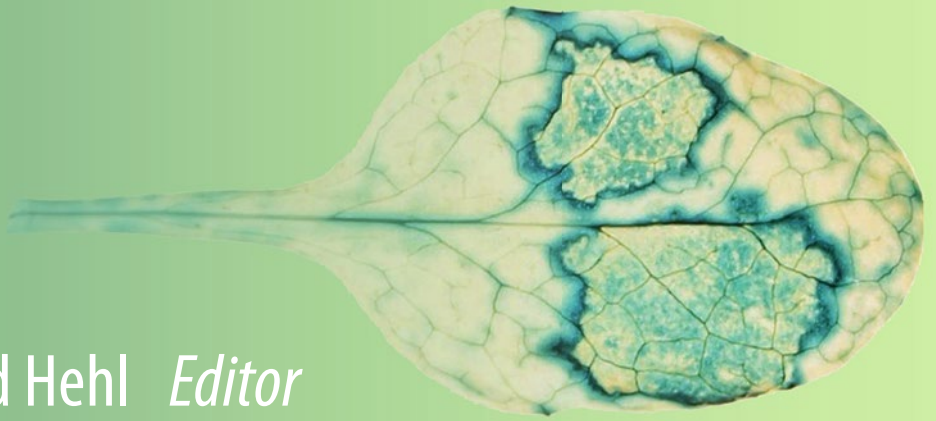


Methods in
Molecular Biology 1482

Springer Protocols



Reinhard Hehl *Editor*

Plant Synthetic Promoters

Methods and Protocols

 Humana Press

METHODS IN MOLECULAR BIOLOGY

Series Editor
John M. Walker
School of Life and Medical Sciences
University of Hertfordshire
Hatfield, Hertfordshire, AL10 9AB, UK

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


Plant Synthetic Promoters

Methods and Protocols

Edited by

Reinhard Hehl

Institut für Genetik, Technische Universität Braunschweig, Braunschweig, Germany

 Humana Press

Editor

Reinhard Hehl
Institut für Genetik
Technische Universität Braunschweig
Braunschweig, Germany

ISSN 1064-3745 ISSN 1940-6029 (electronic)
Methods in Molecular Biology
ISBN 978-1-4939-6394-2 ISBN 978-1-4939-6396-6 (eBook)
DOI 10.1007/978-1-4939-6396-6

Library of Congress Control Number: 2016949073

© Springer Science+Business Media New York 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Humana Press imprint is published by Springer Nature
The registered company is Springer Science+Business Media LLC New York

Preface

The promoter is one of the fundamental elements for spatial and temporal gene expression regulation. Understanding its function has fascinated generations of scientists. Despite many other levels of gene expression regulation at the chromatin and post-transcriptional levels, the main prerequisite for expression is the promoter-driven transcription of the genes.

The detailed understanding of the regulatory elements required for transcription permits the de novo assembly of synthetic promoters by combining *cis*-regulatory elements with minimal promoter elements towards conferring new and specific transcription patterns in plants. Such synthetic promoters can be widely used in basic and applied research.

Fused to a reporter gene, the activity of a synthetic promoter can be monitored over time and space, thus adding to our understanding of promoter function and the function of the transcription factors interacting with specific *cis*-elements. In the applied field, synthetic promoters are useful to drive gene expression specifically for a desired purpose. This could be the expression of resistance genes in response to pathogen infection or the expression of genes for engineering or modifying metabolic pathways.

This book assembles experimental and bioinformatics protocols for the design and experimental testing of synthetic promoters. The identification of *cis*-regulatory elements potentially achieving the desired expression of a gene is at the core of synthetic promoter design. For this, several bioinformatics chapters are presented. The experimental verification of the proposed expression profile conferred by the *cis*-regulatory elements requires the assembly of synthetic promoters. Several chapters are dedicated to the assembly of synthetic promoters, also including specific software tools to facilitate promoter design. Transient and transgenic reporter gene technology is a prominent approach to test the spatial and temporal expression driven by synthetic promoters, and several chapters address this approach. In summary, this book covers all steps required from the identification of *cis*-regulatory elements, over synthetic promoter design, to the experimental analysis of synthetic promoter function.

Braunschweig, Germany

Reinhard Hehl

Contents

<i>Preface</i>	<i>v</i>
<i>Contributors</i>	<i>ix</i>
1 What Have We Learned About Synthetic Promoter Construction? <i>Paul J. Rushton</i>	1
2 Quantitative Analysis of <i>Cis</i> -Regulatory Element Activity Using Synthetic Promoters in Transgenic Plants <i>Geoffrey Benn and Katayoon Dehesb</i>	5
3 The Identification of <i>Cis</i> -Regulatory Sequence Motifs in Gene Promoters Based on SNP Information <i>Paula Korkuć and Dirk Walther</i>	31
4 Quantitative Analysis of Protein–DNA Interaction by qDPI-ELISA. <i>Stefan M. Fischer, Alexander Böser, Jan P. Hirsch, and Dierk Wanke</i>	49
5 Analyzing Synthetic Promoters Using Arabidopsis Protoplasts. <i>Ralf Stracke, Katharina Thiedig, Melanie Kublmann, and Bernd Weisshaar</i>	67
6 Selecting Hypomethylated Genomic Regions Using MRE-Seq <i>Elisabeth Wischnitzki, Kornel Burg, Maria Berenyi, and Eva Maria Sebr</i>	83
7 Spatio-Temporal Imaging of Promoter Activity in Intact Plant Tissues. <i>Tou Cheu Xiong, Frédéric Sanchez, Jean-François Briat, Frédéric Gaymard, and Christian Dubos</i>	103
8 Novel Synthetic Promoters from the Cestrum Yellow Leaf Curling Virus. <i>Dipak Kumar Sahoo, Shayan Sarkar, Indu B. Maiti, and Nrisingha Dey</i>	111
9 Fast and Efficient Cloning of <i>Cis</i> -Regulatory Sequences for High-Throughput Yeast One-Hybrid Analyses of Transcription Factors <i>Zsolt Kelemen, Jonathan Przybyla-Toscano, Nicolas Tissot, Loïc Lepiniec, and Christian Dubos</i>	139
10 The <i>Physcomitrella patens</i> System for Transient Gene Expression Assays <i>Johanne Thévenin, Wenjia Xu, Louise Vaisman, Loïc Lepiniec, Bertrand Dubreucq, and Christian Dubos</i>	151
11 Analysis of Microbe-Associated Molecular Pattern-Responsive Synthetic Promoters with the Parsley Protoplast System <i>Konstantin Kanofsky, Mona Lehmeyer, Jutta Schulze, and Reinhard Hehl</i>	163
12 A Framework for Discovering, Designing, and Testing MicroProteins to Regulate Synthetic Transcriptional Modules <i>Elisa Fiume, Niek de Klein, Seung Yon Rhee, and Enrico Magnani</i>	175
13 Simultaneous Analysis of Multiple Promoters: An Application of the PC-GW Binary Vector Series. <i>Jyoti Dalal</i>	189

14 GenoCAD Plant Grammar to Design Plant Expression Vectors
for Promoter Analysis 219
Anna Coll, Mandy L. Wilson, Kristina Gruden, and Jean Peccoud

15 Bioinformatic Identification of Conserved Cis-Sequences
in Coregulated Genes 233
Lorenz Bülow and Reinhard Hehl

16 In Silico Expression Analysis 247
Julio Bolívar, Reinhard Hehl, and Lorenz Bülow

17 FootprintDB: Analysis of Plant *Cis*-Regulatory Elements,
Transcription Factors, and Binding Interfaces 259
Bruno Contreras-Moreira and Alvaro Sebastian

18 RSAT::Plants: Motif Discovery Within Clusters of Upstream Sequences
in Plant Genomes 279
*Bruno Contreras-Moreira, Jaime A. Castro-Mondragon, Claire Rioualen,
Carlos P. Cantalapiedra, and Jacques van Helden*

19 RSAT::Plants: Motif Discovery in ChIP-Seq Peaks of Plant Genomes 297
*Jaime A. Castro-Mondragon, Claire Rioualen, Bruno Contreras-Moreira,
and Jacques van Helden*

Index 323

Contributors

- GEOFFREY BENN • *Department of Plant Biology, University of California, Davis, CA, USA*
- MARIA BERENYI • *AIT Austrian Institute of Technology GmbH, Tulln, Austria*
- JULIO BOLÍVAR • *Institut für Genetik, Technische Universität Braunschweig, Braunschweig, Germany*
- ALEXANDER BÖSER • *ZHMB Pflanzenbiologie, Universität des Saarlandes, Saarbrücken, Germany*
- JEAN-FRANÇOIS BRIAT • *Biochimie et Physiologie Moléculaire des Plantes, CNRS, INRA, Montpellier SupAgro, Univ. Montpellier, Montpellier, France*
- LORENZ BÜLOW • *Julius Kühn-Institut, Federal Research Centre for Cultivated Plants, Institute for Breeding Research on Agricultural Crops, Quedlinburg, Germany*
- KORNEL BURG • *AIT Austrian Institute of Technology GmbH, Tulln, Austria*
- CARLOS P. CANTALAPIEDRA • *Estación Experimental de Aula Dei-CSIC, Zaragoza, Spain*
- JAIME A. CASTRO-MONDRAGON • *INSERM, Marseille, France; Aix Marseille University, Marseille, France*
- ANNA COLL • *Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia*
- BRUNO CONTRERAS-MOREIRA • *Estación Experimental de Aula Dei-CSIC, Zaragoza, Spain; Fundación ARAID, Zaragoza, Spain*
- JYOTI DALAL • *Department of Crop Science, North Carolina State University, Raleigh, NC, USA*
- KATAYOON DEHESH • *Department of Plant Biology, University of California, Davis, CA, USA*
- NRISINGHA DEY • *Department of Gene Function and Regulation, Institute of Life Sciences, Chandrasekharpur, Bhubaneswar, Odisha, India; Department of Biotechnology, Government of India, Chandrasekharpur, Bhubaneswar, Odisha, India*
- CHRISTIAN DUBOS • *Biochimie et Physiologie Moléculaire des Plantes, CNRS, INRA, Montpellier SupAgro, Univ. Montpellier, Montpellier, France*
- BERTRAND DUBREUCQ • *INRA, Institut Jean-Pierre Bourgin, Versailles, France; AgroParisTech, Saclay Plant Sciences, Institut Jean-Pierre Bourgin, Versailles, France*
- STEFAN M. FISCHER • *ZMBP Pflanzenphysiologie, Universität Tübingen, Tübingen, Germany*
- ELISA FIUME • *Institut Jean-Pierre Bourgin, INRA Centre de Versailles-Grignon, Versailles, France*
- FRÉDÉRIC GAYMARD • *Biochimie et Physiologie Moléculaire des Plantes, CNRS, INRA, Montpellier SupAgro, Univ. Montpellier, Montpellier, France*
- KRISTINA GRUDEN • *Department of Biotechnology and Systems Biology, National Institute of Biology, Ljubljana, Slovenia*
- REINHARD HEHL • *Institut für Genetik, Technische Universität Braunschweig, Braunschweig, Germany*
- JACQUES VAN HELDEN • *INSERM, Marseille, France; Aix Marseille University, Marseille, France*

- JAN P. HIRSCH • *ZHMB Pflanzenbiologie, Universität des Saarlandes, Saarbrücken, Germany*
- KONSTANTIN KANOFSKY • *Institut für Genetik, Technische Universität Braunschweig, Braunschweig, Germany*
- ZSOLT KELEMEN • *INRA, Institut Jean-Pierre Bourgin, Saclay Plant Sciences, Versailles, France; AgroParisTech, Institut Jean-Pierre Bourgin, Saclay Plant Sciences, Versailles, France*
- NIEK DE KLEIN • *Department of Plant Biology, Carnegie Institution for Science, Stanford, CA, USA; Department of Genetics, University Medical Center Groningen, Groningen, The Netherlands*
- PAULA KORUKUĆ • *Max Planck Institute for Molecular Plant Physiology, Potsdam-Golm, Germany*
- MELANIE KUHLMANN • *Genome Research, Bielefeld University, Bielefeld, Germany*
- MONA LEHMEYER • *Institut für Genetik, Technische Universität Braunschweig, Braunschweig, Germany*
- LOÏC LEPINIEC • *INRA, Institut Jean-Pierre Bourgin, Saclay Plant Sciences, Versailles, France; AgroParisTech, Institut Jean-Pierre Bourgin, Saclay Plant Sciences, Versailles, France*
- ENRICO MAGNANI • *Institut Jean-Pierre Bourgin, INRA Centre de Versailles-Grignon, Versailles, France*
- INDU B. MAITI • *KTRDC, College of Agriculture Food and Environment, University of Kentucky, Lexington, KY, USA*
- JEAN PECCOUD • *Department of Chemical & Biological Engineering, Colorado State University, Fort Collins, CO, USA; GenoFAB, LLC, San Francisco, CA, USA*
- JONATHAN PRZYBYLA-TOSCANO • *Biochimie et Physiologie Moléculaire des Plantes, INRA/CNRS/SupAgro-M/UM2, Montpellier, France*
- SEUNG YON RHEE • *Department of Plant Biology, Carnegie Institution for Science, Stanford, CA, USA*
- CLAIRE RIOUALEN • *INSERM, Marseille, France; Aix Marseille University, Marseille, France*
- PAUL J. RUSHTON • *22nd Century Group Inc., Clarence, NY, USA*
- DIPAK KUMAR SAHOO • *KTRDC, College of Agriculture Food and Environment, University of Kentucky, Lexington, KY, USA; Department of Agronomy, Iowa State University, Ames, IA, USA*
- FRÉDÉRIC SANCHEZ • *Biochimie et Physiologie Moléculaire des Plantes, CNRS, INRA, Montpellier SupAgro, Univ., Montpellier, France*
- SHAYAN SARKAR • *Department of Gene Function and Regulation, Institute of Life Sciences, Chandrasekharpur, Bhubaneswar, Odisha, India*
- JUTTA SCHULZE • *Institut für Pflanzenbiologie, Technische Universität Braunschweig, Braunschweig, Germany*
- ALVARO SEBASTIAN • *Evolutionary Biology Group, Adam Mickiewicz University, Poznan, Poland*
- EVA MARIA SEHR • *AIT Austrian Institute of Technology GmbH, Tulln, Austria*
- RALF STRACKE • *Genome Research, Bielefeld University, Bielefeld, Germany*
- JOHANNE THÉVENIN • *INRA, Institut Jean-Pierre Bourgin, Saclay Plant Sciences, Versailles, France; AgroParisTech, Institut Jean-Pierre Bourgin, Saclay Plant Sciences, Versailles, France*
- KATHARINA THIEDIG • *Genome Research, Bielefeld University, Bielefeld, Germany*

- NICOLAS TISSOT • *Biochimie et Physiologie Moléculaire des Plantes, INRA/CNRS/SupAgro-M/UM2, Montpellier, France*
- LOUISE VAISMAN • *INRA, Institut Jean-Pierre Bourgin, Saclay Plant Sciences, Versailles, France; AgroParisTech, Institut Jean-Pierre Bourgin, Saclay Plant Sciences, Versailles, France*
- DIRK WALTHER • *Max Planck Institute for Molecular Plant Physiology, Potsdam-Golm, Germany*
- DIERK WANKE • *ZMBP Pflanzenphysiologie, Universität Tübingen, Tübingen, Germany; ZHMB Pflanzenbiologie, Universität des Saarlandes, Saarbrücken, Germany*
- BERND WEISSHAAR • *Genome Research, Bielefeld University, Bielefeld, Germany*
- MANDY L. WILSON • *Biocomplexity Institute of Virginia Tech, Blacksburg, VA, USA*
- ELISABETH WISCHNITZKI • *AIT Austrian Institute of Technology GmbH, Tulln, Austria*
- TOU CHEU XIONG • *Biochimie et Physiologie Moléculaire des Plantes, CNRS, INRA, Montpellier SupAgro, Univ., Montpellier, France*
- WENJIA XU • *INRA, Institut Jean-Pierre Bourgin, Saclay Plant Sciences, Versailles, France; AgroParisTech, Institut Jean-Pierre Bourgin, Saclay Plant Sciences, Versailles, France*

Chapter 1

What Have We Learned About Synthetic Promoter Construction?

Paul J. Rushton

Abstract

The molecular components of transcriptional regulation are modular. Transcription factors have domains for specific functions such as DNA binding, dimerization, and protein–protein interactions associated with transcriptional activation and repression. Similarly, promoters are modular. They consist of combinations of *cis*-acting elements that are the binding sites for transcription factors. It is this promoter architecture that largely determines the expression pattern of a gene. The modular nature of promoters is supported by the observation that many *cis*-acting elements retain their activities when they are taken out of their native promoter context and used as building blocks in synthetic promoters. We therefore have a large collection of *cis*-acting elements to use in building synthetic promoters and many minimal promoters upon which to build them. This review discusses what we have learned concerning how to use these building blocks to make synthetic promoters. It has become clear that we can increase the strength of a promoter by adding increasing numbers of *cis*-acting elements. However, it appears that there may be a sweet spot with regard to inducibility as promoters with increasing numbers of copies of an element often show increased background expression. Spacing between elements appears important because if elements are placed too close together activity is lost, presumably due to reduced transcription factor binding due to steric hindrance. In many cases, promoters that contain combinations of *cis*-acting elements show better expression characteristics than promoters that contain a single type of element. This may be because multiple transcription factor binding sites in the promoter places it at the end of multiple signal transduction pathways. Finally, some *cis*-acting elements form functional units with other elements and are inactive on their own. In such cases, the complete unit is required for function in a synthetic promoter. Taken together, we have learned much about how to construct synthetic promoters and this knowledge will be crucial in both designing promoters to drive transgenes and also as components of defined regulatory networks in synthetic biology.

Key words Synthetic promoter, *Cis*-acting elements, Synthetic biology, Transgene expression, Plant biotechnology

1 Introduction

This review focuses mainly on the synthetic promoter projects that I have been involved with and serves as a guide to producing the best synthetic promoters. There are general trends, some of which we could not have predicted when we first started to construct

synthetic promoters, but there will always be exceptions to the rules. The reader is urged to use the observations presented here to help them in their own synthetic promoter projects but ultimately it is the activity of the constructed promoter that will decide whether a project is successful. One final note, in biotech projects a synthetic promoter will be used to drive a transgene and it is the best possible transgene expression that decides whether a project has ultimately been successful or not. A synthetic promoter can be used to optimize expression levels so that the transgene is expressed at the right time, in the right place, and to the optimum level. This potential optimization of expression (where, when, and how much) is where the advantage of synthetic promoters lies over native ones.

2 The Modular Nature of Transcriptional Regulation

The modular nature of transcription (and indeed signaling in general) has become apparent. Proteins have specific domains for certain function such as dimerization, ligand binding, nuclear localization, and so on. These domains can often retain their activities in domain swap or addition experiments. With transcription factors, this modularity is very clear. It includes nuclear localization domains, dimerization domains, calmodulin binding domains, protein–protein interaction domains associated with transcriptional activation or repression, and many others [1]. Building synthetic transcription factors with altered activities is therefore possible. For example, adding a repression domain such as an EAR domain from an ERF transcription factor can transform a transcription factor that normally functions as an activator into a dominant negative [2].

Similarly, promoters are modular as they typically contain combinations of *cis*-acting elements that are the binding sites for transcription factors. It is this promoter architecture that largely determines the expression pattern of a gene as it determines the specificity of transcription factor binding to the promoter. At the level of the promoter, binding of the transcription factors to the DNA is accompanied by protein–protein interactions between transcription factors themselves and also interactions with the general transcriptional machinery (general transcription factors, co-activators, and co-repressors) and other proteins that alter chromatin structure [1]. With each promoter containing multiple transcription factor binding sites and also with each transcription factor potentially forming multiple protein–protein interactions, it was originally unclear whether there would be any chance that a *cis*-acting element, when taken out of its native promoter context, could retain its activity [3]. This retention of activity would be a prerequisite for the construction of synthetic promoters.

In the late 1990s, I started a project on constructing pathogen-inducible synthetic promoters. There were many reports from the

literature of defined *cis*-acting elements retaining activity in synthetic promoters but it was unclear how widespread this phenomenon was. Several different types of known pathogen-responsive *cis*-acting elements were tested in synthetic promoters and strikingly, the majority of these elements retained their activity [3]. This included GCC-like boxes, W boxes, and Box D (which is still ill-defined). This showed that transcription at the promoter level is indeed modular and that many of these DNA modules can therefore be used to construct synthetic promoters. A synthetic promoter could therefore be build up from any number or combination of these modular building blocks in a similar way that someone builds something from Lego blocks.

3 Making a Synthetic Promoter

In its simplest form, a synthetic promoter will consist of a minimal promoter (the binding sites for general transcription factors including RNA polymerase II) and a defined *cis*-acting element [4]. The minimal promoter will typically contain a TATA box and a site at which transcription will start but little else as this may influence the expression characteristics of the promoter. Upstream of this minimal promoter are placed the *cis*-acting elements that will determine the expression characteristics of any transgene whose expression is driven by the promoter. These *cis*-acting elements can include any number of copies of an individual *cis*-acting element or combinations of different elements in any order and in any number. The possibilities are seemingly endless. With current advances in DNA technology, it is possible to simply synthesize any given synthetic promoter and this can speed up the process of building a promoter considerably. However, previously synthetic promoters were typically synthesized from ligating oligonucleotides containing the defined *cis*-acting element sequences upstream of a minimal promoter (Fig. 1).

Using technology based on two different restriction endonucleases with compatible sticky ends, this approach has the advantage that the resultant promoters can be used like Lego building blocks to optimize and test synthetic promoters. For this reason, this approach is still valuable today. Briefly, a defined *cis*-acting element is synthesized as two oligonucleotides, one for each strand of the DNA. When annealed together, the double-stranded DNA has sticky ends at both the 5 prime and 3 prime ends that are compatible (for example SpeI and XbaI or BamHI and BglII). The single copy of the *cis*-acting element is inserted into the corresponding restriction enzyme sites just upstream of a chosen minimal promoter to create a synthetic promoter with a single copy of the element (a 1 × construct). The beauty of this strategy becomes apparent when this 1 × construct is used to make other synthetic promoters. The 1 × construct is cut with a restriction enzyme that cuts the backbone

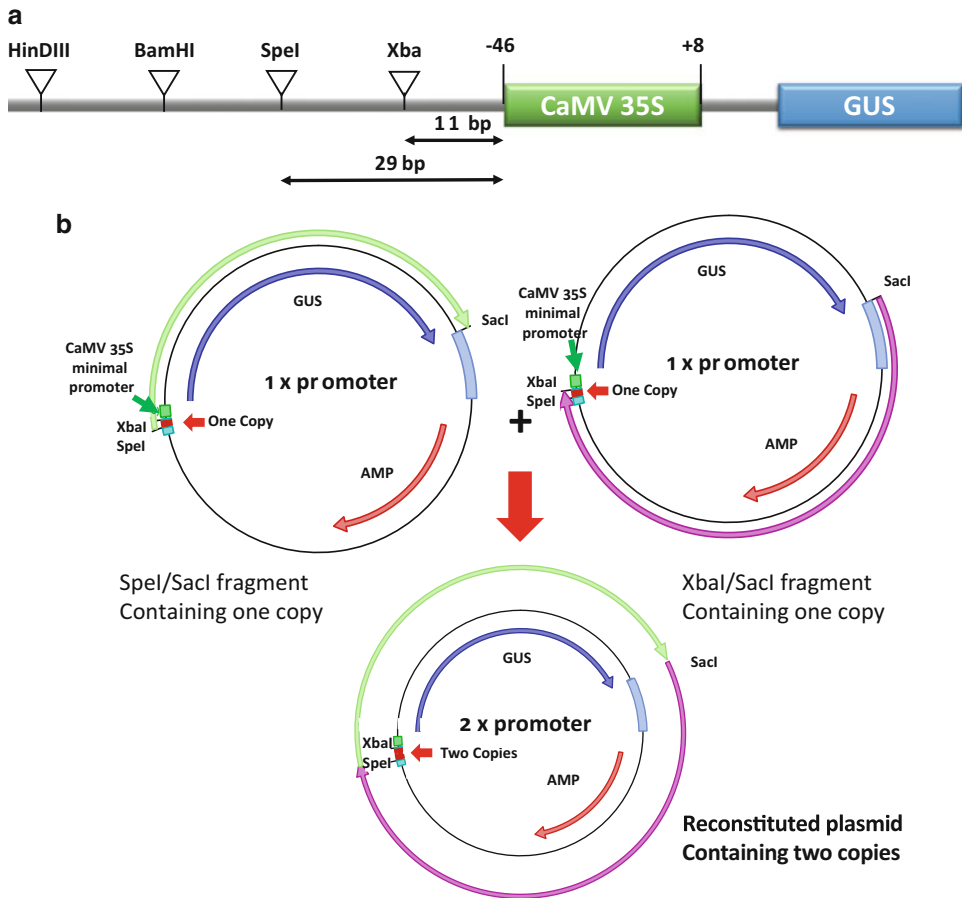


Fig. 1 A system to produce synthetic promoters with any number of *cis*-acting elements in any order. **(a)** The minimal promoter and restriction sites in MS23 (pBT10) [3]. Single-stranded oligonucleotides containing defined *cis*-acting elements with a SpeI sticky end at the 5' end and an XbaI sticky end at the 3' end are annealed and then inserted into SpeI/XbaI double-digested vector DNA 11 bp upstream of the CaMV35S -46 minimal promoter. **(b)** How to make a 2× element promoter construct from 1× element promoters. In two separate restriction digests, the 1× promoter DNA is digested by either SacI and XbaI or SacI and SpeI. In each case the fragment containing the single copy of the *cis*-acting element is gel purified and the SpeI/SacI and XbaI/SacI fragments are then annealed to give a 2× element promoter. The ligation of the SpeI and XbaI sticky ends destroys the restriction site in the middle of the 2× element yielding a SpeI site at the 5' end and an XbaI site at the 3' end. This pattern of restriction sites is identical to the 1× element construct and means that the process can be repeated to yield 4× and then 8× constructs and so on. The beauty of this system is that by using different promoter constructs as starting materials, promoters containing combinations of elements in any number and in any order can be produced

of the plasmid and then either of the two enzymes with compatible sticky ends (for example SpeI or XbaI). In each case, the fragment that contains the *cis*-acting element is then chosen and the two pieces are ligated together. Because each fragment contained a copy of the element, the resulting synthetic promoter contains two copies of the defined *cis*-acting element (a 2× construct). The inventive

step of this approach is that where the two copies of the element come together, they ligate together as they have compatible sticky ends. However, the restriction site between the two elements is not recreated because the two restriction sites are different. The result is one piece of DNA with no internal restriction site but the same 5 prime and 3 prime sites that you started with. This means that the process can be repeated and two copies can become four and then eight. In addition, by choosing different *cis*-acting element monomers, promoters can be constructed with any number of elements in any combination and in any order. Once monomer constructs are available that contain different *cis*-acting elements, they can then truly act as Lego building blocks for building synthetic promoters to the design of the researcher [3, 4].

4 The Effect of *Cis*-acting Element Number on Strength and Inducibility

One of the first questions that I asked when constructing synthetic promoters was “What is the effect of increasing the number of copies of a single *cis*-acting element in a synthetic promoter?”. Figure 2 shows that increasing the number of copies progressively from one to eight increases the strength of the promoter, presumably by providing more transcription factor binding sites. This suggests that an increasing number of transcription factors bound to the

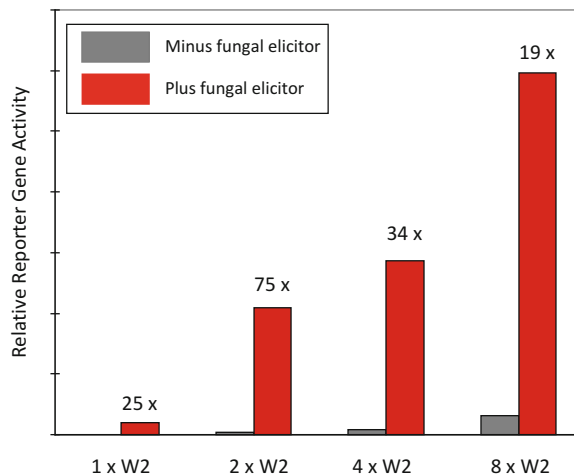


Fig. 2 Increasing the number of *cis*-acting elements in a synthetic promoter increases strength. 1 × W2, 2 × W2, 4 × W2, and 8 × W2 synthetic promoters were tested for induction by a fungal elicitor in a parsley transient expression system [3]. Increasing the number of copies progressively from one to eight increases the strength of the promoter, although the best signal-to-noise ratio is obtained with a 2 × W2 construct due to an increase in background expression with an increasing number of elements

promoter increases the rate of transcription from the synthetic promoter. This increase was observed not only in systems of reduced complexity but also in transgenic plants and has profound consequences for synthetic promoters in general. The fact that we can alter the strength of a synthetic promoter by varying the number of *cis*-acting element building blocks in the promoter means that we can modulate promoter strength by design. This immediately underlines an advantage of synthetic promoters over native ones—with synthetic promoters we can vary strength to find the optimum expression level of a transgene. This modulation is not possible when using a single native promoter.

One additional observation from Fig. 1, and this was also apparent in transgenic lines, is that there is one downside to increasing strength, namely that as the synthetic promoters get stronger the level of background expression in inducible promoters (such as pathogen-inducible promoters) often increases and therefore the fold inducibility is reduced [2]. The exact reason for this is unclear. It may be that more binding sites increase the level of basal transcription or alternatively allow increased binding of transcription factors that may have a lower affinity for the native promoter. Either of these two possibilities may increase the level of transcription in the absence of the signal and lead to increased background expression.

Again, the choice of synthetic promoter will be driven by the choice of transgene and how this transgene is best expressed. For some projects, reasonable levels of background expression could be tolerated (as the expression pattern is still considerably better than constitutive overexpression using, for example, the CaMV 35S promoter). For others, the best inducibility is required such as expression in infected plant tissues but not uninfected ones when using pathogen-inducible synthetic promoters. For the former, a strong promoter with eight copies of a *cis*-acting element might be best, whereas for the latter two copies may be preferred as it shows the best inducibility (signal:noise). The above examples provide a nice example of the value of synthetic promoters—we are designing promoters for specific purposes and different promoters will be suited to different projects.

5 The Effect of Spacing on Promoter Strength

Once the building blocks for a synthetic promoter have been chosen, how do we put them together to make a good synthetic promoter? Well one of the first considerations is spacing. This includes not only the spacing between multiple copies of an element but also spacing with respect to the minimal promoter. When I first started to construct a range of synthetic promoters, I suspected that spacing between elements might be crucial for promoter activity driven by the need for the cognate transcription factors to interact with other proteins in a productive way. However, although

spacing turns out to be important, results suggest that this is not in the way that I had envisaged. Elements such as GCC boxes and W boxes appear to function independently of each other and spacing between the elements themselves appears to have little or no effect. In fact, systematic rotating of *cis*-acting elements relative to each other by one base pair at a time through one complete turn of the DNA helix had a negligible effect on promoter activity. With *cis*-acting elements that function independently, it would appear that the exact distance between them has little or no effect.

However, spacing is crucial to synthetic promoter activity in one crucial respect—if you place the *cis*-acting elements too close together they lose activity. The exact distance will need to be determined experimentally, but in my experience if the core sequences of elements are less than 10–15 bp apart then activity is reduced. This makes sense if one considers the binding of proteins to the short promoter DNA sequence. A transcription factor will require a certain length of DNA to bind to and if this synthetic promoter puts two binding sites too close together then binding to one site will preclude binding of another transcription factor to the next site. This reduction in activity due to steric hindrance is also seen if the promoter puts a *cis*-acting element too close to the minimal/core promoter. In this case, general transcription factors will compete for binding to the promoter with the transcription factors that bind to the *cis*-acting elements that have been added upstream. As a rule of thumb, at least 15–20 bp should be allowed between multiple copies of a *cis*-acting element in a synthetic promoter so that their core sequences are separated by close to 50 bp. In addition, it is best to have at least 50 bp between the TATA Box and the core sequence of the first *cis*-acting element placed upstream of it (Fig. 3).

6 Combinations of *Cis*-acting Elements Appear Best

Some aspects of synthetic promoter technology were not necessarily predictable and only became apparent once a systematic approach was used to design a spectrum of different synthetic promoters [3, 4]. One of these observations is that synthetic promoters that contain more than one type of *cis*-acting element may be better than simpler promoters that contain multiple copies of only one type of element. With pathogen-inducible synthetic promoters, high expression at infection sites coupled to low background expression in non-infected tissues is preferred in order to reduce any negative effects of transgene expression in non-infected tissues. It was observed that promoters that contain more than one type of *cis*-acting element showed the best inducibility coupled with lower background, making them much better suited for transgene expression. It is likely that the reason for this is that multiple different *cis*-acting elements place a synthetic promoter at the end of more

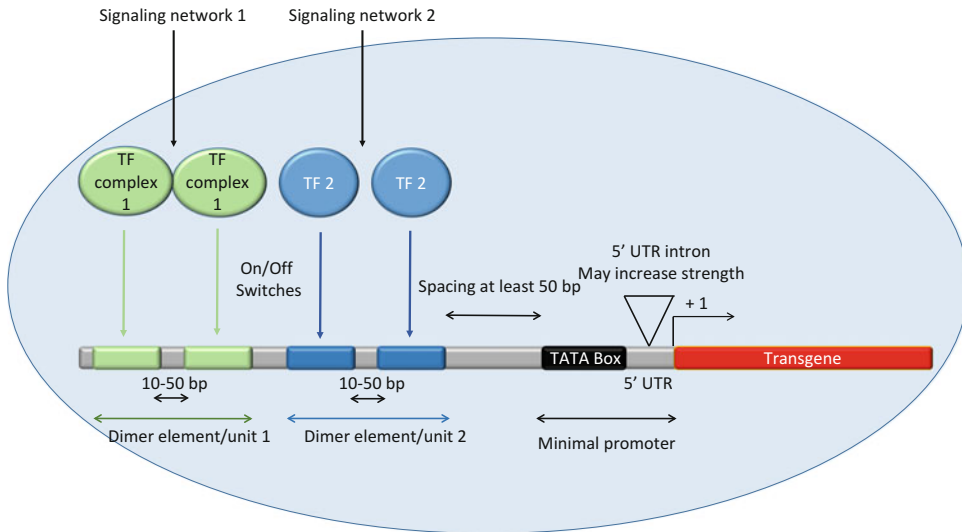


Fig. 3 What we have learned about synthetic promoter construction. Although promoters will need to be optimized for each transgene and each required transgenic plant line, a number of observations have been made concerning the best starting strategies and are summarized in this figure. The best signal:noise ratios appear to be produced from dimers of individual *cis*-acting elements/units. These dimers should be placed with spacing of 10–50 bp between the elements to avoid reduced activity. Synthetic promoters that contain more than one type of *cis*-acting element/unit appear to give better expression characteristics, presumably because this places the promoter at the end point of more than one signaling pathway so that it responds to multiple inputs. The minimal promoter should consist of the TATA Box, the 5' UTR and the start of transcription only. This ensures that it is functional but does not affect promoter characteristics. The exception to this is that certain minimal promoters that contain introns in their 5' UTRs may direct stronger expression. The individual *cis*-acting elements can contain a single element or consist of a functional unit of more than one element

than one signaling pathway. As the constructed promoter takes signals from multiple pathways and multiple transcription factor types, its activity is likely to be more tightly regulated resulting in a better expression pattern.

As we start to understand how to construct synthetic promoters for specific purposes, we can start to combine some of the observations when designing promoters. For example, two copies of an individual *cis*-acting element in a promoter probably give the best signal to noise ratio (Fig. 2) and multiple different *cis*-acting elements also seem to give better inducibility. It is therefore likely that a promoter that contains two copies of several *cis*-acting elements would be among the best promoters in terms of inducibility and that is exactly what was observed in the project reported by Rushton et al. [3]. The best synthetic pathogen-inducible promoter was $2 \times W2/2 \times S/2 \times D$. It combines three different types of *cis*-acting elements, two copies of each element, and at least two different families of transcription factors (WRKY and AP2/ERF) as end points in the signal transduction pathways. In addition, the elements are spaced far enough apart and from the minimal

promoter to avoid loss of activity due to steric hindrance. These observations give important pointers as to how we might use defined *cis*-acting elements to build the best synthetic promoters.

7 The Choice of Minimal Promoter

Many synthetic promoter projects have previously used the minimal CaMV 35S promoter as the start point for synthetic promoter construction. This was probably because the CaMV35S is the best characterized strong promoter and the -46 version shows minimal basal activity in the absence of added *cis*-acting elements. However, it appears that many minimal promoters can be used as the basis for synthetic promoter construction. Some work is however required as a minimal promoter needs to be defined that is active but that also does not affect expression characteristics (this normally contains the start of transcription and a TATA Box but little else). As a rule of thumb, the best minimal promoter to use is probably one from a gene whose expression characteristics already are closest to the desired expression characteristics of the final synthetic promoter. For a drought inducible promoter, this would be from a drought inducible gene, for a wound inducible promoter this would be from a wound inducible gene and so on.

One further observation is important in the production of synthetic promoters that are designed to direct strong expression levels because here the choice of minimal promoter may be more important. The promoters of several genes that show very high expression levels (for example several ubiquitin genes) contain introns in their 5 prime UTRs [5]. These introns appear to contribute to strength, and choosing a minimal promoter that contains such an intron may therefore be an important part of the design strategy when increasing strength.

8 Functional Units Need to Be Kept Together

More recently, we have gained more insights into the production of synthetic promoters. This can be illustrated by work using the GAG fragment from PMT promoters in tobacco [6]. The GAG fragment is so called because it consists of three parts. A G box followed by an AT-rich spacer region and then a GCC-like box. The GAG fragment, like the PMT promoters themselves directs jasmonate and wounding inducible expression (Fig. 4). The expression pattern for synthetic promoters with the GAG fragment is exquisite as wounding of a leaf or jasmonate treatment results in expression in the cortex of the root (tissue-specific expression at a distance). This is also the expression pattern of the native PMT promoters and illustrates that this small GAG fragment is sufficient to drive expression that is similar to the native promoter.

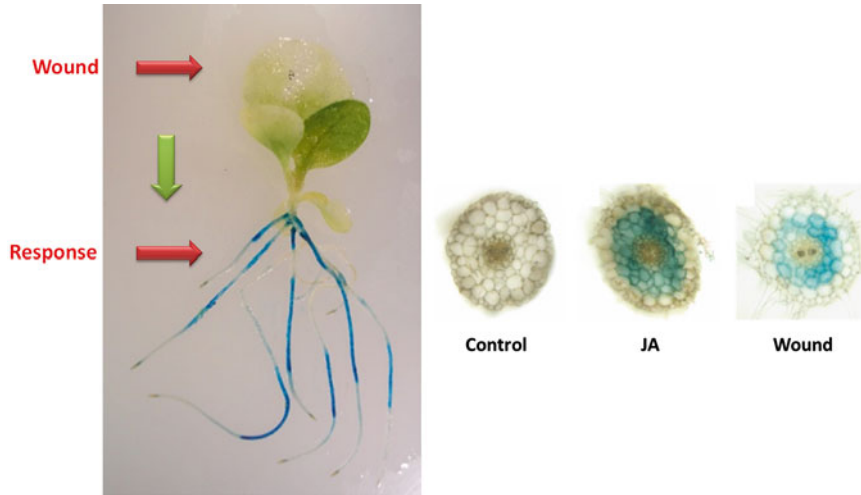


Fig. 4 Transgenic tobacco plants containing a 4 × GAG synthetic promoter show both wound and jasmonate inducible expression at a distance. Wounding or jasmonate treatment of the leaves leads to expression from the 4 × GAG synthetic promoter in the roots. This root-specific expression is concentrated in the cortex, the main site of nicotine biosynthesis

Synthetic promoters containing the GAG fragment and its three constituent parts present important new evidence concerning how to build synthetic promoters. Each one of the three constituent parts of the GAG fragment (the G box, the AT-rich region, and the GCC-like box) is inactive on its own. The G box and GCC-like box, although similar to known *cis*-acting elements that are bound by bHLH, bZIP, and AP2/ERF transcription factors, show no activity if used alone in synthetic promoters [6]. However, if the G box is combined with the GCC-like box, then jasmonate inducibility is restored. It is clear that the two *cis*-acting elements function together as a unit and that at least two transcription factors (a bHLH and an ERF) are required for function. However, the story does not end there because although the G box–GCC-like box unit is active, it is neither as strong nor as inducible as the G box–AT-rich region–GCC-like box unit (Fig. 5). It appears that the AT-rich region is required for full activity and that this activity is not dependent on the sequence of the region because an AT-rich region of different sequence but the same length appears similarly active. Taken together, it is clear that the GAG fragment is a unit consisting of three elements. Two of these elements appear to be binding sites for transcription factors and the third is most likely a spacer region. This suggests that the two transcription factors probably interact directly or indirectly for function of the GAG fragment.

The lesson for synthetic promoter construction is that some promoters consist of functional units with more than one constituent element. In such cases the entire unit needs to be used for activity. Importantly, if one uses each functional unit (for example

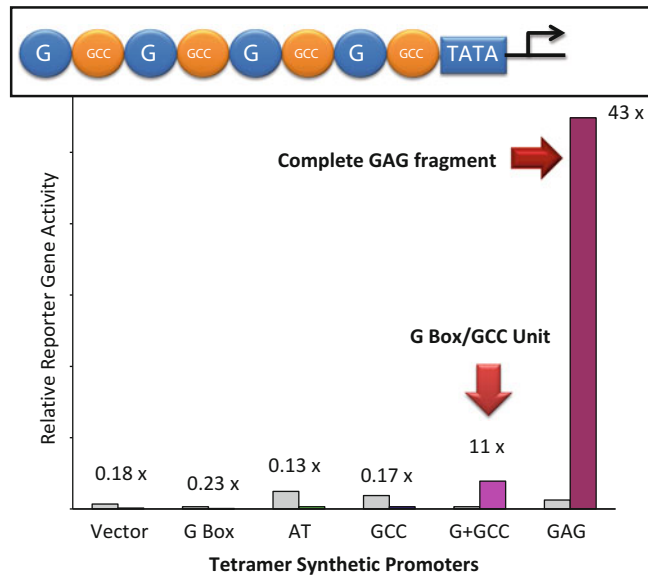


Fig. 5 Functional units consisting of more than one *cis*-acting element require all elements for full activity. Synthetic promoters containing tetramers of the GAG fragment and its constituent individual *cis*-acting elements (the G box, AT-rich region, and GCC-like element) were tested for activity (*gray bars*) and jasmonate inducibility (*colored bars*) in stably transformed BY-2 cells. None of the individual elements were functional alone. However, a tetramer of the G box–GCC-like element showed jasmonate inducibility suggesting that these elements are binding sites for transcription factors (bHLH and ERF) and together form a jasmonate response element in synthetic promoters. However, it is also clear that the GAG fragment is a unit consisting of three elements because synthetic promoters that also contain the AT-rich region (the complete GAG fragment) are both stronger and more inducible by jasmonate than the G box–GCC element. These data show that functional units that consist of more than one *cis*-acting element should be kept together when used in building synthetic promoters

the GAG fragment) in a similar way to *cis*-acting elements that function alone (such as W boxes) then the same rules apply. For example, $4 \times \text{GAG}$ is stronger than $2 \times \text{GAG}$ and so on. The modular nature of promoter technology still applies with some Lego bricks consisting of one *cis*-acting element whereas others consist of units of more than one element.

9 The Best Place to Start

Synthetic promoters have several potential advantages over native promoters. The main advantage is that the strength of the promoter can be altered to produce promoters that are stronger or weaker depending on the number of copies of each element/unit.

Additional potential advantages include the possibility of reducing unwanted expression characteristics by using a single *cis*-acting element from a promoter and eliminating other elements that may direct undesired expression characteristics. This has been a major theme in, for example, pathogen-inducible promoters where expression in uninfected tissues is undesirable. This strategy of eliminating other *cis*-acting elements has, however, met with only limited success. For example, a W box that directs pathogen-inducible expression (desired) often also directs wound-inducible expression (undesired). The likely cause of this is that the cognate transcription factors are involved in both pathogen and wound induced signaling and as a result, induction by the two stimuli cannot be separated.

Despite the potential advantages in using synthetic promoters to fine tune transgene expression in biotechnology projects, the best piece of advice is not to use one at all if a good native promoter is available that drives the desired expression characteristics! The other good piece of advice when choosing where to start, is to start with the native promoters that most closely fit the desired expression characteristics because they will be the best source of *cis*-acting element/unit building blocks to build an improved synthetic promoter. In the case of the GAG fragment, the source of the unit was the tobacco PMT promoters and the GAG fragment drives expression that has similarities to the full-length promoters. In the case of pathogen-inducible synthetic promoters, the best sources were the promoters of pathogenesis-related genes such as PR10s (the sources of various W boxes and Box D) [3].

10 Conclusions and Future Prospects

Many promising transgenes have failed not because of a poor choice of transgene but because of a poor choice of promoter with which to drive it with. I sometimes wonder how many transgenes have been discarded over the years as being unsuitable for improving crop plants based on results using high-level ectopic overexpression using the CaMV 35S promoter or other unsuitable promoters. The choice of promoter can make or break a project and over the years many plant scientists have been unimaginative in their choice of promoter. Synthetic promoters can change this and fine tuning promoter activity using synthetic promoters should be an increasingly important topic in plant biotechnology.

One area which should increase the use of synthetic plant promoters is synthetic biology. As we have seen above, signaling is modular, both at the protein domain level and the *cis*-acting element level with activity often residing in the individual protein domain or transcription factor binding site. Using these building blocks it should be possible to construct complete signaling pathways from the ground up using building blocks from different

proteins and promoters. With this approach, synthetic promoters will be a crucial part of synthetic biology as they represent synthetic end points for synthetic signaling pathways.

Acknowledgements

I would like to thank Peter Bowerman at BASF for critically reading this manuscript and Neal Stewart for inviting me to speak on this topic at the 2014 World Forum on Biology in Atlanta Georgia. This chapter is based on the seminar given there. I would like to thank Imre Somssich for his support and sound advice on the pathogen-inducible synthetic promoter projects. I would particularly like to thank Michael Timko for allowing me to include unpublished data on the GAG fragment in this chapter. I would also like to thank Reinhard Hehl for his patience while I was moving to a new position.

References

1. Liu W, Stewart CN (2016) Plant synthetic promoters and transcription factors. *Curr Opin Biotechnol* 37:36–44
2. Hiratsu K, Matsui K, Koyama T, Ohme-Takagi M (2003) Dominant repression of target genes by chimeric repressors that include the EAR motif, a repression domain, in *Arabidopsis*. *Plant J* 34:733–739
3. Rushton PJ, Reinstädler A, Lipka V, Lippok B, Somssich IE (2002) Synthetic plant promoters containing defined regulatory elements provide novel insights into pathogen- and wound-induced signaling. *Plant Cell* 14:749–762
4. Gurr SJ, Rushton PJ (2005) Engineering plants with increased disease resistance: how are we going to express it? *Trends Biotechnol* 2005(23):283–290
5. Hernandez-Garcia CM, Bouchard RA, Rushton PJ, Jones ML, Chen X, Timko MP, Finer JJ (2010) High level transgenic expression of soybean (*Glycine max*) GmERF and Gmubi gene promoters isolated by a novel promoter analysis pipeline. *BMC Plant Biol* 10:237. doi:[10.1186/1471-2229-10-237](https://doi.org/10.1186/1471-2229-10-237)
6. Sears MT, Zhang H, Rushton PJ, Wu M, Han S, Spano AJ, Timko MP (2014) NtERF32: a non-NIC2 locus AP2/ERF transcription factor required in jasmonate-inducible nicotine biosynthesis in tobacco. *Plant Mol Biol* 84:49–66. doi:[10.1007/s11103-013-0116-2](https://doi.org/10.1007/s11103-013-0116-2)

Quantitative Analysis of *Cis*-Regulatory Element Activity Using Synthetic Promoters in Transgenic Plants

Geoffrey Benn and Katayoon Dehesh

Abstract

Synthetic promoters, introduced stably or transiently into plants, are an invaluable tool for the identification of functional regulatory elements and the corresponding transcription factor(s) that regulate the amplitude, spatial distribution, and temporal patterns of gene expression. Here, we present a protocol describing the steps required to identify and characterize putative *cis*-regulatory elements. These steps include application of computational tools to identify putative elements, construction of a synthetic promoter upstream of *luciferase*, identification of transcription factors that regulate the element, testing the functionality of the element introduced transiently and/or stably into the species of interest followed by high-throughput *luciferase* screening assays, and subsequent data processing and statistical analysis.

Key words *Cis*-regulatory element, Synthetic promoter, *Luciferase* reporter, Stable transformation, Transient transformation

1 Introduction

Identification and characterization of promoter regulatory elements is critical for understanding how cells control the timing, spatial patterning, and levels of gene expression, thereby facilitating the regulation of complex signaling and metabolic networks. As such, this endeavor continues to attract much attention, as it has since the discovery of functional promoter elements [1]. In plants, studies of promoter elements were initially carried out by construction of promoter fragments fused to reporter genes, such as chloramphenicol acetyltransferase (CAT) or β -glucuronidase (GUS), as a proxy for sequence activity [2–5]. The resolution of this approach was subsequently increased through random mutagenesis, or linker scan of the natural promoter, whereby a small (~5–12 bp) fragment of the promoter of interest is either randomly mutagenized or replaced with a linker sequence [6, 7]. While the identification of regulatory elements via analysis of promoter fragments remains viable, this approach is inherently lengthy and

cumbersome. This inefficiency, however, has been addressed by the recent development of a wide variety of computational approaches centered on comparison of promoter sequences, in combination with co-expression analyses, resulting in the rapid discovery of putative regulatory elements [8–11]. Functional verification of a putative *cis*-regulatory element can subsequently be carried out by construction of a synthetic promoter consisting of one to several copies of the element fused to a minimal promoter and reporter gene, followed by introduction of the construct in planta [11–14]. Many studies utilize constructs consisting of these plant synthetic promoters fused to the GUS reporter to examine the activity of putative regulatory elements [12, 14]. However, the inherent stability of GUS limits these analyses of element functionality to single time-point comparisons of levels and spatial patterns of expression, thus precluding detailed profiling of temporal changes in activity. Development of the luciferase reporter system [15] addressed this deficiency and enabled exquisitely detailed time course analysis of synthetic promoter activity in response to signals induced by a wide range of inputs including circadian rhythms, stress treatments, and chemical stimuli [11, 13, 16–18]. In addition, the employment of transient assays using synthetic promoters expanded the analytical capability for testing the functionality of specific transcriptional regulator(s) in controlling the activity of the *cis*-regulatory element [17].

Here we provide a summary of these combinatorial approaches (Fig. 1), which were instrumental in the recognition and characterization of the Rapid Stress Response Element (RSRE) as a general-stress-responsive *cis*-regulatory element [11], and the further elucidation of the non-uniform contribution of different members of the Calmodulin-Binding Transcriptional Activator (CAMTA) family in regulation of this functional response element [17].

2 Materials

All solutions should be prepared with ultrapure water (i.e. MilliQ).

2.1 Stable Transformation of *Arabidopsis thaliana*

1. Luria Broth (LB): 10 g/L tryptone, 5 g/L yeast extract, 5 g/L NaCl, adjust pH to 7.0 with NaOH. Autoclave at 121 °C for 20–35 min and cool to at least 55 °C before adding antibiotics (see Note 1).

Fig. 1 (continued) *4xRSRE:Luciferase* (*4xRSRE:LUC*) reporter construct. (4a) Image of luciferase activity in *Arabidopsis* plants stably expressing the *4xRSRE:LUC* construct. (4b) Image of luciferase activity in an *N. benthamiana* leaf that has been transiently transformed with the *4xRSRE:LUC* reporter construct and its transcriptional activator CAMTA3. (5) Graph showing time course of *LUC* activity in response to wounding in *Arabidopsis* plants stably transformed with the *4xRSRE:LUC* reporter construct

1. ID of regulatory elements by comparative promoter sequence analysis.

Motif 6	
CGCGTT	p = 1.94e-07
CGCGT	p = 7.70e-05
GCGCGT	p = 3.48e-03
CCGCGT	p = 3.32e-04

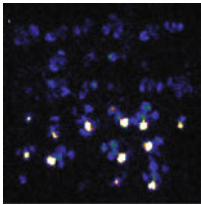
2. Selection of core element by statistical analysis and database searches.



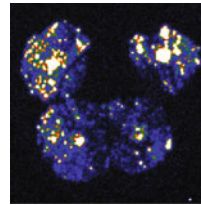
3. Creation of synthetic promoter:luciferase reporter construct.



4a. Stable transformation of synthetic promoter:luciferase construct into *A. thaliana*.



4b. Transient transformation of the synthetic promoter:luciferase construct into *N. benthamiana*.



5. High-throughput screening in CCD camera, followed by data processing and statistical analysis.

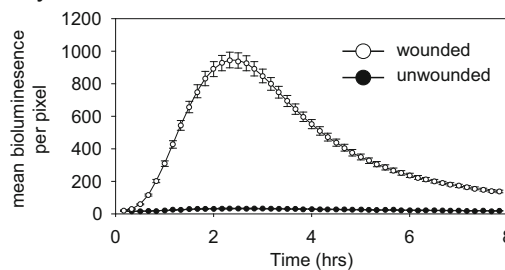


Fig. 1 Summary of protocol. Here, we summarize the steps required for identification of putative *cis*-regulatory elements and quantification of their activity in transient or stably transformed plant species. (1) Sample motif discovery readout. (2) Web Logo depiction of binding site of a TF, CAMTA3. (3) Simplified schematic of the

2. Sucrose solution: 5% (w/v) solution in water. Add silwet L-77 to 0.05% just prior to transformation of plants.
3. SOC media: 0.5% yeast extract, 2% tryptone, 10 mM NaCl, 2.5 mM KCl, 10 mM MgCl₂, 10 mM MgSO₄, 20 mM glucose. Add all ingredients except the glucose and autoclave at 121 °C for 20–35 min and cool to at least 55 °C. Then, using sterile technique, add filter-sterilized glucose to 20 mM.

2.2 *Nicotiana benthamiana* Transient Assay

1. Luria Broth (LB): *see* [Subheading 2.1](#).
2. Infiltration buffer: 50 mM 2-(N-morpholino) ethanesulfonic acid hydrate (MES) pH 5.6, 100 μM acetosyringone, 2 mM Na₃PO₄ (dodecahydrate), 0.5% glucose.
3. SOC media: *see* [Subheading 2.1](#).

2.3 Imaging

1. MS media (½ strength MS with 0.8% phytoagar): 2.22 g/L Murishige and Skoog (MS) basal medium, adjust pH to 5.7 using NaOH (1 M or lower). Add 0.8% phytoagar and autoclave for 20–30 min at 121 °C. Allow solution to cool to ~60 °C and then pour into plates (100 mm × 100 mm × 15 mm) to a depth of 5–6 mm. 1 L MS media is sufficient for preparation of 30 plates.
2. Luciferin working solution: 1 mM luciferin, 0.01% Tween 20. Solution may be prepared in advance and kept at 4 °C for up to 1 month.
3. Imaging system: Andor DU-484BV charge-coupled device (CCD) camera with Andor Solis software (v15). The camera is affixed to a light-tight box. Other imaging systems and software may be substituted.

3 Methods

3.1 Design of the Synthetic Promoter

1. Identify a list of genes from which to identify over-represented promoter motifs—for example a set of genes induced in response to a specific treatment, a set of genes mis-regulated in a mutant, or a set of genes commonly induced by a set of related stimuli.
2. Obtain the promoter sequences of the genes (*see* [Note 2](#)).
3. Perform motif discovery analyses on the set of promoter sequences. Paste promoter sequences into motif discovery tool(s) and analyze for over-represented sequence motifs (*see* [Note 3](#)).
4. Select a motif for further analysis. Selection should be based on the *p*-values derived from different motif discovery tools for a given element. Elements with strongly significant *p*-values are more likely to contribute to the control of gene expression in the conditions under study. Motifs with high scores in the

discovery analysis should also be queried against both the scientific literature and specialized *cis*-element databases, in order to assess whether they are novel motifs or have been previously identified (*see Note 4*).

5. Identify which promoters in the input data set contain the sequence of the chosen motif, then acquire the 10 bp of flanking sequences on either side of each instance of the element. If the number of promoters is small, this may be done manually by searching for the motif and copying and pasting the sequences into a FASTA-formatted file (for larger promoter sets, *see Note 5*).
6. Paste the FASTA file containing the core and flanking sequences into an alignment and visualization tool, such as weblogo (<http://weblogo.berkeley.edu/logo.cgi>) [19]. To design the final sequence employed in further analyses, assess the consensus among the flanking sequences to specifically determine how many bases of the flanking sequence are potentially conserved and therefore should be added to core element (Fig. 2, *see also Note 6*). Finally, design two to four tandem repeats of the putative regulatory sequence for construction of the synthetic promoter.
7. Adequate spacing between the copies of putative functional motif sequences needs to be ensured by addition of short spacer sequences, such as DNA sequence targets of six-base cutter restriction enzymes, that separate the motifs in the synthetic promoter. These spacer sequences should be searched against the literature and regulatory element databases to confirm that they do not constitute functional elements, or that the sequence combination of the spacer and the flanking region do not result in inadvertent construction of previously known functional elements absent in the promoters of interest (*see Note 7*).
8. Design a control version of the synthetic promoter by substituting three to four nucleotides within the core sequence

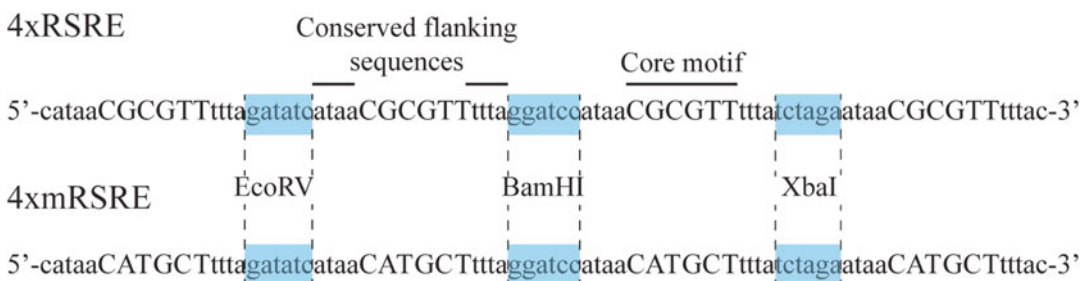


Fig. 2 Design of the tetrameric *RSRE* used in construction of synthetic promoter. Example of a multi-copy functional element (*4xRSRE*) and the mutated control (*4xmRSRE*) separated by recognition nucleotide sequences of different restriction enzymes, as previously described [11, 17]

(Fig. 2, *see* also **Note 8**). This approach will serve as a control for not only nucleotide specificity of the core sequence, but also as a vehicle to exclude inadvertent contribution of combined flanking and spacers sequences to the activity of the synthetic promoter.

3.2 Identification of Transcription Factors Interacting with Synthetic Promoters

1. Perform a literature search to determine if the core motif used in the synthetic promoter matches previously described transcription factor (TF) binding sites (*see* **Note 9**).
2. In parallel with the literature search, look for the motif in databases of known TF-binding sites (*see* **Note 4**).
3. If the core motif is not found in the literature or TF-binding databases, select and employ an alternative approach for identification of TFs that interact with the motif. One approach is to use the synthetic promoter and mutated control as bait in a yeast 1-hybrid screen against a library of cloned TFs [20] (*see* **Note 10**). A second option, which may be run in parallel, is to use plants stably transformed with the synthetic promoter fused to the luciferase reporter (*see* **Subheading 3.3**) in a forward genetic screen (*see* **Note 11**). A third option is to use the synthetic promoter to screen a recombinant expression library (such as λ gt11 cloning system) as previously described [21, 22], followed by validation of specificity using the mutated version of the core sequence as control.

3.3 Stable Transformation of *A. thaliana*

The stable transformation protocol for *A. thaliana* (Arabidopsis) is a slightly modified version of the procedure previously developed by Andrew Bent's lab (<http://www.plantpath.wisc.edu/fac/afb/protocol.html>) [23].

1. Using standard molecular biology techniques, transform the synthetic promoter into an appropriate vector for stable transformation of Arabidopsis. The vector should include bacterial and plant selection genes, a cloning site upstream of a minimal promoter, and the luciferase reporter gene (*see* **Note 12**).
2. Transform the vector into an appropriate strain of *Agrobacterium tumefaciens* (*see* **Note 13**). Thaw 50 μ L competent *A. tumefaciens* on ice, then add 1–2 μ L of vector DNA. Freeze the cells in liquid nitrogen for 5' and then thaw in a 37 °C water bath for 5'. Add 1 mL 30 °C SOC media to tube and incubate in a 30 °C shaker for 2 h. Plate 50 μ L and 200 μ L of transformation reaction on LB plates with appropriate antibiotics and incubate at 28 °C for 2 days. Pick individual colonies and use PCR followed by sequencing to confirm successful transformation.
3. Grow Arabidopsis plants in pots of soil covered with standard charcoal fiberglass window screening or other similar mesh. To

stimulate flowering, plants should be grown under long day conditions (**Note 14**).

4. Cut the first set of inflorescences off and allow the plants to produce additional inflorescences. This increases the number of flowers and as such enhances transformation efficiency.
5. Add appropriate antibiotic to 250 mL LB in a 1 L flask and at room temperature. Inoculate culture with a toothpick that has been scraped across the surface of the previously prepared glycerol stock of the *A. tumefaciens* transformed with the synthetic promoter-luciferase construct (*see Note 15*). Cap culture flask with aluminum foil and incubate in a shaker at 28 °C for 16–18 h.
6. Take a sample from the culture and measure the optical density (OD₆₀₀). The OD₆₀₀ should be 0.8–1.2. Pour cultures into 500 mL centrifuge bottles and spin down at 3000×g for 15 min at 15 °C. Resuspend pellet in 5% sucrose solution to OD₆₀₀=0.8 (*see Note 16*).
7. Add Silwet L-77 to resuspended bacteria, to a concentration of 0.05% and mix by inversion.
8. Wet two paper towels with deionized water and spread them out on the bottom of a standard-sized planting tray.
9. Pour *A. tumefaciens* solution into a shallow container (*see Note 17*). Depth of the solution should be about 4 cm.
10. Dip the bolts and rosettes into the *A. tumefaciens* solution and gently swirl for 3 s. Remove the plants from the culture and gently pat with a paper towel to remove excess *A. tumefaciens* solution (*see Note 18*).
11. Lay each treated pot on its side in the previously prepared planting tray. Cover with a second tray and move to growth chamber (*see Note 19*). After 24 h, upright the pots and move them to a fresh tray. From this point, grow the plants using standard light and watering conditions. Plants should be kept separate by staking up inflorescences. Once the seeds are set, reduce watering, and then withhold water once all seeds have matured.
12. Once plants have fully senesced, harvest seeds. Keep seeds from individual plants separate.
13. Harvest seeds and select transformants using an appropriate antibiotic or herbicide, depending on the vector used (*see Note 20*).

3.4 *Nicotiana benthamiana* Transient Assay

The transient transformation protocol is a modified version of the procedure previously described by Jurgen Denecke's lab (<http://www.plants.leeds.ac.uk/jd/pdf/Agrobacterium%20infiltration.pdf>).

1. Grow *N. benthamiana* (tobacco) plants on soil until 4–6 weeks old (*see* **Note 21**).
2. Using the appropriate standard molecular biology techniques, clone the transcription factor(s) of interest and synthetic promoter(s) into appropriate vectors for transient expression in tobacco (*see* **Note 22**).
3. Transform the vectors into a suitable strain of *A. tumefaciens* (*see* **Note 13**) following the procedure outlined in **Subheading 3.3, step 2**.
4. Culture the transformed *A. tumefaciens* overnight in 3–5 mL of LB medium with appropriate antibiotics (*see* **Note 23**).
5. Centrifuge 1 mL of the cultures at $3000 \times g$ for 5 min at room temperature. Discard the supernatant.
6. Resuspend in 1 mL of infiltration buffer and repeat **step 5**. Repeat once.
7. Measure the OD₆₀₀ of the *A. tumefaciens* and then adjust to an approximate OD₆₀₀ of 0.1 using infiltration buffer (*see* **Note 24**).
8. Move the plants to the lab and select leaves for infiltration. Leaves that are large, but not the oldest on the plant, should be used (*see* **Note 25**).
9. Use a P200 pipette tip to make a small hole in the leaf where each infiltration site will be. We typically do four infiltrations of the same construct on a single leaf. Use a permanent marker to draw a circle, approximately 30 mm in diameter, around each hole. Label the leaf with the constructs with which it will be infiltrated.
10. Mix the cultures of the two corresponding transformed *A. tumefaciens* strains, one with the TF containing strain and the other containing the synthetic promoter, at a 1:1 ratio.
11. Draw up 1 mL of the combined bacterial culture into a 1 mL syringe (without tip). Place a finger over the hole on the upper side of the leaf, then press the syringe into the hole from the bottom side of the leaf and inject bacterial culture into the leaf. This should produce a region of discoloration radiating out from the hole—continue with injection until the discoloration fills the circle drawn on the leaf. This will require approximately 250 μ L of bacterial culture per site. Repeat for the other three sites on the leaf.
12. After all infiltrations have been completed, return the plants to normal growth conditions for 2 days.
13. Measure luciferase activity in the infiltration sites, following the procedure described in **steps 2–9** of **Subheading 3.5** (*see* **Note 26**).

3.5 Imaging and Analysis

1. Plate sterilized *Arabidopsis* seeds (T_3 or subsequent generations) on MS media (*see Note 27*). After a 4 day stratification period at 4 °C, place plates in a growth chamber set to standard conditions suitable for your specific needs. For manageable number of genotypes and treatments, it is advisable to use each plate as an experimental block (i.e. replicate all genotypes and treatments on one plate).
2. Prior to imaging, spray plants with approximately 500 μ L of 1 mM luciferin solution per plate. Carefully spray plants evenly and thoroughly (*see Note 28*).
3. Move plants from growth chamber to a location near the CCD camera imaging box (*see Note 29*). Turn on CCD camera and allow it to cool down to working temperature.
4. Use an old plate of plants to focus the camera. Do this by placing the plate into the camera system, leaving the door of the light-tight imaging box open, and running the camera on a real-time imaging setting. Adjust the focus by turning the focusing ring on the camera lens (*see Note 30*).
5. Remove the plate from the camera, shut the door, and take a background reading (*see Note 31*).
6. Place the experimental plates in the imaging box and image the plants for the desired length of time. Save the resulting file.
7. Use appropriate software to measure bioluminescence (*see Note 32*). In the image processing software, place regions of interest (ROI) over the tissue—these will calculate the average light intensity for the pixels within the region. If interested in local responses to stimuli, place one ROI on the treated tissue (Fig. 3b). If interested in systemic responses to stimuli, multiple ROI may be placed on tissues distal from the treated area, and then averaged together to give one systemic LUC expression value for that plant (Fig. 3c).
8. Reformat the data output from the image analysis software into a layout amenable for visualization and statistical analysis (*see Note 33*).
9. Perform appropriate statistical analyses (*see Note 34*).

4 Notes

1. LB may be prepared in advance of transformation and stored at room temperature. However, to ensure sterility, we recommend making a dedicated batch of LB for the transformation.
2. For *A. thaliana*, we recommend using the promoter retrieval tool from the lab of Matthew Hudson (University of Illinois at Urbana-Champaign): <http://stan.crops.ci.uiuc.edu/prom.php>.

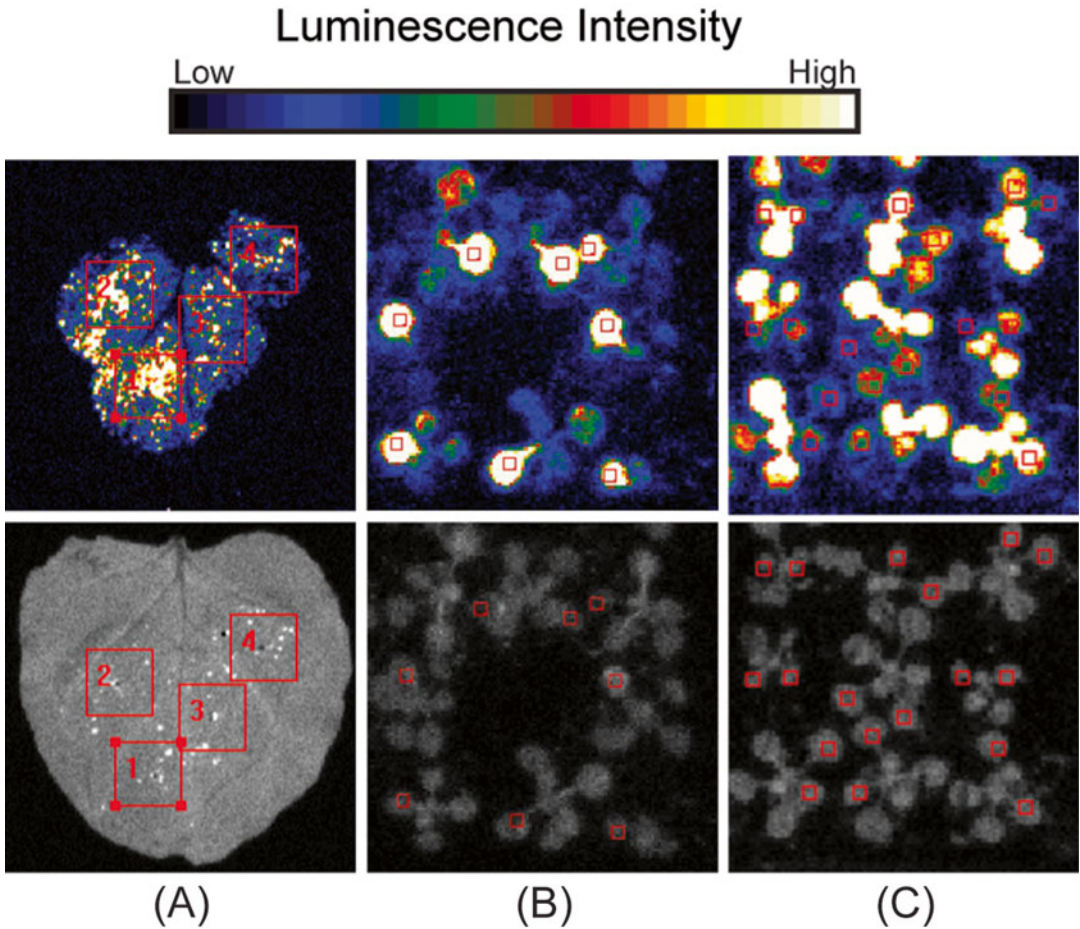


Fig. 3 Selection of regions of interest (ROI) for quantification of luciferase activity is dependent on experimental design. *Top panel* shows images of leaves with luciferase activity, as indicated by the scale above the images. *Bottom panel* shows the same leaves autofluorescing. Each *red box* is a single ROI placed on the image using the Andor Solis software. The software reports an average light intensity for all pixels within each ROI. Images are from experiments described in Benn et al. [17]. (a) ROI for *N. benthamiana* transient assay demonstrating activation of *4xRSRE:LUC* by the transcriptional activator CAMTA3. (b) ROI for measurement of local *4xRSRE:LUC* activity in wounded leaves of *A. thaliana* displaying local response. (c) ROI for measurement of systemic *4xRSRE:LUC* activity in *A. thaliana* plants where treatment of wounded leaves with flg22 has generated a systemic response

This tool retrieves the 2000 base pairs (bp) sequence immediately 5' to the translational start site. Alternatively, sequences may be retrieved from TAIR at <https://www.arabidopsis.org/tools/bulk/sequences/index.jsp>, however these sequences exclude the 5' UTR [24].

3. We used the motif discovery tool from the lab of Matthew Hudson (University of Illinois at Urbana-Champaign): <http://stan.cropsci.uiuc.edu/cgi-bin/sift/sift.cgi> [8]. A variety of other tools exists for motif discovery, such as DREME

(<http://meme.nbcr.net/meme/tools/dreme>) and the TAIR motif discovery tool (<http://www.arabidopsis.org/tools/bulk/motiffinder/index.jsp>) [24, 25]. We recommend running the promoter set of interest through several motif discovery algorithms, to ensure that motifs selected for further analysis are not artifacts of a particular algorithm.

4. To assess whether identified motifs match characterized elements, we recommend using motif discovery with DREME (*see* **Note 3**) followed by application of the TOMTOM (<http://meme-suite.org/tools/tomtom>) tool to search for top elements in the JASPAR database (<http://jaspar.genereg.net/>) [26, 27]. Candidate elements may also be searched against the AGRIS database (<http://arabidopsis.med.ohio-state.edu/AtcisDB/bindingsites.html>) [28].
5. For large number of promoters (>50), we recommend using a script to extract the flanking sequences. An example of a script with this functionality (among others), *cisfinder.v2.pl*, is available at our lab's github page (<https://github.com/DeheshLab/Cis-Element-Tools>).
6. In the case of the *4xRSRE* containing synthetic promoter, 4 base-pairs of flanking sequence was used on either side of the core motif (Fig. 2).
7. In the *4xRSRE* containing synthetic promoter, we used restriction enzyme recognition sites as spacers, specifically EcoRV, BamHI, and XbaI (Fig. 2).
8. For the mutated *4xRSRE* promoter, we altered three base-pairs in the core motif (Fig. 2).
9. In the case of the *RSRE*, a literature search revealed that a similar element, *vCGCGb*, had been identified via an oligo selection experiment as the binding site for a TF known as *CAMTA3* [29]. Furthermore, a specific instance of the *vCGCGb* element, *CCGCGT*, had been shown to be required for *CAMTA3* activation of the *CBF2* gene [30]. Both of these findings strongly suggested *CAMTA3* as a likely binding partner for the *RSRE*.
10. The yeast 1-hybrid approach may be hindered if the synthetic promoter is similar to yeast *cis*-regulatory elements, leading to auto-activation of the synthetic promoter. This was the case with the *4xRSRE* promoter, which was strongly auto-activating in yeast (unpublished data).
11. The forward genetic screen approach was successfully used with the *4xRSRE* promoter, allowing the identification of *CAMTA3* and *MEKK1* as regulators of the element [18].
12. We used pATM-NOS as our cloning vector for stable transformation of Arabidopsis [11]. This vector contains the LUC+ reporter gene, an improved version of firefly luciferase.

13. We used the GV3101 strain of *A. tumefaciens* for both stable transformation of Arabidopsis and transient transformation of *N. benthamiana*.
14. Plants used for the transformation assay are grown in Sunshine mix soil in growth chambers set to 22 °C and 16 h light.
15. It is critical that all aspects of the **step 9** procedure are performed sterilely. The flask, toothpicks, and pipette tips should be autoclaved prior to use and all work should be performed in a sterile hood. However, once the culture has reached the desired OD, all subsequent steps need not be done sterilely.
16. If the OD₆₀₀ of the culture is such that the resuspension volume will exceed that of the centrifuge bottle, you may spin down only part of the culture. Resuspension may be accomplished via vortexing in a small volume, though we find that using a P1000 pipette is sometimes necessary. However, pipetting should be limited to reduce damage to the cells from shearing.
17. The container used for plant transformation should be wide enough to allow dipping of the plants into the solution. We use shallow Tupperware containers. For smaller pots, pipette tip boxes may be used.
18. 3 s of swirling the plants in the culture of transformed *A. tumefaciens* works well in our lab. This exposure time may be varied as needed, however less time may result in fewer transformants, whereas longer times may produce multiple insertions. Patting plants with paper towels will remove excess bacterial culture on the stems and leaves. Flowers should still be visibly wet after this step.
19. Use of an opaque planting tray as a lid is to maintain high humidity but reduce buildup of heat, which in turn can reduce transformation efficiency.
20. If *A. tumefaciens* containing the pATM-NOS vector is used for transformation, we recommend the selection method of Harrison et al. [31]. Sterilize and plate seeds on ½ strength MS media with 50 µg/mL kanamycin. High plating density may reduce the efficiency of selection. Seeds should be evenly distributed and planted at a density of ~100 seeds per plate (100 mm × 100 mm). Stratify at 4 °C for 2 days, expose to light for 6 h, then place the plates in dark for 2 days. Following the 2 day dark treatment, move plates to light for 48 h, after which transformants will display increased growth and almost normal greening, while non-transformants will be smaller and chlorotic.
21. We grow *N. benthamiana* on Sunshine mix soil under light and humidity conditions similar to those used for Arabidopsis in our lab. The optimal developmental stage for transient transformation is just post flowering time.

22. For the transcription factors, we use the pYL436 vector and for the synthetic promoters, we use the pBGWL7 vector [32, 33].
23. Depending on the construct and strain of *A. tumefaciens* used, time to reach the required OD may vary from 18 to 48 h.
24. The ideal OD₆₀₀ used in transformation depends upon the particular construct being used. In the case of *4xRSRE:LUC* construct, which expresses well and is highly responsive to environmental perturbations, an OD₆₀₀ of 0.1 works well. In the case of a less responsive element in a construct that is also expressing poorly, OD₆₀₀ values from 0.2 to 0.5 may be employed.
25. We use up to four leaves from a single plant for transient assays. We recommend using different stages of leaf maturity as a controlled experimental factor—i.e. if the older pair of leaves is used for a particular construct on the first repeat, use younger leaves for the same construct on the second repeat.
26. Leaves are detached prior to imaging. We typically set a single, large, region of interest (ROI) to cover each infiltration site for image quantification (Fig. 3a).
27. We find that conducting experiments on MS media (as opposed to soil) produces more consistent results, likely due to the stress-inducible nature of many of the promoter:*LUC* fusion constructs used in our lab. Seeds should be plated in a sterile hood on MS media made to the standard specifications of the lab (typically ¼ or ½ strength MS with 0.8–1.0% phytoagar).
28. Plants may be sprayed up to 18 h in advance of imaging. For constructs predicted to be wound or touch-inducible, plants should be sprayed the day prior to imaging, so that any spray-mediated induction of the reporter luciferin returns to basal levels prior to the initiation of the experiment.
29. If the synthetic promoters are expected to be highly responsive to environmental conditions, plants should be moved close to the CCD camera setup in advance of the experiment. We find that moving *4xRSRE:LUC* plants 4 h prior to the start of imaging is sufficient.
30. For our system, focusing is done with the following settings: acquisition mode = real time, readout mode = imaging, readout time per pixel (µs) = 1, shutter time (s) = 0.3, external shutter = fully auto, data type = counts.
31. The optimal exposure time must be empirically determined for each construct. The exposure should be at the shortest time that still allows for clear detection of the luciferase signal. In our lab, the *4xRSRE:LUC* activity is imaged using a 5' exposure, while the less active *pHYDROPEROXIDELYASE::LUC* construct requires a 15' exposure. For the *4xRSRE:LUC* in our system, the background reading is taken with the following settings (for

an 8 h run): acquisition mode=kinetic series, readout mode=imaging, readout time per pixel (μs)= $32 \times 2 \times \text{Gain}$, shutter time (s)=300, number of accumulations=1, number in kinetic series=48, kinetic cycle time (s)=600, external shutter=fully auto, data type=counts (bg corrected).

32. To measure bioluminescence in the Andor Solis software, it is first necessary to change the default false coloring settings to allow viewing of the plants. This is done by opening the data histogram tool, followed by changing the mode to “range”, and finally altering the range until plants are visible (we use a range of 0–200 for *4xRSRE:LUC*). Toggling through the available color palettes using the change palette tool enables selection of most suitable color palette for the clearest visualization of the plants. Once the plants are clearly visible, use the ROI tool to place an ROI on the image by clicking the “Add ROI” button. ROI can then be dragged over the tissue to be measured and appropriately resized by clicking and dragging the corners of the ROI box. Repeat this process until ROI have been placed for all plant leaves in a particular experimental unit (i.e. wild-type, untreated) (Fig. 3). Then select all of the data in the ROI window and copy it into a spreadsheet utility, such as Microsoft Excel.
33. Reformatting of the data may be done by manually copying out and pasting the relevant data (i.e. mean luminescence at each time point) from the output of the image analysis software. If many experiments are to be performed, this process can become tedious, and prone to error. As such, we recommend using a processing script to convert the raw data into a form suitable for analysis. We have developed scripts for conversion of the raw Andor Solis ROI data into a format suitable for analysis in R. These scripts (`Andor_parseR.pl` and `Andor_parseR.systemic.pl`) are available at <https://github.com/DeheshLab/CCD-camera-data-analysis>.
34. Depending on the nature of the particular construct, the data may need to be log-transformed prior to analysis. For example, the *4xRSRE:LUC* reporter produces data that can range over several orders of magnitude, necessitating log-transformation to achieve normality and homogeneity of variance. Alternatively, nonparametric statistical tests may be used.

Acknowledgements

We would like to thank Marta Bjornson for her helpful comments on the paper. This work was supported by National Institute of Health (R01GM107311), National Science Foundation (IOS-1036491 and IOS-1352478), and Agriculture experiment station (CA-D-PLB-3510-H) grants awarded to K.D.

References

1. Arditti RR, Scaife JG, Beckwith JR (1968) The nature of mutants in the lac promoter region. *J Mol Biol* 38(3):421–426
2. Timko MP, Kausch AP, Castresana C, Fassler J, Herrera-Estrella L, Van den Broeck G, Van Montagu M, Schell J, Cashmore AR (1985) Light regulation of plant gene expression by an upstream enhancer-like element. *Nature* 318(6046):579–582
3. Ha SB, An G (1988) Identification of upstream regulatory elements involved in the developmental expression of the Arabidopsis thaliana *cab1* gene. *Proc Natl Acad Sci U S A* 85(21):8017–8021
4. Bustos MM, Guiltinan MJ, Jordano J, Begum D, Kalkan FA, Hall TC (1989) Regulation of beta-glucuronidase expression in transgenic tobacco plants by an A/T-rich, cis-acting sequence found upstream of a French bean beta-phaseolin gene. *Plant Cell* 1(9):839–853. doi:10.1105/tpc.1.9.839
5. Siebertz B, Logemann J, Willmitzer L, Schell J (1989) cis-analysis of the wound-inducible promoter *wun1* in transgenic tobacco plants and histochemical localization of its expression. *Plant Cell* 1(10):961–968. doi:10.1105/tpc.1.10.961
6. Bruce WB, Deng XW, Quail PH (1991) A negatively acting DNA sequence element mediates phytochrome-directed repression of *phyA* gene transcription. *EMBO J* 10(10):3015–3024
7. Straub PF, Shen Q, Ho TD (1994) Structure and promoter analysis of an ABA- and stress-regulated barley gene, *HVA1*. *Plant Mol Biol* 26(2):617–630
8. Hudson ME, Quail PH (2003) Identification of promoter motifs involved in the network of phytochrome A-regulated gene expression by combined analysis of genomic sequence and microarray data. *Plant Physiol* 133(4):1605–1616. doi:10.1104/pp.103.030437
9. Harmer SL, Hogenesch JB, Straume M, Chang HS, Han B, Zhu T, Wang X, Kreps JA, Kay SA (2000) Orchestrated transcription of key pathways in Arabidopsis by the circadian clock. *Science* 290(5499):2110–2113
10. Zou C, Sun K, Mackaluso JD, Seddon AE, Jin R, Thomashow MF, Shiu SH (2011) Cis-regulatory code of stress-responsive transcription in Arabidopsis thaliana. *Proc Natl Acad Sci U S A* 108(36):14992–14997. doi:10.1073/pnas.1103202108
11. Walley JW, Coughlan S, Hudson ME, Covington MF, Kaspi R, Banu G, Harmer SL, Dehesh K (2007) Mechanical stress induces biotic and abiotic stress responses via a novel cis-element. *PLoS Genet* 3(10):1800–1812. doi:10.1371/journal.pgen.0030172
12. Rushton PJ, Reinstadler A, Lipka V, Lippok B, Somssich IE (2002) Synthetic plant promoters containing defined regulatory elements provide novel insights into pathogen- and wound-induced signaling. *Plant Cell* 14(4):749–762
13. Whalley HJ, Sargeant AW, Steele JF, Lacoere T, Lamb R, Saunders NJ, Knight H, Knight MR (2011) Transcriptome analysis reveals calcium regulation of specific promoter motifs in Arabidopsis. *Plant Cell* 23(11):4079–4095. doi:10.1105/tpc.111.090480
14. Koschmann J, Machens F, Becker M, Niemeyer J, Schulze J, Bulow L, Stahl DJ, Hehl R (2012) Integration of bioinformatics and synthetic promoters leads to the discovery of novel elicitor-responsive cis-regulatory sequences in Arabidopsis. *Plant Physiol* 160(1):178–191. doi:10.1104/pp.112.198259
15. Millar AJ, Short SR, Chua NH, Kay SA (1992) A novel circadian phenotype based on firefly luciferase expression in transgenic plants. *Plant Cell* 4(9):1075–1087. doi:10.1105/tpc.4.9.1075
16. Harmer SL, Kay SA (2005) Positive and negative factors confer phase-specific circadian regulation of transcription in Arabidopsis. *Plant Cell* 17(7):1926–1940. doi:10.1105/tpc.105.033035
17. Benn G, Wang CQ, Hicks DR, Stein J, Guthrie C, Dehesh K (2014) A key general stress response motif is regulated non-uniformly by CAMTA transcription factors. *Plant J* 80(1):82–92. doi:10.1111/tbj.12620
18. Bjornson M, Benn G, Song X, Comai L, Franz AK, Dandekar AM, Drakakaki G, Dehesh K (2014) Distinct roles for mitogen-activated protein kinase signaling and CALMODULIN-BINDING TRANSCRIPTIONAL ACTIVATOR3 in regulating the peak time and amplitude of the plant general stress response. *Plant Physiol* 166(2):988–996. doi:10.1104/pp.114.245944
19. Crooks GE, Hon G, Chandonia JM, Brenner SE (2004) WebLogo: a sequence logo generator. *Genome Res* 14(6):1188–1190. doi:10.1101/gr.849004
20. Gaudinier A, Zhang L, Reece-Hoyes JS, Taylor-Teeple M, Pu L, Liu Z, Breton G, Pruneda-Paz JL, Kim D, Kay SA, Walhout AJ, Ware D, Brady SM (2011) Enhanced Y1H assays for Arabidopsis. *Nat Methods* 8(12):1053–1055. doi:10.1038/nmeth.1750
21. Vinson CR, LaMarco KL, Johnson PF, Landschulz WH, McKnight SL (1988) In situ

- detection of sequence-specific DNA binding activity specified by a recombinant bacteriophage. *Genes Dev* 2(7):801–806
22. Dehesh K, Bruce WB, Quail PH (1990) A trans-acting factor that binds to a GT-motif in a phytochrome gene promoter. *Science* 250(4986):1397–1399
 23. Clough SJ, Bent AF (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J* 16(6):735–743
 24. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E (2012) The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* 40(Database issue):D1202–D1210. doi:[10.1093/nar/gkr1090](https://doi.org/10.1093/nar/gkr1090)
 25. Bailey TL (2011) DREME: motif discovery in transcription factor ChIP-seq data. *Bioinformatics* 27(12):1653–1659. doi:[10.1093/bioinformatics/btr261](https://doi.org/10.1093/bioinformatics/btr261)
 26. Gupta S, Stamatoyannopoulos JA, Bailey TL, Noble WS (2007) Quantifying similarity between motifs. *Genome Biol* 8(2):R24. doi:[10.1186/gb-2007-8-2-r24](https://doi.org/10.1186/gb-2007-8-2-r24)
 27. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B, Sandelin A, Wasserman WW (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* 42(Database issue):D142–D147. doi:[10.1093/nar/gkt997](https://doi.org/10.1093/nar/gkt997)
 28. Yilmaz A, Mejia-Guerra MK, Kurz K, Liang X, Welch L, Grotewold E (2011) AGRIS: the *Arabidopsis* Gene Regulatory Information Server, an update. *Nucleic Acids Res* 39(Database issue):D1118–D1122. doi:[10.1093/nar/gkq1120](https://doi.org/10.1093/nar/gkq1120)
 29. Yang T, Poovaiah BW (2002) A calmodulin-binding/CGCG box DNA-binding protein family involved in multiple signaling pathways in plants. *J Biol Chem* 277(47):45049–45058. doi:[10.1074/jbc.M207941200](https://doi.org/10.1074/jbc.M207941200)
 30. Doherty CJ, Van Buskirk HA, Myers SJ, Thomashow MF (2009) Roles for *Arabidopsis* CAMTA transcription factors in cold-regulated gene expression and freezing tolerance. *Plant Cell* 21(3):972–984. doi:[10.1105/tpc.108.063958](https://doi.org/10.1105/tpc.108.063958)
 31. Harrison SJ, Mott EK, Parsley K, Aspinall S, Gray JC, Cottage A (2006) A rapid and robust method of identifying transformed *Arabidopsis thaliana* seedlings following floral dip transformation. *Plant Methods* 2:19. doi:[10.1186/1746-4811-2-19](https://doi.org/10.1186/1746-4811-2-19)
 32. Rubio V, Shen Y, Saijo Y, Liu Y, Gusmaroli G, Dinesh-Kumar SP, Deng XW (2005) An alternative tandem affinity purification strategy applied to *Arabidopsis* protein complex isolation. *Plant J* 41(5):767–778. doi:[10.1111/j.1365-313X.2004.02328.x](https://doi.org/10.1111/j.1365-313X.2004.02328.x)
 33. Karimi M, De Meyer B, Hilson P (2005) Modular cloning in plant cells. *Trends Plant Sci* 10(3):103–105. doi:[10.1016/j.tplants.2005.01.008](https://doi.org/10.1016/j.tplants.2005.01.008)

The Identification of *Cis*-Regulatory Sequence Motifs in Gene Promoters Based on SNP Information

Paula Korkuć and Dirk Walther

Abstract

Conservation of particular molecular sequence motifs throughout evolution is a strong indicator of their functional relevance as selective pressure likely prevented the accumulation of mutations. Known as “phylogenetic footprinting”, this rationale has been exploited for the identification of novel functional motifs using sequence information from sequence alignments of diverse species, in particular transcription factor binding site motifs in aligned gene promoter sequences of orthologous genes. With the rapid advances of sequencing technologies, whole genome sequence information is accumulating not only across different species, but increasingly for variants of the same species exhibiting relatively little sequence variability, primarily present as single nucleotide polymorphisms (SNPs). Here, we lay out the basic strategy for the identification of functional *cis*-regulatory motifs in gene promoter regions based on SNP information.

