. UV lamps (15 W UV tubes) can be used in case of low root-soil contrast (e.g., bright-colored soil), taking advantage of root autofluorescence to enhance the contrast between root and soil background. However, homogeneous illumination with UV tubes is challenging due to strong attenuation with distance from the light source. We point out that inhomogeneous illumination can complicate image analysis (root segmentation) and, in this case, requires additional preprocessing and should be minimized as much as possible during imaging.

2. Image acquisition: RGB images are acquired with a Canon EOS 6D digital camera fixed by quick-release plates at the respective top and bottom positions of the black box. In our setup, two images are taken which cover the upper and lower half of the rhizobox. A ruler is attached on each side of the

Table 2
Example camera settings (Canon EOS 6D) for RGB rhizobox imaging with different illumination methods

Fluorescent tubes	UV tubes
Autofocus and stabilizer off	Autofocus and stabilizer off
Mirror lock off	Mirror lock off
Manual mode	Manual model
ISO speed 500	ISO speed 1000
Shutter speed to 13	Shutter speed to 13
Aperture 5.6	Aperture 5.6
Auto white balance	White balance mode: Fluorescence

rhizobox for image stitching and scaling of the image at analysis. Example camera settings of the Canon EOS 6D for rhizoboxes illuminated with fluorescent and UV tubes, respectively, are given in Table 2.

2.4 Hyperspectral Imaging Setup

1. Scanner system: The hyperspectral root imaging system (Fig. 4a) consists of (a) a thermo-electrically cooled 14-bit monochrome Xeva NIR camera (Xenics, BE) with a spectral range from 900 to 1700 nm, 320 by 256 pixels, and a frame rate of 100 Hz and (b) an ImSpector N17E imaging spectrograph (Specim, FI) with a spectral range of 900–2500 nm. A halogen line illumination source (four 50 W halogen spots) is arranged in a 45°/−45° geometry. The imaging sensor is mounted on a two-axis positioning system. The scan window has a size of 240 × 1000 mm; that is, 30 mm at each edge of the rhizobox is not covered by the image. At the top of the system a white standard (Spectralon tile) is mounted for white standard acquisition before each scan.
2. Hyperspectral image acquisition: The scanner system is controlled via a Matlab-based software for setting (a) camera integration time, (b) spatial resolution of the scan (optional: pixel size 0.1 mm—field of view (FOV) 30 mm; pixel size 1.0 mm, FOV 300 mm), and (c) imaging region on the rhizobox (entire box or section).

   Determination of the adequate camera integration time for the rhizobox scan (with soil + root as target objects) and the white standard is done via the Xenics Xeneth camera software. The camera is moved to a position where both roots and soil are within the FOV. The integration time is then adjusted to cover approximately 85% of the full dynamic range of the sensor. Exceeding the maximum range (integration time too

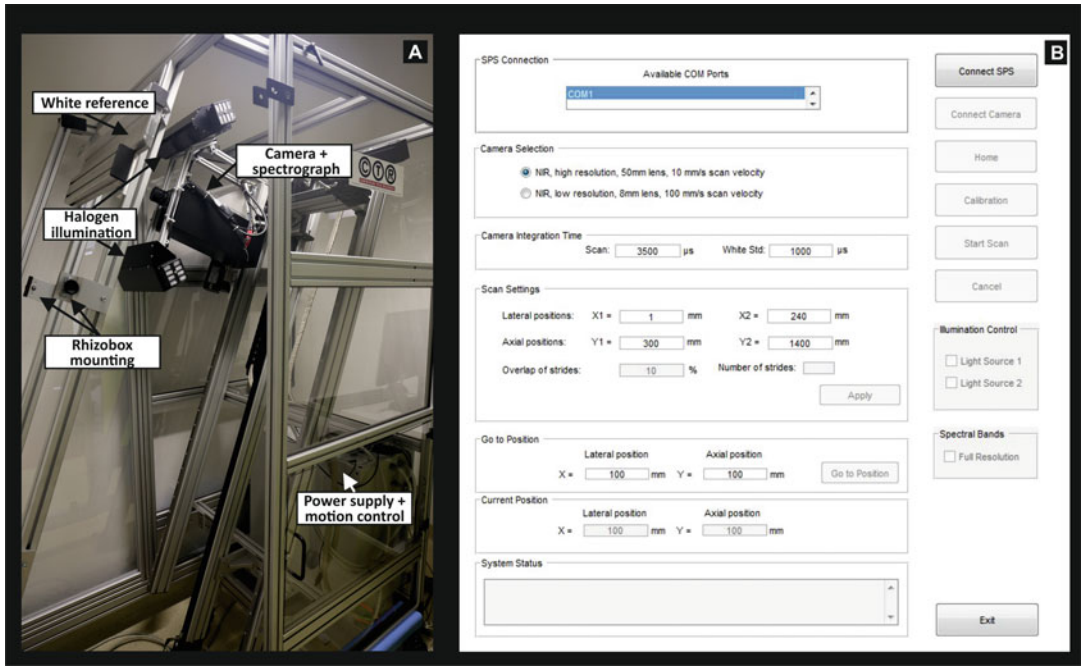


Fig. 4 Hyperspectral scanner of BOKU root phenotyping platform. (**a**) Hardware setup, (**b**) Matlab GUI for scanner settings

high) will result in data losses during image acquisition, while setting integration time too low does not use the full capacity of the camera. The same procedure is followed for the white standard. The respective settings are then transferred into the Matlab-GUI. Figure 4b provides a screenshot of typical settings for scanning bright roots on a dark soil background (*see* **Note 1** on data saving).

3 Methods

Image analysis methods for roots are strongly context specific and dictated by the targeted phenotyping traits (*see* **Note 2**). This is even more the case for chemometric analyses aiming to identify/quantify biochemical image properties from hyperspectral data. Some examples of rhizobox experiments and the respective plant husbandry are given in Table 3.

In case of RGB images, most experimenters aim to gather architectural root descriptors. There is an increasing number of software tools available (for an overview see https://www.quantitative-plant.org). The major challenge in case of rhizobox images is (1) the lack of continuous root axes due to local invisibility when roots grow into the soil and (2) high overlap of root systems at the later development stages that can be achieved in rhizobox experiments.

Table 3
Examples of rhizobox experiments and experimental settings performed at BOKU root imaging platform

Experimental question	Imaging[a]	Species/ plants per rhizobox	Duration	Watering	Fertilization
Genotypic differences in root architecture	RGB	Hexaploid wheat/ one	Flowering	80% PAW[b]	Regular with liquid NPK solution
Root architectural influence on drought resistance	RGB/ HSI	Tetraploid wheat/ one	Flowering	Dry down to 30% PAW at tillering	Regular with NPK until dry-down phase
Placed P-fertilizer effect on root architecture	RGB	Sugar beet/ one	10 leaf stage	80% PAW + 40% PAW	Initial P supply, regular liquid N supply
Mycorrhiza effect on root architecture	RGB	Faba bean/ one	Flowering	80% PAW + 40% PAW	None (sufficient initial soil supply with P and K)
Root decomposition	RGB/ HSI	Cover crop species/ two	16-week decomposition after clipping at full shoot development	80% PAW	Regular with liquid NPK solution

[a]RGB imaging and image analysis base on RGB image data in the VIS range; HSI imaging and image analysis based on hyperspectral image data in the SWIR range
[b]PAW Plant available water

In the following section, we describe example approaches used by the authors for evaluating RGB and hyperspectral datasets from different rhizobox experiments.

3.1 RGB Image Analysis Approaches

Figure 5 gives an overview of the general approach followed for RGB image analysis to obtain root architectural parameters (*see* **Note 3**).

1. Preparatory steps:
   Image stitching: RGB images from upper and lower parts of the rhizobox are merged manually into a single image using Adobe Photoshop. For this purpose, the layer opacity of one image region is reduced to 80% and the overlapping parts of the images are aligned based on the ruler attached at each side of the rhizoboxes. An alternative to manual stitching is automated image registration and stitching methods. For example, we successfully used the approach of [13] based on a discrete fast Fourier transform and subsequent sub-pixel cross-correlation implemented in Matlab.