Key words Phylogenetic footprinting, Transcription factor binding sites, Single nucleotide polymorphism, Gene promoter, Conservation, Gene expression

Abbreviations

TFBS Transcription factor binding site
TSS Transcription start site
SNP Single nucleotide polymorphism

1 Introduction

Sequence conservation in evolution is a powerful indicator of functional relevance of the respective molecule or sequence motif embedded in longer molecules. As conservation across evolutionarily long time frames requires selection against the inevitable accumulation of random mutations, the very property of being conserved alone implies that the respective region or molecule encodes or fulfills a function that proved beneficial. Hence, the identification of evolutionarily conserved sequences can serve as a

powerful discovery tool to identify novel functional molecules or sequence motifs. This rationale has been exploited for the search for novel transcription factor binding site motifs (TFBSs) in gene promoter regions. Assuming that genes found to be conserved in different species (so-called orthologous genes) are in turn regulated by conserved transcription factors that bind to similar (or even identical) binding sites in the gene promoter regions of the orthologous genes, TFBSs can be identified as those sequence motifs that remain unchanged in evolution. This concept was introduced as “phylogenetic footprinting” [1] and has seen increased application with increasing availability of genomic sequence information [2–7]. In order to be discernable as conserved, the considered species ideally have diverged sufficiently in evolution such that conserved motifs stand out in the context of otherwise nonfunctional, and hence randomly mutating, and therefore, variable surrounding sequence regions of gene promoters. However, the sequences must not have diverged beyond the limits of establishing orthology relationships with reasonable confidence. In the case of highly similar sequences or genomes, signal-to-noise (conservation vs. variation) ratios can be increased by adding additional genome sequences from different species. Even though mutation densities derived from pairwise comparisons among all considered related species may be low, when combined vertically across a multiple sequence alignment of all available sequences, the effective mutation density can be increased substantially by mapping onto a common reference sequence. The term “phylogenetic shadowing” has been coined for this approach and was first applied to the identification of functional regions in primate genomes [8] and has since found numerous applications, e.g. in plants [9]. A detailed protocol describing phylogenetic shadowing is available in [10].

In the classical phylogenetic footprinting approach, functional motifs “reveal themselves” by their increased conservation level and motif sequences—including motif lengths—can be derived accordingly from the respective sequence. Alternatively, a set of candidate motifs postulated prior to motif discovery scans can be queried for evidence of conservation across whole genome alignments or alignments of multiple orthologous genomic regions. Developed for the identification of functional motifs in yeast species, this approach yielded numerous novel regulatory motifs by probing a large set of candidate “mini motifs” and merging them into full motifs [11].

Evidently, with shorter evolutionary separation and correspondingly reduced sequence divergence between species and their associated genome sequences, motif identification via evidence of conservation footprints becomes increasingly challenging as passive conservation (simply “no time” for any mutation to occur) versus active conservation (mutations that did occur have been selected out) cannot be easily distinguished. The situation is particularly challenging in the case of sequence variants associated

with one and the same species, where individuals differ by single nucleotide polymorphisms (SNPs) or short insertions and deletions sparsely scattered across the genomic sequence. While, on the one hand, establishing orthology relationships—ensuring that a particular sequence segment serves the same function in different genomes—is facilitated and more likely correct than in the case of very divergent species, conservation-based motif discovery seems almost futile. The available SNP-based sequence variability may be too low to permit motif discovery at sufficient resolution. This challenge can either be met by a massive depth of available sequence information with lots of individual sequences available such that the effective SNP-density is increased as pursued in the phylogenetic shadowing approach or motifs are searched for not only vertically across an alignment but identified via an aggregated conservation measure of all occurrences of postulated candidate motifs.

Here, we lay out a protocol for the discovery of novel cis-regulatory sequence motifs in gene promoter regions based on SNP information following in its basic methodology the example set in [12] for the plant species *Arabidopsis thaliana* (see **Note 1**). The rationale to identify novel TFBSs by searching for sequence motifs that exhibit low variability across many individual sequence variants associated with one species found support by the observation that known TFBSs in *A. thaliana* were found to indeed show significantly lower variability as judged by SNP-density across hundreds of *Arabidopsis* accessions than promoter regions not annotated to be part of TFBSs [12].

With the rapidly growing sequencing capacities targeting not only diverse species, but increasingly sequence variations within species, strategies for the exploitation of the generated sequence information for the purpose of motif discovery can be expected to become increasingly relevant. This protocol for the identification of novel cis-regulatory motifs in gene promoter regions can also be applied toward the discovery of other genomic elements such as enhancers, silencers, and others. However, it is to be expected that transcriptional regulation is modified in response to changing environments. Motifs and associated transcription factors binding to them that have emerged in subgroups exposed to similar environments of accessions/species will likely be missed using the presented approach. Evidently, allowing for sequence variation runs counter to the rationale of identifying motifs using a logic that rests on conservation. To identify such motifs, prior clustering of accessions/species would be necessary and the analysis performed on the subgroups. Here, the remaining sequence variability within the individual clusters will determine as to whether conservation-based approaches can be pursued. Alternatively, we might ask whether the very property of adaptation may be exploited. For protein coding sequences, increased frequencies of

amino acid changing mutations relative to random expectation is frequently taken as evidence for positive selection. Concepts abbreviated as K_a/K_s ratios, where K_a denotes the rate of non-synonymous and K_s the rate of synonymous mutations, have been pursued to identify proteins whose accelerated non-neutral mutation rates can be interpreted as a selection-favored adaptation to changing environments. However, for this approach to work, the proteins as a whole still have to be recognizable, i.e. sufficiently similar in sequence to be identified as orthologs. Therefore, the direct transfer of approaches that rely on sequence conservation—as laid out here—to the identification of subgroup-specific TFBSs remains challenging.

Please note that this protocol assumes the interested reader to possess sufficient programming skills and ideally be experienced in the statistical programming language R to be able to interpret and implement it. Where appropriate, we refer to existing software solutions, R-routines, and our own software solutions made accessible for general use.

2 Materials

1. As the main prerequisite of SNP-based motif detection, whole genome sequence information needs to be available for a given species across as many individuals as possible. Depending on the subject species and associated genetics, individual genomes may also be referred to as accessions/ecotypes as in the case of selfing plants or as strains in the case of bacteria and fungi. Throughout this protocol, we shall use the term “individual sequences”. Typically, relevant genome sequence information is most easily available from the web portals established for the various species that are of scientific or economic interest such as for the plant *Arabidopsis thaliana* (<http://1001genomes.org/> and The Arabidopsis Information Resource (TAIR, [13])) or The International Rice Informatics Consortium for rice (<http://iric.irri.org/>).
2. True sequence variants need to be identified. As all sequencing technologies generate a certain percentage of base call errors, the identification of true sequence variants remains challenging. Programs such as VarScan have been developed to call SNPs from deep sequencing data [14]. A review on strategies and available resources and software solutions for SNP-calling focusing on plant genomes can be found in [15]. However, at the time of deposition into the respective databases, sequences may be taken “as is” with the understanding that a certain percentage of sequence variants may be false. As a measure to safeguard against sequencing errors wrongly interpreting as a polymorphism,

a minimum number of different individual sequences can be required to also show the particular variant base at a given position. This requirement may be introduced as a minimum minor allele frequency threshold, e.g. 5% (*see* **Note 2**).

3. Genome sequence annotation must have been performed and be available with precise start–end positions for all genes. This allows for the identification of gene promoter regions. Genome annotation information is made available in gff-file format (General Feature Format) or can be obtained from dedicated, species-specific resources such as TAIR for *Arabidopsis thaliana* [13] or the Saccharomyces Genome Database (SGD) for yeast [16].
4. If available, gene functional annotation such as captured by gene ontology (GO) terms should be obtained for use in guiding motif consensus formation (*see* Subheading 3.4, **step 2**). Again, species-focused databases may provide the required information.
5. For the purpose of candidate motif validation, expression profiling data sets should be acquired. Ideally, gene expression data sets for the species under study exposed to a large number of different conditions or perturbations and measured using the same expression platform (gene chip or RNAseq using the same sequencing technology) and for as many genes as possible are available. Sources of relevant expression information are the Gene Expression Omnibus (GEO) [17], ArrayExpress [18], or species-specific databases such as NASCArrays for *A. thaliana* [19]. All available expression profiles may have to be transformed (log-transformed) to render the distributions normal and need to be normalized jointly. Here, quantile normalization as implemented as the function `normalize.quantile` from the `preprocessCore` R package can be used (*see* **Note 3**).
6. Information on known TFBS motifs (or the respective alternative motif type of interest) in the given species should be located. Dedicated databases have been established that manage such information such as [20–22]. Increasingly, motif sets are available from large-scale protein binding assays and available from the supplementary material or established databases associated with the respective publication article [23].

3 Methods

Naïvely, it should be possible to identify motifs based on sequence conservation simply by inspecting short sequence intervals and determining their degree of variability across all available and aligned individual sequences (Fig. 1). Candidate motifs determined

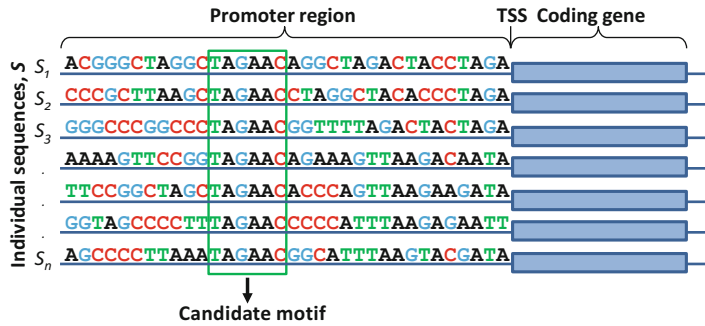


Fig. 1 Basic principle of the phylogenetic footprinting approach for the discovery of functional motifs. Based on an alignment of divergent genomic sequences, sequence portions found to be conserved across the individual upstream gene promoter sequences are considered candidate cis-regulatory motifs. In this hypothetical example, the accumulated sequence divergence is large enough to permit motif discovery by simply searching for consecutive segments of vertically conserved sequence positions. As in the case of SNP-based motif discovery, sequence variation is much less pronounced, alternative approaches need to be pursued (see Fig. 4 for a schematic illustration). At the same time, however, the reliability of the alignments is increased and promoter sequences can be easily aligned by lining up the transcription start site (TSS) positions of the genes found in the different individual genome sequences. In classical phylogenetic footprinting, where different species are compared, the genes must correspond to orthologs. In the case of single species, i.e. SNP-based motif discovery, vertically aligned genomic positions are almost guaranteed to result in correctly matched genes that can be taken as allelic variants of one another

as those with low variability then need to be checked for novelty and the potentially novel motifs validated further. However, given the typical SNP-density, even in the case of hundreds of different individual genotype sequences are available, this naïve approach would result in nearly all sequence stretches to be identified as fully conserved. For example, despite having access to genomic sequences of 350 *Arabidopsis* accessions, the average spacing between two polymorphic sites in 500 bp upstream gene promoter regions was estimated at 56 bp (with SNP positions defined as sites with minor allele frequency above 5%). Thus, the direct vertical screen for sequence conservation—as done in phylogenetic footprinting or shadowing alike—cannot be pursued. Instead, the available variability information needs to be aggregated and motif conservation be assessed from all occurrences of a motif in the genome. This, in effect, increases the SNP-density, but also necessitates the decision on defining candidate motifs beforehand and subsequently probing them for SNP-density across all their genomic locations [11]. Especially the length of candidate motifs needs to be decided upon. Then, the set of candidate motifs of length k , so called k -mers with each k -mer having a different sequence, can be searched for in the genome and k -mer-specific variability statistic be gathered.

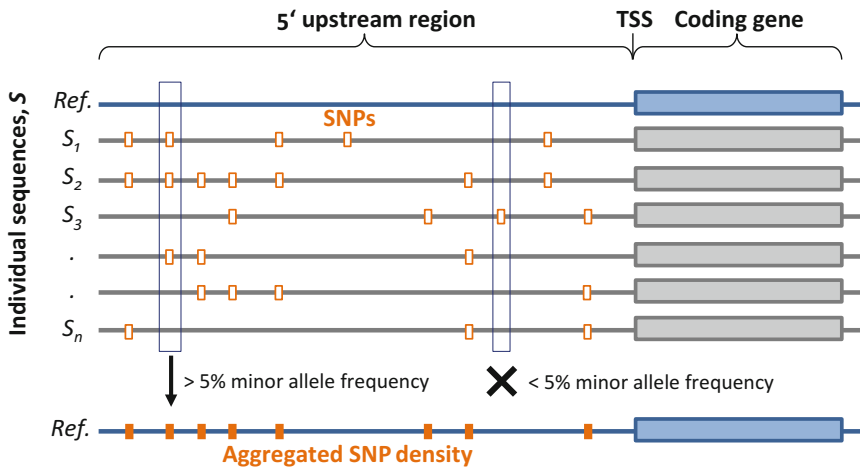


Fig. 2 Alignment of all available individual genomic sequences associated with individuals. As sequence divergence is low—in individual genome sequences of same species, differences exist as SNPs and short insertions/deletions only—alignments can be accomplished by mapping all SNP positions onto a common reference sequence based on sequence positions alone (see Note 9). A minor allele frequency threshold (e.g. 5%) may be introduced to use frequent variants only and to reduce the chance of sequencing errors (see Note 2)

The key phases for SNP-based motif discovery are: (1) Preparatory steps (sequence alignment and definition of relevant sequence interval, definition of individual candidate motifs), (2) Motif mapping (identification of all motif mappings and determination of their SNP-density), (3) Motif filtering (select candidate motifs based on SNP-density, positional profiles, comparison to known motifs), (4) Consensus motif generation from individual candidate motifs, and (5) Motif validation (GO-term enrichment analysis and gene expression statistic). The detailed protocol laid out below follows the published approach in [12] and includes the following steps.

3.1 Preparatory Steps

1. Produce an alignment of all individual genome sequences (Fig. 2). Here, the low sequence variability when using variant sequences from the same species proves helpful as no major rearrangements of whole chromosomal sequence regions are to be expected. Most of the variability will be present as single nucleotide polymorphisms (SNPs) that do not shift the alignment position of one sequence relative to another. Thus, the alignment step would be reduced to tabulating all changes at a given genomic position. However, as (oftentimes short) insertions/deletions do occur, the exact alignment of all sequences—the multiple sequence alignment—remains challenging. A practical solution is to express all considered sequences as variants of one common reference genome sequence. Typically, this reference is the first (historically), and therefore best characterized sequence in the species under study.

Thus, deletions relative to the reference in other individuals can be properly represented, while insertions would not be reflected correctly as this requires introducing gaps in the reference sequence and all other individual sequences not harboring this insertion. However, as this protocol exploits SNPs, the most frequent genomic sequence variant type, this less than ideal multiple sequence alignment has no severe detrimental consequences on the success of motif discovery other than underestimating the variability of motifs due to insertions.

2. The length of the considered gene promoter regions needs to be decided upon. The next upstream gene constitutes a reasonable upper limit of individual gene promoters—as gene regions can be assumed not to harbor *cis*-regulatory motifs. Therefore, all upstream regions should be clipped at the site of the next upstream gene. Note that this gene may lie on the opposite strand. Operationally, the considered length may be set arbitrarily to a reasonable threshold, such as several hundred nucleotides with the understanding that motifs further upstream will be missed. Evidently, this threshold depends on the gene density, and thus, the average intergenic sequence length of the species under study. A SNP-informed decision with regard to promoter length can also be reached by examining the SNP-density as a function of sequence distance from the TSS. As shown in Arabidopsis [12], beyond 500 bp the SNP-density remains constant while it is lower for the 500 bp immediately upstream of the transcription start site (shown schematically in Fig. 3). Hence, 500 bp appears a reasonable choice for promoter length in Arabidopsis as conservation—the hallmark of functional relevance—was detected for this interval.
3. Candidate motifs need to be defined. Most importantly, the length of candidate motifs has to be set unless specific user-defined motifs are to be probed. TFBSs are generally short (6–20 bp with a median of 9 bp for 144 known Arabidopsis motifs [12]). Therefore, assuming a motif length of 6 bp represents a reasonable compromise between motif specificity—the longer the motif, less likely random hits will occur—and number of different candidate motifs that need to be tested. The longer the motif, the more different motifs are possible—with every added position, the number of possible motifs increases by a factor of 4, hence resulting in fewer mapping occurrences per motif and associated weaker SNP-density statistic. Allowing all four base types at all six positions, $4^6 = 4096$ different k-mer sequences (“AAAAAA”, “AAAAAC”, ..., “TTTTTT”) are possible. However, as typically no distinction is made between motif hits in forward and reverse-complement orientation, the number of truly sequence-different k-mers collapses to only 2080 different

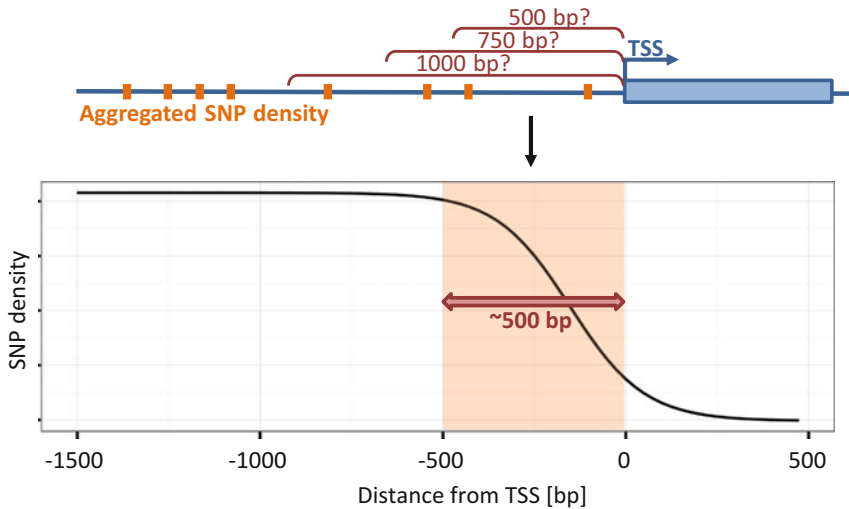


Fig. 3 SNP-frequency-informed definition of the effective promoter length. Taking sequence conservation in upstream regions as indication of promoter activity, the effective average promoter length in the species under study can be determined based on the available aggregated SNP-density in upstream sequence regions. Based on, for example, a logistic function fitted to the SNP-density (shown schematically here as the black line), the upstream interval considered to be the gene promoter can be set based on an appropriately chosen threshold value. The transcription start site (TSS), by definition, marks the 3'-terminus of this interval. Furthermore, upstream genes (possibly also on the opposite strand) can also be taken as effective 5'-terminal positions of individual gene promoters

hexamers comprising 2016 forward and reverse-complement hexamer pairs and 64 palindromic hexamers that are identical in forward- and reverse-complement orientation (*see Note 4*).

3.2 Motif Mapping

1. All 2080 unique hexamer motifs are to be mapped to all promoter sequences (Fig. 4a). For later use as a reference, mappings should also be generated to intergenic sequences further upstream of the region considered the promoter (*see Subheading 3.1, step 2*). As mentioned above, 2016 motifs need to be mapped in both forward- and reverse-complement orientation, but are considered as hits of the same motif. The remaining 64 motifs are palindromic and orientation-invariant. The actual mapping can be performed with the respective string-matching functions available in the programming language of choice (*see Note 5*).
2. SNP-density, SD , of candidate k -mers: Summed up over all nm mapping locations of hexamer candidate motif m , the number, Lm with $Lm=6*nm$, corresponds to the total length of all genomic positions covered by hexamer m . With Sm representing the number of polymorphic sequence positions within all Lm motif hit positions, the SNP-density of motif m computes as $SDm = Sm/Lm$.

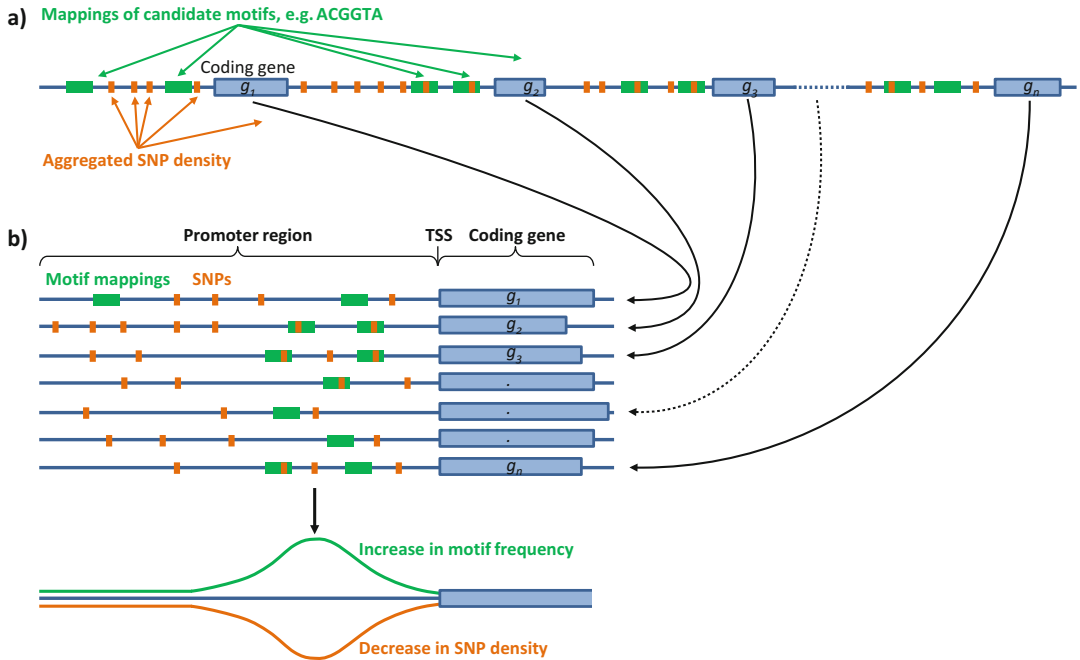


Fig. 4 (a) All candidate motifs are mapped onto the promoter regions of all genes ($g_1 \dots g_n$) positioned on the reference genomes and associated motif-specific SNP-densities determined across all its mapping instances using the aggregated SNP-information collated from all individual sequences (Fig. 1). (b) Based on the determined location of all mapping instances relative to the TSS of the respective downstream gene, motif occurrence density distributions (shown schematically by the *green* frequency distribution) can be computed and compared to the observed SNP-density (shown schematically by the *orange* frequency distribution). Anti-correlation of both frequency distributions—as shown here in an idealized example—serves as a filter criterion to enrich for true positive motif hits

3.3 Motif Filtering

1. Based on conservation: All candidate hexamer motifs could straight be sorted by SNP-density, SD , with the motifs with lowest SD representing promising candidates. However, depending on the actual sequence, random SNP-densities may differ and, hence, a better estimate of conservation would be the comparison of SNP-density in promoter sequences (e.g. 500 bp upstream regions), where motifs are considered to possibly act as TFBSs, relative to random background representing chance occurrences. For example, mappings could be produced to regions 501–1500 bp upstream of the TSS. Then, the sorting of all motifs m would be in ascending order of the ratio of $SD_{m,rel} = SD_{m,500} / SD_{m,501-1500}$, where the index range indicates the considered regional interval upstream of the TSS. At this point, all hexamer motifs are sorted in ascending order of variability and the top-x percent can be pursued further. What portion of motifs is considered further can be either set pragmatically (e.g. top-10%) or by performing shuffling experiments, in which actual upstream sequences are

shuffled repeatedly (>100 times), all mappings re-generated resulting in an estimate of the chance distribution of SDm_{rel} , by which the 5 %-tail could be selected to enrich for truly conserved motifs in promoter segments relative to background.

2. Based on evidence of corroborating positional preference: Assuming that indications of positional preferences of candidate motifs lend further confidence to candidate motifs [24], motif occurrence densities across promoter intervals can be computed. Either, indication of nonuniform distributions can be taken as evidence of preferred motif locations or, in addition, it is checked whether preferred positional intervals of a given candidate motif coincides with lowered SNP-density in this interval. Thus, for the latter, a negative correlation between SNP-density and occurrence-frequency profiles in upstream regions are to be expected and can be imposed as an efficient filter to enrich for true-positive motifs (Fig. 4b). Probing directly for uneven motif location distribution could be accomplished by computing a positional entropy measure, with the considered promoter length interval segmented into even sub-intervals and the occurrence frequency in the subintervals combined into an effective location entropy. Following the rationale of detecting anti-correlation of position and SNP-density profiles, motifs can be selected based on the resulting Pearson correlation coefficients with a threshold applied to the associated p -values properly adjusted for multiple testing as many candidate motifs (the ones passing Subheading 3.3, **step 1**) are tested. Position and SNP-density profiles need to be generated by applying a sliding window (either overlapping or nonoverlapping) to the upstream sequence segments defined as the promoter region. A reasonable window size needs to be selected based on occurrence frequencies ensuring sufficient events per window (this requirement will increase the window size) while at the same time also resulting in high positional resolution (this requirement calls for small window sizes). In addition, for a correlation measure to be reasonable, 5 or more data pairs need to be correlated. For example, assuming a promoter length of 500 bp, subdividing upstream regions into 25 bp nonoverlapping intervals would result in 20 values for both SNP-density and position-frequency. Typically, multiple testing correction is done by adopting the concept of Benjamini-Hochberg False Discovery Rate (FDR) [25] and is conveniently performed in R using the function `p.adjust` (see **Notes 6** and **7**).

3.4 Consensus Motif Generation

Evidently, true motifs are of different length and not limited to a length of 6 bp. Implementing the k-mer strategy essentially follows the logic of identifying motif seeds that are to be assembled into longer motifs based on sequence overlap.

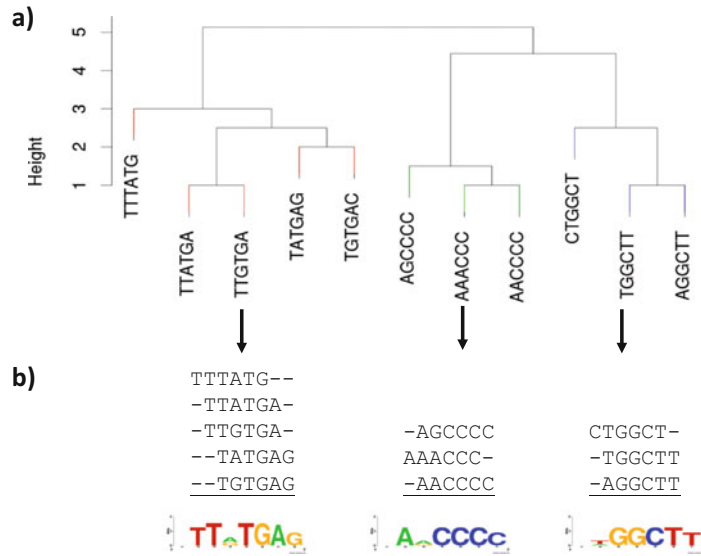


Fig. 5 Exemplary hierarchical clustering of candidate k-mer motifs passing all filter criteria based on Levenshtein distance (a) followed by consensus motif generation using ClustalW (b). The decision about partitioning motifs into distinct clusters can be guided by GO-term enrichment statistics (*see* Subheading 3.4, **step 2**) or can be decided based on established cluster number metrics such as Dunn's Index [26] or visually if the number of candidate motifs is small. Resulting consensus motifs are illustrated as sequence logos reflecting the position-specific nucleotide conservation

1. Candidate k-mers obtained after motif filtering are to be clustered, for example, based on their pairwise Levenshtein edit distance applying an exchange score (0 = match, 1 = mismatch) (Fig. 5a). A reasonable requirement would be to only allow motifs to be extended by a base without allowing any mismatches in central motif positions. Clustering of all motifs can be done applying hierarchical clustering or k-means methods followed by further inspections as the number of distinct clusters needs to be determined. This decision is ideally informed by functional annotation information (Subheading 3.4, **step 2**). Otherwise, for deciding on the correct cluster number based on the distance metric itself, several parameters have been developed in unsupervised machine learning such as Dunn's Index [26] and the reader is referred to [27] for further discussion of the issue. The routine stringDist from the R-package Biostrings can be used for motif clustering.
2. GO-term enrichment as evidence for involvement in similar processes. For every candidate hexamers, gene sets can be generated with one set containing all genes that harbor the given motif in their promoter regions, while the second set of genes does not. Then, applying commonly established GO-term enrichment analysis strategies, e.g. relying on a Fisher exact

test, GO-terms enriched in the candidate hexamer containing set relative to the control set can be determined. Subsequently, k-mer merging into larger consensus motifs can be guided by consistent GO term annotations associated with individual k-mer motifs. A Perl-script for the computation of GO-term enrichment analyses of two different gene sets as described in this paragraph is available at <http://bioinformatics.mpimp-golm.mpg.de/research-projects-publications/supplementary-data/walther>.

3. Once clustered together, the actual consensus motif can be created by generic multiple sequence alignment programs such as ClustalW [28] (Fig. 5b).
4. Already at the stage of individual candidate k-mers or at the stage of consensus motifs, motifs that are already known from previous studies need to be eliminated. Also similar or identical motifs reported in other species would be worth knowing about. Depending on the extent of prior research on cis-regulatory motifs in the species under study and associated existence of databases allowing easy access to the information, the attrition rate of candidate motifs will vary. The search for identical or similar motifs in other species is conveniently done by the tool Tomtom of the MEME suite of programs [29].

3.5 Motif Validation

Definitive validation of functional relevance of candidate motifs can be provided by experimental confirmation only, either via targeted mutation-induced disruption of motifs and loss of gene activation or silencing and associated phenotype or by binding assays revealing that indeed proteins (transcription factors) bind to them. Short of experimental validation, an elegant and powerful *in silico* approach is to exploit available gene expression studies conducted in the species of study.

3.5.1 Validation via Evidence Co-expression

As done for the GO-enrichment analysis (Subheading 3.4, step 2), given a motif, all genes can be partitioned into a set containing the respective motif in their promoters or not. Assuming that the presence of the motif causes the respective downstream genes to be co-expressed—at least under some conditions—pairwise correlation of gene expression levels between genes within the motif-containing set and across all available experimental conditions can be expected to be larger (larger positive values) compared to the respective pairwise correlation coefficients between all genes in the motif-absent gene set. Pearson correlation coefficients can be applied for the pairwise correlation followed by *t*-tests or nonparametric equivalent (Mann–Whitney rank sum test) and/or effect size differences (Cohen’s *d*, *see* ref. 12) applied to the resulting distributions of correlation coefficients obtained from all pairwise correlations in the two respective gene sets. Significantly higher

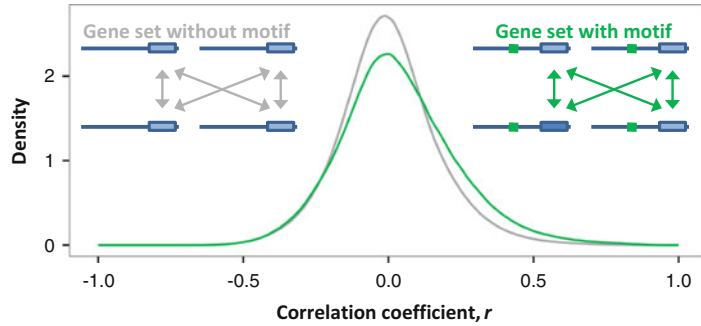


Fig. 6 Illustration of the proposed candidate motif validation scheme based on evidence of co-expression observed for genes harboring a particular candidate motif in their promoter. Assuming that motifs exert their cis-regulatory effect similarly in all genes in which they are present, the associated distribution of all pairwise correlation coefficients, r , across all available and jointly normalized expression samples (*green* density distribution) is expected to be shifted to larger positive values relative to a respective distribution derived from genes not containing this motif (*gray* density distribution). Both density distributions can be checked for significance and magnitude (effect size) of their respective difference (see **Note 7**)

correlations (toward greater positive values) can be taken as affirming the candidate motif to be functional (Fig. 6). An R-script for the computation of expression-based pairwise correlation coefficients within gene sets and the statistical comparison of two different sets as described in this paragraph is available at <http://bioinformatics.mpimp-golm.mpg.de/research-projects-publications/supplementary-data/walther> (see **Notes 3** and **8**).

4 Notes

1. Following, in essence, this protocol, Korkuc et al. [12] identified 17 candidate hexamer motifs collapsing into five consensus motifs in *Arabidopsis thaliana* with each receiving support from annotation information as well as co-expression evidence. Given that about 150 motifs have been reported in the literature, the five novel consensus motifs represent a sizeable, but not dramatic increase of the motif inventory. Recently, a set of several hundred new motifs have been reported from protein binding microarray profiling experiments [30]. Thus, and very likely caused by the low SNP-density and stringent filters to detect true-positives from indirect evidence, SNP-based motif discovery cannot be expected to yield novel motifs as possible with experimental techniques probing the binding of proteins directly.
2. To safeguard against sequencing errors being wrongly interpreted as a polymorphism, we recommend considering SNPs

with a minor allele frequency above a certain threshold (e.g. 5 %) only. Evidently, this reduces the number of polymorphic sites available for analysis. However, false-positive polymorphisms will be efficiently removed and, furthermore, only alleles which have become fixed in a population will be examined. Thus, neither detrimental nor very rare SNPs will be considered.

3. When gene expression data are available for the species under study, ideally all available datasets have been generated using the same expression profiling platform. This facilitates normalization and minimizes the risk of falsely detecting correlated gene expression. However, as our protocol suggests performing a comparison of a set of genes considered to harbor a candidate motif to a set devoid of this motif, proper controls are, in effect, always in place.
4. Due to the helical structure of DNA, TFBS motifs can also be functional as longer sequences with spaced subintervals exposed to the same face of DNA to which transcription factors are binding. Hence, motifs can also be defined as sequences with discontinuous conservation. This protocol was designed to identify contiguous motifs. However, extensions are conceivable. For example, coupled occurrences of two candidate k -mers at fixed spacing intervals between them would be indicative of the two motif seeds actually acting as one.
5. A subtlety of motif mapping concerns whether to allow overlapping motif hits or not. For example, the motif “ACA” is found twice in the sequence “GACACAT” in case of overlapping motif hits, only once otherwise. Per se, there is no strong argument in favor of either one of the two, but the mapping statistics will be different in both cases. Our preference would be to choose non-overlapping mapping hits as it more naturally follows the rationale of one-site–one-use.
6. Requiring corroborating evidence from motif mapping locations can be seen as an effective filter to enrich for true positives. If for a given motif both conservation and location preference is observed to coincide, both sides in combination will act as a powerful filter to select true positive motifs. However, as not all motifs can be assumed to actually possess positional preferences, this criterion may also increase the false-negative rate.
7. The exact location of the transcription start site of genes is oftentimes not known exactly. Furthermore, a single gene may possess multiple different start sites [31]. Thus, the position of the annotated TSS is a source of potential error. Motif mappings may wrongly be considered upstream when, in fact, they are part of the transcribed genic region.
8. The difference between correlation coefficients computed for all genes containing a particular candidate and those obtained

for a set not containing it, should not be expected to be substantial in magnitude (effect size). First, a large set of expression samples measured across different conditions may have assembled with motif action evident in only a subset of them, and furthermore, not all mapping instances can be expected to be truly functional or are modulated by the presence of additional motifs and other cis- and trans-regulatory factors. However, because of the oftentimes many expression samples, significance can typically be established.

9. As mentioned in Methods (Subheading **3.1, step 1**), insertions/deletions present a challenge for the correct multiple sequence alignment. However, when focusing on SNPs as the basis for motif discovery, considered alignment regions can be relatively easily and faithfully identified as regions devoid of any insertions/deletions in all considered sequences relative to the common reference genomic sequence.

References

1. Tagle DA, Koop BF, Goodman M, Slightom JL, Hess DL, Jones RT (1988) Embryonic epsilon and gamma globin genes of a prosimian primate (*Galago crassicaudatus*). Nucleotide and amino acid sequences, developmental regulation and phylogenetic footprints. *J Mol Biol* 203(2):439–455
2. Wasserman WW, Palumbo M, Thompson W, Fickett JW, Lawrence CE (2000) Human-mouse genome comparisons to locate regulatory sites. *Nat Genet* 26(2):225–228. doi:[10.1038/79965](https://doi.org/10.1038/79965)
3. Blanchette M, Tompa M (2002) Discovery of regulatory elements by a computational method for phylogenetic footprinting. *Genome Res* 12(5):739–748. doi:[10.1101/gr.6902](https://doi.org/10.1101/gr.6902)
4. Blanchette M, Schwikowski B, Tompa M (2002) Algorithms for phylogenetic footprinting. *J Comput Biol* 9(2):211–223. doi:[10.1089/10665270252935421](https://doi.org/10.1089/10665270252935421)
5. Blanchette M, Tompa M (2003) FootPrinter: a program designed for phylogenetic footprinting. *Nucleic Acids Res* 31(13):3840–3842
6. McGuire AM, Hughes JD, Church GM (2000) Conservation of DNA regulatory motifs and discovery of new motifs in microbial genomes. *Genome Res* 10(6):744–757
7. Gelfand MS, Koonin EV, Mironov AA (2000) Prediction of transcription regulatory sites in Archaea by a comparative genomic approach. *Nucleic Acids Res* 28(3):695–705
8. Boffelli D, McAuliffe J, Ovcharenko D, Lewis KD, Ovcharenko I, Pachter L, Rubin EM (2003) Phylogenetic shadowing of primate sequences to find functional regions of the human genome. *Science* 299(5611):1391–1394. doi:[10.1126/science.1081331](https://doi.org/10.1126/science.1081331)
9. Hong RL, Hamaguchi L, Busch MA, Weigel D (2003) Regulatory elements of the floral homeotic gene *AGAMOUS* identified by phylogenetic footprinting and shadowing. *Plant Cell* 15(6):1296–1309
10. Boffelli D (2008) Phylogenetic shadowing: sequence comparisons of multiple primate species. *Methods Mol Biol* 453:217–231. doi:[10.1007/978-1-60327-429-6_10](https://doi.org/10.1007/978-1-60327-429-6_10)
11. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* 423(6937):241–254. doi:[10.1038/nature01644](https://doi.org/10.1038/nature01644)
12. Korkuc P, Schippers JH, Walther D (2014) Characterization and identification of cis-regulatory elements in *Arabidopsis* based on single-nucleotide polymorphism information. *Plant Physiol* 164(1):181–200. doi:[10.1104/pp.113.229716](https://doi.org/10.1104/pp.113.229716)
13. Huala E, Dickerman AW, Garcia-Hernandez M, Weems D, Reiser L, LaFond F, Hanley D, Kiphart D, Zhuang M, Huang W, Mueller LA, Bhattacharyya D, Bhaya D, Sobral BW, Beavis W, Meinke DW, Town CD, Somerville C, Rhee SY (2001) The *Arabidopsis* Information Resource (TAIR): a comprehensive database and web-based information retrieval, analysis, and visualization system for a model plant. *Nucleic Acids Res* 29(1):102–105

14. Koboldt DC, Chen K, Wylie T, Larson DE, McLellan MD, Mardis ER, Weinstock GM, Wilson RK, Ding L (2009) VarScan: variant detection in massively parallel sequencing of individual and pooled samples. *Bioinformatics* 25(17):2283–2285. doi:[10.1093/bioinformatics/btp373](https://doi.org/10.1093/bioinformatics/btp373)
15. Kumar S, You FM, Cloutier S (2012) Genome wide SNP discovery in flax through next generation sequencing of reduced representation libraries. *BMC Genomics* 13:684. doi:[10.1186/1471-2164-13-684](https://doi.org/10.1186/1471-2164-13-684)
16. Issel-Tarver L, Christie KR, Dolinski K, Andrada R, Balakrishnan R, Ball CA, Binkley G, Dong S, Dwight SS, Fisk DG, Harris M, Schroeder M, Sethuraman A, Tse K, Weng S, Botstein D, Cherry JM (2002) Saccharomyces Genome Database. *Methods Enzymol* 350:329–346
17. Edgar R, Domrachev M, Lash AE (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res* 30(1):207–210
18. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, Mani R, Rayner T, Sharma A, William E, Sarkans U, Brazma A (2007) ArrayExpress—a public database of microarray experiments and gene expression profiles. *Nucleic Acids Res* 35(Database issue):D747–D750. doi:[10.1093/nar/gkl995](https://doi.org/10.1093/nar/gkl995)
19. Craigon DJ, James N, Okyere J, Higgins J, Jotham J, May S (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res* 32(Database issue):D575–D577. doi:[10.1093/nar/gkh133](https://doi.org/10.1093/nar/gkh133)
20. Wingender E, Dietze P, Karas H, Knuppel R (1996) TRANSFAC: a database on transcription factors and their DNA binding sites. *Nucleic Acids Res* 24(1):238–241
21. Sandelin A, Alkema W, Engstrom P, Wasserman WW, Lenhard B (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res* 32(Database issue):D91–D94. doi:[10.1093/nar/gkh012](https://doi.org/10.1093/nar/gkh012)
22. O'Connor TR, Dyreson C, Wyrick JJ (2005) Athena: a resource for rapid visualization and systematic analysis of Arabidopsis promoter sequences. *Bioinformatics* 21(24):4411–4413. doi:[10.1093/bioinformatics/bti714](https://doi.org/10.1093/bioinformatics/bti714)
23. Weirauch MT, Yang A, Albu M, Cote AG, Montenegro-Montero A, Drewe P, Najafabadi HS, Lambert SA, Mann I, Cook K, Zheng H, Goity A, van Bakel H, Lozano JC, Galli M, Lewsey MG, Huang E, Mukherjee T, Chen X, Reece-Hoyes JS, Govindarajan S, Shaulsky G, Walhout AJ, Bouget FY, Ratsch G, Larrondo LF, Ecker JR, Hughes TR (2014) Determination and inference of eukaryotic transcription factor sequence specificity. *Cell* 158(6):1431–1443. doi:[10.1016/j.cell.2014.08.009](https://doi.org/10.1016/j.cell.2014.08.009)
24. Kielbasa SM, Korbel JO, Beule D, Schuchhardt J, Herzel H (2001) Combining frequency and positional information to predict transcription factor binding sites. *Bioinformatics* 17(11):1019–1026
25. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate—a practical and powerful approach to multiple testing. *J R Stat Soc Series B* 57(1):289–300
26. Dunn J (1973) A fuzzy relative of the ISODATA process and its use in detecting compact well-separated clusters. *J Cybern* 3:32–57. doi:[10.1080/01969727308546046](https://doi.org/10.1080/01969727308546046)
27. Tan P-N, Steinbach M, Kummer V (2006) Cluster analysis: basic concepts and algorithms. In: Tan P-N, Steinbach M, Kummer V (eds) *Introduction to data mining*. Pearson Education, Essex, UK
28. Thompson J, Gibson T, Higgins D (2002) Multiple sequence alignment using ClustalW and ClustalX. *Curr Protoc Bioinformatics* 2–3
29. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. *Web Server Issue* 37(2):W202–W208. doi:[10.1093/nar/gkp335](https://doi.org/10.1093/nar/gkp335)
30. Jothi R, Cuddapah S, Barski A, Cui K, Zhao K (2008) Genome-wide identification of in vivo protein-DNA binding sites from ChIP-Seq data. *Nucleic Acids Res* 36(16):5221–5231. doi:[10.1093/nar/gkn488](https://doi.org/10.1093/nar/gkn488)
31. Alexandrov NN, Troukhan ME, Brover VV, Tatarinova T, Flavell RB, Feldmann KA (2006) Features of Arabidopsis genes and genome discovered using full-length cDNAs. *Plant Mol Biol* 60(1):69–85. doi:[10.1007/s11103-005-2564-9](https://doi.org/10.1007/s11103-005-2564-9)

Quantitative Analysis of Protein–DNA Interaction by qDPI-ELISA

Stefan M. Fischer, Alexander Böser, Jan P. Hirsch, and Dierk Wanke

Abstract

The specific binding of DNA-binding proteins to their cognate DNA motifs is a crucial step for gene expression control and chromatin organization *in vivo*. The development of methods for the identification of *in vivo* binding regions by, e.g. chromatin immunoprecipitation (ChIP) or DNA adenine methyltransferase identification (Dam-ID) added an additional level of qualitative information for data mining in systems biology or applications in synthetic biology. In this respect, the *in vivo* techniques outpaced methods for thorough characterization of protein–DNA interaction and, especially, of the binding motifs at single base-pair resolution. The elucidation of DNA-binding capacities of proteins is frequently done with methods such as yeast one-hybrid, electrophoretic mobility shift assay (EMSA) or systematic evolution of ligands by exponential enrichment (SELEX) that provide only qualitative binding information and are not suited for automation or high-throughput screening of several DNA motifs. Here, we describe the quantitative DNA–protein-Interaction-ELISA (qDPI-ELISA) protocol, which makes use of fluorescent fusion proteins and, hence, is faster and easier to handle than the classical DPI-ELISA. Although every DPI-ELISA experiment delivers quantitative information, the qDPI-ELISA has an increased consistency, as it does not depend on immunological detection. We demonstrate the high comparability between probes and different protein extracts in qDPI-ELISA experiments.

Key words DNA binding, Protein–DNA interaction, Quantitative analysis of DNA binding, Quantitative DNA–protein-Interaction-ELISA (qDPI-ELISA), Transcription factor-DNA binding kinetics, *Cis*-regulatory elements, Synthetic DNA probes, GFP-BPC6

1 Introduction

With fully sequenced genomes at hand, the identification and analysis of DNA-binding proteins has greatly enhanced our understanding in gene expression control. The direct binding of proteins to their target DNA sequences in the genomes can nowadays be investigated by methods such as ChIP or Dam-ID and provides a comprehensive image of *in vivo* binding regions [1–5]. Such information affects almost all fields of basic and applied research in all organisms, both prokaryotes and eukaryotes.

Single DNA-binding domains of proteins recognize their specific DNA motifs, which are usually short (~6 base pairs or even less), by either sequence readout, by local DNA-shape or both [6–9]. Despite decades of research, it is still unclear how DNA-binding proteins identify their cognate binding motif *in vivo* and differentiate between similarly short sequences in a genomic context at highest precision.

For in-depth analyses and predictions of protein–DNA interaction, high-quality binding data are invaluable for the examination of regulatory networks in bioinformatics or synthetic biology. In contrast to the growing number of reports on target region identification *in vivo*, the thorough analysis of high- and low-affinity binding motifs at base pair resolution lags behind. Especially, quantitative data on DNA-binding specificities would provide an additional level of information that is required for a detailed understanding of the dynamic processes during gene expression control at the DNA.

The standard *in vitro* method for the analysis of DNA–protein interaction is the Electrophoretic Mobility Shift Assay (EMSA), which is used to study differential binding to different DNA-probes qualitatively [10–12]. Unfortunately, the inter-comparison between different EMSA experiments or different protein extracts is hardly possible [1, 10]. A successful EMSA essentially relies on the stability of the protein–DNA complex, a low variability in the labeling efficiency between different DNA-probes and the purity of the DNA-binding protein under investigation. These three constraints prevent quantitative readout from EMSA experiments.

Other qualitative techniques, such as the Systematic Evolution of Ligands by Exponential enrichment (SELEX), EMSA-seq, Bind-n-Seq or High-Throughput Sequencing—Fluorescent Ligand Interaction Profiling (HiTS-FLIP) make use of next-generation sequencing (NGS) and require extensive bioinformatics analyses for DNA-motif discovery [13–17]. Hence, those approaches cannot be applied as a simple laboratory routine, but may be reserved for specialists.

So far, only the protein binding microarray (PBM) and the DNA–Protein-Interaction Enzyme-Linked ImmunoSorbant Assay (DPI-ELISA) provide quantitative data on the DNA-motif specificity *in vitro* [1, 6, 18–22]. Both techniques use an array of immobilized DNA-oligonucleotides that are simultaneously probed with a defined DNA-binding protein.

The DPI-ELISA is currently the only laboratory scale technique that provides a quantitative readout at reasonable revenue and expense [1, 22, 23]. As the DNA-probes can be arranged on the ELISA plate in any custom-made fashion, the DPI-ELISA offers a very broad range of downstream applications that employ ELISA microwell plate formats. For example, the DPI-ELISA was used for the automated screening of hundreds of double-stranded

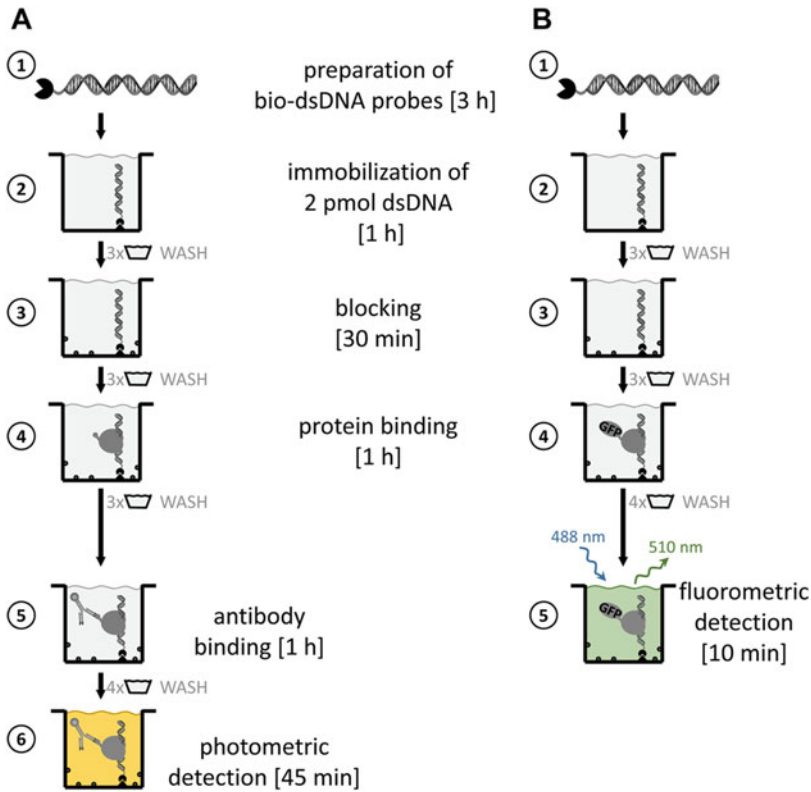


Fig. 1 Comparison of the antibody-based DPI-ELISA and the modified qDPI-ELISA. Overview of the different steps and incubation times of the classical DPI-ELISA protocol (a), that is based upon immunological detection, and the qDPI-ELISA (b), which makes use of fluorescent fusion proteins. Besides a higher robustness of the qDPI-ELISA, the protocol is faster by at least 90 min compared with the classical DPI-ELISA method

DNA-probes to uncover DNA-binding specificities de novo [6, 21]. The DPI-ELISA can also be used to search for compounds that specifically interfere with protein binding or to investigate the formation of higher order protein complexes at the DNA [23, 24]. Similarly, this technique can be utilized for studying the binding of proteins to RNA or single-stranded DNA.

Although the classical DPI-ELISA is highly reproducible and can provide quantitative information even between different DNA-probes and protein extracts [1, 6, 21–23, 25], the immunological detection of the retained proteins increased data variability considerably. Especially, different lots of antibody or varying epitope accessibility led to unanticipated variation that is barely avoidable.

Here, we provide the protocol for the quantitative DNA-protein-Interaction-ELISA (qDPI-ELISA), which uses DNA-binding proteins that were fused to fluorophores, such as GFP (Fig. 1). Thereby, we were able to decrease variability and to increase the linear range of detection by several magnitudes (Fig. 2). In addition, this modified qDPI-ELISA is much faster than the classical one with

immunological detection and results in even higher reproducibility. We also demonstrate that the qDPI-ELISA readout provides preliminary insight into binding kinetics and stoichiometry (Figs. 3 and 4). The qDPI-ELISA protocol (*see* Subheading 3.4) described here is essentially the synthesis of our previous publications [1, 6, 22] and was used before to study protein complex formation at the DNA [23]. We would like to point out that both techniques, the classical DPI-ELISA with immunological detection and the qDPI-ELISA that uses fluorescent proteins, complement each other and are used simultaneously in our laboratory for different purposes. While the qDPI-ELISA provides excellent comparative data on different DNA-probes, the use of immunological detection is still superior for all qualitative analyses, e.g. those experiments where it is unclear whether a protein binds to DNA or not.

2 Materials

Prepare all solutions using ultrapure water (prepared by purifying deionized water to attain a sensitivity of 18 M Ω cm/5.5 μ S/cm at 25 °C) and analytical grade reagents. Prepare and store all reagents at room temperature (unless indicated otherwise). Diligently follow all waste disposal regulations when disposing waste materials.

Do not add sodium azide or DNase to any of your reagents. Always use biotin-free BSA.

Thaw protein extracts immediately before use. Unless indicated otherwise, all solutions with proteins were placed on ice (0 °C).

2.1 Components for DNA-Probes

1. Order complementary sense (5'-biotinylated) and antisense (non-biotinylated) oligonucleotides from a company (*see* Note 1). Oligonucleotides should at least be 16 base pairs in length (*see* Note 2).
2. 1 M Tris-HCl, pH 7.5–8.0: Add some water to a 1 L graduated cylinder. Weigh 121.1 g Tris and transfer to the cylinder. Add water to a volume of 900 mL water. Mix and adjust pH with HCl. Make up to 1 L with water.
3. 2 M MgCl₂: Weigh 19 g MgCl₂ (anhydrous) and transfer to a 100 mL graduated cylinder. Add about 80 mL of water and mix. Make up to 100 mL with water (*see* Note 3).
4. 2 M NaCl: Weigh 11.7 g NaCl and transfer to a 100 mL graduated cylinder. Add about 80 mL of water and mix. Make up to 100 mL with water (*see* Note 3).
5. Annealing Buffer (10 \times): For a final volume of 20 mL, use a 50 mL reaction tube to mix 8 mL Tris-HCl (1 M, pH 7.5–8), 2 mL MgCl₂ (2 M), 5 mL NaCl (2 M) with 5 mL water. Make aliquots of 0.5 mL or 1 mL and store at -20 °C (*see* Note 4).
6. PCR thermocycler.

2.2 Components for Quantification of dsDNA-Probes (Optional)

1. SYBR green nucleic acids stain (*see Note 5*).
2. dsDNA as reference (*see Note 6*).
3. Plate reader that is compatible with the detection of the fluorescent dye and with ELISA plate format (*see Note 7*).
4. Regular flat-bottom black ELISA-microplates (384 wells) for fluorescence detection (*see Note 8*).

2.3 Components for Protein Detection

1. Protein extracts with fluorescent proteins of interest, e.g. GFP-fusion protein (*see Note 9*).
2. Plate reader that is compatible with the detection of the fluorescent fusion protein under investigation and with 384-well ELISA plates (*see Note 7*).
3. Regular flat-bottom black ELISA-microplates (384 wells) for fluorescence detection (*see Note 8*).

2.4 Components for the qDPI- ELISA

1. Protein extracts containing the fluorescent protein of interest, e.g. GFP-fusion protein, or appropriate control proteins, e.g. soluble monomeric GFP (*see Note 9*).
2. TBS buffer (10×): Add about 250 mL of water to a 500 mL graduated cylinder. Weigh 12.1 g Tris and transfer to cylinder. Add 52.6 g NaCl and mix. Adjust pH with HCl to pH 7.5 and make up to 500 mL with water. Dilute to TBS with water prior to use: Mix 100 mL TBS buffer (10×) and 900 mL water.
3. TBS-T: add 1 mL of Tween-20 to 1 L of TBS solution (*see Note 10*).
4. Blocking solution (2% BSA in TBS-T): Weigh 2 g of Biotin-free BSA and transfer to a 100 mL graduated cylinder. Add 10 mL of TBS buffer (10×) and make up to 100 mL with water (*see Note 11*).
5. Black flat-bottom Streptavidin-coated (2 pmol/well) ELISA-microplates (384 wells) for DPI-ELISA use (*see Note 8*).
6. Incubator chamber to provide an equal warmth of 37 °C.

3 Methods

Carry out all procedures at room temperature unless otherwise specified.

3.1 Preparation of Double-Stranded DNA-Probes

1. Dissolve sense and antisense oligonucleotides. Add equivalent amounts of water to each vial to yield 100 μM oligonucleotide stock solutions according to the manufacturer's datasheet and mix. Transfer 10 μL of each 100 μM oligonucleotide solution to a clean reaction tube and add 90 μL water to gain a 10 μM

oligonucleotide solution at working concentration. Store stock solution and leftover working solution at $-20\text{ }^{\circ}\text{C}$ (*see Note 12*).

2. Prepare 100 μL oligonucleotide mix in a PCR tube (*see Note 13*) by adding 20 μL of 5'-biotinylated sense oligonucleotide, 20 μL of non-biotinylated antisense oligonucleotide and 10 μL of annealing buffer (10 \times). Mix and then add 50 μL of water. Mix again by pipetting up and down. Place reaction tube in PCR thermocycler.
3. Operate the following temperature profile to allow annealing of the sense and antisense oligonucleotides: During 3 min at $95\text{ }^{\circ}\text{C}$ secondary structures will resolve and allow for annealing of the different strands during a gradual temperature decrease (*see Note 14*). If possible, use the ramp option of your thermocycler. Decrease the temperature by about $1\text{ }^{\circ}\text{C}$ per minute. After a gradual decrease to $28\text{ }^{\circ}\text{C}$, the annealing is finished and double-stranded (ds) DNA-probes (2 pmol/ μL) are ready-to-use. Store dsDNA-probes at $-20\text{ }^{\circ}\text{C}$ and thaw prior to use.

3.2 Linear Range for Fluorescence Measurements

As the concentration of the fluorescent protein varies between extracts, the fluorescence intensity will also differ. Moreover, total fluorescence of the same probe diverges between different microplate readers. Hence, it is essential to assess the linear range for fluorescence measurements of the microplate reader and to analyze the fluorescence intensity of each protein extract.

1. Make a protein dilution series of 1:10 and/or 1:5 in triplicates. For each 1:10 dilution, transfer three times 10 μL protein extract [30 μL total] into three clean reaction tubes and mix each with 90 μL TBS-T, to produce three technical replicates of 100 μL volume. Mix thoroughly. Make serial dilutions with TBS-T as diluent that cover a range of at least five magnitudes (e.g. no dilution, 1:10, 1:100, 1:1000, 1:10,000, and 1:100,000) (Fig. 2).
2. Transfer 30 μL of each serial dilution into wells of a regular black flat-bottom ELISA-microplate (384 wells) for fluorescence measurement.
3. Measure appropriate wells in a fluorescence microplate reader (*see Note 15*).
4. Compute average and standard error from each triplicate. Display values in a double logarithmic graph (x - and y -axis in logarithmic scale) and derive regression line (Fig. 2). It is anticipated that the detection of fluorescent proteins is about a hundred times more sensitive than the photometric measurements by classical immunological detection (Fig. 2a). Therefore, a linear measurement range over all five magnitudes is expected, e.g. for a GFP fusion protein, which was expressed in *E. coli* (Fig. 2b) (*see Note 16*).

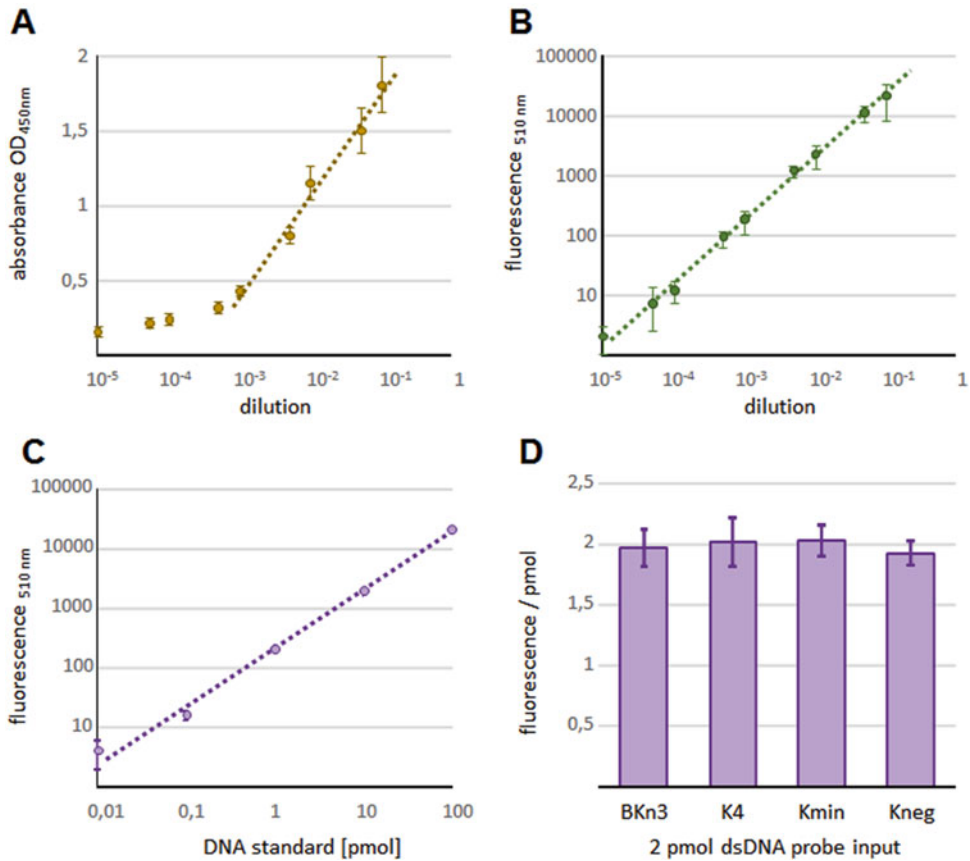


Fig. 2 Quantitative basis of the qDPI-ELISA. **(a)** and **(b)** Comparison of the linear range of the readout between the classical DPI-ELISA **(a)** and the qDPI-ELISA **(b)**, that employs fluorometric readout. The qDPI-ELISA method displays a linear range in readout over at least five magnitudes. Hence, the qDPI-ELISA protocol we describe here is much more useful to obtain dynamic quantitative data on protein binding than the classical DPI-ELISA method. In contrast, immunological detection might be favorable for proteins of low availability or with a lower binding efficiency. Note that the use of fluorescent fusion proteins allows for the possibility of sequential fluorometric and immunological detection, by using an antibody against the fluorophore. **(c)** and **(d)** The basis for a quantitative readout from any DPI-ELISA experiment is the quantitative immobilization of a defined amount of oligonucleotide probes to the plate's surface. **(c)** Standard curve for different dilutions (pmol) of double-stranded DNA stained with SYBR Green. **(d)** Measurement of SYBR Green fluorescence emission (510 nm) validates the successful immobilization of 2 pmol dsDNA-probes per microplate well

- To achieve equal loading between different protein extracts use the least fluorescing probe as reference for normalization. Dilute all extracts with TBS-T to reach equivalent fluorescence in all protein extracts (*see Note 17*).

3.3 Optional Quantification of dsDNA-Probes with SYBR-Green

It might be useful to validate the successful hybridization of the single-stranded oligonucleotides (*see Subheading 3.1*) or to quantify the amount of immobilized dsDNA-probes (*see Subheading 3.4*) in each of the ELISA wells. Therefore, a DNA stain for quantitative fluorescent readout might be beneficial

(see **Note 5**). Because of its high sensitivity and specificity for double-stranded oligonucleotides, we use a SYBR Green I nucleic acid staining protocol.

1. Prepare a standard curve by a series of 1:10 successive dilutions of dsDNA in water or TBS. Either use double-stranded oligonucleotides of known molecular weight directly, or dissolve an appropriate amount of double-stranded control DNA (see **Note 6**). Adjust to 100 ng/ μ L. Make serial dilutions of control DNA (e.g. 10 ng/ μ L, 1 ng/ μ L, 0.1 ng/ μ L, 0.01 ng/ μ L, 0.001 ng/ μ L). For each 1:10 dilution, transfer three times 10 μ L of DNA in solution into three clean reaction tubes and mix each with 90 μ L water or TBS, to produce three technical replicates of 100 μ L volume. Mix thoroughly.
2. Adjust SYBR Green I nucleic acid gel stain solution according to the manufacturers' instructions (see **Note 5**).
3. Transfer 10 μ L of DNA dilution to a black flat-bottom ELISA-microplate (384 wells). Add an appropriate amount of SYBR Green I nucleic acid gel stain solution (see **Note 6**). Fill-up with water as diluent to a final volume of 50 μ L/well.
4. Transfer ELISA plate to a microplate reader that is compatible with the detection of the fluorescent dyes. Use 485 nm excitation wavelength and measure emission at 510 nm.
5. Compute average and standard error from each triplicate. Display values in a double logarithmic graph (x - and y -axis in logarithmic scale) and derive regression line. It might be advisable to scale x -axis in pmol instead of ng, if molecular weight (g/mol) is known (Fig. 2c) (see **Note 18**).
6. Measure in triplicate respective samples of double-stranded DNA (hybridized oligonucleotides or immobilized dsDNA-probe in the DPI-ELISA plate) by adding an appropriate amount of SYBR Green I nucleic acid gel stain to the DNA (see **Note 5**). Fill-up with water as diluent to a final volume of 50 μ L/well and measure emission at 510 nm in microplate reader.
7. Derive the exact amount of double-stranded DNA in the sample from comparison with the regression curve (Fig. 2d) (see **Note 19**).

3.4 qDPI-ELISA Protocol with Fluorescent Proteins

1. For each well, mix 2.5 μ L (5 pmol) of dsDNA-probe (see Subheading 3.1) with 27.5 μ L TBS-T (see **Note 20**). Transfer 30 μ L of such a mixture to each well of a black flat-bottom Streptavidin-coated ELISA-microplate (384 wells). While planning the experiment, be sure to include control wells that do not contain any DNA and/or protein. At this step, add 30 μ L TBS-T to control wells without DNA extract.

2. After distribution of all dsDNA-probes to separate wells, incubate plate for 1 h at 37 °C in an incubator (*see Note 21*).
3. Remove the DNA containing liquid from the plate by rigorously tapping the plate upside down on absorbent papers. Before turning the plate correctly again, tap the inverted plate until the surface appears dry (*see Note 22*).
4. Rinse each well with 50 μ L TBS-T, to wash off residual unbound dsDNA-probes. Again, remove liquid from the plate by rigorously tapping the plate upside down on absorbent papers. Repeat this washing step for two additional rounds (*see Note 23*).
5. Remove residual liquids from the plate surface by tapping on fresh dry absorbent papers. Immediately continue with the protocol, as the plate should not fall dry at any moment (*see Note 24*).
6. Add 50 μ L blocking solution (2% Biotin-free BSA in TBS-T) to each well and incubate 30 min at room temperature (*see Note 25*).
7. Remove blocking solution from the plate by rigorously tapping the plate upside down on absorbent papers. Before turning the plate correctly again, tap the inverted plate until the surface appears dry.
8. Rinse each well with 50 μ L TBS-T and remove buffer from the plate by rigorously tapping the plate upside down on absorbent papers. Repeat this washing step for two additional rounds.
9. Add 30 μ L of protein extract to respective wells. While planning the experiment, be sure to include control wells that do not contain any protein extract. At this step, add 30 μ L TBS-T to control wells without protein extract.
10. After distribution of all protein extracts to separate wells, incubate plate for 1 h at 37 °C in an incubator.
11. Rinse each well with 50 μ L TBS-T and remove buffer thoroughly from the plate by rigorously tapping the plate upside down on absorbent papers. Repeat this washing step once again (*see Note 26*).
12. Rinse each well with 50 μ L TBS and remove buffer thoroughly from the plate by rigorously tapping the plate upside down on absorbent papers. Repeat this washing step once again (*see Note 27*).
13. Add 15 μ L of TBS to each well and transfer DPI-ELISA plate to a microplate reader. For GFP fusion proteins, use 485 nm excitation wavelength and measure emission at 510 nm. Set bandwidth for excitation wavelength to be as narrow as possible (e.g. 2.5 nm). For detection of the emission wavelength, bandwidth should be as broad as possible. A bandwidth of

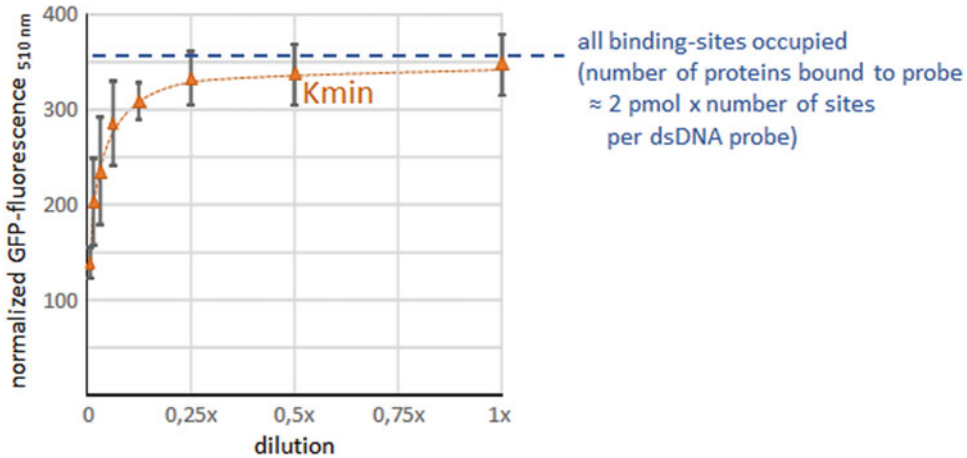


Fig. 3 Saturation curve for a DNA-binding protein showing the relation between the protein concentration and DNA-motif recognition in qDPI-ELISA. Binding of GFP-BPC6 to a dsDNA-probe (Kmin), which contains only one possible GAGA-binding motif. The Kmin dsDNA-probe (5'-TAATGCAGCAAGTAAGAGAACGAGTGTTC-3') was derived from a non-binding negative control probe (Kneg) (see refs. 22, 23). Note: The binding curve displays similarities to a classical Michaelis–Menten kinetics: As the protein concentration gets higher, the binding sites within the dsDNA-probes become saturated with binding proteins and the signal intensity approximates a maximum. For dsDNA-probes with a known number of binding sites, the amount of fluorescent proteins (pmol) can, thus, be estimated (given the proteins do not form multimers without DNA contact)

7 nm is appropriate for GFP-fluorescence (see Note 28). To achieve best results, use “find optimal gain” and a z-scan on a well containing a positive control probe (see Note 29).

14. Repeat the experiment at least three times, to get a sample size good enough for downstream statistic evaluation and to rule out artifacts or bias from handling. Also, use different proteins from different extracts. To rule out artifacts due to production of the plate, repeat the experiment at least once on different microwell plates.

3.5 qDPI-ELISA Data Evaluation

1. A quantitative readout is only possible under circumstances that do not result in saturating conditions. Therefore, settings of the microplate reader should be suitable for all wells. Also, binding sites at the dsDNA-probes might be limiting, as most of the qDPI-ELISA experiments can readily be performed with an excess of binding protein in the extracts (Fig. 3).
2. In those cases, where the number DNA binding motifs per probe is known, a saturation curve might allow to estimate the amount of binding protein in the extract (Fig. 3). Make about 9 serial dilutions (1:2) of the protein extract under investigation and perform a qDPI-ELISA in triplicates against the same dsDNA-probe. Plot average and error bars as saturation curve. Consider that 2 pmol dsDNA-probe is present in each well.

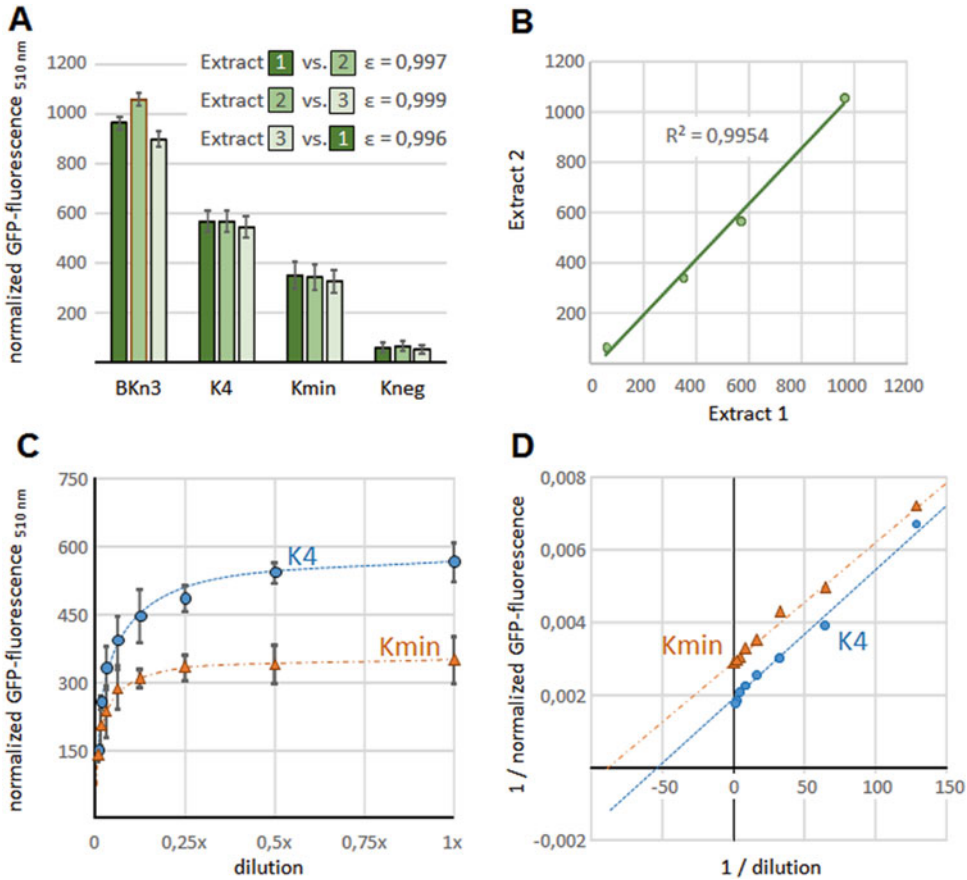


Fig. 4 High reproducibility and comparability of qDPI-ELISA experiments. The qDPI-ELISA is a robust method that allows quantitative comparison between different protein extracts and/or varying DNA-probes. (a) GFP-fluorescence of GFP-BPC6 discloses a linear correlation with the number of possible binding sites, which are contained in different dsDNA-probes (*see refs. 22, 23, 26*). For each measurement, the average and standard error of three technical replicates is shown. Correlation coefficient $\epsilon > 0,99$ between different biological replicates. (b) There is a high degree of reproducibility between different qDPI-ELISA experiments with different protein extracts. (c) Saturation curves for binding of GFP-BPC6 to dsDNA-probes, which contain either three (K4) or only one (Kmin) GAGA-motif (*see refs. 22, 23, 27*). (d) Double-reciprocal transformation of the saturation curves (c) into linear plots, which are similar to a Lineweaver–Burk plot of kinetic data for an enzyme. Such a display might harbor additional information on kinetics or specificity of a binding protein, but is especially valuable to show that both dsDNA-probes are bound by the same affinity—both trajectories are representative of nearly identical slopes

Hence, the plateau of the saturation curve will approximate 2 pmol of bound fluorescent protein, in case a single binding site is present (Fig. 3) (*see Note 30*).

- For statistical analysis, be sure to repeat the experiment several times also with different protein extracts. Generally, technical replicates are not suited for statistical analysis. Therefore, use different protein extracts on different plates as biological replicates wherever possible (Fig. 4a). Compute the average of each

technical replicate and compare values between independent protein extracts. Expect the range of fluctuations to be less than 10% between different extracts on the same dsDNA-probe (Fig. 4a, b).

4. To provide a quantitative insight into the binding specificities, you might like to perform saturation curves for the same protein, but on different dsDNA-probes. Make a dilution series of the protein extract and perform the qDPI-ELISA in triplicates on different dsDNA-probes. For comparative reasons, plot different saturation curves in one diagram (Fig. 4c). Note that the principal nature of protein–DNA interaction can be approximated as single-substrate reaction in a regular Michaelis–Menten kinetics. Hence, linearization of the saturation curves can be achieved in a Lineweaver-Burk-like display (Fig. 4d). Such a display might provide valuable information while screening for pharmaceuticals that compete or enhance binding.

4 Notes

1. Order oligonucleotides as lyophilized and desalted probes. To prevent problems, use quality-checked oligonucleotides of highest purity—especially for probes longer than 20 base pairs. Known binding motifs should be placed more to the centre or distal to the plate's surface. All our experiments were performed with one strand being biotinylated.

Also RNA-binding proteins can be studied by DPI-ELISA techniques, most of which will also accept single-stranded DNA instead of RNA. By using RNA, order dissolved RNA-oligonucleotides and adjust pH in buffers according to your needs (usually more acidic to prevent hydrolysis).

2. For longer probes (>100 base pairs), a 5' biotinylation at both ends of the dsDNA-probes might reduce flexibility and, therefore, increases binding efficiency for some proteins. Likewise, oligonucleotide hybridization efficiency varies and, hence, particularly repetitive sequences that are known to form Z-like DNA-conformations or extended palindromes should be avoided.
3. Only 2 mL of the 2 M stock solution is needed for one share of 20 mL annealing buffer (10×). Make as many 2 mL aliquots from the remaining 98 mL solution as you need and store at –20 °C. When frozen, thaw and warm the stock solution prior to use.
4. When frozen aliquots of the annealing buffer (10×) are used, be sure to thaw them completely and mix briefly. Warm to room temperature prior to use, to avoid precipitation of humidity.

5. We use SYBR[®] Green I nucleic acid gel stain (Invitrogen) according to the manufacturer's description. Keep in mind that this SYBR[®] Green I solution is 10,000× concentrated. We, therefore, prepare a 1:100 stock solution in 1× TE (pH 8) and store it in a foiled 1.5 mL reaction tube at 4 °C. The solution at working concentration requires an additional dilution of 1:100.
6. To show quantitative binding of the oligonucleotides to the 2 pmol Streptavidin on the plate, we use double-stranded oligonucleotides as reference DNA, because it can be ordered at defined quantities (e.g. pmol). Still, all different kinds of DNA can be used, e.g. plasmid or salmon sperm, but they are provided as a weighted volume and not as the absolute amount of a substance. Hence, some conversions might be required.
7. Be sure that the microplate reader is compatible with 384 microplates and equipped with a suitable set of filters compatible with the respective fluorophore. We use a spectrofluorometer with options for automatic adjustment to optimal gain and detection of measurement position (z-scan).
8. Regular flat-bottom black ELISA-microplates are needed for normalization of different fluorescent protein extracts or the measurement of standard curves for proteins (*see* Subheading 3.2) and/or DNA (*see* Subheading 3.3). The qDPI-ELISA is performed in black flat-bottom Streptavidin-coated (2 pmol/well) ELISA-microplates (*see* Subheading 3.4). The protocols described here are calculated for 384 well ELISA-microplates. We experienced that these plates provide a much better signal-to-noise ratio, compared to, e.g. 96-well ELISA plates. Still, using black 96-well ELISA-microplates is possible with 4 times higher volumes for all reagents and an elevated amount of immobilized Streptavidin per plate where appropriate.
9. Express your fluorescent protein of interest and appropriate control proteins in any expression system you like. Prepare native protein extracts according to your expression system. Wherever possible, perform protein extraction under mild and native conditions. Denaturing and renaturing of proteins is sometimes problematic in functional assays like the qDPI-ELISA.

Avoid extraction buffers and protocols that contain DNase, as this will degrade the dsDNA-probes during the assay. Always use biotin-free BSA: Biotin in solution reduces signal intensity, possibly by a low rate of competition with the biotinylated dsDNA-probes [22]. The use of protease inhibitors according to the manufacturers' descriptions and up to 1 mM 1,4-Dithiothreitol (DTT) is recommended.

It is evident that the fluorescent protein under investigation needs to be detectable in the soluble fraction and at considerable amounts. We recommend to do Coomassie staining of your protein gel and Western blot. These analyses can be performed under denaturing conditions. Always use an antibody against your fluorophore, e.g. an anti-GFP-antibody to detect possible degradation or free GFP. All analyses should provide a prominent band of the expected size. Be sure to check your control extracts, too. Also check for fluorescence signals at the respective emission wavelength in your extract.

10. Tween-20 is a non-ionic detergent, which has a lower critical micelle concentration than ionic detergents. Still, when pipetting detergents into a large volume of water, micelles will occur that will take a while to dissolve properly. Therefore, stir gently with magnetic mixer to avoid foam, but for quite a while until properly dissolved. Use a cut end of a blue tip to aspirate Tween-20 easily.
11. Be sure to use Biotin-free BSA. Do not use milk powder or any commercial blocking reagent that might contain Biotin. It has been noted before that residual Biotin could possibly affect quantitative readout (*see ref. 22*).
12. When frozen, thaw and warm the stock solution or working solution to room temperature prior to use, to avoid precipitation.
13. Use heated lid of your PCR thermocycler or cover reaction mix with mineral oil. We noted a better annealing in larger volumes (100 μ L), although smaller volumes might be more appropriate.
14. Note that some oligonucleotides, e.g. long DNA, might require different duration of the initial denaturing step. For probes longer than 30 bp, we recommend to examine successful annealing by the optional quantification of dsDNA-probes with SYBR-Green (*see Subheading 3.3*).
15. We use a spectrofluorometer with options for automatic adjustment to optimal gain and detection of measurement position (z-scan). Use wells with highest protein concentration to compute optimal gain and measurement position. Be sure to set excitation bandwidth as narrow as possible, to minimize overlapping excitation and emission spectra. Also, use these values as reference, for possible normalization.
16. We test our protein of interest for linear detection over a broad range of concentrations. For the standard curve, however, you might want to use exact amounts of commercially available fluorescent dyes or GFP.

17. To achieve the best possible quantitative readout and comparability between different extracts, the fluorescence intensities between different protein extracts should not exceed a twofold difference.
18. We use custom-made oligonucleotides or commercial vectors as reference DNA, which come at exact concentrations. Oligonucleotides are favorable, because their amount of a substance (molecular weight) is known and they can be adjusted to a very defined volume. If the x -axis of the standard curve is displayed in pmol, the amount of immobilized dsDNA can readily be inferred from the diagram without conversion.
19. As a fixed amount of 2 pmol Streptavidin is immobilized at the bottom of each well, one should expect that also exactly 2 pmol of dsDNA-probe are bound via the Biotin. Therefore, calculate the expected molecular weight of your polymerized and double-stranded DNA-probe. Use the following equation to compute the exact molecular weight (MW) of the double-stranded DNA from the number of each nucleotides (#) of only one strand:

$$\text{MW}(\text{g} / \text{Mol}) = \#A(617.4) + \#C(618.4) + \#G(618.4) + \#T(617.4) + 36.0$$

Convert the 2 pmol of your dsDNA under investigation into an equivalent weight and compare this value with your measurement and the standard curve. We usually get a quantitative binding of the dsDNA-probes to the available Streptavidin with only little variabilities (Fig. 2d). Fluorescence intensities that are significantly lower than expected indicate problems during the hybridization of the oligonucleotides or in initial dilution steps. Higher values might indicate an unspecific and unwanted binding to the plate.

20. Make master mixes to reduce variability and pipetting errors. Every dsDNA-probe should be tested in triplicate (3× technical replicates).
21. During this step the dsDNA-probes adhere quantitatively to the 2 pmol immobilized Streptavidin at the plate bottom. You might want to cover the plate with a lid to avoid evaporation. There is no need to shake or agitate the plate during this incubation step.
22. The plate's surface should be dry after removal of liquids, to prevent left-over solutions from flowing back into some wells, which might be a possible cause for contaminations and high background signals.
23. After the last washing step, the protocol can be paused at this point for overnight. The DNA inside the wells should be overlaid with TBS-T. Properly cover plate with a clean lid and store at 4 °C.

24. Any drying of the wells will affect the quantitative readout and increase unnecessarily the variability between replicates or different samples.
25. After the 30 min incubation, the protocol can be paused at this point for overnight. The plate needs to be sealed properly to prevent any evaporation.
26. These two rounds of washing with TBS-T are the most crucial to remove residual unbound protein quantitatively from the wells.
27. We feel that these two rounds of washing with TBS instead of TBS-T appear to increase the quality of the washing procedure and to improve the readout. It is also possible, of course, to wash a total of 4 times with TBS-T.
28. There is a certain possibility of overlapping fluorescence spectra, if bandwidth is larger than 10 nm. For example, excitation (488 nm) and emission (510 nm) maxima for GFP are very close to each other and there might be a noticeable “bleeding” effect from the excitation light into the detectable emission at larger bandwidths. Thus, excitation bandwidth should be as narrow as possible. As a rough rule of thumb, one-third the distance between excitation and emission maxima is an appropriate bandwidth for detection of the fluorescence emission, e.g. for GFP: $510 \text{ nm} - 488 \text{ nm} = 22 \text{ nm} \rightarrow$ divided by 3 is about 7 nm, which we consider an appropriate emission bandwidth to avoid overlapping spectra.
29. A large gain results in high background values of negative control probes and of empty wells. Likewise, a high gain might have a generally bad effect on the signal-to-noise ratio. Therefore, decide for the smallest gain where possible. Always be sure not to measure outside the focal optimum close to the plate’s surface, e.g. a focus point above the buffer or below the bottom of the plate. This might be avoided by a z-scan that will usually uncover the best measuring position.
30. Sometimes the amount of the protein-of-interest in the extracts is too low to reach a saturation. Then use protein concentration spin columns to narrow the total volume. Such concentration columns are available for different molecular exclusion sizes. Be sure to use appropriate ones.

Acknowledgements

We like to thank Luise H. Brand and Klaus Harter for continuous support. We acknowledge Angelika Anna and Sabine Hummel for technical assistance.

References

1. Brand LH, Satbhai SB, Kolukisaoglu Ü, Wanke D (2013) Limits and prospects of methods for the analysis of DNA-protein interaction. In: Berendzen KW (ed.), Kilian J, Wanke D (co-eds.) *The analysis of regulatory DNA: current developments, knowledge and applications uncovering gene regulation*. Bentham Science Publishers, pp. 124–148
2. Gordan R, Hartemink AJ, Bulyk ML (2009) Distinguishing direct versus indirect transcription factor-DNA interactions. *Genome Res* 19:2090–2100
3. Massie CE, Mills IG (2008) ChIPping away at gene regulation. *EMBO Rep* 9:337–343
4. Germann S, Gaudin V (2011) Mapping in vivo protein-DNA interactions in plants by DamID, a DNA adenine methylation-based method. *Methods Mol Biol* 754:307–321
5. Orian A, Abed M, Kenyagin-Karsenti D, Boico O (2009) DamID: a methylation-based chromatin profiling approach. *Methods Mol Biol* 567:155–169
6. Brand LH, Hennes C, Schussler A, Kolukisaoglu HU, Koch G, Wallmeroth N, Hecker A, Thurow K, Zell A, Harter K et al (2013) Screening for protein-DNA interactions by automatable DNA-protein interaction ELISA. *PLoS One* 8, e75177
7. Rohs R, Jin X, West SM, Joshi R, Honig B, Mann RS (2010) Origins of specificity in protein-DNA recognition. *Annu Rev Biochem* 79:233–269
8. Rohs R, West SM, Liu P, Honig B (2009) Nuance in the double-helix and its role in protein-DNA recognition. *Curr Opin Struct Biol* 19:171–177
9. Schroder A, Eichner J, Supper J, Wanke D, Hennes C, Zell A (2010) Predicting DNA-binding specificities of eukaryotic transcription factors. *PLoS One* 5, e13876
10. Fried MG, Bromberg JL (1997) Factors that affect the stability of protein-DNA complexes during gel electrophoresis. *Electrophoresis* 18:6–11
11. Gaudreault M, Gingras ME, Lessard M, Leclerc S, Guerin SL (2009) Electrophoretic mobility shift assays for the analysis of DNA-protein interactions. *Methods Mol Biol* 543:15–35
12. Hellman LM, Fried MG (2007) Electrophoretic mobility shift assay (EMSA) for detecting protein-nucleic acid interactions. *Nat Protoc* 2:1849–1861
13. Wong D, Teixeira A, Oikonomopoulos S, Humburg P, Lone IN, Saliba D, Siggers T, Bulyk M, Angelov D, Dimitrov S et al (2011) Extensive characterization of NF- κ B binding uncovers non-canonical motifs and advances the interpretation of genetic functional traits. *Genome Biol* 12:R70
14. Chai C, Xie Z, Grotewold E (2011) SELEX (Systematic Evolution of Ligands by EXponential Enrichment), as a powerful tool for deciphering the protein-DNA interaction space. *Methods Mol Biol* 754:249–258
15. Roulet E, Busso S, Camargo AA, Simpson AJ, Mermod N, Bucher P (2002) High-throughput SELEX SAGE method for quantitative modeling of transcription-factor binding sites. *Nat Biotechnol* 20:831–835
16. Bonvin AM, Boelens R, Kaptein R (2005) NMR analysis of protein interactions. *Curr Opin Chem Biol* 9:501–508
17. Wang X, Kuwahara H, Gao X (2014) Modeling DNA affinity landscape through two-round support vector regression with weighted degree kernels. *BMC Syst Biol* 8:S5
18. Agius P, Arvey A, Chang W, Noble WS, Leslie C (2010) High resolution models of transcription factor-DNA affinities improve in vitro and in vivo binding predictions. *PLoS Comput Biol* 6, e1000916
19. Berger MF, Bulyk ML (2009) Universal protein-binding microarrays for the comprehensive characterization of the DNA-binding specificities of transcription factors. *Nat Protoc* 4:393–411
20. Bulyk ML (2006) Analysis of sequence specificities of DNA-binding proteins with protein binding microarrays. *Methods Enzymol* 410:279–299
21. Brand LH, Fischer NM, Harter K, Kohlbacher O, Wanke D (2013) Elucidating the evolutionary conserved DNA-binding specificities of WRKY transcription factors by molecular dynamics and in vitro binding assays. *Nucleic Acids Res* 41:9764–9778
22. Brand LH, Kirchler T, Hummel S, Chaban C, Wanke D (2010) DPI-ELISA: a fast and versatile method to specify the binding of plant transcription factors to DNA in vitro. *Plant Methods* 6:25
23. Hecker A, Brand LH, Peter S, Simoncello N, Kilian J, Harter K, Gaudin V, Wanke D (2015) The Arabidopsis GAGA-Binding Factor BASIC PENTACYSTEINE6 Recruits the POLYCOMB-REPRESSIVE COMPLEX1 Component LIKE HETEROCHROMATIN PROTEIN1 to GAGA DNA Motifs. *Plant Physiol* 168:1013–1024
24. Alonso N, Guillen R, Chambers JW, Leng F (2015) A rapid and sensitive high-throughput screening method to identify compounds

- targeting protein-nucleic acids interactions. *Nucleic Acids Res* 43, e52
25. Soyk S, Simkova K, Zurcher E, Luginbuhl L, Brand LH, Vaughan CK, Wanke D, Zeeman SC (2014) The enzyme-like domain of Arabidopsis nuclear beta-amylases is critical for DNA sequence recognition and transcriptional activation. *Plant Cell* 26:1746–1763
 26. Santi L, Wang Y, Stile MR, Berendzen K, Wanke D, Roig C, Pozzi C, Muller K, Muller J, Rohde W et al (2003) The GA octodinucleotide repeat binding factor BBR participates in the transcriptional regulation of the homeobox gene *Bkn3*. *Plant J* 34:813–826
 27. Kooiker M, Airoldi CA, Losa A, Manzotti PS, Finzi L, Kater MM, Colombo L (2005) BASIC PENTACYSTEINE1, a GA binding protein that induces conformational changes in the regulatory region of the homeotic Arabidopsis gene *SEEDSTICK*. *Plant Cell* 17:722–729

Analyzing Synthetic Promoters Using Arabidopsis Protoplasts

Ralf Stracke, Katharina Thiedig, Melanie Kuhlmann,
and Bernd Weisshaar

Abstract

This chapter describes a transient protoplast co-transfection method that can be used to quantitatively study *in vivo* the activity and function of promoters and promoter elements (reporters), and their induction or repression by transcription factors (effectors), stresses, hormones, or metabolites. A detailed protocol for carrying out transient co-transfection assays with Arabidopsis At7 protoplasts and calculating the promoter activity is provided.