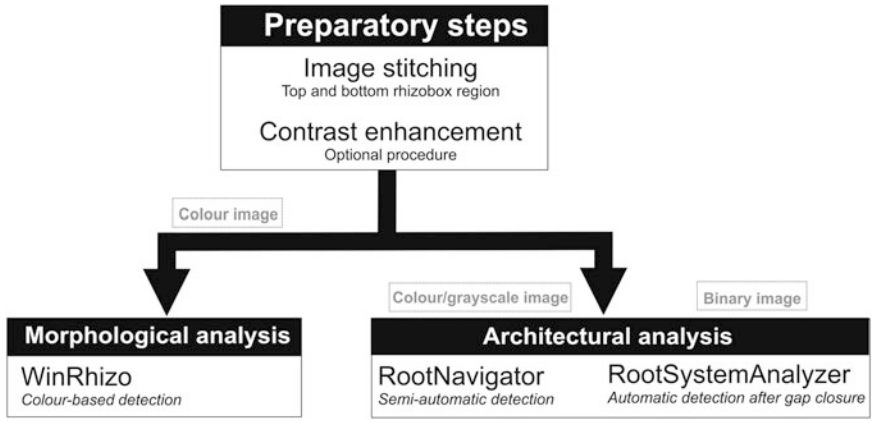


Fig. 5 Processing steps to analyze RGB images from rhizobox experiments for morphological and architectural root descriptors with example software used at the BOKU root imaging platform

Contrast enhancement: In case of images with low contrast between foreground (roots) and background (soil)—e.g., bright-colored soil or pigmented root axes—classification and segmentation accuracy of root vs. soil pixels can be improved by previous contrast enhancement. This is done either by visually shifting brightness/contrast settings in Adobe Photoshop or by using contrast enhancement functions in Matlab (e.g., *imadjust*, *histeq*, *adapthisteq*).

Image format: Depending on the subsequent software usage and segmentation approach, the merged and preprocessed image is saved as true color, grayscale, or binary file. RGB image analysis software often directly includes classification/segmentation algorithms for binarization before quantifying the respective output parameters: for example WinRhizo (Régent Instruments) uses color- or intensity-based thresholding approaches, while RootNav [14] applies an expectation maximization (EM) algorithm [15] for identifying roots. For other software we have used for architectural measurements like Root System Analyzer [16] binary images have to be provided. For this purpose, we use Matlab classification/segmentation approaches also applied for hyperspectral image analysis, i.e., thresholding [e.g., *adaptthresh* and *multithresh*] or clustering [e.g., *kmeans* and fuzzy].

2. Image analysis:

Table 4 provides a list of parameters which can be obtained from RGB root image analysis software applied to rhizobox data. Depending on the complexity of the root system (overlap) and continuous visibility of axes (gaps), analysis targets

Table 4
Parameters for describing root systems from rhizobox images

Parameter	Image section	Software	Automation
Morphological descriptors			
Visible root length	Entire root system	WinRhizo[a]	Full
Visible length distribution over depth	Entire root system	WinRhizo	Full
Fine root proportion	Entire root system	WinRhizo	Full
Architectural descriptors			
Inter-branch distance	Small root system or single axis	Root System Analyzer[b]	Full after manual gap closure
Apical length	Small root system or single axis	Root System Analyzer	Full after manual gap closure
Basal length	Small root system or single axis	Root System Analyzer	Full after manual gap closure
Maximum branching order	Small root system or single axis	Root System Analyzer	Full after manual gap closure
Lateral branching angle	Small root system or single axis	RootNav[c]	Semiautomated
Convex hull	Small root system or single axis	RootNav	Semiautomated
Primary-to-lateral length ratio	Small root system or single axis	RootNav	Semiautomated

[a]Régent Instruments
[b][16]
[c][14]

either the entire root image (eventually after previous manual tracking/gap filling) or only single axes. An example for application in a cover crop experiment is found in [17].

(a) Morphological analysis: The respective parameters are acquired using the commercial software WinRhizo Pro. First roots are separated from the background (root segmentation) and transformed into a binary image based on color classification. The length scale is set using the ruler on the side of the rhizobox image. A calibration file is built by manually marking pixels belonging to root and soil. This file, with the respective color classes for roots and soil (background), is then used for classification and segmentation. The segmentation result is visually judged and, if necessary, the *Debris and Rough Edges* filtering options are adjusted to improve removal of noise from misclassified pixels.

Depth distribution of visible root length is automatically obtained by setting an increment (e.g., 10 cm) to subdivide the analysis outputs of the entire image into single segments. Fine root proportion is calculated for the length of axes with a diameter smaller than a given threshold (e.g., 0.5 mm following Böhm classification [18]) in relation to total visible root length. However, we notice that the accuracy of diameter measurement and the resulting allocation of root length into single-diameter classes strongly depends on image quality and should be taken with care.

(b) Architectural analysis: To the best of our knowledge, currently there is no software for fully automated architectural analysis of rhizobox images, mainly due to the difficulty of dealing with axes frequently interrupted by gaps when roots locally disappear behind soil. Depending on the complexity of a root system (size, overlap) we based root architecture analysis either on the entire visible root system (in case of smaller root systems) or on single visible axes as a representative sample for describing the branching pattern of the entire system.

The software Root System Analyzer [16] is used to analyze the unbranched length from the tip to the start of the branching zone (apical length) and from the end of the branching zone to the origin of the root (basal length). In-between the length of the branching zone is measured and the number of emerging laterals counted, thereby providing inter-branch distance between laterals as a key architectural measure. This analysis is performed for the different root orders present in the root system. Root System Analyzer allows fully automated analysis. However, skeletonization of the root system currently requires continuous axes and accuracy of the automated tracking algorithm of an axis depends on the complexity of the system (overlaps). Thus previous manual gap closure in the image (e.g., using Adobe Photoshop or CorelDraw) is necessary to obtain meaningful results from Root System Analyzer.

RootNav [14] is applied to measure the angle of lateral root branching and to infer on the size of the rooted zone via the convex hull area. The user marks primary and lateral axis origins and tip positions and the axes are then tracked automatically (with user correction in case of improper direction at overlaps). Thereafter a number of morphological and architectural measures for the single axes and the entire root system are outputted (partially overlapping with the parameters obtained from

Root System Analyzer). We notice that recently a new version of RootNav has been published for automated root tracking based on deep learning [19], which is currently tested for different species and might provide in future a substantial increase in throughput of root system analysis.

3.2 Hyperspectral Image Analysis Approaches

An overview of the steps followed for processing and analysis of hyperspectral rhizobox images is provided in Fig. 6. All steps are performed using Matlab scripts (*see* **Note 4**).

A major challenge upon hyperspectral image processing and analysis is the size of images. Therefore, strategies for size reduction are an essential part in each processing and analysis step.

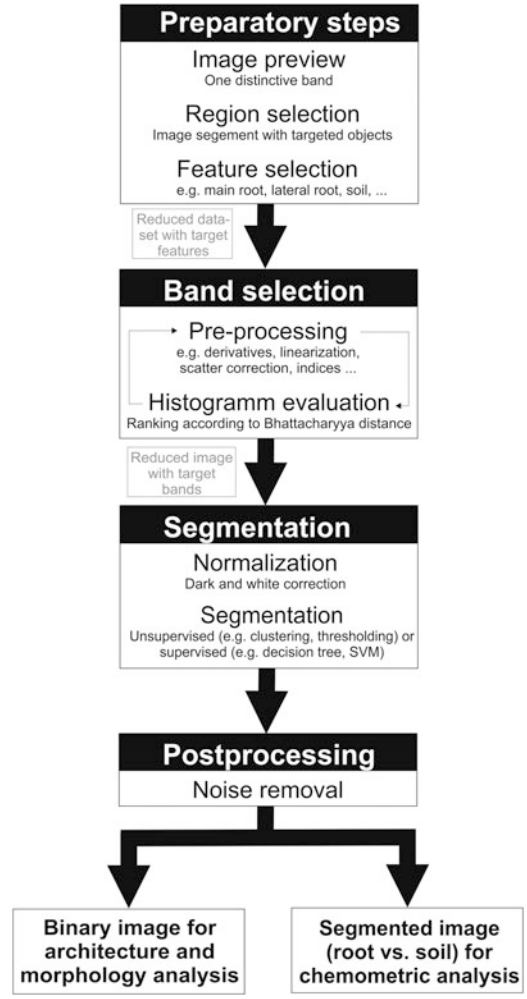


Fig. 6 Steps for processing and analysis of hyperspectral images of rhizobox-grown plant roots

1. Preparatory steps: During the preparatory steps a hyperspectral data subsample is selected from the rhizobox image that contains all objects of interest (i.e., in most cases root axes and soil background; in particular cases roots of different order, age, etc.).

   The rhizobox strides stored during scanning (36 files per rhizobox) are fused to provide a preview image of the entire rhizobox. This is done at one band (=wavelength) with good distinction of root and soil. This band for previewing is selected visually by scrolling through all bands in one stored file (typically the top part of an image stride where many roots can be expected). From the preview image the area on the rhizobox is defined which is used for the subsequent steps.

   This specified part on the rhizobox is now opened (i.e., extracted from the compressed SIF files) with all 256 bands. A freehand selection tool is then used to mark specific regions of interest (ROIs) and save them with a unique file name (e.g., ROI_root, ROI_soil, ROI_debris). Optionally the datasets are also labeled at this stage (i.e., 1 for root, 2 for soil) in case of later usage as training dataset for supervised classification approaches.

2. Band selection: Band selection aims to identify the most informative wavelength for classification and segmentation between the features of interest while removing all non-informative and noisy bands. The manually selected ROI datasets are fused into a common dataset to search for distinctive bands.

   The first step in band selection is preprocessing to enhance foreground (root) vs. background (soil) contrast and reduce scattering in the dataset. There are several approaches of image preprocessing via chemometric pretreatments (*see* ref. 20 for an overview). For most preprocessing methods (derivatives, scatter correction, baseline correction), we use the free Matlab mdatools toolbox [21]. Besides simple bands, also a combination of bands (indices) can strongly enhance image contrast (e.g., difference and ratios between two bands [22]). Combining bands strongly increases the dataset size (e.g., for difference spectra of 256 bands, dimensionality becomes $256 \times 256 = 65{,}536$).

   Band selection is performed via histogram evaluation using Bhattacharyya distance as quantitative measure [23]. In the dataset the band or band combination is searched that maximizes the separation between histogram peaks resulting from the respective pixels of spectrally distinctive features (e.g., root vs. soil).

   The loop of preprocessing and band selection via Bhattacharyya distance maximization is repeated for each transformation of the dataset to be tested in order to finally detect the best

combination of preprocessing method and informative bands that maximize spectral separation between the features contained in the dataset. As a result, the files composing an entire rhizobox are reduced to few informative bands and subsequent steps are performed with the reduced spectral image containing the relevant bands for efficient feature classification only.

3. Segmentation: At this step, the position of pixels belonging to roots vs. soil (and other features with spectrally distinctive characteristics) is identified, pixels are classified accordingly, and the image is segmented into the respective objects.

   Here we also perform normalization of the spectrally reduced image between the white (maximum reflectance) and dark (systemic noise) standards recorded for each scan. Normalization previous to band selection of the full 256 band 3D image would strongly increase computation time and is not recommended.

   There are numerous approaches to image classification and segmentation with performance strongly depending on the classification/segmentation problem (i.e., type of targeted objects, image quality, object and image size, availability, and size of labeled objects).

   For root classification and segmentation from a soil background, we have implemented both unsupervised and supervised approaches. Results are compared both visually and using some quantitative metrics (e.g., entropy, skewness). Table 5 gives a list of methods that have been successfully used in various rhizobox experiments with some comments on their performance. Overall, from the results obtained so far for different root systems and various soil backgrounds (i.e., classification problems) a subjective ranking of classification/segmentation methods according to accuracy and computational time would be thresholding $\ll$ support vector machine $<$ classification tree $<$ kmeans clustering $<$ fuzzy clustering $=$ Frangi filtering.

4. Post-processing: Post-processing aims to remove the remaining noise (i.e., misclassified pixels) from the image after segmentation. Noise removal is done with the images transformed into binary datasets and using Matlab morphological operators. Noise is identified as small (e.g., $<$10 pixels) and isolated objects. Additionally, the shape can be used for detecting non-root objects via the length-to-width ratio (i.e., defining a circle rather than a rootlike line shape). Depending on the thresholds set for size and shape, the strength of noise removal is adjusted.

Table 5
Classification and segmentation algorithms for hyperspectral rhizobox images

Method (*Matlab command*)	Type	Comment
Thresholding • Global: *graythresh* • Multilevel: *multithresh* • Adaptive: *imbinarize (I,'adaptive')*	Unsupervised	For 2D images (one spectral band). Good results only in case of high contrast between root and soil and low image scattering. Therefore strongly dependent on preprocessing. Low computational time
Clustering • kmeans: *kmeans* • Fuzzy: *SFCM2D*[a]	Unsupervised	Fuzzy clustering algorithm for 2D, kmeans for 2D and 3D (i.e., all informative bands) images. Very good results with fuzzy clustering, also good results with kmeans. Computational time fuzzy > kmeans
Frangi Vesselness filter[b] (*FrangiFilter2D, FrangiFilter3D*)	Unsupervised	For 2D and 3D images. Recognizes vessel-type structures in the image. Good results after empirically finding the optimum settings of Frangi parameters. Intermediate computational time
Error-correcting output codes (*fitcecoc*) • Classification tree: *templateTree* • Support vector machine: *templateSVM*	Supervised	2D and 3D images with two or more objects for classification. Result strongly dependent on the quality of labeled training dataset. Different learning methods can be selected (good results with classification tree and SVM with reasonable computational time)

[a]A [24]
[b][25]

We note that we also tested algorithms for gap closure to connect root axes locally disappearing behind the soil substrate. However, in most cases, results have not been sufficiently reliable for the more complex and overlapping rooting systems imaged from rhizobox systems. Thus, in case of using image analysis software which requires continuous axes (like Root System Analyzer), still manual gap closure is required.

5. Image analysis: Binary images can be analyzed for morphological and architectural parameters with the same approaches as used for RGB images. However, due to the lower resolution of the hyperspectral camera, the main applications of hyperspectral root analysis are chemometric information beyond the root morphological/architectural descriptors attainable also from RGB images. So far, we used hyperspectral images for two main objectives: inference on (a) soil water content around root axes and (b) root decay/decomposition. Both applications require previous segmentation between roots and soil. Then chemometric analysis is applied to the soil pixels in the first case, while in the second case the target object is the root pixels.

   The first step in chemometric analysis is to recover the full spectral information of the target objects. For this purpose, the indices of the segmented root pixels are stored. In case of

Fig. 7 Rhizobox with defined water contents (hydraulically separated by Styrodur sheets; at dry range germination paper strips additionally help to keep fine soil particles in the respective compartment) for spectral calibration to predict soil water content

targeting soil, the root-related pixels are darkened (i.e., set to a value of zero), while in case of further analyzing roots the same is done for the non-root pixels.

Example of analyzing soil water content: For this purpose, a regression-based model for water content has been used. This is obtained from a rhizobox filled with the same soil as for the experiment, which is set to known water contents (Fig. 7).