Key words *Arabidopsis thaliana*, At7 cell culture, Protoplast co-transfection assays, Reporter gene

1 Introduction

For high-level constitutive expression, or for precise control of transgene activity in response to a specific stimulus, promoters are the key for successful genetic engineering strategies. Therefore, detailed knowledge about the concerted action of both *cis*- and *trans*-acting elements is necessary. Defining *cis*-acting elements and the characterization of transcription factors which bind and/or regulate a promoter of choice is a standard experimental approach. Since *in vitro* DNA-binding assays might not reflect the situation in a living cell, *in vivo* assays are favored. The yeast one-hybrid assay [1] is the method of choice for rapid detection of protein–DNA interactions. However, this experimental approach has some limitations for the analysis of plant transcription factors and promoters. Among these are different conditions inside the yeast nucleus compared to the situation in plants, additional transcriptional start sites in promoters larger than approximately 300 bps in length, and false positives due to regulatory proteins with high affinity to unspecific DNA regions [2]. For the analysis of (synthetic) plant promoters, the use of plant cell systems avoids

most of the disadvantages of yeast systems. Furthermore, plant systems provide necessary perception and signaling systems for the study of signal-induced processes.

Transient protoplast co-transfection systems are used for the study of unbiased activity and function of promoter elements, as well as for promoter activation or repression by a single or multiple transcription factors or several stresses or hormone treatments. In this respect, the transfection assay is focused on the analysis of factors that control the activity of a given promoter, and the role of proximal 5'-upstream *cis*-regulatory elements or sequences. These elements are important and central components for accurate and refined synthetic promoter design. Transient protoplast co-transfection assays allow fast access to results, especially when compared with stable transformation. In addition, they are unaffected by position effects caused by features of the site of transgene integration and by the copy number of inserted transgenes. The drawback is that cell type specificity or developmental control of promoter activity can usually not be studied in protoplasts from cultured cells.

The protoplast assay system employs purified plasmid DNA introduced into the cells via PEG-mediated DNA uptake. Various plasmids can be introduced at the same time (co-transfection). The protoplast assay system relies on assessing the level of gene expression for an engineered promoter construction that drives the expression of a reporter gene. The level of expression of the reporter, in the case described here, β -glucuronidase (*uidA*, GUS) is taken as a measure for promoter activity. Thus, transient protoplast co-transfection assays have provided a wealth of information about *cis* elements required for promoter function, transcription factors, and signaling proteins that regulate expression of genes and signals regulating inducible gene expression [3–8].

2 Materials

2.1 Cells, Buffers, and Solutions

1. At7 cell culture: hypocotyl-derived *A. thaliana* Columbia cell culture [9], maintained at 26 °C in darkness on a rotary shaker, weekly subcultured.
2. *dam*- *E. coli*: Methylation-deficient *dam* and *dcm* *E. coli* strain K12 ER2925 (NEB).
3. 1000 \times 2,4-Dichlorophenoxyacetic acid (2,4-D): 1 mg/mL.
4. MS medium: 4.3 g Murashige and Skoog Basal Salt Mixture (MS, Sigma-Aldrich), 1 mL 1000 \times 2,4-D, 10 mL 1000 \times Gamborg's Vitamin Solution (Sigma-Aldrich), 30 g sucrose. Adjust to pH 5.7 with 1 M KOH and bring to 1 L with deionized H₂O. Autoclave.

5. Cellulase-mazerozyme solution: 1.16% (w/v) cellulase Onozuka R-10 (Serva), 0.27% mazeroyzyme R-10 (Serva). Solve in 240 mM CaCl₂, stir cautiously until enzymes are dissolved (1–1.5 h). Pass enzyme solution through a folded filter paper, then filter sterilize.
6. B-5 floating medium (B5 solution): 3.1 g Gamborg's B-5 Basal Salt Mixture (Sigma-Aldrich), 1 mL 1000×2,4-D, 136 g sucrose. Adjust to pH 5.7 with 1 M NaOH and bring to 1 L with deionized H₂O. Filter sterilize.
7. 240 mM CaCl₂. Autoclave.
8. PEG solution: 125 g PEG 6000, 11.8 g Ca(NO₃)₂×4H₂O, 41 g mannitol. Adjust to pH 9 with 1 M KOH and bring to 0.5 L with deionized H₂O. Filter sterilize and store in 5 mL aliquots at –20 °C.
9. 275 mM Ca(NO₃)₂: Adjust to pH 6.0 with 1 M KOH. Autoclave.
10. 0.1 M K₂HPO₄ and 0.1 M KH₂PO₄. Autoclave.
11. 0.1 M Potassium phosphate: Mix the appropriate volumes of 0.1 M K₂HPO₄ and 0.1 M KH₂PO₄ for a desired pH of 7.0. Store at 4 °C up to 1 month.
12. Protein extraction buffer: 100 mM potassium phosphate, 1 mM DTT (Sigma-Aldrich). Filter sterilize and store at 4 °C.
13. 2× Luciferase assay stock solution: 40 mM tricine, 2.14 mM Mg(CO₃)₄Mg(OH)₂×5H₂O, 5.34 mM MgSO₄×7H₂O, 0.2 mM EDTA. Store at 4 °C.
14. Luciferase substrate solution: 1× luciferase assay stock solution, 33.3 mM DTT (Sigma-Aldrich), 270 μM CoA trilithium salt (Sigma-Aldrich), 470 μM luciferin (Roche), 570 μM ATP (Sigma-Aldrich). CoA trilithium salt and luciferin are light-sensitive, keep in the dark. Check pH and adjust to pH 7.5 if necessary. Filter sterilize and store in 5 mL aliquots at –80 °C in light-tight tubes.
15. 0.5 M Na₂HPO₄ and 0.5 M NaH₂PO₄. Autoclave.
16. 0.5 M Sodium phosphate buffer: Mix the appropriate volumes of 0.5 M Na₂HPO₄ and 0.5 M NaH₂PO₄ for a desired pH of 7.0. Store at 4 °C up to 1 month.
17. GUS buffer: 50 mM sodium phosphate buffer, 1 mM EDTA pH 8.0, 0.1% (v/v) Triton X-100, 10 mM β-mercaptoethanol.
18. 4-MUG substrate solution: 20 mM 4-methylumbelliferyl-β-D-glucopyranosiduronic acid (4-MUG, Sigma-Aldrich). Solve in GUS buffer. Filter sterilize and store in 15 mL aliquots at –20 °C.
19. MU stock solution: 10 mM 4-methylumbelliferone (MU, Sigma-Aldrich). Solve in ethanol. Store at 4 °C.

20. MU dilution series: Dilute the MU stock solution with GUS buffer to 0, 2.5, 5.0, 12.5, 25, 50, 100, 150, 200, and 250 μM .
21. BSA dilution series: Dilute a 10 mg/mL BSA stock solution with protein extraction buffer to 0, 2, 4, 8, and 16 μg per 10 μL buffer.
22. Protein assay dye reagent: Dilute the protein assay dye reagent concentrate (Bio-Rad) 1:5 with deionized H_2O .

2.2 Equipment

1. Laminar flow cabinet.
2. 50 mL Falcon tubes.
3. 13 mL centrifuge tubes.
4. 1.5 mL reaction tubes.
5. Cell-Saver tips (Biozyme Scientific).
6. Petri dishes (145 mm).
7. Folded filter paper.
8. Sterile filter units with 0.22 μm pore size.
9. Incubator.
10. Rotary shaker.
11. Centrifuge with swing-out rotor; programmable settings should include the specification of acceleration/deceleration rates (*see Note 8*).
12. Benchtop centrifuge.
13. Plasmid purification kit with prepacked gravity-flow anion-exchange columns in maxi (500 μg) or mega (2.5 mg) scale; e.g. Plasmid Mega Kit (Qiagen), JETstar Plasmid Purification MIDI Kit (Genomed).
14. Fluid aspiration system.
15. Hemocytometer.
16. Vortexer.
17. FLUOstar Optima microplate reader (BMG Labtech) equipped with an on-board syringe injector (*see Note 1*).
18. 96-well microplates white LUMITRAC 200 (Greiner) (*see Note 1*).
19. 96-well microplates black FLUOTRAC 200 (Greiner) (*see Note 1*).
20. Nunc™ MicroWell™ 96-well microplates (Nunc) (*see Note 1*).
21. Equipment for maintenance of a cell suspension culture.
22. Nylon net filter with 70 μm pore size (optional, *see Note 10*).

2.3 Plasmids

1. Reporter constructs
Reporter constructs are based on the vector pBT10GUS [5] or the Gateway-compatible derivative pDISCO [10]. In both

cases the promoter (fragment) to analyze is inserted or recombined upstream of the *uidA* ORF:nos terminator cassette (see **Notes 2** and **3**).

In case of testing the activity of a transcriptional repressor, the construction of a weakly active synthetic promoter is needed (see **Note 4**).

2. Effector expression constructs

Effectors were expressed under the control of the strong constitutive CaMV 35S promoter (positions -417 to +8), inserted into classical pBT vectors [3] or the Gateway-compatible derivative pBTdest [6].

3. LUC standardization plasmid

The LUC standardization plasmid used in the co-transfection assays contains a *Photinus pyralis* luciferase (LUC) encoding open reading frame [11] under the control of the constitutive *Petroselinum crispum* UBI4-2 promoter [12] in pBT (see **Notes 5** and **6**).

4. Filling plasmid

The promoter-deleted standardization plasmid pBT10- Δ -LUC, which in protoplasts leads to no detectable luciferase activity, is added to keep the total amount of the transfected plasmid DNA constant (25 μ g).

3 Methods

This protocol uses *A. thaliana* At7 cells from a cell suspension culture which is maintained at 26 °C in the dark on a rotary shaker at 105 rpm. Cells are subcultured once a week by transferring approximately 2.8 g of cells to 40 mL fresh MS-medium. For protoplast isolation, additional Erlenmeyer flasks with 40 mL MS-medium were inoculated 5 days prior to the harvesting of the cells.

3.1 Protoplast Preparation

1. For each subcultivated Erlenmeyer flask of At7 cells with 40 mL cell culture prepare 60 mL of cellulase-mazerozyme solution (see **Note 7**).
2. Transfer each 5-day-old At7 cell suspension subculture to a 50 mL Falcon tube.
3. Centrifuge 5 min at $130\times g$ with moderate acceleration (5/9) and deceleration (3/9) in a swing-out rotor (see **Note 8**).
4. Discard the supernatant carefully.
5. Detach cell pellet with caution by soft tapping.
6. Add 50 mL 240 mM CaCl_2 and resuspend the cells by gentle inversion of the tube.
7. Centrifuge 5 min at $130\times g$ with moderate acceleration (5/9) and deceleration (3/9) in a swing-out rotor.

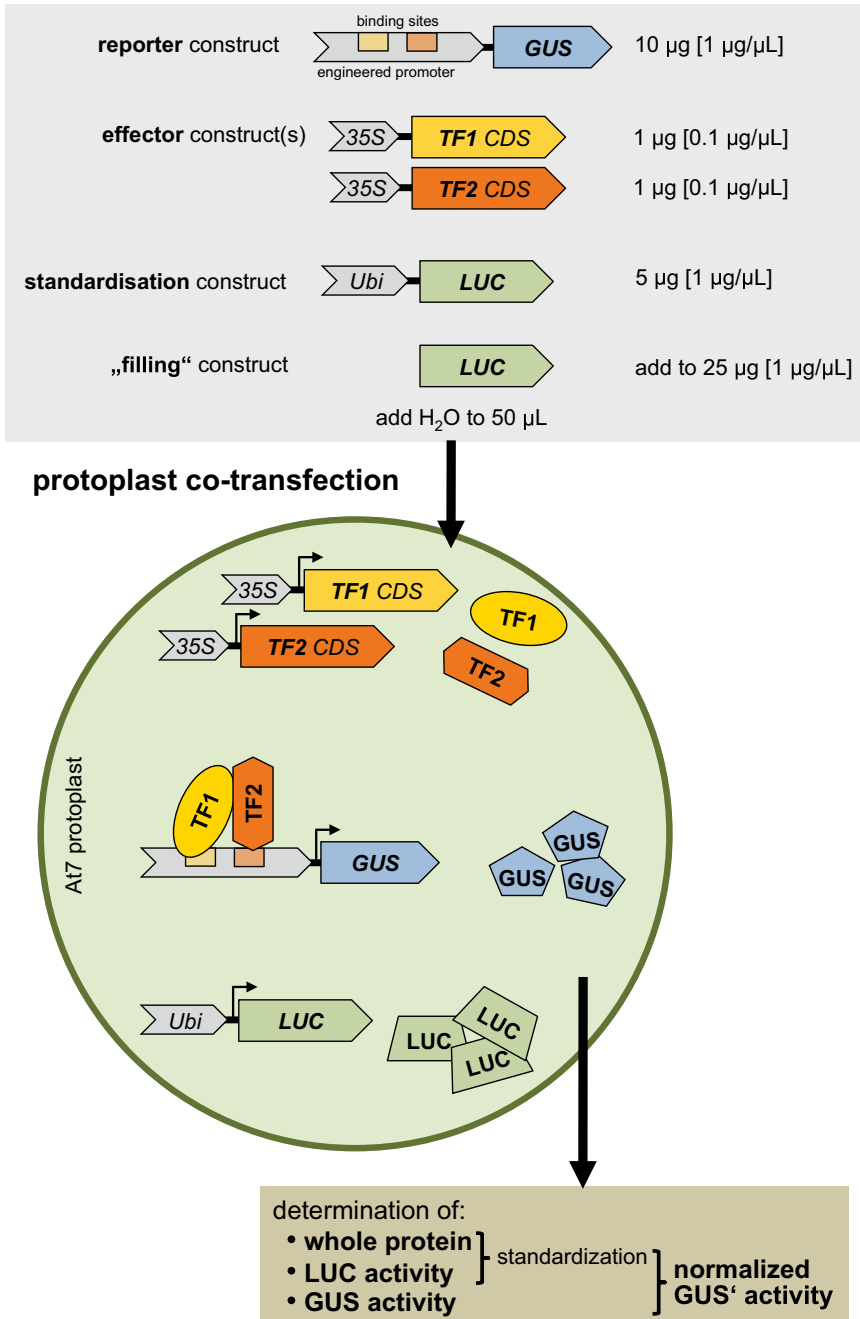


Fig. 1 Schematic depiction of transient protoplast expression assay with co-transfected At7 protoplasts. The *gray box* at the *top* gives the plasmids (constructs) used in co-transfection experiments with the amounts and concentrations specified. The middle part depicts a transfected At7 protoplast with plasmids (*italics*) and the produced proteins (*normal letters*) and their interactions. The “filling” plasmid is not considered. The *brown box* at the *bottom* summarizes the calculation of normalized specific GUS' activity, as a measure for promoter activity. Abbreviations: 35S cauliflower mosaic virus 35S promoter, CDS coding sequence, GUS β-glucuronidase, LUC luciferase, TF transcription factor, Ubi *ubiquitin4-2* promoter

8. Discard the supernatant carefully.
9. For each harvested At7 cell suspension flask, prepare two 145 mm Petri dishes each with 10 mL cellulase-mazerozyme solution (*see Note 7*).
10. Little by little, add the residual 40 mL of the enzyme solution to the cell pellet and gently resuspend by inversion. Avoid cell clumping.
11. Add half of the cell suspension (about 20 mL) to each of the prepared Petri dishes.
12. Incubate overnight at 26 °C in the dark, shake at 20 rpm.
13. Intensify the shaking to 40 rpm for no longer than 20 min.
14. From each Petri dish, carefully transfer the protoplasts into a 50 mL Falcon tube (*see Note 9*).
15. Centrifuge 5 min at $90 \times g$ with moderate acceleration (5/9) and deceleration (3/9) in a swing-out rotor.
16. Discard the supernatant carefully.
17. Detach the pellet with caution by soft tapping.
18. Wash the protoplasts by little and little adding 25 mL 240 mM CaCl_2 and resuspend the protoplasts by gentle inversion of the tube (*see Note 10*).
19. Centrifuge 5 min at $90 \times g$ with moderate acceleration (5/9) and deceleration (3/9) in a swing-out rotor.
20. Discard the supernatant carefully.
21. Detach the pellet with caution by soft tapping.
22. Little by little add 20 mL B-5 floating medium to each pellet and combine two resuspended pellets from 50 mL Falcon tubes in one tube (*see Note 11*).
23. Centrifuge 5 min at $130 \times g$ with maximal acceleration (9/9) and minimal deceleration (1/9) in a swing-out rotor.
24. Transfer the floating protoplasts (2–5 mL) with a Cell-Saver tip into a new 50 mL Falcon tube (*see Note 12*).
25. Cautiously fill the Falcon tube with B-5 floating medium.
26. Centrifuge 5 min at $130 \times g$ with maximal acceleration (9/9) and minimal deceleration (1/9) in a swing-out rotor.
27. Pool all floating protoplasts in a 13 mL centrifuge tube using a Cell-saver tip.
28. Assess the quality of the protoplast suspension (*see Note 13*).
29. Use the protoplasts immediately for transfection. 200 μL of protoplasts (containing $1\text{--}2 \times 10^6$ At7 protoplasts) are needed per co-transfection (*see Note 14*).

3.2 Preparing DNA for Co-transfection

1. Plasmids to be used in protoplast co-transfection experiments are retransformed into the *dam* and *dcm* methylation-deficient *E. coli* strain K12 ER2925 (NEB) (*see Note 15*).
2. Plasmid DNA is prepared from methylation-deficient *E. coli* strain using a plasmid purification kit with prepacked gravity-flow anion-exchange columns in maxi (500 µg) or mega (2.5 mg) scale according to manufacturer's instructions.
3. The high concentrated *dam*- plasmid DNA should be stored at 4 °C (*see Note 16*).
4. For co-transfection experiments, plasmid dilutions have to be made: reporter constructs, standardization constructs and "filling" constructs with 1 µg/µL, effector constructs with 0.1 µg/µL (Fig. 1).
5. Combined plasmid DNA solutions for co-transfection should be prepared in advance to enable a smooth execution of the protocol. All combined DNA solutions should contain an equal amount of plasmid DNA (25 µg) in a volume of 50 µL (Fig. 1). The use of a positive control (35S::GUS reporter construct) and a negative control (TATA::GUS reporter construct, containing only the truncated -46 minimal promoter (TATA)) is recommended in the experimental setup (*see Note 17*).

3.3 Protoplast Co-transfection

1. Thaw PEG solution at room temperature (*see Note 18*).
2. For each co-transfection transfer 200 µL of the protoplast suspension with a Cell-Saver tip into a 13 mL centrifuge tube.
3. Pipet the prepared DNA solutions (25 µg in 50 µL) onto the protoplasts (*see Note 19*).
4. Add 200 µL PEG solution and mix thoroughly but gently by soft shaking and tapping of the tube rack (*see Note 20*).
5. Incubate the protoplast-DNA-PEG mixture at room temperature for 15 min (*see Note 21*).
6. Stop the transfection reaction by stepwise adding 5 mL 275 mM Ca(NO₃)₂ (*see Note 22*).
7. Centrifuge 5 min at 90×g with maximal acceleration (9/9) and moderate deceleration (5/9) in a swing-out rotor.
8. Discard the supernatant carefully.
9. Stepwise add 7 mL B5 solution (*see Note 23*).
10. Incubate at 26 °C for approximately 20 h in the dark, keeping the tubes in an almost horizontal position (*see Note 24*).

3.4 Harvesting of Transfected Protoplasts

1. Prepare 50 mL Falcon tubes with 20 mL of cold 240 mM CaCl₂ (4 °C) for each co-transfection.
2. Add the protoplasts within the B5 solution to the tubes by decanting.

3. Centrifuge for 10 min at $300 \times g$ and $4\text{ }^{\circ}\text{C}$ in a swing-out rotor.
4. Remove the supernatant with a fluid aspiration system down to ca. 1 mL.
5. Resuspend the pellet in the remaining supernatant and transfer the protoplast suspension into a 1.5 mL reaction tubes using Cell-Saver tips.
6. Centrifuge 30 s at $10,000 \times g$ and $4\text{ }^{\circ}\text{C}$.
7. Remove the supernatant using the fluid aspiration system and instantly freeze the protoplast pellet in liquid nitrogen.
8. Store the frozen protoplasts at $-80\text{ }^{\circ}\text{C}$ until use or thaw on ice for following protein extraction.

3.5 Extract Proteins from Protoplasts

1. Thaw the protoplasts on ice.
2. Add 750 μL of protein extraction buffer.
3. Resuspend the pellet by vortexing rigorously for a minimum of 30 s.
4. Centrifuge for 10 min at full speed on $4\text{ }^{\circ}\text{C}$ in a benchtop centrifuge.
5. Keep the reaction tubes on ice; the supernatant is used for the protein and reporter gene assays (*see Note 25*).

3.6 Reporter Gene Assays

In this assay the luciferase activity is quantified as a measure for transfection efficiency (*see Note 25*).

3.6.1 Luciferase Assay

1. Pipet 10 μL of the protoplast protein extracts to the wells of a white LUMITRAC 96-well microplate.
2. Adjust the instrument settings to luminescence detection.
3. Fill the syringe injector of the FLUOstar Optima microplate reader with 100 μL luciferase substrate solution for each sample and start the luminescence measurement (*see Note 26*).
4. Measure the produced light (relative light unit, RLU) during 10 s using the FLUOstar Optima microplate reader.
5. Calculate the specific luciferase activity LUC_i [RLU/s/ μg] for each co-transfection by dividing the measured light [RLU/s/ μL] by the protein concentration [$\mu\text{g}/\mu\text{L}$].

$$\text{LUC}_i = \frac{\frac{\text{RLU}}{\text{s} \times \mu\text{L}}}{\frac{\mu\text{g protein}}{\mu\text{L}}}$$

3.6.2 GUS Assay

Fluorometric analysis allows quantification of GUS activity. In the presence of GUS, MUG is hydrolyzed to the fluorescent product 4-methylumbelliferone (MU). After the reaction, total fluorescence is measured and product concentration is calculated based on a MU standardization curve.

1. Pipet 100 μL of the protein extracts into the wells of a black FLUOTRAC 200 96-well microplate.
2. Add 100 μL of 4-MUG substrate solution to each protein extract.
3. Pipet the MU dilution series to the microplate as well. This series is used to generate a MU standardization curve.
4. Set excitation to 365 nm and read the sample emission at 455 nm after 20, 40, and 60 min at 37 °C in the FLUOstar Optima microplate reader.
5. Determine the average change in measured MU fluorescence (ΔE_{455}) from 20 to 40 min, and from 40 to 60 min (*see Note 27*).
6. Calculate the line of best fit for the MU fluorescence in the MU dilution series (MU standardization curve). Determine the slope (m) of the MU standardization curve.
7. The specific β -glucuronidase activity GUS_i [pmol/min/mg] is determined by the formula:

$$\text{GUS}_i = \frac{\Delta E_{455} \times 1000 \frac{\mu\text{g}}{\text{mg}} \times \frac{200\mu\text{L}}{20\mu\text{L}}}{20\text{min} \times \frac{m}{\text{pmol}} \times \mu\text{g protein}}$$

The GUS activity measurement uses 20 μL of a 200 μL sample (consisting of 100 μL protein extract and 100 μL 4-MUG substrate solution).

The amount of protein [μg] in 100 μL protein extract is given.

3.7 Protein Concentration Measurement

To determine the protein concentration in the protein extract samples, a Bradford assay [13] is performed.

1. Pipet 10 μL of the protein extracts to the wells of a Nunc™ MicroWell™ 96-well microplate.
2. Pipet 10 μL of the BSA dilution series to the microplate as well. This series is used to generate a protein standardization curve.
3. Mix with 200 μL protein assay dye reagent.

4. Incubate for 5 min at 37 °C.
5. Measure OD₅₉₅ in the FLUOstar Optima microplate reader.
6. Calculate the protein concentrations in the extracts with help of the BSA dilution series.
7. Determine the amount of protein [μg] in 10 μL protein extract and calculate the protein concentration [μg/μL].

**3.8 Calculation
of the Normalized
Specific GUS' Activity
(See Note 28)**

1. We normally repeat each co-transfection experiment (with the same combination of plasmids) six times, with six independent (*i*) co-transfections with three different protoplast preparations, giving an “experimental block”. A “whole experiment”, including controls and all related experiments to answer a biological question, consists of several experimental blocks.
2. Calculate the average of all specific LUC_{*i*} values (LUC_{*M*}) from a whole experiment.

$$LUC_M = \frac{1}{n} \times \sum LUC_i$$

n: sum of all co-transfections in a whole experiment

3. For standardization, a specific correction factor *F_i* for each individual co-transfection experiment is determined by dividing LUC_{*M*} by the specific LUC_{*i*} value.

$$F_i = \frac{LUC_M}{LUC_i}$$

4. The standardized, corrected GUS activity (GUS_{*ki*}) is obtained by multiplying the specific correction factor *F_i* with the specific GUS activity GUS_{*i*}.

$$GUS_{ki} = F_i \times GUS_i$$

5. The average of specific GUS_{*ki*} values of an experimental block of six co-transfections is calculated as standardized GUS activity (GUS').

$$GUS' = \frac{1}{6} \times \sum GUS_{ki}$$

6. The standard deviation of GUS' (SD(GUS')) is determined by the formula:

$$SD(GUS') = \frac{1}{\sqrt{6(6-1)}} \times \sqrt{\sum (GUS_{ki} - GUS')^2}$$

4 Notes

1. It is also possible to use alternative reaction containers and instruments for the protein, LUC, and GUS measurements.
2. The transfection rate can be variable and depends on the plasmids used [5]. Generally, small vector sizes achieve higher transformation rates. The pBT vectors have a size-minimized backbone of 1989 bp containing colE1-ori and amp^r [3].
3. *cis*-acting elements often contain palindromic sequences. This has been interpreted as a reflection of the fact that DNA-binding proteins are in many cases active as dimers or tetramers, binding symmetrically to DNA matching the symmetry of the binding site [14]. Otherwise, several asymmetric *cis*-elements are known which are often recognized by heterodimeric *trans*-acting factors. Aiming to test the interaction of synthetic *cis*-elements and DNA binding proteins, we found it helpful to insert such binding elements in both orientations single copy, as a dimer, and as a tetramer fused to the GUS coding sequence in the pBT10GUS vector.
4. A promoter with appreciable activity in At7 protoplasts in the absence of added effectors has been reported. This reporter construct contains the region between -90 to +8 of the CaMV 35S promoter (-90 CaMV 35S promoter), including an activation sequence factor 1 binding site (positions -83 to -63) [15], fused to *uidA* encoding GUS. Binding domains (*cis*-elements) to analyze can be cloned immediately upstream of the -90 CaMV 35S promoter in pBT10GUS [4].
5. Addition of the LUC standardization plasmid in the co-transfection is used to determine specific LUC activity in a given sample to estimate the transfection rate (efficiency of transfection).
6. The *Photinus pyralis* luciferase (LUC) encoding ORF contains three silent point mutations which remove XbaI, EcoRI, and ClaI sites [16].
7. We observed strong differences between different lots of cellulase in terms of success in protoplasting. Each lot has to be tested and the amount of cellulase in the cellulase-mazerozyme solution has to be adjusted for successful protoplasting. Mazerozym is a multi-component enzyme mixture containing activities of pectinase, α amylase and hemicellulase; cellulase proteins hydrolyze 1,4- β -D-glucosidic bonds.
8. We are using a Multifuge 1s (Heraeus) centrifuge with a TTH400 swing-out rotor (Heraeus) or a Multifuge 3s-r (Heraeus) centrifuge with a TTH750 swing-out rotor (Heraeus). These centrifuges have the option to specify acceleration and deceleration in ramps from 1 to 9. These parameters have been optimized to protect the delicate living samples.

9. When harvesting the protoplasts from the Petri dishes, transfer the cell suspension into the Falcon tube by placing the brim of the Petri dish centrally over the tube and then decant carefully. Work above the inverted lid of the Petri dish to be able to recover eventually spilled protoplast suspension.
10. When adding solutions to the protoplasts, never just pour the liquid on top of the protoplasts, rather hold the tube at a flat angle and let the liquid slowly run along the side of the tube. Slow rotation of the tube helps dissolving solid pellets.
11. By washing with B5 floating medium, the living protoplasts are separated from the debris of broken cells. The high sugar concentration in the B5 floating medium causes floating of intact protoplasts.
12. When pipetting living protoplasts, always use pipetting equipment which reduces the shearing force, such as Cell-Saver tips or pipettes with a wide tip orifice.
13. When the floating protoplast suspension contains a high amount of cell clusters (caused by inefficient enzyme treatment for cell wall removal), it is possible to filter the protoplasts through a nylon net filter with 70 μm pore size. Although this filtering step drastically reduces the number of cells, the obtained protoplast suspension contains only the desired protoplasts.
14. Protoplasts number is determined using a hemocytometer.
15. The use of *dam*- plasmid DNA results in significantly reduced background activity, as shown for parsley protoplasts [17].
16. In our hands, high-concentrated ($>1 \mu\text{g}/\mu\text{L}$) *dam*- plasmid DNA in TE (10 mM Tris-HCl pH 8, 1 mM EDTA) stored at 4 °C is stable for more than 10 years, and can be used in co-transfection experiments without known limitations.
17. We recommend to generate and use a detailed pipetting plan.
18. Thawing of PEG solution at room temperature could take a few hours.
19. In order to ensure a simultaneous start of co-transfections for all experiments, do not mix the DNA and protoplasts at this point. If you are performing a small-scale experiment this is not as critical as if you are handling 40 co-transfection reactions at the same time.
20. When using a dispenser to add the PEG solution, it is recommended to hold the tubes or the rack with the tubes at a flat angle and pipet to the side of the tube.
21. Avoid to move the tubes during the incubation time.
22. We found that splitting the pipetting of $\text{Ca}(\text{NO}_3)_2$ into two steps worked well to achieve a stopping of all reactions at

around the same time. First pipet 1 mL to all the tubes and then add the remaining 4 mL.

23. Split the pipetting of the 7 mL in at least two steps and first pipet 1 mL to all tubes. If you add the whole volume of B5 solution at once, it might result in clumping of the protoplasts.
24. Almost horizontal incubation avoids contact of the liquid to the cap of the centrifugation tube, which might result in protoplasts stuck to the cap.
25. The luciferase assay should be performed immediately after protein extraction as the luciferase is degraded in the extract. The recommended order in which the following measurements have to be performed is (1) luciferase assay, (2) GUS assay, and (3) protein concentration measurement.
26. The luminescence measurement should immediately start after the addition of luciferase substrate solution. We implemented a program on the FLUOstar Optima microplate reader that adds 100 μ L of luciferase substrate solution to the samples immediately before the luminescence is measured.
27. Only assays with linear increase of the E_{455} values are taken into account.
28. We recommend the generation and use of an *Excel* sheet for the calculations.

References

1. Fields S, Song O (1989) A novel genetic system to detect protein-protein interactions. *Nature* 340:245–247
2. Dobi KC, Winston F (2007) Analysis of transcriptional activation at a distance in *Saccharomyces cerevisiae*. *Mol Cell Biol* 27: 5575–5586
3. Weisshaar B, Armstrong GA, Block A, da Costa e Silva O, Hahlbrock K (1991) Light-inducible and constitutively expressed DNA-binding proteins recognizing a plant promoter element with functional relevance in light responsiveness. *EMBO J* 10:1777–1786
4. Jin H, Cominelli E, Bailey P, Parr A, Mehrstens F, Jones J, Tonelli C, Weisshaar B, Martin C (2000) Transcriptional repression by AtMYB4 controls production of UV-protecting sunscreens in *Arabidopsis*. *EMBO J* 19:6150–6161
5. Sprenger-Haussels M, Weisshaar B (2000) Transactivation properties of parsley proline rich bZIP transcription factors. *Plant J* 22(1):1–8
6. Baudry A, Heim MA, Dubreucq B, Caboche M, Weisshaar B, Lepiniec L (2004) TT2, TT8, and TTG1 synergistically specify the expression of *BANYULS* and proanthocyanidin biosynthesis in *Arabidopsis thaliana*. *Plant J* 39:366–380
7. Zimmermann IM, Heim MA, Weisshaar B, Uhrig JF (2004) Comprehensive identification of *Arabidopsis thaliana* MYB transcription factors interacting with R/B-like BHLH proteins. *Plant J* 40:22–34
8. Stracke R, Favory J-J, Gubler H, Bartelniewöhner L, Bartels S, Binkert M, Funk M, Weisshaar B, Ulm R (2010) The *Arabidopsis* bZIP transcription factor HY5 regulates expression of the *PFG1/MYB12* gene in response to light and ultraviolet-B radiation. *Plant Cell Environ* 33:88–103
9. Trezzini GF, Horrichs A, Somssich IE (1993) Isolation of putative defense-related genes from *Arabidopsis thaliana* and expression in fungal elicitor-treated cells. *Plant Mol Biol* 21:385–389
10. Stracke R, Jahns O, Keck M, Tohge T, Niehaus K, Fernie AR, Weisshaar B (2010) Analysis of production of flavonol glycosides-dependent flavonol glycoside accumulation in *Arabidopsis thaliana* plants reveals MYB11-

- MYB12- and MYB111-independent flavonol glycoside accumulation. *New Phytol* 188: 985–1000
11. Luehrsen KR, de Wet JR, Walbot V (1992) Transient expression analysis in plants using firefly luciferase reporter gene. *Methods Enzymol* 216:397–414
 12. Kawalleck P, Somssich IE, Feldbrügge M, Hahlbrock K, Weisshaar B (1993) Polyubiquitin gene expression and structural properties of the *ubi4-2* gene in *Petroselinum crispum*. *Plant Mol Biol* 21:673–684
 13. Bradford MM (1976) A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal Biochem* 72: 248–254
 14. Takeda Y, Ohlendorf DH, Anderson WF, Matthews BW (1983) DNA-binding proteins. *Science* 221:1020–1026
 15. Katagiri F, Lam E, Chua N-H (1989) Two tobacco DNA-binding proteins with homology to the nuclear factor CREB. *Nature* 340: 727–730
 16. Hartmann U, Valentine WJ, Christie JM, Hays J, Jenkins GI, Weisshaar B (1998) Identification of UV/blue light-response elements in the *Arabidopsis thaliana* chalcone synthase promoter using a homologous protoplast transient expression system. *Plant Mol Biol* 36:741–754
 17. Tovar Torres J, Block A, Hahlbrock K, Somssich IE (1993) Influence of bacterial strain genotype on transient expression of plasmid DNA in plant protoplasts. *Plant J* 4:587–592

Selecting Hypomethylated Genomic Regions Using MRE-Seq

Elisabeth Wischnitzki, Kornel Burg, Maria Berenyi, and Eva Maria Sehr

Abstract

Here, we describe a method capable of filtering the hypomethylated part of plant genomes, the so-called hypomethylome. The principle of the method is based on the filtration and sequence analysis of small DNA fragments generated by methylation-sensitive four-cutter restriction endonucleases, possessing ^(5me)CpG motifs in their recognition sites. The majority of these fragments represent genes and their flanking regions containing also regulatory elements—the gene space of the genome. Besides the enrichment of the gene space, another advantage of the method is the simultaneous depletion of repetitive elements due to their methylated nature and its easy application on complex and large plant genomes. Additionally to the wet lab procedure, we describe how to analyze the data using bioinformatics methods and how to apply the method to comparative studies.

Key words Plant, Hypomethylome, Gene space, DNA methylation, Reduced representation libraries, MRE-seq, Non-model organisms, De novo assembly, Reference-based assembly, Comparative analysis

1 Introduction

Epigenetic modifications like DNA methylation influence gene expression without changing the underlying DNA sequence. The methylation of cytosines in the DNA is a reversible process common in plants. However, this methylation does not occur randomly in the genome but appears in pattern of methylated stretches in the genome. It has been observed that the gene space (genes and their flanking regions) is showing low methylation levels (hypomethylated) while cytosine methylation is found predominantly in repetitive elements (e.g. transposable elements). Not only the methylation of the gene body influences the expression but especially methylation pattern in the promoter regions has been associated with differential expression indicating a direct role of DNA methylation in the regulation of gene expression [1, 2]. Thus, investigating a plant's methylome (the methylated part of the genome) or hypomethylome (the non-methylated part of the genome) is an

increasingly popular strategy for understanding the genetic and environmental interactions behind biological processes. Methylation-sensitive restriction enzyme-based genome digestion used for creating reduced representation libraries allows the enrichment of gene space-related sequences by selecting for the hypomethylome [3–9]. The obtained libraries represent not only exons but also potential regulatory regions where regulatory sequences like transcription factor binding sites may reside. More precisely, the identified regions contain additionally introns and gene-flanking regions both up- and downstream. A combination of these libraries with next-generation sequencing (NGS) for their characterization is called MRE-seq (Methylation-sensitive Restriction Enzyme followed by sequencing) [10, 11].

Here, we detail a modified MRE-seq technique designed for the isolation and characterization of a plant's hypomethylome. The method is based on the restriction digestion of total genomic DNA with methylation-sensitive frequent cutter restriction endonucleases resulting in short DNA fragments of hypomethylated regions. PCR-based size selection, next-generation sequencing, and bioinformatics analysis of these short genomic fragments provide a comprehensive sequence representation and characterization of the hypomethylome [8, 9]. In this chapter, we describe both the filtration method and the bioinformatics procedure to analyze the data and give recommendations for performing a comparative analysis between samples.

Our method provides an easy tool to produce reduced representation libraries enriched for gene space omitting repetitive elements from small amounts of genomic DNA samples and opens the way for comparative analysis of genetic and epigenetic variation among genotypes or tissues even in a larger set of samples.

2 Materials

2.1 Plant Material

1. Leaves of the selected plant species were used for the analysis. We recommend fresh material but any material that will yield high-molecular-weight genomic DNA can be used (*see Note 1*).
2. The experiments presented are based on rice, *Oryza sativa* sp. *indica* variety SHZ-2A (seeds are kindly provided by R. Mauleon, IRRI International Rice Research Institute, Los Banos, Philippines) and Norway spruce, *Picea abies* (L.) H. Karst (twigs kindly provided by S. Schüller, Department of Forest Genetics, Austrian Research Centre for Forests, Vienna, Austria).

2.2 Buffers, Enzymes, Adapters, and PCR Primer

1. PCR grade water was prepared in the laboratory by UV irradiating Millipore Synergy generated ion exchanged water in Stratagene UV Stratalinker 2400 for 20 min and then sterile filtered.

2. CTAB lysis buffer: 140 mM Sorbitol, 220 mM Tris-HCl pH 8.0, 22 mM EDTA, 880 mM NaCl, 1 % Sarcosyl and 0.8 % CTAB.
3. Phenol-chloroform-isoamyl alcohol: 25:24:1 saturated with 10 mM Tris-HCl, pH 8.0, 1 mM EDTA.
4. Chloroform-isoamyl alcohol: 24:1 mixture of chloroform and isoamyl alcohol.
5. Isopropyl alcohol.
6. 70 % EtOH, ethanol diluted by distilled water.
7. Absolute EtOH, 100 % ethanol.
8. Liquid N₂.
9. 10 % SDS, Sodium Dodecyl Sulfate dissolved in distilled water.
10. 3.0 M NaOAc pH 5.2, 3 M sodium acetate, pH adjusted with acetic acid.
11. 5 mM Tris-HCl pH 8.0, Tris-HCl dissolved in distilled water, pH adjusted with NaOH.
12. 10 mM Tris-HCl pH 8.0.
13. (10×) restriction enzyme buffer.
14. PCR buffer (10×).
15. MgCl₂ (25 mM), Magnesium chloride dissolved in distilled water.
16. dNTP (20 μM), Mixture of the four deoxy nucleotide triphosphates.
17. BSA, Bovine Serum Albumin.
18. 20 mM EGTA.
19. Based on our experience the four-cutter restriction endonucleases HpaII, AciI, and Bsh1236I with recognition sites containing a CpG motif as methylation site provide the best results for the isolation of the hypomethylome of plant genomes [9] (*see* Table 1).
20. T4 DNA ligase (5 U/μL) and buffer.

Table 1
Restriction enzymes

Restriction enzyme	Cut site	End	Buffer	°C	Source
HpaII	C ^{5me} CGG	5'GC	Neb1	37	New England Biolabs
AciI	C ^{5me} CGC	5'GC	Neb3	37	New England Biolabs
Bsh1236I	CG ^{5me} CG	CG blunt	Neb4	37	Thermo Fischer Scientific
PmeI	GTTTAAAC	A blunt	Neb4	37	New England Biolabs

Table 2
Adapter and PCR primer

	Designation	Sequence of oligo^a	Feature
Adapter A	PmeI_CGWA	5' GCACGACTGTTTAAA	
Adapter B	PmeI_CGB	5p' CGTTTAAACAGTCGT	5' Phosphorylation
Adapter B blunt	PmeI_CGBlunt	5p' TTTAAACAGTCGT	5' Phosphorylation
PCR primer	PmeI_CG17	5' CACGACTGTTTAAACGG	

^aThe adapter sequences are designed to not restore the original restriction site (*see Note 5*)

21. HotStar Taq DNA Polymerase.
22. RNase A: 100 mg/mL diluted in water.
23. Proteinase K: 20 mg/mL purchased as liquid.
24. Bal3126. Adapters and PCR primer are listed in Table 2.

2.3 Equipment

1. Retsch mill: Retsch MM301, 25 mL jar, 15 mm steel ball.
2. High speed centrifuge, e.g. Sorvall RC6, SS34 rotor.
3. Eppendorf centrifuge, e.g. Eppendorf Centrifuge 5415D.
4. Vortex mixer, e.g. Scientific Industries Vortex Genie.
5. Thermocycler MJ research.
6. Agilent Bioanalyzer.
7. Glass rod; 1–1.5 mm diameter glass capillary pipet with melted, closed tips.
8. –20 °C freezer.
9. Waterbath.
10. Micropipettes 2, 20, 200, 1000 µL.

2.4 Disposables

1. 35 mL Polyallomer Nalgene Conical Oak Ridge centrifuge tubes.
2. 500 µL, 1 mL, and 2 mL Eppendorf tubes.

2.5 Recommended Analysis Software

The recommended tools (*see Table 3* and Subheadings 3.6–3.9) reflect our experience and are not exclusive. They can be exchanged with appropriate alternatives depending on the data, the analysis, or new technological developments. Most of the tools are command-line based and require a Unix system. In general the analysis will be much faster, if the separate steps can be run in parallel. For the de-novo assembly, we recommend a compute cluster with a high amount of RAM.

Table 3
Recommended analysis software

Software	References	Analysis
CutAdapt	[16]	Removal of adapter sequences (<i>see</i> Subheading 3.6)
Trimmomatic	[17]	Removal of low-quality and short sequences (<i>see</i> Subheading 3.6)
Bowtie2	[18]	Removal of other sequences (<i>see</i> Subheading 3.6) Reference-based assembly (<i>see</i> Subheading 3.7) De novo assembly (<i>see</i> Subheading 3.7) De novo assembly (<i>see</i> Subheading 3.7) Mixed approach (<i>see</i> Subheading 3.7) Comparative sequence analysis (<i>see</i> Subheading 3.9)
samtools	[21]	Reference-based assembly (<i>see</i> Subheading 3.7) De novo assembly (<i>see</i> Subheading 3.7) Mixed approach (<i>see</i> Subheading 3.7) Comparative sequence analysis (<i>see</i> Subheading 3.9)
Bedtools	[22]	Reference-based assembly (<i>see</i> Subheading 3.7) De novo assembly (<i>see</i> Subheading 3.7) Mixed approach (<i>see</i> Subheading 3.7) Annotation (<i>see</i> Subheading 3.8) Comparative sequence analysis (<i>see</i> Subheading 3.9)
Trinity	[23, 24]	De novo assembly (<i>see</i> Subheading 3.7) Mixed approach (<i>see</i> Subheading 3.7)
FLASH	[36]	De novo assembly (<i>see</i> Subheading 3.7)
Blast	[25, 26]	Mixed approach (<i>see</i> Subheading 3.7) Annotation (<i>see</i> Subheading 3.8) Comparative sequence analysis (<i>see</i> Subheading 3.9)
InterproScan	[27, 28]	Annotation (<i>see</i> Subheading 3.8)
Blast2GO	[29, 30]	Annotation (<i>see</i> Subheading 3.8)
cd-hit	[31, 32]	Comparative sequence analysis (<i>see</i> Subheading 3.9)
IGV	[33, 34]	Comparative sequence analysis (<i>see</i> Subheading 3.9)

3 Methods

The method consists of the following different procedures which are further detailed in the Subheadings: 3.1. Preparation of high-molecular-weight genomic DNA, 3.2. Adapter preparation, 3.3. Enzymatic digestion of genomic DNA and ligation of adapters, 3.4. Amplification of adapter ligated DNA, 3.5. Sequencing, 3.6. Data processing and quality control of the raw reads, 3.7. Identification of hypomethylated regions detailing the different assembly strategies, 3.8. Different annotation methods. An additional Subheading 3.9 is focusing on the application of comparative analysis between different samples.

3.1 Preparation of High-Molecular-Weight Genomic DNA

Genomic DNA from *Oryza sativa* and *Picea abies* was prepared with a modified protocol from Janice Keller and Ian Bancroft [12].

1. Grind plant material (0.5 g) to fine powder in liquid N₂ with steel balls in a Retsch mill (Retsch MM301, 25 mL jar, 15 mm steel ball).
2. Melt the frozen powder in 7 mL of 65 °C CTAB lysis buffer and 1 mL of 10% SDS (35 mL Polyallomer Nalgene Conical Oak Ridge centrifuge tubes).
3. Incubate samples at 65 °C in a waterbath for 30 min with occasional vigorous vortex shaking.
4. After incubation extract the samples twice with 9 mL of chloroform-isoamyl alcohol (24:1) and centrifuge at 8000×g for 10 min (Sorvall RC6, SS34 rotor).
5. Transfer the upper aqueous phase into a new tube (same as above).
6. Precipitate with 0.8 volumes of isopropyl alcohol.
7. Incubate for 10 min at room temperature.
8. Centrifuge the samples at 15,000×g for 20 min.
9. Wash pellets with 70% EtOH, dry and dissolve in 600 µL 5 mM Tris-HCl pH 8.0 containing 300 ng of RNase A.
10. Incubate samples at 37 °C for 1 h.
11. Add 30 µg of Proteinase K and incubate at 37 °C for an additional hour.
12. Extract the samples twice with equal volumes of phenol-chloroform-isoamyl alcohol and twice with equal volumes of chloroform-isoamyl alcohol (24:1) in 2 mL Eppendorf tubes.
13. Precipitate the extracted samples by adding 0.1 volumes of 3 M NaOAc pH 5.2 and 2 volumes of absolute EtOH.
14. Roll out the high-molecular-weight genomic DNA with a glass rod.
15. Wash the samples with 70% EtOH and let it dry.
16. Dissolve the DNA overnight in 100 µL of 5 mM Tris-HCl pH 8.0.

3.2 Adapter Preparation

1. Dissolve the lyophilized adapters and primers at 100 µM concentration in sterile PCR grade water.
2. Dilute an aliquot of both A and B (or B blunt) adapters (see Table 2) to 10 µM concentration.
3. Mix in a 1:1 ratio.
4. Anneal in thermocycler by heating the mix to 95 °C for 5 min and subsequently cooling it stepwise by 5 °C/5 min to 25 °C.
5. Store the annealed adapters at -20 °C aliquoted in 500 µL Eppendorf tubes (see Note 2).

6. The oligo pairs PmeI_CGWA and PmeI_CGB oligos are used for HpaII and AciI enzyme site adapters, while for Bsh1236I enzyme site PmeI_CGWA and PmeI_CGBlunt oligos are used (*see* Table 2).

3.3 Digestion and Ligation of Genomic DNA

1. Use a single reaction for the restriction digestion and adapter ligation of the genomic DNA. The 50 μ L reaction mix contains 300 ng purified genomic DNA (*see* Note 3), 5 μ L (10 \times) restriction enzyme buffer (*see* Table 1), 4 μ L (10 μ M) annealed adapter, 40 units of restriction enzyme, 10 units of T4 ligase, and fill up with water to 50 μ L.
2. Incubate the digestion-ligation reaction overnight at 37 $^{\circ}$ C.
3. After heat inactivation at 65 $^{\circ}$ C for 20 min in the thermocycler, dilute the samples 1:1 with 50 μ L PCR grade water and extract subsequently by equal volumes of phenol–chloroform–isoamyl alcohol, then by chloroform–isoamyl alcohol (24:1).
4. Precipitate the extracted DNA with 2.5 volumes of absolute EtOH in the presence of 0.3 M NaOAc pH 5.2, wash with 70% EtOH. Dry and dissolve in 100 μ L 5 mM Tris–HCl pH 8.0.

3.4 Amplification of the Adapter-Ligated Genomic Fragments

1. Amplification of the adapter-ligated DNA (*see* Note 4): For Illumina sequencing, attain fragments by PCR amplification of the restriction-digested and adapter-ligated genomic DNA samples. Use 5 μ L of digested and adapter-ligated DNA (about 15 ng) and 4 μ L of 10 μ M amplification primer PmeI_CG17 in a 50 μ L PCR reaction containing 5 μ L PCR buffer (10 \times), 1 μ L MgCl₂ (25 mM), 1 μ L dNTP (20 μ M), 0.5 μ L HotStar Polymerase (2.5 units) and add PCR grade water to 50 μ L. Initialize the PCR at 95 $^{\circ}$ C for 15 min followed by 25–30 cycles of 95 $^{\circ}$ C 30 s/55 $^{\circ}$ C 40 s/72 $^{\circ}$ C 50 s and finish by 72 $^{\circ}$ C for 5 min. The exact number of cycles has to be evaluated experimentally (*see* Note 6). The selected three restriction endonucleases are performing equally well for the filtration (*see* ref. 9 and Fig. 1). Note that the sizes of the predominant fragments (bands) are characteristic both for the restriction enzyme (*see* Fig. 1) and for the analyzed plant genome (not shown) and are reproducible (*see* Fig. 2).
2. Removal of the adapter sequences: To increase the length of the usable sequence information, the majority of the adapter sequence should be removed. This can be achieved by PmeI digestion, because its rare cut site GTTTAAAC is present in the adapter sequence. Digest 1 μ g of the PCR amplicates with PmeI (NEB) in NEB4 buffer, in two steps under the presence of 100 ng/ μ L BSA. Perform the first digestion in a 50 μ L reaction volume, containing 100 U PmeI enzyme on 37 $^{\circ}$ C for 2 h followed by a subsequent volume increase to 100 μ L in 1 \times NEB4 buffer including

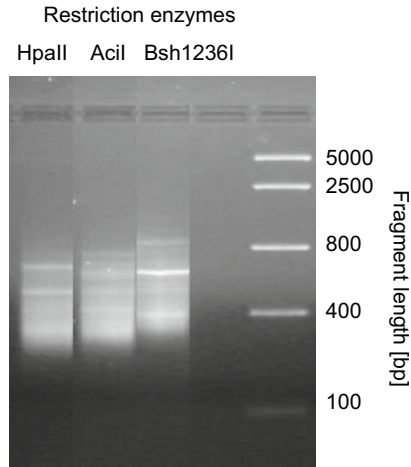


Fig. 1 The results of the size-selective amplification of *Oryza sativa*-digested genomic DNA show characteristic predominant fragment sizes for the different restriction enzymes (0.8 μ M PmeI_CG17 primer concentration; 27 PCR cycles)

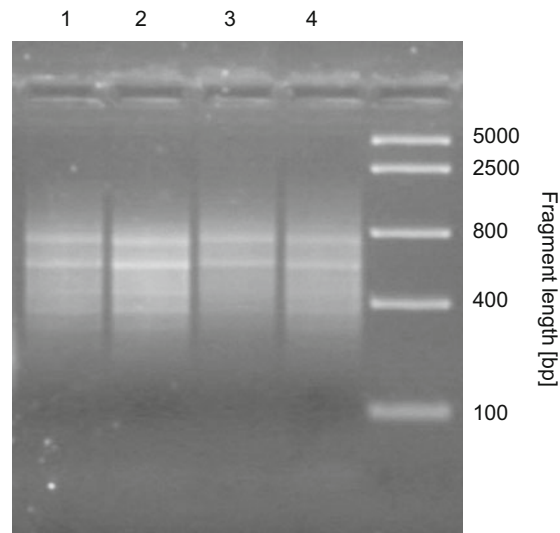


Fig. 2 Filtrated fragments of genomic DNA show reproducible predominant fragment sizes. The results for four different *Picea abies* samples representing different phenotypes are shown

additional 50 U PmeI and incubate for additional 2 h. Finally, stop the reaction at 65 °C for 20 min. Purify the samples (100 μ L) with 100 μ L phenol–chloroform–isoamyl alcohol (25:24:1), then twice with the same volume of chloroform–isoamyl alcohol. Pipet the upper phase into a new 1.5 mL Eppendorf tube, and precipitate the reaction by adding 0.1 volumes of 3.0 M NaOAc pH 5.2 and

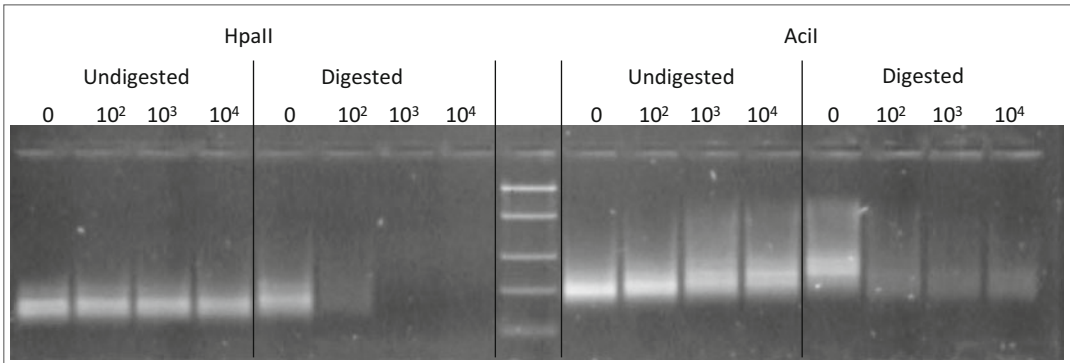


Fig. 3 Test for the efficiency of adapter removal. Samples after PmeI digestion show strongly reduced or no amplification with the PmeI_CG17 primer due to the removed sequence (Digested). In comparison in undigested samples the amplification is still possible for all dilutions (Undigested). All samples were diluted up to 10^4 fold and subsequently amplified with the PmeI_CG17 primer

2.5 volumes of absolute EtOH for 1 h at $-20\text{ }^{\circ}\text{C}$. Centrifuge the precipitated DNA at $4\text{ }^{\circ}\text{C}$ with $16,000\times g$ in an Eppendorf centrifuge (full speed) for 20 min then wash with 70% EtOH. Centrifuge for 10 min at $15,000\times g$. Discard the supernatant and let the pellet dry at room temperature. Dissolve in 5 mM Tris-HCl pH 8.0 to reach a concentration of about 200 ng/ μL . Store at $-20\text{ }^{\circ}\text{C}$. Test the effectivity of the adapter removal (*see Note 7* and Fig. 3).

3.5 Sequencing

1. During our analysis we tried different sequencing technologies (Illumina HiSeq2000, MiSeq and Ion Torrent) and discovered that the method is suitable for different sequencing technologies as the obtained results were comparable [9]. However, the data created with the Ion Torrent technology are similar to those created by the Illumina technology, but show a much lower coverage. Therefore, the decision for a specific technology is more a matter of availability and personal preferences. However, we recommend to base the decision on the size of the expected fragments (for the expected read length) and on the desired coverage (for the technology). We have been able to enrich for the gene space and deplete repetitive sequences in *Picea abies* with a rather low genome wide coverage of $0.1\times$. This represents a theoretical coverage of the gene space of $\sim 4\times$ as only about 2.4% of the spruce genome is described as gene and gene-like sequences [13]. Similar results were also achieved in *Crocus sativus* [9]. For more information about the recommended coverage *see Note 8*.
2. The sequencing technology of Illumina requires the fragments to be sufficiently different in the first few bases to be able to distinguish the so-called sequence clusters within one lane. If the

sequences are identical, the technology cannot distinguish between actually different fragments and the sequencing run will yield no results. Due to the amplification adapter and the identical genomic cut site, this is the case for the isolated fragments, even if the amplification adaptor is removed. However, there are different methods to approach this technical issue. (1) Adding PhiX fragments as recommended by Illumina for amplicon sequencing did not yield satisfying results in our experiments. A test showed that we had to add 30% of PhiX in order to receive satisfying data. This presents a practical loss of information and can be circumvented by applying one of the next proposed solutions. (2) Dilute the samples with random genomic fragments properly sized to Illumina sequencing instead of PhiX. This way an additional unfiltered genomic reference dataset is obtained (*see Note 9*). (3) Removing the amplification adaptor and the genomic cut site with an exonuclease will result in random sequence ends of the fragments suitable for Illumina sequencing. This method can be performed additionally or instead of the removal of the amplification adapter in the protocol. In our hands, it was successfully applied to samples from the saffron crocus [9]. For the removal of a few base pairs at each end of the fragments, the exonuclease Bal31 was used for the digestion (<https://www.neb.com/products/m0213-nuclease-bal-31>), which has already been applied in a number of studies for the controlled length reduction of linear double-stranded DNA, including studies focusing on telomere truncation [14, 15]. Bal31-driven shortening of the fragments, however, needs optimization for the analyzed samples. Set a 50 μL reaction containing 2 μg of PCR amplified fragments, 1 \times Bal31 reaction buffer and 1 U of Bal 31 enzyme. Incubate at 30 $^{\circ}\text{C}$ and take samples of 5 μL at 15, 30, 60, 120, etc. seconds, up to 6 min. Mix the removed samples immediately with 5 μL of 40 mM EGTA pre-warmed to 65 $^{\circ}\text{C}$ and incubate for 10 min to stop the reaction. To visualize the progression of the shortening, the samples are diluted 1:5 with 10 mM Tris-HCl pH 8.0 and loaded to an Agilent Bioanalyzer. A size shift of about 30–40 bps is expected for a proper removal of the uniform parts at the fragment ends. Using the results of the time-course experiment, do estimate the necessary digestion time to reach this goal. The size reduction is fragment length dependent, since longer fragments are stronger affected. Therefore, we highly recommend separate optimization for each studied set of samples/species. To prepare fragments for the sequencing set up the same reaction and use the identified time for optimal shortening.

3.6 Data Processing

All sequence reads should be cleaned following the recommended procedure. Especially if the method is used for comparative studies, all samples should be treated in the same way in order to

guarantee high-quality data and comparability of the datasets. Some de novo assembly tools include preprocessing but we recommend filtering the reads before further analysis steps. Based on our experience this yields the best results and presents the best basis for comparative analyses.

1. Removal of adapter sequences: In case the digestion of the adapter sequences was not complete, there might be still parts of it present in the dataset. These smaller parts of the adapter sequence should be removed from the read sequences as they are artificially added during the procedure and do not reflect the plant genome. Any available tools that can handle not only a complete adapter sequence but also its substrings can be used for this process, e.g. CutAdapt [16]. This step is still recommended even if the adapter removal step was replaced by the exonuclease digestion (*see* Subheading 3.5).
2. Removal of low-quality regions and short sequences: Low-quality regions with low base call accuracy can affect the mapping and de novo assembly of the fragments by introducing wrong information. Also very short reads should be removed as they may not be mapped uniquely to the reference or may affect the assembly. Therefore, the removal of those sequences is a necessary step to ensure the quality of the results. Any available preprocessing tool is able to perform this procedure, e.g. Trimmomatic [17]. The threshold might depend on the data and the analysis. We recommend applying a Q-value threshold of 30, representing a base call accuracy of 99.9%, and a minimal length threshold of 50 bp. The latter could be set lower if other sequencing methods are used producing shorter reads.
3. Removal of other sequences (optional, *see* Notes 10 and 11): Depending on the aims of the analysis it is useful to remove certain sequences prior to further analysis steps, e.g. repetitive elements, ribosomal, chloroplast, or mitochondrial sequences. The filtering can be performed with different tools. We use bowtie2 [18] for filtering the processed reads. We recommend treating each read of a pair separately and applying option: “--un”. This will result in a file containing all reads that did not create a sufficient alignment. As reference general databases can be used like, e.g. REdat for repetitive elements [19] or customized datasets depending on your data and research question. If a reference of the chloroplast or the mitochondrion is lacking for the studied species, we recommend using the sequences of close relatives and adjust the threshold to allow more variation, if necessary (e.g. option “--local”).

3.7 Identifying Hypomethylated Regions

The procedure for the identification of hypomethylated regions depends on the studied species. More precisely the deciding factor

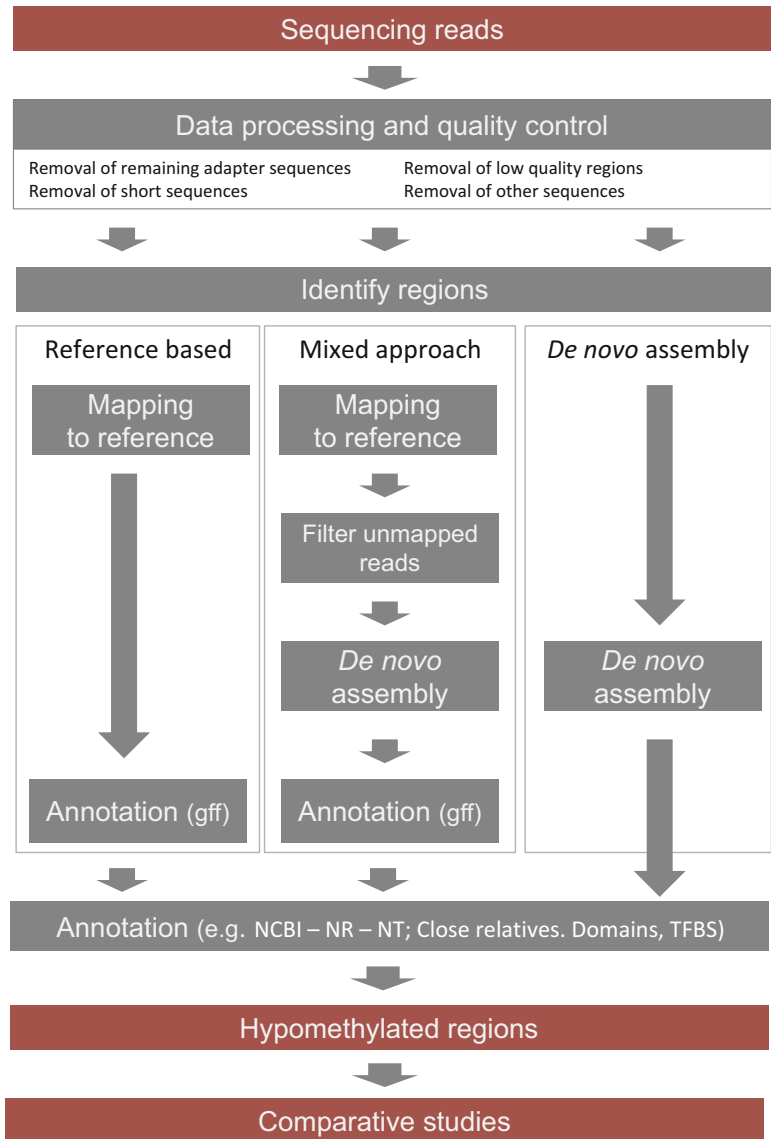


Fig. 4 Bioinformatics workflow for the identification and analysis of hypomethylated regions

is whether a reference sequence is available or not. If a reference sequence is available, a reference-based assembly can be performed. If no reference is present, a *de novo* assembly approach has to be applied. A combination of reference-based and *de novo* assembly is a favorable option to yield the best results for the analyzed dataset. Those different analysis procedures are detailed in the following sections (*see* Fig. 4).

1. Reference-based assembly: The reference-guided assembly is performed by assigning all high-quality reads to the reference

sequences. There are different read mapping tools available that are suitable for this procedure. The choice should depend on the analysis and is personal preference. For a review *see* [20]. We recommend bowtie2 with default settings [18] as it allows a bit more variation in the alignment as other tools which is preferable for comparative studies.

Furthermore, we apply additional filter for retaining reliable results and to further reduce the possibility for false-positive sequences. We recommend selecting only regions which are hit by at least five reads (bowtie2 [18], samtools [21], bedtools [22], developed perl-scripts). This threshold will depend on the analysis and the datasets but will increase the reliability of the results.

2. De novo assembly: The de novo assembly for each dataset is performed using Trinity [23, 24] (*see* **Note 12**) with minimal contig length of 100 bp. This parameter can be adjusted if necessary based on the filtered fragment length and obtained read length, e.g. if only 300 bp reads were obtained this can be set higher. For all technologies and datasets we tested, this value was sufficient. The analysis of various tools showed that Trinity yields the best results. For the purpose of our method—to enrich for the gene space—an assembly method focusing on the transcriptome is best suited. Based on our experience Trinity also performs well for fragments not derived from the gene body [9].

The resulting contigs should be evaluated by mapping the high-quality reads used for the assembly to the assembled sequences using, e.g. bowtie2 [18] and only contigs consisting of at least five reads should be retained similar to the reference-based assembly.

3. Mixed approach: Whether a reference or de novo approach is applied depends on the organism. If a genome sequence is available apply the reference-based approach first. If a high fraction of reads could not be aligned to the reference genome, use the de novo approach. However, a mixed approach is also possible or even recommendable in some cases. Perform a reference-based assembly first and extract the reads that could not be aligned to the reference (similar to the data processing step “Removal of other sequences” using, e.g. bowtie2 option `-un` [18]) and perform a de novo assembly with this subset to identify specific sequences that were either not present in the reference or differ too much from the reference to be aligned properly. The de novo sequences can afterward be compared to the reference using blast [25, 26] to identify the potential locations in the reference that differ between your samples and the reference [9] (*see* **Note 13**). De novo-assembled contigs that could not be located in the genome should be subjected to separate annotation to determine their origin.

3.8 Annotation

1. In the case a reference-based assembly was performed, the coordinates of the identified regions should be compared with the available annotation (usually available in gff or gff3 format) using, e.g. bedtools [22]. An additional similarity search is also recommended (*see Note 14*).
2. The annotation of the de novo-assembled contigs can be performed by similarity searches using blast [25, 26] against the databases NR and NT from NCBI (<http://www.ncbi.nlm.nih.gov/>). If close relatives have been sequenced, we recommend a separate blast run against those as well (*see Note 14*).
3. Further annotation using, e.g. known transcription factor binding sites (*see Note 15*), InterproScan [27, 28] or Blast2GO [29, 30] can additionally provide useful information and is highly recommended.

3.9 Comparative Sequence Analysis

The strategies for comparative studies depend again on the studied species and the availability of a reference. We present some general suggestions but the exact downstream workflow will depend on the specific question to be answered (*see Note 16*). However, regardless which method is used it is important to treat all samples subjected to the comparative analysis the same way to ensure comparable results.

1. In case a reference has been published, all samples should be aligned separately to this reference and annotated as described for the reference based assembly. The coordinates of the identified regions can be compared between the samples using, e.g. bedtools [22] to identify regions that are unique to samples or conditions (e.g. `intersectbed -v`). This can also be applied to identify parts of larger hypomethylated regions that are differentially methylated as it might occur if a regulatory element is methylated in the promoter in one sample but not in the other (e.g. `subtractBed`).
2. For samples with no reference there are several possibilities. Either all reads are used to create a reference hypomethylome by de novo assembly which is further used as common reference for a reference-based assembly and the subsequent analysis, as described above. Or the samples are de novo-assembled separately and clustered using, e.g. `cd-hit` [31, 32] to identify unique sequences. For the identification of overlapping sequences, we recommend post-processing the clustering results and creating separate alignments for the exact identification of the differentially methylated parts. Also similarity searches using, e.g. blast [25, 26] are recommended to identify similar sequences. Here again further global alignments might be recommendable to study differences in more detail.

3. For a mixed assembly method as described above, we recommend to add the additional de novo-assembled contigs to the previously known reference. This extended reference will present a more complete basis for the identification of differentially methylated regions. We recommend performing a reference-based analysis with this extended reference as described above.
4. Visualization: If a common reference is present the read alignments can be prepared using, e.g. samtools [21] as bam and bai files and visualized together with the bed files indicating the locations of the identified regions (output of the analysis described in the previous steps) with, e.g. the Integrated Genome Viewer (IGV) [33, 34].

4 Notes

1. Select the DNA isolation method suitable for the chosen plant species, resulting in high-molecular-weight DNA with OD 260/280 nm > 1.8 and 260/230 nm ratio > 2 to warrant the proper digestibility by the restriction endonuclease. The described method was successfully applied in rice, Norway spruce, banana, sweet potato, saffron, pepper-bark tree and in the 1RS chromosome arm of rye. Store high-molecular-weight genomic DNA at 4 °C to reduce the fragmentation by frequent freezing-melting.
2. Aliquot the annealed adapters in amounts, which may be used up in one ligation reaction and store at -20 °C. Use a single tube for a single restriction-ligation reaction and always discard the rest of the adapters. The annealed adapters are stable at -20 °C for at least 6 months.
3. The amount of DNA used in the reaction is dependent on the genome size of the selected species. The larger the genome the more input DNA is needed. 300 ng of input DNA is equivalent to about 15,000 Norway spruce or about 6×10^5 rice genomes. It is possible to perform the method with less than 20 ng genomic DNA, but think of the genome size! Do not use more than 300 ng DNA per digestion ligation reaction because of the increasing possibility of partial digestion. Always use high concentration restriction enzymes. Think of inhibitory effects of, e.g. glycerine if too much volume of the restriction enzyme is used.
4. Optimization of the PCR: Short fragment amplification is not favored by use of a single adapter and primer, since it allows looping of the short fragments, thereby prohibiting their amplification. However, this handicap may be overcome by

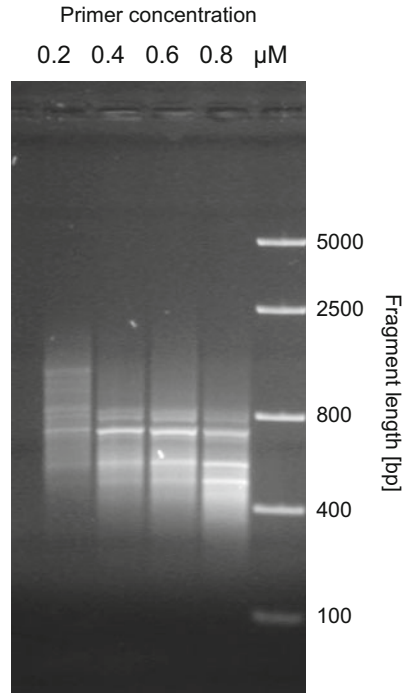


Fig. 5 Amplification of adapter-ligated genomic DNA of *Oryza sativa* produces fragments between 200 and 800 bp for Pmel_CG17 primer concentrations of 0.6–0.8 μM (digestion with 10 ng HpaII restriction enzyme, amplification is shown for different concentrations between 0.2 and 0.8 μM Pmel_CG17 amplification primer)

adjusting the amplification primer concentration [35]. Therefore, the optimization of primer concentration to obtain properly sized PCR products for the Illumina HiSeq sequencing is necessary. The results in Fig. 5 show that increasing the primer concentration in the PCR reaction favors the amplification of shorter amplicons. In the presented case 0.8 μM primer concentration resulted in approx. 200–800 bp fragments, fulfilling the size requirements of both, the hypomethylome filtration and the Illumina HiSeq sequencing.

5. We have observed that the amount of fragments derived from transposable elements can be further decreased by digesting the PCR-amplified fragments with the same restriction endonuclease. The previous methylated sites become unmethylated after PCR amplification and thus digestible. The designed adapters do not restore the original cut site and are therefore not affected. After digestion, the intact fragments—having adapter sequences on both ends—can be further amplified by PCR. Note that this procedure will not only reduce the amount of fragments derived from transposable elements but also affect methylated sites in, e.g. genes. However, this is only a minor fraction (unpublished data).

6. Do not over-amplify the PCR. Make reactions with different cycle number to evaluate the optimal amplification. The high abundant fragments should not be showing up obtrusively on the agarose gel. We recommend parallel amplifications to get enough material for the subsequent sequencing.
7. It is important to remove as much adapter sequences from the fragments as possible before sequencing (see also Subheading 3.5). Test the level of adapter removal with re-amplification of the digested samples. Dilute the digested samples with 5 mM Tris-HCl pH 8.0 to 100, 1000 and 10,000 fold and amplify as before. Compare to non-digested samples (*see* Fig. 3). Repeat the digestion if necessary.
8. Recommended coverage: Based on the results obtained in rice we performed an *in silico* simulation to estimate the minimal coverage necessary to identify the hypomethylome of the whole genome. Reads were randomly selected to represent different coverage thresholds and allocated to the genome sequence. The data show that with a genome coverage of 3× still 92% of the regions were identified. This corresponds to a coverage of the gene-space of ~6× [9]. Therefore, we recommend a minimal coverage of 6–7× of the gene space.
9. The dilution with random genomic fragments instead of adding PhiX sequence provides the additional advantage of having an unfiltered dataset as reference. This may be the preferable alternative if such a reference dataset is advantageous for the analysis. Furthermore, this option is well suited for sequencing on an Illumina HiSeq machine which provides sufficient coverage for the filtered fragments even if 30–50% of the output derived from the random fragments. However, the use of machines with a lesser sequence output (e.g. MiSeq) is not recommended.
10. Mitochondrial and chloroplast DNA are most likely still present in the data. Depending on the questions you want to answer, it is advantageous to remove those sequences before further analysis. However, depending on the used enzyme, sequences from genomic chloroplast and mitochondrial regions can still be present even after removal of the mitochondrial and chloroplast sequences. This can be due to alternative methylation mechanisms or DNA modification which causes the enzyme to cut despite the methylation. This issue has been discussed in [9].
11. Removal of PhiX-sequence from Illumina datasets: We noticed that in most sequencing datasets a limited amount of PhiX-DNA is present even without adding it specifically. This varies between 0.1 and 5% depending on the dataset but is on average about 1%. We recommend removing read sequences derived from the PhiX-genome from the datasets before further analysis (see Subheading 3.6).

12. Computational power necessary for a de novo assembly: Depending on the amount of reads integrated into the analysis the RAM demand of Trinity or any other de novo assembler, might be rather high. According to the manual Trinity needs as a basic recommendation approximately ~1G of RAM per ~1 M pairs of Illumina reads. Complex datasets might require even more (<https://github.com/trinityrnaseq/trinityrnaseq/wiki/Trinity-Computing-Requirements>). Keep in mind that the assembly process will take some time. To reduce time and RAM demand, we recommend using the “--normalize_reads” parameter. Furthermore, paired reads can be tested prior to assembly whether they overlap and could be combined into one longer sequence using, e.g. FLASH [36].
13. Comparing de novo-assembled contigs to genomes using blast: Be aware that similar hits can occur especially in polyploid genomes or genomes with large-scale duplication events in the past. In the rice genome about 10% of the de novo-assembled contigs produced multiple occurrences with identical hit-statistics [9]. Also large gene families with small sequence divergence within the family can lead to multiple similar blast hits. For a more detailed analysis, we recommend looking at the read alignment for those regions and try to discriminate whether different haplotypes, gene copies, or different family members are present.
14. Additional annotation of identified regions: Running a similarity search against NR and NT is always recommended even if an annotation is available for the used reference. A lot of genes are annotated as “hypothetical protein” or with similar rather uninformative descriptions. This can provide at least a bit more information about the gene function. Also the available information might have changed since the annotation of the reference and new data is available. We recommend an additional search with an e-value threshold of $1e-5$ to obtain additional information about the identified regions.
15. Identifying known regulatory elements in sequence data produces a high amount of false-positive data, depending on the analyzed elements. The majority of transcription factor binding sites are very short and occur everywhere in the genome simply by chance. Therefore, these data should be additionally confirmed or otherwise treated with care. However, if those elements are identified in differentially methylated regions this might give important hints to differences in the regulation of the affected gene.
16. Comparative studies: For the detection of small sequence differences between samples (e.g. SNPs, InDels, SSRs, etc.) that may hint to differential methylation of alleles between the samples or that may cause differential methylation, we recommend

performing a de novo assembly for each dataset and compare the interesting sequences separately. This approach may identify differences, which may have been excluded in a pure reference based analysis.

Acknowledgements

This work was supported by the AIT Austrian Institute of Technology GmbH.

References

- Rabinowicz PD, Citek R, Budiman MA et al (2005) Differential methylation of genes and repeats in land plants. *Genome Res* 15:1431–1440. doi:10.1101/gr.4100405
- Wang J, Marowsky NC, Fan C (2014) Divergence of gene body DNA methylation and evolution of plant duplicate genes. *PLoS One* 9, e110357. doi:10.1371/journal.pone.0110357
- Springer NM, Xu X, Barbazuk WB (2004) Utility of different gene enrichment approaches toward identifying and sequencing the maize gene space. *Plant Physiol* 136:3023–3033. doi:10.1104/pp.104.043323
- Palmer LE, Rabinowicz PD, O’Shaughnessy AL et al (2003) Maize genome sequencing by methylation filtration. *Science* 302:2115–2117. doi:10.1126/science.1091265
- Rabinowicz PD, Palmer LE, May BP et al (2003) Genes and transposons are differentially methylated in plants, but not in mammals. *Genome Res* 13:2658–2664. doi:10.1101/gr.1784803
- Raleigh EA, Murray NE, Revel H et al (1988) McrA and McrB restriction phenotypes of some *E. coli* strains and implications for gene cloning. *Nucleic Acids Res* 16:1563–1575
- Whitelaw CA, Barbazuk WB, Pertea G et al (2003) Enrichment of gene-coding sequences in maize by genome filtration. *Science* 302:2118–2120. doi:10.1126/science.1090047
- Berenyi M, Mauleon RP, Kopecky D et al (2009) Isolation of plant gene space-related sequence elements by high C+G patch (HCGP) filtration: model study on rice. *Plant Mol Biol Report* 27:79–85. doi:10.1007/s11105-008-0063-2
- Wischnitzki E, Sehr EM, Hansel-Hohl K et al (2015) How to isolate a plant’s hypomethylome in one shot. *Biomed Res Int* 2015:570568. doi:10.1155/2015/570568
- Li D, Zhang B, Xing X, Wang T (2015) Combining MeDIP-seq and MRE-seq to investigate genome-wide CpG methylation. *Methods* 72:29–40. doi:10.1016/j.ymeth.2014.10.032
- Zhang B, Zhou Y, Lin N et al (2013) Functional DNA methylation differences between tissues, cell types, and across individuals discovered using the M&M algorithm. *Genome Res* 23:1522–1540. doi:10.1101/gr.156539.113
- Keller J, Bancroft I. 3_CTAB_DNA_extraction. https://www.arabidopsis.org/download_files/Protocols/compleat_guide/3_CTAB_DNA_extraction.pdf
- Nystedt B, Street NR, Wetterbom A et al (2013) The Norway spruce genome sequence and conifer genome evolution. *Nature* 497:579–584. doi:10.1038/nature12211
- Ballal RD, Saha T, Fan S et al (2009) BRCA1 localization to the telomere and its loss from the telomere in response to DNA damage. *J Biol Chem* 284:36083–36098. doi:10.1074/jbc.M109.025825
- Dlaska M, Anderl C, Eisterer W, Bechter OE (2008) Detection of circular telomeric DNA without 2D gel electrophoresis. *DNA Cell Biol* 27:489–496. doi:10.1089/dna.2008.0741
- Martin M (2011) Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17:10. doi:10.14806/ej.17.1.200
- Bolger AM, Lohse M, Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30:2114–2120. doi:10.1093/bioinformatics/btu170
- Langmead B, Salzberg SL (2012) Fast gapped-read alignment with Bowtie 2. *Nat Methods* 9:357–359. doi:10.1038/nmeth.1923
- Nussbaumer T, Martis MM, Roessner SK et al (2013) MIPS PlantsDB: a database framework

- for comparative plant genome research. *Nucleic Acids Res* 41:D1144–1151. doi:[10.1093/nar/gks1153](https://doi.org/10.1093/nar/gks1153)
20. Reinert K, Langmead B, Weese D, Evers DJ (2015) Alignment of next-generation sequencing reads. *Annu Rev Genomics Hum Genet* 16:133–151. doi:[10.1146/annurev-genom-090413-025358](https://doi.org/10.1146/annurev-genom-090413-025358)
 21. Li H, Handsaker B, Wysoker A et al (2009) The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 25:2078–2079. doi:[10.1093/bioinformatics/btp352](https://doi.org/10.1093/bioinformatics/btp352)
 22. Quinlan AR, Hall IM (2010) BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26:841–842. doi:[10.1093/bioinformatics/btq033](https://doi.org/10.1093/bioinformatics/btq033)
 23. Grabherr MG, Haas BJ, Yassour M et al (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29:644–652. doi:[10.1038/nbt.1883](https://doi.org/10.1038/nbt.1883)
 24. Haas BJ, Papanicolaou A, Yassour M et al (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 8:1494–1512. doi:[10.1038/nprot.2013.084](https://doi.org/10.1038/nprot.2013.084)
 25. Altschul SF, Gish W, Miller W et al (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410. doi:[10.1016/S0022-2836\(05\)80360-2](https://doi.org/10.1016/S0022-2836(05)80360-2)
 26. Camacho C, Coulouris G, Avagyan V et al (2009) BLAST+: architecture and applications. *BMC Bioinformatics* 10:421. doi:[10.1186/1471-2105-10-421](https://doi.org/10.1186/1471-2105-10-421)
 27. Jones P, Binns D, Chang H-Y et al (2014) InterProScan 5: genome-scale protein function classification. *Bioinformatics* 30:1236–1240. doi:[10.1093/bioinformatics/btu031](https://doi.org/10.1093/bioinformatics/btu031)
 28. Mitchell A, Chang H-Y, Daugherty L et al (2014) The InterPro protein families database: the classification resource after 15 years. *Nucleic Acids Res*. doi:[10.1093/nar/gku1243](https://doi.org/10.1093/nar/gku1243)
 29. Conesa A, Götz S, García-Gómez JM et al (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 21:3674–3676. doi:[10.1093/bioinformatics/bti610](https://doi.org/10.1093/bioinformatics/bti610)
 30. Conesa A, Götz S (2008) Blast2GO: a comprehensive suite for functional analysis in plant genomics. *Int J Plant Genomics* 2008:1–12. doi:[10.1155/2008/619832](https://doi.org/10.1155/2008/619832)
 31. Li W, Godzik A (2006) Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics* 22:1658–1659. doi:[10.1093/bioinformatics/btl158](https://doi.org/10.1093/bioinformatics/btl158)
 32. Fu L, Niu B, Zhu Z et al (2012) CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics* 28:3150–3152. doi:[10.1093/bioinformatics/bts565](https://doi.org/10.1093/bioinformatics/bts565)
 33. Robinson JT, Thorvaldsdóttir H, Winckler W et al (2011) Integrative genomics viewer. *Nat Biotechnol* 29:24–26. doi:[10.1038/nbt.1754](https://doi.org/10.1038/nbt.1754)
 34. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14:178–192. doi:[10.1093/bib/bbs017](https://doi.org/10.1093/bib/bbs017)
 35. Shagin DA, Lukyanov KA, Vagner LL, Matz MV (1999) Regulation of average length of complex PCR product. *Nucleic Acids Res* 27, e23
 36. Magoc T, Salzberg SL (2011) FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* 27:2957–2963. doi:[10.1093/bioinformatics/btr507](https://doi.org/10.1093/bioinformatics/btr507)

Spatio-Temporal Imaging of Promoter Activity in Intact Plant Tissues

Tou Cheu Xiong, Frédéric Sanchez, Jean-François Briat, Frédéric Gaymard, and Christian Dubos

Abstract

Localization and quantification of expression levels of genes help to determine their function. Localization of gene expression is often achieved through the study of their promoter activity. Three main reporter genes β -glucuronidase (GUS), green fluorescent protein (GFP), and luciferase (LUC) have been intensively used to characterize promoter activities, each having its own specificities and advantages. Among them, the *LUC* reporter gene is best suitable for the analysis of the promoter activity of genes in intact living plants. Here, we describe a LUC-based method that allows to precisely localize and quantify promoter activity at the whole plant level, and to study the mechanisms that are involved in long-distance regulation of gene expression in *Arabidopsis thaliana*. Imaging LUC signals with a low-light CCD camera allows monitoring promoter activity in time and space in the transgenic plant harboring the promoter fused with the *LUC* gene. In addition, it allows quantifying change of promoter activities in plant during several hours.

Key words Luciferase, Bioimaging, Ferritin, *AtFer1*, Iron, *Arabidopsis thaliana*

1 Introduction

The control of gene expression is complex and relies on numerous molecular mechanisms, i.e. requiring the binding of regulatory proteins to specific DNA sequences. In this regard, DNA sequences localized upstream of the transcribed region, also called promoter regions, are of central importance. For instance, it is through promoter regions that transcription factors regulate the expression of their target genes by directly and specifically interacting with distinct types of *cis*-regulatory sequences thus acting as activators, repressors, or both.

The development of tools allowing the analysis of gene expression has substantially contributed to the functional characterization of numerous genes. Amongst these tools, several reporter genes have been used to monitor the activity of gene promoters allowing the tissue, cellular, and subcellular determination of gene

expression in a quantitative and dynamic manner. Such tools have also permitted identifying several key regulatory elements (i.e. *cis*-regulatory sequences) that are involved in the control of gene expression. These methods rely on the use of three main reporter genes [1–3].

The β -glucuronidase reporter gene (*uidA*), or GUS, is a sensitive system for identifying promoter activity at the tissue and cellular levels. In addition, GUS activity allows precise in vitro quantification of promoter activity [2, 4]. The main drawback of this method is that treatments which must be applied to the samples in order to reveal GUS activity lead to the death of the cells or the destruction of the sample itself (i.e. tissue grinding for quantitative analysis). As a consequence, GUS is not a suitable reporter gene to monitor in vivo promoter activity of genes in a dynamic manner. In contrast, green fluorescent protein (GFP), as well as other related fluorescent proteins (e.g. RFP, YFP, CFP), is a nondestructive reporter allowing accurate localization of gene expression at the tissue, cellular, and subcellular levels [1]. The main disadvantage of this method is that GFP requires UV excitation in order to emit fluorescence, which can cause some damages to the samples during long exposure (e.g. phototoxicity on DNA). GFP photobleaching and autofluorescence of plant tissues are additional limiting parameters that have to be considered. The luciferase reporter gene relies upon photon emission, allowing the in situ study of promoter activity in a dynamic manner. LUC light emission is catalyzed by a cofactor (luciferin) that is provided to the samples prior analysis. Indeed, LUC activity is also routinely used for quantitative monitoring of gene expression over time (e.g. study of circadian rhythm) [5, 6]. Beside the fact that LUC analysis requires a dedicated device for low-light signal detection, which could limit its application, this reporter gene is ideal for gene expression studies at the whole plant level.

In this chapter, we describe a method allowing spatio-temporal imaging of promoter activity in intact plants, together with quantitative analysis of gene's promoter activity and long distance signaling studies (e.g. from roots to shoots). As an illustration we followed the LUC activity driven by the *AtFer1* (*At5g01600*) promoter whose activity is induced in response to iron treatment [7–9]. *AtFer1* encodes the most abundant ferritin protein—mainly involved in the transient storage of iron in chloroplasts—present in *Arabidopsis thaliana* vegetative tissues [7–9].