The rhizobox is scanned and a regression-based calibration between spectral information and measured water content is developed. Different approaches can be used: (1) simple (or multiple) regression (Matlab: *regress*) with specific spectral bands directly related to the physical SWIR absorption/reflection properties of water (e.g., 1240, 1450 nm; [26]); (2) partial least squares regression (PLSR; Matlab: *plsregress*) or principal component regression (PCR; Matlab: *pca* and *regress*) models encoding several spectral regions into latent/composite variables related to the measured water content; and (3) support vector machine regression (SVMR; Matlab: *fitrsvm*), where the different water content regions are labeled as training dataset and then a trained multilevel classifier is applied to predict water content from the spectral pattern of the testing data.

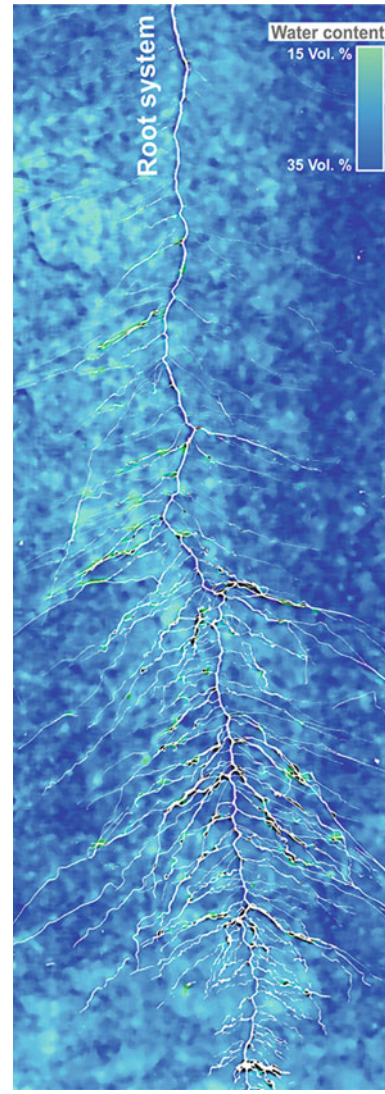


Fig. 8 False color image mapping of the water content around a rhizobox-grown sugar beet root. Water content calibration from rhizobox shown in Fig. 7

The calibration model is then applied to predict the water content for each soil pixel of the rhizobox (Fig. 8). Note that in case of preprocessing (e.g., scatter correction) evidently also the calibration model is developed from data with the same preprocessing as the prediction image. With the position of root axes defined by their indices, the water content depletion around roots can be calculated by pixelwise increase of the root indices and subsequent extraction of the respective local water content information.

Example analysis of root decay: For this purpose, chemometric analysis is used to identify distinctive spectral features encoding the decay of roots after different times from clipping their shoots. Here

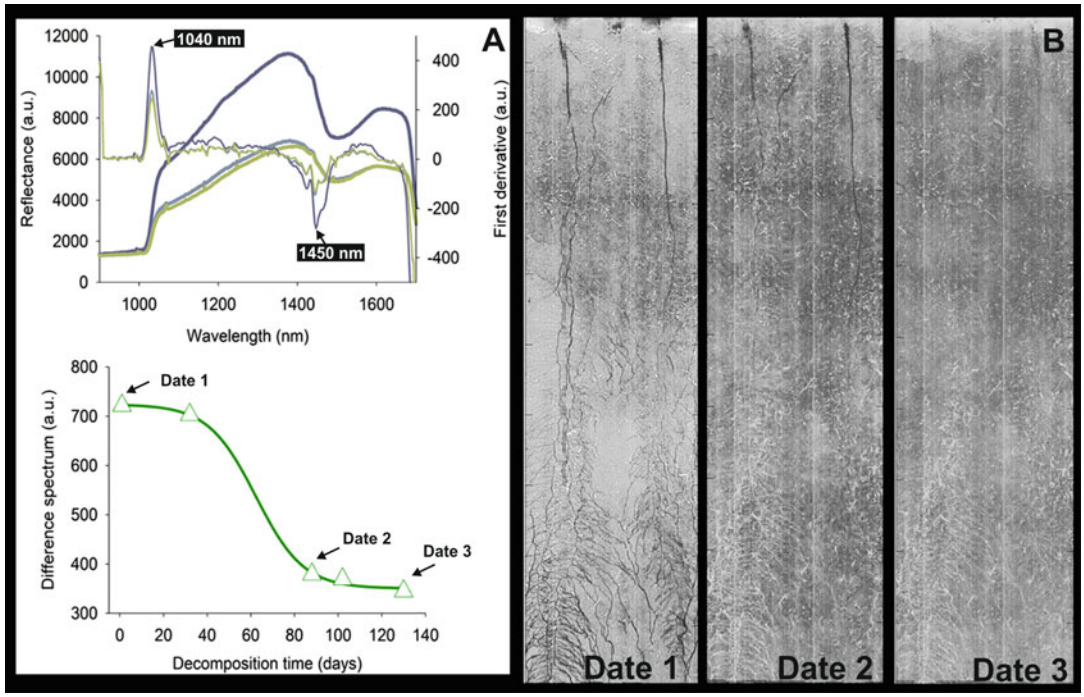


Fig. 9 Analysis of root decomposition of an oilseed radish cover crop from hyperspectral images. (**a**) Top: Raw spectra and first derivatives for three time points (clipping of shoot at first date 26th November; Date 2: 21st February; Date 3: 4th April). Bottom: Decay function based on difference spectra of the first derivative at 1040 and 1450 nm. (**b**) Example images for the three times with spectra shown in **a** (top)

temporal changes in the spectral pattern are directly targeted for inference on root decomposition without any directly measured chemical properties (e.g., lignin, water content) for prediction.

The root indices of rhizobox scans at five times after clipping are retrieved from segmentation and a subsample of root spectral data of similar size (10,000 pixels) is randomly selected (Matlab: *datasample*). The five subsamples are then combined to one dataset and evaluated for histogram distance, as described above, using Bhattacharyya distance as quantitative measure (Fig. 9).

Again, preprocessing (e.g., derivatives), spectral indices (e.g., difference spectra), as well as composite variables from PCA are used to find the spectral bands maximizing the distance between histogram peaks from different time points.

For the selected spectral measures (i.e., single bands, indices, principal components), a function of time after clipping is fit through the data points describing the dynamics of root decay (in Fig. 9a a sigmoidal function). This function of the stage of root decomposition might also be mapped on the root pixels at a given time point to infer on the different potential functionality of young (apical) vs. old (basal) parts of the root system.

4 Notes

1. Data size and saving: When starting a scan, the hyperspectral imaging sensor first acquires the dark and white standards. This is recommended before each imaging campaign, e.g., once a day. The dark standard represents the camera noise (dark current), while the white standards give the maximum reflectivity. These data are later used for normalization at image processing. Due to the huge size of an entire rhizobox image and to avoid problems during data saving, (a) the image is subdivided into nine separate strides (30 × 1000 mm and 256 spectra) with 10% overlap, (b) each stride is further subdivided into four parts (three of 300 mm length, one of 100 mm length), and (c) all files are saved as compressed SIF files.

   Thus, in total one rhizobox is saved as 36 SIF files. Each file has a unique stamp consisting of stride number (1–9) and part (1–4) as well as date and time (YYYY.MM.DD HH:MM:SS). Disk space required for an entire rhizobox scan is 13.7 GB. Scanning duration for high-resolution settings (FOV 30 mm) is approximately 16 min. It is also recommended (a) to save files composing one rhizobox in a separate folder with folder name defining the object (e.g., cultivar name_treatment_replicate) and (b) to include a text file with an accurate metadata description of the image within the same folder where the scanned strides are located.

2. Application scope: There is no unique solution to root phenotyping. The protocol presented here for soil-filled rhizoboxes constitutes a root imaging method under field-near conditions that allows in situ analysis of root architecture (RGB) and functioning (hyperspectral). From the point of view of throughput it is situated between classical (low throughput) research methods for root-soil interactions and phenotyping applications with high-throughput requirements. It is not suitable as a screening method for high numbers of cultivars. Its application in a breeding context is mostly in pre-breeding tests of novel germplasm as well as in the evaluation of germplasm preselected with high-throughput methods (e.g., germination paper; [27]) under field-near conditions.