2 Materials

2.1 Plant Material

A transgenic *Arabidopsis thaliana* line expressing the promoter fused with the reporter gene LUC generated by Duc et al. [8] was used in this method. The promoter of the *AtFer1* ferritin gene

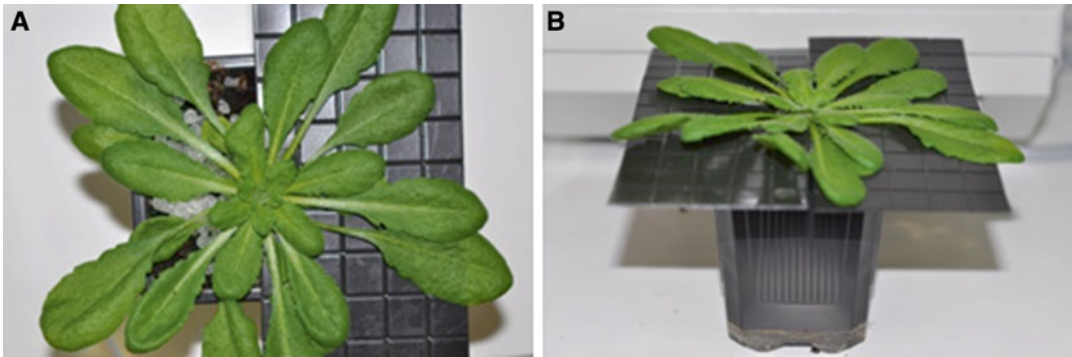


Fig. 1 Example of plant system for bioimaging. Prior to the measurements, two pieces of opaque and semi-rigid black plastic sheets are placed between plant leaves and soil to avoid any contact. (a) Picture with half of the pot covered with the black plastic sheets (on the *right side*). (b) Side view of the pot when the two sheets are in place

(At5g01600, 1.4 kb DNA fragment upstream from the start codon) fused to the firefly luciferase (LUC+) reporter gene was introduced in *Arabidopsis thaliana* (Col-0) (for details *see ref. 8*) (*see Note 1*).

2.2 Solutions

Solutions are prepared with deionized water.

1. Luciferin stock solution: 100 mM Luciferin-EF™ (endotoxin-free) (Promega) (*see Note 2*).
2. Luciferin working solution: 1 mM Luciferin-EF™ in 0.01% Triton X-100.
3. Iron solution: 2 g/L Fe-EDDHA solution (Toner PS).

2.3 Labware, Equipment, and Software

1. Back-illuminated CCD camera (Hamamatsu C4880-30) (*see Note 3*).
2. Hipic32 5.1.0 software (Hamamatsu Photonic) (*see Note 4*).
3. Opaque and semi-rigid black plastic sheets (Fig. 1).
4. ImageJ freeware (1.50a) [10] (*see Note 5*).
5. BG subtraction from ROI plugin for ImageJ.
6. PC with Windows® XP.
7. Plant growth chamber for *Arabidopsis thaliana*.

3 Methods

3.1 Plant Growth Conditions

1. Seeds of stably transformed transgenic plants containing luciferase promoter fusion are sown on soil (*see Note 6*).
2. Plants are grown for 6 weeks prior measurements in a growth chamber under short day condition (8 h light, 23 °C; 16 h darkness, 16 °C) with 250 μmol/m²/s of light intensity and 70% of relative humidity.

3.2 Bioluminescence Imaging

1. Prepare 5 mL of Luciferin working solution from 50 μ L of Luciferin stock solution.
2. 24 h before measurements, place an opaque and semi-rigid black plastic sheet between soil and plant leaves to prevent any contact. Then spray the plant with 5 mL of Luciferin working solution (*see Note 7*).
3. After 24 h, transfer the transgenic plant under the CDD camera in the dark (*see Note 8*).
4. Wait 1 h prior measurement (*see Note 9*).
5. Measurements are performed during the night period. At the beginning of this period, the stimulus is applied. Here, 75 mL iron solution per plant is provided to the roots, by pouring the solution directly onto the soil. Luciferase signals were imaged with 1 min exposure time each minute during 16 h (Fig. 2a–e) (*see Note 10*).
6. An image of the plant under white light is taken at the end of the measurements (Fig. 2f) (*see Note 11*).

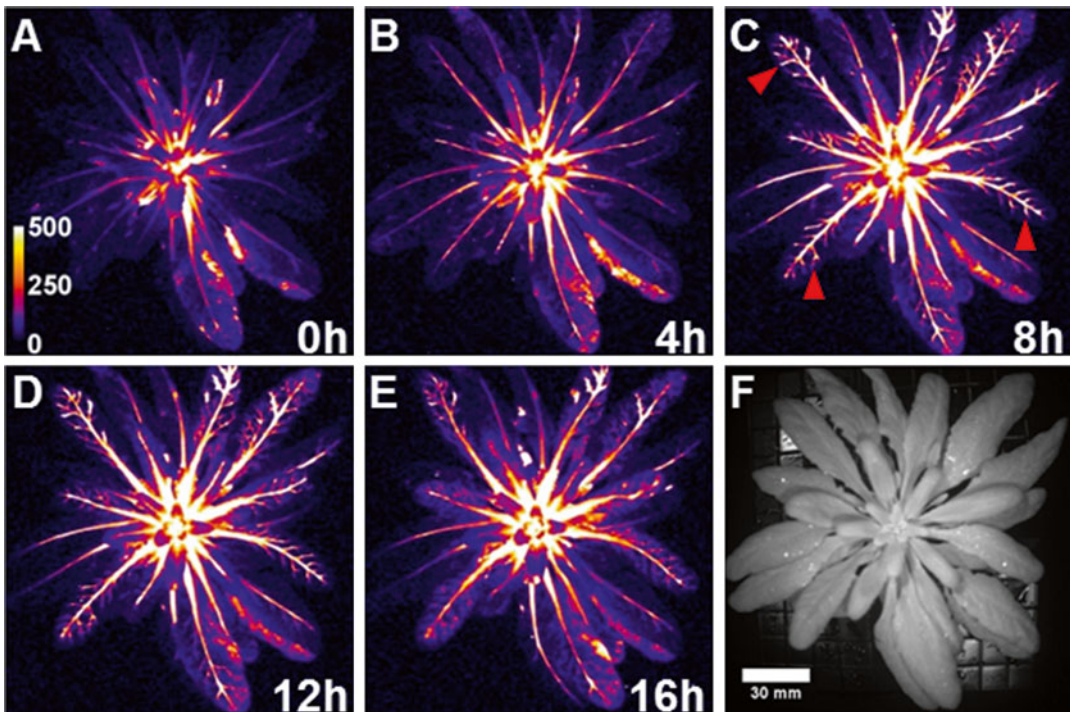


Fig. 2 Induction of *pAtFer1::LUC* expression in plant leaves in response to iron treatment. Time-series images in (a), (b), (c), (d), and (e) show the luciferase signals in intact *pAtFer1::LUC* plant leaves every 4 h during 16 h upon iron (2 g/L Fe-EDDHA solution) application on roots. The luciferase signal is collected with 1 min exposure time. Red arrows are showing induction of luciferase signals (c). The corresponding time (in hours) for each panel is indicated in the bottom left part of the picture, and refers to the time when iron was applied (0 h). Signal intensity is represented in false color (calibration scale is shown in (a)). Scale bar = 30 mm

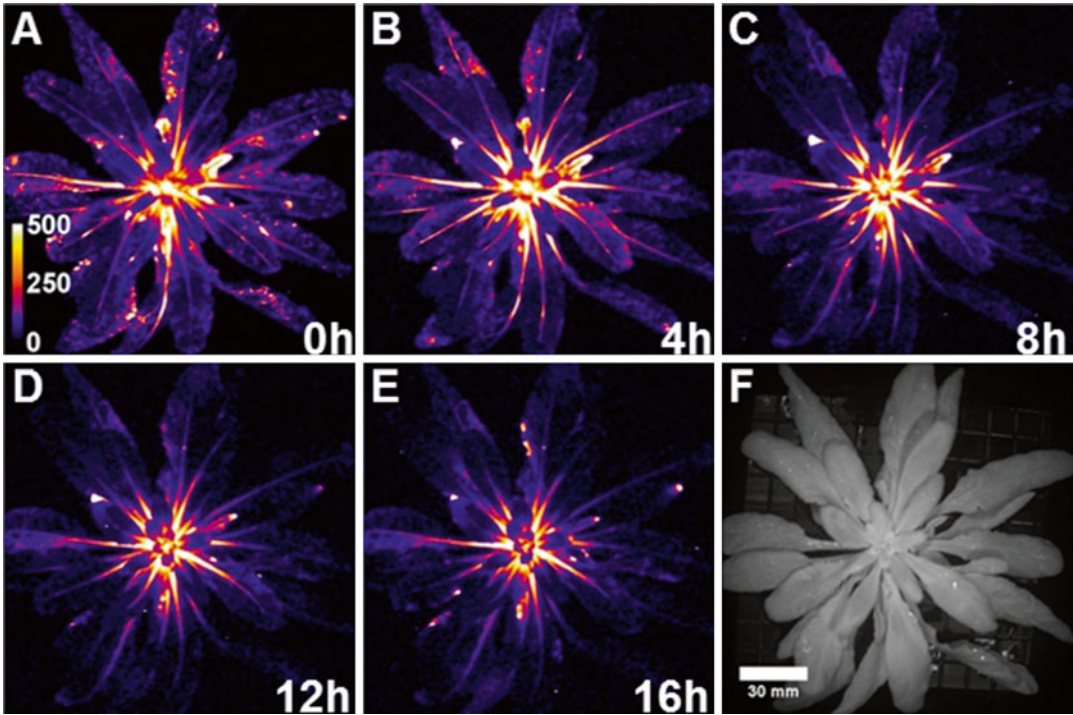


Fig. 3 Induction of *pAtFer1::LUC* expression in plant leaves in response to water application. Time-series images in (a), (b), (c), (d), and (e) show the luciferase signals in intact *pAtFer1::LUC* plant leaves every 4 h during 16 h upon water (deionized) application on roots. The luciferase signal is collected with 1 min exposure time. Water is not inducing *AtFer1* promoter activity. The corresponding time (in hours) for each panel is indicated in the *bottom left part* of the picture, and refers to the time when iron was applied (0 h). Signal intensity is represented in false color (calibration scale is shown in (a)). Scale bar = 30 mm

7. At the end of the experiment, an image without the plant is taken as background reference with the same exposure time (*see Note 12*).
8. The same protocol (**steps 1–7**) is then repeated for the control condition using deionized water (Fig. 3).

3.3 Image Corrections

1. Open all images as a sequence image in ImageJ by dragging the folder that contains all the images and dropping it onto the ImageJ menu bar. The entire sequence image open as stack image.
2. Correct background of the stack image by subtracting the reference background image. To achieve this step select on the ImageJ bar menu *Process*, then *Image calculator*. A new window opens where the stack image (to be corrected) is selected as Image 1, *Subtract* as Operation, and the background reference image as Image 2. Subtraction is effective once the *Ok* button is pressed.

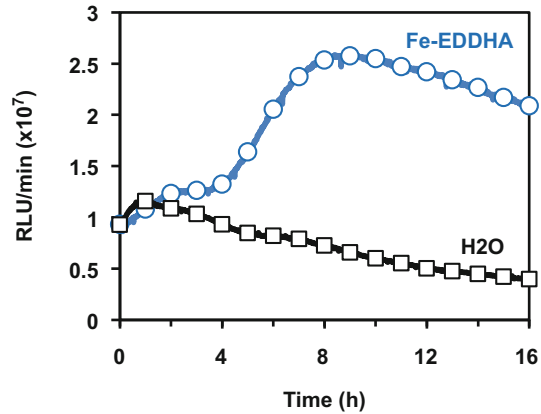


Fig. 4 Quantification of luciferase signals in *pAtFer1::LUC* plants treated or not with iron solution. Luciferase signals from the entire plant leaves shown in Figs. 2 and 3 were collected with 1 min exposure time every min during 16 h. In accordance with previous reports [7–9], *AtFer1* gene is induced in plant leaves after iron supply (Fe-EDDHA) (see **Note 1**)

3. If the background between the different images is fluctuating, the level of the background must be normalized on the entire image stack using the ImageJ plugin *BG subtraction from ROI* (region of interest). To do this, first select a square ROI of non-plant pixels and then launch the plugin (see **Note 13**).

3.4 Signals Quantification at the Whole Plant Level

1. Open ROI manager by selecting on the ImageJ bar menu *Analyse*, then *Tools*, then *ROI manager*.
2. On the corrected image stack, select the entire image as ROI (see **Note 14**).
3. Save the ROI in *ROI manager* by clicking on *Add* (see **Note 15**).
4. Select the parameters to measure by selecting *Analyse* on the bar menu, then *Set measurements*. A new window opens with all the parameters ImageJ could measure (*Integrated density*, *Mean*, *Median*, etc.). *Integrated density* is selected for the quantification.
5. In ROI manager window, select the ROI to measure.
6. Click on *More* and select on the list menu *Multi Measure* (see **Note 16**).
7. The data appear in a new window.
8. Save data or copy and paste in applicable software.
9. Plot the data intensity vs. time. Results are represented as Relative Light Unit (RLU) per exposure time (1 min) (Fig. 4).

4 Notes

1. *pAtFer1::LUC* line was generated and used for forward screening approach [8]. Real time quantification of *AtFer1* promoter induction in living plant with LUC activity are unpublished data.
2. Endotoxin-free Luciferin-EFTM is exclusively provided by Promega. Store 100 mM stock solution (50 μ L aliquotes) at -20 °C in the dark.
3. This is a low-light detection camera. Bioluminescence signals are collected with 1 min exposure time.
4. This software is used to control the Hamamatsu C4880-30 CDD camera.
5. The freeware ImageJ can be download at <http://imagej.nih.gov/ij/download.html>.
6. Here we use as an example a transgenic line described in [8] that contains a fusion with the promoter of the *Arabidopsis thaliana AtFer1* gene (At5g01600; *pAtFer1::LUC*) whose activity depends on iron availability [7–9].
7. Plants were sprayed just after the light/dark transition. Time of the day must be carefully chosen if the promoter is light-dependent.
8. During measurement plants should be kept in the dark.
9. Dark adaptation is important to minimize autoluminescent background from plant chlorophyll. The first 5 min could not be used for quantification due to luminescence emission decay of chlorophyll.
10. After 5 min, exposure time can be adjusted to improve the signal.
11. An image should also be taken at the beginning of the experiment.
12. Background image reference is required for image correction and quantification.
13. This step has to be applied if the background is variable over the time due to technical problem such as an unstable temperature of the camera that might affect the background level. Any fluctuations of the background might affect the result and mislead the interpretation made from the luciferase signals, especially if the obtained signals are weak.
14. The ROI could be selected on sub-region of the plant, e.g. individual leaf to study promoter activity on the selected ROI.
15. Several ROIs can be saved and re-used with ROI manager.
16. *Multi Measure* will measure the selected ROI on the entire stack. *Measure* button on ROI manager is measuring only on the current image.

Acknowledgment

This work was supported by INRA BAP STARTER Grant (ST07 FerROS-RACINE).

References

1. Haseloff J, Siemering KR, Prasher DC, Hodge S (1997) Removal of a cryptic intron and sub-cellular localization of green fluorescent protein are required to mark transgenic *Arabidopsis* plants brightly. *Proc Natl Acad Sci U S A* 18:2122–2127
2. Jefferson RA, Kavanagh TA, Bevan MW (1987) GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO J* 6:3901–3907
3. Ow DW, DE Wet JR, Helinski DR, Howell SH, Wood KV, Deluca M (1986) Transient and stable expression of the firefly luciferase gene in plant cells and transgenic plants. *Science* 234:856–859
4. Xu W, Grain D, Le Gourrierc J, Harscoët E, Berger A, Jauvion V, Scagnelli A, Berger N, Bidzinski P, Kelemen Z, Salsac F, Baudry A, Routaboul JM, Lepiniec L, Dubos C (2013) Regulation of flavonoid biosynthesis involves an unexpected complex transcriptional regulation of TT8 expression, in *Arabidopsis*. *New Phytol* 198:59–70
5. Millar AJ, Carré IA, Strayer CA, Chua NH, Kay SA (1995) Circadian clock mutants in *Arabidopsis* identified by luciferase imaging. *Science* 267:1161–1163
6. Tissot N, Przybyla-Toscano J, Reyt G, Castel B, Duc C, Boucherez J, Gaymard F, Briat JF, Dubos C (2014) Iron around the clock. *Plant Sci* 224:112–119
7. Petit JM, van Wuytswinkel O, Briat JF, Lobréaux S (2001) Characterization of an iron-dependent regulatory sequence involved in the transcriptional control of *AtFer1* and *ZmFer1* plant ferritin genes by iron. *J Biol Chem* 276:5584–5590
8. Duc C, Cellier F, Lobréaux S, Briat JF, Gaymard F (2009) Regulation of iron homeostasis in *Arabidopsis thaliana* by the clock regulator time for coffee. *J Biol Chem* 284:36271–36281
9. Petit JM, Briat JF, Lobréaux S (2001) Structure and differential expression of the four members of the *Arabidopsis thaliana* ferritin gene family. *Biochem J* 359:575–582
10. Schneider CA, Rasband WS, Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. *Nat Methods* 9:671–675

Novel Synthetic Promoters from the Cestrum Yellow Leaf Curling Virus

Dipak Kumar Sahoo, Shayan Sarkar, Indu B. Maiti, and Nrisingha Dey

Abstract

Constitutive promoters direct gene expression uniformly in most tissues and cells at all stages of plant growth and development; they confer steady levels of transgene expression in plant cells and hence their demand is high in plant biology. The gene silencing due to promoter homology can be avoided by either using diverse promoters isolated from different plant and viral genomes or by designing synthetic promoters. The aim of this chapter was to describe the basic protocols needed to develop and analyze novel, synthetic, nearly constitutive promoters from Cestrum yellow leaf curling virus (CmYLCV) through promoter/leader deletion and activating *cis*-sequence analysis. We also describe the methods to evaluate the strength of the promoters efficiently in various transient expression systems like agroinfiltration assay, gene-gun method, and assay in tobacco protoplasts. Besides, the detailed methods for developing transgenic plants (tobacco and *Arabidopsis*) for evaluation of the promoter using the *GUS* reporter gene are also described. The detailed procedure for electrophoretic mobility shift assay (EMSA) coupled with super-shift EMSA analysis are also described for showing the binding of tobacco transcription factor, TGA1a to *cis*-elements in the CmYLCV distal promoter region.

Key words Synthetic promoter, *Caulimovirus*, Transgenic plants, *GUS*, *Arabidopsis*, Tobacco, Gene expression

1 Introduction

Among the viral promoters, the CaMV35S promoter from the cauliflower mosaic virus is one of the most widely utilized promoters for basic research and the development of transgenic plants [1]. The constitutive expression of the CaMV35S promoter tends to be relatively high in different tissues of many plants and seems to result from an additive effect of multiple tissue-specific elements [2]. The Peanut chlorotic streak virus, Dahlia mosaic virus, Strawberry vein banding virus, Mirabilis mosaic virus, Cassava vein mosaic virus, Figwort mosaic virus, and Cestrum yellow leaf curling virus have been used to identify regulatory elements that would drive high levels of constitutive gene expression in plants [3–19]. Chimeric

promoters that drive constitutive gene expression are created by combining elements from these viral-derived sequences other than the 35S promoter and such chimeric promoters created through shuffling of regulatory elements and inclusion of plant-derived or other viral-derived sequences have shown high levels of transgene expression in several plant species [6, 19–27]. Numerous transgenic plants have exhibited the phenomenon of homology-dependent gene silencing and such gene silencing mediated by promoter homology occurs at the level of transcription resulting in meiotically heritable alterations in methylation and gene activity [28]. The gene silencing due to promoter homology can be avoided by either using diverse promoters isolated from different plant and viral genomes or by designing synthetic promoters [3–10, 13, 26, 29].

Synthetic promoters comprise consensus DNA sequences of common elements of natural promoter regions and they differ tremendously from native promoters as they can provide expression profiles that do not exist in nature. In addition, synthetic promoters are useful for functional validation of promoter sequences. Synthetic promoters are classically constructed by combinatorial engineering of *cis*-elements, which include enhancers, activators, or repressors directly upstream of the core promoter sequence [30]. Moreover, arrangement of *cis*-elements within a synthetic promoter can result in very precise transgene expression while nonspecific expression resulting from the additional elements present within the “full-length” promoter sequences is avoided [30, 31].

Several approaches are reported for designing synthetic promoters [30]. One of the approaches to develop controllable promoters is to modify known, natural promoters so that they contain novel regulatory units [32]. We recently designed and evaluated Cestrum yellow leaf curling virus (CmYLCV) full-length transcript promoter through promoter/leader deletion and activating *cis*-sequence analysis to develop synthetic promoters with enhanced activity [16]. In this chapter, we describe in detail the methods for developing and characterizing useful synthetic plant promoters from Cestrum yellow leaf curling virus (CmYLCV). We describe the detailed promoter/leader 5'- and 3'-end deletion analysis of the full-length transcript promoter from CmYLCV virus, to define the optimal boundaries required for the maximum promoter activity of the CmYLCV promoter. We describe various transient expression systems to evaluate the strength of these promoters efficiently: by agroinfiltration assay, gene-gun method, and assay in tobacco protoplasts. To evaluate the strength of these promoters in transgenic plants (tobacco and Arabidopsis), the *GUS* reporter gene employing both fluorimetric and histochemical assays is used. We quantify the *uidA*-mRNA level in transgenic plants expressing *GUS* under various synthetic promoters by real-time PCR (qRT-PCR). We also confirm the binding of tobacco transcription factor, TGA1a to *cis*-elements in the CmYLCV distal promoter

region by electrophoretic mobility shift assay (EMSA) coupled with super-shift EMSA analysis. Due to availability of kits from different commercial sources for many of the approaches, we recommend to follow manufacturer's protocols when commercial kits are used. We only briefly describe portions of the protocols from commercial kits that are used during the study.

2 Materials

2.1 Plasmids and Primers

1. The protoplast expression vector pUCPMAGUS [9], pUC119, and plant expression vector pKYLX71GUS [9, 33] are used to clone various *CmYLCV* promoter fragments, whereas pUCPMA-Lux-*CaMV35S* [34] is used as an internal control and pUCPMAGUS-*CaMV35S* [5] is used as a positive/reference control in GUS assay experiments. The plasmid pUCPMA-*CaMV35S*-GFP with GFP reporter is used to monitor tobacco protoplast transformation efficiency.
2. Various *CmYLCV* promoter fragments are generated by PCR amplification using appropriately designed PCR primers to introduce an *EcoRI* site at the 5'-end and a *HindIII* site at the 3'-end of the amplified products (Table 1).
3. All synthetic oligonucleotides and PCR primers used are obtained from a commercial supplier (e.g. Integrated DNA Technologies, Skokie, IA, USA).

2.2 PCR, Cloning and Transformation

1. Taq DNA polymerase and dNTPs: PfuUltra high-fidelity Taq DNA polymerase (Stratagene, USA) and dNTP mix (10 mM each). All PCR reagents are stored at -20°C .
2. *EcoRI* and *HindIII* restriction enzymes and T4 DNA ligase are obtained from a commercial supplier (e.g. NEB, USA).
3. Ultrapure low melting agarose.
4. 50× electrophoretic Tris-acetate-EDTA (TAE) buffer: 242 g/L Tris base, 100 mL/L of 0.5 M EDTA (pH 8.0), 57.1 mL/L glacial acetic acid. Autoclave and store at room temperature. Dilute to 1× TAE in milli-Q water for regular use. A horizontal electrophoresis system is used for gel electrophoresis.
5. Gel Extraction Kit.
6. *E. coli* strain K12 TB1 (F-ara Δ (lac-proAB) [Φ 80dlac Δ (lacZ) M15] rpsL(StrR) thi hsdR) (NEB, USA).
7. Luria–Bertani (LB) medium: sodium chloride (NaCl, 10 g/L), Bacto tryptone (10 g/L), and Bacto yeast extract (5 g/L). Add 15 g/L Bacto agar to LB medium to make LB agar plate. Dissolve in deionized water and autoclave.
8. Plasmid isolation kit for *E. coli*.

Table 1
Sequence of oligonucleotide primers used in the study

Sl. No.	Sequence 5'–3'	Name of the CmYLCV constructs in which below primers are used
1	GCGGGCGAATTCTGCAGAAGAAATAA	Fw primer for construct CmYLCV1.11 and CmYLCV1.10
2	GCGGGCGAATTCTAGAAGGTGTGTAT	Fw primer for construct CmYLCV2.11
3	GCGGGCGAATTCAAGGTTTGGTATCA	Fw primer for construct CmYLCV3.11
4	GCGGGCGAATTCAGAAATTGAAGATG	Fw primer for construct CmYLCV4.11
5	GCGGGCGAATTCTCAAAGCCATGGAA	Fw primer for construct CmYLCV5.11
6	GCGGGCGAATTCAAGATTCTTTGCCA	Fw primer for construct CmYLCV6.11
7	GCGGGCGAATTCGTGCAAATCCGAG	Fw primer for construct CmYLCV7.11 and CmYLCV7.10
8	GCGGGCGAATTCAGTCAGAAGACGA	Fw primer for construct CmYLCV8.11
9	GCGGGCGAATTCGTGGCAGACATAC	Fw primer for construct CmYLCV9.11
10	ATGCAGAAGCTTTTTCTTCTTCCTGG	Rv primer for constructs CmYLCV1.10 and CmYLCV7.10
11	ATGCAGAAGCTTAGCTCTTACCTG	Rv primer for constructs CmYLCV1.11 to CmYLCV 9.11
12	GCGGGCGAATTCCTGGCAGACAAAG	Fw primer for constructs CmpC and CmpS
13	ATGCAGAAGCTTTTGCTCCCTTAACA	Rv primer for construct CmpC
14	ATGCAGAAGCTTCTACTTCTAGGCTA	Rv primer for construct CmpS
15	GCGGGCGAATTCCTAACAAACATC	Fw primer for constructs CmYLCV1, CmYLCV10 to CmYLCV13
16	GCGGGCGAATTCGACGAAGACTTTTC	Fw primer for construct CmYLCV2
17	GCGGGCGAATTCATACTGTCCCACA	Fw primer for construct CmYLCV3
18	GCGGGCGAATTCATGCGTCTGACA	Fw primer for construct CmYLCV4
19	GCGGGCGAATTCGGTCCCTACCACGA	Fw primer for construct CmYLCV5
20	GCGGGCGAATTCGAACAAATAAGATT	Fw primer for construct CmYLCV6
21	GCGGGCGAATTCCTTCAGACTCCAA	Fw primer for construct CmYLCV7
22	GCGGGCGAATTCAGGGTAGTTTGG	Fw primer for construct CmYLCV8
23	GCGGGCGAATTCCTCCCTCATTG	Fw primer for construct CmYLCV9
24	ATGCAGAAGCTTTCTCTAGGACTATC	Rv primer for construct CmYLCV10
25	ATGCAGAAGCTTAGATTTTGCTCCCT	Rv primer for construct CmYLCV11
26	ATGCAGAAGCTTGCTAAGTATTTATA	Rv primer for construct CmYLCV12

(continued)

Table 1
(continued)

Sl. No.	Sequence 5'–3'	Name of the <i>CmYLCV</i> constructs in which below primers are used
27	ATGCAGAAGCTTTACTTATTTTCGTAA	Rv primer for construct <i>CmYLCV13</i>
28	ATGCAGAAGCTTTTCGCGCGTTC	Rv primer for constructs <i>CmYLCV1</i> to <i>CmYLCV9</i>

Fw forward, *Rv* reverse

9. Reagents and instruments for DNA sequencing: GenomeLab DTCS quick start kit, GenomeLab separation gel, GenomeLab separation buffer (Beckman Coulter, USA), and Beckman Coulter CEQ-8000 sequencer. Store the DTCS quick start kit at -20°C , and the separation buffer and gel are stored at 4°C .
10. Thermal cycler (Bio-Rad).

2.3 Tobacco Cell Suspension Culture, Protoplast Isolation and Electroporation

1. Cell suspension culture medium (Murashige and Skoog's medium) containing potassium phosphate (KH_2PO_4 ; 204 mg/L), pyridoxine HCl (0.5 mg/L), nicotinic acid (0.5 mg/L), thiamine HCl (0.5 mg/L), 2,4-dichlorophenoxyacetic acid (2,4-D; 0.2 mg/L), and kinetin (0.1 mg/L). Adjust medium pH to 5.8 with 1 M NaOH before autoclaving.
2. MMC solution: Mannitol (91.1 g/L), 2-morpholinoethane sulfonic acid (MES, 1.95 g/L), and calcium chloride ($\text{CaCl}_2 \cdot 2\text{H}_2\text{O}$, 1.47 g/L). Adjust pH to 5.6 with 1 M NaOH before autoclaving and store at 4°C .
3. Enzyme solution: Cellulase Onozuka RS (0.75%; Yakult Honsa, Japan) and pectinase (0.075%; Sigma). Dissolve the enzymes in MMC solution on a stirring plate and filter sterilize using Millipore filter (0.22 μm size; Millipore, USA) (*see Note 1*).
4. Sucrose solution: Sucrose (25 g/100 mL), MES (1.95 g/L), and $\text{CaCl}_2 \cdot 2\text{H}_2\text{O}$ (1.47 g/L). Adjust pH to 5.6 with 1 M NaOH before autoclaving and store at 4°C .
5. Electroporation buffer: Mannitol (91.1 g/L), potassium chloride (KCl, 5.21 g/L), and MES (975 mg/L). Adjust pH to 5.6 with 1 M NaOH before autoclaving and store at 4°C .
6. Protoplast culture medium: Mannitol (91.1 g/L), MES (1.95 g/L), $\text{CaCl}_2 \cdot 2\text{H}_2\text{O}$ (1.47 g/L), KH_2PO_4 (27 mg/L), KNO_3 (101 mg/L), MgSO_4 (120 mg/L), KI (2 mg/L), and sucrose (30 g/L). Adjust pH to 5.6 with 1 M NaOH before autoclaving and store at room temperature.
7. Mannitol agarose solution: Dissolve 9 g of mannitol and 1 g of low melting agarose in 100 mL milli-Q water and autoclave. 1 mL of this solution is used to coat a 25-mm sterile Petri dish.

8. Electroporator: Gene Pulser II apparatus with the capacitance extender II (model 165–2107; Bio-Rad, USA).
9. Tobacco cell suspension cultures (*Nicotiana tabacum* L. cv. Xanthi).

2.4 Biolistic Onion Peel Transient Assay

1. Cells from epidermal peels of onion (*Allium cepa* cv. Tango bulbs).
2. Expression plasmid vector pUCPMAGUS containing CmYLCV4, CmYLCV9.11, CmYLCV13, CmpC, CmpS, and CaMV35S promoter fragments.
3. Olympus SZX12 bright-field microscope.
4. Gold particles 0.6 μm or 1.0 μm in diameter.
5. 100% ethanol.
6. 2.5 M CaCl_2 .
7. 0.1 M Spermidine.
8. Solid medium: 1 \times Murashige and Skoog (1962) salts [35], 30 g/L sucrose, adjust to pH 5.7 with 1 M NaOH and then add 2% agar.
9. PDS-1000/He System (Bio-Rad, USA).
10. Microcarriers M17, tungsten, 1.1 μm (for DNA coating).
11. Macrocarriers (for spotting microcarriers for bombardment).
12. Rupture disks (to build up pressure for bombardment) (1100 psi).
13. Stopping screens (metal mesh to stop the rupture disk and macrocarrier to be shot out of the stage and hit [destroy] the sample).

2.5 β -Glucuronidase (*uidA* or *GUS*) Assay

1. QB: For 100 mL add 5 mL of 2 M KPO_4 (pH 7.8), 200 μL of 0.5 M EDTA (pH 8.0), 1 mL of Triton X-100, 12.5 mL 80% Glycerol and 15.4 mg DTT in 81.1 mL dH_2O and store at -20°C .
2. GUS assay reagent: In 25 mL QB add 50 μL DTT, 22 mg MUG (Methylumbelliferyl β -D-glucuronide) and store at -20°C .
3. MU calibration stock: For 1 mM MU, add 9.9 mg of MU in 50 mL of dH_2O and for making 1 μM MU, add 10 μL 1 mM MU in 10 mL dH_2O . For making 50 nM MU calibration solution, add 100 μL of 1 μM MU in 1.9 mL of Na_2CO_3 Stop solution. Make fresh immediately before each use.
4. 0.2 M Na_2CO_3 Carbonate Stop Buffer: Dissolve 21.2 g of Na_2CO_3 in sterile water and adjust the volume to 1000 mL.
5. 2 M KPO_4 (pH 7.8): Dissolve 63.2 g of K_2HPO_4 and 5.0 g of KH_2PO_4 in sterile water, adjust pH to 7.8 with 1 M KOH and make up the volume to 200 mL.

- 0.5 M EDTA (pH 8.0): Dissolve 46.52 g of EDTA in sterile water and adjust pH to 8.0 by adding 5 pellets of NaOH (or alternatively with 10 N NaOH) and adjust the volume to 250 mL (*see Note 2*).
- 10 N NaOH: Dissolve 100 g of NaOH in sterile water and adjust the volume to 250 mL. Store at RT in a plastic bottle (NaOH will react with glass).
- 1 M DTT: Dissolve 1.545 g of DTT and add 0.01 M sodium acetate, NaOAc (0.01 M NaOAc is 33 μ L of 3 M NaOAc pH~5.2 in 9.67 mL dH₂O), adjust pH to 5.2 and adjust volume to 10 mL, filter sterilize. Aliquots of 1 mL are stored at -80 °C.
- Fluorometer (TKO 100 fluorometer, Hoefer Scientific Instruments, USA).
- Reagents and instrument for luciferase assay: Luciferase assay system (Promega, USA); Luminometer (Turner Designs, USA).
- Spectrophotometer (Unico UV-2000 Spectrophotometer, SpectraLab Scientific Incorporation, USA).
- The Bradford kit (Quick Start™ Bradford Protein Assay Kit, Bio-Rad, USA).

2.6 Histochemical GUS Staining

- 2 \times 0.1 M Phosphate buffer pH 7: Dissolve 12.0 g NaH₂PO₄ and 14.19 g Na₂HPO₄ in 800 mL dH₂O. Adjust pH to 7.0 with 1 M NaOH or 1 M HCl and adjust volume to 1 L.
- Fixative: 4% formaldehyde, prepared fresh from paraformaldehyde, in 1 \times phosphate buffer. Prepare the fixative fresh on the day it is going to be used and do not store for later use.
- X-gluc substrate solution: Dissolve 1 mg 5-bromo-4-chloro-3--indolyl β -D-Glucuronide (X-Gluc) in 0.1 mL methanol, add 1 mL 2 \times phosphate buffer, 20 μ L 0.1 M potassium ferrocyanide, 20 μ L 0.1 M potassium ferricyanide, 10 μ L 10% (w/v) solution of Triton X-100, 0.85 mL water.
- 70% (v/v) ethanol.
- 50% (v/v) and 100% glycerol.
- Vacuum pump and desiccators.
- Bright-field microscope (e.g. Olympus SZX12).

2.7 Generation of Transgenic Tobacco and Arabidopsis Plants

- Plant material: Wild type 4- to 8-week-old tobacco plant (*Nicotiana tabacum* L. cv Samsun) and Arabidopsis plants (*Arabidopsis thaliana* ecotype Columbia-0).
- Agrobacterium strain: *Agrobacterium tumefaciens* strain C58C1:pGV3850.
- B5 vitamin: Dissolve 50.0 g of myo-inositol, 5.0 g of thiamin-HCl, 0.5 g of nicotinic acid, 0.5 g of pyridoxine-HCl in 1 L of water, freeze as 2 mL aliquots at -20 °C.

4. 6-Benzyl-aminopurine (BAP): Take 125 mg of BAP in 4 mL of water, add 4 mL of 1 M NaOH to dissolve it and make volume to 100 mL with water. Make aliquots of 1 mL (containing 1.25 mg BAP) and keep at -20°C .
5. Indole-3-Acetic Acid (IAA): Dissolve 100 mg IAA in 100 mL of 50% ethanol, Make aliquots of 1.0 mL (containing 1.0 mg IAA) and keep at -20°C .
6. TOM (Callus and Regeneration Tomato) Medium: Dissolve 4.31 g of Murashige and Skoog Salts, 2 mL (2.5 mg) of Benzylaminopurine (BAP) stock solution, 1 mL (1 mg) of Indole-3-Acetic Acid (IAA) stock solution, 30.0 g of sucrose and 2 mL of B5 vitamins in 800 mL of water, adjust pH to 5.7–5.9 with 1 M NaOH or 1 M HCl, adjust the volume to 1 L and add 8.0 g of Bacto-agar (Difco, USA) before autoclave.
7. TKM Medium: TOM medium with 250 mg/L kanamycin and 500 mg/L cefotaxime.
8. Rooting (T-) Medium: Dissolve 4.31 g of Murashige and Skoog Salts, 30.0 g of sucrose and 2 mL of B5 vitamins in 800 mL of water, adjust pH to 5.7–5.9 with 1 M NaOH or 1 M HCl, adjust the volume to 1 L and add 7.0 g of Bacto-agar (Difco, USA) before autoclave.
9. LB medium: Dissolve 10 g of bacto-tryptone, 5 g of bacto-yeast extract, 10 g NaCl in 800 mL water and adjust volume to 1 L. For solid LB medium add 18 g bacto-agar in 1 L of the medium and autoclave before use.
10. YEB medium: Dissolve 1 g of bacto-yeast extract, 5 g of beef extract, 5 g of peptone and 5 g of sucrose in 800 mL water and make volume to 1 L. For solid YEB medium add 18 g bacto-agar in 1 L of the medium and autoclave before use.
11. Surfactant Silwet L-77 (Lehle Seeds) 0.025%, 0.05%, 0.075% and 0.1% (v/v).
12. Tools needed: Petri dishes (100 mm \times 15 mm), forceps, surgical blades, parafilm, biohazard bags, and permanent markers.
13. Kanamycin (50 $\mu\text{g}/\text{mL}$): Dissolve 0.5 g of kanamycin into 10 ml of ddH₂O. Filter through a 0.22 μm filter to sterilize. Aliquot and store at -20°C .
14. 95% (v/v) Ethanol.
15. 10% Clorox (v/v) solution: For disinfection, a fresh 10% (v/v) solution of Clorox bleach in deionized water is recommended.
16. 0.5 \times MS/0.8% tissue culture medium: Dissolve 2.15 g of Murashige and Skoog Salts in 1 L of dH₂O and add 8.0 g of Bacto-agar before autoclave.

2.8 RNA Isolation and Quantitative Real-Time PCR (qRT-PCR)

1. Plant RNA extraction kit (e.g. Plant RNeasy extraction kit, Qiagen, CA, USA).
2. cDNA synthesis kit (e.g. iScript cDNA synthesis kit, BioRad, USA).
3. iTaq universal SYBR Green one-step kit (BioRad, USA).
4. Real-Time PCR System (Step One Real-Time PCR System, Applied Biosystems).

2.9 Nuclear Extract Preparation

1. 0.5 M dithiothreitol (DTT) in 0.01 M sodium acetate pH 5.2, store at -20°C .
2. 0.1 M phenylmethanesulfonylfluoride (PMSF) in isopropanol, store at -20°C .
3. Extraction Buffer: 50 mM Tris-HCl (pH 7.5), 5 mM MgCl_2 , 0.1 mM EDTA, 0.3 M sucrose, 15 mM KCl. Store at 4°C . Immediately before use, supplement the extraction buffer with 1 mM DTT, 0.2 mM PMSF, and 10 mg protease inhibitor/mL (containing equal amounts of leupeptin and pepstatin).
4. Nuclear protein isolation buffer: 10 mM Tris-HCl (pH 7.5), 0.4 mM NaCl. Immediately before use, supplement the isolation buffer with 1 mM DTT, 0.2 mM PMSF, and 10 mg protease inhibitor/mL (containing equal amounts of leupeptin and pepstatin).

2.10 Recombinant TGA1a Expression and Purification

1. Expression vector for protein expression pET-29b (Invitrogen).
2. Plasmid pTf-16 encoding chaperone Trigger Factor (tig) (Takara, Japan).
3. Competent cells *Escherichia coli* strain BL21 (DE3).
4. LB medium (*see* Subheadings 2.2, item 5 and 2.6, item 7).
5. Selective antibiotics: kanamycin 50 mg/mL stock in double distilled water (ddH₂O), chloramphenicol 25 mg/mL stock, sterile filter (0.2 μm), store at -20°C .
6. 1 M Isopropyl- β -D-1-thiogalactoside (IPTG): dissolve 2.38 g of IPTG in deionized water, make the final volume up to 10 mL, and sterilize the IPTG stock solution by filtering, store at -20°C .
7. Arabinose (200 mg/mL) stock in double distilled water (ddH₂O), sterile filter (0.2 μm), store at 4°C .
8. HisPur™ Cobalt resin (Thermo Scientific).
9. Plastic columns 2 mL capacity (Thermo Scientific).
10. Protein extraction reagent: B-PER™ Bacterial Protein Extraction Reagent (Thermo Scientific).
11. 1.0 M Sodium Phosphate Buffer, pH 7.4: Prepare by adding 1 M $\text{Na}_2\text{HPO}_4 \cdot 7\text{H}_2\text{O}$, 268.07 g per 1 L and 1 M $\text{NaH}_2\text{PO}_4 \cdot \text{H}_2\text{O}$, 34.45 g per 250 ml, until pH 7.4 is reached; sterilize by autoclaving.

12. 5 M sodium chloride (NaCl): Dissolve 146.1 g NaCl in ~350 ml H₂O, then bring up to volume with dH₂O. Sterilize by autoclaving.
13. Imidazole stock (1 M): Dissolve 68 g in 1 L of sterile water. Sterile filter using 22 µm filter.
14. Equilibration/Wash Buffer: 50 mM sodium phosphate, 300 mM sodium chloride, 10 mM imidazole; pH 7.4. Sterile-filter. Store at 4 °C.
15. Elution Buffer: 50 mM sodium phosphate, 300 mM sodium chloride, 150 mM imidazole; pH 7.4. Sterile-filter. Store at 4 °C.
16. Dialysis buffer: 20 mM HEPES pH 7.9 (pH adjusted with 1 M KOH), 40 mM KCl, 2 mM MgCl₂, 0.5 mM DTT, 0.5 mM EDTA (pH 8.0), 10% v/v glycerol.
17. Amicon[®] Ultra 4 mL Centrifugal Filters (10 kDa cut-off; Millipore).
18. Ammonium persulfate: dilute to a 10% stock solution in water. Store at 4 °C.
19. Tetramethylethylenediamine (TEMED), store at room temperature.
20. 30% acrylamide stock solution: Several manufacturers (e.g. Bio-Rad) sell electrophoresis grade acrylamide that is free of contaminating metal ions. For making the stock solution, add 29.0 g of acrylamide and 1.0 g of bisacrylamide to 100 ml of H₂O. Filter the stock solution through Whatman filter paper and store at 4 °C. Prepare fresh stock acrylamide solution every few weeks.
21. 1.5 M Tris-HCl pH 8.8: Dissolve Tris base 18.15 g in 50 mL dH₂O, adjust pH 8.8 with 1 N HCl and adjust the volume to 100 mL with sterile water.
22. 0.5 M Tris-HCl pH 6.8: Dissolve Tris base 3.0 g in 25 mL dH₂O, adjust pH 6.8 with 1 N HCl and adjust the volume to 50 mL with sterile water.
23. Make 10% SDS-polyacrylamide gel using a gel handcast system (Mini-PROTEAN[®] Tetra handcast systems, Bio-Rad, USA) by mixing 2.5 mL 30% acrylamide stock solution, 2.8 mL 1.5 M Tris-HCl pH 8.8 buffer, 2.2 mL dH₂O, 75 µL 10% SDS, 25 µL 10% APS, 5 µL TEMED as the resolving mix followed by the addition of stacking mixture composed of 1.34 mL of 30% acrylamide solution, 2.6 mL 0.5 M Tris-HCl pH 6.8 buffer, 5.86 mL dH₂O, 100 µL 10% SDS, 100 µL 10% APS, 10 µL TEMED.
24. Bioruptor (Diagenode, Luik, Belgium).

2.11 Labeling DNA Probes and EMSA

1. Labeled nucleotide γ -³²P-ATP with specific activities of 111 TBq/mmol. Caution: ³²P emits high-energy β radiation.
2. T4 polynucleotide kinase.

3. Kinase labeling buffer: 70 mM Tris-HCl pH 7.6, 10 mM MgCl₂, 5 mM DTT.
4. Gel solubilization buffer: 0.5 M ammonium acetate, 10 mM magnesium acetate, 1 mM EDTA pH 8.0, 0.1% (w/v) SDS.
5. 1× Binding Buffer: 20 mM Hepes pH 7.9, 40 mM KCl, 2 mM MgCl₂, 0.5 mM DTT, 0.5 mM EDTA, 5% v/v glycerol, 0.1–0.2 µg/µL poly(dI-dC), 0.05% NP-40.
6. Poly (deoxyinosinic–deoxycytidylic) acid sodium salt (dI-dC), stock solution of 200 mg/mL in water stored at –20 °C.
7. 1× Supershift Binding Buffer: 10 mM Tris-HCl, pH 7.5, 25 mM NaCl, 0.5 mM MgCl₂, 0.5 mM EDTA, 4% glycerol.
8. 40% (w/v) Acrylamide/bis-acrylamide stock solution (19:1); solution stored at 4 °C.
9. TBE 0.5×: Tris 6.05 g, boric acid 2.75 g, EDTA 0.465 g, add water up to 1 L and autoclave.
10. 3MM Whatman paper.
11. Gel dryer.
12. X-ray film (BioMax MR-film, Kodak) and cassettes for autoradiography.

3 Methods

3.1 PCR, Gel Purification of the PCR Products, Cloning in pUCPMAGUS Protoplast Expression Vector, and Transformation into *E. coli* Cells

1. For the 3'-end deletion analysis of the CmYLCV promoter, an 865-bp synthetic segment of CmYLCV genome (genomic coordinates 5700–6565, GenBank Accession No. NC004324) is designed and then synthesized (<https://www.thermofisher.com/us/en/home/lifescience/cloning/gene-synthesis/geneart-gene-synthesis.html>). This synthesized fragment is cloned subsequently at the corresponding *Eco*RI and *Hind*III sites of the protoplast expression vector pUCPMAGUS [9] using T4 DNA Ligase following manufacturer's instructions and is named as pUCPMAGUS-CmYLCV.
2. Respective CmYLCV promoter fragments are generated by PCR amplification using appropriately designed PCR primers (Table 1) to introduce an *Eco*RI site at the 5'-end and a *Hind*III site at the 3'-end of amplified products.
3. PCR amplifications are carried out in a total volume of 50 µL containing 50 ng plasmid DNA (pUCPMAGUS-CmYLCV), 0.4 µM primer pair for each specific fragments (Table 1), 200 µM dNTPs, and 2 U of PfuUltra high-fidelity DNA polymerase. PCR amplification is performed with an initial denaturation (94 °C for 2 m) followed by 30 cycles under the following conditions: denaturation (94 °C for 30 s), annealing (55 °C for 40 s) and extension (68 °C for 1 m), with a final extension at 68 °C for 10 m.

4. Each PCR fragment is gel-purified using a gel extraction kit (QIAGEN, USA), digested with *Eco*RI and *Hind*III restriction enzymes for 1 h and, subsequently cloned in the cloning vector pUC119 at the corresponding sites. Sequence integrity is confirmed before further use. The following 5'-end and 3'-end deletion plasmids are generated, with the 5'- and 3'-coordinates of promoter fragments with respect to the transcription start site (TSS) provided in parentheses: CmYLCV1.11 (-729 to +137), CmYLCV2.11 (-679 to +137), CmYLCV3.11 (-629 to +137), CmYLCV4.11 (-579 to +137), CmYLCV5.11 (-529 to +137), CmYLCV6.11 (-479 to +137), CmYLCV7.11 (-429 to +137), CmYLCV8.11 (-379 to +137), CmYLCV9.11 (-329 to +137), CmYLCV1.10 (-729 to +102), CmYLCV7.10 (-429 to +102), CmYLCV1 (-405 to +72), CmYLCV2 (-371 to +72), CmYLCV3 (-321 to +72), CmYLCV4 (-271 to +72), CmYLCV5 (-221 to +72), CmYLCV6 (-171 to +72), CmYLCV7 (-121 to +72), CmYLCV8 (-71 to +72), CmYLCV9 (-21 to +72), CmYLCV10 (-405 to +30), CmYLCV11 (-405 to +10), CmYLCV12 (-405 to -21), CmYLCV13 (-405 to -71), CmpC (-341 to +5), and CmpS (-341 to +59) [16, 19]. All PCR primers are listed in Table 1.
5. Subsequently, *Eco*RI and *Hind*III digested PCR products are ligated at the corresponding sites of the protoplast expression vector pUCPMAGUS [9] using T4 DNA Ligase following manufacturer's instructions.
6. For transformation, 10 μ L of the ligated product is added to 100 μ L of chemically competent TB1 cells, incubated on ice for 30 m, heat shocked for 2 m in a 37 °C water bath, and kept for 2 m on ice. Then 1 mL of LB medium is added to the tube and incubated at 37 °C with continuous shaking (200 rpm). After 1 h the cells are pelleted in a microcentrifuge at 3500 $\times g$ for 5 m, plated on LB plate containing 100 μ g/mL ampicillin, and placed in a 37 °C incubator overnight.
7. For making pKYLX71GUS-CmYLCV promoter fragment constructs, the CmYLCV promoter fragments from pUCPMAGUS are gel purified (as *Eco*RI-*Hind*III fragment) and ligated at the corresponding sites of a plant expression vector pKYLXGUS [9, 33] using T4 DNA Ligase following manufacturer's instructions.
8. For transformation, 10 μ L of the ligated product is added to 100 μ L of chemically competent TB1 cells, incubated on ice for 30 m, heat shocked for 2 m in a 37 °C water bath, and kept for 2 m on ice. Then 1 mL of LB medium is added to the tube and incubated at 37 °C with continuous shaking (200 rpm). After 1 h the cells are pelleted in a microcentrifuge at 3500 $\times g$ for 5 m, plated on LB plate containing 15 μ g/mL tetracycline and placed in a 37 °C incubator overnight.

3.2 Plasmid Isolation and Sequencing of the *CmYLCV* 3' and 5' Deletion Plasmid Constructs

1. The transformants are grown in 2 mL of LB medium containing 100 µg/mL ampicillin at 37 °C overnight. Plasmids are isolated from the overnight cultures and dissolved in water.
2. For confirming the sequence of the various pUCPMAGUS-*CmYLCV* promoter constructs, sequencing is done using, e.g. GenomeLab DTCS quick start kit (Beckman Coulter, USA). PCR is performed in a total volume of 20 µL containing 300 ng of plasmid, 1 µL of primer (10 µM solution), and 8 µL of reaction mix (GenomeLab DTCS quick start kit). The PCR products are purified and dissolved in sample loading solution according to the manufacturer's instructions. Sequencing to verify the *CmYLCV* 3' and 5' deletion fragments is performed in Beckman Coulter sequencer CEQ-8000. The sequencing data are analyzed and plasmids with *CmYLCV* 3' and 5' deletion fragments are selected for protoplast electroporation.

3.3 Isolation of Tobacco Cell Suspension Protoplasts and Electroporation of Plasmids

1. Tobacco cell suspension cultures (cultivar Xanthi) are maintained by subculturing to fresh medium at 4-day intervals (5 mL of culture is transferred to 50 mL medium). Three-day-old cell culture is used for protoplast isolation (*see Note 3*). The cells are harvested in a 50 mL tube by centrifugation at $110\times g$ for 4 m.
2. The culture medium is replaced with 30 mL enzyme solution and the cell suspension is transferred to a 250 mL flask, incubated at 26 °C in dark with slow shaking (50 rpm). After 2 h, the cell suspension is transferred to a 50 mL tube, centrifuged at $110\times g$ for 3 m, and the enzyme solution is carefully removed (*see Note 4*).
3. The protoplasts are washed once with 20 mL of MMC solution and then resuspended in 10 mL of MMC solution. Protoplasts are subsequently layered carefully on 25% sucrose solution and centrifuged for 4 m at $110\times g$. Upon settling, protoplasts form a ring at the interface of sucrose and MMC solution, which is carefully recovered. They are transferred to a fresh tube and resuspended in electroporation buffer.
4. An aliquot of 750 µL, containing approximately 2×100000 protoplasts in an electroporation cuvette (0.4 cm gap; 200 V and 950 µF; Bio-Rad, USA), is electroporated with 5–10 µg each of the reporter (GUS) plasmid and 5 µg of the internal control plasmid (pUCPMA-Lux-*CaMV35S*) [34]. A plasmid containing the firefly luciferase coding sequence under the control of the *CaMV35S* promoter and *rbcS* terminator (pUCPMA-Lux-*CaMV35S*) is co-electroporated as an internal control [34] and a plasmid containing GUS under the control of the *CaMV35S* promoter and *rbcS* terminator (pUCPMAGUS-*CaMV35S*) is electroporated as a positive/reference control [5]. Protoplasts are transferred to a new tube, centrifuged at $200\times g$ for 3 m, and electroporation buffer is carefully removed.

5. The protoplasts are resuspended in 1 mL of culture medium and plated onto a 25 mm plate coated with agarose. After incubation at 26 °C for 20–22 h, the protoplasts are harvested to measure GUS and luciferase activities [34].
6. The protoplasts are harvested in 1.5 mL tubes by centrifugation at $1000\times g$ for 3 m in a microcentrifuge and the culture medium is carefully removed.
7. To each tube, 100 μ L of $1\times$ lysis buffer is added, vortexed for 30 s to break the cells, centrifuged for 2 m at maximum speed, and the supernatant is carefully transferred to a fresh tube.
8. Aliquot 50–100 μ L of GUS assay reagent to each test tube, place tubes into the 37 °C water bath, add 10–20 μ L of protein sample to the pre-warmed test tube at 37 °C and note time. Add next sample at a convenient time interval (*see* Subheading 3.7.1 for details).
9. Add 1.0 mL Stop solution to each tube after sufficient time has elapsed. Add the Stop solution at the same time intervals at which the protein was added to the GUS assay reagent containing tubes (*see* Subheading 3.7.1 for details). The fluorescence is measured using a fluorometer (e.g. TKO 100 fluorometer, Hoefer Scientific Instruments, USA) (*see* Subheading 3.7.1 for details).
10. Luciferase activity in transfected protoplasts is measured using a luciferase assay system (e.g. Promega, Madison, USA) following manufacturer's instructions. The cell lysate (10–20 μ L) (prepared in previous **step 7**) is added to luciferase assay reagent (50–100 μ L) and the luminescence is measured in a luminometer (e.g. Turner Designs, USA).
11. GUS activity is normalized against luciferase activity and expressed as fold activation relative to control. All constructs are tested in at least three independent experiments.
12. We use pUCPMA-CaMV35S-GFP [5] to calculate protoplast transformation efficiency by counting the protoplasts expressing GFP using hemocytometer under fluorescent microscope [5, 36]. We usually have transformation efficiency for tobacco protoplasts (approximately 2×1000.00 protoplasts taken for electroporation) ranging from 70% to 80% using 10 μ g of the reporter (GFP) plasmid (pUCPMA-CaMV35S-GFP) and 5 μ g of the internal control plasmid (pUCPMA-Lux-*CaMV35S*) [34].

3.4 Biolistic-Onion Peel Transient Assay

Onion tissues are prepared and bombarded with the expression vector pUCPMAGUS containing CmYLCV4, CmYLCV9.11, CmYLCV13, CmpC, CmpS, and CaMV35S promoters following a standard protocol [37]. After 2 days, transient GUS expression is detected by a histochemical method [38, 39] and visualized under an Olympus SZX12 bright-field microscope.

3.4.1 Preparation of Particles

1. Weigh out 10 mg of gold particles (for nine shots) in a microfuge tube. The gold particles can be either 0.6 μm or 1.0 μm in diameter (Biorad, USA).
2. Add 1.0 mL of 100% ethanol into it and vortex for at least 10 s (three times) and incubate in ice for 30 s.
3. Centrifuge for 5 m at $4500\times g$ and remove supernatant. Add 115 μL of 100% ethanol and vortex for 1 m with speed setting at 5, push tube far into holder to particles stay more in bottom.
4. Divide into 35 μL aliquots. Gold must be mixed continuously between aliquots. Centrifuge for 10 s at $4500\times g$.

3.4.2 Precipitation of Gold/DNA

1. Add carefully 1.0 mL sterile water per aliquot (for three shots) without suspending, centrifuge for 5 m at $850\times g$ and remove the supernatant.
2. Vortex for 10 s with speed setting at 3 while adding each of ingredient 12.5 μL DNA (4–5 μg), 220 μL sterilized water, 250 μL 2.5 M CaCl_2 , 100 μL 0.1 M spermidine. Keep in ice for 2 m. Vortex for 10 m with speed setting at 2 and centrifuge for 5 m at $70\times g$. Remove supernatant and quickly add 100 μL of 100% ethanol.
3. Do not allow pellet to dry and vortex for 1 m with speed setting at 3 and centrifuge for 1 m at highest speed. Discard supernatant and add 36 μL of 100% ethanol, mix and incubate for 1 h on ice.
4. Resuspend with pipettor before use. Use 10 μL particle suspension per macrocarrier.

3.4.3 Preparation of Onion Epidermal Peels

1. *Allium cepa* (e.g. cv. Tango) bulbs are purchased from a local farmer's market.
2. Under sterile conditions, inner epidermal peels (2×2 cm) are placed on solid agar plates.
3. Peels are bombarded within 1 h of transfer to agar plates.

3.4.4 Preparation of Disks

1. This is a good time to place macrocarriers on filter paper kept in Petri dish. Use 2 or 3 macrocarriers per construct; use millipore forceps.
2. Place 10 μL particle suspension on each macrocarrier, there should be enough particles for three shots. It is usually a good idea to do three shots per construct in case one goes wrong.
3. Cover macrocarrier dishes and let ethanol evaporate and as soon as the ethanol is evaporated you are ready to shoot.

3.4.5 Bombardment

1. You will need rupture disks, stopping screens, prepared macrocarriers, prepared tissues, millipore forceps.

2. To open helium tank, open large gray screw by 1/4 turn, screw in brass regulator until secondary pressure reaches 1300 psi, i.e. ~200 psi above the value of the rupture disks.
3. Start vacuum pump and turn on gene gun.
4. The macrocarrier assembly should be on the top shelf and the sample tray should be on the second from the bottom. Remove macrocarrier assembly, load rupture disk. Make sure holder is firmly tightened or it will blow too soon. Open macrocarrier assembly (big screw on top), remove macrocarrier holding ring and place macrocarrier in ring, particles should be facing up. Place stopping screen in bottom of assembly invert and replace ring with macrocarrier, particles should be facing down. Slide macrocarrier assembly into top shelf, place sample on lower shelf. Do not forget to open the dish, close door.
5. Set vacuum switch to top position (VAC), pull vacuum to about 27.5, flip vacuum switch to lowest position (HOLD). Press (and hold) FIRE switch. Pressure in upper chamber will rise to ~1100 psi before disk ruptures.
6. Release FIRE switch, vacuum switch to BLEED, open chamber and remove sample, macrocarrier, stopping screen and rupture disk.
7. Repeat for other samples. Second and third shots for the same constructs can be done with the same stopping screen; it is helpful to invert it between shots and duplicates can be bombarded into the same tissue sample.
8. After bombardment, return all onions to Petri dishes. Cover the Petri dishes with lids. Wrap them with Parafilm to prevent drying. Incubate the onion pieces at room temperature in the dark for 16–48 h.
9. Make sure to turn off gas and release pressure before detaching the gene gun from the helium tank. The steps to shut down are: first clean the gene gun, pull vacuum to ~25 psi and hold. Close big gray screw on helium tank, loosen brass regulator until it turns freely press FIRE on gene gun to release pressure. Keep an eye on the manometers of the helium tank, only the secondary should drop down to zero. Release the vacuum by switching to BLEED, set open door, set vacuum switch to VAC, turn off vacuum pump, vacuum switch to bleed and turn off gene gun.

3.4.6 Observation

1. After incubating bombarded onion pieces for 16–20 h at room temperature, you can proceed to the GUS histochemical staining to observe expression of GUS.
2. Use forceps with flat ends to slowly peel a single cell layer off the inner epidermis of the onion (the layer that directly faces the gun during the bombardment) and proceed for GUS histochemical staining (follow Subheading 3.7.2 for the detailed steps).

3. Under bright field of the microscope, observe histochemically stained cells for GUS localization and expression.

3.5 Transient Expression of CmYLCV Promoters in Tobacco Leaves by Agro-Infiltration

3.5.1 Preparation of *Agrobacterium* Cultures for Agroinfiltration

The CmYLCV promoter fragments are cloned as *EcoRI*–*HindIII* fragments into the plant expression vector pKYLX71GUS [9, 33]. The pKYLX71GUS-based plant expression vectors are mobilized into the *Agrobacterium tumefaciens* strain C58C1:pGV3850 [6]. Suspensions of *Agrobacterium* strains that harbored individual plant expression constructs are infiltrated into leaves of *Nicotiana benthamiana* as described previously [40]. The details are described below:

1. The pKYLX71GUS-CmYLCV promoter fragment constructs (Subheading 3.1, step 7) are introduced into *Agrobacterium tumefaciens* strain C58C1:pGV3850 by freeze–thaw method [6]. Grow recombinant *A. tumefaciens* overnight at 28 °C in 100 mL conical flask containing 10 mL of LB medium supplemented with 50 µg/mL kanamycin.
2. Aliquote of 50 µL of this overnight culture is used for inoculation of 10 mL of LB medium supplemented with 10 mM MES buffer, pH 5.7, 50 µg kanamycin per mL and 150 µM acetosyringone (3,5-dimethoxy-4'-hydroxy-acetophenone) [40].
3. Grow the precultures overnight at 28 °C in a shaker and harvest cells by centrifugation and resuspend to a final concentration corresponding to an optical density (OD) of 1.0 [40] at 600 nm in a solution containing 10 mM MgCl₂, 10 mM MES pH 5.7, and 150 µM acetosyringone (unless stated otherwise).
4. Incubate cultures at room temperature for 3 h before infiltration. Infiltrate two to three top-leaves per plant with a 2 mL syringe without a needle. Leaves can be superficially wounded with a needle to improve infiltration. Three plants are agroinfiltrated for each construct.
5. After 2 days of agro-infiltration, the transient GUS expression is evaluated by the histochemical GUS staining method [38, 39]. The infiltrated leaves are taken for GUS histochemical staining (follow Subheading 3.7.2 for detailed steps).

3.6 Preparation of the Plant Transformation Vector and the Generation of Transgenic Tobacco and *Arabidopsis* Plants

The CmYLCV promoter fragments are cloned as *EcoRI*–*HindIII* fragments into the previously described plant expression vector pKYLX71GUS [9, 33]. The pKYLX71GUS-based plant expression vectors are mobilized into the *Agrobacterium tumefaciens* strain C58C1:pGV3850. Tobacco (*Nicotiana tabacum* cv. Samsun NN) leaf discs are transformed with engineered *Agrobacterium* as described previously [6]. *Arabidopsis* transformation is performed using pKYLX71GUS-based CmYLCV9.11 promoter construct using the floral dip method [41]. The generation of transgenic *Arabidopsis* plants (*Arabidopsis thaliana* ecotype Columbia-0) and

tobacco transgenic lines (*Nicotiana tabacum* cv Samsun NN) and their maintenance are performed following the published procedures as described previously [18, 22, 26].

3.6.1 Transformation of Tobacco Plants

1. Culture a single colony of engineered *Agrobacterium* in liquid LB + 50 mg/L kanamycin + 100 mg/L rifampicin and shaking at 300 rpm for 2 days at 28 °C in the dark. Do a mini-plasmid-prep with 5 mL of the culture and check by restriction digestion.
2. Collect bacterial cells by a centrifuge at 2500 × *g* for 1 m, discard the liquid and suspend the bacterial pellet in liquid TOM medium. Dilute the suspension to an OD₆₀₀ of 0.5–0.8 and pour into an empty sterile Petri plate.
3. For explants preparation, take healthy fully expanded leaves from 4 to 5 weeks old, pre-sterilized tobacco plants (aseptically grown plants in condos) with sterile forceps and scalpels and cut into 0.6–0.8 cm² (or can use a cork borer, which is about 1.0 cm diameter) in the *Agrobacterium* suspension.
4. Transfer the explants on TKM medium (TOM medium with 250 mg/L kanamycin for transformant selection and 500 mg/L cefotaxime for killing excess *Agrobacterium* around the culture). Subculture the explants to fresh selection medium twice during the first week and once per week starting the following week.
5. Transfer the explants (abaxial surface of the explants in contact with the medium), to the TOM plates for 2 days without antibiotics at 25 °C. This enhances tissue infection.
6. Callus should appear 3 weeks after the initial infection, with plantlets developing soon after. Once the plantlets are large enough transfer the whole explants together with the shoots to selection rooting-medium (T- + 250 mg/L kanamycin + 500 mg/L cefotaxime) for rooting. Culture the shoots at a 16 h photoperiod for 3 weeks and after roots have generated, transfer the rooted plants to soil in the greenhouse.
7. For each promoter construct, approximately ten independent plant lines should be generated by kanamycin (250 mg/L) selection and grown in greenhouse conditions.

3.6.2 The Generation of Transgenic Arabidopsis Plants

1. Grow healthy *Arabidopsis* plants (*Arabidopsis thaliana* ecotype Columbia-0) until they are flowering under long days in pots. To encourage proliferation of many secondary bolts, clip first bolts. Plants will be ready roughly 4–6 days after clipping. Clipping can be repeated to delay plants. Optimal plants should have many immature flower clusters and not many fertilized siliques (it is recommended to remove the siliques), although a range of plant stages can be successfully transformed.

2. Prepare *Agrobacterium tumefaciens* strain carrying gene of interest on a binary vector. Grow a large liquid culture at 28 °C in LB with antibiotics to select for the binary plasmid. The mid-log cells or a recently stationary culture can be used. Spin down *Agrobacterium*, resuspend to $OD_{600} = 0.8$ (can be higher or lower) in 5% Sucrose solution (if made fresh, no need to autoclave). You will need 100–200 mL for each two or three small pots to be dipped, or 400–500 mL for each two or three 3.5" (9 cm) pots.
3. Before dipping, add Silwet L-77 to a concentration of 0.05% (500 μ L/L) and mix well. If there are problems with L-77 toxicity, use 0.02% or as low as 0.005%. Dip above-ground parts of plant in *Agrobacterium* solution for 10–15 s, with gentle agitation. You should then see a film of liquid coating plant. Some investigators dip inflorescence only, while others also dip rosette to hit the shorter axillary inflorescences.
4. Place dipped plants under a dome or cover for 16–24 h to maintain high humidity (plants can be laid on their side if necessary). Water and grow plants normally, tying up loose bolts with wax paper, tape, stakes, twist-ties, or other means. Stop watering as seeds become mature. Harvest dry seed.
5. Select for transformants using antibiotic or herbicide selectable marker. Transformants are usually all independent, but are guaranteed to be independent if they come off of separate plants. After sterilizing (steps are: soak seeds in water for 30 m, 95% in ethanol for 5 m, 10% Clorax solution for 5 m and rinse in water for 1 m for 4 times) plate 40 mg = 2000 seeds (resuspended in 4 mL 0.1% agarose) on 0.5 \times MS/0.8% tissue culture Agar plates with 50 μ g/mL kanamycin, cold treat for 2 days, and grow under continuous light (50–100 microEinsteins) for 7–10 days. Transplant putative transformants to soil to grow and test by GUS assay and molecular analysis (*see Note 5*).

3.7 β -Glucuronidase (*uidA* or *GUS*) Assay and Histochemical *GUS* Staining

In this study, the *GUS* reporter gene is used to monitor and analyze the synthetic CmYLCV promoter activity in both stable and transient systems. Fluorometric *GUS* enzymatic assays for measuring *GUS* activities in tobacco protoplast extracts, *Arabidopsis*, and tobacco plant extracts are performed as described previously [38]. The total protein content in protoplast and plant extracts is estimated by the Bradford method using BSA as a standard [42].

3.7.1 β -Glucuronidase (*uidA* or *GUS*) Assay

1. Label all tubes. Prepare solutions and have ready at hand. Remove the tissue from the –80 °C freezer and thaw on ice. If the tissue is fresh, keep on ice (or alternatively work in a cold room).
2. Place tissue in a mortar and pestle and add ~2 mL of QB/g tissue, grind and transfer it into a microfuge tube (*see Note 6*).

3. Spin samples at top speed in the microfuge (4 °C for 15 m) and transfer the liquid supernatant into a second (new) microfuge tube and store samples in the -80 °C.
4. Remove the protein samples from the -80 °C freezer and thaw on ice, label all test tubes in which assay will be performed and set circulating water bath to 37 °C. Aliquot 400 µL of GUS assay reagent to each test tube, place tubes into the 37 °C water bath, add 5 µL of protein sample to the pre-warmed test tube at 37 °C and note time. Add next sample at a convenient time interval. (For example: add protein sample #1 to the first test tube containing GUS assay reagent and at 15 s add protein sample #2 to the second tube containing GUS assay reagent and at 30 s add protein sample #3 to the third tube containing GUS assay reagent, etc. until the entire set to be analyzed has been added to the GUS assay reagent containing tubes).
5. Add 1.6 mL Stop solution to each tube after sufficient time has elapsed. Add the Stop solution at the same time intervals at which the protein was added to the GUS assay reagent containing tubes (For example if protein sample was added to the tubes containing GUS assay reagent at 0, 15, 30, 45, 60 ... s, add stop solution at 15 m 0 s, 15 m 15 s, 15 m 30 s, 15 m 45 s, ... likewise).
6. Turn on the fluorometer (e.g. TKO 100 fluorometer, Hoefer Scientific Instruments, San Francisco, USA) 15 m (or more) before use. Fill the glass cuvette with 1.9 mL carbonate stop buffer. Place the cuvette into the sample chamber, close the lid and adjust the scale knob until the display reads zero.
7. Take 100 µL from each of the different MU standard solutions/incubated sample into the cuvette containing the 1.9 mL carbonate stop solution and mix (by inversion or up and down pipetting). Place the cuvette into the chamber and take readings to plot the standard curve. The time interval can be variable (the final result is expressed as a function of time: pmol product (4MU) released per minute per mg of protein) but should be linear over time; i.e., the relative fluorescence at 10 m should be two times the relative fluorescence at 5 m. In order to accomplish this requirement it may be necessary to dilute the protein sample an order of magnitude or more (i.e., 10× dilution, 100× dilution, etc.). In order to determine that the GUS assay is linear over time, it is necessary to perform a trial assay before beginning with the assay for all samples.
8. Bradford protein concentration determination assays: Measure the protein concentration in the extract using, e.g. Quick Start Bradford Protein assay kit (Bradford, USA) following manufacturers' instructions, based on the dye-binding assay of Bradford [42]. Dilute the Bradford reagent fivefold in dH₂O (1 part Bradford:4 parts dH₂O). Add 5–20 µL of the protein extract to

1 mL of the diluted reagent and mix. Measure the blue color formed at 595 nm using a Spectrophotometer. Use disposable plastic cuvettes to prevent the formation of a blue film. Prepare a standard curve using a serial dilution series (0.1–1.0 mg/mL) of a known protein sample concentration, e.g. BSA dissolved in QB. Determine the protein concentration of the plant extract from the regression curve of the known sample. Express the results as pmol product released per minute per mg of protein.

3.7.2 Histochemical *GUS* Staining Procedure

Histochemical *GUS* staining is carried out in plants following the published protocol [22, 38], and photographs are taken under a bright-field microscope. The detailed steps are described below:

1. Fix for 30 m in ice cold fixative, shaking occasionally. Wash for 30–60 m in several changes of ice cold 1× buffer. Alternatively, take fresh tissue for proceeding next step.
2. Vacuum infiltrate or incubate (for fresh tissues) in the X-gluc substrate medium in dark at room temperature or at 37 °C for several hours or overnight or until distinct blue staining appears (no longer than 24 h).
3. Rinse in distilled water.
4. Incubate green objects in 70% ethanol until the chlorophyll is removed, then transfer to distilled water again.
5. Optional: place specimens in 50% glycerol, for 1 h, then transfer to pure glycerol, again leave for 1 h or more. Mount objects in 100% glycerol on microscope slides, examine under bright-field microscope (Olympus SZX12).

3.8 RNA Isolation and Quantitative Real-Time PCR (qRT-PCR)

1. Total RNA is isolated from 4-week-old (R_1 progeny, second generation) tobacco seedlings using a Plant RNA extraction kit (Plant RNeasy extraction kit, Qiagen, CA, USA) following the manufacturer's specification.
2. Two μg of total RNA in 20 μL reaction volume is taken for synthesizing cDNA by using cDNA synthesis kit (iScript cDNA synthesis kit, BioRad, USA).
3. For quantitative real-time PCR (qRT-PCR), manufacturers' instructions of iTaQ universal SYBR Green one-step kit (BioRad, USA) are followed essentially for 20 μL reaction volume. The qRT-PCR analysis is performed for relative quantification of *GUS*-specific transcript using *GUS*-specific forward (5'-d-TTACGTCCTGTAGAAACCCCA-3') and reverse (5'-d-ACTGCCTGGCACAGCAAT TGC-3') primers. The PCR reaction is performed with four replicates and is repeated with three biological samples. The tobacco α -tubulin gene-specific forward 5'-d-ATGAGAGAGTGCATATCGAT-3' and reverse 5'-d-TTCACTGAAGAAGGTGTTGAA-3' primers are used to normalize the amount of total mRNA in all samples.

4. The comparative threshold cycle (Ct) method (Applied Biosystems bulletin, part No. 4376784 Rev. C, 04/2007) is used to evaluate the relative expression levels of the transcripts. The threshold cycle is automatically determined for each reaction by the system set with default parameters (Step One Real-Time PCR System, Applied Biosystems). The specificity of the PCR is determined by melting curve analysis of the amplified products using the standard method installed in the system (Step One Real-Time PCR System, Applied Biosystems).

3.9 Preparation of Tobacco Nuclear Extracts for EMSA

1. Grind approximately 5 g of freshly collected tobacco (*Nicotiana tabacum* cv. Samsun NN) seedlings or leaves in a chilled mortar with 1 g of acid-washed sand as abrasive and 2.5 volumes of extraction buffer.
2. Filter the homogenate through two layers of Miracloth (Calbiochem) into a chilled centrifuge tube.
3. Centrifuge the filtrate at $4300 \times g$ for 10 m at 4 °C.
4. Resuspend the pellet gently, containing the nuclei, in 500 μ L of protein isolation buffer (100 μ L per gram of fresh weight tissue).
5. Maintain the resuspended pellet at 4 °C for 40 m, stirring every 6 m by vortexing for 5 s.
6. Aliquot into Eppendorf tubes and centrifuge at $12,000 \times g$ for 15 m at 4 °C. Recover the supernatant by discarding the pellet.
7. Dialyze the extract two times in dialysis buffer for overnight.
8. Concentrate the sample using an Amicon® Ultra 4 mL Centrifugal Filters.
9. Take out the crude nuclear protein extract (supernatant) and adjust it to 20% (v/v) glycerol and divide the resulting supernatants containing nuclear proteins into small aliquots and store at -70 °C until use.
10. Determine protein concentrations using a Quick Start Bradford Protein assay kit with BSA as the standard (as mentioned in Subheading 3.7.1, step 8).

3.10 Expression of Recombinant TGA1a Protein

A synthetic gene encoding *Nicotiana tabacum* TGACG sequence-specific DNA binding protein (TGA1a; Accession No. X16449), optimized with an *Escherichia coli* (*E. coli*) bias codon and a C-terminal 6 \times his-tag (5'-*Nde*I-TGA1a-6 \times his-tag-*Sst*I-3'), is cloned into the *E. coli* expression vector pET-29b (Invitrogen) at *Nde*I/*Sac*I sites (see Note 7). Plasmid pTf-16 encoding chaperone Trigger Factor (*tig*) is purchased from Takara (Japan), which carried an origin of replication derived from pACYC and a chloramphenicol resistance gene (Cm^r), and the chaperone gene is located at downstream of the *araB* promoter. At the first stage chaperone plasmid is separately transformed into the chemically competent BL21 (DE3) cells, subsequently

TGA1a expression plasmid (pET-29b- TGA1a) is also transformed into the cells bearing chaperone expression plasmids and plated on LB-agar containing 50 µg/mL kanamycin and 20 µg/mL chloramphenicol. The resultant clone is designated TGA1a/pTf16.

1. Pick up a single colony from the transformed BL21- TGA1a/pTf16 plate and inoculate into 5 mL of LB broth, supplemented with 50 µg/mL kanamycin and 20 µg/mL chloramphenicol.
2. Grow overnight at 37 °C under shaking at 200 rpm.
3. Dilute 1/50 into 500 mL of LB broth supplemented with the same antibiotics.
4. At the same time, add 2 mg/mL L-arabinose for induction of tig chaperone encoded by the pTf-16 plasmid.
5. The culture is grown at 37 °C shaking with 200 rpm.
6. When OD₆₀₀ reaches 0.4 reduce the temperature to 25 °C (*see Note 8*).
7. At mid-log phase at OD₆₀₀ = 0.6, induce the protein expression by adding 0.4 mM IPTG (from 1 M stock solution).
8. Incubate for 8 h at 25 °C.
9. Harvest the bacterial cells by centrifugation 10 m at 5000 × *g* at 4 °C. From here on perform all steps at 4 °C or on ice.
10. Resuspend the pellet in 10 mL B-PER™ Bacterial Protein Extraction Reagent (Thermo Scientific) supplemented with lysozyme (0.2 mg/mL), and lyse the cells by sonication in a Bioruptor with ten cycles of 40 s at amplitude 40 and 20 s rest on ice between cycles.
11. The lysate is centrifuged at 27,000 × *g* for 45 m at 4 °C.
12. Keep the clear supernatant by passage through a filter with 0.45-µm pore size.
13. Pack the 2 mL column with an appropriate amount of HisPur™ cobalt resin (depending on the lysate volume). Allow the storage buffer to drain from resin by gravity flow (*see Note 9*).
14. Mix the protein extract with an equal volume of Equilibration/Wash Buffer.
15. Equilibrate the resin with two resin-bed volumes of Equilibration/Wash Buffer. Allow buffer to drain from resin, flow rate should be 0.5–1 mL/m.
16. Add the prepared protein extract (from **step 14**) to the equilibrated resin in the tube and mix on an end-over-end rotator for 30 m. Collect the flow-through fraction in a tube.
17. Wash resin three times with two resin-bed volumes of Equilibration/Wash Buffer and collect the flow-through fractions. Repeat this step using a new collection tube until the absorbance of the flow-through fraction at 280 nm approaches baseline.

18. Elute His-tagged proteins from the resin with two resin-bed volumes of Elution Buffer. Repeat this step twice, collecting each fraction in a separate tube.
19. Pool all the eluted fractions into one. Dialyze the pooled eluted fractions overnight at 4 °C against dialysis buffer.
20. Concentrate the sample using an Amicon® Ultra 4 mL Centrifugal Filters to a concentration of 6–7 mg/mL.
21. The quality of TGA1a can be verified by SDS-PAGE, the quantity can be determined with the Bradford protein assay.

3.11 Electrophoretic Mobility Shift Assay

3.11.1 Labeling of Probe Through PCR

1. Mix 5 pmol of forward primer (5'-TGAAGGCATCTTCAGACTCC-3') with 2 µL of [γ -³²P] ATP (specific activity 3000 Ci/mmol), 1 unit of T4 polynucleotide kinase in the labeling buffer, in a final volume of 10 µL.
2. Incubate the tube at 37 °C for 1 h. Stop the reaction by adding 1 µL of 0.5 M EDTA–NaOH (pH 8.0).
3. Unincorporated nucleotides are removed by gel chromatography (G25 spin columns; Pharmacia).
4. PCR amplifications are carried out in a total volume of 50 µL containing 50 ng plasmid DNA (CmYLCV9.11-GUS), 5 pmol labeled forward primer, 10 pmole of unlabeled reverse primer (5'-GTATTTATAGACTGACGGGTGAGTGG-3'), 200 µM dNTPs, PCR buffer (containing MgCl₂) and 2 U of Taq DNA polymerase. PCR amplification is performed on a Peltier thermal cycler (MJ Research, USA) with an initial denaturation (95 °C for 5 m) followed by 35 cycles under the following conditions: denaturation (95 °C for 30 s), annealing (55 °C for 30 s) and extension (72 °C for 1 m), with a final extension at 72 °C for 10 m.

3.11.2 Probe Isolation from PAGE

1. Clean the glass plates with water (and with soap, if necessary) and then with absolute ethanol.
2. Assemble glass plates with 1.5 mm spacers.
3. Prepare a 6% polyacrylamide (acrylamide: bisacrylamide = 19:1) gel by mixing 30 mL of gel mix (or as much as needed for the particular set-up used), 0.5× TBE with 50 µL TEMED and 500 µL 10% (w/v) APS.
4. Mix by swirling and pour between the plates.
5. Insert a 1.5 mm thick comb with 8 mm wide teeth.
6. Let the gel polymerize for 30 m and remove the comb.
7. Mount the gel in an electrophoresis apparatus and add 0.5× TBE to the electrophoresis tanks.
8. To the PCR amplified DNA, add 0.25 volumes 5× loading mix and load in separate wells.
9. Run the gel at 100 V.

10. After running, open the plates and transfer the gel onto Whatman paper 3MM, cover with a Saran Wrap and expose it to an X-ray film in an autoradiography cassette mounted with a tungstate intensifying screen for 5 m.
11. Develop the X-ray film in Developer and Fixer.
12. After drying the X-ray film, align the X-ray film where the PCR-amplified bands appear with that of the gel mounted on Whatman paper 3MM.
13. Cut out the bands of interest with a scalpel/razor blade.
14. Incubate the acrylamide slices with the fragments of interest in a 1.5 mL Eppendorf tube with 0.5 mL of gel solubilization buffer with gentle shaking overnight.
15. Next day centrifuge the mixture at maximum speed for 10 m at room temperature.
16. Transfer the supernatant to a new Eppendorf tube, discarding the gel-pieces.
17. Add 1 mL 95% ethanol and centrifuge for 5 m at $15,000 \times g$.
18. Remove liquid and add 0.5 mL 70% ethanol to the pellet.
19. Centrifuge for 5 m at $15,000 \times g$.
20. Remove the liquid.
21. Air-dry the pellet at room temperature or dry it in a speedvac apparatus without heating.
22. Resuspend the pellet in 20 μ L nuclease-free water.
23. Count probe and then put in the radioactive storage box at -20 °C.

3.11.3 EMSA

1. Prepare a 6% native PAGE as described in Subheading 3.11.2, **step 3**.
2. The tobacco protein extracts or purified TGA1a are incubated for 20 m on ice with $1 \times$ binding buffer before the addition of radiolabeled probe (20,000 cpm) and incubation is further continued for 30 m at room temperature.
3. To test the specificity of the DNA–protein-binding reactions, add excess unlabeled PCR-amplified DNA (5- to 100-fold) to the reaction mixture 15 m before adding the labeled probe (*see Note 10*).
4. Adjust the volume of each tube to 20 μ L of reaction volume with binding buffer.
5. The gel is pre-run for 30 m at 120 V (= 10 V/cm).
6. Subject DNA–protein complexes to electrophoresis in 6% PAGE with $0.5 \times$ TBE buffer at 120 V.

7. The presence of TGA1a in the shifted band for tobacco nuclear extract can be demonstrated by a super shift induced by adding 2 μg of rabbit polyclonal TGA1a antiserum (in-house made rabbit antiserum against bacterial purified TGA1a). Antibodies are added to the nuclear extract either overnight/20 m before or 20 m after the ^{32}P -labeled probe. Samples are incubated for a total of 30 m at room temperature in 1 \times supershift binding buffer.
8. The reaction products are then loaded on a 4.5% polyacrylamide gel (1.5 mm thick) containing 0.5 \times TBE and the gel is run at 120 V in 4 $^{\circ}\text{C}$.
9. Gels are dried using a vacuum dryer for 1 h at 70 $^{\circ}\text{C}$.
10. Place the dried gel in a plastic folder and expose to an X-ray film in an autoradiography cassette mounted with a tungstate intensifying screen in a -80 $^{\circ}\text{C}$ freezer overnight.

4 Notes

1. Use a freshly prepared enzyme solution for protoplast isolation.
2. EDTA will not completely go into solution until the pH approaches 8.0 and the H_2O is almost at final volume. Essentially, the pH needs to be continuously adjusted as the EDTA dissolves.
3. A 3-day-old tobacco cell suspension is ideal for getting good protoplasts.
4. Protoplasts are delicate and must be handled carefully. Centrifugation at high speeds will damage the protoplasts.
5. For higher rates of Arabidopsis transformation, plants may be dipped two or three times at 7-day intervals. We suggest one dip 2 days after clipping, and a second dip 1 week later. Do not dip less than 6 days apart.
6. For small (<1 g) quantities of tissue, prepare a pestle by flaming the end of a blue pipette tip and sealing the end by gently smashing it into a microfuge tube while working in the fume hood. Prepare as many pestles as tissue samples to be isolated and grind tissue in QB.
7. For bacterial expression of TGA1a, we have employed pET-29b (Invitrogen) as the protein expression vector. However, selection of the protein expression vector is not that important as long as it is compatible with the BL21 strain.
8. Temperature selection for expressing recombinant TGA1a in *E. coli* is very important for growing the bacteria at low temperature and better yield of the proteins.

9. Fresh Co²⁺ resin bead is recommended for the purification of His-tagged TGAla protein for the highest protein yield.
10. During competitive EMSA analyses, unlabeled DNA duplex of the same length may be recommended for use for detecting the level of nonspecific protein binding.

References

1. Odell JT, Nagy F, Chua NH (1985) Identification of DNA-sequences required for activity of the cauliflower mosaic virus-35S promoter. *Nature* 313:810–812
2. Lam E, Chua NH (1989) ASF-2—a factor that binds to the cauliflower mosaic virus-35S promoter and a conserved GATA motif in cab promoters. *Plant Cell* 1:1147–1156
3. Pattanaik S, Dey N, Bhattacharyya S et al (2004) Isolation of full-length transcript promoter from the Strawberry vein banding virus (SVBV) and expression analysis by protoplasts transient assays and in transgenic plants. *Plant Sci* 167:427–438
4. Maiti IB, Shepherd RJ (1998) Isolation and expression analysis of peanut chlorotic streak caulimovirus (PCISV) full-length transcript (FLt) promoter in transgenic plants. *Biochem Biophys Res Commun* 244:440–444
5. Sahoo DK, Ranjan R, Kumar D et al (2009) An alternative method of promoter assessment by confocal laser scanning microscopy. *J Virol Methods* 161:114–121
6. Sahoo DK, Stork J, DeBolt S et al (2013) Manipulating cellulose biosynthesis by expression of mutant *Arabidopsis* gene in transgenic tobacco. *Plant Biotechnol J* 11:362–372
7. Sahoo DK, Sarkar S, Raha S et al (2015) Analysis of Dahlia mosaic virus full-length transcript promoter-driven gene expression in transgenic plants. *Plant Mol Biol Rep* 33:178–199
8. Banerjee J, Sahoo DK, Raha S et al (2015) A region containing an as-1 element of Dahlia mosaic virus (DaMV) subgenomic transcript promoter plays a key role in green tissue- and root-specific expression in plants. *Plant Mol Biol Rep* 33:532–556
9. Dey N, Maiti IB (1999) Structure and promoter/leader deletion analysis of mirabilis mosaic virus (MMV) full-length transcript promoter in transgenic plants. *Plant Mol Biol* 40:771–782
10. Bhattacharyya S, Pattanaik S, Maiti IB (2003) Intron-mediated enhancement of gene expression in transgenic plants using chimeric constructs composed of the Peanut chlorotic streak virus (PCISV) promoter-leader and the antisense orientation of PCISVORF VII (p7R). *Planta* 218:115–124
11. Bhattacharyya S, Dey N, Maiti IB (2002) Analysis of cis-sequence of subgenomic transcript promoter from the Figwort mosaic virus and comparison of promoter activity with the cauliflower mosaic virus promoters in monocot and dicot cells. *Virus Res* 90:47–62
12. Kroumova ABM, Sahoo DK, Raha S et al (2013) Expression of an apoplast-directed, T-phylloplanin-GFP fusion gene confers resistance against *Peronospora tabacina* disease in a susceptible tobacco. *Plant Cell Rep* 32:1771–1782
13. Verdaguer B, de Kochko A, Fux CI et al (1998) Functional organization of the cassava vein mosaic virus (CsVMV) promoter. *Plant Mol Biol* 37:1055–1067
14. Sahoo DK, Raha S, Hall JT et al (2014) Overexpression of the synthetic chimeric native-T-phylloplanin-GFP genes optimized for monocot and dicot plants renders enhanced resistance to blue mold disease in tobacco (*N. tabacum* L.). *ScientificWorldJournal* 2014, Article ID 601314. doi:10.1155/2014/601314
15. Sahoo DK, Dey N, Maiti IB (2014) pSiM24 is a novel versatile gene expression vector for transient assays as well as stable expression of foreign genes in plants. *PLoS One* 9:e98988. doi:10.1371/journal.pone.0098988
16. Sahoo DK, Sarkar S, Raha S et al (2014) Comparative analysis of synthetic DNA promoters for high-level gene expression in plants. *Planta* 240:855–875
17. Maiti IB, Ghosh SK, Gowda S et al (1997) Promoter/leader deletion analysis and plant expression vectors with the Figwort mosaic virus (FMV) full length transcript (FLt) promoter containing single or double enhancer domains. *Transgenic Res* 6:143–156
18. Sahoo DK, Maiti IB (2014) Biomass derived from transgenic tobacco expressing the *Arabidopsis CESA3ixr1-2* gene exhibits improved saccharification. *Acta Biol Hung* 65:189–204
19. Stabolone L, Kononova M, Pauli S et al (2003) Cestrum yellow leaf curling virus (CmYLCV)

- promoter: a new strong constitutive promoter for heterologous gene expression in a wide variety of crops. *Plant Mol Biol* 53:703–713
20. Patro S, Kumar D, Ranjan R et al (2012) The development of efficient plant promoters for transgene expression employing plant virus promoters. *Mol Plant* 5:941–944
 21. Patro S, Maiti IB, Dey N (2013) Development of an efficient bi-directional promoter with tripartite enhancer employing three viral promoters. *J Biotechnol* 163:311–317
 22. Kumar D, Patro S, Ranjan R et al (2011) Development of useful recombinant promoter and its expression analysis in different plant cells using confocal laser scanning microscopy. *PLoS One* 6:e24627. doi:[10.1371/journal.pone.0024627](https://doi.org/10.1371/journal.pone.0024627)
 23. Ranjan R, Patro S, Pradhan B et al (2012) Development and functional analysis of novel genetic promoters using DNA shuffling, hybridization and a combination thereof. *PLoS One* 7:e31931. doi:[10.1371/journal.pone.0031931](https://doi.org/10.1371/journal.pone.0031931)
 24. Acharya S, Ranjan R, Pattanaik S et al (2014) Efficient chimeric plant promoters derived from plant infecting viral promoter sequences. *Planta* 239:381–396
 25. Acharya S, Sengupta S, Patro S et al (2014) Development of an intra-molecularly shuffled efficient chimeric plant promoter from plant infecting *Mirabilis* mosaic virus promoter sequence. *J Biotechnol* 169:103–111
 26. Banerjee J, Sahoo DK, Dey N et al (2013) An intergenic region shared by At4g35985 and At4g35987 in *Arabidopsis thaliana* is a tissue specific and stress inducible bidirectional promoter analyzed in transgenic *Arabidopsis* and tobacco plants. *PLoS One* 8. doi: [10.1371/journal.pone.0079622](https://doi.org/10.1371/journal.pone.0079622)
 27. Venter M (2007) Synthetic promoters: genetic control through cis engineering. *Trends Plant Sci* 12:118–124
 28. Park YD, Papp I, Moscone EA et al (1996) Gene silencing mediated by promoter homology occurs at the level of transcription and results in meiotically heritable alterations in methylation and gene activity. *Plant J* 9:183–194
 29. Bhullar S, Chakravarthy S, Advani S et al (2003) Strategies for development of functionally equivalent promoters with minimum sequence homology for transgene expression in plants: cis-elements in a novel DNA context versus domain swapping. *Plant Physiol* 132:988–998
 30. Rushton PJ, Reinstadler A, Lipka V et al (2002) Synthetic plant promoters containing defined regulatory elements provide novel insights into pathogen- and wound-induced signaling. *Plant Cell* 14:749–762
 31. Hernandez-Garcia CM, Finer JJ (2014) Identification and validation of promoters and cis-acting regulatory elements. *Plant Sci* 217:109–119
 32. McWhinnie RL, Nano FE (2014) Synthetic promoters functional in *Francisella novicida* and *Escherichia coli*. *Appl Environ Microbiol* 80:226–234
 33. Schardl CL, Byrd AD, Benzion G et al (1987) Design and construction of a versatile system for the expression of foreign genes in plants. *Gene* 61:1–11
 34. Yang Z, Patra B, Li R et al (2013) Promoter analysis reveals cis-regulatory motifs associated with the expression of the WRKY transcription factor CrWRKY1 in *Catharanthus roseus*. *Planta* 238:1039–1049
 35. Murashige T, Skoog F (1962) A revised medium for rapid growth and bio-assays with tobacco tissue cultures. *Physiol Plant* 15(3):473–497
 36. Di Sansebastiano GP, Paris N, Marc-Martin S et al (1998) Specific accumulation of GFP in a non-acidic vacuolar compartment via a C-terminal propeptide-mediated sorting pathway. *Plant J* 15:449–457
 37. Lu Y, Chen X, Wu Y et al (2013) Directly transforming PCR-amplified DNA fragments into plant cells is a versatile system that facilitates the transient expression assay. *PLoS One* 8, e57171. doi:[10.1371/journal.pone.0057171](https://doi.org/10.1371/journal.pone.0057171)
 38. Jefferson RA, Kavanagh TA, Bevan MW (1987) GUS fusions-beta-glucuronidase as a sensitive and versatile gene fusion marker in higher-plants. *EMBO J* 6:3901–3907
 39. Jefferson RA, Klass M, Wolf N et al (1987) Expression of chimeric genes in *Caenorhabditis elegans*. *J Mol Biol* 193:41–46
 40. Voinnet O, Rivas S, Mestre P et al (2003) An enhanced transient expression system in plants based on suppression of gene silencing by the p19 protein of tomato bushy stunt virus. *Plant J* 33:949–956
 41. Clough SJ, Bent AF (1998) Floral dip: a simplified method for *Agrobacterium*-mediated transformation of *Arabidopsis thaliana*. *Plant J* 16:735–743
 42. Bradford MM (1976) Rapid and sensitive method for quantitation of microgram quantities of protein utilizing principle of protein-dye binding. *Anal Biochem* 72:248–254

Fast and Efficient Cloning of *Cis*-Regulatory Sequences for High-Throughput Yeast One-Hybrid Analyses of Transcription Factors

Zsolt Kelemen, Jonathan Przybyla-Toscano, Nicolas Tissot, Łoic Lepiniec, and Christian Dubos

Abstract

Yeast one-hybrid (Y1H) assay has been proven to be a powerful technique to characterize *in vivo* the interaction between a given transcription factor (TF), or its DNA-binding domain (DBD), and target DNA sequences. Comprehensive characterization of TF/DBD and DNA interactions should allow designing synthetic promoters that would undoubtedly be valuable for biotechnological approaches. Here, we use the ligation-independent cloning system (LIC) in order to enhance the cloning efficiency of DNA motifs into the pHISi Y1H vector. LIC overcomes important limitations of traditional cloning technologies, since any DNA fragment can be cloned into LIC compatible vectors without using restriction endonucleases, ligation, or *in vitro* recombination.

Key words Transcription factor, *Cis*-element, Yeast one-hybrid, Ligase-independent cloning

1 Introduction

Regulation of gene expression is central to all organisms. This regulation is coordinated by a number of different molecular mechanisms requiring sequence-specific DNA binding of regulatory proteins (e.g. DNA methylation, chromatin organization, transcription) [1, 2]. Amongst these proteins, transcription factors (TFs) play a central role by modulating gene expression in response to environmental (e.g. abiotic and biotic stresses) and internal (e.g. hormones or nutrition) signals [3]. TFs modulate the expression of their targets by acting as transcriptional activators, repressors, or both. TFs possess a modular structure generally comprising a regulatory or sensing domain together with a DNA-binding domain (DBD). This is through this DBD that TFs interact with specific DNA *cis*-regulatory sequences usually localized upstream of the transcribed region of their targets [4].

During the evolution of the green lineage this group of regulatory protein has expanded to represent about 8% of the protein coding sequences that have been categorized into 58 different TF families. For example, in the model plant *Arabidopsis thaliana* 2296 TFs have been identified out of the 27,000 protein coding genes [5]. Such expansion reflects the importance of these proteins in controlling plant growth and development in an environment that is by nature highly variable. In contrast with the central role TFs play in plants little data on their DNA binding properties, that can be resumed by the DNA sequences their DBDs recognized, have been gathered over the years [6]. The concomitant identification of these *cis*-elements and the determination of the interacting TFs (or DBDs) is thus an essential step toward a comprehensive understanding of the transcriptional regulatory code occurring in plants. Similarly, extensive characterization of DBD and DNA interactions should allow designing synthetic promoters that would undoubtedly be valuable for plant biotechnological approaches aiming at improving crops.

Various *in vitro* and *in vivo* approaches have been developed in order to characterize the interaction occurring between TFs (or DBDs) and their target DNA sequences, each having their own advantages and limitations [7, 8]. If most *in vitro* methods (e.g. CASTing, SELEX, Surface Plasmon Resonance analysis, protein-binding microarrays) allow accurate determination of TF/DNA or DBD/DNA interaction properties, they necessitate the production of recombinant proteins, which has been proven to be the most critical step. In contrary *in vivo* methods allow taking into account the cellular and nuclear context (e.g. transient expression assays in plants or protoplasts, ChIP-CHIP or ChIP-Seq) but remain difficult to use for the analysis of large set of interactions. In this regard yeast one-hybrid (Y1H) experiment offers some flexibility allowing high-throughput screening of TF/DNA or DBD/DNA interactions. However, one of the limiting steps remains the cloning of large sets of DNA fragments (e.g. *cis*-regulatory sequences or synthetic promoters).

Here, we use the ligation-independent cloning system (LIC) to clone known or putative *cis*-regulatory sequences into the pHISi yeast one-hybrid vector. LIC was developed to facilitate complex cloning and sub-cloning strategies [9]. LIC overcomes important limitations of traditional cloning technologies, since any PCR product can be cloned into LIC compatible vectors without using restriction endonucleases and ligation or recombination. The LIC method takes advantage of the 3' exonuclease activity of T4 DNA polymerase to create complementary 12- to 15-nucleotide overhangs in the vector and PCR product. Upon transformation into *Escherichia coli* cells, the host repair enzymes ligate at the vector-insert junction; thus, LIC achieves fast, cheap, and efficient cloning with minimal non-recombinant background.

As a proof of concept, we converted the pHIS vector into a LIC vector for cloning of *cis*-regulatory sequences in order to carry Y1H experiments.

2 Materials

Unless otherwise specified, all reactions are carried out in standard nuclease-free PCR tubes and incubated in a thermal cycler. Instead of thermal cyclers, water baths or incubators can be used. All solutions are sterilized prior use. *Escherichia coli* and yeast manipulation must be achieved under aseptic conditions.

2.1 LIC Vector Conversion

1. LIC forward (5'-AATTCGACAAGAACACGTGCTCTTCT TCAA-3') and reverse (5'-CTAGTTGAAGAAGAGCACG TGTTCTTGTCG-3') oligonucleotides diluted in sterile water (1 $\mu\text{g}/\mu\text{L}$).
2. Purified pHISi (Clontech) vector (300 ng/ μL).
3. EcoRI and XbaI, restriction enzymes.
4. Gel Extraction Kit.
5. T4 polymerase.
6. T4 ligase.
7. *Escherichia coli* competent cells.
8. Plasmid extraction kit.

2.2 Cloning of Cis-regulatory Sequences

1. DNA fragments containing the 5'-CGACAAGAACAC-3' and 5'-GAAGAAGAGCAC-3' sequences at their 5'-end in the sense and antisense orientation, respectively.
2. PmlI restriction enzyme.
3. Gel Extraction Kit.
4. dTTP.
5. T4 polymerase.
6. Plasmid extraction kit.
7. LB medium: 10 g/L tryptone, 5 g/L yeast extract, 10 g/L NaCl, pH 7.0.

2.3 Yeast Transformation

1. ApaI restriction enzyme.
2. YPDA (yeast peptone dextrose adenine) media: 20 g/L Difco peptone, 10 g/L yeast extract, 20 g/L glucose, 0.03 g/L L-Adenine hemisulfate, 20 g/L agar (for plates only), pH 6.5.
3. Yeast resuspension solution: 0.1 M Lithium acetate, 1 mM Tris-HCl pH 7.4, 0.1 mM EDTA pH 8.

4. Yeast transformation solution: 0.1 M Lithium acetate, 1 mM Tris-HCl pH 7.4, 0.1 mM EDTA pH 8, 40% Polyethylene Glycol (PEG) 4000.
5. Carrier DNA.
6. Synthetic defined (SD) minimal yeast media: 6.7 g/L yeast nitrogen base without amino acids, 1× Dropout Supplement, 20 g/L agar (for plates only), pH 5.8.
7. 10× Dropout (DO) Supplement -Ura: 200 mg/LL-Adenine hemisulfate salt, 200 mg/LL-Arginine HCl, 200 mg/LL-Histidine HCl monohydrate, 300 mg/LL-Isoleucine, 1000 mg/LL-Leucine, 300 mg/LL-Lysine-HCl, 200 mg/LL-Methionine, 500 mg/LL-Phenylalanine, 2000 mg/LL-Threonine, 200 mg/LL-Tryptophan, 300 mg/LL-Tyrosine, 1500 mg/LL-Valine.

2.4 Yeast One-Hybrid

1. Ordered prey library for yeast one-hybrid screenings, for example the REGIA transcription factor library [10].
2. YPDA (yeast peptone dextrose adenine) media: 20 g/L Difco peptone, 10 g/L yeast extract, 20 g/L glucose, 0.03 g/LL-Adenine hemisulfate, 20 g/L agar (for plates only), pH 6.5.
3. *Saccharomyces cerevisiae* strains EGY48 and YM4271.
4. Microplate replicator.
5. Synthetic defined (SD) minimal yeast media: 6.7 g/L yeast nitrogen base without amino acids, 1× Dropout Supplement, 20 g/L agar (for plates only), pH 5.8.
6. 10× Dropout (DO) Supplement -Ura: 200 mg/LL-Adenine hemisulfate salt, 200 mg/LL-Arginine HCl, 200 mg/LL-Histidine HCl monohydrate, 300 mg/LL-Isoleucine, 1000 mg/LL-Leucine, 300 mg/LL-Lysine-HCl, 200 mg/LL-Methionine, 500 mg/LL-Phenylalanine, 2000 mg/LL-Threonine, 200 mg/LL-Tryptophan, 300 mg/LL-Tyrosine, 1500 mg/LL-Valine.
7. 10× Dropout (DO) Supplement -Trp: 200 mg/LL-Adenine hemisulfate salt, 200 mg/LL-Arginine HCl, 200 mg/LL-Histidine HCl monohydrate, 300 mg/LL-Isoleucine, 1000 mg/LL-Leucine, 300 mg/LL-Lysine-HCl, 200 mg/LL-Methionine, 500 mg/LL-Phenylalanine, 2000 mg/LL-Threonine, 300 mg/LL-Tyrosine, 200 mg/LL-Uracil, 1500 mg/LL-Valine.
8. 10× Dropout (DO) Supplement -His-Trp-Ura: 200 mg/LL-Adenine hemisulfate salt, 200 mg/LL-Arginine HCl, 300 mg/LL-Isoleucine, 1000 mg/LL-Leucine, 300 mg/LL-Lysine-HCl, 200 mg/LL-Methionine, 500 mg/LL-Phenylalanine, 2000 mg/LL-Threonine, 300 mg/LL-Tyrosine, 1500 mg/LL-Valine.
9. 3-aminotriazol (3-AT).

3 Methods

3.1 Annealing of Oligonucleotides

1. Mix 5 μL of both complementary LIC oligonucleotides in a standard PCR tube.
2. Heat to 95 $^{\circ}\text{C}$ and leave at 95 $^{\circ}\text{C}$ for 2 min, then heat to 55 $^{\circ}\text{C}$ and incubate for 5 min.
3. Briefly spin the tubes in a microfuge and store on ice or at 4 $^{\circ}\text{C}$ until use. For long-term storage, annealed oligonucleotides can be kept at -20 $^{\circ}\text{C}$. (*see* **Note 1**)

3.2 Conversion of the pHISi Vector into a LIC Vector

1. Set up the following reaction on ice: 1 μg pHISi vector, 2 μL 10 \times restriction enzyme buffer (*see* **Note 2**), 10 units of both EcoRI and XbaI restriction enzymes, and nuclease-free water to 20 μL final volume. Mix by pipetting up and down.
2. Digest for 2 h at 37 $^{\circ}\text{C}$.
3. Separate the reaction on agarose gel (1%).
4. Purify digested pHISi vector with a gel extraction kit according to the manufacturer's instructions.
5. Anneal complementary LIC-oligos (*see* Subheading 3.1 and **Note 3**).
6. Mix 1 μg linearized vector and 30 ng annealed oligonucleotides in a 1.5 mL microcentrifuge tube. Add 2 μL 10 \times T4 DNA ligase reaction buffer and nuclease-free water to 19 μL final volume. Mix well by pipetting up and down. Add 1 μL T4 ligase and mix well by pipetting up and down. Perform ligation overnight at 16 $^{\circ}\text{C}$.
7. Transform 2 μL ligation reaction into competent *E. coli* cells according to the manufacturer's instructions.
8. Purify pHISi-LIC vector from a single colony grown in the presence of ampicillin (50 $\mu\text{g}/\text{mL}$) antibiotics on an LB agar plate using a plasmid extraction kit according to the manufacturer's instructions.

3.3 Cloning of Cis-regulatory Sequences (Fig. 1)

1. Synthesize the *cis*-regulatory sequences as hexamers, as both sense and antisense oligonucleotides (diluted in sterile water at 1 $\mu\text{g}/\mu\text{L}$). The 5'-CGACAAGAACAC-3' and 5'-GAAGAAGAGCAC-3' sequences must be added to the 5'-end of the sense and antisense oligonucleotides, respectively (*see* **Note 4**).
2. Anneal oligos: *see* Subheading 3.1.
3. Set up the following reaction in a microcentrifuge tube on ice: 20 μg pHISi-LIC vector, 5 μL 10 \times restriction enzyme buffer, 100 units PmlI restriction enzyme and nuclease-free water to 50 μL final volume. Mix by pipetting up and down.

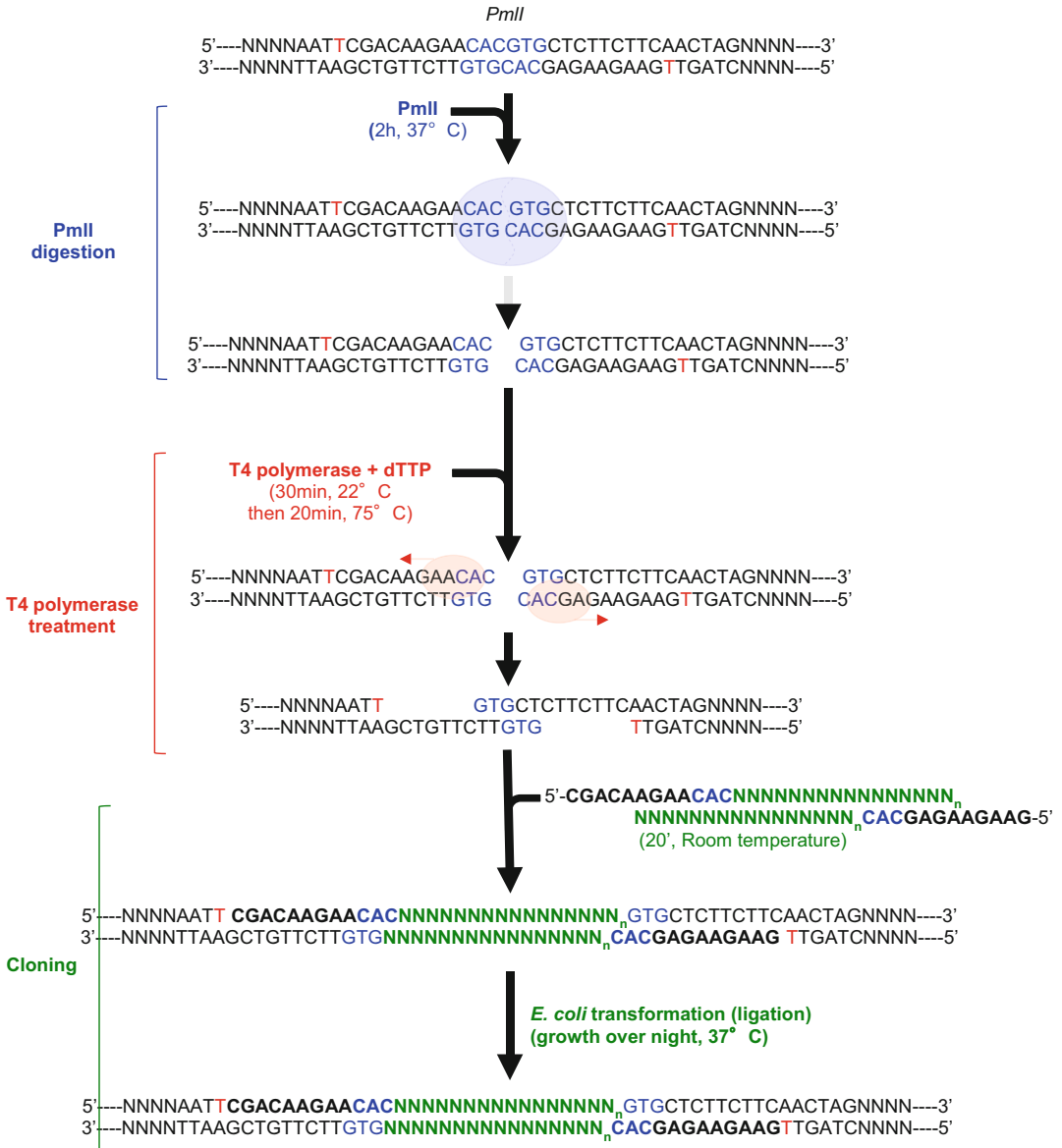


Fig. 1 Cloning of *cis*-elements using ligase-independent cloning (LIC) method. During the LIC-cloning process the pHSi-LIC vector is linearized with *PmlI* restriction enzyme, and the T4 polymerase treatment in the presence of dTTP nucleotides results in long cohesive ends. Fragments containing compatible cohesive ends are combined with the treated vector and transformed directly into *E. coli* competent cells. Upon transformation, the host repair enzymes ligate at the vector-insert junction

4. Digest for 2 h at 37 °C.
5. Separate the reaction on agarose gel (1 %).
6. Purify the digested vector with a gel extraction kit according to the manufacturer’s instructions.

7. Set up the following reaction on ice: 1 μg linearized pHis-LIC vector, 2 μL 10 \times T4 DNA polymerase buffer 2.1, 5 μL 100 mM dTTP, 1 unit T4 DNA polymerase and nuclease-free water to 20 μL final volume. Mix by pipetting up and down.
8. Incubate for 30 min at 22 $^{\circ}\text{C}$.
9. Stop the reaction by heat inactivation for 20 min at 75 $^{\circ}\text{C}$.
10. Mix 100 ng vector and 30 ng fragment (corresponding to the studied *cis*-element) and adjust to 5 μL final volume with nuclease-free water.
11. Incubate the mixture at room temperature for 20 min.
12. Transform the whole reaction into competent *E. coli* cells according to the manufacturer's instructions.
13. Purify pHis-LIC vector containing the *cis*-element from a single colony grown in the presence of ampicillin (50 $\mu\text{g}/\text{mL}$) antibiotics on an LB agar plate using a plasmid extraction kit according to the manufacturer's instructions (*see Note 5*).

3.4 Yeast Transformation

1. Set up the following reaction in a microcentrifuge tube on ice (*see Note 6*): 5 μg pHis-LIC containing the *cis*-element, 2 μL 10 \times restriction enzyme buffer, 10 units ApaI restriction enzyme and nuclease-free water to 20 μL final volume. Mix by pipetting up and down.
2. Digest for 4 h at 25 $^{\circ}\text{C}$.
3. Store the reaction mix at -20°C until use (*see Note 7*).
4. Grow the appropriate yeast strain (*see Note 8*) on 90 mm petri dish filled with YPDA at 28 $^{\circ}\text{C}$ for 2 days.
5. Collect the culture from one petri dish and resuspend it in 1 mL sterile water in a 1.5 mL microcentrifuge tube.
6. Centrifuge at maximum speed for 15 s in a microcentrifuge, and remove the supernatant.
7. Resuspend the pellet in 1 mL Yeast resuspension solution (*see Note 9*).
8. Mix 5 μg ApaI-digested pHis-LIC construct with 5 μL carrier DNA, 0.1 mL yeast cell solution and 0.6 mL Yeast transformation solution in a 1.5 mL microcentrifuge tube.
9. Vortex for 30 s.
10. Incubate the transformation mixture for 30 min at 28 $^{\circ}\text{C}$ with occasional mixing in an incubator.
11. Perform heat-shock at 42 $^{\circ}\text{C}$ for 25 min in a water bath.
12. Centrifuge in a 1.5 mL microcentrifuge tube at maximum speed for 15 s.
13. Resuspend the pellet in 250 μL sterile water and plate onto selective SD plates deprived of uracil (SD -U; *see Note 10*).

14. Perform self-activation test: grow yeast cells harboring *cis*-element constructs on SD plates deprived of uracil and histidine (SD -U -H) containing 5, 15, 30, 45, and 60 mM 3-aminotriazol (3-AT). For screening use the lowest 3-AT concentration where no growth can be observed. If self-activation is observed at 60 mM 3-AT then the corresponding experiment should be removed from the study (*see* **Note 11**).

3.5 Yeast One-Hybrid (Y1H) Experiments

1. Grow ordered transcription factor library on SD plates deprived of tryptophan (SD -W) for 1–2 days at 28 °C. Resuspend about 10 µL yeast from each clone in 100 µL sterile water in a 96-well microtiter plate (*see* **Note 8**). With replicator, transfer about 5 µL solution onto YPDA plates, let it dry under the sterile hood.
2. Grow reporter yeast strain (containing the *cis*-elements) on 90 mm SD -U plates for 1–2 days at 28 °C. Collect the culture from a 90 mm petri dish and resuspended it in 10 mL sterile water. Distribute the yeast solution into the wells of a 96-well microtiter plate (80 µL/well). With replicator transfer about 5 µL solution on top of the library strains.
3. Let it dry under the sterile hood and incubate for 1–2 days at 28 °C.
4. With replicator, transfer the yeast cells onto SD -U -W plates to select diploid yeast cells and incubate for 1–2 days at 28 °C.
5. To analyze interactions, transfer diploid yeast cells to 100 µL sterile water (in a 96-well plate) with replicator, mix well.
6. With replicator, transfer 5 µL yeast solutions onto selective media (*see* **Note 12**). Incubate for 3–5 days at 28 °C. Diploid colonies growing on a medium lacking the histidine amino acid and in the presence of various concentrations (from 15 to 60 mM) of 3-AT are considered as positive clones expressing the candidate transcription factor interacting with the studied DNA motif (Fig. 2).

4 Notes

1. Alternatively, oligos can be annealed in a water bath. Boil water in a large glass beaker on a hotplate. Incubate the tube of oligonucleotides in the boiling water for 2 min. Turn off the hotplate, leaving the oligonucleotides in the beaker on the hotplate to slowly cool to room temperature. This would take several minutes.
2. Buffer ensuring 100% activity for both restriction enzymes.
3. LIC-oligos are complementary sequences designed to contain, after annealing, EcoRI- and XbaI-compatible cohesive ends as well as a unique PmlI restriction site (*see* Subheading 2).

a

	1	2	3	4	5	6	7	8	9	10	11	12
A	AtMYB030	AtMYB031	AtMYB060	AtMYB094	AtMYB096	AtMYB013	AtMYB014	AtMYB015	AtMYB058	AtMYB063	AtMYB003	AtMYB004
B	AtMYB007	AtMYB032	AtMYB123	AtMYB075	AtMYB090	AtMYB113	AtMYB114	AtMYB011		AtMYB111	AtMYB016	AtMYB017
C	AtMYB106	AtMYB009	AtMYB039	AtMYB107	AtMYB041	AtMYB074	AtMYB102	AtMYB028	AtMYB029	AtMYB034	AtMYB051	AtMYB076
D	AtMYB122	AtMYB050	AtMYB055	AtMYB061	AtMYB086	AtMYB036	AtMYB037	AtMYB038	AtMYB068	AtMYB084	AtMYB087	AtMYB000
E	AtMYB023	AtMYB066	AtMYB018	AtMYB019	AtMYB045	AtMYB033	AtMYB065	AtMYB081	AtMYB097	AtMYB101	AtMYB104	
F	AtMYB021	AtMYB024	AtMYB002	AtMYB062	AtMYB078	AtMYB108	AtMYB112	AtMYB116	AtMYB052	AtMYB054	AtMYB056	AtMYB069
G	AtMYB105	AtMYB110	AtMYB117	AtMYB044	AtMYB070	AtMYB073	AtMYB077	AtMYB001	AtMYB025	AtMYB109	AtMYB053	AtMYB092
H	AtMYB093	AtMYB022	AtMYB064	AtMYB100	AtMYB115	AtMYB118		AtMYB005	AtMYB026	AtMYB035	AtMYB046	AtMYB057

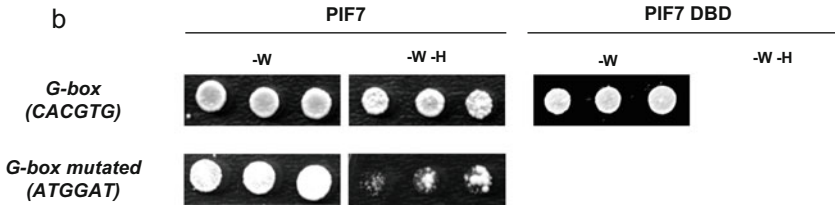
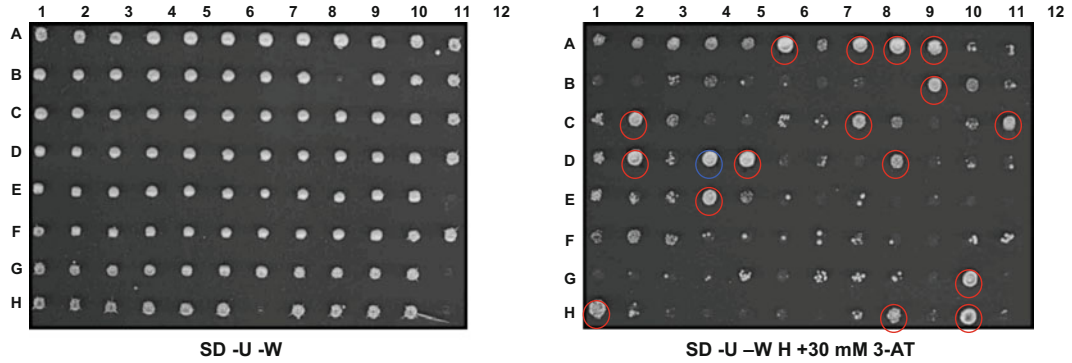


Fig. 2 Example of yeast one-hybrid assays using ligase- independent cloning (LIC) method. Yeast strains harboring the *HIS3* reporter gene under the control of the chimeric promoter were mated with strains containing pDEST22 expression vector allowing the expression of Arabidopsis R2R3-MYB transcription factors [11]. Diploid cells were plated on appropriate media to maintain the expression of both vectors (SD -U -W) and to test the activation of the *HIS3* reporter gene (SD -U -W -H with 30 mM 3-AT). *U* uracil, *W* tryptophan, *H* histidine, *3-AT* 3-Amino-1,2,4-triazole. (a) An ordered R2R3-MYB cDNA library composed of 93 individual genes was made and assayed against the *AC-I cis*-regulatory (*CACCTACC*) sequence that was cloned using the LIC method [12]. Previous studies have demonstrated using various methods (including Y1H assays) that *AC-I* is a direct target of AtMYB61 [8]. Here we confirmed the interaction between AtMYB61 (blue dotted lines) and the *AC-I* regulatory sequence, and identified 16 additional interactions (red dotted lines). Table: position of the R2R3-MYB proteins on petri dishes. (b) PIF7 (PHYTOCHORME INTERACTING FACTOR 7, which encode a bHLH transcription factor) or its DNA-binding domain (DBD) were assayed with a *G-box* regulatory sequence or a mutated version that were both cloned using the LIC method. As previously found in Y1H experiments, PIF7 (or its DBD) can interact with its cognate DNA target (i.e. *G-box*) and not with a mutated version [13]

- After annealing, the fragments contain sticky ends that are compatible with the overhangs of the T4 polymerase-treated vector. Here we use *cis*-regulatory sequences synthesized as hexamers, however this method can also be used with DNA

fragments from other sources (e.g. PCR reaction, digested cloning vectors).

5. The pHISi-LIC construct can be verified either by restriction digest or by sequencing. Double digest with PmlI and XhoI restriction enzymes should produce two fragments: one 1 kbp and one 5.8 kbp fragment. For sequencing use M13 forward sequencing primer (5'-GTAAAACGACGGCCAGT-3') or T7 universal primer (5'-TAATACGACTCACTATAGGG-3').
6. The pHISi-LIC construct must be linearized at ApaI site in order to direct the insertion into the URA3 locus. **Steps 1–3** can be carried out any time prior yeast transformations.
7. If the pHISi-LIC vector containing the DNA fragment of interest is used for yeast transformation the same day as it is digested, the reaction mix can be kept on ice until use.
8. Transform pHISi-LIC construct containing the *cis*-element into *Saccharomyces cerevisiae* α -type mating strain EGY48 at the URA3 locus as the pHISi vector carries the URA3 selection marker with a unique ApaI restriction site. Use the yeast α -type mating strain YM4271 for the transformation of the transcription factor library in pDEST22 (Invitrogen).
9. **Steps 6** and **7** can be repeated twice in order to increase the purity of the yeast suspension.
10. Two different 90 mm petri dish are used to select transformed yeast cells on which 50 and 200 μ L are plated.
11. 3-Aminotriazole (3-AT) is a competitive inhibitor of the *HIS3* gene product allowing to overcome leaky expressions.
12. Use SD deprived of histidine (His), tryptophan (Trp), and uracil (Ura).

Acknowledgment

This work was supported by two projects, STREG (Plant-KBBE) and CERES (ANR-2010-BLAN-1238).

We acknowledge “Ministère de l'enseignement supérieur et de la recherche” for supporting N.T.

References

1. Bell JT, Pai AA, Pickrell JK, Gaffney DJ, Pique-Regi R, Degner JF, Gilad Y, Pritchard JK (2011) DNA methylation patterns associate with genetic and gene expression variation in HapMap cell lines. *Genome Biol* 12:R10
2. Hübner MR, Eckersley-Maslin MA, Spector DL (2013) Chromatin organization and transcriptional regulation. *Curr Opin Genet Dev* 23:89–95
3. Todeschini AL, Georges A, Veitia RA (2014) Transcription factors: specific DNA binding and specific gene regulation. *Trends Genet* 30: 211–219
4. Wittkopp PJ, Kalay G (2012) Cis-regulatory elements: molecular mechanisms and evolu-

- tionary processes underlying divergence. *Nat Rev Genet* 13:59–69
5. Jin JP, Zhang H, Kong L, Gao G, Luo JC (2014) PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res* 42: D1182–D1187
 6. Charoensawan V, Wilson D, Teichmann SA (2010) Lineage-specific expansion of DNA-binding transcription factor families. *Trends Genet* 26:388–393
 7. Fornalé S, Shi X, Chai C, Encina A, Irar S, Capellades M, Fuguet E, Torres JL, Rovira P, Puigdomènech P, Rigau J, Grotewold E, Gray J, Caparrós-Ruiz D (2010) ZmMYB31 directly represses maize lignin genes and redirects the phenylpropanoid metabolic flux. *Plant J* 64: 633–644
 8. Prouse MB, Campbell MM (2013) Interactions between the R2R3-MYB transcription factor, AtMYB61, and target DNA binding sites. *PLoS One* 8, e65132
 9. Aslanidis C, de Jong PJ (1990) Ligation-independent cloning of PCR products (LIC-PCR). *Nucleic Acids Res* 18:6069–6074
 10. Paz-Ares J, The REGIA Consortium (2002) REGIA, an EU project on functional genomics of transcription factors from *Arabidopsis Thaliana*. *Comp Funct Genomics* 3:102–108
 11. Dubos C, Stracke R, Grotewold E, Weisshaar B, Martin C, Lepiniec L (2010) MYB transcription factors in Arabidopsis. *Trends Plant Sci* 15:573–581
 12. Patzlaff A, Newman LJ, Dubos C, Whetten RW, Smith C, McInnis S, Bevan MW, Sederoff RR, Campbell MM (2003) Characterisation of Pt MYB1, an R2R3-MYB from pine xylem. *Plant Mol Biol* 53:597–608
 13. Dubos C, Kelemen Z, Sebastian A, Bülow L, Huep G, Xu W, Grain D, Salsac F, Brousse C, Lepiniec L, Weisshaar B, Contreras-Moreira B, Hehl R (2014) Integrating bioinformatic resources to predict transcription factors interacting with *cis*-sequences conserved in co-regulated genes. *BMC Genomics* 15:317

The *Physcomitrella patens* System for Transient Gene Expression Assays

Johanne Thévenin, Wenjia Xu, Louise Vaisman, Loïc Lepiniec, Bertrand Dubreucq, and Christian Dubos

Abstract

Transient expression assays are valuable techniques to study *in vivo* the transcriptional regulation of gene expression. These methods allow to assess the transcriptional properties of a given transcription factor (TF) or a complex of regulatory proteins against specific DNA motifs, called *cis*-regulatory elements. Here, we describe a fast, efficient, and reliable method based on the use of *Physcomitrella patens* protoplasts that allows the study of gene expression in a qualitative and quantitative manner by combining the advantage of GFP (green fluorescent protein) as a marker of promoter activity with flow cytometry for accurate measurement of fluorescence in individual cells.

Key words *Physcomitrella patens*, Protoplasts, Transient expression, Transcription factor, Subcellular localization

1 Introduction

Plant growth and development necessitate the tight and coordinated expression of several hundreds of genes. Transcription factors (TFs) play a central role in this process by activating or repressing the transcription of their target genes. Such regulations occur through the direct interaction between TF DNA-binding domains (DBDs) and specific DNA sequences known as *cis*-regulatory elements. These interactions, that are characteristic to each TF, have been studied through various methods, all having their own advantages and limitations.

In vitro methods (e.g. CASTing, SELEX, Surface Plasmon Resonance analysis, protein-binding microarrays) have allowed accurate determination of numerous TF/DNA or DBD/DNA interaction properties [1, 2]. The main drawback of these approaches, beside the need of recombinant proteins, is that they are by definition not taking

Johanne Thévenin and Wenjia Xu contributed equally with all other contributors.

into account the molecular context found in living cells. For these reasons, numerous *in vivo* strategies were developed. The yeast one-hybrid system is a simple and efficient method allowing high-throughput analyses [3]. However, in this heterologous system, only a small number of TFs (usually up to three) can be assayed simultaneously against a DNA target [4]. Various efficient plant systems relying on the use of either tissues or cell transformation with *Agrobacteria* or protoplast transfection have also been developed [5–8]. However, several limitations associated with these plant systems (e.g. efficiency of transformation, maintenance of cell cultures, time-consuming and labor-intensive preparation of protoplasts) may hamper their use.

Here, we describe a fast, efficient, and reliable method based on the use of protoplasts generated from the moss *Physcomitrella patens*. This procedure allows the study of gene expression in a qualitative and quantitative manner by combining the advantage of GFP as a marker of promoter activity together with flow cytometry for accurate measurement of fluorescence in individual cells [9]. The method described herein allows very rapid sample processing, as only 2–3 days are sufficient from the harvesting of *P. patens* protonema and protoplasts production to the final results. Moreover, this method allows the study at least four TFs and a target promoter within the same experiment permitting the study of complex transcriptional mechanisms [9]. Finally, this protocol can also be used to decipher the subcellular localization of proteins.

2 Materials

Prepare all solutions with ultrapure water. Culture media are derived from the recipes described in [10, 11].

2.1 Moss Culture and Protoplast Isolation and Transformation Medium

1. PPNH₄ medium: macro elements 0.8 g/L CaNO₃·4H₂O, 0.25 g/L MgSO₄·7H₂O, 0.0125 g/L FeSO₄·7H₂O; micro elements 0.055 mg/L CuSO₄·5H₂O, 0.055 mg/L ZnSO₄·7H₂O, 0.614 mg/L H₃BO₃, 0.389 mg/L MnCl₂·4H₂O, 0.055 mg/L CoCl₂·6H₂O, 0.028 mg/L KI, 0.025 mg/L Na₂MoO₄·2H₂O; other elements 250 mg/L KH₂PO₄, 500 mg/L ammonium tartrate (*see Note 1*). Prepare a 100× stock solution for each ingredient of macro elements, autoclave, store at 4 °C and add 10 mL per liter of medium. Prepare a 1000× stock solution for each ingredient of micro elements, autoclave, store at 4 °C and add 1 mL per liter of PPNH₄ medium. Dissolve 25 g KH₂PO₄ in 100 mL water and titrate to pH 7 with KOH to make a 1000× stock solution, autoclave, store at 4 °C and add 1 mL per liter of medium. Dissolve 18.415 g Ammonium tartrate in 100 mL water to make 1 M stock solution, store at 4 °C and add 2.7 mL per liter of PPNH₄ medium.
2. Solid PPNH₄ medium: add 7.2 g agar per liter of PPNH₄ medium, and add 200 mg/mL cefotaxime to suppress bacterium contamination.

3. 8.5% mannitol solution: add 85 g mannitol per liter of PPNH_4 medium.
4. PPNH_4 + 6.6% mannitol solution: add 66 g mannitol per liter of PPNH_4 medium.
5. MMM solution: 8.5 g/100 mL mannitol, 0.305 g/100 mL $\text{MgCl}_2 \cdot 6\text{H}_2\text{O}$, 0.1 g/100 mL MES. Dissolve all powders and adjust the pH to 5.6 with KOH.
6. PEG solution: 7 g/100 mL mannitol, 2.36 g/100 mL $\text{CaNO}_3 \cdot 4\text{H}_2\text{O}$, 0.12114 g/100 mL Tris, 40 g/100 mL PEG4000. All chemicals are directly dissolved in 65 mL ultra-pure water in a graduated bottle (for a 100 mL final volume), mix. Heating at 50–60 °C can help dissolving PEG.
7. Driselase solution: 2 g/100 mL driselase, 8.5 g/100 mL mannitol. Dissolve the driselase at 4 °C with gentle shaking during at least 2 h for complete solvation, then centrifuge at $5000 \times g$ for 5 min, filtrate (0.2 μm) the clear supernatant, dispense into 10-mL aliquots and store at -20 °C.

Solutions 1, 2, 3, 4, 5, and 6 should be autoclaved at 120 °C for 20 min, then 1–5 stored at room temperature and 6 stored at -20 °C. Once open, keep them at 4 °C to avoid contamination.

2.2 Labware

1. 140 μm , 80 μm , 40 μm sieves adapted to a 100-mL beaker.
2. Cellophane disks.
3. 30-mL glass tubes with round base.
4. 3 M™ Micropore™ surgical tape.
5. Hematocytometer.
6. Spatula.
7. 13-mL tubes with round base.
8. 0.2 μm syringe filters.
9. 50 μm cell strainer.
10. Homogenizer.
11. Flow Cytometer (Sysmex-Partec S.A.R.L., Sainte Genevieve des Bois, France), with a 488-nm solid sapphire 20-mW laser for excitation.
12. FLOMAX Software (Sysmex-Partec S.A.R.L., Sainte Genevieve des Bois, France).
13. Calibration beads (Green, Sysmex-Partec Ref: 05-4006).
14. Sheath fluid (Sysmex-Partec reference 04-4007).
15. Epifluorescence microscope equipped with an HBO burner.

Materials 1, 2, and 3 should be autoclaved at 120 °C for 20 min, then stored at room temperature.

2.3 Vector Constructions

1. *pBS Tpp-A* vector, described in [9].
2. *pBS Tpp-B* vector, described in [9].
3. Binary vector contains the 35S cauliflower mosaic virus promoter.
4. Phusion DNA polymerase.
5. Gel Extraction Kit.
6. *Escherichia coli* competent cells.
7. Plasmid extraction kit.
8. BP gateway® clonase.
9. LR gateway® clonase.

3 Methods

Moss culture and protoplast isolation and transformation should be done under a horizontal laminar flow hood.

3.1 Moss Culture

1. Cut 7–10 days old moss protonemal tissues with a homogenizer, until no big chunk is visible.
2. Spread uniformly 2–3 mL of a freshly fragmented protonema per 90-mm sterile petri dish poured with solid PPNH₄ culture medium and overlaid with cellophane disks (*see Note 2*).
3. Incubate the plates at 24 °C with a light regime of 16 h light : 8 h darkness at 80 μmol/m²/s (adapted from ref. 12).
4. Collect protonema by gently scraping the surface of the cellophane disks after 7–10 days of growth (before culture starts to turn brown) for either liquid storage or protoplast preparation (*see Note 3*).

3.2 Protoplast Preparation

1. Collect protonema from 7 to 10 days old moss from 2 to 3 PPNH₄ plates with a sterile spatula.
2. Transfer into a petri dish containing 1 % driselase diluted in an 8.5 % mannitol solution (*see Note 4*).
3. Incubate at room temperature for 1–3 h with gentle shaking (*see Note 5*).
4. Filter the protoplast suspension successively through two overlaid sieves, 80 μm and 40 μm, respectively. Protoplasts are collected in a beaker and then transfer into 30 mL sterile tubes with round base.
5. Centrifuge the filtered suspension at 300 × *g* for 5 min at room temperature, and discard the supernatant.
6. Wash the pellet with 10 mL 8.5 % mannitol, centrifuge at 300 × *g* for 5 min at room temperature, and discard the supernatant.

7. Wash the pellet with 10 mL 8.5% mannitol, and take 15 μL suspension before centrifugation to determine protoplasts concentration with a hemacytometer.
8. Suspend the pellet gently in MMM solution and adjust volume to reach the protoplasts concentration of $0.5\text{--}0.8 \times 10^6$ protoplasts/mL. At this stage protoplasts are ready for transformation (*see Note 6*) (Fig. 1).

3.3 Transformation

Water bath must be at 45 °C before the beginning of the whole procedure.

1. Add 4.5 μg of each purified DNA plasmid into a 13-mL sterile tube, in a total volume of no more than 30 μL to maintain optimal conditions for transformation (*see Note 7*).
2. Add 300 μL of protoplast suspension into the tube containing the plasmid DNAs and mix (*see Note 8*).
3. Add 300 μL of PEG solution into the protoplast/DNA mixture (*see Note 9*).
4. Incubate the protoplast/DNA/PEG mixture in a water bath at 45 °C for 7 min (*see Note 10*).
5. Incubate at room temperature for 10 min (*see Note 11*).
6. Progressively dilute the samples with 6.5 mL $\text{PPNH}_4 + 6.6\%$ mannitol solution: add first 1 mL $\text{PPNH}_4 + 6.6\%$ mannitol solution into each sample, mix gently, then add the rest of the solution into each sample, mix gently.
7. Incubate the samples in darkness for 36–48 h at 24 °C (Fig. 1).

3.4 Detection of GFP Fluorescence by Flow Cytometry

1. Remove 5 mL supernatant of each sample.
2. Filter the transformed protoplast suspension with a 50 μm cell strainer (*see Note 12*).
3. GFP quantification in living protoplasts is performed on a PARTEC CyFlow Space instrument, using FLOMAX acquisition and analysis software, and a 488-nm solid sapphire 20-mW laser for excitation (*see Note 13*).
4. GFP fluorescence is detected with a 527-nm/30-nm band-pass filter (FL1 channel). Red chlorophyll-based fluorescence from living protoplasts is detected with a 610-nm/30-nm band-pass filter (FL2 channel). The side light scatter (SSC) detector high voltage was set to 161.5 V. The photomultiplier tube voltages were adjusted to 275 V for FL1 and 475 V for FL2 (*see Note 14*) (Fig. 2).

3.5 Subcellular Localization Analyzed by Epifluorescence Microscopy

1. Add 2 μL 4',6-Diamidino-2-Phenylindole (DAPI, 100 mg/mL) into 100 μL of concentrated protoplasts, then incubate at room temperature for 5 min.

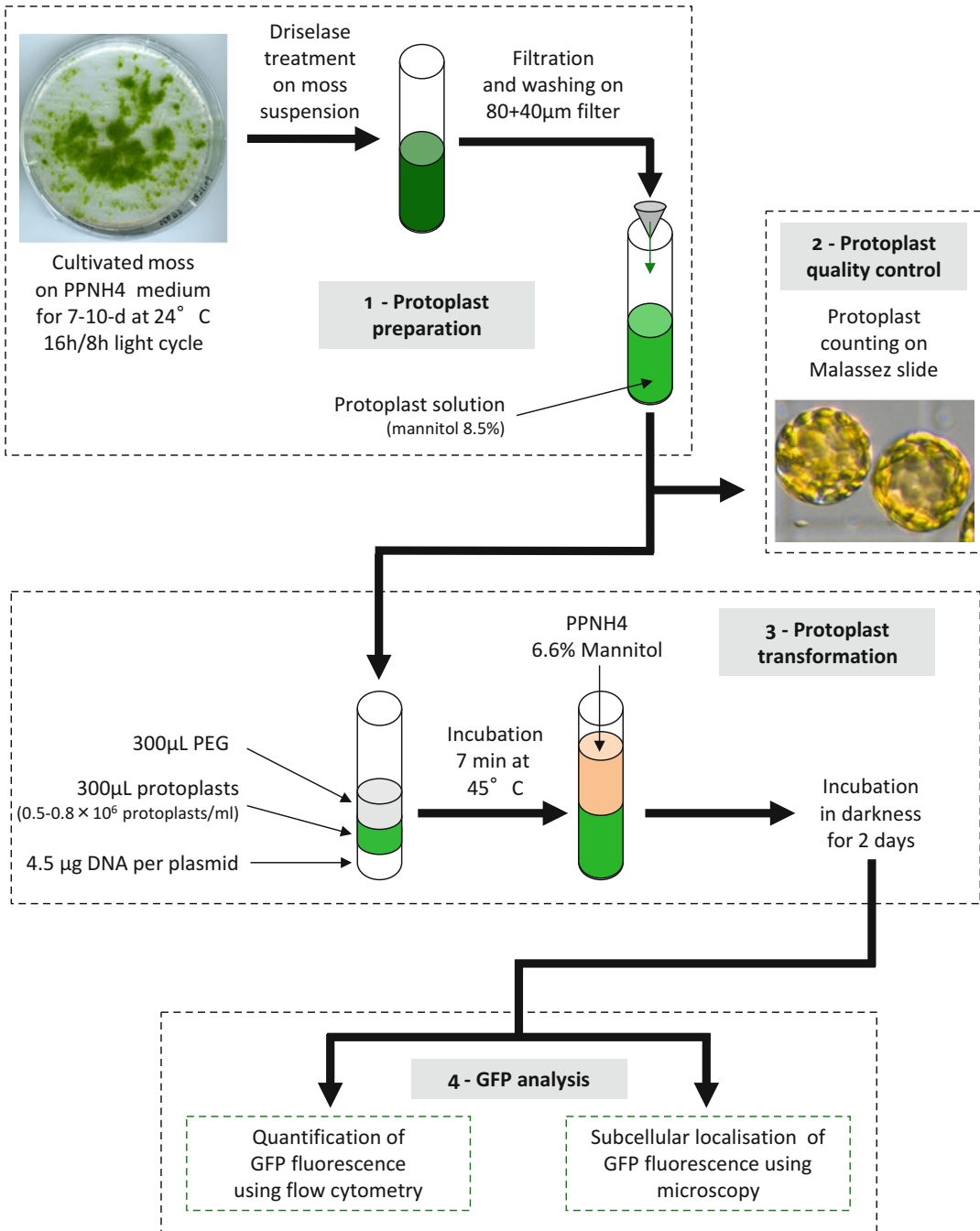


Fig. 1 Schematic representation of the procedure used to produce, transform, and analyze *Physcomitrella patens* protoplasts

- Images of GFP positives are obtained on a Zeiss Axioplan II epifluorescence microscope equipped with an HBO burner and using a GFP bandpass (459–490 nm BP 515–565 nm), a long-pass filter (459–490 nm LP 520 nm), and a DAPI filter (365–395 nm LP 397 nm) (Fig. 3).

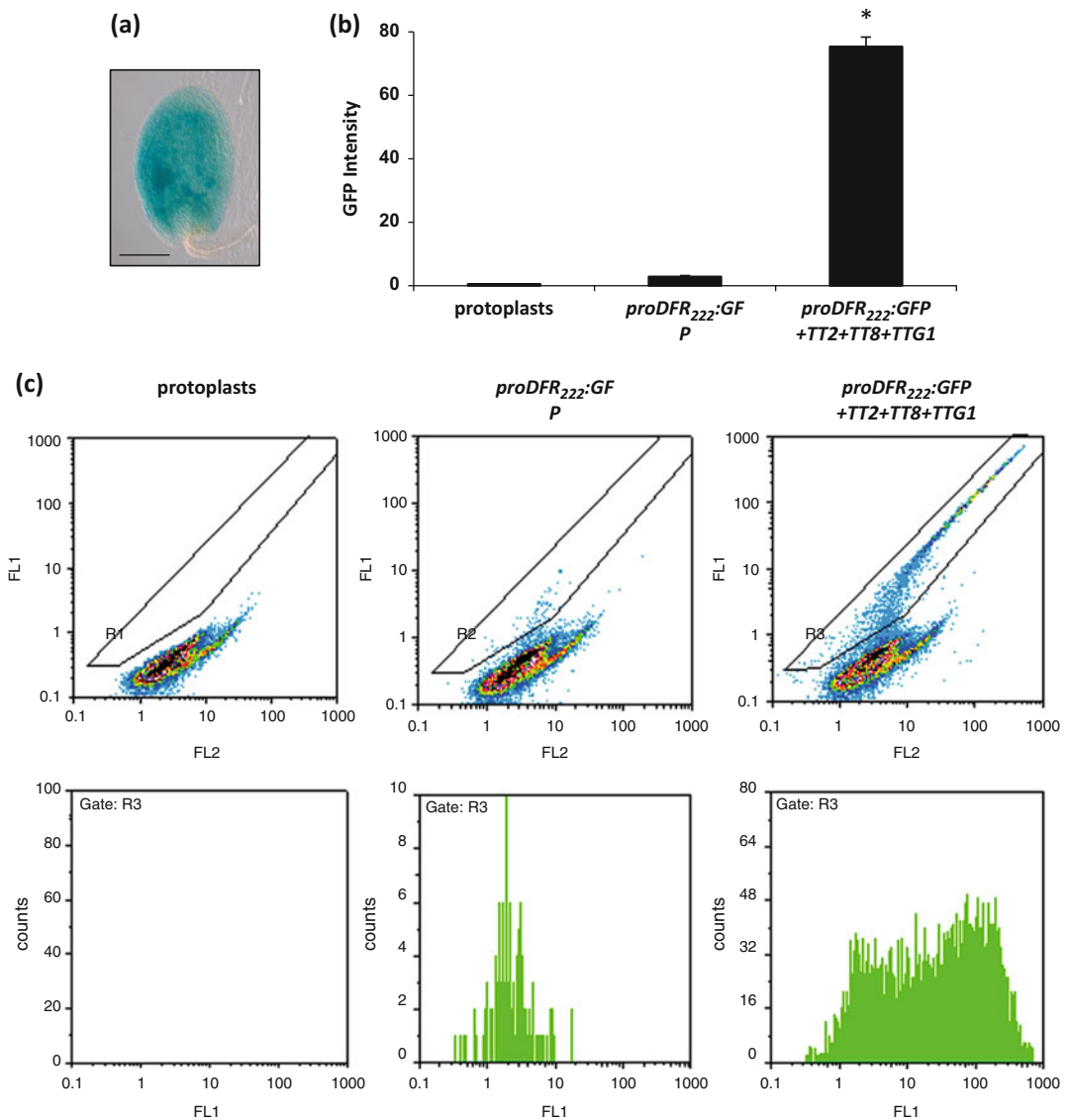


Fig. 2 Transient gene expression assays using the *Physcomitrella patens* system. In *Arabidopsis thaliana* seeds, the biosynthetic pathway leading to proanthocyanidin (condensed tannins) accumulation has been extensively studied [21]. *Dihydroflavonol-4-reductase* (*DFR*) encodes a key enzyme involved in this biosynthetic process. *DFR* expression is directly controlled at the transcriptional level by several MYB-bHLH-WD40 ternary protein complexes amongst which AtTT2/AtMYB123 (R2R3-MYB), AtTT8/AtbHLH042 (bHLH), and TTG1 (WD40 repeat containing protein) play a preponderant role. **(a)** The *Arabidopsis thaliana proDFR222* promoter fragment (–350 to –128 bp prior to the start codon) fused to the minimal 35S promoter (cauliflower mosaic virus) is functional in seeds, as revealed by β -glucuronidase (*uidA/GUS*) activity (whole mount seeds). **(b)** The TT2-TT8-TTG1 complex activates *GFP* expression from *proDFR222* (fused to the minimal 35S promoter) in transient expression assays using *P. patens* protoplasts. Transactivation activity was monitored by GFP fluorescence from three biological repetitions. Error bars: \pm SE. *t*-test significance: *, $P < 0.001$. **(c)** GFP fluorescence signals from flow cytometry are displayed and analyzed in scatter plot diagrams. The presented results are comparing non-transfected (*Left column*) and transfected protoplasts with *proDFR222:GFP* alone (*Middle column*) or in combination with AtTT2, AtTT8, and AtTTG1 (*Right column*). Analyses were carried out in a logarithmic amplification mode (four decades range). (*Top row*) Scatter plots displaying GFP signal (FL1) and red fluorescence (FL2, highlighting chlorophylls fluorescence): R3 window corresponds to the area where living protoplasts with high GFP fluorescence are gathering (i.e. gated protoplasts region). (*Bottom row*) Number of protoplasts (counts) to each of the GFP intensities that have been measured in the R3 window (FL1)

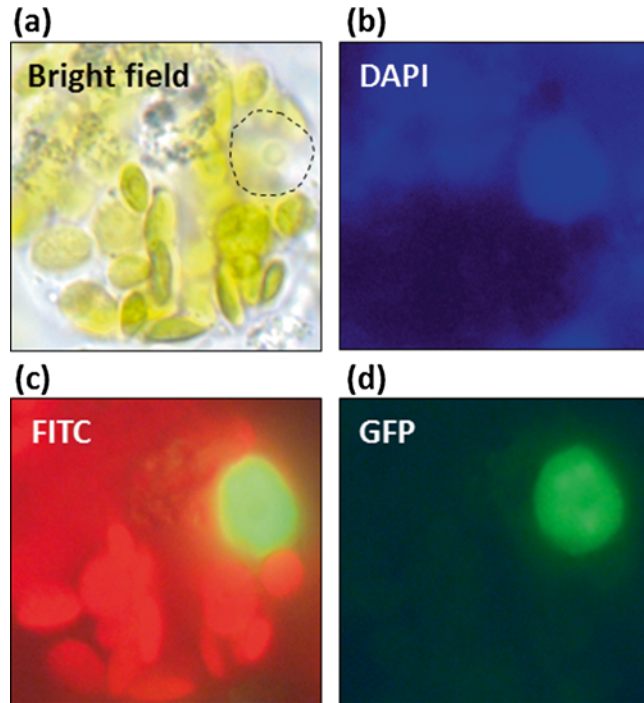


Fig. 3 Subcellular localization of At4g25210, a nucleolar localized protein [22]. *p35S:At4g25210:GFP* (*pGWB5* binary vector) expression vector was transformed into *P. patens* protoplasts revealing the nuclear localization of the encoded chimeric protein. (a) Protoplast visualization under bright field illumination (*dashed lines*: nucleus delimitation). (b) Nucleus localization under U.V. following DAPI (4',6-diamidino-2-phenylindole) staining of DNA using an exc. 365 nm/em. LP-420 nm filter. (c) FITC: fluorescence detection of chlorophylls (highlighting chloroplasts in *red*) and GFP (*green*) using an exc. 470–40 nm/em LP-515 nm filter. (d) Specific GFP fluorescence is detected using an exc. 470–40 nm/em. BP-525-50 filter

4 Notes

1. Ammonium tartrate allows the cultivation of moss protonema predominantly in the chloronema stage. PPNH_4 medium without ammonium tartrate is also possible for the yield of caulonemata and spores, and the culture can be kept in the growth room for up to 5 weeks [10].
2. 3 M micropore tape is recommended to be used for sealing petri dishes, as air exchange might be beneficial to the culture. The use of parafilm considerably reduces growth. Distribute uniformly the moss culture on the plates to avoid cell death. Cellophane disks are used to stop mosses growing down into the media. It is recommended to wet the cellophane disks in sterile ultrapure water before covering the media in order to avoid creases.

3. PPNH₄ liquid medium is recommended for moss storage, which keeps moss available for culture during a period of at least 6 months. Cefotaxime (200 mg/mL) is added to suppress bacterium contamination.
4. One tube of 10-mL driselase is used for 1 petri dish of moss culture, which produces best yield of protoplasts. In average the amount of moss protoplasts issued from 1 petri dish allows around 15 transformations. Handling a maximum of 30 transformations at one time is recommended in order to ensure stable repetitions.
5. The time for digestion of moss may vary depending on the moss strain and age, but longer exposure to driselase (up to 3 h) does not appear to affect protoplast viability greatly, but its viability may decrease as incubation time increases above 3 h.
6. Warm up MMM solution to room temperature before transformation. The competent protoplasts in MMM solution can be kept at room temperature for a couple of hours.
7. Appropriate positive and negative controls are recommended. Moss protoplasts transformed with the promoter of interest fused to GFP alone is used as negative control to remove background [9, 13–17]. For transient expression assays, dedicated vectors using the gateway[®] technology have been generated in order to (1) clone target promoters (endogenous or synthetic) or regulatory sequences (*pBS Tpp-A*) and to (2) constitutively express the assayed regulatory proteins (*pBS Tpp-B*). Subcellular localization can be achieved by using any binary vector that contains the 35S cauliflower mosaic virus promoter to drive the expression of the studied gene fused to GFP. In this regard, *pGWB5* or *pGWB6* [18] were successfully used. 10 µg of binary vector are used for subcellular localization.
8. The use of tip whose extremity has been cut off is recommended in order to avoid protoplast disruption.
9. Warm up PEG solution to room temperature before use. Heating at 60 °C can help to defreeze the solution, however the solution must be kept at room temperature once fully defrosted. Mix immediately after adding PEG solution to the samples in order to avoid high osmotic pressure to protoplasts that are partially in contact with the solution. Mixing will also protect DNA in case of nuclease/DNase contamination in the suspension [19].
10. This step is adapted from: <http://raizadalab.weebly.com/climb-protocols.html>.
11. Heat-shocked cells are competent for about 20–30 min, however transformation efficiency is best when protoplasts are left 10–15 min in concentrated PEG after heat shock [20].

12. After filtering cell suspension, keep strainers in water and wash immediately after analysis for recycling use, as cell debris clumps are very easily stuck into the micropores, especially when getting dry.
13. Beads and sheath may vary accordingly to the equipment that is used.
14. For visualization of protoplasts with flow cytometry, we routinely display outputs into two scatter plots: plot 1: FL1 vs. FL2 and plot 2: counts vs. FL1. For each parameter, logarithmic amplification is advisable as broad ranges of intensities are to be analyzed (Fig. 2).

Acknowledgement

We thank the “Plateforme de cytologie et imagerie végétale (PCIV)” of the Plant Observatory from the Institut Jean-Pierre Bourgin (IJPB) for excellent technical support. We also thank F. Charlot and F. Nogué (IJPB) for their help and advice in setting up this protocol and for providing *P. patens* spores. We acknowledge China Scholarship Council (CSC) for supporting W.X. This work was supported by two projects, STREG (Plant-KBBE) and CERES (ANR-2010-BLAN-1238).

References

1. Franco-Zorrilla JM, Lopez-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc Natl Acad Sci U S A* 111:2367–2372
2. Prouse MB, Campbell MM (2013) Interactions between the R2R3-MYB transcription factor, AtMYB61, and target DNA binding sites. *PLoS One* 8, e65132
3. Li JJ, Herskowitz I (1993) Isolation of ORC6, a component of the yeast origin recognition complex by a one-hybrid system. *Science* 262:1870–1874
4. Baudry A, Heim MA, Dubreucq B, Caboche M, Weisshaar B, Lepiniec L (2004) TT2, TT8, and TTG1 synergistically specify the expression of *BANYULS* and proanthocyanidin biosynthesis in *Arabidopsis thaliana*. *Plant J* 39:366–380
5. Berger B, Stracke R, Yatusovich R, Weisshaar B, Flügge UI, Gígolashvili T (2007) A simplified method for the analysis of transcription factor–promoter interactions that allows high-throughput data generation. *Plant J* 50:911–916
6. Fukuda H, Ito M, Sugiyama M, Komamine A (1994) Mechanisms of the proliferation and differentiation of plant cells in cell culture systems. *Int J Dev Biol* 38:287–299
7. Hartmann U, Valentine WJ, Christie JM, Hays J, Jenkins GI, Weisshaar B (1998) Identification of UV/blue light-response elements in the *Arabidopsis thaliana* chalcone synthase promoter using a homologous protoplast transient expression system. *Plant Mol Biol* 36:741–754
8. Marion J, Bach L, Bellec Y, Meyer C, Gissot L, Faure JD (2008) Systematic analysis of protein subcellular localization and interaction using high-throughput transient transformation of *Arabidopsis* seedlings. *Plant J* 56:169–179
9. Thévenin J, Dubos C, Xu W, Le Gourrierc J, Kelemen Z, Charlot F, Nogué F, Lepiniec L, Dubreucq B (2012) A new system for fast and quantitative analysis of heterologous gene expression in plants. *New Phytol* 193:504–512
10. Ashton NW, Cove DJ (1977) The isolation and preliminary characterisation of auxotrophic and analogue resistant mutants of the

- moss *Physcomitrella patens*. Mol Gen Genet 154:87–95
11. Ashton NW, Grimsley N, Cove DJ (1979) Analysis of gametophytic development in the moss, *Physcomitrella patens*, using auxin and cytokinin resistant mutants. Planta 144:427–435
 12. Schaefer DG, Zrýd JP (1997) Efficient gene targeting in the moss *Physcomitrella patens*. Plant J 11:1195–1206
 13. Dubos C, Kelemen Z, Sebastian A, Bulow L, Huep G, Xu W, Grain D, Salsac F, Brousse C, Lepiniec L, Weisshaar B, Contreras-Moreira B, Hehl R (2014) Integrating bioinformatic resources to predict transcription factors interacting with *cis*-sequences conserved in co-regulated genes. BMC Genomics 15:317
 14. Schaart JG, Dubos C, Romero De La Fuente I, van Houwelingen AM, de Vos RC, Jonker HH, Xu W, Routaboul JM, Lepiniec L, Bovy AG (2013) Identification and characterization of MYB-bHLH-WD40 regulatory complexes controlling proanthocyanidin biosynthesis in strawberry (*Fragaria* × *ananassa*) fruits. New Phytol 197:454–467
 15. Xu W, Grain D, Le Gourrierc J, Harscoet E, Berger A, Jauvion V, Scagnelli A, Berger N, Bidzinski P, Kelemen Z, Salsac F, Baudry A, Routaboul JM, Lepiniec L, Dubos C (2013) Regulation of flavonoid biosynthesis involves an unexpected complex transcriptional regulation of *TT8* expression, in Arabidopsis. New Phytol 198:59–70
 16. Xu W, Grain D, Bobet S, Le Gourrierc J, Thévenin J, Kelemen Z, Lepiniec L, Dubos C (2014) Complexity and robustness of the flavonoid transcriptional regulatory network revealed by comprehensive analyses of MYB-bHLH-WDR complexes and their targets in Arabidopsis seed. New Phytol 202:132–144
 17. Xu W, Lepiniec L, Dubos C (2014) New insights toward the transcriptional engineering of proanthocyanidin biosynthesis. Plant Signal Behav 9, e28736
 18. Nakagawa T, Kurose T, Hino T, Tanaka K, Kawamukai M, Niwa Y (2007) Development of series of gateway binary vectors, pGWBs, for realizing efficient construction of fusion genes for plant transformation. J Biosci Bioeng 104:34–41
 19. Maas C, Werr W (1989) Mechanism and optimized conditions for PEG mediated DNA transfection into plant protoplasts. Plant Cell Rep 8:148–151
 20. Berges T, Barreau C (1988) Heat shock at an elevated temperature improves transformation efficiency of protoplasts from *Podospora anserina*. J Gen Microbiol 135:601–604
 21. Xu W, Dubos C, Lepiniec L (2015) Transcriptional control of flavonoid biosynthesis by MYB-bHLH-WDR complexes. Trends Plant Sci 20:176–185
 22. Pendle AF, Clark GP, Boon R, Lewandowska D, Lam YW, Andersen J, Mann M, Lamond AI, Brown JW, Shaw PJ (2004) Proteomic analysis of the Arabidopsis nucleolus suggests novel nucleolar functions. Mol Biol Cell 16:260–269

Analysis of Microbe-Associated Molecular Pattern-Responsive Synthetic Promoters with the Parsley Protoplast System

Konstantin Kanofsky, Mona Lehmeyer, Jutta Schulze, and Reinhard Hehl

Abstract

Plants recognize pathogens by microbe-associated molecular patterns (MAMPs) and subsequently induce an immune response. The regulation of gene expression during the immune response depends largely on *cis*-sequences conserved in promoters of MAMP-responsive genes. These *cis*-sequences can be analyzed by constructing synthetic promoters linked to a reporter gene and by testing these constructs in transient expression systems. Here, the use of the parsley (*Petroselinum crispum*) protoplast system for analyzing MAMP-responsive synthetic promoters is described. The synthetic promoter consists of four copies of a potential MAMP-responsive *cis*-sequence cloned upstream of a minimal promoter and the *uidA* reporter gene. The reporter plasmid contains a second reporter gene, which is constitutively expressed and hence eliminates the requirement of a second plasmid used as a transformation control. The reporter plasmid is transformed into parsley protoplasts that are elicited by the MAMP Pep25. The MAMP responsiveness is validated by comparing the reporter gene activity from MAMP-treated and untreated cells and by normalizing reporter gene activity using the constitutively expressed reporter gene.

Key words MAMP, Parsley protoplasts, Plant–pathogen interaction, Transient reporter gene assays, Synthetic promoter

1 Introduction

Plants are infected by a wide variety of different pathogens. To induce an immune response, plants recognize the invader by pattern recognition receptors interacting with microbe-associated molecular patterns (MAMPs). The receptors activate a signaling cascade leading to basal immunity [1]. A basal immune response is associated with the upregulation of pathogen-responsive genes. These genes are regulated by transcription factors binding to specific *cis*-elements. Such *cis*-elements can be employed to generate synthetic promoters for specific expression of either reporter genes for basic research or for the expression of genes involved in the plant immune response [2, 3]. When those genes confer a hypersensitive

response upon pathogen infection, low background expression is essential otherwise plants tend to develop spontaneous necrosis [4]. Therefore, a large number of *cis*-sequences need to be isolated and tested in the context of synthetic promoters.

The identification of *cis*-elements for the development of synthetic promoters is greatly facilitated by bioinformatics. This approach assumes the conservation of specific *cis*-sequences in the promoters of genes upregulated by diverse pathogenic stimuli. This approach has been widely used for many different stress responses [5–9].

For fast and easy testing of such *cis*-sequences, transient gene expression systems are particularly helpful [10, 11]. The parsley protoplast system has been developed more than 20 years ago and has been established as a robust system to test MAMP-responsive gene expression [12]. Subsequent transformation of promoter reporter gene constructs into parsley protoplasts, the protoplasts are subjected to the MAMP Pep25, an oligopeptide from a surface glycoprotein of the phytopathogen *Phytophthora sojae* [13, 14]. After transformation, reporter gene expression is measured in the presence and absence of the MAMP. To account for differences in transformation efficiencies between different experiments, a transformation control consisting of a plasmid constitutively expressing a second reporter gene is cotransformed with the promoter reporter gene constructs [10, 11]. In this case it is important to accurately establish a constant proportion between both plasmids to avoid large variabilities between experiments.

To facilitate the control for transformation efficiency, a new plasmid was recently established which harbors a constitutively expressed second reporter gene on the same plasmid [7]. This plasmid was designated pBT10GUS-d35SLUC and is shown in Fig. 1. pBT10GUS-d35SLUC is based on pBT10GUS [10] and contains the luciferase reporter gene (LUC) expressed by a double 35S Cauliflower Mosaic Virus (CaMV) promoter (d35S) from plasmid p70S-ruc [15]. Despite the addition of a novel gene, pBT10GUS-d35SLUC maintains the possibility of easy multimerization of cloned *cis*-sequences using specific restriction enzyme sites upstream of the CaMV minimal promoter which drives the *uidA* (GUS) reporter gene [7, 10]. The inclusion of a constitutively expressed reporter gene on the same plasmid in which *cis*-responsive sequences are tested with a different reporter gene posed the question if the regulatory sequences of the d35S promoter influence the expression of the GUS gene. Testing this plasmid without *cis*-sequences cloned upstream of the GUS gene in parsley cells showed no background GUS activity [7]. Furthermore, a control promoter harboring four copies of the D-box upstream of the GUS gene did not show GUS activity in the absence of the MAMP but strong GUS activity in the presence of Pep25 [7]. The D-box was identified in the parsley *PR2* promoter [11, 16, 17]. The low background

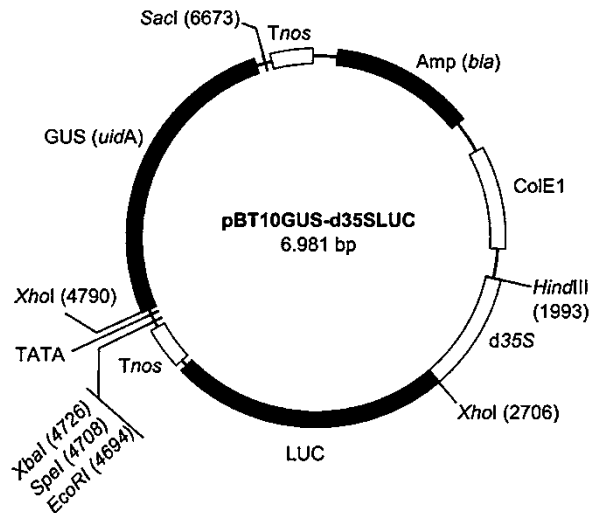


Fig. 1 The pBT10GUS-d35SLUC vector [7]. Four copies of a potential *cis*-element are cloned upstream of the minimal promoter (TATA) and the GUS reporter gene. The double CaMV 35S promoter (d35S) regulates the expression of the LUC reporter gene. Furthermore, the vector contains two nopaline synthase terminator (Tnos), an origin of replication (ColE1), and an ampicillin resistance gene (Amp)

activity of plasmid pBT10GUS-d35SLUC and the efficient way of normalization for transformation efficiency has led to the analysis of many MAMP-responsive *cis*-sequences using pBT10GUS-d35SLUC in the parsley protoplast system [7, 18–20].

This chapter illustrates how the classic parsley protoplast system can be used together with pBT10GUS-d35SLUC to analyze potential MAMP-responsive *cis*-sequences. The protocol includes the cultivation of the parsley suspension culture, generation of protoplasts, transformation of protoplasts, and measurement and normalization of reporter gene activity.

2 Materials

2.1 Equipment and Material

1. Laminar flow hood.
2. 100 mL Erlenmeyer flasks and corresponding shaker (e.g. GFL shaker 3005; GFL Gesellschaft für Labortechnik mbH, Burgwedel, Germany).
3. Plant growth chamber (e.g. Percival CU-36L4; CLF Plant Climatics, Wertingen, Germany).
4. Sodium spoon (Copper woven with wooden handle. Total length 185 mm. Spoon diameter 26.35 mm, mesh aperture 1 mm; Carl Roth GmbH + Co. KG, Karlsruhe, Germany).

5. 50 mL centrifuge tubes.
6. 15 mL centrifuge tubes.
7. 1.5 mL microcentrifuge tubes.
8. Centrifuge for 15 and 50 mL tubes and corresponding swing bucket rotor (e.g. Beckman GPKR centrifuge and Beckman GH 3.7 Swing Bucket Rotor; Beckman Instruments GmbH, München, Germany).
9. Microcentrifuge.
10. 0.2 µm sterile filters.
11. 150 mm Ø petri dishes.
12. Parafilm.
13. Celloshaker Variospeed (e.g. Renner GmbH, Dannstadt-Schauernheim, Germany).
14. 3 mL transfer pipette.
15. Microscope.
16. Microliter pipette with sterile tips.
17. Cell-saver pipette tips.
18. Vortexer.
19. Mixer 5432 (e.g. Eppendorf; Wesseling-Berzdorf, Germany).
20. Transparent, white, and black 96-well microtiter plates.
21. TriStar® LB 941 microplate reader (Berthold Technologies GmbH & Co. KG, Bad Wildbad, Germany).

2.2 Protoplast Transformation and Cell Culture Reagents

Prepare all solutions using ddH₂O and store at room temperature (*see* **Note 1**).

1. Plasmids: pBT10GUS-d35SLUC and 4D-pBT10GUS-d35SLUC (*see* **Note 2**).
2. Parsley (*Petroselinum crispum*) suspension culture Pc 5/3.
3. HA medium: 2500 mg/L KNO₃, 171 mg/L CaCl₂·2H₂O, 250 mg/L MgSO₄·7H₂O, 134 mg/L (NH₄)₂SO₄, 150 mg/L NaH₂PO₄·H₂O, 0.75 mg/L KI, 3 mg/L H₃BO₃, 11.2 mg/L MnSO₄·H₂O, 3 mg/L ZnSO₄·7H₂O, 0.25 mg/L Na₂MoO₄·2H₂O, 0.39 mg/L CuSO₄·5H₂O, 0.25 mg/L CoCl₂·6H₂O, 13.9 mg/L FeSO₄·7H₂O, 18.6 mg/L Na₂-EDTA, 100 mg/L myo-inositol, 1 mg/L nicotinic acid, 1 mg/L pyridoxine-HCl, 10 mg/L thiamine-HCl, 1 mg/L 2,4-dichlorophenoxyacetic acid, and 20 g/L of sucrose, adjust to pH 5.5 with 1 M KOH. Sterilize by autoclaving.
4. CaCl₂ solution: 0.24 M CaCl₂. Sterilize by autoclaving.
5. Enzyme solution: 0.5% (w/v) Cellulase Onozuka R-10 (Duchefa Biochemie, Haarlem, The Netherlands), 1.08% (w/v) Macerozyme R-10 (Duchefa Biochemie, Haarlem,

The Netherlands) in 0.24 M CaCl₂. Add 0.15 g cellulase and 0.325 g macerozyme to a 50 mL tube. Dissolve in 30 mL 0.24 M CaCl₂ by stirring for approximate 2 h. Sterilize by filtering (*see Note 3*).

6. P5 medium: 3.164 g/L Gamborg B5 medium including vitamins (Duchefa Biochemie, 2003 RV Haarlem, the Netherlands), 0.28 M sucrose, 1 mg/L 2,4-dichlorophenoxyacetic acid, adjust to pH 5.5 with 1 M KOH. Sterilize by autoclaving.
7. PEG solution: 25 % (w/v) polyethylene glycol 6000, 100 mM Ca(NO₃)₂, 45 mM mannitol, adjust to pH 9.0 with 0.1 M KOH. Sterilize by filtering. Store at -20 °C.
8. Ca(NO₃)₂-MES-solution: 275 mM Ca(NO₃)₂, 2 mM MES (2-(N-morpholino)ethanesulfonic acid), adjust to pH 6.0 with 1 M KOH. Sterilize by autoclaving.
9. Pep25: 100 µg/mL synthetic Pep25 (DVTAGAEVWNQPV RGFKVYEQTMT) from a commercial supplier. Store at -20 °C.
10. LUC extraction buffer: 0.1 M NaH₂PO₄, 1 mM dithiothreitol, adjust to pH 7.8 with 1 M NaOH.
11. Bradford solution: 2 mL 5xRoti[®]-Quant Reagenz (Carl-Roth, Karlsruhe, Germany), 5.5 mL ddH₂O.
12. LUC reaction buffer: 15 mM MgSO₄, 25 mM glycylglycin, adjust to pH 7.8 with 1 M KOH, 5 mM ATP. Prepare a stock solution without ATP and store at 4 °C. For measuring LUC activity add 5 mM ATP from a 100 mM ATP stock stored at -20 °C.
13. Luciferin substrate: 0.2 mM luciferin, 25 mM glycylglycin, adjust to pH 7.8 with 1 M KOH. Store a 25 mM glycylglycin solution, pH 7.8 at 4 °C and add 0.2 mM luciferin from a 10 mM stock solution stored at -20 °C before measuring LUC activity.
14. GUS reaction buffer: 50 mM NaPO₄, pH 7.0, 10 mM Na₂EDTA, 0.1 % (v/v) Triton X-100, 0.1 % (w/v) N-lauryl sarcosine. This solution can be stored at 4 °C. Before measuring GUS activity add 10 mM β-mercaptoethanol and 1 mM 4-methylumbelliferyl-β-D-glucuronide.

3 Methods

3.1 Cultivation of the Parsley Suspension Culture

Carry out all following procedures in a laminar flow hood unless otherwise specified.

1. Parsley suspension culture Pc 5/3 is cultivated at 23 °C by shaking at 160 rpm in the dark in a 100 mL Erlenmeyer flask containing 40 mL HA medium in a plant growth chamber.

2. Every seventh day approximate 3 mL cell culture are transferred into 40 mL fresh HA medium.
3. For isolation of protoplasts a subculture is required which will be obtained by transferring cells with one full sterile sodium spoon from a 7-day-old culture into 40 mL fresh HA medium and cultivated for 5 days (*see Note 4*).

3.2 Isolation of Protoplasts

1. Divide a 5-day-old parsley suspension culture into two 50 mL centrifuge tubes and centrifuge at $300\times g$ for 5 min at room temperature.
2. Discard the supernatant.
3. Resuspend each cell pellet in 15 mL enzyme solution and fill up to 45 mL with CaCl_2 solution.
4. Transfer each cell suspension carefully into a 150 mm \varnothing petri dish, seal with Parafilm, and shake (Celloshaker Variospeed) for 20 h at 15 rpm and 23°C in a plant growth chamber and darkness, and subsequently for 20 min at 40 rpm (*see Note 5*).
5. Transfer each suspension slowly with a 3 mL transfer pipette into 50 mL tubes and centrifuge for 2 min at $300\times g$ and room temperature.
6. Discard the supernatant gently and fill up to 30 mL with CaCl_2 solution, resuspend the protoplast suspension (do not vortex!) and repeat the centrifugation.
7. Discard the supernatant and fill up to 25 mL with P5 medium and resuspend the protoplasts. Combine both protoplast suspensions into one new 50 mL tube and centrifuge for 5 min at $300\times g$ and room temperature.
8. After the centrifugation, intact protoplasts float on the surface of the medium (*see Note 6*). Transfer these protoplasts gently with a 3 mL transfer pipette to a new 50 mL tube and fill it up to 50 mL with P5 medium, invert the tube carefully, and centrifuge for 5 min at $300\times g$ and room temperature.
9. Divide the floating protoplasts into two 15 mL tubes, fill up to 15 mL with P5 medium and centrifuge as before.
10. Combine the floating protoplasts into one 15 mL tube, fill up to 15 mL with P5 medium and centrifuge as before.
11. Transfer the floating protoplasts into a new 15 mL tube (*see Note 7*). Obtained protoplasts can be used for transformation.

3.3 Transformation of Parsley Protoplasts and MAMP-Treatment

1. Mix 10 μg plasmid DNA (20 μL 0.5 $\mu\text{g}/\mu\text{L}$) with 200 μL PEG-solution in a 15 mL tube (*see Note 8*).
2. Add 200 μL of parsley protoplasts and mix gently (*see Note 9*).
3. Incubate protoplast-PEG-DNA-solution for 20 min at room temperature in the dark.

4. To stop the transformation add 5 mL $\text{Ca}(\text{NO}_3)_2$ -MES-solution (*see Note 10*).
5. Invert the 15 mL tube and collect the transformed protoplasts by centrifugation at $150\times g$ and room temperature for 7 min.
6. Discard the supernatant and resuspend the pellet with 6 mL P5 medium carefully.
7. Transfer 3200 μL of the suspension into a new 15 mL tube and add 9 μL Pep25 to a final concentration of 300 ng/mL (*see Note 11*).
8. Incubate both tubes with the transformed protoplasts for 24 h at 23 °C in the dark in the plant growth chamber.
9. To harvest the protoplasts fill up to 12 mL with CaCl_2 , invert and centrifuge for 10 min at $1400\times g$ and room temperature.
10. Take of the supernatant and leave 1 mL in the tubes. Resuspend the protoplast pellets in the remaining mL and transfer each suspension into a 1.5 mL tube.
11. Collect the protoplasts by centrifugation for 30 s at $16,000\times g$ and room temperature.
12. Take of the supernatant, freeze the pellet in liquid nitrogen, and store at $-80\text{ }^\circ\text{C}$.

3.4 Protein Extraction

For the following procedures, a laminar flow hood is not needed.

1. Put the frozen samples on ice, open the tube, and incubate for 5 min.
2. Add 150 μL cold LUC extraction buffer to each frozen sample, resuspend by vortexing, and shake (mixer 5432; Eppendorf) for 20 min at 4 °C.
3. Centrifuge for 10 min at 4 °C and $25,000\times g$.
4. Transfer the supernatant into a new 1.5 mL tube and store on ice.

3.5 Protein Quantification

1. To calculate the protein concentrations prepare a 1:10 dilution by mixing 30 μL of protein solution with 270 μL ddH₂O.
2. Each protein dilution will be measured in three replicates. Therefore, transfer three times 50 μL of protein dilution into wells of a transparent 96-well microtiter plate and add 200 μL Bradford solution. For blank value mix 50 μL LUC extraction buffer with 200 μL Bradford solution.
3. Incubate the plate for 5 min in darkness.
4. Measure the absorbance of each sample with excitation 590 nm (0.1 counting time; 13,000 lamp energy) in the TriStar® LB 941 (Fig. 2).



Fig. 2 The TriStar® LB 941 microplate reader used for reporter gene assays

5. The protein concentration will be determined by using a standard curve (*see Note 12*).
6. For each sample prepare a 200 μL protein solution with a protein concentration of 80 $\mu\text{g}/\text{mL}$ in LUC extraction buffer.

3.6 LUC Assay

1. The LUC activity will be measured in duplicates. Therefore transfer two times 50 μL of diluted sample (4 μg protein) into wells of a white 96-well microtiter plate (on ice) and insert into the TriStar® LB 941 (Fig. 2). 50 μL of LUC extraction buffer is used as blank (*see Note 13*).
2. Via one injector 175 μL of LUC reaction buffer is added to each well while another injector dispenses 50 μL of luciferin substrate. The luminescence is measured for 15 s.
3. For determining the LUC activity each integral will be corrected by the blank value. The LUC activity is calculated in RLU/s/mg protein (*see Note 14*).

3.7 GUS Assay

1. The GUS activity will be measured in duplicates. Transfer twice 25 μL of diluted protein solution (2 μg protein) into wells of a black 96-well microtiter plate.
2. Add 200 μL GUS reaction buffer to each well.
3. Insert the plate into the TriStar® LB 941 preheated to 37 $^{\circ}\text{C}$ (Fig. 2).
4. After incubation at 37 $^{\circ}\text{C}$ for 10 min the GUS activity is measured for 3 h and 37 $^{\circ}\text{C}$. Each well is measured 12 times every 15 min for 1 s (excitation 360 nm; emission 460 nm).

5. The GUS activity is calculated in pmol 4-Methylumbelliferone (4-MU)/min/mg protein by dividing the linear regression of the measured fluorescence over the time and the slope of the standard curve with 4-MU in a concentration range from 0 μ M up to 75 μ M (*see Note 15*).
6. All GUS values are normalized by corresponding LUC values from Pep25-untreated cells (*see Note 16*).
7. For normalization of the GUS values, one LUC value (without Pep25 elicitor) was selected and all other LUC values without elicitor were divided by this selected LUC value. The obtained quotients were used to divide corresponding GUS values with and without elicitor. Standard deviations were calculated from these normalized GUS values (*see Note 17*).
8. For each synthetic promoter the normalized GUS values of Pep25 treated and untreated samples will be compared to evaluate the Pep25 responsiveness of examined *cis*-elements.

4 Notes

1. Protoplasts are sensitive to osmotic stress. Thus all solutions should be prepared very carefully.
2. The plasmids pBT10GUS-d35SLUC (Fig. 1) and 4D-pBT10GUS-d35SLUC have been described before [7]. In case novel *cis*-sequences are being tested for MAMP responsiveness, forward and reverse oligonucleotides should correspond to a monomer of this *cis*-sequence. Design complementary oligonucleotides with partial *SpeI* and *XbaI* sites. The forward oligonucleotide should have the sequence 5'-CTAGT_(x)T-3', and the reverse oligonucleotide should have the sequence 5'-CTAGAN_(x)A-3'. N_(x) designates the length of the complementary oligonucleotides. These oligonucleotides should not have *SpeI*, *SacI*, and *XbaI* restriction sites because these enzymes are used during multimerization. Also the forward oligonucleotide should not end with GA, otherwise a methylation site (GATC) will be generated. If using an *E. coli* strain containing the *dam* gene, the *XbaI* recognition sequence will be methylated and *XbaI* cannot cut any longer. It is recommended that oligonucleotides are not too short, because often combinatorial elements are required for MAMP-responsive gene expression [19, 21]. Cloning and multimerization of the oligonucleotides have been described before [7, 10]. When testing novel *cis*-sequences, plasmids pBT10GUS-d35SLUC and 4D-pBT10GUS-d35SLUC can be used as negative and positive controls, respectively.
3. Longer stirring time facilitates sterile filtration or use a filtration unit with a pre-filter.

4. Protoplast yield strongly depends on culture conditions of the cell suspension.
5. Check the isolated protoplasts under a microscope. Here you can see if the incubation time of the suspension culture in the enzyme solution was too short, how many cells died, or if there is any contamination. If a contamination occurred the experiment should be stopped and restarted with a new parsley culture.
6. If the protoplasts are not floating, check your prepared P5 medium. The correct sucrose concentration is required for floating protoplasts.
7. The centrifugation steps before (**steps 7–10**) are required to get rid of dead protoplasts and to increase the density of intact protoplasts. The volume of intact protoplasts after the last centrifugation should be 4–6 mL. If the obtained volume is higher than 6 mL, the density of protoplasts is too low. Another centrifugation step should be done. When carefully transferring floating protoplasts the volume should be as small as possible.
8. For plasmid preparation it is recommended to use a commercial kit and to use always the same method. Impurities of plasmid preparations can result in large differences in transformation efficiencies.
9. To transfer parsley protoplasts use cell-saver pipet tips.
10. We transform several reactions in a 30 s time interval. To stop one reaction we first add 1 mL $\text{Ca}(\text{NO}_3)_2$ -MES-solution and after all reactions are stopped, we add additional 4 mL $\text{Ca}(\text{NO}_3)_2$ -MES-solution.
11. At this step you have two protoplast suspensions from one transformation, treated and untreated with the MAMP Pep25.
12. The protein standard curve is obtained by preparing several protein dilutions with different concentrations from 0 $\mu\text{g}/\text{mL}$ to 100 $\mu\text{g}/\text{mL}$. Each protein concentration will be measured in three replicates. Therefore, transfer three times 50 μL of protein dilution into wells of a transparent 96-well microtiter plate and add 200 μL Bradford solution. Incubate the plate for 5 min in darkness and measure the absorbance of each sample with excitation 590 nm (0.1 counting time; 13,000 lamp energy) in the TriStar[®] LB 941. It is recommended to renew the calibration once in a while.
13. It is important to measure the LUC-activity before GUS-activity, because the luciferase is not as stable as β -glucuronidase.
14. The TriStar[®] LB 941 can be programmed to perform these calculations by using the provided MikroWin software. The MikroWin software defines the instrument settings and evaluates the measured data. The command “KITG(MES)”

calculates the integral of each sample to determine the LUC activity.