3. Architectural analysis shortcomings: Full automation of architectural analysis with sufficient accuracy and reliability of outputs for complex root systems of rhizobox-grown plants is still not resolved. Software selection depends on hypotheses, visual observations of interesting patterns, as well as feasibility considerations in an experiment and the related trait requirements. While the analysis of visual root morphology can be computed with different software solutions (e.g., WinRhizo), for

architectural analysis the experimenter has to decide for (semi-automated) manual tracking and/or analysis based on single fully visible axes as representative subsample for the entire root system.

4. Hyperspectral image analysis: Due to the only very recent publication and advance in the application of hyperspectral imaging in root phenotyping [28], there are still no easy-to-use software solutions. Experimenters have to rely on tailor-made solutions, which generally require programming skills. We welcome any request for the Matlab scripts used for the hyperspectral image analysis presented in this protocol. However, we underline that the scripts are written problem specific and, although extensively commented, do not constitute a comprehensive software applicable without Matlab knowledge.

References

1. Hammer GL, Dong Z, McLean G, Doherty A, Messina C, Schussler J, Zinselmeier C, Paszkiewicz S, Cooper M (2009) Can changes in canopy and/or root system architecture explain historical maize yield trends in the US corn belt? Crop Sci 49:299–312
2. Siddique KHM, Chen YL, Rengel Z (2015) Efficient root system for abiotic stress tolerance in crops. Procedia Environ Sci 29:295
3. Nakhforoosh A, Grausgruber H, Kaul H-P, Bodner G (2014) Wheat root diversity and root functional characterization. Plant Soil 380:211–229
4. Palta JA, Chen X, Milroy SP, Rebetzke GJ, Dreccer MF, Watt M (2011) Large root systems: are they useful in adapting wheat to dry environments? Funct Plant Biol 38:347–354
5. Tardieu F (2011) Any trait or trait-related allele can confer drought tolerance: just design the right drought scenario. J Exp Bot 63:25–31
6. Bodner G, Nakhforoosh A, Kaul H-P (2015) Management of crop water under drought: a review. Agron Sustain Dev 35:401–442
7. Atkinson JA, Pound MP, Bennett MJ, Wells DM (2019) Uncovering the hidden half of plants using new advances in root phenotyping. Curr Opin Biotechnol 55:1–8
8. Kuijken RCP, van Eeuwijk FA, Marcelis LFM, Bouwmeester HJ (2015) Root phenotyping: from component trait in the lab to breeding. J Exp Bot 66:5389–5401
9. Pandey P, Ge Y, Stoerger V, Schnable JC (2017) High throughput in vivo analysis of plant leaf chemical properties using hyperspectral imaging. Front Plant Sci 8:1348
10. Hobley E, Steffens M, Bauke SL, Kögel-Knabner I (2018) Hotspots of soil organic carbon storage revealed by laboratory hyperspectral imaging. Sci Rep 8:1–13
11. Nagel KA, Putz A, Gilmer F, Heinz K, Fischbach A, Pfeifer J, Faget M, Blossfeld S, Ernst M, Dimaki C (2012) GROWSCREEN-Rhizo is a novel phenotyping robot enabling simultaneous measurements of root and shoot growth for plants grown in soil-filled rhizotrons. Funct Plant Biol 39:891–904
12. Passioura JB (2006) The perils of pot experiments. Funct Plant Biol 33:1075–1079
13. Guizar M (2020) Efficient subpixel image registration by cross-correlation. MATLAB Central File Exchange. https://www.mathworks.com/matlabcentral/fileexchange/18401-efficient-subpixel-image-registration-by-cross-correlation. Accessed 21 Mar 2020
14. Pound MP, French AP, Atkinson JA, Wells DM, Bennett MJ, Pridmore T (2013) RootNav: navigating images of complex root architectures. Plant Physiol 162:1802–1814
15. Dempster AP, Laird NM, Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm. J R Stat Soc B Methodol 39:1–22
16. Leitner D, Felderer B, Vontobel P, Schnepf A (2014) Recovering root system traits using image analysis exemplified by two-dimensional neutron radiography images of lupine. Plant Physiol 164:24–35
17. Bodner G, Loiskandl W, Hartl W, Erhart E, Sobotik M (2019) Characterization of cover crop rooting types from integration of

rhizobox imaging and root atlas information. Plants 8:514

18. Böhm W (1979) Methods of studying root systems. Springer, New York, pp 125–138
19. Pound MP, Atkinson JA, Townsend AJ, Wilson MH, Griffiths M, Jackson AS, Bulat A, Tzimiropoulos G, Wells DM, Murchie EH (2017) Deep machine learning provides state-of-the-art performance in image-based plant phenotyping. GigaScience 6:gix083
20. Esquerre C, Gowen AA, Burger J, Downey G, O'Donnell C (2012) Suppressing sample morphology effects in near infrared spectral imaging using chemometric data pre-treatments. Chemometrics Intell Lab Syst 117:129–137
21. Kucheryavskiy S (2019). https://github.com/svkucheryavski/mdatoolsm. Accessed 21 Mar 2020
22. Kim DM, Zhang H, Zhou H, Du T, Wu Q, Mockler TC, Berezin MY (2015) Highly sensitive image-derived indices of water-stressed plants using hyperspectral imaging in SWIR and histogram analysis. Sci Rep 5:15919
23. Bhattacharyya A (1946) On a measure of divergence between two multinomial populations. Sankhyā 7:401–406
24. ABing (2020) Spatial fuzzy clustering and level set segmentation. MATLAB Central File Exchange. https://www.mathworks.com/matlabcentral/fileexchange/31068-spatial-fuzzy-clustering-and-level-set-segmentation. Accessed 21 Mar 2020
25. Dirk-Jan Kroon (2020) Hessian based Frangi Vesselness filter. MATLAB Central File Exchange. https://www.mathworks.com/matlabcentral/fileexchange/24409-hessian-based-frangi-vesselness-filter. Accessed 19 Mar 2020
26. Bruning B, Liu H, Brien C, Berger B, Lewis M, Garnett T (2019) The development of hyperspectral distribution maps to predict the content and distribution of nitrogen and water in wheat (*Triticum aestivum*). Front Plant Sci 10:1380
27. Gioia T, Galinski A, Lenz H, Müller C, Lentz J, Heinz K, Briese C, Putz A, Fiorani F, Watt M (2017) GrowScreen-PaGe, a non-invasive, high-throughput phenotyping system based on germination paper to quantify crop phenotypic diversity and plasticity of root traits under varying nutrient supply. Funct Plant Biol 44:76–93
28. Bodner G, Nakhforoosh A, Arnold T, Leitner D (2018) Hyperspectral imaging: a novel approach for plant root phenotyping. Plant Methods 14:84

Chapter 18

Light Drones for Basic In-Field Phenotyping and Precision Farming Applications: RGB Tools Based on Image Analysis

Federico Pallottino, Simone Figorilli, Cristina Cecchini, and Corrado Costa

Abstract

Plant phenotyping has garnered major attention in recent years, leading to developing new strategies to measure and assess plant traits of interest. For data acquisition of large fields, devices and sensors are required that deliver detailed and reproducible temporal and spatial information on the cultivated crop. This work proposes the potential use of low-cost light drones for in-field phenotyping applications on cereal crops. The proposed method allows to obtain precise measurements of color and height of the plants for the individual plots. The method is based on a color calibration algorithm (TPS-3D interpolating function) and a 3D ortho image reconstruction. The method has been applied on an experimental field with durum and soft wheat parcels obtaining information on real color (with an error lower than 12/256) and height for each single plot.