15. A computer with MikroWin software linked to the TriStar® LB 941 can calculate the linear regression of the measured fluorescence over the time by using the command “KSLP(MES)”. To plot a standard curve, a calibration of fluorescence units with defined amounts of 4-MU was performed in the TriStar® LB 941. A linear increase of fluorescence units with 4-MU concentrations has been observed up to at least 75 µM.
16. Only LUC values from Pep25-untreated cells are used, because the MAMP Pep25 has an effect of the LUC activity. The transformation efficiency of treated and untreated cells should be the same, because protoplasts of the same transformation are divided before treatment.
17. It is recommended to perform at least three independent experiments (biological replicates) with two measurements for each experiment with and without Pep25 (technical replicates). The standard deviation of the same experiment will be calculated from all measurements.

Acknowledgements

This work was supported by the KWS SAAT AG, Einbeck, Germany.

References

1. Tena G, Boudsocq M, Sheen J (2011) Protein kinase signaling networks in plant innate immunity. *Curr Opin Plant Biol* 14(5):519–529
2. Gurr SJ, Rushton PJ (2005) Engineering plants with increased disease resistance: how are we going to express it? *Trends Biotechnol* 23(6):283–290
3. Liu W, Stewart CN Jr (2016) Plant synthetic promoters and transcription factors. *Curr Opin Biotechnol* 37:36–44. doi:10.1016/j.copbio.2015.10.001
4. Niemeyer J, Ruhe J, Machens F, Stahl DJ, Hehl R (2014) Inducible expression of p50 from TMV for increased resistance to bacterial crown gall disease in tobacco. *Plant Mol Biol* 84:111–123. doi:10.1007/s11103-013-0122-4
5. Harb A, Krishnan A, Ambavaram MM, Pereira A (2010) Molecular and physiological analysis of drought stress in *Arabidopsis* reveals early responses leading to acclimation in plant growth. *Plant Physiol* 154(3):1254–1271. doi:10.1104/pp.110.161752
6. Zou C, Sun K, Mackaluso JD, Seddon AE, Jin R, Thomashow MF, Shiu SH (2011) Cis-regulatory code of stress-responsive transcription in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A* 108(36):14992–14997. doi:10.1073/pnas.1103202108
7. Koschmann J, Machens F, Becker M, Niemeyer J, Schulze J, Bülow L, Stahl DJ, Hehl R (2012) Integration of bioinformatics and synthetic promoters leads to the discovery of novel elicitor-responsive cis-regulatory sequences in *Arabidopsis*. *Plant Physiol* 160:178–191. doi:10.1104/pp.112.198259
8. Liu W, Mazarei M, Peng Y, Fethe MH, Rudis MR, Lin J, Millwood RJ, Arelli PR, Stewart CN Jr (2014) Computational discovery of soybean promoter cis-regulatory elements for the construction of soybean cyst nematode-inducible synthetic promoters. *Plant Biotechnol J* 12(8):1015–1026. doi:10.1111/pbi.12206
9. Dubos C, Kelemen Z, Sebastian A, Bülow L, Huep G, Xu W, Grain D, Salsac F, Brousse C, Lepiniec L, Weisshaar B, Contreras-Moreira B,

- Hehl R (2014) Integrating bioinformatic resources to predict transcription factors interacting with cis-sequences conserved in co-regulated genes. *BMC Genomics* 15(1):317. doi:[10.1186/1471-2164-15-317](https://doi.org/10.1186/1471-2164-15-317)
10. Sprenger-Haussels M, Weisshaar B (2000) Transactivation properties of parsley proline-rich bZIP transcription factors. *Plant J* 22(1): 1–8
 11. Rushton PJ, Reinstadler A, Lipka V, Lippok B, Somssich IE (2002) Synthetic plant promoters containing defined regulatory elements provide novel insights into pathogen- and wound-induced signaling. *Plant Cell* 14(4): 749–762
 12. Hahlbrock K, Scheel D, Logemann E, Nürnberger T, Parniske M, Reinold S, Sacks W, Schmelzer E (1995) Oligopeptide elicitor-mediated defense gene activation in cultured parsley cells. *Proc Natl Acad Sci U S A* 92: 4150–4157
 13. Nürnberger T, Nennstiel D, Jabs T, Sacks WR, Hahlbrock K, Scheel D (1994) High affinity binding of a fungal oligopeptide elicitor to parsley plasma membranes triggers multiple defense responses. *Cell* 78(3): 449–460
 14. Rushton PJ, Torres JT, Parniske M, Wernert P, Hahlbrock K, Somssich IE (1996) Interaction of elicitor-induced DNA-binding proteins with elicitor response elements in the promoters of parsley PR1 genes. *EMBO J* 15(20):5690–5700
 15. Stahl DJ, Kloos DU, Hehl R (2004) A sugar beet chlorophyll a/b binding protein promoter void of G-box like elements confers strong and leaf specific reporter gene expression in transgenic sugar beet. *BMC Biotechnol* 4(1):31
 16. van de Löcht U, Meier I, Hahlbrock K, Somssich IE (1990) A 125 bp promoter fragment is sufficient for strong elicitor-mediated gene activation in parsley. *EMBO J* 9(9): 2945–2950
 17. Kirsch C, Takamiya-Wik M, Schmelzer E, Hahlbrock K, Somssich IE (2000) A novel regulatory element involved in rapid activation of parsley ELI7 gene family members by fungal elicitor or pathogen infection. *Mol Plant Pathol* 1(4):243–251. doi:[10.1046/j.1364-3703.2000.00029.x](https://doi.org/10.1046/j.1364-3703.2000.00029.x)
 18. Machens F, Becker M, Umrath F, Hehl R (2014) Identification of a novel type of WRKY transcription factor binding site in elicitor-responsive cis-sequences from *Arabidopsis thaliana*. *Plant Mol Biol* 84:371–385. doi:[10.1007/s11103-013-0136-y](https://doi.org/10.1007/s11103-013-0136-y)
 19. Lehmeyer M, Kanofsky K, Hanco EKR, Ahrendt S, Wehrs M, Machens F, Hehl R (2016) Functional dissection of a strong and specific microbe-associated molecular pattern-responsive synthetic promoter. *Plant Biotechnol J* 14(1):61–71. doi:[10.1111/pbj.12357](https://doi.org/10.1111/pbj.12357)
 20. Lehmeyer M, Hanco EKR, Roling L, Gonzalez L, Wehrs M, Hehl R (2016) A cis-regulatory sequence from a short intergenic region gives rise to a strong microbe-associated molecular pattern-responsive synthetic promoter. *Mol Genet Genomics* 291(3):1155–1165. doi:[10.1007/s00438-016-1173-4](https://doi.org/10.1007/s00438-016-1173-4)
 21. Singh KB (1998) Transcriptional regulation in plants: the importance of combinatorial control. *Plant Physiol* 118(4):1111–1120

Chapter 12

A Framework for Discovering, Designing, and Testing MicroProteins to Regulate Synthetic Transcriptional Modules

Elisa Fiume, Niek de Klein, Seung Yon Rhee, and Enrico Magnani

Abstract

Transcription factors often form protein complexes and give rise to intricate transcriptional networks. The regulation of transcription factor multimerization plays a key role in the fine-tuning of the underlying transcriptional pathways and can be exploited to modulate synthetic transcriptional modules. A novel regulation of protein complex formation is emerging: microProteins—truncated transcription factors—engage in protein–protein interactions with transcriptional complexes and modulate their transcriptional activity. Here, we outline a strategy for the discovery, design, and test of putative miPs to fine-tune the activity of transcription factors regulating synthetic or natural transcriptional circuits.

Key words microProteins, Synthetic promoter, Transcription factor, Multimerization, Protein complex, Protein–protein interaction domain

1 Introduction

Transcription factors (TFs) often work as protein complexes in transcriptional regulation networks and signal transduction pathways. The formation of TF dimers or oligomers can affect their function by changing, for example, DNA-binding specificity or localization [1]. Therefore, the regulation of protein multimerization is a fundamental step for the modulation of the underlying cellular events and can be exploited in synthetic biology. A novel layer of transcriptional regulation by microProteins (miPs)—truncated transcription factors—is emerging and might have considerable implication in the design of synthetic transcriptional networks [2]. miPs are TFs carrying a protein–protein interaction domain but lacking a DNA-binding domain. miPs play critical regulatory roles by engaging in protein–protein interactions in transcriptional complexes [2]. They provide positive or negative feedback controls to fine-tune the action of their target TFs. miPs have been shown

to prevent the formation of TF complexes, titrate TFs in the cytosol, or work as cofactors in active transcriptional protein complexes [2]. The development and application of a software program, *miP Prediction Program* (miP3), to predict miPs across genomes has recently revealed a potentially ubiquitous layer of miP regulation that has expanded considerably in plants [2, 3].

Here we present a framework for the discovery, design, and preliminary test of miPs to target TFs engaged in the regulation of synthetic or natural promoters. Since miPs have been predicted to exist widely in plants, it is convenient to start searching for natural miPs targeting a TF or a class of TFs of interest [2]. We show how to use the miP3 software to predict miPs potentially regulating a TF of interest from any genome. In addition, we present guidelines for designing synthetic miPs starting from a TF sequence. Finally, we describe a method for the *in vivo* test of miP action on the TF-promoter module of interest.

2 Materials

2.1 *Agrobacterium tumefaciens* Transformation and Growth

1. Liquid LB-medium: For 1 L of LB medium, dissolve 10 g NaCl, 10 g tryptone, and 5 g yeast extract in 950 mL deionized water. Adjust the pH to 7.0 using 1 N NaOH and bring volume up to 1 L with deionized water. Sterilize the medium by autoclave for 20 min at 15 psi and store at room temperature.
2. Antibiotics stock solutions: 50 mg/mL gentamicin in deionized water. 50 mg/mL rifampicin in methanol. The antibiotic stock solution for the selection of the binary vector (*see* below) depends on the vector employed. Filter sterilize the solution through a 0.2 μm membrane and store at $-20\text{ }^{\circ}\text{C}$.
3. 100 mM Acetosyringone (5'-Dimethoxy-4'-hydroxyacetophenone) stock solution in ethanol. Filter sterilize the solution through a 0.2 μm membrane and store at $-20\text{ }^{\circ}\text{C}$.
4. 0.5 M MES stock solution in deionized water. Adjust the pH to 5.6 with KOH, sterilize the medium by autoclave for 20 min at 15 psi and store at room temperature.
5. Infiltration solution: 10 mM MgCl_2 and 10 mM MES (pH 5.6) solution in deionized water. Sterilize the medium by autoclave for 20 min at 15 psi and store at room temperature. Add acetosyringone to the solution to a final 100 μM concentration (*see* **Note 1**).
6. 1.5 mL, 15 mL, and 50 mL tubes.
7. Heated shaker.
8. Centrifuge for 50 mL tubes.
9. Spectrophotometer.
10. 1 mL syringes.

11. *Agrobacterium tumefaciens* strain *GV3101* hosting the synthetic promoter plasmid (SP plasmid), a binary vector carrying the synthetic promoter of interest upstream of the reporter gene *uidA*, also known as *GUS*, which encodes the enzyme β -glucuronidase (*see* **Notes 2** and **3**).
12. *A. tumefaciens* strain *GV3101* hosting the transcription factor plasmid (TF plasmid), a binary vector carrying the gene encoding the transcription factor of interest downstream of the constitutive cauliflower mosaic virus 35S promoter (35S).
13. *A. tumefaciens* strain *GV3101* hosting the empty TF plasmid, the backbone of the TF plasmid without the gene encoding the transcription factor of interest.
14. *A. tumefaciens* strain *GV3101* hosting the microProtein plasmid (miP plasmid), a binary vector carrying the gene encoding the microProtein of interest downstream of the 35S promoter.
15. *A. tumefaciens* strain *GV3101* hosting the empty miP plasmid, the backbone of the miP plasmid without the gene encoding the miP of interest.
16. *Nicotiana benthamiana* plants grown for 2–4 weeks under controlled conditions (24 °C and 40–65 % relative humidity) and a long-day photoperiod (14 h light and 10 h dark, with illumination of 130–150 μ E).

2.2 Quantitative Measurement of GUS Activity

1. GUS extraction solution: 50 mM NaHPO₄ (pH 7.0), 10 mM β -mercaptoethanol, 10 mM EDTA, 0.1 % (w/v) sodium lauryl sarcosine, and 0.1 % (w/v) Triton X-100 in deionized water. Prepare a fresh solution every time.
2. Bradford reagent.
3. Bovine serum albumin (BSA).
4. Carbonate stop solution (5 \times stock solution): 1 M Na₂CO₃ in deionized water. Sterilize the solution by autoclave for 20 min at 15 psi and store at room temperature.
5. GUS assay solution: 2 mM 4-MUG (4-methylumbelliferyl β -D-glucuronide) in GUS extraction solution (*see* **Note 4**). Prepare a fresh solution every time and keep it on ice protected from light (*see* **Note 5**).
6. 4-MU calibration stock solution: 1 mM 4-MU (7-hydroxy-4-methylcoumarin) in deionized water. Store at 4 °C protected from light.
7. Plastic pestles that fit 1.5 mL tubes.
8. Centrifuge for 1.5 mL tubes.
9. Fluorimeter.

2.3 miP3 Prediction

1. A computer with an operating system that has at least 4GB of RAM and that can install Python 2.7.7.
2. miP3 software from <https://github.com/npklein/miP3/releases/> [3]. This includes all input files needed to run miP3 with *Arabidopsis thaliana* transcription factors.
3. Python 2.7.7 from <http://www.python.org/download/releases/2.7/>.
4. Biopython module from <http://biopython.org/wiki/Download> [4].
5. BLAST 2.2.29+ from <http://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/2.2.29/> [5].
6. SOAPpy from <http://www.aheil.de/2013/08/17/soap-with-python/>.

3 Methods

3.1 Using the miP3 Software to Predict miPs

The miP3 software can predict natural miPs from any genome of interest. The performance of the software has been tested for the prediction of miPs targeting TFs in *A. thaliana* using the default parameters described below [3]. We provide instructions on how to change the parameters and test miP3 against other genomes.

1. Download and install on the Desktop the software components listed in Subheading 2 (*see Note 6*).
2. Open the command line and change the directory to match the location of miP3.py on the desktop by using the following command:


```
cd ~/Desktop/miP3_version_2/
```
3. Run miP3 from the command line (with default parameter values) with the following command (*see Notes 7, 8, and 9*):


```
python2.7 miP3.py -p TAIR10_pep_20101214
-i arabidopsis_transcription_factors.fasta -f Pfam.txt -o miP_output.csv
-b ncbi_blast_2.2.29+/bin/
```
4. Change the parameters (*see Note 10*) as needed or desired. *See Table 1* for detailed information about each parameter. An example command with all the parameters is provided below:


```
python2.7 miP3.py -p TAIR10_pep_20101214
-i arabidopsis_transcription_factors.fasta -f Pfam.txt -o miP_output.csv
-b ncbi_blast_2.2.29+/bin/ -s 200 -a 500 -e 0.0001 -z 0.1 -x 0.01
```

Table 1
miP3 parameters [3]

Parameter	Input	Example value
-p	FASTA formatted file with all protein sequences of the organism of interest (<i>see Note 11</i>).	<i>A. thaliana</i> proteome
-i	FASTA formatted file with all TF proteins in the organism of interest (<i>see Note 12</i>).	<i>A. thaliana</i> TF proteins
-f	Newline-delimited file with all domains that are important for the functioning of the protein class.	DNA-binding domains from Interpro database [6].
-o	Name of the output file.	miP_output.csv
-b	Path to the BLAST/bin/ folder that contains the blastp and makeblastdb program (<i>see Note 13</i>).	ncbi_blast_2.2.29+/bin/
-s	Maximum size of small proteins. It is more difficult to find homologs of smaller proteins, so proteins below this size have a different method for homolog detection.	200
-a	Maximum size of all proteins. Modify the parameter if you are only interested in proteins up to a certain size.	550
-e	E-value used in the BLASTP search against all proteins (<i>see Note 14</i>).	0.0000001
-z	E-value used in the BLASTP search against small proteins (<i>see Note 14</i>).	0.5
-x	E-value used in the reBLASTP search (<i>see Note 14</i>).	0.1

- Retrieve the results. The results are written in a tab-delimited file (*see Note 15*) (Fig. 1, Table 2) and can be opened in Excel or other text-editing software.
- Test any predicted miP as described in Subheading 3.3 (*see Note 16*).

3.2 Designing Synthetic miPs

- If miP3 predicts miPs targeting members of the TF family of interest but not the TF under study, retrieve such putative miP protein sequences (*see Note 16*). Perform a multiple sequence alignment of the TF of interest and the miP protein sequences. Paste the miP and TF sequences in the input box at the CLUSTAL OMEGA software website (<http://www.ebi.ac.uk/Tools/msa/clustalo/>), keeping the default settings, and run the job [7]. Select the protein region of the TF that aligns significantly to the miPs and amplify its encoding DNA sequence by PCR using a forward primer carrying a start codon at the 5' end and a reverse primer carrying a stop codon at the 5' end (both start and stop codons must be in frame with the gene sequence to be amplified). Clone these sequences in the miP plasmid (*see Subheading 2*) and test them as putative miPs as described in Subheading 3.3.

	A	B	C	D
1	miP	TFs	miP domains	miP length
2	AT3G25950.1	AT5G14280.1	ipr006634	251
3	AT4G04740.1	AT4G12020.2	ipr018247, ipr011009, ipr000719	520
4	AT3G50390.1	AT4G25440.1	ipr001680, ipr020472, ipr015943	469
5	AT5G12090.1	AT4G12020.2	ipr002290, ipr008271, ipr011009	369
6	AT2G07180.2	AT4G12020.1	ipr011009, ipr000719, ipr001245	442
7	AT4G19110.2	AT4G12020.1	ipr011009, ipr000719, ipr008271	464
8	AT5G46490.1	AT4G12020.1	ipr003593, ipr002182, ipr027417	357
9	AT2G30980.1	AT4G12020.1	ipr011009, ipr000719, ipr008271	412
10	AT2G46990.1	AT5G62000.3	ipr011525, ipr003311	175

Fig. 1 miP3 output file

Table 2

Explanation of the data in the miP3 output file

Column	Column name	Explanation of column value
A	miP	Gene identifier of the predicted miP
B	TFs	Gene identifiers of the transcription factors predicted to be targeted by the putative miPs in column A
C	miP domains	Interpro domain identifiers of the domains predicted to be in the putative miP in column A
D	miP length	The length of the predicted miP in column A

2. If miP3 does not predict any miP for the TF family of interest, design a synthetic miP. Identify protein domains present in the TF by pasting the TF protein sequence in the input box of the “sequence search” page at the Pfam website (<http://pfam.xfam.org/>), keeping the default settings, and run the job [8].
 - (a) If Pfam detects a protein–protein interaction (PPI) domain, amplify its encoding DNA sequence by PCR using a forward primer carrying the start codon at the 5′ end and a reverse primer carrying a stop codon at the 5′ end (both start and stop codon must be in frame with the gene sequence to be amplified). Clone these sequences in the miP plasmid (*see* Subheading 2) and test them as putative miPs as described in Subheading 3.3.
 - (b) If Pfam detects a DNA-binding (DB) domain but not a PPI domain, amplify the DNA sequences encoding the region

upstream or downstream of the DB by PCR using primers carrying the start and stop codons as described above. Clone these sequences in the miP plasmid (*see* Subheading 2) and test them as putative miPs as described in Subheading 3.3.

- (c) If Pfam does not detect any PPI or DB domain, predict any coiled coil region (a potential PPI domain) in the TF [9]. Paste the TF protein sequence in the input box at the COILS website (http://www.ch.embnet.org/software/COILS_form.html), keeping the default settings, and run the job [10]. Select any putative coiled coil region and amplify its encoding DNA sequence by PCR using primers carrying the start and stop codons as described above. Clone these sequences in the miP plasmid (*see* Subheading 2) and test them as putative miPs as described in Subheading 3.3.

3.3 Testing miPs

3.3.1 *A. tumefaciens* Culture

1. Inoculate a single colony of each of the five *A. tumefaciens* strains carrying the SP, TF, empty TF, miP or empty miP plasmids in 50 mL tubes containing 5 mL LB solution, 50 µg/mL gentamicin, 25 µg/mL rifampicin and the appropriate antibiotic for the selection of each plasmid (*see* Note 17). Put the culture tubes in a shaker preheated at 28–30 °C and shake them overnight at 200 rpm.
2. Pipet 1 mL of each *A. tumefaciens* overnight culture into 1.5 mL tubes.
3. Centrifuge the tubes at 2200×*g* for 5 min. Discard the supernatant and resuspend the pellet in 1 mL infiltration solution.
4. Repeat **step 4** two more times (*see* Note 18).
5. Transfer the 1 mL culture in a 15 mL tube and add 4 mL infiltration solution.
6. Use an aliquot of the culture to measure the optical density at 600 nm (OD₆₀₀) in a spectrophotometer.
7. Dilute the culture with infiltration solution to reach an OD₆₀₀ of 0.15.
8. Prepare the following combination of cultures in a 15 mL tube and mix them gently:
 - (a) 1 mL SP, 1 mL empty TF and 1 mL empty miP plasmid culture (control sample).
 - (b) 1 mL SP, 1 mL TF and 1 mL empty miP plasmid culture (TF sample).
 - (c) 1 mL SP, 1 mL empty TF and 1 mL miP plasmid culture (miP sample).
 - (d) 1 mL SP, 1 mL TF and 1 mL miP plasmid culture (TF-miP sample).

3.3.2 *N. benthamiana* Leaf Infiltration [11]

1. Aspirate 1 mL of each of the four combination of cultures from above with a 1 mL syringe without the needle.
2. Turn a *N. benthamiana* true leaf (around 10 cm long) in order to expose its abaxial (bottom) side (see **Note 19**) (Fig. 2a). Gently press the syringe against the abaxial side of the leaf in a region between two thick leaf veins to seal the syringe hole (Fig. 2b). Gently exert a counter-pressure with a gloved finger on the other side of the leaf (adaxial/top).
3. Press the nozzle of the syringe and inject the culture slowly into the leaf. The infiltrated area turns dark (Fig. 2c). Stop when the infiltrated area is about 1.5 cm in diameter. Infiltrate the same leaf four times with all four cultures. Infiltrate four leaves and rotate the position of each culture on the leaf every time in order to have each culture infiltrated in each leaf position (see **Note 20**).
4. Mark the borders of the infiltrated area with a permanent marker (see **Note 21**).
5. Grow infiltrated plants for 2–3 days under controlled conditions (24 °C and 40–65% relative humidity) and a long-day photoperiod (14 h light and 10 h dark, with illumination of 130–150 μE).

3.3.3 Quantitative GUS Assay [12]

1. Cut a leaf disk (~0.5 cm in diameter) with a hole puncher inside all infiltrated leaf areas and put them in 1.5 mL tubes on ice.
2. Quickly add 500 μL of cold (4 °C) GUS extraction solution to each tube.
3. Grind the tissue with a plastic pestle inside the tube until homogenized.
4. Centrifuge the samples at $8000\times g$ for 10 min in a refrigerated centrifuge at 4 °C.

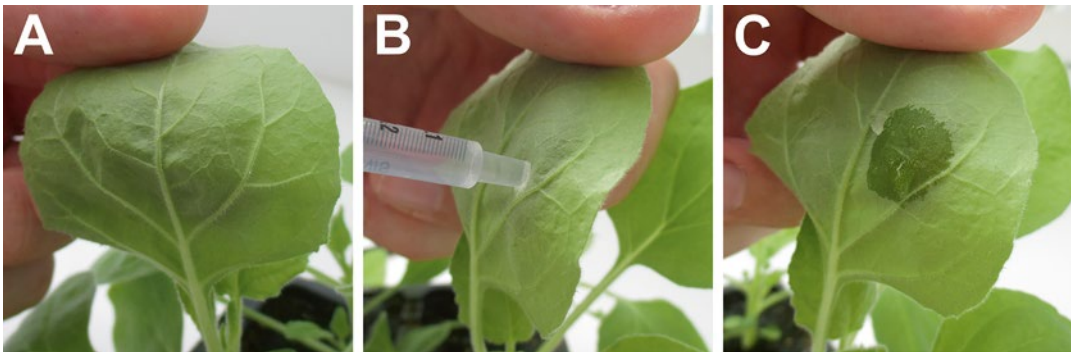


Fig. 2 *N. benthamiana* leaf infiltration. (a) A *N. benthamiana* true leaf turned to expose its abaxial side. (b) A syringe positioned on the abaxial side of the leaf ready for the infiltration. (c) A dark infiltrated area

5. Transfer the supernatants (protein extracts) in new 1.5 mL tubes and keep them on ice (*see Note 22*).
6. Transfer 50 μL of each protein extract into a new 1.5 mL tube and determine the total protein concentration (mg protein/mL) through the Bradford protein assay [13]:
 - (a) Dilute the Bradford reagent fivefolds in deionized water.
 - (b) Add 10 μL of each protein extract to 1 mL of the diluted Bradford reagent and mix. Make triplicate replicates for each protein extract.
 - (c) Prepare a serial dilution series of BSA in GUS extraction solution (0.1–10 mg/mL). Make triplicate replicates for each BSA dilution.
 - (d) Add 10 μL of each BSA dilution to 1 mL of the diluted Bradford reagent and mix.
 - (e) Measure the absorbance of the BSA and protein extract samples in the diluted Bradford reagent at 595 nm in a spectrophotometer.
 - (f) Prepare a standard curve using the BSA serial dilution series.
 - (g) Determine the concentration of the protein extract from the BSA standard curve.
7. For each protein extract prepare two tubes containing 500 μL of GUS assay solution and 450 μL of GUS extraction solution (assay tubes) each. Prepare also two tubes containing 500 μL of GUS assay solution and 500 μL of GUS extraction solution (blank tubes) each. Mix the solutions in the assay and blank tubes and preheat at 37 °C for 2 min.
8. Add 50 μL aliquots of each protein extract to two assay tubes (reaction tubes) and mix them (*see Note 23*).
9. Immediately transfer 100 μL of the solution of each reaction and blank tube into tubes containing 2 mL carbonate stop solution (stopped reaction and stopped blank tubes). Mix and keep the stopped reaction and blank tubes on ice.
10. Put all reaction and blank tubes at 37 °C and keep note of the time.
11. Transfer 100 μL aliquots from each reaction and blank tube after 20, 40, and 60 min of incubation at 37 °C into tubes containing 2 mL carbonate stop solution (stopped reaction and blank tubes). Mix and keep the stopped reaction and blank tubes on ice.
12. Dilute the 4-MU calibration stock solution into the carbonate stop solution to obtain five tubes containing 2 mL of a 10 nM, 20 nM, 40 nM, 60 nM, 80 nM, or 100 nM 4-MU calibration solution. These samples will be the 4-MU standard dilutions.

13. Set a fluorimeter with emission and excitation filters at 455 nm and 365 nm, respectively. Blank the machine by using 2 mL carbonate stop solution. Measure fluorescence intensity (FI) of all stopped reaction and blank tubes and 4-MU standard dilutions.
14. Use the FI data from the 4-MU standard dilutions to draw a calibration curve. Plot FI data versus pmol 4-MU (2 mL of a 10 nM 4-MU standard dilution equals 20 pmol). From the calibration curve calculate FI/pmol 4-MU.
15. Average the FI values of the two repetitions of stopped blank and reaction tubes for each sample and time interval. Subtract averaged FI values of stopped blank tubes from averaged FI values of stopped reaction tubes for each sample and time interval. For each sample, plot the FI data versus time and calculate FI/min.
16. Calculate β -glucuronidase activity of each sample in pmol 4-MU per minute per mg protein according to the following equation:

$$\begin{aligned} & \beta \text{ glucuronidase activity of extract (pmol 4 MU / min/ mg protein)} \\ &= (\text{FI} / \text{min}) / (\text{FI} / \text{pmol 4 MU}) \\ & \times (2.1 \text{ mL reaction volume}) / (0.05 \text{ mL sample volume}) \\ & \times (1 / 0.1 \text{ mL test volume}) \times [1 / (\text{mg total protein} / \text{mL})] \end{aligned}$$

17. Compare the β -glucuronidase activity of the control versus the TF sample to determine the TF transcriptional activity (*see Note 24*). Compare the β -glucuronidase activity of the control versus the miP sample to make sure that the miP does not exert any transcriptional activity (*see Note 25*). Compare the β -glucuronidase activity of the TF versus TF-miP sample to identify the miP effect on the TF transcriptional activity.

4 Notes

1. Acetosyringone is temperature and light-sensitive and should be added to the infiltration solution after autoclaving and immediately before infiltrating.
2. The *A. tumefaciens* strain *GV3101* mediates the transformation of *N. benthamiana* plants with high efficiency. Other *A. tumefaciens* strains can be used but they have to be compatible with *N. benthamiana* transformation.

3. If the TF of interest is a transcriptional activator, the promoter of interest alone should drive no or little GUS expression (in the absence of the TF of interest) in tobacco leaf epidermal cells. By contrast, if the TF of interest is a transcriptional repressor, the promoter of interest alone should drive GUS expression in tobacco leaf epidermal cells. A minimal 35S promoter can be added upstream of the promoter under study to drive expression in leaf epidermal cells.
4. Higher or lower concentrations of 4-MUG might be necessary depending on the strength of expression of the GUS reporter gene. Dilute accordingly.
5. If 4-MUG is not kept at -20°C it can decompose and lead to high background fluorescence readings in blank samples.
6. All software programs must be installed in the same directory. For practical reasons, we advise installing them on the desktop.
7. For clarity, we broke up the lines of the command and pasted them under each other. The command line should be written in one line.
8. When running miP3 on Windows the full path to python.exe has to be given. For example, if Python is installed under “C:/Program Files” as Python-2.7 the command has to be:

```
C:/Program\Files/Python_2.7/python2.7.exe miP3.py  
-p TAIR10_pep_20101214  
-i arabidopsis_transcription_factors.fasta -f Pfam.txt -o miP_  
output.csv  
-b ncbi_blast_2.2.29+/bin/
```
9. The local BLAST tool crashes if there are spaces in the path to any of the input files. To avoid this, make sure that the full path of all input files does not contain any spaces. For example, the input files cannot be located in ~/example folder/miP3/, instead they have to be in ~/example_folder/mip3/.
10. Changing the miP3 parameters requires a full understanding of the miP3 software [3]. For first-time users, we advise using the default parameters. For more information on miP3 see reference [3].
11. To retrieve all the proteins of another plant, go to Phytozome (<http://contacts.jgi-psf.org/registration/new>) and register an account [14]. After logging in, go to <http://phytozome.jgi.doe.gov/pz/portal.html>, click “Download”, go to the newest Phytozome version, and select the species of interest. Finally, select the assembly, download the .fa.gz file, and unpack it where you want to save the .fa file.

12. To retrieve all transcription factors of another plant, go to PlantTFDB (<http://planttfdb.cbi.pku.edu.cn/>), click “Downloads” and select the species of interest to download the FASTA file [15].
13. The option “-b” has to indicate the path to where the NCBI BLAST tools have been installed. In the example “-b ncbi_blast_2.2.29+/bin/”, the BLAST tools are installed in the miP3 software folder. If the BLAST tools are installed elsewhere, change the path accordingly. For example, if the BLAST tools are installed in the Program Files, change the path to “-b C:/Program\ Files/ncbi_blast_2.2.29+/bin/”.
14. Higher e-values allow miP3 to retrieve more candidate miPs but increase the number of false positives. By contrast, lower E-values result in fewer candidate miPs but increase the number of false negatives.
15. miP3 does not rank miPs based on their probability to be true positives.
16. miP and target TF may share a PPI domain that drives the TF homo-dimerization. In this scenario, miPs affect the TF homo-dimerization. Alternatively, miP and target TF might share a PPI domain that drives hetero-dimerization of the TF with a third protein. In this case, the miP affects the interaction of the TF with the third partner. To correctly search for or design miPs, it is of paramount importance to know or test the properties of the PPI domain of the TF of interest. If the TF of interest homo-dimerizes, use the sequence of TF itself to search for or design putative miPs targeting the TF of interest. If the TF of interest hetero-dimerizes, use the sequence of the partner protein to search for or design putative miPs targeting the TF of interest.
17. Gentamicin selects for the *A. tumefaciens* strain *GV3101* virulence plasmid while rifampicin selects for the *A. tumefaciens* strain *GV3101* itself.
18. It is important to wash the *A. tumefaciens* culture from the antibiotics because they might affect the viability of plant cells during the infiltration.
19. Do not use oldest and youngest leaves because transformation and expression efficiency are highly variable.
20. The leaf proximal-distal and medial-lateral polarity might influence the results of the experiment. Therefore, it is important to randomize the effect of the leaf position by testing each sample in each position.
21. The infiltrated area is visible right after infiltration but disappears after few hours.
22. Use the protein extract immediately or store it at $-80\text{ }^{\circ}\text{C}$. Do not store the extract at $-20\text{ }^{\circ}\text{C}$ because enzyme activity is lost at $-20\text{ }^{\circ}\text{C}$.

23. It might be necessary to reduce the amount of the protein extract assayed if it results in fluorescence intensity (FI) higher than the upper limit of the fluorimeter. A calibration procedure can be performed with different quantities of protein extract before conducting the experiment.
24. *A. tumefaciens* infiltration of *N. benthamiana* epidermal cells is a highly robust and efficient method of transformation and, if conducted correctly, does not require accounting for transformation efficiency when comparing independent transformations. Nevertheless, if variability among biological repetitions is too high, add 1 mL of *A. tumefaciens* strain *GV3101* culture hosting a binary vector carrying the gene encoding for the green fluorescent protein (GFP) downstream of the 35S promoter to all culture mixtures. Cut a fragment of the leaf infiltrated area and count the number of cells expressing GFP per unit of surface (an index of the transformation efficiency) using a confocal laser-scanning microscope.
25. If the putative miP exerts transcriptional activity on the promoter of interest in the absence of the TF of interest, it might indicate that the miP affects the activity of a TF expressed in *N. benthamiana* epidermal cells that binds to the promoter of interest. In this case, test the miP in another plant tissue. Alternatively, the putative miP might carry a DB domain that binds to the promoter of interest. To overcome this problem, search for and eliminate putative DBs or test the PPI domain alone as described in Subheading 3.2. In the case in which the DB and PPI domain overlap, perform site-directed mutagenesis to try to inactivate the DB domain.

References

1. Perica T, Marsh JA, Sousa FL, Natan E, Colwell LJ, Ahnert SE, Teichmann SA (2012) The emergence of protein complexes: quaternary structure, dynamics and allostery. Colworth Medal Lecture. *Biochem Soc Trans* 40:475–491
2. Magnani E, de Klein N, Nam HI, Kim JG, Pham K, Fiume E, Mudgett MB, Rhee SY (2014) A comprehensive analysis of microProteins reveals their potentially widespread mechanism of transcriptional regulation. *Plant Physiol* 165:149–159
3. de Klein N, Magnani E, Banf M, Rhee SY (2015) microProtein Prediction Program (miP3): a software for predicting microProteins and their target transcription factors. *Int J Genomics* 2015:ID 734147
4. Cock PJ, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, de Hoon MJ (2009) Biopython: freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics* 25:1422–1423
5. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215:403–410
6. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, Bateman A, Bernard T, Binns D, Bork P, Burge S, de Castro E, Coggill P, Corbett M, Das U, Daugherty L, Duquenne L, Finn RD, Fraser M, Gough J, Haft D, Hulo N, Kahn D, Kelly E, Letunic I, Lonsdale D, Lopez R, Madera M, Maslen J, McAnulla C, McDowall J, McMenamin C, Mi H, Mutowo-Muellenet P, Mulder N, Natale D, Orengo C, Pesseat S, Punta M, Quinn AF, Rivoire C, Sangrador-Vegas A, Selengut JD, Sigrist CJ, Scheremetjew M, Tate J, Thimmajananathan

- M, Thomas PD, Wu CH, Yeats C, Yong SY (2012) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res* 40:D306–D312
7. Goujon M, McWilliam H, Li W, Valentin F, Squizzato S, Paern J, Lopez R (2010) A new bioinformatics analysis tools framework at EMBL-EBI. *Nucleic Acids Res* 38:W695–W699
 8. Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M (2014) Pfam: the protein families database. *Nucleic Acids Res* 42:D222–D230
 9. Liu J, Zheng Q, Deng Y, Cheng CS, Kallenbach NR, Lu M (2006) A seven-helix coiled coil. *Proc Natl Acad Sci U S A* 103(42):15457–15462
 10. Lupas A, Van Dyke M, Stock J (1991) Predicting coiled coils from protein sequences. *Science* 252:1162–1164
 11. Neuhaus JM, Boevink P (2001) *Plant cell biology*. Oxford University Press, Oxford
 12. Jefferson RA, Kavanagh TA, Bevan MW (1987) GUS fusions: beta-glucuronidase as a sensitive and versatile gene fusion marker in higher plants. *EMBO J* 6:3901–3907
 13. Hammond JB, Kruger NJ (1988) The Bradford method for protein quantitation. *Methods Mol Biol* 3:25–32
 14. Goodstein DM, Shu S, Howson R, Neupane R, Hayes RD, Fazo J, Mitros T, Dirks W, Hellsten U, Putnam N, Rokhsar DS (2012) Phytozome: a comparative platform for green plant genomics. *Nucleic Acids Res* 40:D1178–D1186
 15. Jin J, Zhang H, Kong L, Gao G, Luo J (2013) PlantTFDB 3.0: a portal for the functional and evolutionary study of plant transcription factors. *Nucleic Acids Res* 42:D1182–D1187

Simultaneous Analysis of Multiple Promoters: An Application of the PC-GW Binary Vector Series

Jyoti Dalal

Abstract

With the advances in the field of synthetic biology, there is an increasing demand for multi-gene cloning technologies. Molecular cloning to generate multi-gene constructs can be performed by restriction digestion, or by recombination-based cloning strategies such as Gateway[®]. This chapter details cloning, transformation, and selection procedures involved in generation of multi-gene expressing transgenic plants. Methods are described for cloning five distinct promoter-reporter fusion constructs into the PC-GW-BAR vector (from the PC-GW vector series) using Gateway[®] technology and meganuclease sites. Further, transformation and selection methods are described for the biofuel crop *Camelina sativa* from the Brassicaceae family. These methods would be constructive toward generating multi-gene expressing plants for simultaneous expression analysis of five promoters in a short time period.

Key words Multi-gene cloning, Gateway[®], Meganuclease, PC-GW-BAR, *Agrobacterium*, *Camelina*

1 Introduction

Our increased understanding of molecular pathways affecting plant function has enabled development of transgenic plants that take advantage of multiple foreign genes. Multi-gene expressing plants express foreign genes and genetic pathways for various applications ranging from increasing photosynthetic efficiency [1, 2] to improving vitamin E synthesis [3]. Multi-gene expressing constructs can also be used to test the activity and expression of multiple plant promoters simultaneously. Of the various methods available for stable nuclear transformation in plants, T-DNA insertion is among the most widely used.

The lateral DNA transfer from the soil bacterium, *Agrobacterium tumefaciens* to plant nuclear DNA is a fast and efficient method to generate transgenic plants. A binary vector system is often used to shuttle genes between *E. coli* and *Agrobacterium*. Excellent reviews on this DNA integration process have been previously published [4, 5]. In this two plasmid system, the *Agrobacterium* contains a helper T_i

plasmid, that contains the genes for *Agrobacterium* virulence, and a wide-host-range small replicon, that contains the T-DNA. This latter plasmid is the destination vector in the cloning process, and is referred to as the “binary vector”. Transgenes to be cloned into the plant DNA are housed in the T-DNA region of the binary vector. T-DNA as a vector for plant transformation has an upper size limit, as T-DNAs larger than 50 kb are unstable and fall prey to simultaneous deletions [5]. But with less than 50 kb of transgenic material, this technology is reliable and powerful. Quite often, all transgenes in the same T-DNA are integrated into the host DNA at the same locus. This makes it simple to simultaneously identify individuals transformed with or homozygous for all the transgenes.

Generation of multi-gene constructs involves careful prior consideration of gene elements and the overall construct design. Each gene to be cloned contains certain gene elements, selection of which depends on the research application. In general, there is an upstream promoter to guide gene expression, a downstream terminator, and in between a coding sequence with or without introns. Careful determination of all the gene elements in the construct is critical. Cloning methodology depends on the number of genes and their sizes, the choice of promoters and terminators, restriction sites in the sequences and availability of resources. If the genes are cloned from very distant organisms, the need for organism-specific codon optimization may be assessed. Repetitive sequences, such as promoters, terminators, or transit peptide sequences (for targeting proteins to chloroplasts) that may appear in the construct repetitively should be avoided as they may contribute toward homology-based recombination and gene silencing [6].

When up to four genes are to be cloned together, the Gateway® technology is ideal due to its speed and accuracy. Gateway® is a site-specific cloning technology based on the recombination method used by bacteriophage lambda to integrate its DNA in the *E. coli* chromosome [7]. By attaching the *att* sites at the 5' and 3' flanks of the gene, a maximum number of four genes can be simultaneously cloned into a Gateway®-compatible destination vector in an overnight recombination reaction. Because the recombination is based on the *att* sequences, any gene can be readily cloned by Gateway®. For cloning more than four genes into the vector, additional approaches can be used, such as restriction digestion. If the restriction sites from the vector's multiple cloning site are absent in the gene, or if they can be removed from the gene by codon optimization during synthesis or PCR, restriction digestion is a reliable approach to clone genes in the vector. Another variant of this approach is the use of meganuclease sites [8]. Meganucleases have large recognition sites (18–40 nucleotides) and generally occur only once, if at all, in most plant genomes. Since the likelihood of a meganuclease site appearing in a gene of interest is low, cloning can be performed often without any codon optimization.

In this chapter, an application of PC-GW-BAR vectors [9] is described to simultaneously test the activity of five plant promoters *in vivo* using the biofuel crop *Camelina sativa*. The methods require little to no optimization to be applied to *Arabidopsis*, and can be readily adjusted for use in other plant species. The promoter sequences that the researcher wishes to test will be cloned upstream of reporter genes, the signals of which can be independently studied. In this example, four fluorescent reporters are used: mCherry, enhanced green fluorescent protein (EGFP), enhanced yellow fluorescent protein (EYFP), and monomeric cyan (teal) fluorescent protein (mTFPI), and one chemical reporter β -glucuronidase (GUS).

2 Materials

2.1 Design of Constructs

1. Vector NTI^R (Thermo Fisher Scientific) or equivalent software.

2.2 Multisite Gateway[®] and Meganuclease Site Cloning

1. Powder-free nitrile examination gloves.
2. PCR strip tubes 1.5 mL, 2 mL Eppendorf tubes.
3. Pipette tips (VWR[®] Signature Low-Binding Tips).
4. Spatula.
5. Plasmid isolation kit.
6. Autoclaved distilled water.
7. *EcoRI*.
8. *PstI*.
9. *I-CeuI* (New England BioLabs).
10. *ZraI*.
11. *BamHI*.
12. DNA Polymerase I, Large (Klenow) Fragment.
13. Alkaline Phosphatase, Calf Intestinal (CIP).
14. T4 DNA ligase.
15. T4 DNA Ligase Reaction Buffer.
16. 10× CutSmart buffer (NEB).
17. NanoDrop 1000 spectrophotometer (Thermo Fisher Scientific).
18. DNA Clean & Concentrator[™] kit (Zymo Research).
19. Zymoclean[™] Gel DNA Recovery Kit (Zymo Research).
20. MultiSite Gateway[®] Pro Plus kit (Thermo Fisher Scientific)—includes BP Clonase[®] II enzyme mix, LR Clonase[®] II enzyme mix, proteinase K, pDONR[™] vectors, One Shot[®] Mach1[™] T1^R *E. coli* cells, TE buffer and S.O.C. medium.

21. MultiSite Gateway® Pro manual (Thermo Fisher Scientific).
22. One Shot® ccdB Survival™ 2 T1R Competent Cells (Thermo Fisher Scientific).
23. Heatable water bath.
24. LB/kan-Difco™ LB Agar, Miller, 40 g/L with 50 mg/L kanamycin.
25. Autoclave.
26. Round petri dishes.
27. Kanamycin sulfate 50 mg/mL stock.
28. Incubator (37 °C).
29. Laminar flow hood.

2.3 Identification of Recombinant Clones

1. AccuPower PCR premix (Bioneer).
2. Forward Primer (10 pmol/μL), Reverse Primer (10 pmol/μL).
3. PCR strip tubes.
4. PCR thermocycler.
5. Molecular grade agarose.
6. 2-Log DNA ladder (NEB).
7. Glycerol.
8. 15 mL and 50 mL sterile tubes.
9. Gel-Doc EZ system (Bio-Rad).
10. Long-wavelength UV light lamp and mask.
11. Refrigerator (4 °C) and freezers (-20 °C, -80 °C).

2.4 Plant Transformation via Agrobacterium

1. Six-week-old budding *Camelina sativa* plants.
2. Chemically competent *Agrobacterium tumefaciens* strain GV3101.
3. Infiltration medium (MS salts (1/2X), 5% (w/v) sucrose, 1× Gamborg's B5 vitamins, 0.044 μM BAP, 500 μL/L Silwet L-77).
4. LB Broth, Miller (Luria-Bertani).
5. Aluminum foil.
6. Black (trash) bags.
7. Desiccator-Secador® Techni-Dome® 360 (Terra Universal. Inc.).
8. Vacuum pump.
9. Vacuum pressure gauge.

2.5 Transgenic Selection by Herbicide Resistance

1. Finale™ or basta, Bayer Crop Science LP.
2. Phosphinothricin (Gold Biotechnology).
3. Seed Sterilization Solution: 70% ethanol, 10% Bleach (Clorox).

4. Vortex.
5. Centrifuge.
6. ½ MS plates: Murashige and Skoog basal salts with 1 g/L MES, 0.8% agar pH 5.7 adjusted with KOH.
7. Agar, plant cell culture tested.
8. Square Petri dishes.
9. Petri Dish 150 mm × 20 mm.
10. Parafilm.
11. Growth chamber (22 °C, 12 h photoperiod, PAR ~ 400 μmol/m²/s).
12. Soil.
13. Trays for plant growth (at least 3 in. deep, holes at the bottom).

2.6 Confirmation of Gene Integration

1. Plant DNAzol[®] Reagent (Thermo Fisher Scientific).
2. Mini bead-beater.
3. Liquid nitrogen.
4. Additional primers: Forward Primer (10 pmol/μL), Reverse Primer (10 pmol/μL).

2.7 Promoter Analysis by RNA Expression

1. Ceramic mortar and pestle.
2. Liquid nitrogen.
3. Metal spatula.
4. 1.5 mL microcentrifuge tubes.
5. TRIzol reagent (Life Technologies).
6. Chloroform.
7. Shaker.
8. Micro-centrifuge.
9. Centrifuge with rotors compatible with 15–50 mL tubes.
10. Isopropanol.
11. Ethanol (75%).
12. Nuclease-free water.

2.8 Promoter Analysis by Fluorescence

1. Fluorescence dissection microscope fitted with mCherry filter (560 nm excitation, 630 nm emission).
2. Confocal microscope.

3 Methods

3.1 Design of Constructs

3.1.1 Identify the Promoter and Reporter Sequences

Plan the cloning experiment by determining the cloning methodologies, entry and destination vectors, and set up a general plan, as seen in Table 1.

In this example, we clone five promoters, *P1–P5*. Promoters *P1–P4* drive expression of fluorescent marker genes *mCherry* [10], enhanced green fluorescent protein *EGFP* [11], enhanced yellow fluorescent protein *EYFP* [12], and monomeric cyan (teal) fluorescent protein *mTFP1* respectively [12, 13] (see Fig. 1). The constructs made with these four promoters will be cloned by Gateway® to the binary vector PC-GW-BAR. After this cloning, the construct with promoter *P5*, which drives the expression of chemical reporter gene *GUS* [11], will be cloned into the PC-GW-BAR vector by meganuclease cloning (see Fig. 2).

All the fluorescent proteins selected here have distinct excitation and emission spectra (see Table 2) [14]. This would enable clear discrimination between the expressions of the various promoters (see Note 1).

3.1.2 Identify Terminator Sequences

Identify five terminator sequences to place at the 3' end of the reporter coding sequences. In this example, we use the sequences of the following five terminator: *CaMV* 35S terminator (35S) [15],

Table 1
Planning the cloning experiment

Entry vector design							
Promoters for in vivo expression analysis	Reporter gene	Terminator	Cloning methodology	Construct flanking sequences		Entry vector	Destination binary vector
				5' end	3' end		
<i>P1</i>	<i>mCherry</i>	<i>35S</i>	Gateway®	<i>attB1</i>	<i>attB5r</i>	pDONR™ 221 P1-P5r	PC-GW-BAR
<i>P2</i>	<i>EGFP</i>	<i>nos</i>		<i>attB5</i>	<i>attB4</i>	pDONR™ 221 P5-P4	
<i>P3</i>	<i>EYFP</i>	<i>ocs</i>		<i>attB4r</i>	<i>attB3r</i>	pDONR™ 221 P4r-P3r	
<i>P4</i>	<i>mTFP1</i>	<i>hsp</i>		<i>attB3</i>	<i>attB2</i>	pDONR™ 221 P3-P2	
<i>P5</i>	<i>GUS</i>	<i>ubi3</i>	Meganuclease site	<i>I-CeuI</i>	<i>I-CeuI</i>	pUC57	

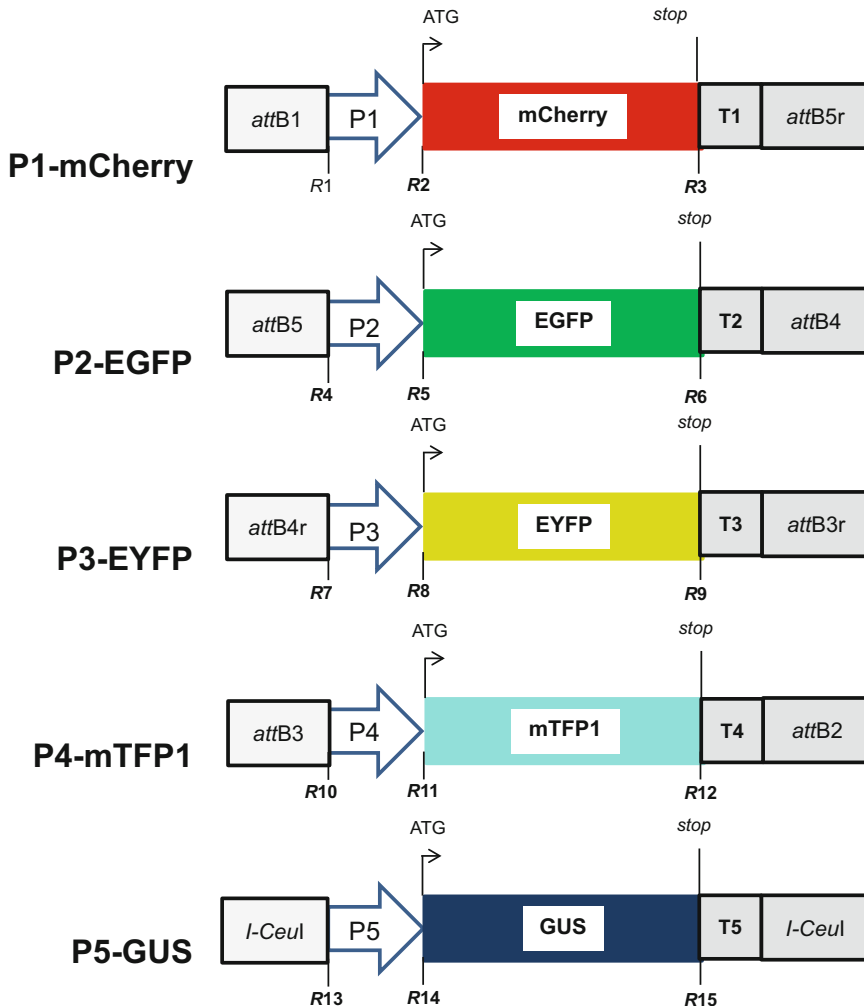


Fig. 1 Entry vector construct design. Five promoters (P1–P5) drive expression of five reporter genes *mCherry*, *EGFP*, *EYFP*, *mTFP*, and *GUS*. Promoters P1–P4 drive expression of fluorescence markers whereas P5 drives expression of chemical reporter gene *GUS*. End of genes are marked by stop codons and transcriptional terminators T1–T5 (see Table 1). Constructs with promoters P1–P4 are flanked with *att* sites for multi-site Gateway® recombination. The construct with promoter P5 is flanked by *I-CeuI* sites for cloning by restriction digestion. All five constructs will be cloned into PC-GW-BAR. In addition, 15 unique restriction sites (R1–R15) were identified which are absent in the PC-GW-BAR vector. The sites were introduced in the entry vectors at the start and end of each promoter, and at the end of each reporter gene coding sequence. The sites were silenced from all other locations in all the constructs. These sites provide modularity to the construct

tobacco *nopaline synthase* gene terminator (*nos*) [16], *octopine synthase* terminator (*ocs*) [17], *heat shock protein 18.2* terminator (*hsp*) [18], and potato *ubiquitin-3* terminator (*ubi3*) [19] (see Table 1).

3.1.3 Assemble the Sequences Using Software

Assemble the sequences of the promoter–reporter constructs using Notepad, MS-Word, or sequence analysis software such as Vector NTI. Further, assemble the sequences as they would appear in

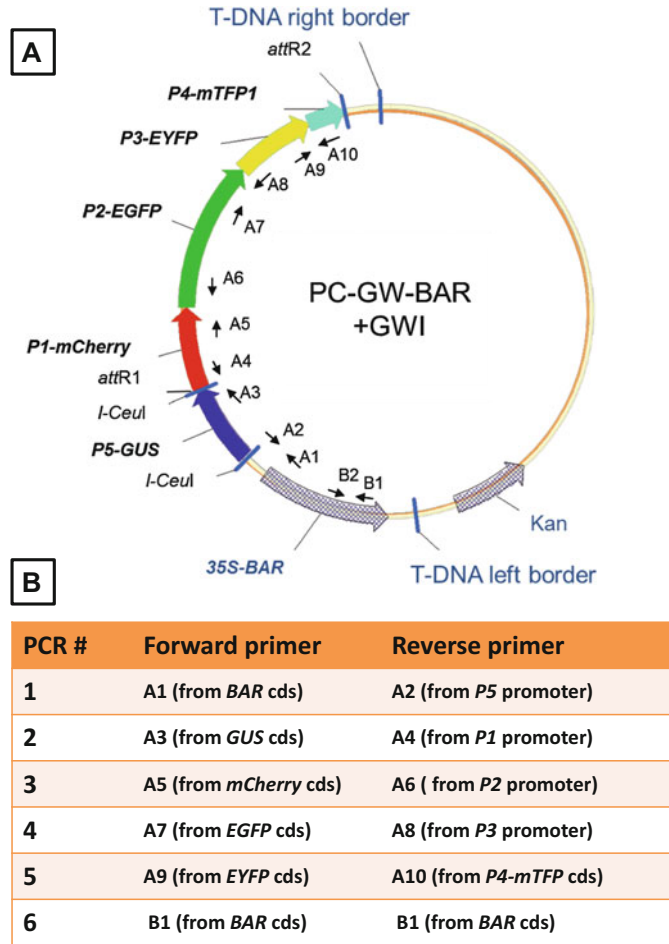


Fig. 2 PC-GW-BAR vector map with all five constructs. (a) In this experiment, five promoters, namely P1–P5 are driving expression of five reporter genes. Using VectorNTI™ software, all the sequences were assembled together to visualize the complete construct at the end of the cloning experiment. The complete construct PC-GW-BAR+GWI has PC-GW-BAR backbone, with four genes in the gateway region and one between the *I-CeuI* sites. (b) Primers were generated to confirm the presence and position of all the gene elements. Primer pair B1 and B2 amplifies the BAR gene

destination binary vector, PC-GW-BAR (see Fig. 2). Such an assembly is intended to help with primer design and planning restriction digestion experiments to verify the construct once it is made.

3.1.4 Design Modularity into the Construct

To make the gene constructs modular, identify fifteen unique restriction sites R1–R15, which are absent in the PC-GW-BAR vector (GenBank accession number KP826773.1). In this example, we will get the constructs synthesized using commercial gene synthesis services. Within the sequence assembly in silico, identify these restriction sites in all five constructs and silence them by

Table 2
Fluorescent reporter genes – excitation and emission spectra

Fluorescent gene	Fluorescence	Maximum excitation wavelength (nm)	Maximum emission wavelength (nm)	Brightness (% EGFP)	GenBank accession number for coding sequence
<i>mCherry</i>	Red	587	610	47	KJ541669
<i>EGFP</i>	Green	488	507	100	EF212308
<i>EYFP</i>	Yellow	514	527	151	EF212303
<i>mTFP1</i>	Teal	462	492	162	FJ530950

(Information from <http://www.microscopyu.com/articles/livecellimaging/fpintro.html>)

using alternative codons during gene synthesis. Place the restriction sites at the start of each promoter, just before the start codon on the reporter genes and just after the stop codon of the reporter genes, as shown in Fig. 1. This step makes it possible to replace any promoter or reporter gene from the final destination vector at any time by a simple restriction digestion experiment. For example, as shown in Fig. 3, we can replace the reporter marker driven by the promoter *P4* to fluorescent protein *mOrange* [13] (see Note 2).

3.1.5 Generate the Promoter–Reporter–Terminator Fusion Constructs Using Gene-Synthesis Services

Generate the five promoter–reporter–terminator fusion constructs with flanking restriction/recombination sites as shown in Table 1. These constructs may be generated by gene synthesis using commercial services. They may also be generated by fusion PCR [20], which is enabled by designing overlapping primers amplifying the promoter, reporter, and terminator sequences. PCR primers will also be used in this case to attach the flanking recombination and restriction sites. In this example, we discuss a scenario where the promoter–reporter fusion constructs are commercially synthesized with the appropriate recombination sites (as given in the MultiSite Gateway® Pro manual) and restriction sites (as given on the website of New England BioLabs Inc.). The genes are synthesized by the company GenScript and cloned by them in between the *EcoRI* and *HindIII* sites of the pUC57 vector (see Note 3) requesting the *EcoRI*, *HindIII*, and R1–R15 sites to be silenced in all other locations in the constructs.

3.2 Multisite Gateway® Cloning

In this section, we clone constructs *P1-mCherry*, *P2-EGFP*, *P3-EYFP*, and *P4-mTFP1* into entry vectors (see Table 1), and then into PC-GW-BAR using the MultiSite Gateway® Pro Plus kit.

3.2.1 Generation of Entry Vectors

1. Obtain plasmids of the four Gateway® constructs from the synthesis service. If the plasmids were transformed into

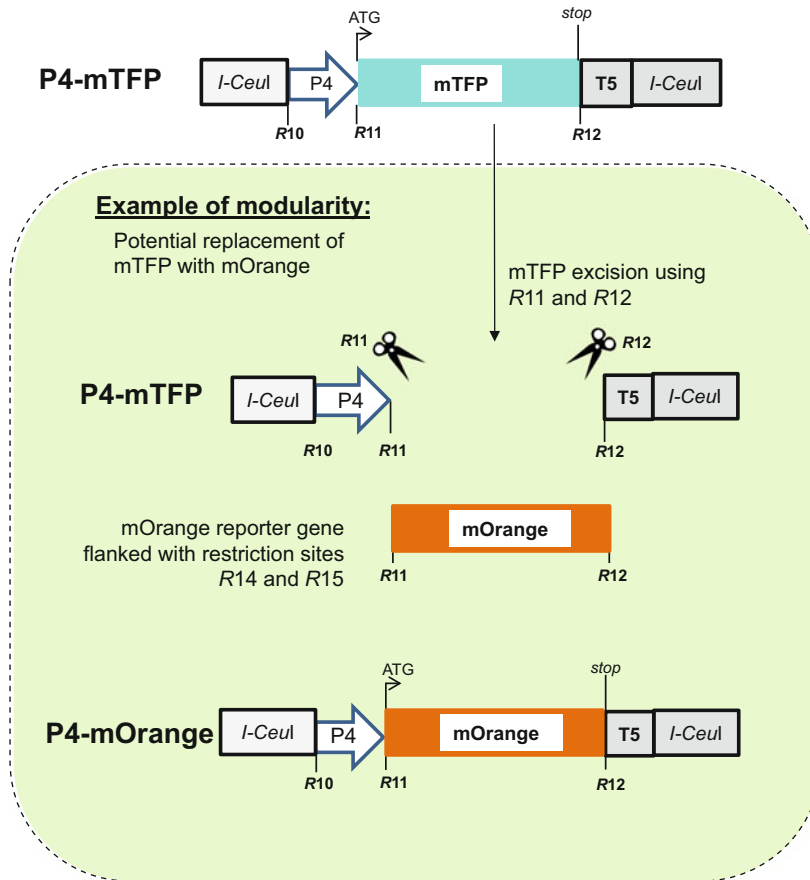


Fig. 3 Using restriction sites to replace gene elements. An example is given where the reporter gene *mTFP* was excised from the P4-mTFP construct using *R11* and *R12* restriction sites. The reporter gene *mOrange* was cloned in between the *R11* and *R12* sites to generate the construct *P4-mOrange*. Using sites *R1*–*R15*, individual gene elements can be replaced at the entry vector stage or when all the genes are already cloned into PC-GW-BAR

E. coli, use 10 mL of overnight bacterial culture for plasmid isolation using a home-made [21] or commercial kit for plasmid isolation.

2. Digest the purified plasmids with *EcoRI* to linearize them. Set up the reaction as shown in Table 3.
3. Incubate the reaction for 2 h at 37 °C.
4. After 2 h, purify the linearized plasmid using a PCR purification kit (*see Note 4*).
5. Check the purity and concentration of the eluted DNA. The linearized plasmid DNA should be of high quality and concentration. Proceed with BP cloning (*see Note 5*), using the positive and negative control reactions recommended by the MultiSite Gateway® Pro Plus manual.

Table 3**Set up of digestion reaction using restriction enzyme *EcoRI***

Reaction component	Volume
10× CutSmart buffer (NEB)	10 μL
<i>EcoRI</i>	2 μL
Plasmid DNA	1–5 μg
Water	up to 88 μL
Total reaction volume	100 μL

6. In separate reactions, add 20–50 fmoles (*see Note 6*) of each of the linearized plasmids with *attB* sites (Table 1). The DNA should be concentrated enough to contain 50 fmoles DNA in 7 μL volume or lesser (*see Note 7*).
7. Add 50 fmoles of the appropriate pDONRTM vectors (1 μL of 150 ng/ μL DNA). The pDONRTM vectors are supplied with the MultiSite Gateway[®] Pro Plus kit.
8. Add TE buffer (pH 8.0) to a final volume of 8 μL .
9. Add 2 μL of BP Clonase[®] II enzyme, and incubate at 25 °C for 1 h.
10. After 1 h, add 1 μL of the Proteinase K solution (provided in MultiSite Gateway[®] Pro Plus kit) to each reaction and incubate the reactions for 10 min at 37 °C.
11. After 10 min, use 2 μL of each reaction to transform chemically competent One Shot[®] Mach1TM T1R *E. coli* cells (Life Technologies). Since the pDONRTM vectors have kanamycin selection in bacteria, plate the transformed cells on plates with LB agar medium containing 50 mg/L kanamycin sulfate (LB/kan). Incubate the plates at 37 °C overnight.
12. The next day, if bacterial colonies are present on the plates including those from a positive control transformation and absent on the plates from a negative control transformation, proceed with identification of recombinant clones.

3.2.2 Identification of Recombinant Clones

1. Pick about ten colonies per cloning reaction for testing. Pick large, well-formed round colonies which do not touch other colonies on the plate. Circle and number the selected colonies at the back of the plate.
2. In a laminar hood, add 15 μL of sterile water in autoclaved PCR strip tubes. Prepare and label one tube per colony.
3. Using 10 μL capacity autoclaved pipette tip, touch the colony in the center. Rinse the tip in the corresponding 15 μL water tube to create bacterial suspension. Discard the used tip.

4. Once all ten colonies are picked, close the strip tubes and put them on ice. Proceed with preparing the PCR reaction.
5. My lab uses the AccuPower PCR premix. Use this or an equivalent PCR system and add about 5–10 pmoles of forward and reverse primer each (*see Note 8*). Add water to a final volume of 13 μL . Prepare one reaction per colony tested.
6. Add 7 μL of the 15 μL of the bacterial suspension into each PCR reaction bringing the total volume to 20 μL .
7. Proceed with PCR. The first denaturation should be long (10 min) to ensure all bacterial cells are lysed. The steps following that can be designed according to the DNA polymerase used, annealing temperature of the primers, and the size of the amplicon. Keep the number of cycles between 25 and 30.
8. Resolve the PCR product on an agarose gel and observe for correct sized bands. The colonies that result in strong positive bands are often recombinant (*see Note 9*).
9. Select two to four colonies with the heaviest bands of the right size on agarose gel. Pipette the remaining 8 μL of the bacterial suspension into 5 mL LB/Kan medium in a 50 mL tube (*see Note 10*). Shake the tube at 250 rpm at 37 °C overnight.
10. The next morning, resuspend 100 μL of culture in 300 μL of 30% glycerol (sterile) and save the glycerol stock at –20 °C or –80 °C for future use. Use the rest for plasmid mini-prep.
11. Repeat the PCR using about 5–10 ng of plasmid DNA. Discard all stocks of colonies negative for PCR.
12. For plasmids testing positive for the gene of interest by PCR, check the sequence by restriction digestion and DNA sequencing. Verified plasmids from the four constructs are the four Gateway® “**entry vectors**”. After the successful BP reaction, the four constructs *P1-mCherry*, *P2-EGFP*, *P3-EYFP*, and *P4-mTFPI* are now flanked by the recombination sites *attL1* and *attR5*, *attL5* and *attL4*, *attR4* and *attR3*, and *attL3* and *attL2* respectively.
13. Save the plasmid for future use. Also, save glycerol stock of one verified colony per construct for future use.

3.2.3 Isolation of PC-GW-BAR Plasmid

1. Isolate plasmid of the binary destination vector PC-GW-BAR (plasmid mini-prep). Since in this example we are testing promoters of the researcher’s choice to drive reporter genes, we do not need the built-in *35S* promoter in the PC-GW-BAR vector. At this time, digest the PC-GW-BAR plasmid with *PstI* and 1 \times CutSmart buffer for 2 h at 37 °C.

The *35S* promoter in the PC-GW-BAR vectors is flanked by *PstI* sites on both ends, so it gets cut out of the vector.

2. To separate the 35S promoter band from the vector DNA, resolve the products of restriction digestion on an agarose gel.
3. Using a long-wavelength UV light lamp and mask, observe the bands resolved on the agarose gel. Using a clean scalpel, cut the bands that are the correct size for PC-GW-BAR vector minus the 35S promoter. The size of PC-GW-BAR is 11,483 bp. After the digestion with *Pst*I, the 35S promoter (780 bp) gets excised from the vector backbone (10,703 bp). Cut the band resolving to 10,703 bp and put the gel slice in a new Eppendorf tube.
4. The gel slices in the tubes can be stored at 4 °C for up to 3 days. When ready, proceed with gel extraction. My lab uses a commercial kit (Zymoclean™ Gel DNA Recovery Kit) to purify the DNA from the gel. The user may use this or an equivalent method to extract DNA from the gel.
5. Once purified, the DNA can be stored at -20 °C or used directly for ligation.
6. Proceed with re-ligation of the vector. Set up the ligation reaction as shown in Table 4.
7. Incubate the reaction at 4 °C overnight.
8. The next morning, use 2–10 µL of the ligation reaction to transform chemically competent One Shot® Mach1™ T1R *E. coli* cells (Life Technologies).
9. Identify clones where 35S promoter region has been excised. This can be done by designing screening primers that flank the vector regions around the 35S promoter.
10. Save the plasmid that has the 35S promoter region excised, and save the corresponding glycerol stock. In this experiment, this is the “**destination vector**”.
11. Plasmid DNA should be high quality and concentrated enough to contain about 20 fmoles of DNA in ≤1 µL volume.

Table 4
Set up of ligation reaction to re-circularize PC-GW-BAR

Reaction component	Volume
10× Ligation buffer	2 µL
T4 DNA Ligase	2 µL
PC-GW-BAR plasmid DNA	100 ng to 1 µg
Water	up to 16 µL
Total reaction volume	20 µL

3.2.4 Multisite Gateway® Recombination Reaction

The method described here is to clone four promoter–reporter constructs that are already in entry vectors simultaneously into Gateway®-compatible binary vectors, using the method described in the MultiSite Gateway® Pro manual. In this case, the Gateway® binary vector being used is PC-GW-BAR [9].

1. Isolate plasmids from the four entry vectors. Use 10 fmoles of each entry vector DNA in the cloning reaction (*see Note 6*).
2. Using a restriction site present in the vector backbone but absent in the target gene, linearize the entry vectors. This step is especially important in the case where the entry vector has the same selection in bacteria as the destination vector (kanamycin^R). For small entry vectors where the selectable marker in entry vector is different than the one in the destination vector, plasmid DNA may be used directly without linearizing. Plasmid DNA or purified linearized DNA should be of high quality and concentrated enough to contain 10 fmoles of DNA in a few μL , and no more than 7 μL for all the entry vectors combined.
3. Add 20 fmoles of PC-GW-BAR plasmid DNA ($\leq 1 \mu\text{L}$). Place the four entry vector plasmids (10 fmoles each) and the binary destination vector PC-GW-BAR plasmid together in one reaction. Use TE buffer (pH 8.0) to bring the volume to 8 μL .
4. Use the LR Clonase® II Plus enzyme according to the manufacturer's directions (*see Note 11*). After removing from $-80 \text{ }^\circ\text{C}$, vortex twice for 2 s. Then add 2 μL to the 8 μL reaction described above.
5. After 16 h of incubation at room temperature, add 1 μL of the Proteinase K solution (provided in MultiSite Gateway® Pro Plus kit) to the reaction and incubate at $37 \text{ }^\circ\text{C}$ for 10 min.
6. After 10 min, use 2 μL of the reactions to transform chemically competent One Shot® Mach1™ T1^R *E. coli* cells. Since the PC-GW vectors have kanamycin selection in bacteria, plate the transformed cells on LB/kan plates. Incubate the plates at $37 \text{ }^\circ\text{C}$ overnight.
7. The next day, if bacterial colonies are seen on the plates, cloning has been successful (*see Notes 12 and 13*). The resulting plasmid has the four Gateway® constructs cloned into the PC-GW-BAR plasmid, and will be designated as “**PC-GW-BAR+GW**”.

3.3 Cloning Using Meganuclease I-CeuI

Using restriction sites to clone genes into a construct is an established approach for molecular cloning. Here, a method is described for cloning a gene into the meganuclease sites of PC-GW-BAR+GW vector (*see Note 14*). PC-GW-BAR has two meganuclease sites on each side of the Gateway® cassette. These sites can be used to clone a single gene between *I-CeuI* and *I-SceI* sites and another gene between *PI-PspI* and *PI-SceI* sites. But a single gene can also be

cloned into each of these restriction sites by having the same site on both ends. The orientation of the ligated product in this case would matter when the cloned gene has its own promoter and terminator. However, when the 5' 35S promoter is intended to be used, orientation of ligation may be controlled by using unique restriction sites on 5' and 3' ends, or validated by PCR using a gene-specific primer and a vector-specific primer. The procedure of cloning a gene into the *I-CeuI* sites of the PC-GW-BAR+GW vector is described below:

1. In this cloning, the P5-GUS construct (“insert”) is cloned into the PC-GW-BAR+GW construct. Obtain the P5-GUS plasmid from the sequencing company. If the plasmids were cloned into *E. coli*, use 10 mL of overnight bacterial culture for plasmid isolation.
2. Digest both the P5-GUS plasmid and PC-GW-BAR+GW plasmid with *I-CeuI* (see Note 15).
3. Prepare 100 μ L reactions for each digestion (see Table 5).
4. Incubate the reaction overnight at 37 °C (see Note 16).
5. The next morning, store the reaction with entry vector in the freezer (–20 °C).
6. To the tube with the destination vector, add 2 μ L alkaline phosphatase (CIP) to dephosphorylate the sticky ends and prevent self-ligation. Incubate at 37 °C for 1 h.
7. After the incubation, the destination vector may be stored in the freezer with the entry vector reaction, or both the reactions could be resolved on an agarose gel (see Note 17).
8. Using a long-wavelength UV light lamp and mask, observe the bands resolved on the agarose gel.
9. Using a clean scalpel, cut the band that is the correct size for the insert. Also, using a fresh scalpel cut the band representing the linearized PC-GW-BAR+GW plasmid. Place each sliced gel in a

Table 5

Set up of digestion reaction using meganuclease *I-CeuI*

Reaction component ^a	Reaction component ^b	Volume
10× CutSmart buffer (NEB)	10× CutSmart buffer (NEB)	10 μ L
<i>I-CeuI</i> (NEB)	<i>I-CeuI</i> (NEB)	2 μ L
P5-GUS plasmid	PC-GW-BAR+GW plasmid	1–5 μ g
Water	Water	up to 88 μ L
Total reaction volume	Total reaction volume	100 μ L

^aFor generating insert

^bFor generating digested PC-GW-BAR+GW vector

- new Eppendorf tube. The gel slices in the tubes can be stored at 4 °C for up to 3 days. When ready, proceed with gel extraction.
10. Once purified, the DNA can be stored at -20 °C or used directly for ligation.
 11. Based on the sizes of the gene of interest cut out from the entry vector and the size of the linearized PC-GW-BAR+GW vector, calculate the amount of gel-purified DNA to use for ligation. The molar ratio of 1:3 for vector: insert is preferred.
 12. Prepare the following ligation reaction (*see* Table 6).
 13. Incubate the reaction at 4 °C overnight.
 14. The next morning, use 2–10 µL of the ligation reaction to transform chemically competent One Shot® Mach1™ T1^R *E. coli* cells.
 15. Screen recombinant clones as described in Subheading 3.1.1 using gene-specific primers for the P5-GUS construct.
 16. Select three independent colonies testing positive for P5-GUS by PCR. Isolate the plasmids from these colonies. At this stage, perform PCR with gene-specific constructs from all the five inserts, as shown in Fig. 2. These primers can be designed using the vector maps assembled in Subheading 3.1.1 (*see* Note 13).
 17. Recombinant clones that test positive for all the transgenes here are the complete construct “**PC-GW-BAR+GWI**” (Fig. 2).

3.4 Plant Transformation via *Agrobacterium*

Multiple methods are available to introduce transgenes into plant cells depending on the goals of the research. Here, methods are described for *Agrobacterium* transformation with plasmid of interest [22], and *Agrobacterium*-mediated stable transformation of camelina plants [23].

Table 6

Set up of ligation reaction to clone *I-Ceul*-flanked insert into PC-GW-BAR+GW

Reaction component	Volume
10× Ligation buffer (NEB)	2 µL
T4 DNA Ligase (NEB)	2 µL
Purified DNA (P5-GUS insert)	100 ng to 1 µg
Purified DNA (digested PC-GW-BAR+GW vector)	100 ng to 1 µg ^a
Water	up to 16 µL
Total reaction volume	20 µL

^aInsert: vector molar ratio should be 1:3

3.4.1 *Agrobacterium* Transformation

The vector PC-GW-BAR+GWI is a binary vector, which can be transformed into *Agrobacterium tumefaciens*, a soil bacterium routinely used to genetically transform plants. Transform the PC-GW-BAR+GWI construct and empty vector (see **Note 14**) separately into *Agrobacterium* strain GV3101.

1. Obtain chemically competent *Agrobacterium* strain GV3101 cells and thaw on ice.
2. Add about 1 μg PC-GW-BAR+GWI plasmid DNA (should be suspended in ≤ 5 μL volume) to a tube containing 50–100 μL chemically competent *Agrobacterium* cells, and mix by tapping.
3. Freeze the tube in liquid nitrogen and then thaw at 37 °C for 5 min.
4. Add 1 mL of LB broth to each tube and transfer the contents to 15 mL sterile tubes. Incubate for 2 h at 30 °C with shaking (lay the tube flat in the shaker).
5. Pour the contents into 1.5 mL Eppendorf tube and centrifuge for 5 min at 2000 $\times g$. Remove supernatant and resuspend pellet in 100 μL of LB broth.
6. Plate 20 μL and 50 μL of suspension on LB/kan plates and incubate for 2 days at 28–30 °C in dark. At the end of 2 days, multiple colonies containing the PC-GW-BAR+GWI construct should emerge. Pick one colony and grow it in 5 mL LB broth. Use 1 μL of *Agrobacterium* culture to test the presence of PC-GW-BAR+GWI by PCR using primers for the plasmids described in the previous section. The *Agrobacterium* cells can be used for PCR directly without prior plasmid isolation, with two important changes in the PCR method (see Table 7). First, the first initial denaturation step should be 10 min instead to ensure complete lysis of the *Agrobacterium* cells. Second, at least 35 amplification cycles should be employed, because of

Table 7

Agrobacterium colony/culture PCR

Step	Temperature	Duration	
Initial denaturation	95 °C	10 min	
Denaturation	95 °C	15–30 s	} 35 cycles
Annealing	50–60 °C	15–30 s	
Extension	72 °C	1 min per 1 kb amplicon	
Final extension	72 °C	5–7 min	

the low copy number of plasmid in *Agrobacterium*. If the PCR results are positive, make glycerol stock of the PC-GW-BAR+GWI transformed *Agrobacterium* colony.

3.4.2 *Agrobacterium*-Mediated Transformation of Camelina Plants by Floral Dip Method

1. To genetically transform camelina plants, use 6-week-old camelina plants that have just bolted. The buds should be visible and separate (see Fig. 4).
2. Two days prior to the transformation, inoculate 3 mL LB/kan medium in a foil-wrapped 50 mL tube (see Note 18), with the glycerol stock of the *Agrobacterium* transformed with the relevant construct (PC-GW-BAR+GWI in this case). Grow the cells while shaking in a shaker incubator set at 28 °C. This is the pre-culture.
3. The next day, inoculate 150 mL of LB medium in a foil-covered sterile flask with 1 mL of the pre-culture. Grow overnight while shaking at 28 °C.
4. The following day, the O.D. (600 nm) of the culture should be about 0.8 indicating optimal *Agrobacterium* growth. Transfer

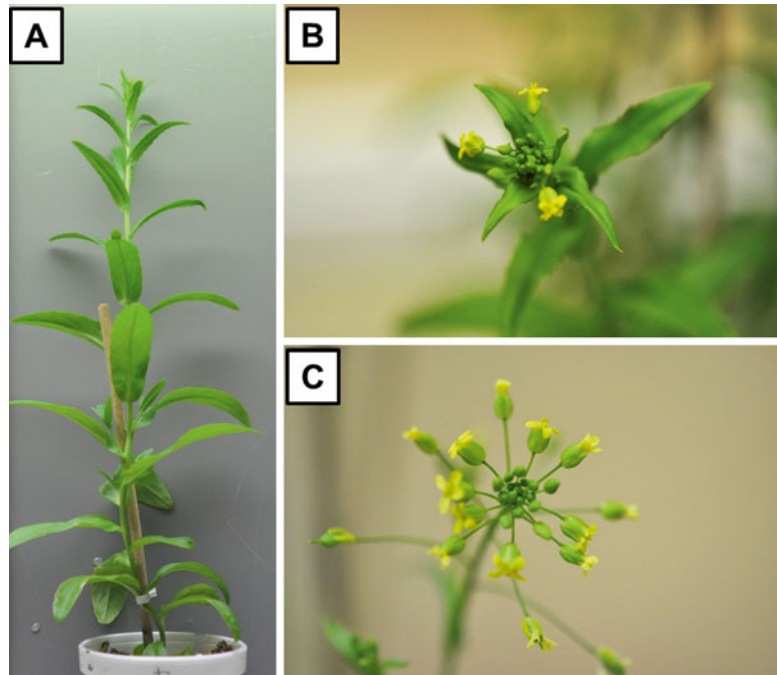


Fig. 4 Identifying the right age of camelina plants for transformation by floral dipping. For maximum rate of transformation by floral dip, camelina plants should be bolting, as seen in (a). There may be a few flowers open but most of the inflorescence should be in bud stage, as seen in (b). Inflorescences where most of the flowers have opened (and most likely self-fertilized), as seen in (c), would give very low transformation rates

the liquid culture to sterile 50 mL tubes and centrifuge for 30 min at $2000\times g$.

5. Resuspend the pellet in 300 mL of infiltration medium.
6. Transfer the suspension to a convenient vessel for placing in a large desiccator (*see Note 19*) fitted with a vacuum pump.
7. Place 6-week-old plants in the desiccator and dip the buds into the *Agrobacterium* transformation solution (*see Fig. 5*). As the plants need to fit in the desiccator, it is best that they be grown in small pots for transformation.
8. Connect the desiccator to a vacuum pump and evacuate at 20 psi for 5 min.
9. Release the vacuum slowly.
10. After infiltration, place the plants in trays with 2 in. of water and cover them with black bags (large trash bags) to maintain dark and high humidity overnight. The following morning, remove the bags (*see Note 20*), wash the plants with water (*see Note 21*) and return them to the growth chamber or greenhouse.
11. Allow the flowers to set seed and the seed to mature (another 6–8 weeks in camelina). Each week after transformation, clip out new branches to avoid seed development from untransformed buds formed after the transformation event.



Fig. 5 Camelina floral dip transformation apparatus. A large desiccator is plugged in with a vacuum pump and a pressure gauge. The infiltration solution (with *Agrobacterium*) is placed in a beaker inside the chamber. Young inflorescences are dipped inside this solution and the chamber is evacuated until the pressure of 20 psi is reached. The pressure is maintained for 5 min, after which the vacuum is slowly released

12. Harvest the seed and proceed with selection of transgenic plants. Using this method, generally a transformation rate of 1–6% is observed.

The method of selection of transgenic plants will depend on the selectable marker on the destination vector. Various tags, such as resistance to antibiotics (e.g. kanamycin^R, hygromycin^R), herbicide (e.g. phosphinothricin), and fluorescence can be used to identify transgenics among populations.

3.5 Transgenic Selection by Herbicide Resistance

Transgenic plants with PC-GW-BAR can be selected using resistance to the herbicide phosphinothricin (ppt). Selection may be done on plates or in soil. Here, both methods are described.

3.5.1 On Soil Selection

To select transformed seed based on ppt resistance in soil, proceed with the following steps.

1. Fill soil in a tray (at least 3 in. deep) with holes at the bottom for water uptake. Irrigate the soil and let unabsorbed water drip through the holes.
2. Spread seed from floral dipped plants (**T0 plants**) into the soil such that each seed is allotted at least 5 mm² surface area on the tray.
3. Place the trays in a growth chamber or greenhouse with desired growth conditions.
4. Cover the trays on top with saran wrap for 2 days to maintain humidity and to aid in seed germination.
5. After 2 days remove the saran wrap. Seedlings should be seen emerged at this time. Let seedlings grow until they are 8 days old.
6. In a spray bottle, mix herbicide (FinaleTM or equivalent) into distilled water to a final concentration of 0.045% ppt. Pour the diluted herbicide in a spray bottle.
7. Evenly spray with the plants FinaleTM once. After 1 day, spray the plants with FinaleTM once more. Do not spray the same plant twice on the same day.
8. Let plants recover. After a week, only resistant and therefore transgenic plants should be seen growing (*see* Fig. 6).
9. Replant the transgenic plants into new pots. In a few weeks, collect tissue for confirmation of transgene integration and expression.

3.5.2 On Plate Selection

To select transformed seed based on ppt resistance on plates with MS medium, proceed with the following steps.

1. Prepare plates with ½ MS medium and 0.8% agar and 5–15 mg/L phosphinothricin and pour into sterile petri plates. Also prepare medium with ½ MS medium, 0.35% agar, and

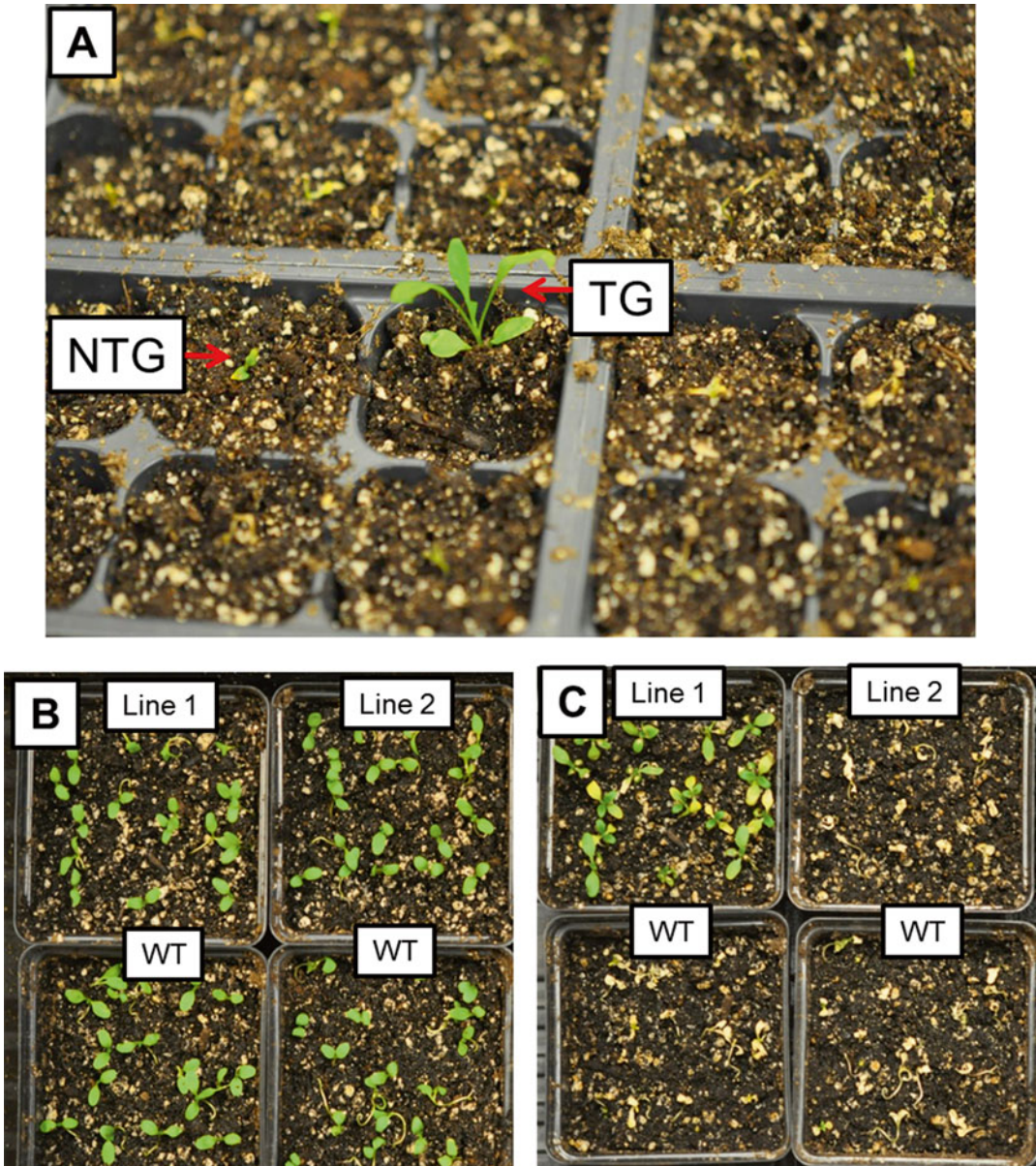


Fig. 6 On-soil phosphinothricin (ppt) selection using Finale™. Seeds harvested from T_0 plants were germinated on soil. Eight-day-old T_1 seedlings were sprayed with Finale™ (0.045 % ppt) to select transgenics. (a) Seedlings from transformed camelina (T_0) were sprayed with Finale™. One plant was identified as transgenic (TG) while the others were non-transgenic (NTG). (b) Two lines of T_2 generation potential transgenic plants were grown on soil along with wild-type plants (WT). At 8 days old, plants were sprayed with Finale™. (c) After 2 days, non-transgenic plants (line 2 and WT) wilted while transgenic plants (line 1) remained green and viable

5–15 mg/L phosphinothricin. Keep this medium in a warm water bath (55 °C) until use. Prepare sterile water to wash seed during surface sterilization. Proceed with surface sterilization.

2. Take about 500 seed in 50 mL sterile tube.

3. Add 30 mL 70% ethanol and vortex the seed for 30 s. Allow the seed to settle and decant the ethanol, along with the seed floating on the surface. The total time in ethanol should not exceed 2 min to maintain seed viability.
4. Add 30 mL 10% bleach, and shake vigorously. Immediately spin the tube in a centrifuge and spin at $2000 \times g$ for a few seconds. Open the tube in a laminar flow hood and decant the bleach solution along with any seed floating on the surface. The total time in bleach solution should not exceed 10 min to maintain seed viability.
5. Add 30 mL sterile water to the tube and vigorously shake the tube to wash off the bleach and remaining ethanol. Centrifuge the tube for a few seconds at $2000 \times g$, take the tube back into the laminar flow hood and decant the water. Repeat the washing steps four times. After the final wash, add 25 mL of MS medium with 0.35% agar prepared in **step 1** into the tube. Proceed with seed plating (*see Note 22*). Mix in seed by inverting the tube a few times and pour all of the solution on a single large petri dish (150 mm \times 20 mm) such that the seed are evenly spaced.
6. Plate the petri dishes into a growth chamber with 12 h-long day period.
7. The next day, seedlings start emerging. In 4 days, transgenic seedlings are seen to grow while non-transgenic seedlings are seen as yellow and wilting (*see Fig. 7*).