Key words Plant phenotyping, Light drones, Ortho-image, Color calibration, Height estimation, RGB

1 Introduction

Nowadays, plant phenotyping is achieving a great success and importance in agriculture. Plant traits are crucial not only to evaluate genotype–environment interactions and to increase crop performance in terms of yields and resistance to pathogens, but also to improve crop management strategies (e.g., fertilization and irrigation) [1].

Plant phenotyping strategies take advantage of noninvasive and digital technologies to allow the measurement and assessment of complex plant traits (e.g., growth, development, tolerance, resistance, architecture, physiology, ecology, yield) which is affected by genetic variation as well as environmental interaction [2–4].

The application of plant phenotyping in the field has strong links with precision agriculture sharing its purposes and part of the technologies. There is growing interest in adapting agricultural machinery and electronic sensors for field-based high-throughput

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9_18,

phenotyping [5, 6] and various vehicle-based high-throughput systems have been used or proposed for phenotyping plants in field [7].

For data acquisition of large fields, devices and sensors are required that deliver detailed and reproducible temporal and spatial information on the cultivated crop. Although most of these systems have elements that would improve the acquisition of phenotypic data, none seems capable of providing the needed throughput for the multiple data types that are essential for efficient field-based plant phenotyping. In-field phenotyping platform for the precise and accurate recording of agronomic traits and crop monitoring results to be few and expensive; thus, it represents a bottleneck for further advancements in knowledge and development of crop and varieties [8].

Fortunately, the latest technologies provide an unprecedented array of hardware and software characteristics that appear capable of providing the required high throughput. Examples for instruments largely relate to the increasing power of electronic components through greater integration of functions and reduction in size, with concomitant reductions in cost and power consumption. Along with these characteristics, sensor resolution and data stream are increasingly creating new potential applications. When applied on field vehicle-mounted sensors allow real-time control mechanisms, such as GPS-enabled automatic tractor or implement steering, or the control of agricultural machineries. Remote sensing approaches based on satellite and aerial vehicles rely, on the opposite, on the capacity to acquire huge surface but this comes at the expense of spatial and temporal much lower resolution.

Our approach to sensor deployment falls in the middle, in the category of remote sensing but far from the approaches of remote aerial or satellite platforms [9]. Indeed, drones are used close to the ground to enhance ground details. In this context, a plant phenotyping system allows for frequent deployment of replicated sets, eventually the use of different sensors in close proximity to the plants, enabling simultaneous phenotyping observations of several plant traits and multiple adjacent plots. This not only enables to record multiple types of data in a single pass increasing the throughput, but would also allow a more accurate and comprehensive phenotype description [6, 10]. Fanigliulo et al. [11] showed the potential use of a low-cost light drone (DJI Spark) application to assess soil roughness and cloddiness. The drone use was paired to RGB 3D image analysis techniques to evaluate different tillage methods (ploughed, harrowed, and grassed) in comparison to traditional methods such as laser profile meter and manual sieving. Light drone application is able to replicate the results scored by the traditional methods but with consistent advantages in terms of time, repeatability, and surface analyzed while reducing the

human error during the data collection. Additionally, the use of light drones benefits from reduced regulation constraints, and ease of transport and use.

The proposed methodology shows the potential use of low-cost light drones for in-field phenotyping applications on cereal crops.

2 Materials

For the field phenotyping application proposed the following materials and software were used:

- As a carrier and for image acquisition the low-cost light drone DJI™ SPARK™ was chosen (Table 1).
- For flight planning to flight over the experimental field using a waypoint, the missions were prepared using the open-source software Mission Planner (License GPLv3).
- The orthoimage reconstructions were conducted using the software "3DF Zephyr" (Zephyr 3DFLow 2018, Verona, Italy).
- For color calibration the color checker GretagMacbeth (24 patches) was used with known color patch values.

The flights were carried out on the CREA-IT experimental farm in Montelibretti (42°07′47.71″N, 12°38′31.01″E, Rome, Italy). The field flown over includes the tests of the national durum wheat and soft wheat network. The plants were arranged in a randomized block design of 10 m^2 plots with a sowing density of 350 seeds/m^2 for durum wheat and 450 seeds/m^2 for soft wheat following a multifactor repeated measurement design with three replications. The block design for both durum and soft wheat was framed by edges of barley plots. A total of 30 × 3 (90) durum wheat and 36 × 3 (108) soft wheat parcels has been analyzed. Fertilization and plant protection were performed to ensure optimal plant growth. Figure 1 shows the orthoimage reconstruction of the wheat plots analyzed.

3 Methods

3.1 Light Drone Flight and Image Acquisition

Images were taken on April 9th, 2019, using the drone lightened to a weight <300 g according to the Italian Civil Aviation Authority (ENAC) regulation ("Aeromobili a pilotaggio remoto," Edition 2, Amendment 4, Art. 12, comma 5. May 5th, 2018) on the use of drones without specific pilot license. Mission planner was used to calculate and plan the surface to acquire respecting a series of parameters such as correct image overlap/sidelap, ground resolution following the (ground sample distance) GSD, and image

Table 1
Specifications of the unmanned aerial vehicle (UAV) DJI™ SPARK™ and of the camera equipped for image acquisition

Details	Items	Specifications
Light drone	Weight	297 g
	Dimensions	143 × 143 × 55 mm
	Max speed	50 km/h
	Satellite positioning systems	GPS/GLONASS
Digital camera	Camera focal length	4.5 mm
	Sensor dimensions ($W \times H$)	6.17 × 4.56 mm
	Sensor resolution	12 megapixels
	Image sensor type	CMOS
	Capture formats	MP4 (MPEG-4 AVC/H.264)
	Still image formats	JPEG
	Video recorder resolutions	1920 × 1080 (1080p)
	Frame rate	30 frames per second
	Still image resolutions	3968 × 2976
GIMBAL	Control range inclination	from −85° to 0°
	Stabilization	Mechanical 2 axes (inclination, roll)
	Obstacle detection distance	0.2–5 m
	Operating environment	Surfaces with diffuse reflectivity (>20%) and dimensions greater than 20 × 20 cm (walls, trees, people, etc.)
Remote control	Operating frequency	5.8 GHz
	Max operating distance	1.6 km
Battery	Supported battery configurations	3S
	Rechargeable battery	Rechargeable
	Technology	Lithium polymer
	Voltage provided	11.4 V
	Capacity	1480 mAh
	Run time (up to)	16 min
	Recharge time	52 min

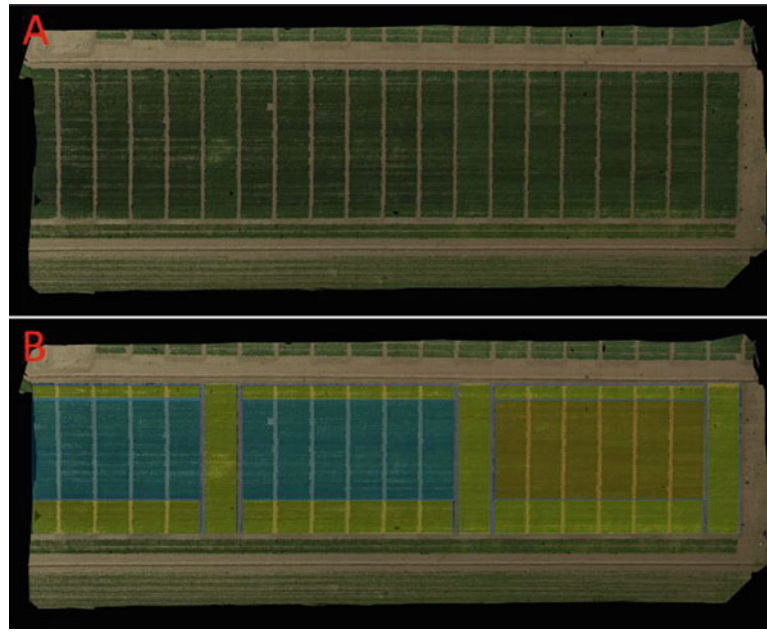


Fig. 1 (**a**) Orthoimage of the experimental field; (**b**) wheat crops evidenced using different colors: yellow for barley; orange for durum wheat; blue for soft wheat

acquisition sync scheduling depending on the flight speed and flight time. The Android app used to control the flight and pilot the UAV was Litchi that can load waypoints from delimited csv file (Fig. 2) for predefined mission flight [11].