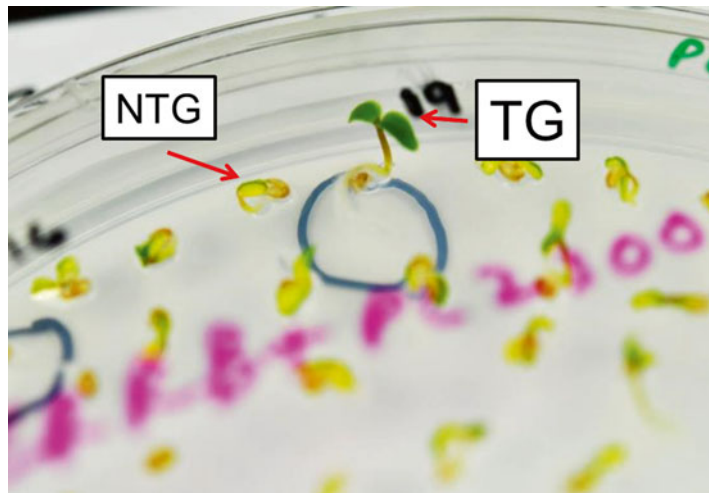


Fig. 7 On-plate phosphinothricin (ppt) selection. Seeds harvested from T_0 plants were surface-sterilized. Then they were resuspended in 0.35% agar and poured on $\frac{1}{2}$ MS plates (0.8% agar, 15 mg/L ppt). The seed spread evenly on the plate. After 4 days, transgenic seedlings (TG) could be identified over non-transgenic seedlings (NTG)

8. Plant the transgenic seedlings into new pots. In a few weeks, collect tissue for confirmation of transgene integration and expression.

3.6 Confirmation of Transgenic Plants by PCR

While the PC-GW-BAR+GWI T-DNA integration can be indirectly tested by seedlings surviving ppt selection, it is critical to test the integration of the entire T-DNA in transgenic plants. This helps in identification of wild-type plants that may have escaped selection or plants that got transformed with truncated T-DNA.

The following steps describe testing the DNA of transgenic plants by PCR. However, the user can use alternate methods of tissue collection and DNA isolation.

1. Grow plants surviving phosphinothricin selection in individual pots in soil. These are the T₁ plants. Label each plant with an identification number, such as a barcode. Since the insertion of T-DNA in their genomes is distinct, they are notated as distinct transgenic “lines” for this construct.
2. Generate at least ten transgenic lines. Repeat transformation if necessary.
3. Once the T₁ plants have adapted to soil and have grown at least two secondary leaves, the analysis can be performed.
4. At this time, prepare one 1.5 mL Eppendorf tube per plant by placing two sterile solid glass beads in each sterile Eppendorf tube. Label the tubes with the identification of one T₁ plant each.
5. Wear nitrile lab gloves (or equivalent). Open the mouth of the tube near the corresponding plant, insert about 5 mm² leaf material into the mouth of the tube, and snap the lid close. Let the lid excise the leaf at the mouth of the tube. Place the tube in liquid nitrogen immediately.
6. Using a mini bead beater, homogenize the tissue. Place the tube in liquid nitrogen again.
7. When ≤ 24 samples have been homogenized, take out one tube at a time from liquid nitrogen and add 300 μ L of Plant DNAzol[®] Reagent. Following the manufacturer’s protocol, precipitate, wash, and solubilize the DNA in water.
8. Using isolated DNA from each transgenic line, conduct PCR using gene-specific primers for *PI* through *P5* and their reporter genes, as shown in Fig. 2. These primers can be designed using the vector maps assembled in Subheading 3.1.1. This is to ensure that the T-DNA did not get truncated. Note that only primer pair B1-B2 would work for the empty vector-transformed plants. Use an endogenous gene primer pair as a positive control. Use a primer pair from PC-GW-BAR outside of the T-DNA region as a negative control. Use 10–50 ng DNA for each PCR.

9. Observe the sizes of the bands to identify lines testing positive for all the genes.
10. For lines testing positive for the genes of interest, collect a fully expanded leaf when plants have at least 15 fully expanded leaves. Store the leaf in $-80\text{ }^{\circ}\text{C}$ for future experiments, such as Southern Hybridization.
11. For positive lines, collect seed and propagate them in the T_2 generation. Repeat selection on soil or plate as described in Subheading 3.5. Observe the segregation ratio. For a single locus transgenic, in the T_2 generation the ratio of selected vs. non-selected seedlings should be approximately 3:1.
12. The T_2 plants can be readily used for visualizing the expression of various promoters.

3.7 Promoter Expression Analysis

The reader may use any method of choice to analyze the promoter activities. Two methods are described here in brief, RNA expression analysis and visual analysis by fluorescence or GUS activity.

3.7.1 Reporter Gene RNA-Expression Analysis

For each plant tested, collect a variety of tissues, as shown in Table 8. Isolate RNA and synthesize cDNA from each stage [24]. Using semi-quantitative PCR, test the cDNA for expression of all the reporter genes [25], as shown in Table 8. Make such a table for every line tested.

The RNA expression may be tabulated as 0 (for absent or undetectable) and 1 (for detectable). In this hypothetical example,

Table 8
Sample table (hypothetical) to record RNA expression in transgenic and control plant lines

Plant tissue	cDNA PCR positive (?)				
	mCherry	EGFP	EYFP	mTFP1	GUS
Seedling (whole)	0	1	1	1	0
Seedling (hypocotyl)	0	1	1	0	0
Seedling (root)	0	1	1	1	0
Young leaf	0	0	1	0	0
Fully expanded leaf	0	0	1	0	1
Stem	0	0	1	0	1
Root (mature)	0	0	1	0	0
Flowers	0	0	0	0	0
Young seed pod	1	0	0	0	0
Mature seed	1	0	0	0	0

mCherry expression is only observed in young seed pods and mature seed, indicating that *P1* may be a seed-specific promoter. EGFP expression is only seen in roots and hypocotyls of young seedlings, and whole seedlings indicating that *P2* promoter is active only at the seedling stage. EYFP transcript is expressed in all tissues except flowers and young seed pods and mature seed, indicating that *P3* activity may be ubiquitous in vegetative tissues. The expression of mTFP1 is only observed in young seedlings (whole) and seedling roots. This indicates that like *P2*, *P4* is also active only at seedling stage. However, unlike *P2*, *P4* is root-specific. The GUS transcript is only observed in mature leaves and stems, indicating the expression domain of *P5* activity.

3.7.2 Reporter Gene Visual Analysis

While RT-PCR can give us an indication of the tissues in which the promoters are active, these data do not provide information about the cell type specificity of the promoter. Fluorescence microscopy and GUS staining can be used to identify the cellular domains of promoter expression. In this example, we demonstrate the visualization of *P1* activity in the mature seed. We use a dissection microscope fitted with mCherry filter (560 nm excitation, 630 nm emission using ET-mCherry filter (Nikon or equivalent)). The transgenic seed with mCherry fluorescence can be readily identified against non-transgenic seeds (*see* Fig. 8).

The *P1* expression is observed on the entire seed coat of the transgenic seed. Cross-section analysis of transgenic seed would reveal domains of *P1* expression within the seed.

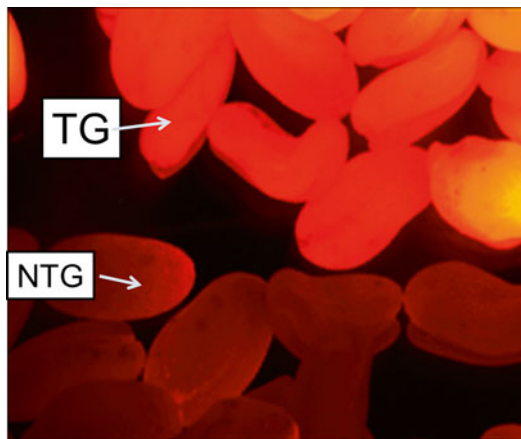


Fig. 8 *P1*-mCherry expression in seed. The expression of mCherry fluorescent protein was observed exclusively in the seed of the transgenic (TG) plants. This indicates that promoter *P1* may have a seed-specific expression domain

4 Notes

1. The fluorescent marker EYFP emits yellow fluorescence when excited with a blue or violet laser. When the fluorescent markers EGFP and EYFP are used in the same plant, they are both excited with 488 nm blue laser, and their emission ranges overlap when using 525/50-nm band pass filter. However, their fluorescence can be distinguished by using a combination of two bandpass filters – 510/20-nm to detect EGFP and 550/30-nm to detect EYFP with a short pass dichroic 525-nm mirror between them. Alternatively, a 405-nm violet laser with a single 500–550 nm range bandpass filter may be used, as it excites EGFP but not EYFP, thereby distinguishing their expression domains [26].
2. If the replacement promoter or reporter gene has the restriction sites needed to place it in the construct, the sites may be silenced by PCR. Primers can be designed to introduce site-specific changes into the sequence. Alternatively, the sequence can be synthesized.
3. Once the plasmids are received from the synthesis company, they should be cloned into a suitable bacterial strain, such as DH5 α and glycerol stocks should be maintained to provide plasmids as needed for the research. The plasmids isolated from the glycerol stocks should be tested by PCR using gene-specific primers and sequenced for validating the sequence prior to any downstream cloning procedures.
4. During the final step of DNA purification, elute the DNA in water instead of TE buffer in order to avoid excess salt in the eluted DNA. Excess salt may in some cases inhibit recombination reaction and/or the subsequent cloning—also *see* **Notes 7** and **15**.
5. The MultiSite Gateway[®] Pro manual is a great source for preparing BP and LR recombination reactions.
6. Using the size of the target gene in base pairs, the ng of DNA to give 20–50 fmoles per reaction may be calculated using methods described in the MultiSite Gateway[®] Pro manual, or online utilities such as (http://www.molbiol.edu.ru/eng/scripts/01_07.html).
7. When DNA is resuspended in water, the fmoles/ μ L concentration can be increased simply by evaporating the water. However, when DNA is eluted in TE, this method of concentration would lead to increased salt content in the DNA. In that case, DNA can be re-precipitated from the solution using 100% ethanol and re-dissolved in lesser volume of water.
8. The primers should amplify about 300–1000 bp of the transgene. The primer sequences should be specific to the transgene.

being tested. In case the same restriction site is being used to insert the DNA into the vector on both 5' and 3' ends, orientation of insertion can be tested using one vector-specific primer and one insert-specific primer.

9. Colony PCR can yield false-positive results, if the tip touches the plate during picking up the colony. The LB plates have some amount of entry vector DNA on them, which is often enough to test positive by PCR. However, there would be a difference in the band intensities of false positives and real positive colonies. In any case, colonies testing positive for the transgene after colony PCR must be tested further at plasmid level.
10. Bacterial growth is optimal when there is good aeration of the medium during the incubation. Therefore, shaking while incubating in bigger tubes is recommended.
11. In my lab's experience, commercial Gateway® enzymes such as LR Clonase® do not last past their warranty, usually 6 months. Within this time period they are very efficient. Therefore, it is advisable to have all the entry vectors prepared and sequence-verified before the LR Clonase® enzyme is purchased. If multiple cloning experiments are planned, prepare all the individual entry vectors before purchasing the enzyme.
12. Many Gateway®-compatible vectors, including the PC-GW vectors, contain *ccdB* gene in between the attR sites. Bacteria containing *ccdB* gene can only propagate in *ccdB*-safe cells, such as “One Shot® *ccdB* Survival™ 2 T1R” Competent Cells. When the cloning reaction is propagated in other bacterial strains, such as the “One Shot® Mach1™ T1R” *E. coli* cells, the non-recombined destination vector with kanamycin selection cannot propagate. At the same time, the entry vectors have been linearized by restriction enzymes prior to cloning, so they are unable to propagate in LB medium containing kanamycin. Therefore, any colonies formed on the LB/kan medium have a strong chance of being recombinational clones.
13. Primers may also be used at this stage to verify the gene sequence of the vector by sequencing. Further, restriction digestion with multiple enzymes can be conducted to confirm the order of the genes in the construct. Restriction enzymes and the corresponding sizes of bands resulting from the digested plasmid may be obtained from VectorNTI™ as well.
14. The Gateway® region in un-recombined PC-GW vectors encodes *ccdB* gene. Therefore, if the researcher does not need Gateway®, it should be excised using the *ZraI* and *BamHI* sites, blunt ended using DNA Polymerase I, Large (Klenow) Fragment, and ligated back into a circular molecule before using it exclusively for restriction-based cloning. This construct can also serve as an “empty vector” for *Agrobacterium* transformation.

15. In my lab's experience, the enzyme *I-CeuI* is very salt-sensitive. We saw little digestion in overnight restriction reactions where there was leftover salt in plasmid DNA preparations. However, the issue was resolved by cleaning the DNA of excess salt by re-precipitation or on column-clean up.
16. When the reaction is incubated overnight at 37 °C, wrap the reaction tube with parafilm to avoid evaporation due to loosening of the tube lid at 37 °C overnight.
17. The amount of agarose in the agarose gel electrophoresis determines the resolution of bands. In general, 1% gel is ideal to resolve bands between 100 and 3000 bp. If smaller sized bands (<500 bp) or larger sized bands (>2000 bp) are to be resolved more clearly, the concentration of agarose in the gel may be modified to a higher or lower level respectively. The time of gel electrophoresis as well as the voltage depends on the sizes of the bands expected to be resolved.
18. *Agrobacterium* cells grow best in dark. Therefore, unless the shaker incubator is in a dark room, it is advisable to cover the tubes or flasks used to grow them in liquid cultures with aluminum foil.
19. While floral dipping can be performed using a small desiccator fitted with a vacuum sealing mechanism, it is much easier to do using a large desiccator, especially when the plants are large such as camelina.
20. It is important for optimum growth and infection of *Agrobacterium* to incubate the infiltrated floral buds in dark and high humidity overnight up to 18 h. However, due to the stress of transformation and dark incubation, the plants must be returned to light the following day. In our lab setting, leaving plants in dark for 2 days causes most infiltrated buds to die.
21. Washing of the plants should be done in an enclosed tank. After washing, the water in the tank should be sterilized using bleach. All bags, paper towels, etc. should be autoclaved to block *Agrobacterium* escape and contamination of soil/water resources.
22. Once sterilized, camelina seed should be plated immediately. Plating the seed a day later significantly hampers the growth of the seedlings.

Acknowledgements

This work was conducted at departments of Crop Science and Plant & Microbial Biology at North Carolina State University, and was supported by DOE ARPA-E grant DE-AR0000207. The author thanks Dr. Ron Qu and Dr. Roopa Yalamanchili for review of the manuscript.

References

1. Kebeish R, Niessen M, Thiruveedhi K, Bari R, Hirsch HJ, Rosenkranz R, Stabler N, Schonfeld B, Kreuzaler F, Peterhansel C (2007) Chloroplastic photorespiratory bypass increases photosynthesis and biomass production in *Arabidopsis thaliana*. *Nat Biotechnol* 25(5):593–599
2. Dalal J, Lopez H, Vasani N, Hu Z, Swift J, Yalamanchili R, Dvora M, Lin X, Xie D, Qu R, Sederoff H (2015) A photorespiratory bypass increases plant growth and seed yield in bio-fuel crop *Camelina sativa*. *Biotechnol Biofuels* 8:175. doi:10.1186/s13068-015-0357-1
3. Raclaru M, Gruber J, Kumar R, Sadre R, Luhs W, Zarhloul MK, Friedt W, Frentzen M, Weier D (2006) Increase of the tocopherol content in transgenic *Brassica napus* seeds by overexpression of key enzymes involved in prenylquinone biosynthesis. *Mol Breeding* 18(2):93–107. doi:10.1007/s11032-006-9014-5
4. Lee LY, Gelvin SB (2008) T-DNA binary vectors and systems. *Plant Physiol* 146(2):325–332. doi:10.1104/pp.107.113001
5. Roy SC (2015) Gene transfer in higher plants for the development of genetically modified crops (GM crops). *Int J Curr Adv Res* 4(6):132–148
6. Luff B, Pawlowski L, Bender J (1999) An inverted repeat triggers cytosine methylation of identical sequences in *Arabidopsis*. *Mol Cell* 3(4):505–511
7. Hartley JL, Temple GF, Brasch MA (2000) DNA cloning using in vitro site-specific recombination. *Genome Res* 10(11):1788–1795. doi:10.1101/Gr.143000
8. Epinat JC, Arnould S, Chames P, Rochaix P, Desfontaines D, Puzin C, Patin A, Zanghellini A, Paques F, Lacroix E (2003) A novel engineered meganuclease induces homologous recombination in yeast and mammalian cells. *Nucleic Acids Res* 31(11):2952–2962
9. Dalal J, Yalamanchili R, La Hovary C, Ji M, Rodriguez-Welsh M, Aslett D, Ganapathy S, Grunden A, Sederoff H, Qu R (2015) A novel gateway-compatible binary vector series (PC-GW) for flexible cloning of multiple genes for genetic transformation of plants. *Plasmid* 81:55–62. doi:10.1016/j.plasmid.2015.06.003
10. Shaner NC, Campbell RE, Steinbach PA, Giepmans BN, Palmer AE, Tsien RY (2004) Improved monomeric red, orange and yellow fluorescent proteins derived from *Discosoma* sp. red fluorescent protein. *Nat Biotechnol* 22(12):1567–1572. doi:10.1038/nbt1037
11. Kong Y, Zhu Y, Gao C, She W, Lin W, Chen Y, Han N, Bian H, Zhu M, Wang J (2013) Tissue-specific expression of SMALL AUXIN UP RNA41 differentially regulates cell expansion and root meristem patterning in *Arabidopsis*. *Plant Cell Physiol* 54(4):609–621. doi:10.1093/pcp/pct028
12. Kothke S, Kock M (2011) The *Solanum lycopersicum* RNaseLER is a class II enzyme of the RNase T2 family and shows preferential expression in guard cells. *J Plant Physiol* 168(8):840–847. doi:10.1016/j.jplph.2010.11.012
13. Geldner N, Denervaud-Tendon V, Hyman DL, Mayer U, Stierhof YD, Chory J (2009) Rapid, combinatorial analysis of membrane compartments in intact plants with a multicolor marker set. *Plant J* 59(1):169–178. doi:10.1111/j.1365-313X.2009.03851.x
14. Mylle E, Codreanu MC, Boruc J, Russinova E (2013) Emission spectra profiling of fluorescent proteins in living plant cells. *Plant Methods* 9(1):10. doi:10.1186/1746-4811-9-10
15. Pietrzak M, Shillito RD, Hohn T, Potrykus I (1986) Expression in plants of two bacterial antibiotic resistance genes after protoplast transformation with a new plant expression vector. *Nucleic Acids Res* 14(14):5857–5868
16. Christensen AH, Quail PH (1996) Ubiquitin promoter-based vectors for high-level expression of selectable and/or screenable marker genes in monocotyledonous plants. *Transgenic Res* 5(3):213–218
17. De Greve H, Dhaese P, Seurinck J, Lemmers M, Van Montagu M, Schell J (1982) Nucleotide sequence and transcript map of the *Agrobacterium tumefaciens* Ti plasmid-encoded octopine synthase gene. *J Mol Appl Genet* 1(6):499–511
18. Nagaya S, Kawamura K, Shinmyo A, Kato K (2010) The HSP terminator of *Arabidopsis thaliana* increases gene expression in plant cells. *Plant Cell Physiol* 51(2):328–332. doi:10.1093/pcp/pcp188
19. Belknap W, Rockhold D, McCue K (2008) pBINPLUS/ARS: an improved plant transformation vector based on pBINPLUS. *Biotechniques* 44(6):753–756. doi:10.2144/000112731
20. Yon J, Fried M (1989) Precise gene fusion by PCR. *Nucleic Acids Res* 17(12):4895
21. Li JF, Li L, Sheen J (2010) Protocol: a rapid and economical procedure for purification of plasmid or plant DNA with diverse applications in plant biology. *Plant Methods* 6(1):1. doi:10.1186/1746-4811-6-1

22. Wise AA, Liu Z, Binns AN (2006) Three methods for the introduction of foreign DNA into *Agrobacterium*. *Methods Mol Biol* 343:43–53. doi:[10.1385/1-59745-130-4:43](https://doi.org/10.1385/1-59745-130-4:43)
23. Lu C, Kang J (2008) Generation of transgenic plants of a potential oilseed crop *Camelina sativa* by *Agrobacterium*-mediated transformation. *Plant Cell Rep* 27(2):273–278. doi:[10.1007/s00299-007-0454-0](https://doi.org/10.1007/s00299-007-0454-0)
24. Dalal J, Land E, Vasani N, He L, Smith C, Rodriguez-Welsh M, Perera IY, Sederoff H (2015) Methods for RNA profiling of gravi-responding plant tissues. *Methods Mol Biol* 1309:91–117. doi:[10.1007/978-1-4939-2697-8_9](https://doi.org/10.1007/978-1-4939-2697-8_9)
25. Khurana N, Chauhan H, Khurana P (2013) Wheat chloroplast targeted sHSP26 promoter confers heat and abiotic stress inducible expression in transgenic *Arabidopsis* plants. *PLoS One* 8(1), e54418. doi:[10.1371/journal.pone.0054418](https://doi.org/10.1371/journal.pone.0054418)
26. Marcus A, Raullet DH (2013) A simple and effective method for differentiating GFP and YFP by flow cytometry using the violet laser. *Cytometry A* 83(11):973–974. doi:[10.1002/cyto.a.22347](https://doi.org/10.1002/cyto.a.22347)

Chapter 14

GenoCAD Plant Grammar to Design Plant Expression Vectors for Promoter Analysis

Anna Coll, Mandy L. Wilson, Kristina Gruden, and Jean Peccoud

Abstract

With the rapid advances in prediction tools for discovery of new promoters and their *cis*-elements, there is a need to improve plant expression methodologies in order to facilitate a high-throughput functional validation of these promoters in planta. The promoter-reporter analysis is an indispensable approach for characterization of plant promoters. It requires the design of complex plant expression vectors, which can be challenging. Here, we describe the use of a plant grammar implemented in GenoCAD that will allow the users to quickly design constructs for promoter analysis experiments but also for other in planta functional studies. The GenoCAD plant grammar includes a library of plant biological parts organized in structural categories to facilitate their use and management and a set of rules that guides the process of assembling these biological parts into large constructs.

Key words Synthetic biology, GenoCAD, Plant grammar, Plant expression vectors, Plant promoters

1 Introduction

The study of plant transcriptional regulation is essential not only in basic research, to understand the function of genes and their control, but also in applied research. Since promoters are important tools in plant genetic engineering, their identification and characterization is crucial in order to supply more diversity and for finer regulation of gene expression at the transcriptional level.

Extensive efforts have been directed to the discovery of new promoters and their *cis*-elements. The availability of whole plant genome sequences and a huge collection of plant transcriptomic data have allowed large-scale prediction analysis of promoters and their regulatory elements. For example, the study of gene expression patterns under different biotic and abiotic stress led to the discovery of more than 1000 putative *cis*-regulatory elements in *Arabidopsis* [1]. *Arabidopsis* microarray data were also used by Yamamoto et al. [2] to predict *cis*-regulatory elements for ABA, auxin, brassinolide, cytokinin, ethylene, jasmonic acid, salicylic acid, and hydrogen

peroxide. However, validation tools allowing high-throughput analysis and characterization of these novel promoters and functional regulatory elements identified through transcriptomics and genomic analyses are still a challenge. The biological roles of the predicted promoters and their *cis*-elements have to be experimentally validated in planta. For this, full-length isolated promoter sequences or synthetic promoters containing putative *cis*-elements are placed upstream of reporter genes and are transiently or stably expressed in plants to determine their functionality.

The first critical step of the functional validation of promoter sequences is the design of complex expression vectors. DNA sequence editing is a time-consuming process with a high risk of introducing errors. Moreover, to store and manage promoter sequences and other biological parts is becoming more difficult as the number of parts for synthetic biology increases. Therefore, there is a need for software tools that help plant synthetic biologists through the design of application-specific expression vectors, including vectors for functional characterization of plant promoters.

2 Software

GenoCAD is a Computer-Aided Design (CAD) software for synthetic biology that relies on the concept of context-free grammars [3]. The grammars implemented in GenoCAD guide the design process of application-specific expression vectors. It is a freely available, web-based tool (www.genocad.com) that provides a system for managing genetic parts, organized according to functional categories, and which guides the user through the design by means of a set of rules that describe how to assemble these genetic parts to produce valid and functional constructs. It also allows the user to customize their workspace according to the requirements of their projects.

Originally released with a default basic *E. coli* grammar, today GenoCAD includes other brand new grammars developed by GenoCAD users [4–7]. Among them, we can find a plant grammar organized into three different modules according to the application of the final design, specifically promoter analysis, protein localization, and protein–protein interaction (PPI) studies [8]. In this chapter, we will focus on the design of constructs for in planta promoter analysis.

3 Plant Grammar

In the following section, we will describe the procedure to design plant expression vectors using GenoCAD (additional guidance is available at <http://solutions.genocad.com/support/home>). We will first present how to import and modify the plant grammar in order

The screenshot shows the GenoCAD homepage. At the top left is the GenoCAD logo. On the top right, there is a user greeting 'Welcome, Guest' with a 'Sign Up' button (highlighted with a red box) and a 'Log In' link. Below this is a search bar and a row of icons representing different tools. The main content area is divided into three columns: 'PARTS AND GRAMMARS' (with a DNA double helix icon), 'DESIGN CONSTRUCT' (with a circular DNA map icon), and 'SIMULATE' (with a molecular structure icon). Each column contains a title, a short description, and a button. Below this is a large banner announcing 'GenoCAD has moved!' with details about the platform's migration. The footer includes the GenoFAB logo and a list of links: 'About GenoCAD | Privacy Policy | Terms of Use | Tutorials | Mailing List | Solutions | Support'.

Fig. 1 GenoCAD homepage

to allow users to customize it according to the requirements of their projects. We will then construct a plant expression vector for promoter analysis studies as an example of how a plant vector can be designed using the GenoCAD plant grammar.

Launching GenoCAD opens a window that offers three options illustrated as a flow diagram (Fig. 1). In this chapter, we will focus on the “Parts and Grammars” section, which provides tools to import and edit the grammar and to manage the collection of genetic parts, and the “Design Construct” section.

Registration is not required to use GenoCAD, but it is recommended because it will make it possible for the user to import grammars, store parts, and save his/her constructs. To create an account, click the “Sign up” link on the upper right side (Fig. 1) and fill in and submit the form.

3.1 Importing the Plant Grammar into GenoCAD

The plant grammar is publically available in GenoCAD; therefore, we can use it to design our construct. However, before starting with the design, the users may prefer to customize the grammar according to their needs; the GenoCAD grammar editor makes this relatively easy. The public grammars are not editable, thus the

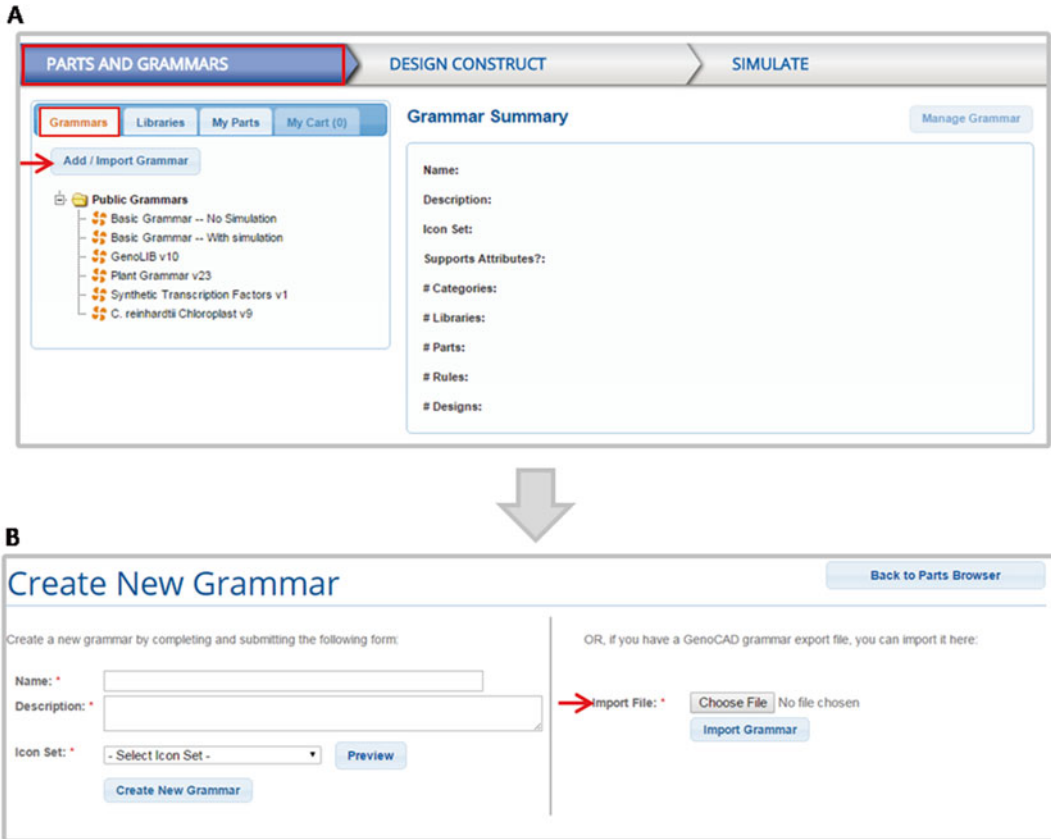


Fig. 2 Importing grammars. (a) On the “Grammar” section, the “Add/Import grammar” button is marked with the red arrow. (b) The tab to import the grammar is highlighted with a red arrow

first step is to copy the public grammar, if already available in the system, or to import it into GenoCAD.

The plant grammar can be both copied or imported because it is available in GenoCAD but also can be downloaded from Figshare [8]. Once it is saved to your computer, it is very simple to import it into GenoCAD. Log into GenoCAD.com, and head for the “Parts and Grammars” section. Then click on the “Grammars” button, and afterward select the “Add/Import Grammar” tab (Fig. 2a). In the newly displayed window click the “Choose File” tab, open the *SIFile.genocad* file previously downloaded from Figshare, and click “Import Grammar” (Fig. 2b). In this example, the imported grammar will be named “Customized Plant Grammar”.

3.2 Customizing the Grammar

Using GenoCAD grammar editor, advanced users can customize the plant grammar by adding/deleting rules and modifying their parts library in order to meet their specific needs.

3.2.1 Editing the Rules of the Grammar

By default, the promoter route (*pro*) of the plant grammar includes a set of rules that allow us to design plant expression vectors for

promoter analysis application purposes. However, the grammar does not provide the option of designing a simple expression cassette segment without the whole vector. In this demonstration, we will add a new rule to the previously imported plant grammar, i.e. “Customized Plant Grammar”, with the aim of introducing the option of designing an expression cassette suitable for promoter studies.

In the “Grammars” section, we select the grammar we want to modify and click the “Manage Grammar” tab (Fig. 3a). The grammar editor (Fig. 3b) consists of three main sections: the “Categories”, the “Category Details”, and the “Category Rules”. From the “Categories” section, we select the category we want to modify, i.e. “Promoter Analysis” (PROA). In the “Category rules”, we can see that the category includes one single rule, *npcas* which indicates that the expression cassette is composed of a native promoter along with the vector where it is inserted. To modify this route, we click the “Add Rule” button in the “Category rules” section (Fig. 3b).

A window that allows us to define a new rule is displayed (Fig. 4). After giving a code to the new rule, we drag and drop categories from the list on the left to the right in order to edit the rule. Categories can also be removed from a rule by clicking the delete button. We will introduce the option of designing an expression cassette for promoter studies; therefore we only need to select the category NPCT, which is the expression cassette that includes native promoter, transcribed region, and terminator. Finally, click “save”.

Back to the grammar editor, we can test to see if the new rule was properly added. In Fig. 5, we can see that the *pro* route offers now two design options: an expression cassette along with a vector, or an expression cassette segment (the new rule added).

3.2.2 Adding Genetic Parts

Currently, the plant library includes several general plasmid features commonly used for the design of expression vectors suitable for the three in planta functional studies incorporated in the grammar. Under gene and promoter categories, it includes specific sequences from *Solanum tuberosum* group *Phureja DMI-3* [9] as an example. However, non-expert users can easily add sequences of genetic parts according to their needs. As an example, we will add the sequence of the nopaline synthase (NOS) promoter.

In the “Library” section we select the library we want to edit, in our case the library from the “Customized Plant Library”. With that, a listing of the parts from our grammar’s part library, along with their descriptions, is shown on the right side (Fig. 6a). We click the “Add New Part” button, and a new window is displayed where we can add the name, description, sequence, and the category of the part (Fig. 6b). On the other hand, we can also add a list of parts at the same time by importing a tab-delimited text file or FASTA file.

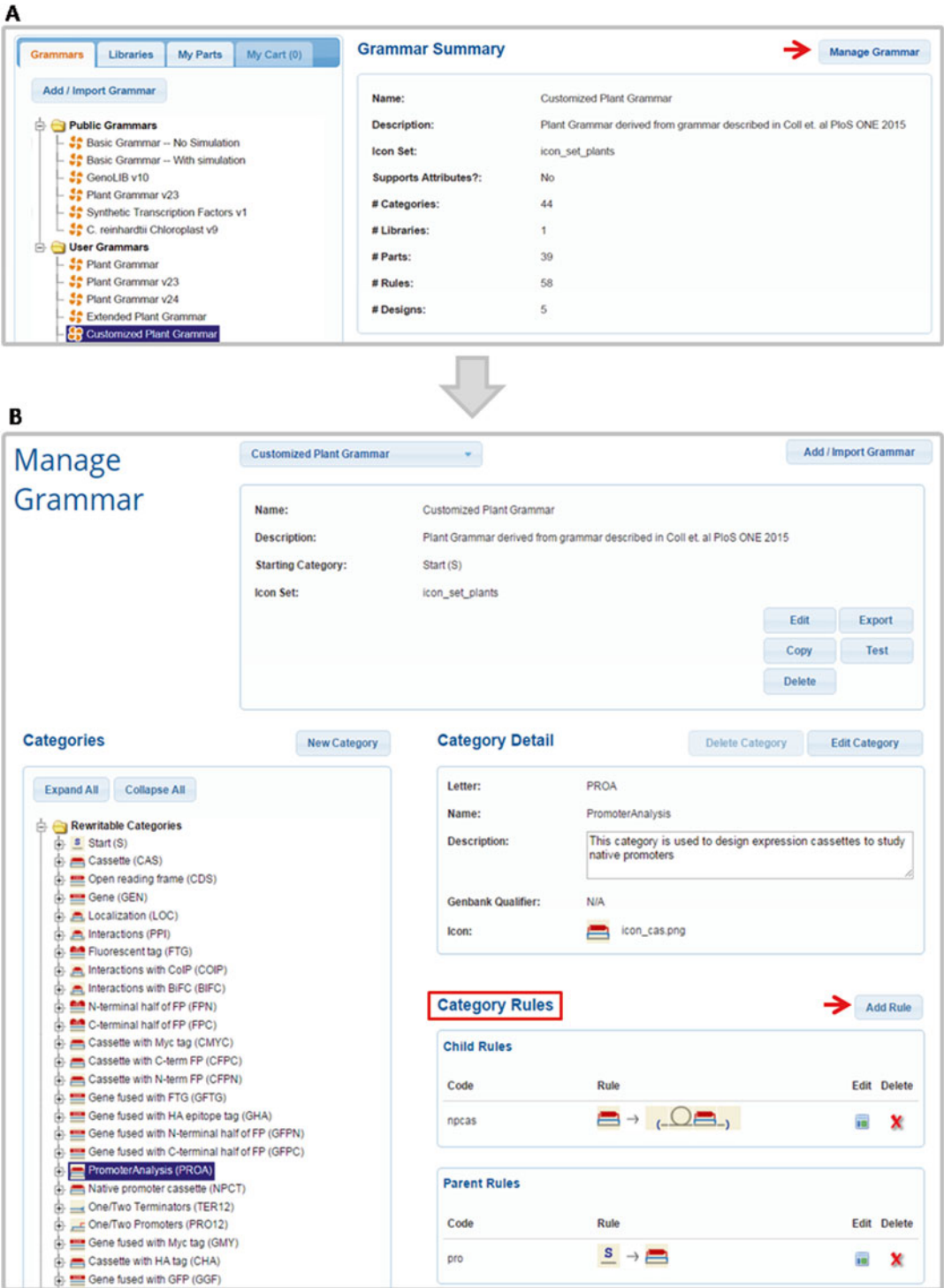


Fig. 3 Grammar editor. (a) To enter the grammar editor, click the “Manage Grammar” tab indicated by the *red arrow*. (b) The grammar editor includes three different sections; the “Category Rules” section is marked with a *red square*. Click on the “Add Rule” button (*red arrow*) to start editing the grammar

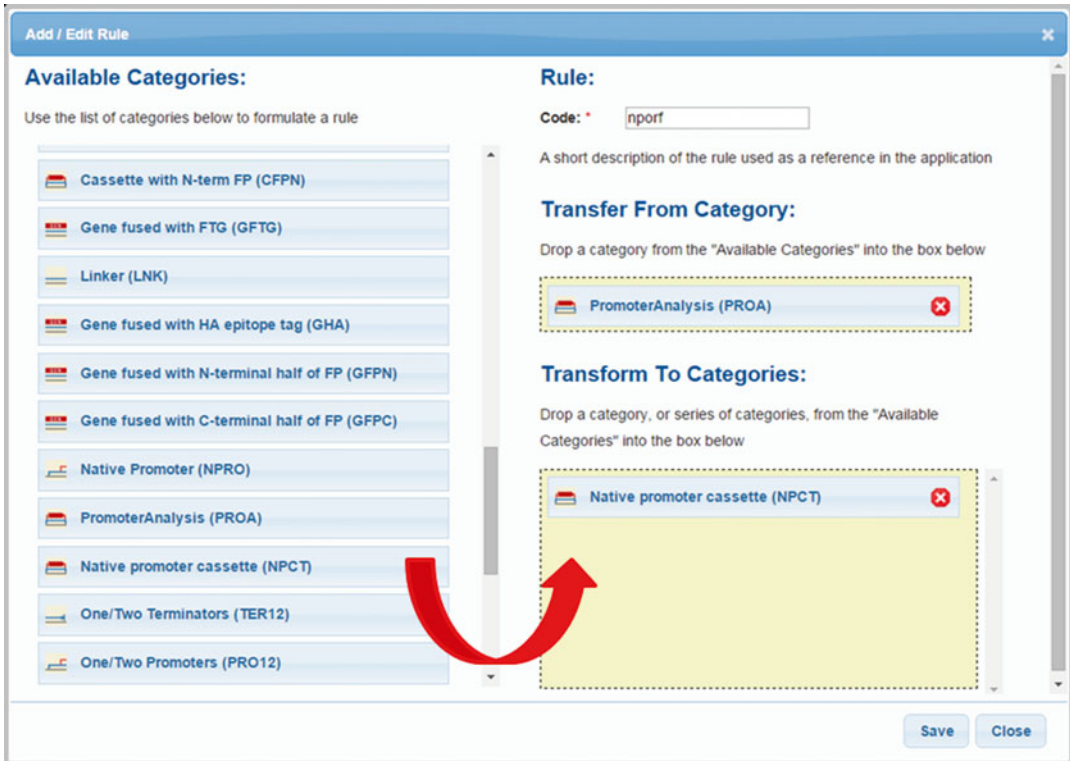


Fig. 4 Add/Edit rules. To edit a new rule, drag and drop the selected categories of DNA parts

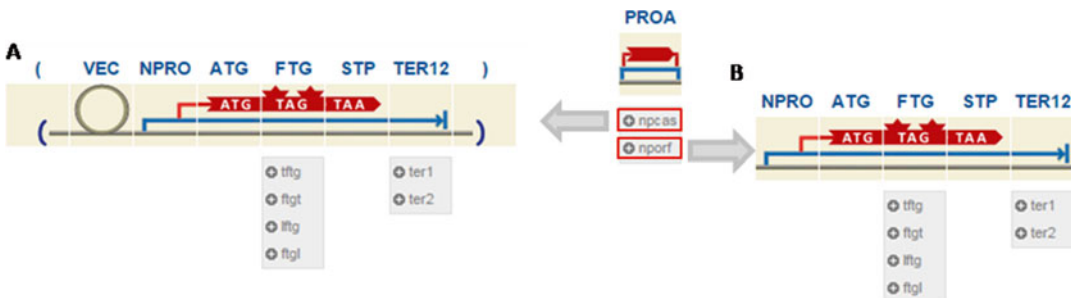


Fig. 5 Two rules define now the *pro* route. (a) Design of an expression vector for promoter analysis. (b) We added the option of designing only an expression cassette segment with the same purposes

3.3 Designing Plant Expression Vectors Using GenoCAD

Here, we will illustrate how to design constructs for in planta functional analysis studies. As it was previously mentioned, the grammar we developed allows the user to design constructs for three categories of experiments, i.e. promoter analysis, protein localization, and PPI studies. As an example, we will describe, step-by-step, how to design plant expression vectors suitable for promoter-reporter analysis. We will demonstrate how the set of designed rules implemented in the GenoCAD grammar guides

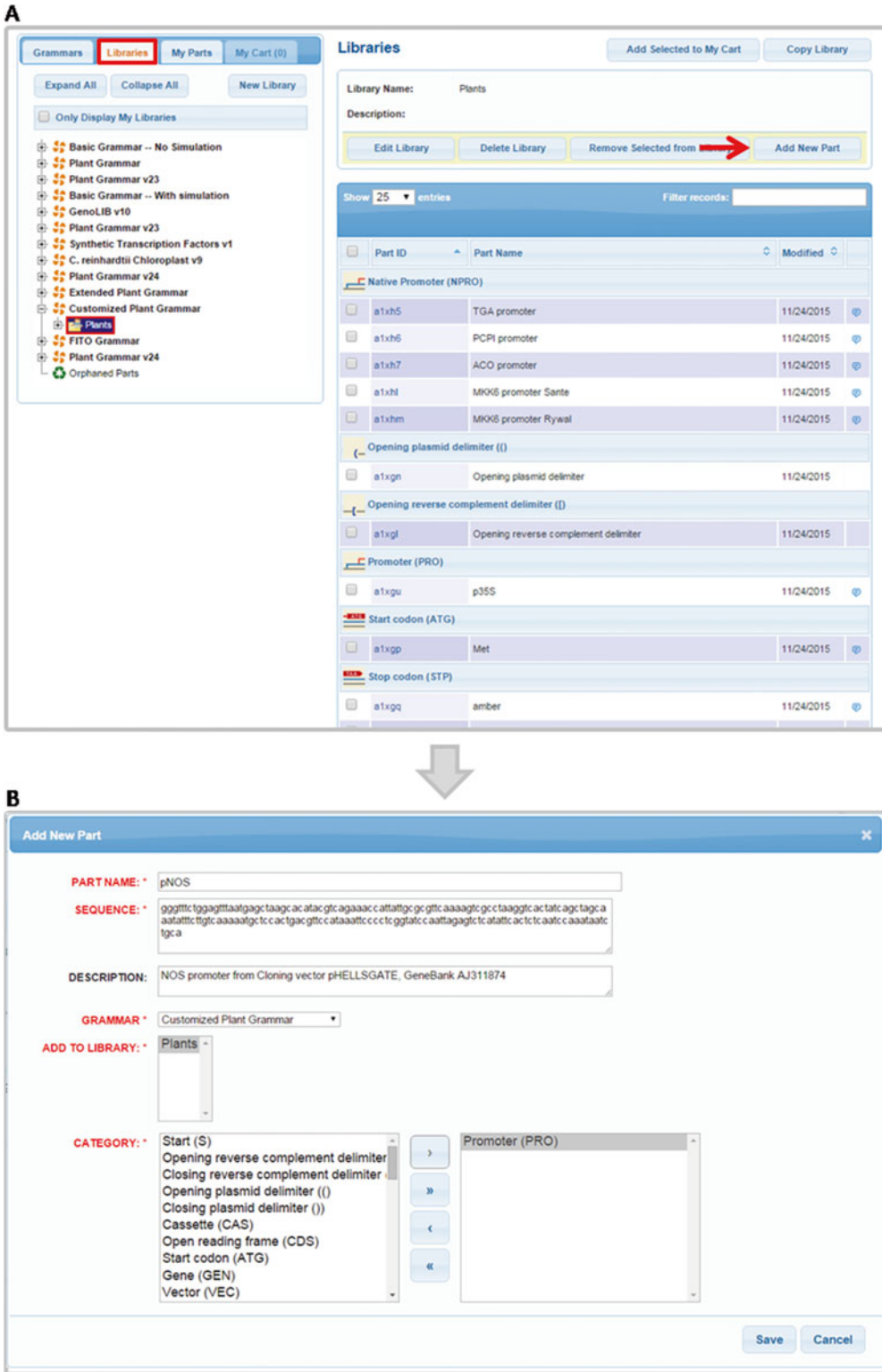


Fig. 6 Add a genetic part. (a) The “Library” section shows a list and characteristics of the genetic parts grouped into functional categories. The “Add New Part” button is marked with a red arrow. (b) Window that allows users to manually add a new genetic part sequence



Fig. 7 “Design Construct” tab. The plant grammar and library have to be selected; afterward GenoCAD will guide the user through the design, offering first three sets of rules grouped according to the application. Notice that rules changing the structure of the construct are shown in *grey squares*. We present here the design of plant expression vectors for promoter analysis studies, therefore we select the route *pro* (*red arrow*)

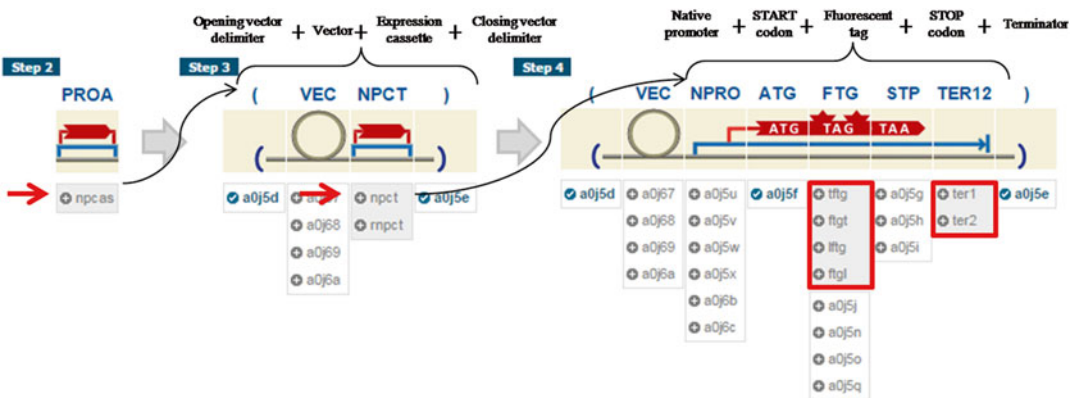


Fig. 8 Step-by-step design of a plant expression vector for promoter analysis studies

non-expert users through the design of a plant expression vector to characterize mitogen-activated protein kinase 6 (MKK6) from *S. tuberosum* cv. Santé [10].

From the GenoCAD main page, we go to the design wizard by clicking on the “Design Construct” link. Following that, we first have to select the grammar we would like to use (Fig. 7). We select our “Customized Plant Grammar”. In the next step, the appropriate library needs to be chosen. Currently, only one library exists for the plant grammar, therefore this library is automatically selected. Now we are ready to start the design; from the first step, we can immediately decide which route we want to follow according to the aim of the final construct: localization studies, PPI or promoter analyses. To show in detail how the design works, using GenoCAD we select, as an example, the *pro* route (Fig. 7).

Once we start with the promoter analysis route (*see* Fig. 8 to follow the design step-by-step), we first see the rule *npcas*, which

defines that our construct will enclose a vector with an expression cassette. Subsequently (Fig. 8, step 3), two rules can be chosen: npct or rnpct. Both rules constrain the user to design an expression cassette with a native plant promoter (NPRO) fused to a reporter gene (FTG) which is the minimal requirement of promoter-reporter systems. The rule rnpct allows the user to clone the expression cassette in reverse orientation. We select here npct rule, which breaks down the expression cassette into a NPRO followed by a FTG and a terminator (TER12). The start (ATG) and stop (STP) codons are not part of the gene sequence. They are considered as categories in order to facilitate the design of fusion proteins. Therefore, the open reading frame includes a fluorescent protein with start and stop codon. To incorporate flexibility into the design, rules *tftg*, *ftgt*, *lftg*, and *ftgl* allow the user to add tags and/or linker domains at both sides of reporter protein. Moreover, there are two rules that can be used to rewrite the category TER12, i.e. rule *ter1* is used to add a single terminator, and *ter2* allows the user to add a double terminator (step 4, Fig. 8).

In our example, we will fuse the fluorescent protein with an epitope tag (ETG) at its C-terminus for immunoprecipitation purposes, and thus we click on the *ftgt* rule. Moreover, by selecting the *ftgl* rule we include a linker between the FTG and the ETG. The final construct is shown in Fig. 9. We can always step back through the history of the design process to make any changes.

Once all the categories of genetic parts are selected according to the application and the needs of the user, the final step is to select the part sequences for each category (Fig. 9a). Each part sequence has a unique GenoCAD ID. When the users drag the mouse over the ID, characteristics of the sequence will be shown in order to facilitate the selection.

At the moment, the plant library allows the user to select between four different vector backbones to clone the assembled parts. All of them are pCAMBIA [11] minimal selection vectors compatible with *Agrobacterium*-mediated plant transformation. They contain minimal heterologous sequences for plant transformation and differ in the bacterial selection (allowing the user to choose between chloramphenicol or kanamycin) and the plant selection (hygromycin B or kanamycin). All vectors were opened at the multiple cloning site (MCS) with *SalI* and *BamHI* restriction enzymes. By selecting GenoCAD ID a1xhh, we chose here pCAMBIA1200 [11] containing chloromphenicol and hygromycin resistance. Since the aim of this example is to characterize the potato promoter of MKK6 (StMKK6), we select then the promoter of this gene (ID a1xhl) among the five plant promoters currently available in the library, and we fused it with yellow fluorescent protein (eYFP, ID a1xgt) as a reporter protein in this example. However, the FTGs added in the GenoCAD parts library include enhanced YFP (eYFP), green fluorescent protein (eGFP), cyan fluorescent

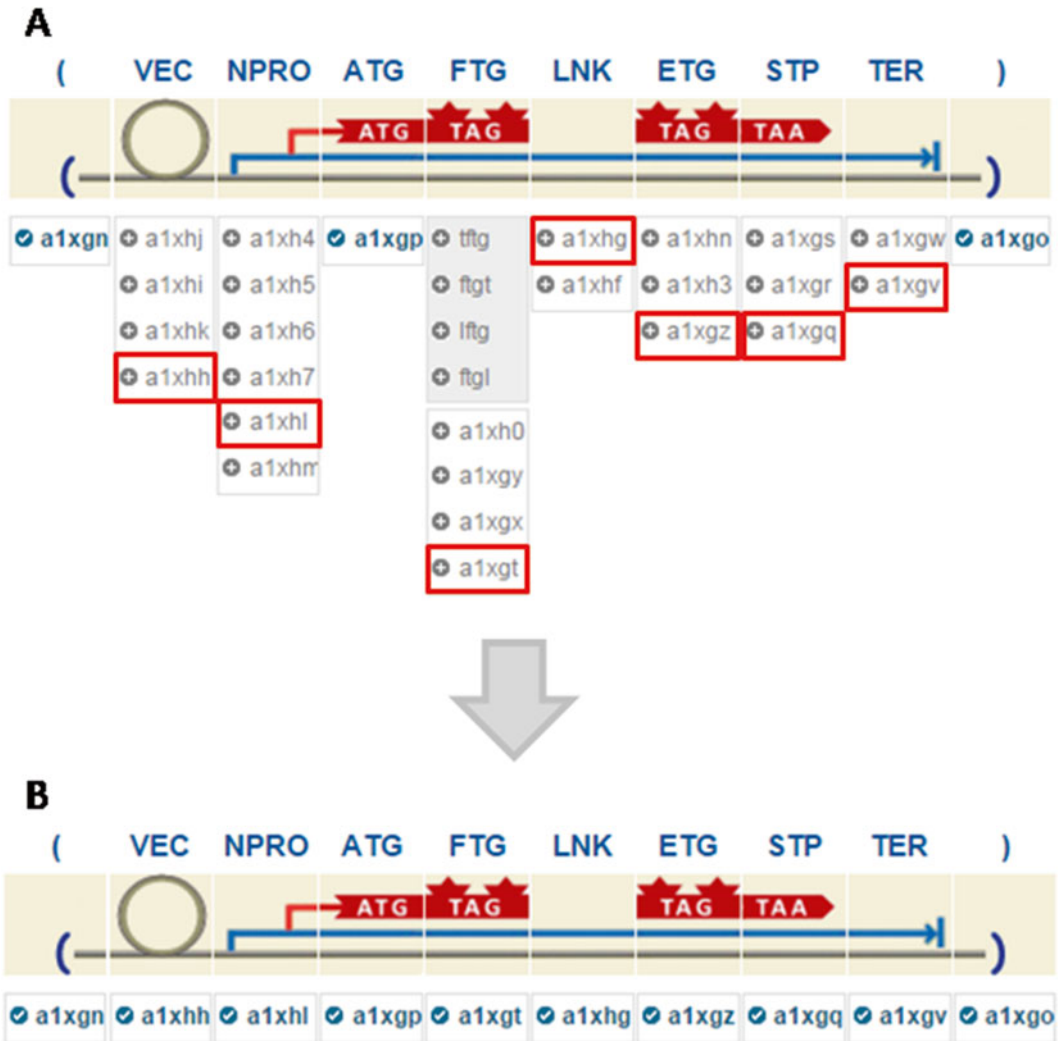


Fig. 9 Design for promoter analysis. (a) The construct includes a native promoter that controls the expression of a fluorescent protein (minimal requirement for a construct with promoter-reporter analysis purposes); moreover, at the C terminal of the fluorescent tag, we included an epitope tag fused by a linker. The part sequences for each category selected in this example are marked with a *red square*. (b) Selecting the part sequence of each category, we obtain the final design

protein (eCFP) and mCherry, and all have been previously tested in *Nicotiana benthamiana* and *S. tuberosum* leaves. We then include a short linker (ID a1xgt) between the fluorescent protein and the ETG and choose myc (ID a1xgz) as ETG. The final design is shown in Fig. 9b.

The design can be saved in GenoCAD for further work with it. By clicking on the “Generate Sequence” tab, the sequence of the designed construct can be exported in three different formats: a GenBank file, a FASTA file, or a Plain Sequence file (Fig. 10).

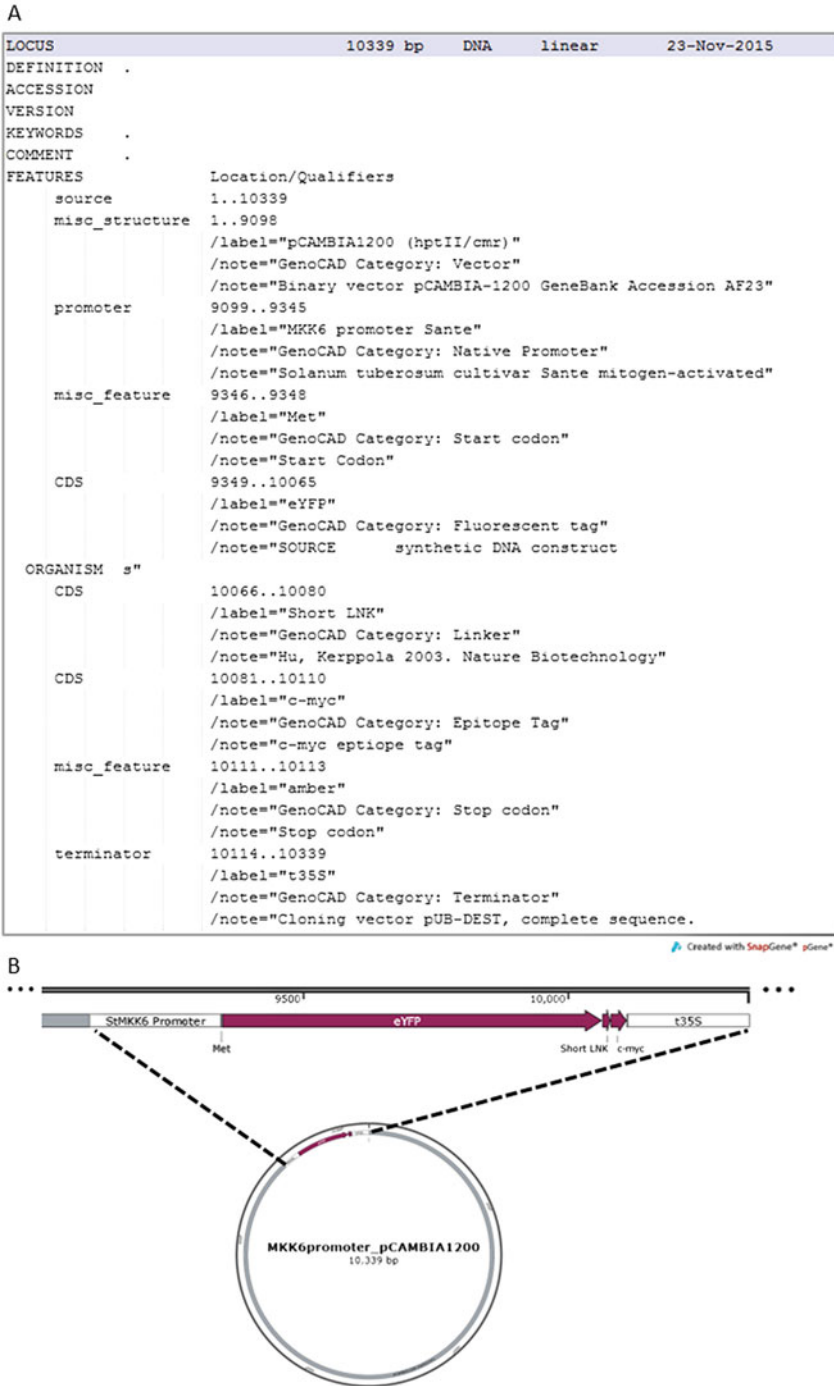


Fig. 10 Final output of GenoCAD. (a) Sequence of the construct has been exported as GenBank file and (b) visualized using SnapGene software (from GSL Biotech; available at snapgene.com)

This sequence can now be synthesized using the service of a DNA synthesis company, or used for designing cloning primers in line with the chosen cloning strategy.

4 Conclusions

This chapter details how to use the plant grammar implemented in GenoCAD, which guides the user through the design of expression vectors for in planta functional analyses. We focused here on the design of constructs for promoter analysis purposes. However, the grammar covers two more types of applications, i.e. protein localization and PPI studies, and our aim is to extend it with other functional applications interesting for plant biologists.

In our lab, we are now in the process of validation of constructs presented in the grammar. Although all library parts have been tested, the complete final plasmids have not yet been experimentally verified for functionality.

There is no doubt that the plant grammar will reduce time and cost of our experiments by decreasing the probabilities of errors and facilitating the design of complex plant expression vectors that can then be obtained using sequence-independent methods.

Acknowledgments

This work was supported by NSF Awards 1241328, the Slovenian Research Agency Program P4-0165, and Project N4-0026.

References

1. Zou C, Sun K, Mackaluso JD et al (2011) Cis-regulatory code of stress-responsive transcription in *Arabidopsis thaliana*. *Proc Natl Acad Sci U S A* 108:14992–14997. doi:[10.1073/pnas.1103202108](https://doi.org/10.1073/pnas.1103202108)
2. Yamamoto YY, Yoshioka Y, Hyakumachi M et al (2011) Prediction of transcriptional regulatory elements for plant hormone responses based on microarray data. *BMC Plant Biol* 11:39. doi:[10.1186/1471-2229-11-39](https://doi.org/10.1186/1471-2229-11-39)
3. Cai Y, Hartnett B, Gustafsson C, Peccoud J (2007) A syntactic model to design and verify synthetic genetic constructs derived from standard biological parts. *Bioinformatics* 23:2760–2767. doi:[10.1093/bioinformatics/btm446](https://doi.org/10.1093/bioinformatics/btm446)
4. Wilson ML, Okumoto S, Adam L, Peccoud J (2014) Development of a domain-specific genetic language to design *Chlamydomonas reinhardtii* expression vectors. *Bioinformatics* 30:251–257. doi:[10.1093/bioinformatics/btt646](https://doi.org/10.1093/bioinformatics/btt646)
5. Overend C, Yuan L, Peccoud J (2012) The synthetic futures of vesicular stomatitis virus. *Trends Biotechnol* 30:497–498. doi:[10.1016/j.tibtech.2012.06.002](https://doi.org/10.1016/j.tibtech.2012.06.002)
6. Adames NR, Wilson ML, Fang G et al (2015) GenoLIB: a database of biological parts derived from a library of common plasmid features. *Nucleic Acids Res* 43:4823–4832. doi:[10.1093/nar/gkv272](https://doi.org/10.1093/nar/gkv272)
7. Purcell O, Peccoud J, Lu TK (2014) Rule-based design of synthetic transcription factors in eukaryotes. *ACS Synth Biol* 3(10):737–744. doi:[10.1021/sb400134k](https://doi.org/10.1021/sb400134k)
8. Coll A, Wilson ML, Gruden K, Peccoud J (2015) Rule-based design of plant expression vectors using GenoCAD. *PLoS One* 10, e0132502. doi:[10.1371/journal.pone.0132502](https://doi.org/10.1371/journal.pone.0132502)

9. The Potato Genome Sequencing Consortium (2011) Genome sequence and analysis of the tuber crop potato. *Nature* 475:189–195. doi:[10.1038/nature10158](https://doi.org/10.1038/nature10158)
10. Lazar A, Coll A, Dobnik D et al (2014) Involvement of potato (*Solanum tuberosum* L.) MKK6 in response to potato virus Y. *PLoS One* 9, e104553. doi:[10.1371/journal.pone.0104553](https://doi.org/10.1371/journal.pone.0104553)
11. Hajdukiewicz P, Svab Z, Maliga P (1994) The small, versatile PZP family of *Agrobacterium* binary vectors for plant transformation. *Plant Mol Biol* 25:989–994

Bioinformatic Identification of Conserved *Cis*-Sequences in Coregulated Genes

Lorenz Bülow and Reinhard Hehl

Abstract

Bioinformatics tools can be employed to identify conserved *cis*-sequences in sets of coregulated plant genes because more and more gene expression and genomic sequence data become available. Knowledge on the specific *cis*-sequences, their enrichment and arrangement within promoters, facilitates the design of functional synthetic plant promoters that are responsive to specific stresses. The present chapter illustrates an example for the bioinformatic identification of conserved *Arabidopsis thaliana* *cis*-sequences enriched in drought stress-responsive genes. This workflow can be applied for the identification of *cis*-sequences in any sets of coregulated genes. The workflow includes detailed protocols to determine sets of coregulated genes, to extract the corresponding promoter sequences, and how to install and run a software package to identify overrepresented motifs. Further bioinformatic analyses that can be performed with the results are discussed.

Key words Coregulated genes, Promoter sequences, Overrepresented motifs, *Cis*-regulatory sequences, *Cis*-elements, PathoPlant, TAIR, BEST, AthaMap

1 Introduction

Regulation of gene expression in plants is required to trigger plant development and environmental responses. Gene expression is regulated by a wide array of mechanisms to increase or decrease the production of specific gene products. Almost any step of gene expression can be regulated, e.g. transcriptional initiation, post-transcriptional RNA processing, or post-translational modification of a protein. The most prominent functional elements in gene regulation are transcription factors (TFs) and their corresponding TF binding sites located within the promoters of their target genes [1]. The TF binding sites are short *cis*-regulatory sequences that are targeted by specific TFs regulating gene expression at transcriptional level. In order to engineer a functional synthetic promoter triggering the specific expression of a gene, it is crucial to include the necessary *cis*-regulatory sequences for controlled gene expression [2]. Potential *cis*-regulatory sequences can be identified as

overrepresented motifs in promoter sequences within a set of genes expressed under specific conditions [3]. The present protocol showcases an example describing the identification of potential *cis*-regulatory sequences as overrepresented motifs in promoter sequences from a set of coregulated genes. It includes the identification of sets of coregulated genes, the extraction of the corresponding promoter sequences, a protocol how to install and run software to identify overrepresented motifs, and further analyses that can be performed with the results. The approach is based on the computational analysis of experimental expression and sequence data. It has been successfully employed to identify numerous plant *cis*-regulatory sequences responsive to biotic stresses [4] and to identify *cis*-sequences enriched in promoters of genes responsive to abiotic stresses [5].

2 Bioinformatic Identification of Conserved *Cis*-Sequences in Coregulated Genes

2.1 Identification of Coregulated Genes

To determine genes being coregulated, microarrays can be employed to simultaneously measure the expression levels of large numbers of plant genes [6, 7]. The development of RNA-Seq technology even enables a whole transcriptome shotgun quantification of gene expression [8]. For the model plant species *Arabidopsis thaliana*, numerous genome-wide gene expression profile experiments have been conducted and the results are stored in publicly available databases. While TAIR Microarray Experiments [9] and NASCArrays [10] represent databases that have been focusing on *A. thaliana* microarray data, NCBI's Gene Expression Omnibus (GEO; <http://www.ncbi.nlm.nih.gov/geo/>; [11]) is a public repository for high-throughput microarray and next-generation sequence expression data from all organisms. TAIR Microarray Experiments and NASCArrays no longer accept new submissions of *A. thaliana* gene expression datasets and have moved their data to NCBI's GEO. TAIR, NASCArrays, and GEO can be used as resources to retrieve expression data to be used in any approach for the identification of coregulated genes.

Such data were, for example, used to be implemented within the PathoPlant database [12]. PathoPlant (<http://www.pathoplant.de/>) harbors *A. thaliana* microarray expression data and focuses on biotic and abiotic stress experiments [13, 14]. PathoPlant's Microarray expression tool is available at <http://www.pathoplant.de/microarray.php> and enables queries for genes coregulated under specific stresses [13]. In the present example, *A. thaliana* genes being coregulated under drought stress conditions will be identified using PathoPlant to subsequently determine overrepresented motifs within their promoters. These may act as drought-specific *cis*-regulatory sequences. In the present example, a drought experiment with shoots being analyzed after 6 h of

drought treatment of *A. thaliana* seedlings from a series of drought experiments performed by the AtGenExpress consortium is used for the analysis [15]. Figure 1 shows PathoPlant's Microarray expression tool at <http://www.pathoplant.de/microarray.php> where specific stresses and parameters can be selected. The upper part of Fig. 1 illustrates the selected stress (drought-stressed shoots 6hr). The other search parameters were left as defaults with the expression parameter induction factor being at least >4 fold for the means from the replicates in order to identify only highly induced genes, the Boolean operator being AND, which has no effect when only one stress is selected, as well as not excluding genes regulated by smallRNA. The query will be submitted by selecting the Search button. PathoPlant returns the query results as a table specifying the coregulated genes that match the search criteria (*see Note 1*). In this result table (not shown), the genes are identified by their locus identifier and the table additionally displays a short gene description, the induction factor for a single experiment, the mean induction factors from replicate experiments, the corresponding stresses, external links and statistical information. The number of records is displayed directly below the Search button and in this example says "Display of 33 records representing 18 loci" (Fig. 1). The number of records is higher than the number of loci (genes) due to replicate experiments. By selecting the link "18 loci", a table with the locus identifiers is displayed. This table is shown in the

The screenshot shows the PathoPlant web interface in a Mozilla Firefox browser. The URL is <http://www.pathoplant.de/microarray.php?type=stimuli&s1=Drought-stressed+shoots+6hr&s2=&s3=>. The page title is "PathoPlant".

The search parameters are:

- Search by stimulus: Drought-stressed shoots 6hr
- Expression parameter induction factor: mean from replicates
- Boolean operator: AND
- Exclude genes regulated by smallRNA: (unchecked)

The search results show "Display of 33 records representing 18 loci". A table of results is displayed:

Locus	Number of records
At1g45145	2
At1g52410	2
At1g62540	1
At1g70700	2
At2g33380	2
At2g34930	1
At2g39030	2
At2g39330	2

On the right side, there is a legend for the induction factor ranges:

- factor < -4
- 4 <= factor < -2.5
- 2.5 <= factor < -1.5
- 1.5 <= factor < 1.5
- 1.5 <= factor < 2.5
- 2.5 <= factor < 4
- factor >= 4

Fig. 1 Identification of 18 drought-induced *Arabidopsis thaliana* genes using PathoPlant

lower part of Fig. 1. The gene identifiers of the set of 18 coregulated genes under drought stress can be copied from this table.

Further drought-induced gene sets can be retrieved in a similar way by selecting another of the 14 drought-stress conditions annotated to PathoPlant, by combining different drought-stress conditions using the Boolean operator AND, by varying the minimum induction factor, or by altering any of the other query parameters. A comprehensive set of 179 overrepresented motifs was identified within the promoters of an array of 32 different drought-induced gene sets each one consisting of 7–34 coregulated genes [5]. In this way, PathoPlant’s Microarray expression tool can be employed for identification of coregulated genes for all annotated microarrays (http://www.pathoplant.de/documentation_microarrays.php).

2.2 Extraction of Promoter Sequences

In order to screen the promoters of coregulated genes for overrepresented motifs, the corresponding promoter sequences have to be known. *A. thaliana* was the first flowering plant whose complete genomic sequence was published [16], and since the year 2000, the sequences of more than 100 plant genomes have been released [17]. The genome of *A. thaliana* exhibits a high gene density with relatively short intergenic regions [16]. A study analyzing SNPs within the upstream regions of genes resulted in an estimate of 500 bp upstream of the transcription start site (TSS) for an effective average promoter length [18]. Nonetheless, functional *cis*-elements are also be found farther away from the TSS and thus individual promoter lengths will vary [18]. Furthermore, the TSS is not known for every *A. thaliana* gene, and in order to also account for the space of the 5’UTR, 1000 bp upstream of the coregulated genes were extracted from the genome as promoter sequences in the drought stress study cited above ([5]; see **Note 2**).

The Arabidopsis Information Resource (TAIR) offers a tool available at <https://www.arabidopsis.org/tools/bulk/sequences/> to download *A. thaliana* promoter sequences by submitting sets of locus identifiers [19]. Following the example described above, the locus identifiers of the 18 coregulated genes under drought stress are pasted into the textbox, the dataset is set to “TAIR10 Loci Upstream Seq - 1000bp”, the “Search against” parameter is set to “Get one sequence per locus (representative gene model/splice form only)”, and Fasta is selected as output format (Fig. 2). By pressing the Get Sequences button, the corresponding promoter sequences are displayed (Fig. 3; see **Note 1**). The sequences are copied to a text editor and saved as a flat text file with Fasta .fas extension (droughtshoots6hr.fas; see **Note 3**).

2.3 Screening for Overrepresented Motifs

For the drought stress study [5], the Binding-site Estimation Suite of Tools (BEST) software package [20] was employed for *de novo* identification of overrepresented motif sequences within the promoters of coregulated genes under drought stress. The advantage of BEST consists in the integration of the four different

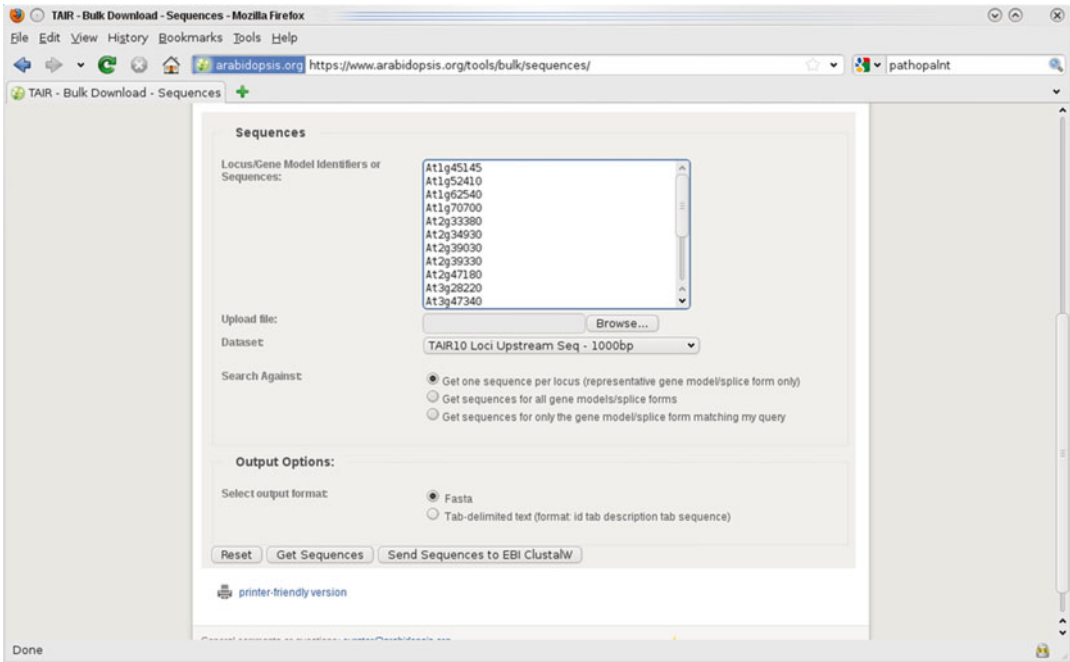


Fig. 2 Extraction of promoter sequences from drought-induced genes using TAIR

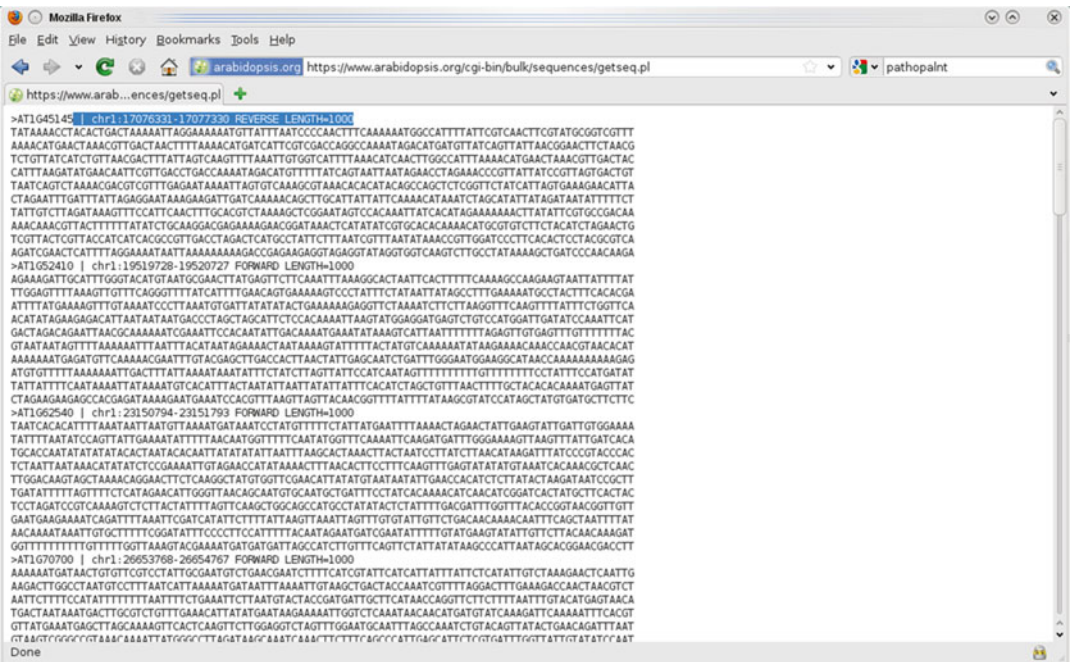


Fig. 3 Fasta-formatted promoter sequences from drought-induced *Arabidopsis thaliana* genes

motif-finding programs MEME [21], AlignACE [22], CONSENSUS [23], and BioProspector [24], with BioOptimizer that takes the output of the four programs and performs an optimization step on the results [25]. Furthermore, BEST provides a

graphical user interface (GUI), which makes its usage easy and intuitive when locally installed on a Linux operating system. A download of the BEST software package is provided at <http://www.people.fas.harvard.edu/~junliu/BEST/> including the BEST documentation with installation instructions.

To install and run BEST, a Linux operating system is required. To be compatible with different Linux versions, 3 versions of BEST are provided for download (<http://www.people.fas.harvard.edu/~junliu/BEST/>). Suitable Linux versions are indicated but other or newer Linux versions should work as well. The set of 18 promoter sequences derived from *Arabidopsis thaliana* genes induced under drought stress was screened using 2 BEST versions installed on two different operating systems: BEST1.0.1 from the file BEST1.0.1.tar.gz on a Linux SuSE 9.2 system and BEST source code version from the file BEST.tar.gz on a Linux SuSE 9.0 system. All following steps are described only for the latter version of BEST (see **Note 4**). After downloading and copying the file BEST.tar.gz to a directory, e.g. the user's home directory, of a Linux SuSE 9.0 system, a terminal window of the corresponding directory can be opened by selecting "Open Terminal" from the file manager "Tools" menu. All files of the current directory will be displayed by typing "ls" and the downloaded file BEST.tar.gz will appear. By typing "gunzip BEST.tar.gz", the file will be unzipped to the archive file BEST.tar. Extraction of the individual files to a newly created BEST directory will be performed by the command "tar -xvf BEST.tar". The command "cd BEST" changes to the newly created BEST directory, and the BEST software package will be installed by typing "./INSTALL". During the process of installation, one shall answer "n" as "no" when prompted.

Before running BEST, the droughtshoots6hr.fas file with the promoter sequences generated previously is copied to a newly created directory named "drought" within the BEST directory. To run BEST, a terminal window from the BEST directory is opened and the command "./BEST" is typed resulting in a new window displaying the BEST GUI (Fig. 4). BEST is subsequently run with default parameters and motif lengths of five to ten nucleotides. The application of these parameters had previously shown to yield optimal results with promoter sequences from *A. thaliana* [4, 5]. This is in accordance with the finding that most eukaryotic transcription factor binding sites span 5–8 bp, whereas TF footprints are typically 10–20 bp in size [26]. The BEST GUI displays five buttons named "AlignACE", "BioProspector", "Consensus", "MEME", and "BioOptimizer" to set individual screening parameters (Fig. 4). Pressing the button named "AlignACE" will result in a new window to set the parameters for the AlignACE screening algorithm (Fig. 5). In the described example, droughtshoots6hr.fas within the BEST/drought directory is selected as input file and the number of columns to align, which means the motif length, is

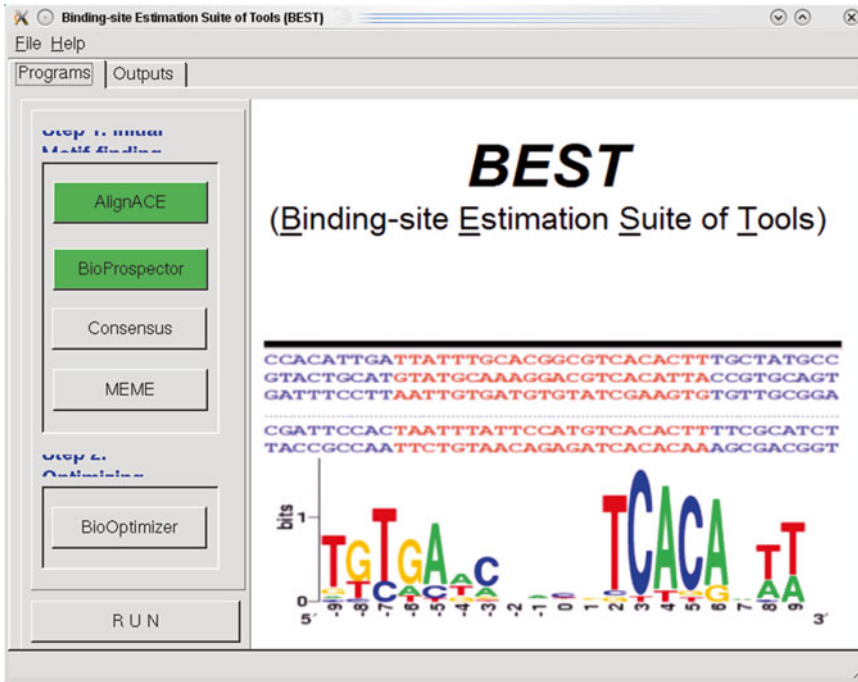


Fig. 4 Main graphical user interface of BEST

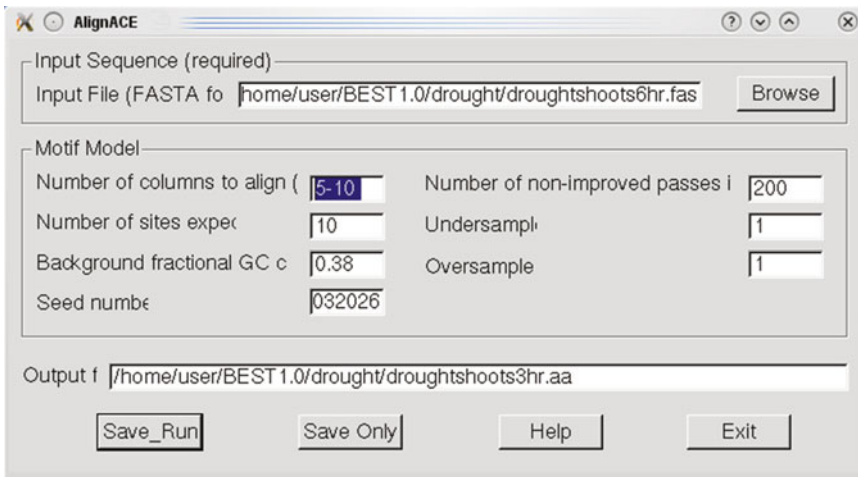
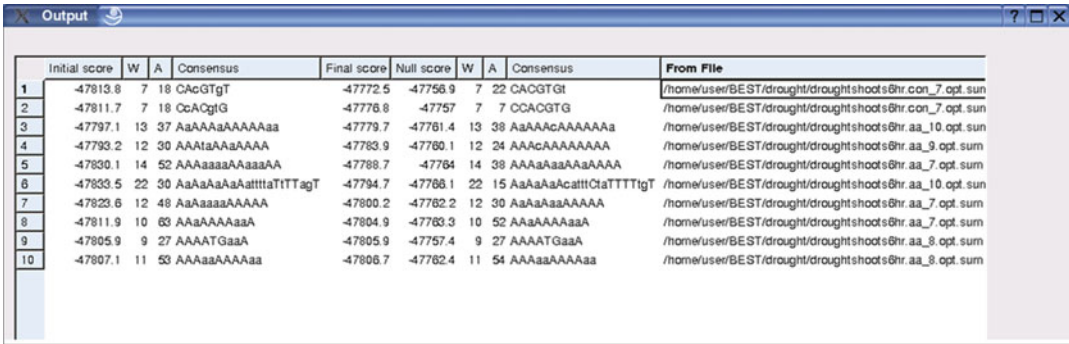


Fig. 5 Setting of BEST screening parameters

set to 5–10 (highlighted in Fig. 5; see Note 5). The AlignACE parameters are saved by selecting the “Save Only” button. By opening the remaining parameter windows for BioProspector, Consensus, MEME, and BioOptimizer and selecting the “Save Only” button, BEST takes over the parameters from AlignACE for the other three screening programs and for BioOptimizer and settings are completed. Once parameters are set for an individual



	Initial score	W	A	Consensus	Final score	Null score	W	A	Consensus	From File
1	-47813.8	7	18	CACGTgT	-47772.5	-47756.9	7	22	CACGTGT	/home/user/BEST/drought/droughtshoots6hr.con_7.opt.sum
2	-47811.7	7	18	CcACgtG	-47776.8	-47757.7	7	7	CCACGTG	/home/user/BEST/drought/droughtshoots6hr.con_7.opt.sum
3	-47797.1	13	37	AaAaAaAAAAAaa	-47779.7	-47761.4	13	38	AaAAcAAAAAaaa	/home/user/BEST/drought/droughtshoots6hr.aa_10.opt.sum
4	-47793.2	12	30	AAAtaAAAAAA	-47783.9	-47760.1	12	24	AAAcAAAAAaaa	/home/user/BEST/drought/droughtshoots6hr.aa_9.opt.sum
5	-47830.1	14	52	AAAAaaaAAAAAA	-47788.7	-47784.14	38	AAaAaaaAAAAAA	/home/user/BEST/drought/droughtshoots6hr.aa_7.opt.sum	
6	-47803.5	22	30	AaAaAaAaAaattttaTTtagT	-47794.7	-47766.1	22	15	AaAaAaAcatttCtaTTTTgT	/home/user/BEST/drought/droughtshoots6hr.aa_10.opt.sum
7	-47823.6	12	48	AaAaaaaAAAAA	-47800.2	-47762.2	12	30	AaAaAaaAAAAA	/home/user/BEST/drought/droughtshoots6hr.aa_7.opt.sum
8	-47811.9	10	63	AAAAAaaaaA	-47804.9	-47763.3	10	52	AAAAAaaaaA	/home/user/BEST/drought/droughtshoots6hr.aa_7.opt.sum
9	-47805.9	9	27	AAAATGaaA	-47805.9	-47757.4	9	27	AAAATGaaA	/home/user/BEST/drought/droughtshoots6hr.aa_8.opt.sum
10	-47807.1	11	53	AAAAAaaaaA	-47806.7	-47762.4	11	54	AAAAAaaaaA	/home/user/BEST/drought/droughtshoots6hr.aa_8.opt.sum

Fig. 6 BEST Output window displaying the screening results

screening program, the corresponding button of the main BEST GUI turns green. Five green buttons indicate that parameters for all screening program and for BioOptimizer have been set, and the screening is started by selecting the “RUN” button (*see Note 6*).

After finishing the screening, BEST opens an Output window displaying the top ten motifs (Consensus) identified within the promoter sequences from genes induced upon drought stress (Fig. 6). BEST produces this output table of overrepresented motifs sorted according to their BioOptimizer scores [20]. For each of the top ten motifs, the table displays the motif scores, widths (W), number of predicted sites (A) and consensus sequences from both the original motif-finding program (left columns) and BioOptimizer (right columns) (for “From File” *see Note 7*). By default, the BEST/drought directory indicated as input file directory also harbors all the BEST screening results. The results from the Output table displayed in Fig. 6 are automatically saved to the file droughtshoots6hr_BEST.Summary. The first two motifs of the Output table within the right Consensus column show CACGTGt and CCACGTG consensus sequences, whereas the remaining eight motifs are comprised of polyA consensus sequences (Fig. 6). These polyA motifs are certainly overrepresented within the promoter sequences but are unlikely to be specific to promoters of genes responsive to drought stress since plant promoter sequences are generally AT-rich [27]. The first two motifs also harbor a common CACGTG G-box core motif [28]. The single sequences underlying the first two motifs from the Output table are extracted from the file droughtshoots6hr.con_7.opt.all (*see Note 7*) and complemented with Fasta headers to construct alignment matrices of the motifs designated >drought1 and >drought2:

```
>drought1
CACGTGT
CACGTGT
CACGTGT
```

CACGTGC
CACGTGG
CACGTGT
CACGTGT
CACGTGG
CACGTGG
CACGTGG
CACGTGT
CACGTGT
CACGTGT
CACGTGT
CACGTGT
CACGTGT
CACGTGT
CACGTGG
CACGTGT
CACGTGT
CACGTGC
CACGTGG
CACGTGG
>drought2
CCACGTG
CCACGTG
CCACGTG
CCACGTG
CCACGTG
CCACGTG
CCACGTG
CCACGTG

3 Further Analyses

There are many different options for further bioinformatics analyses using the identified alignment matrices. The two alignment matrices can be used to compare the underlying *cis*-sequences or motifs with already known ones from different databases. STAMP is a web tool that performs such an analysis [29]. It determines motif similarities among different submitted motifs and a comparative analysis with known motifs and *cis*-regulatory sequences from an array of publicly available databases [29]. STAMP is accessed at <http://www.benoslab.pitt.edu/stamp/> to paste the alignment

matrices of the motifs. The databases with known plant *cis*-elements that can be selected within STAMP are AGRIS [30], AthaMap [31], and PLACE [32]. At the Similarity Matching section, one of the three implemented plant motif databases (AGRIS, AthaMap, or PLACE) is selected. The default values are used for all other alignment parameters and the motifs are submitted to STAMP. The analysis reveals that the two motifs are significantly similar to a G-box motif and to abscisic acid-responsive elements (not shown) that are known bZIP transcription factor binding sites and that are associated with the plant's abiotic stress response [28, 33–35]. The two identified motifs therefore constitute probably functional *cis*-regulatory elements responsive to drought stress and possibly responsive to other abiotic stresses.

As the putative transcription factors recognizing the identified motifs belong to the bZIP family of transcription factors, another analysis on the architecture of the promoters is performed in order to reveal the arrangement of bZIP binding sites within the promoter sequences. For promoters from *Arabidopsis thaliana*, such an analysis can easily be performed using AthaMap's Gene Analysis function that is available at http://www.athamap.de/search_gene.php [36]. The seven unique genes with promoters harboring the identified motifs are extracted from the file droughtshoots6hr.con_7.opt.all (AT1G70700, AT2G33380, AT2G47180, AT3G28220, AT3G47340, AT4G15210, AT4G23600, separated by carriage returns) and pasted into the Genes (AGI) form at AthaMap's Gene Analysis function. The Upstream region is set to -1000, the Downstream region is set to 0, and only bZIP is selected as transcription factor family to be considered. Selection of the Search button starts the analysis and results in a table of all bZIP binding sites annotated to AthaMap. A graphical overview is displayed by selecting the Show Gene Analysis graphical display link resulting in the graphical representation of the seven promoters including the positions of bZIP binding sites [37]. It reveals that most promoters harbor three or four bZIP binding sites with relatively long distances between them (not shown). Knowledge on the specific *cis*-elements as well as their arrangement within promoters will facilitate the design of functional synthetic promoters.

4 Notes

1. PathoPlant and other online databases are being updated on a regular basis and data for the present example were retrieved in December 2015. The results may vary when databases will be queried at a later point in time.
2. For plant species with a lower gene density than *A. thaliana*, effective promoter lengths may exceed 1000 bp upstream of the coregulated genes and should be extended accordingly.