The digital noninterchangeable camera, included in the UAV, was used to collect still images every 2 s using a shutter speed of 1/2000 s and 100 ISO of sensitivity. The camera technical specifications are described in Table 1. Images were collected using the UAV with the digital camera at 0.5 m/s at 3 m aboveground level (AGL). The details of the experimental flight are shown in Table 2. The images were acquired based on a time-lapse function of the RGB camera vertically oriented that took one image every 2 s ensuring around 75% overlapping ratio. It used a sidelap of 70%.

3.2 3D Ortho Image Reconstruction

After the acquisition at each date, the collected pictures were analyzed to reconstruct the orthoimages with 3DF Zephyr [12] through the following steps: project creation selecting the pictures needed; camera orientation and sparse point cloud generation present at high accuracy with the images at 100% resolution (no resize); dense point cloud generation; mesh extraction; textured mesh generation; and export outcome files including the digital surface model (DSM), the digital terrain model (DTM), and the orthoimage (Fig. 3a). Subtracting the DSM to the DTM it was possible to obtain the metric z values (i.e., the heights) [11].

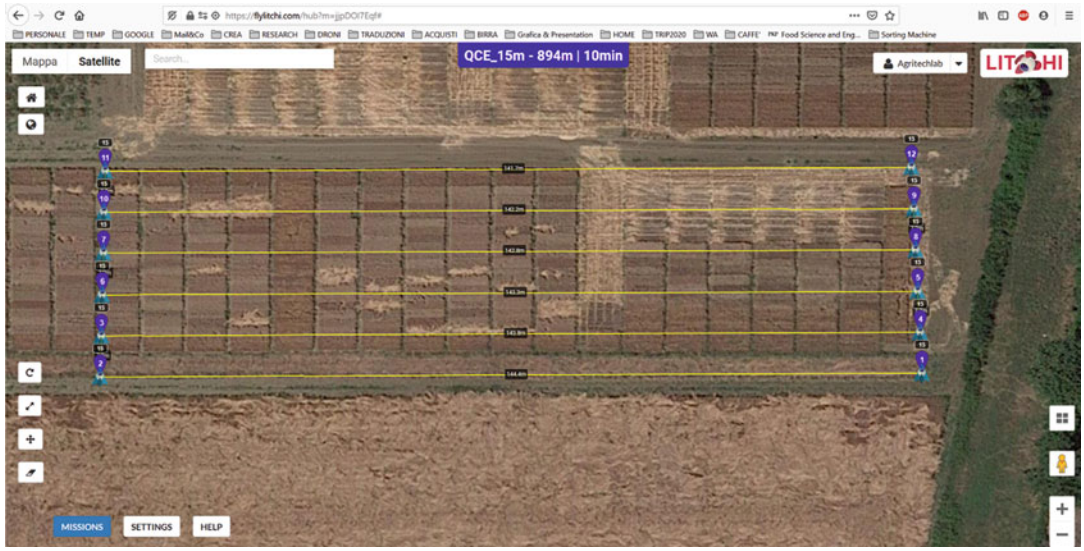


Fig. 2 Litchi Web app showing the mission elaborated using the software open-source Mission Planner

Table 2
Experimental unmanned aerial vehicle (UAV) flight details

Flight date	Image number	Flight altitude (m)	Flight speed (m/s)	Ground resolution (cm/px)	Illumination
Mar. 27, 2019	246	15	2	0.52	Natural light
Apr. 9, 2019	323	15	2	0.52	Natural light
Apr. 15, 2019	251	15	2	0.52	Natural light
May 5, 2019	266	15	2	0.52	Natural light
Jun. 6, 2019	254	15	2	0.52	Natural light

3.3 Color Calibration

Warping an image is a transformation which involves pixels mapping from source positions to other destination positions [13]. A commonly used technique to fit the data is the TPS method, which is useful due to its insensitivity to data noise and its capability to minimize the bending energy of a thin-shell object [14]. The name thin plate spline refers to a physical analogy involving the bending of a thin sheet of metal. In the physical setting, the deflection is in the z direction, orthogonal to the plane. In order to apply this idea to the problem of coordinate transformation, one interprets the lifting of the plate as a displacement of the x or

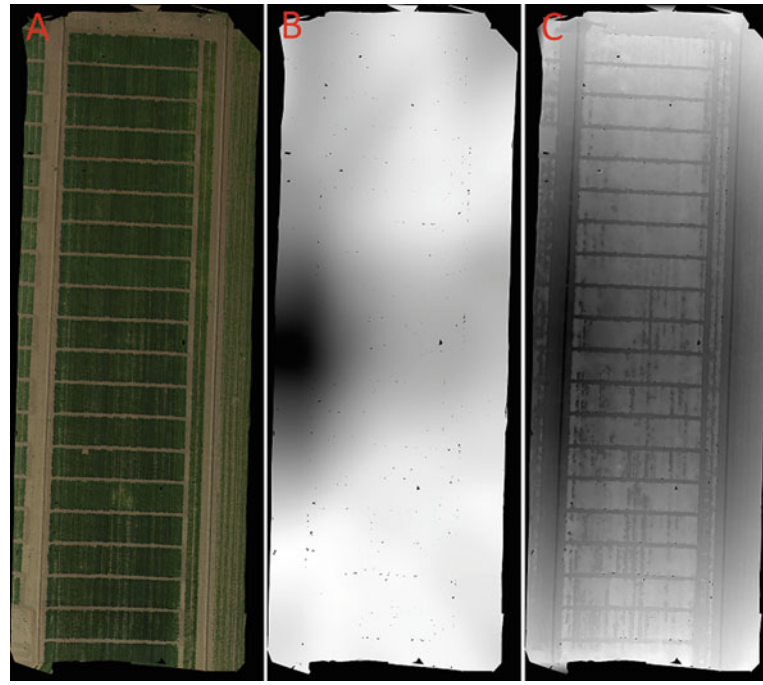


Fig. 3 The reconstructed orthoimage (**a**), DTM (**b**), and DSM (**c**)

y coordinates within the plane. In 2D cases, given a set of K corresponding points, the TPS warp is described by $2(K + 3)$ parameters, which include six global affine motion parameters and $2K$ coefficients for correspondences of the control points [15]. These parameters are computed by solving a linear system; in other words, TPS has a closed-form solution. Only a slight modification is necessary to produce interpolation functions for three-dimensional thin-plate splines [16, 17]. Given two configurations of homologous landmarks, the thin-plate spline is a map from plane to plane that maps each landmark to its correspondent. It can be defined briefly, although not quite rigorously, as the interpolation that has the least bending energy, where bending energy is defined to be the integral of the sum of squared second derivatives. Bending energy is zero precisely when the map is affine [18]. The thin-plate spline interpolation algorithm Matlab code to calibrate colors in sRGB space was reported by Menesatti et al. [17].

In the present work the measured ColorChecker sRGB coordinates within each image (i.e., considering its whole field) were warped (transformed) into the reference coordinates of the same ColorChecker. This transformation was performed through the TPS interpolation function, modified for the three-dimensional space. The three-dimensional sRGB color space is an additive color model in which red, green, and blue lights are added together

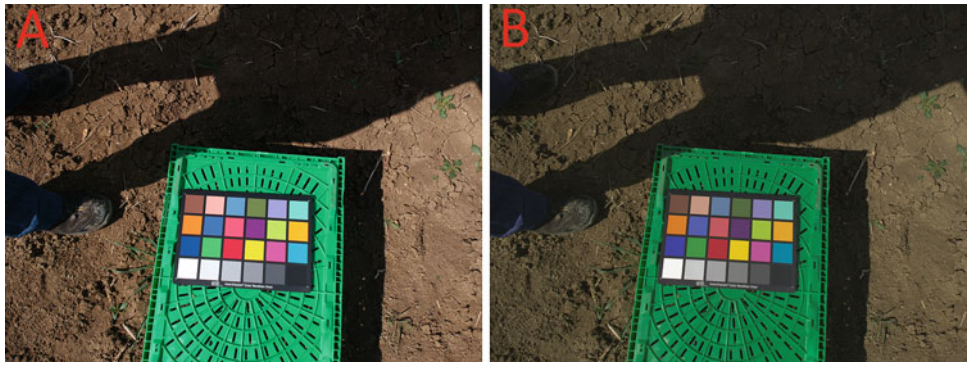


Fig. 4 (**a**) Original acquired image from light drone DJI™ SPARK™ with the color checker GretagMacbeth (24 patches) and (**b**) the resulting calibrated one

in various ways to reproduce a broad array of colors [19]. The procedure was reported by Menesatti et al. [17]. This code, given the set of measured ColorChecker RGB coordinates within the image and the reference coordinates of the same ColorChecker, transforms the RGB value of each pixel of the image following the TPS-3D interpolating function. The code could also be applied to warp 3D images (such as *x*, *y*, *z* references or hyperspectral images) by substituting the colorimetric coordinates with 3D space coordinates.

At the beginning of the flight on the terrain an image was taken including the color checker GretagMacbeth with known color patch values. Figure 4 shows the original acquired image (a) and the resulting calibrated one (b).

4 Notes

The proposed method represents an effective tool allowing a precise color measurement per single parcel (i.e., cultivar) using the light low-cost DJI Spark drone reported in Subheading 2. The applied methodology, pairing fast image acquisition with advanced color calibration methodology, as reported in Subheading 3, scored a precision lower than 12/256 in the RGB color space [14]. The method was used to measure the color and height of each replication of the 30 durum wheat and 36 soft wheat varieties (198 plots) grown within the cereal national network. As an example, in this note we report a visual representation of the color variation for two durum wheat cultivars (Marco Aurelio and Monastir) at the five sampling times as reported in Fig. 5.

Marco Aurelio and Monastir represent medium-early cycle varieties with medium-height plants. While the first variety has a good protein content and greater resistance to Septoria, the second is characterized by good production and less susceptibility to brown rust.

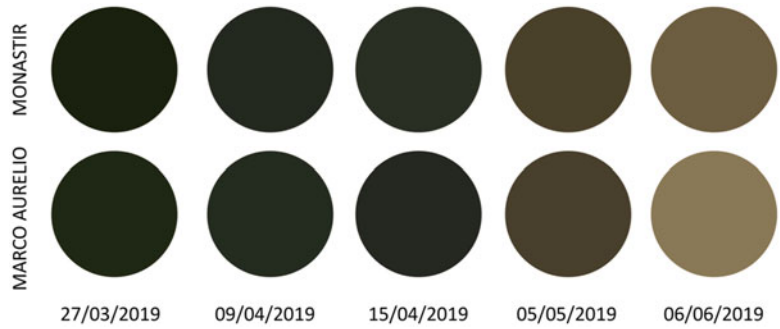


Fig. 5 Mean color for two durum wheat cultivars (Marco Aurelio and Monastir) at the five sampling times

The 3D reconstructions were used to characterize the varieties on the base of their growth (elevation) relative to the specific phenological stages acquired. Considering the negligible error on the *z*-axis produced by the workflow (acquisition and elaboration) the presented methodology could be helpful in evaluating the characteristics affecting the plant emergence and subsequent growth. These are interesting traits normally evaluated by field phenotyping approaches that using conventional visual assessments is very work intensive. The proposed method heavily reduces the need for specialized work hours offering an increased approach standardization.

Acknowledgment

This chapter was funded with the contribution of the Italian Ministry of Agricultural, Food, Forestry and Tourism Policies (MiPAAFT) sub-project "Tecnologie digitali integrate per il rafforzamento sostenibile di produzioni e trasformazioni agroalimentari (AgroFiliere)" (AgriDigit program) (DM 36503.7305.2018 of 20/12/2018).

References

1. Cobb JN, DeClerck G, Greenberg A, Clark R, McCouch S (2013) Next-generation phenotyping: requirements and strategies for enhancing our understanding of genotype–phenotype relationships and its relevance to crop improvement. Theor Appl Genet 126:867–887
2. Fiorani F, Schurr U (2013) Future scenarios for plant phenotyping. Annu Rev Plant Biol 64:267–291
3. Li L, Zhang Q, Huang D (2014) A review of imaging techniques for plant phenotyping. Sensors 14(11):20078–20111
4. Costa C, Schurr U, Loreto F, Menesatti P, Carpentier S (2019) Plant phenotyping research trends, a science mapping approach. Front Plant Sci 9:1933
5. Montes JM, Melchinger AE, Reif JC (2007) Novel throughput phenotyping platforms in plant genetic studies. Trends Plant Sci 12:433–436
6. White JW, Andrade Sanchez P, Gore MA, Bronson KF, Coffelt TA, Conley MM et al (2012) Field-based phenomics for plant genetics research. Field Crop Res 133:101–112

7. Pallottino F, Antonucci F, Costa C, Bisaglia C, Figorilli S, Menesatti P (2019) Optoelectronic proximal sensing vehicle-mounted technologies in precision agriculture: a review. Comput Electron Agric 162:859–873
8. Minervini M, Scharr H, Tsaftaris SA (2015) Image analysis: the new bottleneck in plant phenotyping [applications corner]. IEEE Signal Process Mag 32:126–131
9. Fussell J, Rundquist D, Harrington J Jr (1986) On defining remote sensing. Photogramm Eng Remote Sens 52(9):1507–1511
10. Scotford IM, Miller PCH (2004) Combination of spectral reflectance and ultrasonic sensing to monitor the growth of winter wheat. Biosyst Eng 87(1):27–38
11. Fanigliulo R, Antonucci F, Figorilli S, Pochi D, Pallottino F, Fornaciari L, Grilli R, Costa C (2020) Light drone-based application to assess soil tillage quality parameters. Sensors 20:728
12. Anwar N, Izhar MA, Najam FA (2018) Construction monitoring and reporting using drones and unmanned aerial vehicles (UAVs). In: Proceedings of the 10th international conference on construction in the 21st century (CITC-10), Colombo, Sri Lanka, 2–4 July
13. Glasbey CA, Mardia KV (1998) A review of image warping methods. J Appl Stat 25:155–171
14. Duchon J (1977) Splines minimizing rotation-invariant semi-norms in Sobolev spaces. Constr Theory Funct Sev Var Lect Notes Math 571:85–100
15. Bookstein FL (1989) Principal warps: thin-plate splines and the decomposition of deformations. IEEE Trans Pattern Anal Mach Intell 11:567–585
16. Adams DC, Rohlf FJ, Slice DE (2004) Geometric morphometrics: ten years of progress following the "revolution". Ital J Zool 71:5–16
17. Menesatti P, Angelini C, Pallottino F, Antonucci F, Aguzzi J, Costa C (2012) RGB color calibration for quantitative image analysis: the "3D Thin-Plate Spline" warping approach. Sensors 12:7063–7079
18. Bookstein FL (1991) Morphometric tools for landmark data: geometry and biology. Cambridge University Press, Cambridge, UK
19. Pascale D (2003) A review of RGB color space: from xyY to R′G′B′. The Babel Color Company, Montreal, QC, p 35

INDEX

Pasquale Tripodi (ed.), *Crop Breeding: Genetic Improvement Methods*, Methods in Molecular Biology, vol. 2264, https://doi.org/10.1007/978-1-0716-1201-9, © Springer Science+Business Media, LLC, part of Springer Nature 2021