3. Since the additional information added to the Fasta headers of the promoter sequence file will not be needed for the subsequent analysis, it is recommended to trim the Fasta headers and leave the sole locus identifiers by manually removing the information on the chromosomal position and on the length of the sequence (highlighted in Fig. 3 for the first promoter sequence).
4. Apart from the BEST source code version, the other two versions can be installed accordingly by changing the file names within the commands from “BEST” to “BEST1.0” or to “BEST1.0.1”, respectively.
5. By setting the motif length to 5–10 bp when running BEST (Number of columns to align), also motifs longer than 10 bp may be detected in the case that nucleotides neighboring the motif turn out to be conserved as well.
6. Depending on the number and lengths of the sequences and the hardware performance, the screening using BEST takes minutes to hours to be completed. The progress of the screening can be monitored within the terminal window from where BEST was started. To perform multiple screenings at once, several instances of BEST can be run in parallel just by opening an additional terminal window from the BEST directory, typing the command “./BEST”, setting the screening parameters within the BEST GUI and starting the screening by selecting the “RUN” button.
7. The specific motif-finding programs can be deduced from the file names indicated in the last column of the BEST Output table. *meme_* means MEME, *con_* means CONSENSUS, *aa_* means AlignACE, and *biop_* means BioProspector. The directory and the file name also indicate where additional information on the motif can be found. The given file ending in *.opt.sum* (boxed in Fig. 6) contains the same information stated in the Output table. An additional file ending in *.opt.all* contains detailed information on the motif including a nucleotide frequency matrix of the motif and the genes, positions, orientations, and the sequences which the motif is based on.

Acknowledgements

This work was supported by the Federal Ministry for Education and Research of Germany (BMBF) through grants 0315037B and 0315459A.

References

1. Hehl R, Wingender E (2001) Database-assisted promoter analysis. *Trends Plant Sci* 6:251–255
2. Liu W, Stewart CN Jr (2015) Plant synthetic promoters and transcription factors. *Curr Opin Biotechnol* 37:36–44. doi:[10.1016/j.copbio.2015.10.001](https://doi.org/10.1016/j.copbio.2015.10.001)
3. Hehl R, Bülow L (2008) Internet resources for gene expression analysis in *Arabidopsis thaliana*. *Curr Genomics* 9:375–380
4. Koschmann J, Machens F, Becker M, Niemeyer J, Schulze J, Bülow L, Stahl DJ, Hehl R (2012) Integration of bioinformatics and synthetic promoters leads to the discovery of novel elicitor-responsive cis-regulatory sequences in *Arabidopsis*. *Plant Physiol* 160:178–191. doi:[10.1104/pp.112.198259](https://doi.org/10.1104/pp.112.198259)
5. Dubos C, Kelemen Z, Sebastian A, Bülow L, Huep G, Xu W, Grain D, Salsac F, Brousse C, Lepiniec L, Weisshaar B, Contreras-Moreira B, Hehl R (2014) Integrating bioinformatic resources to predict transcription factors interacting with cis-sequences conserved in co-regulated genes. *BMC Genomics* 15(1):317. doi:[10.1186/1471-2164-15-317](https://doi.org/10.1186/1471-2164-15-317)
6. Redman JC, Haas BJ, Tanimoto G, Town CD (2004) Development and evaluation of an *Arabidopsis* whole genome Affymetrix probe array. *Plant J* 38(3):545–561. doi:[10.1111/j.1365-313X.2004.02061.x](https://doi.org/10.1111/j.1365-313X.2004.02061.x)
7. Ma L, Chen C, Liu X, Jiao Y, Su N, Li L, Wang X, Cao M, Sun N, Zhang X, Bao J, Li J, Pedersen S, Bolund L, Zhao H, Yuan L, Wong GK, Wang J, Deng XW, Wang J (2005) A microarray analysis of the rice transcriptome and its comparison to *Arabidopsis*. *Genome Res* 15(9):1274–1283. doi:[10.1101/gr.3657405](https://doi.org/10.1101/gr.3657405)
8. Wang Z, Gerstein M, Snyder M (2009) RNA-Seq: a revolutionary tool for transcriptomics. *Nat Rev Genet* 10(1):57–63. doi:[10.1038/nrg2484](https://doi.org/10.1038/nrg2484)
9. Reiser L, Rhee SY (2005) Using the *Arabidopsis* Information Resource (TAIR) to find information about *Arabidopsis* genes. *Curr Protoc Bioinformatics* Chapter 1(1):Unit 1.11
10. Craighton DJ, James N, Okyere J, Higgins J, Jotham J, May S (2004) NASCArrays: a repository for microarray data generated by NASC's transcriptomics service. *Nucleic Acids Res* 32:D575–D577
11. Barrett T, Willhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 41(Database issue):D991–D995. doi:[10.1093/nar/gks1193](https://doi.org/10.1093/nar/gks1193)
12. Bülow L, Schindler M, Choi C, Hehl R (2004) PathoPlant: a database on plant-pathogen interactions. *In Silico Biol* 4:529–536
13. Bülow L, Schindler M, Hehl R (2007) PathoPlant: a platform for microarray expression data to analyze co-regulated genes involved in plant defense responses. *Nucleic Acids Res* 35:D841–D845
14. Hehl R, Bolívar JC, Koschmann J, Brill Y, Bülow L (2013) Databases and web-tools for gene expression analysis in *Arabidopsis thaliana*. In: Neri C (ed) *Advances in genome science: probing intracellular regulation*, vol 2. Bentham Science Publishers, Sharjah, UAE, pp 176–193. doi:[10.2174/97816080575661130201](https://doi.org/10.2174/97816080575661130201)
15. Kilian J, Whitehead D, Horak J, Wanke D, Weigl S, Batistic O, D'Angelo C, Bornberg-Bauer E, Kudla J, Harter K (2007) The AtGenExpress global stress expression data set: protocols, evaluation and model data analysis of UV-B light, drought and cold stress responses. *Plant J* 50(2):347–363
16. The *Arabidopsis* Genome Initiative (2000) Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*. *Nature* 408(6814):796–815
17. Michael TP, VanBuren R (2015) Progress, challenges and the future of crop genomes. *Curr Opin Plant Biol* 24:71–81. doi:[10.1016/j.pbi.2015.02.002](https://doi.org/10.1016/j.pbi.2015.02.002)
18. Korkuc P, Schippers JH, Walther D (2014) Characterization and identification of cis-regulatory elements in *Arabidopsis* based on single-nucleotide polymorphism information. *Plant Physiol* 164(1):181–200. doi:[10.1104/pp.113.229716](https://doi.org/10.1104/pp.113.229716)
19. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E (2012) The *Arabidopsis* Information Resource (TAIR): improved gene annotation and new tools. *Nucleic Acids Res* 40(Database issue):D1202–D1210. doi:[10.1093/nar/gkr1090](https://doi.org/10.1093/nar/gkr1090)
20. Che D, Jensen S, Cai L, Liu JS (2005) BEST: binding-site estimation suite of tools. *Bioinformatics* 21(12):2909–2911
21. Bailey TL, Elkan C (1995) The value of prior knowledge in discovering motifs with

- MEME. Proc Int Conf Intell Syst Mol Biol 3:21–29
22. Roth FP, Hughes JD, Estep PW, Church GM (1998) Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. Nat Biotechnol 16(10):939–945
 23. Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. Bioinformatics 15(7–8):563–577
 24. Liu X, Brutlag DL, Liu JS (2001) BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes. Pac Symp Biocomput:127–138
 25. Jensen ST, Liu JS (2004) BioOptimizer: a Bayesian scoring function approach to motif discovery. Bioinformatics 20(10):1557–1564
 26. Wray GA, Hahn MW, Abouheif E, Balhoff JP, Pizer M, Rockman MV, Romano LA (2003) The evolution of transcriptional regulation in eukaryotes. Mol Biol Evol 20(9):1377–1419. doi:10.1093/molbev/msg140
 27. Kanhere A, Bansal M (2005) Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. Nucleic Acids Res 33(10):3165–3175. doi:10.1093/nar/gki627
 28. Giuliano G, Pichersky E, Malik VS, Timko MP, Scolnik PA, Cashmore AR (1988) An evolutionarily conserved protein binding sequence upstream of a plant light-regulated gene. Proc Natl Acad Sci U S A 85(19):7089–7093
 29. Mahony S, Benos PV (2007) STAMP: a web tool for exploring DNA-binding motif similarities. Nucleic Acids Res 35(Web Server issue):W253–W258
 30. Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E (2003) AGRIS: Arabidopsis Gene Regulatory Information Server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. BMC Bioinformatics 4(1):25
 31. Hehl R, Norval L, Romanov A, Bülow L (2016) Boosting AthaMap Database content with data from protein binding microarrays. Plant Cell Physiol 57(1), e4. doi:10.1093/pcp/pcv156
 32. Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. Nucleic Acids Res 27(1):297–300
 33. Guiltinan MJ, Marcotte WR Jr, Quatrano RS (1990) A plant leucine zipper protein that recognizes an abscisic acid response element. Science 250(4978):267–271
 34. Iwasaki T, Yamaguchi-Shinozaki K, Shinozaki K (1995) Identification of a cis-regulatory region of a gene in Arabidopsis thaliana whose induction by dehydration is mediated by abscisic acid and requires protein synthesis. Mol Gen Genet 247(4):391–398
 35. Jakoby M, Weisshaar B, Dröge-Laser W, Vicente-Carbajosa J, Tiedemann J, Kroj T, Parcy F (2002) bZIP transcription factors in Arabidopsis. Trends Plant Sci 7(3):106–111
 36. Galuschka C, Schindler M, Bülow L, Hehl R (2007) AthaMap web-tools for the analysis and identification of co-regulated genes. Nucleic Acids Res 35:D857–D862
 37. Bülow L, Engelmann S, Schindler M, Hehl R (2009) AthaMap, integrating transcriptional and post-transcriptional data. Nucleic Acids Res 37(Database issue):D983–D986

In Silico Expression Analysis

Julio Bolívar, Reinhard Hehl, and Lorenz Bülow

Abstract

Information on the specificity of *cis*-sequences enables the design of functional synthetic plant promoters that are responsive to specific stresses. Potential *cis*-sequences may be experimentally tested, however, correlation of genomic sequence with gene expression data enables an in silico expression analysis approach to bioinformatically assess the stress specificity of candidate *cis*-sequences prior to experimental verification. The present chapter demonstrates an example for the in silico validation of a potential *cis*-regulatory sequence responsive to cold stress. The described online tool can be applied for the bioinformatic assessment of *cis*-sequences responsive to most abiotic and biotic stresses of plants. Furthermore, a method is presented based on a reverted in silico expression analysis approach that predicts highly specific potentially functional *cis*-regulatory elements for a given stress.

Key words In silico validation, *Cis*-regulatory sequences, *Cis*-elements, Expression analysis, Promoter sequences, PathoPlant

1 Introduction

Specific gene expression enables organisms to respond to endogenous and environmental cues by modifying growth, metabolism, and developmental processes. Gene expression in eukaryotes is regulated at the transcriptional level by the binding of transcription factors to specific *cis*-regulatory sequences, which constitute short sequence motifs mainly located within the promoters of the target genes. In plants, such motifs when being overrepresented in promoters of coregulated genes can be identified as potential *cis*-regulatory sequences [1]. This approach has been, for example, applied to promoter sequences of *Arabidopsis thaliana* genes expressed under specific biotic and abiotic stress conditions [2, 3].

Signal transduction is mainly triggered by MAP kinase cascades that are involved in many different responses to various biotic and abiotic stresses, in hormone signaling, in cell division, and in developmental processes. However, the genome of *A. thaliana* harbors only 20 MAP kinases, 10 MAP kinase kinases, and 60 MAPK kinase kinase kinases [4] resulting in a signal transduction

bottleneck and in the necessity for the plant to converge signaling pathways. Pathway crosstalks and overlaps in plants have been reported by identifying several transcription factors and kinases playing roles in different signaling pathways [5]. Crosstalks occur between salicylic acid, jasmonic acid, ethylene, and other phytohormones signaling pathways [6]. Also, pathways related to abiotic stress tolerance and biotic resistance significantly overlap [7]. Thus, different biotic and abiotic stresses trigger expression of overlapping gene sets, and in consequence, overrepresented motifs in promoters of genes coregulated under a specific stress are likely to be potential *cis*-regulatory elements responsive not only to the given stress but also to further stress conditions. This should be considered when employing *cis*-regulatory elements to engineer synthetic plant promoters.

The present chapter provides a protocol to bioinformatically propose the specific stresses a given *cis*-regulatory element may be responsive to. The method is based on the correlation of promoter sequence data with expression data from a wide array of stresses. It is called “in silico expression analysis” and has been implemented with *A. thaliana* sequence information and microarray expression data as an online tool of the PathoPlant database [8, 9]. Furthermore, a method is presented that predicts potentially functional *cis*-regulatory elements highly specific for a given stress by a reverted in silico expression analysis approach.

2 In Silico Expression Analysis

Potential *cis*-regulatory elements may be experimentally tested for specificity with reporter genes but this is quite laborious and time-consuming. The in silico expression analysis constitutes a tool to bioinformatically propose the function of given *cis*-sequences by correlating occurrences of the submitted *cis*-sequences within promoters with the corresponding gene expression data from *Arabidopsis thaliana*. The tool identifies all genes harboring a submitted sequence within a defined promoter region and compares the expression of these genes with an array of stress-related expression data. This results in a ranking of abiotic and biotic stress conditions to which these genes are most likely responsive [9]. The in silico expression analysis constitutes an online tool of the PathoPlant database and is available at http://www.pathoplant.de/expression_analysis.php.

The sequence TACCGACAT corresponds to the Drought-Responsive Element (DRE) that is a well-described element responsive to cold, drought, and salt stress [10]. This should be kept as is because the demo sequence is used in both, the in silico expression analysis online tool as well as the present chapter. When accessing http://www.pathoplant.de/expression_analysis.php, the input form

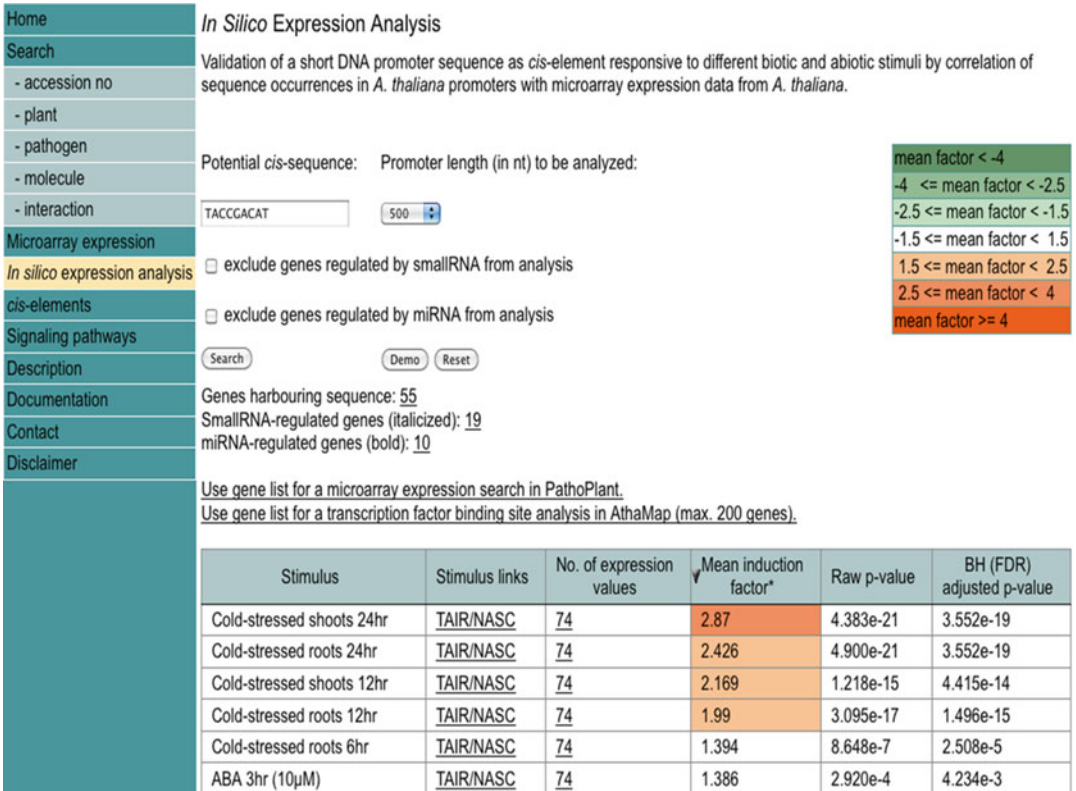


Fig. 1 In silico expression analysis of the *cis*-sequence TACCGACAT with default settings

of the in silico expression analysis will be displayed (upper part of Fig. 1). The sequence TACCGACAT is typed or pasted into the field for the potential *cis*-sequence. The remaining parameters are left as defaults with a promoter length of 500 nt and not excluding genes regulated by smallRNA or by miRNA from the analysis. The analysis starts by selecting the Search button and a table listing the results is displayed (lower part of Fig. 1). The table lists the stresses (Stimulus) and the means of the corresponding induction factors of the genes that harbor the submitted sequence within the given promoter region. By default, the table is sorted by the mean induction factors and displays the stresses with the highest induction at the top of the table. These stresses also exhibit the lowest Student's *t*-test raw *p*-values and Benjamini-Hochberg (BH) false discovery rate (FDR) adjusted *p*-values [11] indicating the most probable stresses the submitted sequence is associated with when occurring within promoters. The table can be resorted by the stress (Stimulus), mean induction factor (*see* Note 1), raw *p*-value and BH (FDR) adjusted *p*-value by selecting the respective column header. As expected, cold stress conditions are the most probable stresses the submitted sequence is associated with followed by an ABA (abscisic acid) stress condition (Fig. 1) and further abiotic

stresses with a lower likelihood (not shown). This indicates that the given sequence may be highly specific for cold stress responsiveness and less responsive to other abiotic stresses. Crosstalk with biotic stress responses seems not to play a role in the case of the submitted sequence (*see Note 2*).

The submitted sequence is present within the promoter regions of 55 genes. This number is given above the result table (Fig. 1, *see Note 3*) and can be selected in order to display these 55 genes. This table of genes (not shown) can be further expanded to display the gene descriptions. 19 genes are annotated as potentially regulated by smallRNAs [12] and 10 genes potentially regulated by microRNAs [13] (Fig. 1). These genes are shown in the gene tables in italics (smallRNAs) and in bold (microRNAs), respectively (not shown). The corresponding gene lists can be displayed by selecting the corresponding numbers of genes (Fig. 1). These genes sets can also be excluded from the *in silico* expression analysis by selecting the corresponding check boxes to analyze exclusively transcriptionally regulated genes (upper part of Fig. 1). When excluding both gene sets, the number of genes is reduced to 31 and the result table indicates higher mean induction values for cold and other abiotic stresses (data not shown). By selecting a link located above the result table (Fig. 1), the gene list can easily be exported to PathoPlant's microarray expression tool that determines the expression profile of the gene set under additional induction conditions [14]. For further analysis of the promoter architecture, the gene list can be exported to AthaMap's Gene analysis tool [15]. This tool identifies the positions of predicted transcription factor binding sites and graphically reveals their spatial distribution patterns, which is very useful information when designing functional synthetic promoters.

The result table of the *in silico* expression analysis (Fig. 1) furthermore provides links (Stimulus links) to the original source of the expression data of each stress and also states the number of expression values used for calculating the mean induction factors and *p*-values (*see Note 4*). The number of expression values can be selected to show detailed information about the genes and their individual expression values for a given stress. By selecting the "74" of the "Cold-stressed shoots 24hr" stress condition in the first row of the result table shown in Fig. 1, a new window shown in Fig. 2 opens that displays the submitted sequence, the specific stress condition, the selected promoter length, and the number of genes present on the respective microarray. A table with the genes, positional information from the promoter screening, and gene expression details is also displayed (Fig. 2). In this table, the orientation and relative distance refers to the distance of the first match position to the point of reference that can either be the transcription start site (TSS), if known, or otherwise the translation start site (ATG). The individual and mean induction factors of each gene are

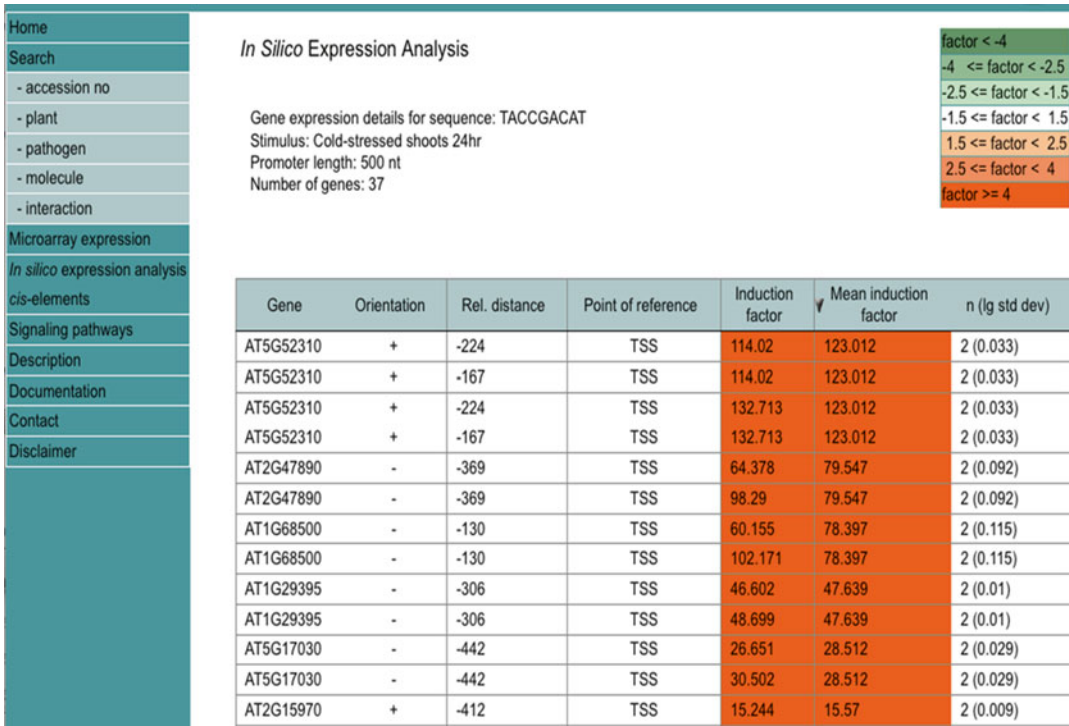


Fig. 2 Gene expression and positional details for the *cis*-sequence TACCGACAT under cold stress (shoots 24hr)

given as well as the number of replicates (n) and the base-10 logarithm of the standard deviation for mean induction factor calculation of each gene. By default, the table entries are sorted by the mean induction factors. By selecting the respective table headers, the table can be resorted by any of the parameters in the columns.

3 Identifying Specific *Cis*-Elements in Stress Responsive Genes

Another approach to identify *cis*-elements specifically associated with abiotic or biotic stress conditions is the selection of a particular stress condition and the identification of *cis*-elements enriched in the promoters of the upregulated genes. The newly developed PathoPlant web tool “*cis*-elements” permits such an analysis. To generate this web tool, a comprehensive analysis using the in silico expression analysis algorithm [9] was conducted for a complete screening of all possible 8mer, 9mer, and 10mer nucleotide sequences [16] (see Note 5). Each sequence was considered to be significantly responsive to a certain stress when displaying a positive mean induction factor and a raw p -value below 10^3 in the in silico expression analysis. These specificities were annotated to the PathoPlant database. It turned out that only a small percentage of

sequences is specifically responsive to a single stress, while most sequences show responsiveness to more than one stress. There are sequences that are responsive to several abiotic stresses only, or others that are exclusively responsive to fungal stresses. Other sequences are less specific showing responsiveness to abiotic and biotic stresses [16]. In PathoPlant, all stresses were categorized in one out of four groups: Abiotic stresses, fungal stresses, biotic stresses excluding fungal ones, and other stresses (*see Note 6*).

Within PathoPlant, the online tool “*cis*-elements” is available at <http://www.pathoplant.de/ciselements.php> and enables a reversion of the *in silico* expression analysis tool described before in order to identify potential *cis*-elements specifically responsive to an individual selected stress or a group of stresses. Figure 3 shows the result with the *cis*-elements tool when “Cold-stressed shoots 24hr (abiotic)” was the selected stress under the default setting. When using the default settings, sequences with lengths of 8, 9, or 10 nt are identified (*see Note 7*). Additionally, selection of default settings will identify sequences not only responsive to the selected stress but that might also be responsive to further stresses belonging to the same group of stresses. The example of Fig. 3 identified 1848 sequences with lengths of 8–10 nt as potentially responsive to the “Cold-stressed shoots 24 hours” stress treatment. These sequences are not exclusively responsive to the cold stress treatment of shoots (24 hours) but may also exhibit responsiveness to further

Home		cis-elements							
Search		Prediction of 8mer, 9mer, 10mer <i>cis</i> -elements responsive to a given stimulus.							
- accession no									
- plant									
- pathogen		Stimulus [Group]: Cold-stressed shoots 24hr (abiotic)							
- molecule		<input checked="" type="radio"/> default settings							
- interaction		<input type="radio"/> advanced settings							
Microarray expression		<input type="text" value="Search"/>		<input type="button" value="Demo"/> <input type="button" value="Reset"/>					
In silico expression analysis		1848 sequences responsive to Cold-stressed shoots 24hr							
cis-elements		Further sequence specificity for abiotic stimuli, not for biotic stimuli, not for fungi stimuli, not for other stimuli							
Signaling pathways		Sequence length: 8-10 nt							
Description		Sequence rev. compl.	Number of genes	Mean	▼ pValue	Number of abiotic stimuli	Number of biotic stimuli	Number of fungi stimuli	Number of other stimuli
Documentation									
Contact		accgacgtg cacgtcgg	36	3.893	1.385e-28	14	0	0	0
Disclaimer		accgacatca tgatgtcgg	21	5.355	1.196e-21	12	0	0	0
		gtcggctca tagaccgac	40	4.046	4.554e-21	6	0	0	0
		cacgtcgg ccgacgtg	119	1.901	9.390e-19	5	0	0	0
		tgtcggtaa ttaccgaca	59	2.679	3.124e-18	6	0	0	0
		atgtcgtca tgaccgacat	22	4.98	2.232e-16	13	0	0	0
		gaccgacata tatgtcggtc	22	4.165	3.363e-15	13	0	0	0
		ccaaaatatac gatattttgg	129	1.762	5.506e-14	5	0	0	0
		ctttgccgac gtcggcaaa	36	2.881	6.437e-14	13	0	0	0
		gatattttca tgaataatatac	178	1.59	1.533e-13	5	0	0	0
		acgtcggc gccgacgt	101	1.872	1.951e-13	6	0	0	0

Fig. 3 Identification of specific *cis*-elements responsive to cold stress (shoots 24hr) with default settings

abiotic stresses. Therefore, only sequences with further cross-responsiveness specificity for abiotic stresses, but not for biotic, fungal, or other stresses (as summarized above the table) are shown with the individual results in the result table (Fig. 3). This is also illustrated by the table displaying the individual results stating the cross-responsivenesses of the identified sequences in the four right columns (“Number of abiotic stimuli”, “Number of biotic stimuli”, “Number of fungi stimuli”, “Number of other stimuli”) by giving the total number of stresses the sequence or its reverse complement sequence (Sequence rev. compl.) is responsive to (Fig. 3). In addition to the sequence and its reverse complement sequence, the table displays the number of genes that harbor the sequence within their promoters, the mean induction factor (mean) for the selected stress as well as the raw p -value calculated by the in silico expression analysis. The most probable specific sequences appear at the top of the table as it is sorted by the p -values by default, but it can be resorted by any other of the column parameters by selecting the corresponding column header.

The advanced settings of the “*cis*-elements” online tool can be used to identify potential *cis*-sequences that are highly specific to only one given stress. When selecting “Cold-stressed shoots 24hr (abiotic)” as stress and selecting the advanced settings, “exclusive for selected stimulus” can be selected (all “Further sequence specificities” shall be deselected before), and when leaving 8–10 nt as sequence length, 728 sequences exclusively responsive to the stress condition “Cold-stressed shoots 24hr” will be displayed (Fig. 4). The table shows that all sequences are responsive to only one abiotic stress, i.e. the submitted cold stress (shoots 24hr), and are not responsive to any further stress (Fig. 4, right columns of the table).

The *cis*-elements online tool’s advanced settings may also be employed to search for sequences that display an intended cross-specificity in the responsiveness with other stresses. For example, abiotic stresses are associated with the plant hormone ABA [17, 18], and the ABA stress treatment belongs to the group of the other stresses within the *cis*-elements online tool. When selecting “Cold-stressed shoots 24hr (abiotic)” as stress and selecting abiotic stress (abiotic stimuli) and additionally other stresses (other stimuli) as further sequence specificity as well as leaving the 8–10 nt as sequence length (Fig. 5), 565 sequences will be displayed that are responsive to the selected cold stress condition and are also specific for at least one of the other stresses (Fig. 5), a group which is mainly comprised of plant hormone stress treatments (*see Note 6*). The sequence TACCGACAT that was described as sample sequence of the in silico expression analysis before is found among the resulting sequences (highlighted in Fig. 5). The sequences of the table are directly linked to the in silico expression analysis tool to display the individual stresses the sequence is responsive to. Selecting the

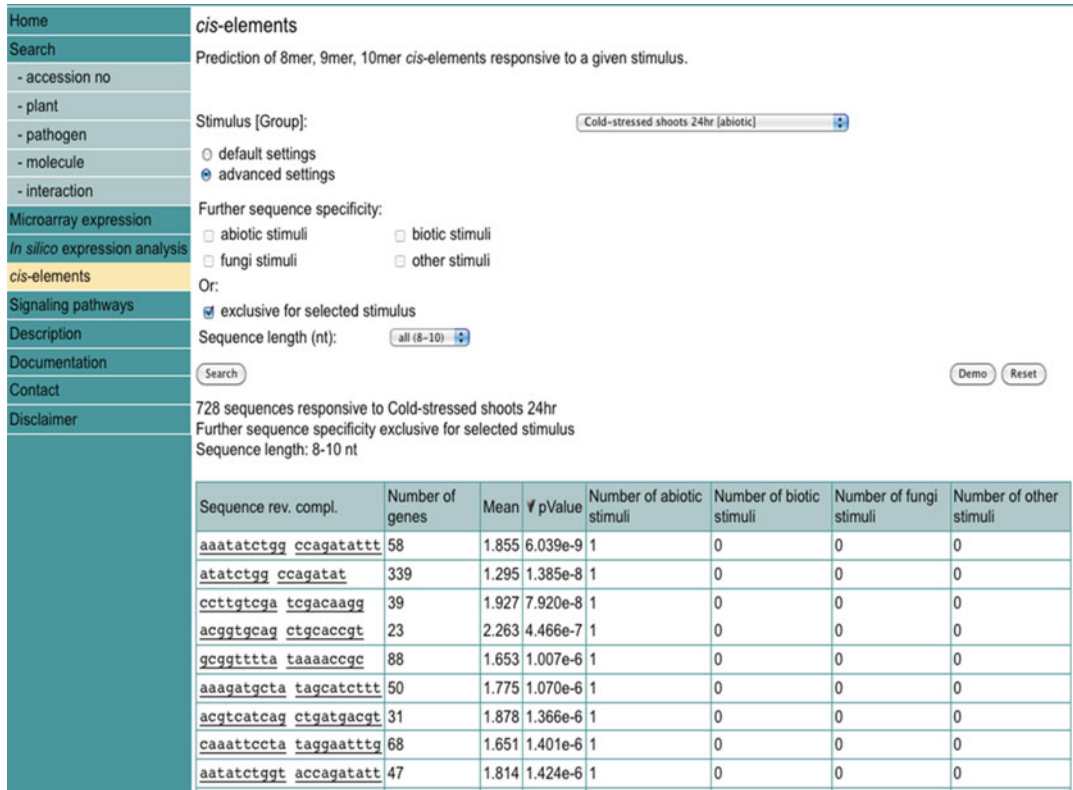


Fig. 4 Identification of specific *cis*-elements responsive to cold stress (shoots 24hr) with settings for highest specificity

sequence TACCGACAT will result in the identification of cold stresses, ABA stress, and further abiotic stresses as shown in Fig. 1.

Although the sequences of the *cis*-elements online tool is based on the in silico expression analysis of all possible 8mer, 9mer, and 10mer nucleotide sequences, the tool only identifies those sequences occurring within *Arabidopsis thaliana* promoters that result in gene sets for which the stress-specific expression can be determined.

4 Notes

1. The mean values of the result table of the in silico expression analysis are normalized in order to adjust the varying overall expression levels resulting from the individual experimental stress conditions. A table indicating the normalization factors of the stress experiments can be displayed by selecting a link at the bottom of the result table (not shown).
2. Although PathoPlant’s expression data cover a wide array of 146 different stresses, it cannot be excluded that there may

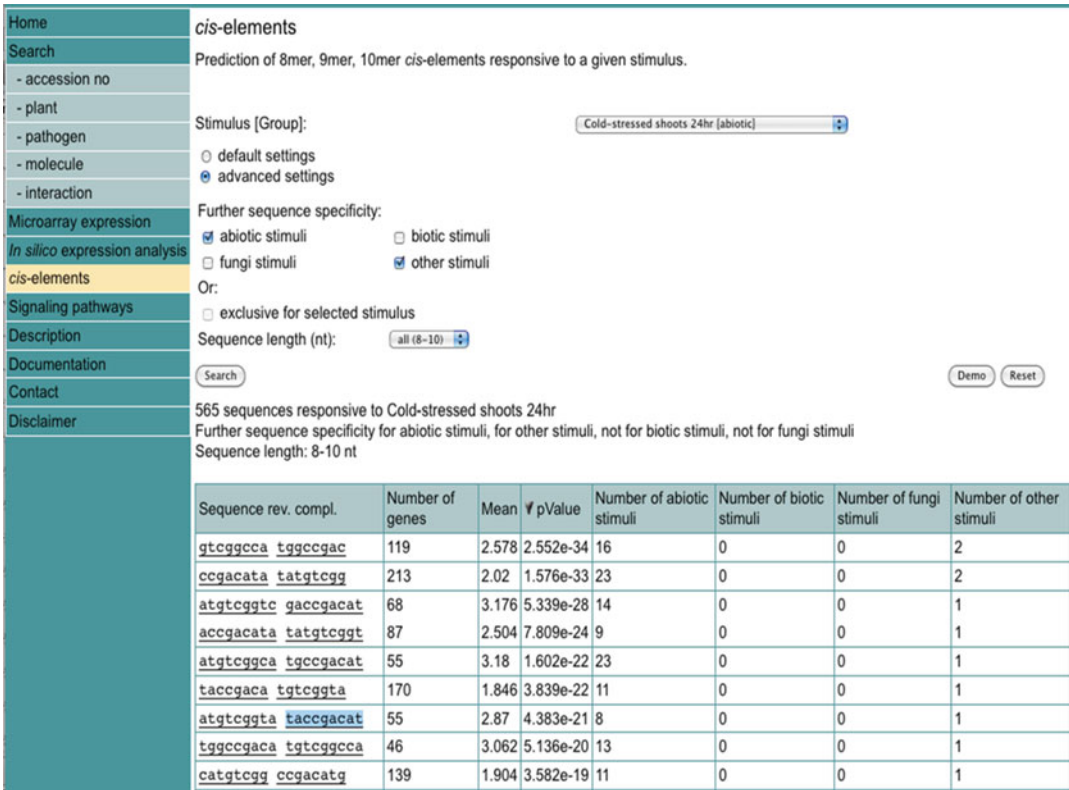


Fig. 5 Identification of specific *cis*-elements responsive to cold stress (shoots 24hr) with settings for cross-responsiveness to other stresses

exist crosstalk with further stresses that have so far not been experimentally analyzed or have not yet been annotated. The Documentation of PathoPlant lists all microarray experiments available for the analysis at http://www.pathoplant.de/documentation_microarrays.php.

- When shorter or AT-rich sequences are submitted to the in silico expression analysis, promoters of too many genes may be identified resulting in an unreliable analysis due to the reduced stress specificity of large gene sets. In such a case, the promoter length shall be reduced to 250 nt to yield fewer genes to be analyzed. In the opposite case, the number of genes to be identified can be increased by selecting 1000 nt as promoter length.
- The number of expression values used to perform the in silico expression analysis often exceeds the number of genes with promoters harboring the submitted sequence due to experimental replicates. In the example shown in Fig. 1, the 55 genes that harbor the sequence to be analyzed within their promoters correspond to 74 expression values for most stresses. These replicates increase the analysis accuracy since

most gene-stress combinations are represented by more than one expression value.

5. For 8mer nucleotide sequences, $4^8 = 65,536$ possibilities, for 9mers $4^9 = 262,144$, and for 10mers $4^{10} = 1,048,576$ possible sequences exist. As the *in silico* expression analysis considers occurrences of sequences in promoters in sense and antisense orientation, direct and reverse complement sequences yield the same results which halves the number of sequences to be screened. For the screening, the promoter length to be analyzed was set to 500 nt.
6. Fungal stresses outgroup from other biotic stresses and were consequently classified separately. The group of other biotic stresses is comprised of bacterial and viral stresses. Abiotic stresses are primarily drought, cold, salt, osmotic, and heavy metal stresses. Mainly plant hormones belong to the group of other stresses. The classification of the stresses is stated within squared brackets in the “Stimulus (Group)” drop-down list of the *cis-elements* tool (upper part of Fig. 3).
7. To compile the data for the *cis-elements* online tool, only 8mer, 9mer, and 10mer nucleotide sequences were screened since *cis*-regulatory elements are typically eight to ten nucleotides long. The online tool allows the identification of sequences with all these lengths (8–10 nt) by default, but the search can also be restricted to a length of 8, 9, or 10 nucleotides in order to yield length-specific sequences.

Acknowledgements

The authors thank Artyom Romanow for implementation of the “*cis-elements*” web interface. This work was supported by the Federal Ministry for Education and Research of Germany (BMBF) through grants 0315037B and 0315459A.

References

1. Hehl R, Bülow L (2008) Internet resources for gene expression analysis in *Arabidopsis thaliana*. *Curr Genomics* 9:375–380
2. Koschmann J, Machens F, Becker M, Niemeyer J, Schulze J, Bülow L, Stahl DJ, Hehl R (2012) Integration of bioinformatics and synthetic promoters leads to the discovery of novel elicitor-responsive *cis*-regulatory sequences in *Arabidopsis*. *Plant Physiol* 160:178–191. doi:10.1104/pp.112.198259
3. Dubos C, Kelemen Z, Sebastian A, Bülow L, Huep G, Xu W, Grain D, Salsac F, Brousse C, Lepiniec L, Weisshaar B, Contreras-Moreira B, Hehl R (2014) Integrating bioinformatic resources to predict transcription factors interacting with *cis*-sequences conserved in co-regulated genes. *BMC Genomics* 15(1):317. doi:10.1186/1471-2164-15-317
4. Rodriguez MC, Petersen M, Mundy J (2010) Mitogen-activated protein kinase signaling in plants. *Annu Rev Plant Biol* 61:621–649. doi:10.1146/annurev-arplant-042809-112252
5. Fujita M, Fujita Y, Noutoshi Y, Takahashi F, Narusaka Y, Yamaguchi-Shinozaki K, Shinozaki K (2006) Crosstalk between abiotic and biotic stress responses: a current view from the points

- of convergence in the stress signaling networks. *Curr Opin Plant Biol* 9(4):436–442. doi:[10.1016/j.pbi.2006.05.014](https://doi.org/10.1016/j.pbi.2006.05.014)
6. Bostock RM (2005) Signal crosstalk and induced resistance: straddling the line between cost and benefit. *Annu Rev Phytopathol* 43:545–580. doi:[10.1146/annurev.phyto.41.052002.095505](https://doi.org/10.1146/annurev.phyto.41.052002.095505)
 7. Mauch-Mani B, Mauch F (2005) The role of abscisic acid in plant-pathogen interactions. *Curr Opin Plant Biol* 8(4):409–414. doi:[10.1016/j.pbi.2005.05.015](https://doi.org/10.1016/j.pbi.2005.05.015)
 8. Bülow L, Schindler M, Choi C, Hehl R (2004) PathoPlant: a database on plant-pathogen interactions. *In Silico Biol* 4:529–536
 9. Bolívar JC, Machens F, Brill Y, Romanov A, Bülow L, Hehl R (2014) ‘*In silico* expression analysis’, a novel PathoPlant web-tool to identify abiotic and biotic stress conditions associated with specific *cis*-regulatory sequences. *Database (Oxford)* 2014:bau030
 10. Yamaguchi-Shinozaki K, Shinozaki K (1994) A novel *cis*-acting element in an *Arabidopsis* gene is involved in responsiveness to drought, low-temperature, or high-salt stress. *Plant Cell* 6(2):251–264. doi:[10.1105/tpc.6.2.251](https://doi.org/10.1105/tpc.6.2.251)
 11. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testings. *J R Stat Soc B* 57(1):289–300
 12. Bülow L, Engelmann S, Schindler M, Hehl R (2009) AthaMap, integrating transcriptional and post-transcriptional data. *Nucleic Acids Res* 37:D983–D986
 13. Bülow L, Bolívar JC, Ruhe J, Brill Y, Hehl R (2012) ‘MicroRNA Targets’, a new AthaMap web-tool for genome-wide identification of miRNA targets in *Arabidopsis thaliana*. *BioData Min* 5:7
 14. Bülow L, Schindler M, Hehl R (2007) PathoPlant: a platform for microarray expression data to analyze co-regulated genes involved in plant defense responses. *Nucleic Acids Res* 35:D841–D845
 15. Galuschka C, Schindler M, Bülow L, Hehl R (2007) AthaMap web-tools for the analysis and identification of co-regulated genes. *Nucleic Acids Res* 35:D857–D862
 16. Bolívar JC (2014) *In silico* expression analysis to identify potentially functional plant *cis*-regulatory elements. Dissertation, Technische Universität Braunschweig, Braunschweig
 17. Yoshida T, Fujita Y, Sayama H, Kidokoro S, Maruyama K, Mizoi J, Shinozaki K, Yamaguchi-Shinozaki K (2010) AREB1, AREB2, and ABF3 are master transcription factors that cooperatively regulate ABRE-dependent ABA signaling involved in drought stress tolerance and require ABA for full activation. *Plant J* 61(4):672–685. doi:[10.1111/j.1365-313X.2009.04092.x](https://doi.org/10.1111/j.1365-313X.2009.04092.x)
 18. Kim JS, Mizoi J, Yoshida T, Fujita Y, Nakajima J, Ohori T, Todaka D, Nakashima K, Hirayama T, Shinozaki K, Yamaguchi-Shinozaki K (2011) An ABRE promoter sequence is involved in osmotic stress-responsive expression of the DREB2A gene, which encodes a transcription factor regulating drought-inducible genes in *Arabidopsis*. *Plant Cell Physiol* 52(12):2136–2146. doi:[10.1093/pcp/pcr143](https://doi.org/10.1093/pcp/pcr143)

Chapter 17

FootprintDB: Analysis of Plant *Cis*-Regulatory Elements, Transcription Factors, and Binding Interfaces

Bruno Contreras-Moreira and Alvaro Sebastian

Abstract

FootprintDB is a database and search engine that compiles regulatory sequences from open access libraries of curated DNA *cis*-elements and motifs, and their associated transcription factors (TFs). It systematically annotates the binding interfaces of the TFs by exploiting protein–DNA complexes deposited in the Protein Data Bank. Each entry in footprintDB is thus a DNA motif linked to the protein sequence of the TF(s) known to recognize it, and in most cases, the set of predicted interface residues involved in specific recognition. This chapter explains step-by-step how to search for DNA motifs and protein sequences in footprintDB and how to focus the search to a particular organism. Two real-world examples are shown where this software was used to analyze transcriptional regulation in plants. Results are described with the aim of guiding users on their interpretation, and special attention is given to the choices users might face when performing similar analyses.

Key words Bioinformatics, Transcription factor, DNA binding, DNA motif, PSSM, Protein domain, Promoter, *Cis*-element, Database, Open-access

1 Introduction

Transcription is a central process in gene expression. It is modulated primarily by the binding of regulatory proteins called transcription factors (TFs) to short DNA sequences, called *cis*-regulatory elements. DNA recognition is a flexible mechanism, since most TFs can usually distinguish a collection of non-identical DNA binding sites (DBSs), which in turn define a DNA-binding motif (DBM). Position-specific scoring matrices (PSSMs) are a common way of representing DBMs, which in their simplest form tally the observed nucleotide frequencies at each position of the motif [1]. DBMs are also frequently plotted as sequence logos, which graphically summarize the binding preferences of TFs [2], as shown in Fig. 1. These are convenient models that hide some known complexities of TFs but are still useful. For instance, columns in a DBM might be correlated and thus not accurately modeled by a PSSM,

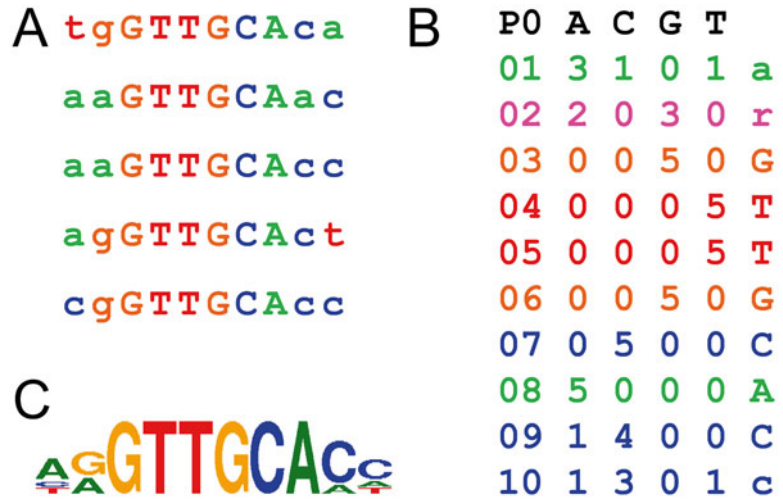


Fig. 1 Typical representations of DNA motifs. (a) Multiple alignment of DNA binding sites recognized by a TF, usually cis-elements located in different promoters. The motif core is in upper-case. (b) Position-specific scoring matrix in TRANSFAC-like notation. The last column is the consensus. (c) Sequence logo, with base heights proportional to conservation across sites

which usually features independent columns. Moreover, the alignment of cis-elements in order to derive motifs has several pitfalls: (1) both DNA strands must be considered; (2) short DBSs are easily misaligned, particularly if structural constraints are not considered [3]; and (3) mismatches are common due to TF binding plasticity. Nevertheless, because of their simplicity, PSSMs are usually the preferred representation of protein–DNA binding models.

Experimental methods to identify DBSs are technically challenging and have been traditionally limited to determining cis-regulatory sites for one TF at a time. Among such protocols are DNA footprinting, chromatin immunoprecipitation (ChIP) or electrophoretic mobility shift assays, which yield high quality data despite their low throughput [4–6]. These approaches are being replaced by higher throughput protocols such as protein binding microarrays, HT-SELEX, ChIP-chip, or ChIP-Seq techniques [7–10]. These procedures often produce large volumes of raw sequence data, which must be pre-processed and filtered in order to derive DBMs employing a variety of recipes [11, 12]. Eventually, resulting PSSMs are collected and annotated in databases.

In addition, a number of algorithms have been developed in order to predict and annotate DBSs within genomic sequences. Some of them try to discover them *de novo* by detecting overrepresented DNA motifs [13–15]. Others use previously known experimental DNA binding data to localize similar regions in genomes by sequence and PSSM alignments or machine learning techniques [16–19].

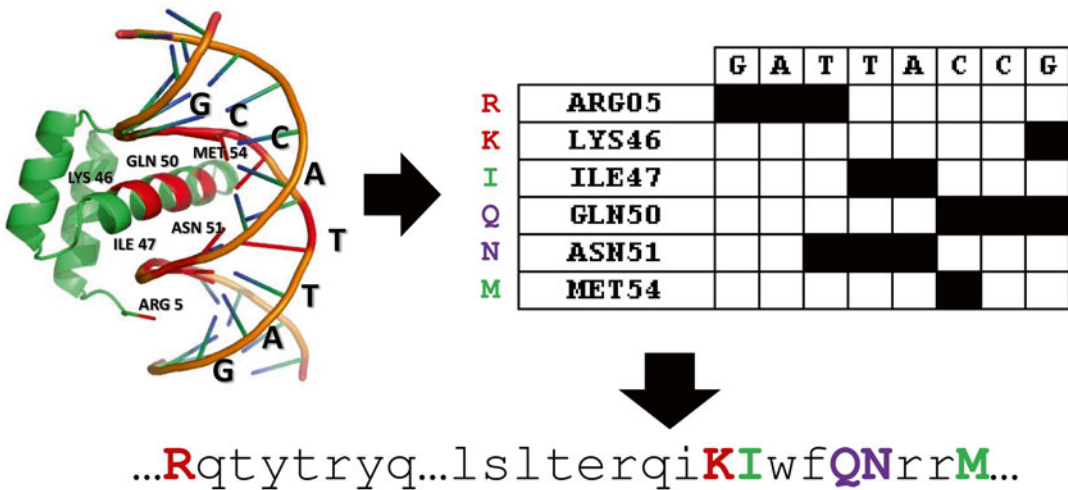


Fig. 2 DNA-binding interface of PDB complex 9ANT (Antennapedia in complex with a cis-element) as annotated by 3D-footprint. Inter-atomic distances are calculated among atoms of amino acid side chains and nitrogen bases, and a matrix of interactions generated. Interface residues, in upper case, are extracted to show the DNA-binding interface core

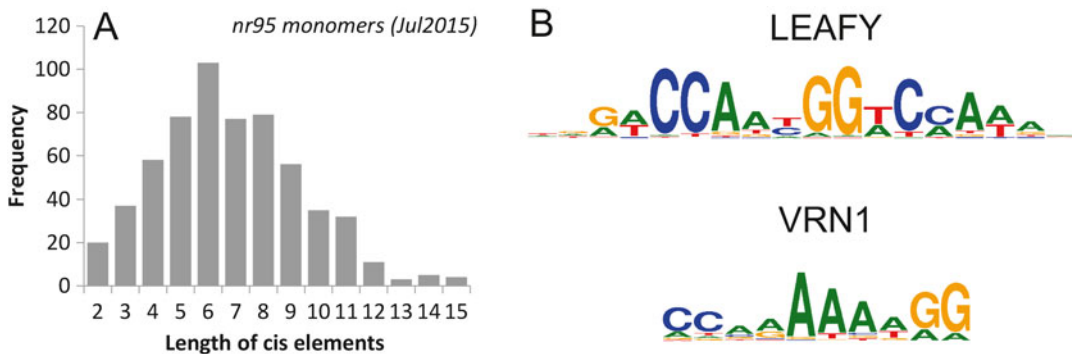


Fig. 3 (a) Histogram of the length of cis-elements recognized by non-redundant monomeric proteins in TF-DNA complexes deposited in the PDB and annotated in 3D-footprint as of July 2015. (b) Sequence logos of experimentally determined binding sites of plant transcription factors LEAFY (*top*) and VRN1 (*bottom*) (available in footprintDB). Note that both PSSMs contain highly conserved sub-motifs with interleaved degenerate sequences

Other experimental approaches focus on characterizing the interface residues of TFs, those in charge of recognizing the nucleotide bases of DBSs (Fig. 2). Besides site-directed mutagenesis [20, 21], the most accurate methods are X-ray crystallography and NMR studies of protein-DNA complexes. The resulting structures are maintained and published at the Protein Data Bank (PDB), and can also be exploited to infer structure-based DBMs [22–27]. An exhaustive analysis of these complexes shows that individual DNA-binding proteins typically bind a nucleotide segment three to ten bases long (Fig. 3a). However, TFs usually identify target sites in conjunction with other proteins. For this reason, biologically

relevant motifs most often correspond to protein multimers or multi-domain TFs, which bind longer, contiguous regions in the DNA sequence as shown in Fig. 3b for two plant TFs.

To facilitate the analysis of DNA–protein interactions, researchers can take advantage of in silico tools for designing experiments and engineering DNA-recognition. In this chapter, we describe footprintDB, a database that compiles experimental data of thousands of TFs and their DNA motifs. FootprintDB has two main applications: (1) the prediction of TFs able to bind novel DNA motifs and (2) the prediction of DNA motifs for uncharacterized TFs. The first search type is illustrated with two protocols and their application to a real research problem concerning transcription factors predicted to regulate a set of co-expressed *Arabidopsis thaliana* promoters. For the second kind of query a generic protocol is also presented and then applied to the study of a stress-related promoter sequence in rice.

2 Materials

The only resources required to replicate the analysis described in this chapter are an Internet connection and a web browser. These will suffice to learn how to use footprintDB and associated tools, which are now presented.

2.1 FootprintDB

FootprintDB is a meta-database that integrates several open access repositories of curated cis-elements, DNA motifs, and TFs into a unique repository (<http://floresta.cead.csic.es/footprintdb>, Fig. 4) [28]. The May 2015 release includes the following databases:

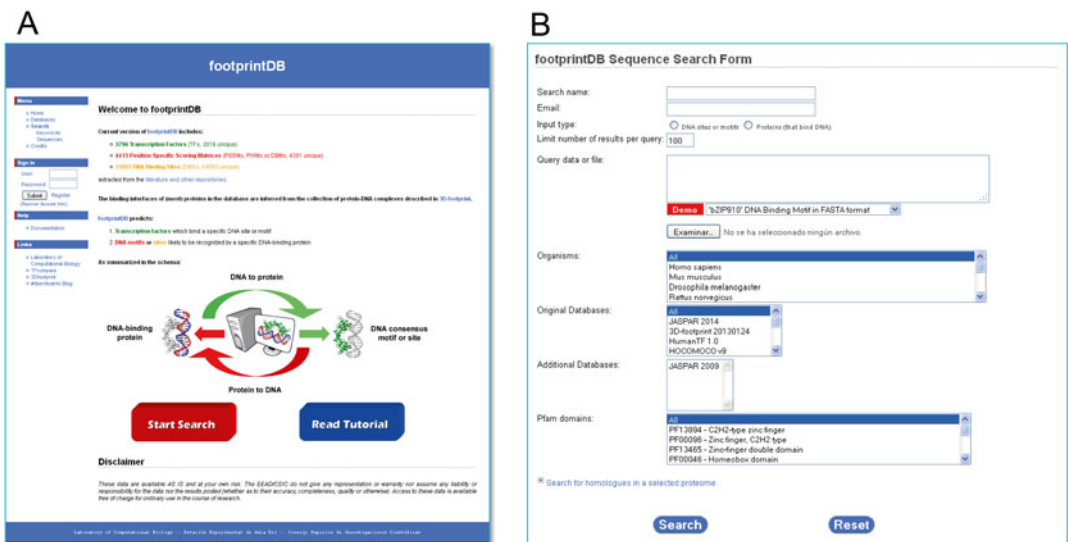


Fig. 4 FootprintDB web interface. (a) Main page. (b) Search form

JASPAR, 3D-footprint, HumanTF, HOCOMOCO, FlyZincFinger, UniPROBE, ArabidopsisPBM, AthaMap, DBTBS, RegulonDB, DrosophilaTF, and EEADannot. JASPAR is the reference source of curated TFs and DNA motifs derived from published collections of experimentally defined DBS for eukaryotes [29]. 3D-footprint annotates cis-elements captured in protein–DNA complexes deposited in the PDB [26]. UniPROBE (Universal PBM Resource for Oligonucleotide Binding Evaluation) hosts data generated by universal protein binding microarrays (PBM) with proteins from a diverse collection of organisms [30]. The remaining repositories provide experimentally supported DBMs for specific organisms and taxa, such as HumanTF and HOCOMOCO for human [31, 32], RegulonDB and DBTBS for bacteria [33, 34], AthaMap and ArabidopsisPBM for *A. thaliana* [35, 36], or DrosophilaTF and FlyZincFinger for fruit fly [37, 38]. Finally, EEADannot is a manually curated set of plant data compiled in our laboratory.

Available DNA-binding data for plant TFs are scarce, and for this reason AthaMap and ArabidopsisPBM collections, as well as EEADannot, are valuable resources for plant promoter analysis and cis-element discovery. We note that commercial database TRANSFAC annotates also a repertoire of plant DNA motifs [39], but a subscription fee is required. Other valuable plant-specific resources such as PLACE and AGRIS were considered. However, while the former contains single DBSs without annotated binding TFs, most of the data in the latter are already annotated by other resources like JASPAR and AthaMap [40, 41].

FootprintDB handles redundant data deposited in several repositories by annotating unique entries with multiple references to the original sources. The underlying database can model complex scenarios in which a single TF binds to several DBMs or where the same cis-element is targeted by multiple TFs.

All in all, footprintDB currently contains 3095 unique TFs, 4646 PSSMs, and 18,840 DBSs (July 2015). Each data entry can be searched by name, identifier, sequence, or other descriptors and visualized in a data sheet format including footprintDB annotations (*see* examples in Fig. 5). Of these, the current release contains 275 non-redundant DNA motifs from plant TFs, which have also been recently included in the RSAT::Plants server (*see* **Note 1**, <http://plants.rsat.eu>). RSAT::Plants is a software suite that integrates a series of modular computer programs designed for the detection of regulatory signals in non-coding sequences [42]. The ArabidopsisPBM and HumanTF collections have also been exported to the MEME suite, which includes a compendium of similar tools (<http://meme.nbcr.net>) [43].

FootprintDB supports two kinds of queries: (1) TF sequences and (2) DNA motifs or sites (Fig. 6). TF sequence searches, using protein sequences as input, retrieve all similar TFs found in the database by performing local sequence alignments with BLASTP [44].

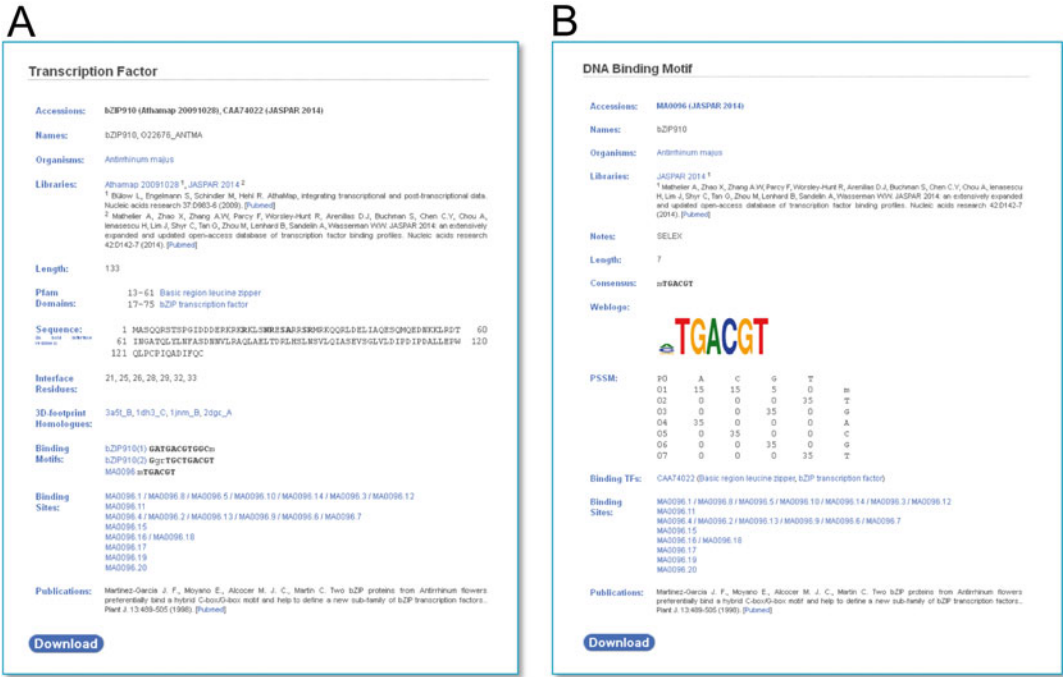


Fig. 5 Example of typical footprintDB data entries: (a) Transcription factor bZIP910. (b) A DNA motif of *cis*-elements recognized by bZIP910

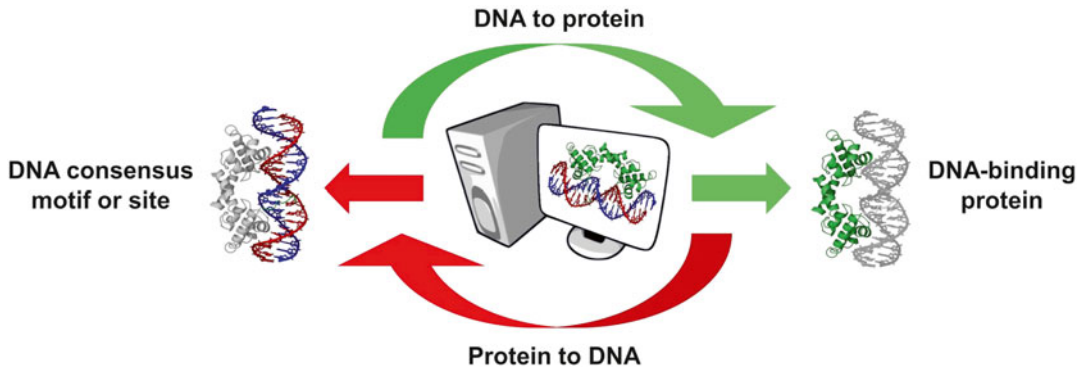


Fig. 6 The two main queries supported by the footprintDB search engine

DNA queries, giving as input DNA motifs in PSSM format or individual DNA sequences, retrieve all similar motifs found in the database using STAMP tool for the alignments [45]. Additionally, search results can be filtered by source database, organism or protein domain, and can be saved for later use.

Users can also upload collections of TFs and DNA motifs into footprintDB, keeping the data for private use (default) or making it publicly available. The footprintDB search engine can also be accessed programmatically using a SOAP web services interface. These and more possibilities are further explained in the software documentation.

Only few organisms have available experimental TF-DNA binding data in the literature. For this reason, footprintDB searches can be extended to third-party organism proteomes, allowing search results to be extrapolated to a particular organism of interest. For example, we can use a known *A. thaliana* TF that might bind conserved cis-elements in promoters of overexpressed genes and then find its homologues in the proteome of *Oryza sativa*, if we are interested in conducting an experiment with rice. According to our benchmarks, homologous TFs from different species are more likely to bind a similar DNA motif when their interfaces are similar [28].

In order to evaluate the contents of footprintDB, the initial 2013 release was compared to subscription-based TRANSFAC 2012.1, and a high degree of redundancy was found among TFs and DBMs stored in both repositories, as detailed in **Note 2** [28].

TF protein sequences have their DNA interfaces annotated in footprintDB. That means that residues involved in the recognition of nucleotide bases are predicted based on homology to proteins that take part on three-dimensional protein–DNA complexes annotated in the 3D-footprint database (*see Note 3*, <http://floresta.ead.csic.es/3dfingerprint>) [26]. FootprintDB lists the PDB accessions for these complexes, making it possible to produce structure-based alignments of cis-elements and to compare binding interfaces with the TFcompare web server (*see Note 4*) [3]. DNA binding protein domains are also annotated by scanning sequences against PFAM domains [46]. Check Fig. 5 for the content of a typical entry of the database.

3 Methods

3.1 DNA Motif Search in FootprintDB

The first run mode of footprintDB takes a DNA motif, cis-element or DNA site as input. This kind of search is useful when we have data about DNA sequences (or a single sequence) recognized by an unknown DNA-binding protein and we want to predict TFs able to interact with them. In the next lines we explain how to feed a DNA motif or site in footprintDB and how to interpret the results using as example the *Antirrhinum majus* motif called “bZIP910” (Fig. 5b), originally annotated in the JASPAR database [29, 47].

1. First, if you have a footprintDB account, log in to store your searches and reuse them.
2. Click on the “Start Search” button or in the “Search Sequences” link on the left menu at <http://floresta.ead.csic.es/footprintdb>.
3. Enter a name for the job and, optionally, an email address if you desire to receive the results by email.
4. Choose as input type “DNA sites or motifs” and the number of results that you desire in “Limit number of results per query” field.

5. Enter your DNA sites or motifs in the text area or upload them from a file. The accepted formats are FASTA and TRANSFAC (some simplifications of the original TRANSFAC format are also accepted). In the present example, we will use the example data by choosing “bZIP910 DNA Binding Motif in TRANSFAC-like format” and clicking the “Demo” button. A matrix with “bZIP910” TF binding preferences will be shown automatically in the text area.
6. To start the search, press the “Search” button.

Additionally, options on the footprintDB search form such as “Order results by” or “Color results using twilight thresholds” can be changed (*see Note 5*). Search can also be limited by organism, source database or Pfam domains (*see Note 6*). Clicking on “Search for homologues in a selected proteome” displays additional parameters which will be explained later in the Subheading 3.2.

Among the obtained results, shown in Fig. 7, we notice that the first one is the query itself (the demo “bZIP910” is a regular

footprintDB results for Demo

Query: bZIP910 (JASPAR CORE) mTGACGT

footprintDB template	Source	Organisms	STAMP e-value	Motif similarity	footprinDB PWM / Consensus	Binding proteins	Interface sequences	Pfam domains
MA0096: bZIP910	JASPAR 2014	Anthrimum majus	1.0e-12	7.00 / 7	ACGTCAk ACGTCAk	Show proteins	Show interfaces	Show domains
TGA2_2: TGA2	ArabidopsisPBM 20140210	Arabidopsis thaliana	3.0e-10	6.97 / 7	-ACGTCak-- kACGTCakCa	Show proteins	Show interfaces	Show domains
MF0002: bZIP CREB/IG-box-like subclass	JASPAR 2014	METAMODEL	8.9e-10	5.99 / 6	ACGTCAk ACGTCA-	METAMODEL		
CREM_ft: CREM	HOCOMOCO v9	Homo sapiens Mus musculus	1.8e-09	6.78 / 7	-ACGTCak-- GACGTcAbys	Show proteins	Show interfaces	Show domains
XBP1_DBD_2: XBP1	HumanTF 1.0	Homo sapiens	9.6e-09	6.85 / 7	-----ACGTCak- wzkGmCAGGTCakc	Show proteins	Show interfaces	Show domains
XBP1_DBD_1: XBP1	HumanTF 1.0	Homo sapiens	1.1e-08	6.76 / 7	---ACGTCak- GaTGAGGTCatc	Show proteins	Show interfaces	Show domains
MA0588: TGA1	JASPAR 2014	Arabidopsis thaliana	1.4e-08	6.64 / 7	--ACGTCak-- hyACGTCabsm	Show proteins	Show interfaces	Show domains
TGA2: TGA2	ArabidopsisPBM 20140210	Arabidopsis thaliana	1.8e-08	6.52 / 7	---ACGTCak rTGAGTCAY	Show proteins	Show interfaces	Show domains
MA0266: CST6	JASPAR 2014	Saccharomyces cerevisiae	2.6e-08	6.37 / 7	--ACGTCak tkACGTCAY	Show proteins	Show interfaces	Show domains
TGA1: TGA1	Athamap 20091028	Arabidopsis thaliana	2.8e-08	6.65 / 7	--ACGTCak--- taaCGTCabsw	Show proteins	Show interfaces	Show domains

Fig. 7 Example of results for the “bZIP910” DNA motif search in footprintDB. Note that slightly different versions of *A. thaliana* motifs TGA 1 and 2 are reported. Results can vary in future versions of footprintDB

footprintDB entry), together with a list of transcription factors that are reported to bind very similar DNA motifs. Each row contains additional information about the motif alignment (E-value and similarity scores, *see* **Notes 5** and **7**), source organism and links to the original source, the motif datasheet in footprintDB, the annotated binding proteins, their binding interface residues and their binding protein domains. By clicking on “Show interfaces” and “Show domains” in the result table it can be seen that the results share a common DNA-binding interface (R[KL]x[SQK]NR[ev][SA]Axx[SCA]RxRK) within the Basic Leucine Zipper domain. Note that predicted binding residues are in uppercase in the consensus.

3.2 Proteome-Specific DNA Search in FootprintDB

As illustrated in the previous section, a regular DNA search can give us valuable information about a novel DNA cis-element or motif by comparing it to similar motifs annotated in footprintDB. However, often we are interested in finding a list of transcription factors that most likely bind that DNA motif in a specific organism. For example, we might know an abiotic stress cis-element in the *Antirrhinum majus* genome (bZIP910) and we want to test rice TFs that potentially bind this sequence. For this purpose, we can extend the footprintDB search by selecting a specific target proteome:

1. Repeat the **steps 1–5** of DNA motif search explained on Subheading **3.1**.
6. Click on the link “Search for homologues in a selected proteome” to expand proteome search options and select from the list of available proteomes “*Oryza sativa*—MSU6.1” or upload a FASTA file with the desired proteome.
7. Click on the “Search” button to start the search.

Results look like a regular DNA motif search, but footprintDB entries with homologous proteins within the rice proteome are shown on the top of the list, and entries without rice homologues at the bottom with the legend “NO HITS”. Click on the link “Show *Oryza sativa*—MSU6.1 homologues” displayed below high-ranking results on the leftmost column to expand a list of significant BLASTP hits (homologues) found in the proteome. Each one contains information about the binding interface and alignment scores with the footprintDB TF sequence. In the bZIP910 example, rice homologues show conserved interface residues and common DNA-binding domains as expected. Following these results, experiments can now be designed to test whether these TFs actually bind a bZIP910-like motif in our conditions of interest, thus reducing the search space from around fifty thousand rice transcripts to a few dozens of footprintDB predictions.

3.3 Protein Sequence Search in FootprintDB

For most putative TFs in sequence databases, there is little or no information regarding their DNA binding preferences. In these cases, we can look for homologous TFs annotated in footprintDB and transfer their curated DNA binding data. To illustrate this type of search in footprintDB we will take the amino acid sequence of bZIP910 TF from *A. majus* (Fig. 5a), extending the examples of the previous two sections.

1. Repeat the **steps 1–3** listed in Subheading 3.1.
4. Choose as input type “Proteins” and set the wished number of results in the “Limit number of results per query” box.
5. Enter the protein sequences in the text area or upload them from a file in FASTA format. In the present example, we will use the demo data by choosing “bZIP910 Protein Sequence in FASTA format” and clicking the “Demo” button. The protein sequence will be shown automatically in the text area.
6. Press the “Search” button.

As previously, search can be limited by organism, source database or Pfam domains (*see Note 6*). Additionally, results can be ordered by E-value (*see Note 7*) or interface similarity (*see Note 8*). As expected, the first reported TF is bZIP910 and all remaining TFs belong to the same “bZIP domain” family, share similar interface residues and recognize G-box (CACGTG) and C-box (GACGTC) motifs related to the cognate bZIP910 consensus: a G-box/C-box hybrid (GACGTG) [29, 47]. As illustrated in Subheading 3.2, an organism-specific proteome can be chosen or uploaded to retrieve homologous proteins only from desired species.

3.4 In Silico Prediction of Transcription Factors for Co-expressed Gene Promoters in *A. thaliana*

Let us review a recent study where 32 co-expressed drought-responsive *A. thaliana* gene clusters were analyzed [48]. One of the aims of this work was the identification of transcription factors involved in drought response regulation and their experimental validation. This is an application of the protocol in Subheading 3.2. To achieve this goal, gene expression profiles in a large *A. thaliana* microarray set were clustered and upstream sequences (1 Kb from transcription start site) of genes in each cluster searched for significantly overrepresented short DNA sequences. The chosen tool for motif discovery was BEST (<http://www.people.fas.harvard.edu/~junliu/BEST>), a meta-predictor which combines different motif-finding programs [49]. Out of 179 motifs discovered, 15 putative regulatory sequences were selected to scan a large library of *A. thaliana* TFs – called REGIA (*see Note 9*)—in a yeast one-hybrid experiment. Here the first of these sequences (shoots1hr_9, CTCCACGTGC) is further used to demonstrate the performance of footprintDB when looking in the *A. thaliana* proteome for TFs binding similar DNA targets:

1. Repeat the **steps 1–4** of DNA motif search explained in Subheading 3.1. Set “Limit number of results per query” to 100. Paste the DNA sequence (*see Note 10*).
5. Select from the list of available proteomes “Arabidopsis thaliana—regia” and set the BLASTP E-value threshold to 1E-10 (*see Note 7*) in order to reproduce the search strategy of the work of Dubos et al. [48]. Note that TRANSFAC 2012.1 was also used as a database on that project, but unfortunately that repository cannot be offered to footprintDB users due to license restrictions.
6. Click on the “Search” button to launch the job.

If the specific proteome search option is turned off, the first plant TF is ranked 10th (MYC4 from JASPAR database, but it can vary in future versions). However, when the REGIA proteome is selected MYC4 is ranked first on the list of results, together with other TFs that in the preliminary search appeared lower in the rankings (*see Note 11*). Homologous TFs from REGIA can be visualized by expanding the link “Show Arabidopsis thaliana—regia homologues”, with binding interfaces highlighted. In this example, the top MYC4 homologues share the interface motif [ns]HV[ev]AE[rk][qr]RReklN(X)12[vi][st][kr]Mdk.

Beyond this example, an important result of the Dubos et al. study [48] was the systematic comparison of computer predictions and the Y1H results for all 15 *cis*-elements under study. First, it turned out that *in silico* TF predictions failed to correctly identify binding proteins whenever footprintDB contained no proteins with significantly similar DNA motifs (STAMP E-value > 1E-3, *see Note 7*). In other words, footprintDB can successfully predict TFs for input DNA motifs only if significantly similar motifs are already annotated in the databases. Second, in five cases where this occurred, footprintDB included the experimentally determined TFs among the predictions. However, it also incorporated a number of TFs which are false positives. In summary, as long as the interface is similar, footprintDB will retrieve TFs from the same family, even if they are not expressed or they do not bind the *cis*-element under the studied experimental conditions. Figure 8 illustrates the agreement between *in silico* and yeast one-hybrid TF predictions for shoots1hr_9 as in the original article by Dubos et al. (including TRANSFAC search results) [48].

3.5 Prediction of OsEREBP1 and OsEREBP2 Regulatory Sites Within the OsRMC Promoter

In this last example, taken from Serra et al. [50], we show how footprintDB can be used to identify target *cis*-elements of a TF of interest. This is an application of the protocol in Subheading 3.3. In that work some experiments were performed to unveil the regulation of rice gene OsRMC under high salinity conditions. Thus, a salt-induced rice cDNA expression library was constructed and subsequently screened using the yeast one-hybrid system and the

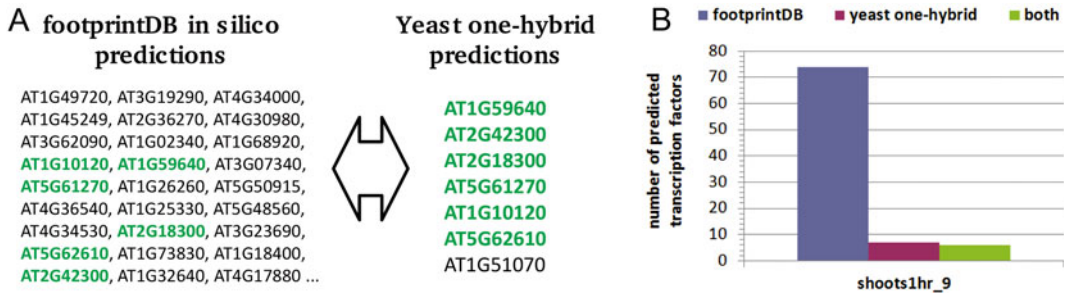


Fig. 8 Comparison of footprintDB in silico predictions of binding TFs and yeast one-hybrid experimental results within a library of cloned *A. thaliana* TFs for the cis-element “shoots1hr_9”. (a) Example TF predictions by both methods, with common ones in *green*. (b) Chart comparing all unique and common TF predictions obtained with both methods

OsRMC promoter as bait. As a result, OsEREBP1 and OsEREBP2, two putative TFs of the AP2/ERF family, were identified to bind a 648 bp region of the OsRMC promoter. In a first approach, the OsRMC promoter sequence was scanned with DNA motifs in the PlantPAN database, which included curated data from PLACE, TRANSFAC 7, AGRIS, and JASPAR 2008 databases [40, 41, 51–53]. Several potential DBSs were identified in the promoter window, but none of them contained the GCC box typical of ERF specificity. This section explains how footprintDB was used to design subsequent experiments which confirmed the precise locations of GCC-like cis-elements in this promoter:

1. Visit UniProt (<http://www.uniprot.org>) and download the amino acid sequences of rice proteins OsEREBP1 and OsEREBP2, which correspond to accessions Q6K7E6 and Q5N965 respectively. OsEREBP2 has only been partially annotated and for that reason appears as “AP2 domain transcription factor-like”. Scroll down the respective UniProt pages, locate the “Sequence” section and download the sequences in FASTA format.
2. Repeat the **steps 1–4** of Subheading 3.3.
6. Paste the downloaded sequences in FASTA format, name the job and select “Athamap” and “3D-footprint” as databases (*see Note 6*).
7. Press the “Search” button.

The best candidate motif retrieved from Athamap and 3D-footprint are ERF4 (plus other ERF-like TFs) and 1gcc_A, respectively. Both are related motifs that contain the GCC box and their cognate TFs have all their interface residues conserved. OsEREBP1, OsEREBP2, ERF4, and 1gcc_A have the following common consensus interface: [iv]R[qk]RpWg[kr]xaaEiRdp(x)4-5RvWlgt. The DNA motifs of ERF4 and 1gcc_A TFs were

subsequently used to locate the most likely cis-elements within the OsRMC promoter using tools from the RSAT::Plants web server.

1. Click on ERF4 and Igcc_A links in the “footprintDB PWM” column from the previous results and download their PSSMs in TRANSFAC-like format.
2. Download the OsRMC gene promoter region from the original Serra et al. paper [50] (*see Note 12*).
3. Go to RSAT::Plants server (<http://plants.rsat.eu>), and select from the left menu: “Pattern-matching—matrix-scan (quick)”.
4. Paste the OsRMC promoter sequence in FASTA format and the two motif matrices selecting TRANSFAC format from the menu.
5. Select “Background model estimation method—Organism-specific” and “Oryza sativa IRGSP” with sequence type “upstream-noorf”. Leave other parameters with default values.
6. Click on “GO” button to start the calculations.

Among the obtained results, the cis-sequence TGCCTGCTC, found by both input motifs with coordinates -481 and -473 , showed binding activity to OsEREBP1 and OsEREBP2 proteins in electrophoretic mobility shift assays (EMSA) [50].

3.6 Conclusions and Perspectives

In this chapter we have presented some examples and real case studies in plants of the possibilities of footprintDB, a unified and open-access online database designed for the analysis of transcription factors and their genomic DNA targets. The main value of footprintDB is probably the integration of a variety of libraries of curated DNA motifs and their associated TFs, which have been increased and updated since the original publication, and in addition, the systematic annotation of interfaces residues of the corresponding TFs. FootprintDB has an open source philosophy, encouraging scientists to contribute with their DNA-binding data to the expansion of the database.

The footprintDB search engine allows querying the database for unknown TFs that are likely to bind input DNA motifs, and also the opposite (Fig. 6). In silico predictions can save time and money when designing laboratory experiments to probe TF DNA binding specificities as in the study cases shown. Search results can be valuable as reported or taken for further analysis in external tools such as RSAT::Plants, TFcompare, etc.

While footprintDB stores data for a variety of organisms, here we have demonstrated typical use cases on plants, reviewing real-world problems that we have encountered with our collaborators. We hope the examples described here can aid other users with related research problems.

4 Notes

1. Due to the drastic increase of available genomes and to improve maintenance and update tasks, RSAT has recently divided in taxon-specific servers, one of them plant specific [42].
2. TRANSFAC (BIOBASE) is a subscription database with curated annotations of transcription factors, experimentally proven binding sites, and the corresponding PSSMs [39]. Additionally, TRANSFAC contains annotation of miRNA and their target sites, together with functional annotations of TFs, predicted promoter binding sites, and additional software tools for DBS prediction and discovery. A comparison between TRANSFAC version 2012.1 and footprintDB initial version showed a high degree of data redundancy between both databases, which shared around 71% of motifs (STAMP $E\text{-value} \leq 1E-10$) and 56% of TFs (% sequence identity ≥ 90). Additionally, some internal redundancy was detected in both databases, accounting for 20–25% of DBMs and 43–45% of TFs. These values indicate that a large proportion of the underlying experimental studies focus on a small number of regulation-related protein families, and that probably there are still many families and cis-elements to be discovered.
3. 3D-footprint is a database which provides estimates of binding specificity for all protein–DNA complexes available at the Protein Data Bank [27]. Each complex in the database is dissected to draw interface graphs and footprint logos, and two complementary algorithms are employed to characterize binding specificity. Moreover, oligonucleotide sequences extracted from literature abstracts are reported in order to show the range of variant sites bound by each protein and other related proteins. 3D-footprint is updated and curated on a weekly basis.
4. If three-dimensional structures of two DNA–protein complexes of the same family are available, it is possible to compare their binding interfaces by a structure-based alignment. This can be done by feeding TFcompare server (<http://floresta.cead.csic.es/tfcompare>) with the PDB identifiers from two protein–DNA complexes [3]. It first extracts individual DNA-binding protein domains to calculate their optimal fit and then returns their structural alignments. The superposition of protein chains is used to generate the structure-based alignment of the bound DNA sequences. As a result, nucleotides that are recognized by equivalent interface residues in both complexes are aligned together. The resulting DNA alignment does not rely in the nucleotide sequences and differs in many cases from a pure sequence alignment. This kind of alignment makes sense only for binding domains, not for the whole protein, and for this reason the original structures are trimmed according to

their PFAM-defined domain boundaries. All DNA-contacting domains from the first structure are aligned to those in the second and produced alignments are scored in terms of the number of identical superposed nucleotides and the sum of N9 (nitrogen 9 in purines) and N1 (nitrogen 1 in pyrimidines) atom pairs within 3.5 Å.

5. Results are sorted by default by STAMP E-value, but can also be sorted on computed DNA similarity. By default rows in the results table are colored after the alignment quality thresholds reported previously [3]. Motifs returned to the user are shown with a green background when the obtained score is above the twilight zone; otherwise the background is in red. While the twilight cut-offs were calculated on a large set of alignments, it is still possible that a correct alignment turns out in red.
6. “Multiple Organisms” can be selected by pressing the Ctrl key, as well as “Original Databases” and “Pfam domains”. The option of restricting the search by “Organisms” should be used with caution because some TFs and DNA motifs are not associated to a specific species.
7. The Expect value (E-value) is an estimate of the number of false-positive results we can expect by chance when searching a database. For example, an E-value of 1 can be interpreted as meaning that in a database of the current size one might expect to see 1 match with a similar score simply by chance. The lower the E-value, or the closer it is to zero, the more probable that a match is a true positive and not a random result. Short sequences have a higher probability of occurring in the database purely by chance, that is why short DNA searches usually yield larger E-values than protein ones.
8. Interface similarity can be useful to score candidate protein sequences within a proteome of interest. This variable takes values from 0 to 100% and captures the sequence similarity of residues possibly involved in specific DNA recognition. In our benchmarks with human, *A. thaliana* and *E. coli* TFs [28], the interface similarity of correct predictions was significantly higher ($\approx 70\%$) than the similarity of all aligned TFs ($\approx 50\%$).
9. REGIA (REGulatory Gene Initiative in Arabidopsis) was an EU-funded project involving 29 European laboratories with the objective of determining the function of virtually all transcription factors from the model plant *A. thaliana* [54]. The REGIA consortium provided a normalized full size TF library (more than 800 full length ORFs cloned) available to the scientific community for screening for additional interactions, particularly with non-TF proteins.
10. In the experiment by Dubos et al. [48], two types of input were tested: single cis-elements and PSSMs compiled from a

group of aligned cis-elements. In this context single elements yielded fewer false positives, although our benchmarks suggest that PSSMs produce more accurate alignments than individual sequences [3]. We therefore recommend trying both inputs if possible.

11. Looking for homologous proteins in *A. thaliana* (or any other target species) helps filtering results, as it favors higher rankings for TFs with homologues in the chosen proteome, which are more likely to be true predictions.
12. OsRMC corresponds to gene OS04T0659300 of the International Rice Genome Sequencing Project (IRGSP, <http://rgp.dna.affrc.go.jp/IRGSP>). The promoter fragment used in the original Serra et al. paper can be retrieved with help from RSAT::Plants (<http://plants.rsat.eu>). On the left menu select “Sequence tools—retrieve sequence”, set “Organism” to “*Oryza sativa* IRGSP”, choose “Gene” selection and paste “OS04T0659300-01”. Leave all other options with default values except “From” and “To”, which should be set to -1321 and -674, respectively.

Acknowledgments

We would like to thank our colleagues C. Dubos, L. Bülow, N. Saibo, T. Serra and J. van Helden for past and current collaborations. This work was funded by grant Euroinvestigación EUI2008-03612 under the framework of the Transnational (Germany, France, Spain) Cooperation within the PLANT-KBBE Initiative.

References

1. Stormo GD (2000) DNA binding sites: representation and discovery. *Bioinformatics* 16(1):16–23
2. Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 18(20):6097–6100
3. Sebastian A, Contreras-Moreira B (2013) The twilight zone of cis element alignments. *Nucleic Acids Res* 41(3):1438–1449. doi:10.1093/nar/gks1301
4. Galas DJ, Schmitz A (1978) DNase footprinting: a simple method for the detection of protein-DNA binding specificity. *Nucleic Acids Res* 5(9):3157–3170
5. Garner MM, Revzin A (1981) A gel electrophoresis method for quantifying the binding of proteins to specific DNA regions: application to components of the *Escherichia coli* lactose operon regulatory system. *Nucleic Acids Res* 9(13):3047–3060
6. O’Neill LP, Turner BM (1996) Immunoprecipitation of chromatin. *Methods Enzymol* 274:189–197
7. Ren B, Robert F, Wyrick JJ, Aparicio O, Jennings EG, Simon I, Zeitlinger J, Schreiber J, Hannett N, Kanin E, Volkert TL, Wilson CJ, Bell SP, Young RA (2000) Genome-wide location and function of DNA binding proteins. *Science* 290(5500):2306–2309. doi:10.1126/science.290.5500.2306
8. Berger MF, Bulyk ML (2006) Protein binding microarrays (PBMs) for rapid, high-throughput characterization of the sequence specificities of DNA binding proteins. *Methods Mol Biol* 338:245–260. doi:10.1385/1-59745-097-9:245
9. Johnson DS, Mortazavi A, Myers RM, Wold B (2007) Genome-wide mapping of in vivo protein-DNA interactions. *Science* 316(5830):1497–1502. doi:10.1126/science.1141319

10. Ogawa N, Biggin MD (2012) High-throughput SELEX determination of DNA sequences bound by transcription factors in vitro. *Methods Mol Biol* 786:51–63. doi:[10.1007/978-1-61779-292-2_3](https://doi.org/10.1007/978-1-61779-292-2_3)
11. Machanick P, Bailey TL (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 27(12):1696–1697. doi:[10.1093/bioinformatics/btr189](https://doi.org/10.1093/bioinformatics/btr189)
12. Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J (2011) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res*. doi:[10.1093/nar/gkr1104](https://doi.org/10.1093/nar/gkr1104)
13. Hertz GZ, Stormo GD (1999) Identifying DNA and protein patterns with statistically significant alignments of multiple sequences. *Bioinformatics* 15(7–8):563–577. doi:[bt069](https://doi.org/10.1093/bioinformatics/bt069) [pii]
14. Bailey TL, Williams N, Misleh C, Li WW (2006) MEME: discovering and analyzing DNA and protein sequence motifs. *Nucleic Acids Res* 34(Web Server issue):W369–W373. doi:[10.1093/nar/gkl198](https://doi.org/10.1093/nar/gkl198) [pii]
15. Frith MC, Fu Y, Yu L, Chen JF, Hansen U, Weng Z (2004) Detection of functional DNA motifs via statistical over-representation. *Nucleic Acids Res* 32(4):1372–1381. doi:[10.1093/nar/gkh299](https://doi.org/10.1093/nar/gkh299)
16. Chen QK, Hertz GZ, Stormo GD (1995) MATRIX SEARCH 1.0: a computer program that scans DNA sequences for transcriptional elements using a database of weight matrices. *Comput Appl Biosci* 11(5):563–566
17. Mahony S, Auron PE, Benos PV (2007) DNA familial binding profiles made easy: comparison of various motif alignment and clustering strategies. *PLoS Comput Biol* 3(3), e61. doi:[10.1371/journal.pcbi.0030061](https://doi.org/10.1371/journal.pcbi.0030061)
18. Turatsinze JV, Thomas-Chollier M, Defrance M, van Helden J (2008) Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. *Nat Protoc* 3(10):1578–1588. doi:[10.1038/nprot.2008.97](https://doi.org/10.1038/nprot.2008.97)
19. Bailey TL, Gribskov M (1998) Combining evidence using p-values: application to sequence homology searches. *Bioinformatics* 14(1):48–54
20. Shortle D, DiMaio D, Nathans D (1981) Directed mutagenesis. *Ann Rev Genet* 15:265–294. doi:[10.1146/annurev.ge.15.120181.001405](https://doi.org/10.1146/annurev.ge.15.120181.001405)
21. O'Neill M, Dryden DT, Murray NE (1998) Localization of a protein-DNA interface by random mutagenesis. *EMBO J* 17(23):7118–7127. doi:[10.1093/emboj/17.23.7118](https://doi.org/10.1093/emboj/17.23.7118)
22. Morozov AV, Havranek JJ, Baker D, Siggia ED (2005) Protein-DNA binding specificity predictions with structural models. *Nucleic Acids Res* 33(18):5781–5798. doi:[10.1093/nar/gki875](https://doi.org/10.1093/nar/gki875)
23. Alamanova D, Stegmaier P, Kel A (2010) Creating PWMs of transcription factors using 3D structure-based computation of protein-DNA free binding energies. *BMC Bioinformatics* 11:225. doi:[10.1186/1471-2105-11-225](https://doi.org/10.1186/1471-2105-11-225)
24. Contreras-Moreira B, Collado-Vides J (2006) Comparative footprinting of DNA-binding proteins. *Bioinformatics* 22(14):e74–e80. doi:[10.1093/bioinformatics/btl215](https://doi.org/10.1093/bioinformatics/btl215)
25. Angarica VE, Perez AG, Vasconcelos AT, Collado-Vides J, Contreras-Moreira B (2008) Prediction of TF target sites based on atomistic models of protein-DNA complexes. *BMC Bioinformatics* 9:436. doi:[10.1186/1471-2105-9-436](https://doi.org/10.1186/1471-2105-9-436)
26. Contreras-Moreira B (2010) 3D-footprint: a database for the structural analysis of protein-DNA complexes. *Nucleic Acids Res* 38(Database issue):D91–D97. doi:[10.1093/nar/gkp781](https://doi.org/10.1093/nar/gkp781)
27. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The Protein Data Bank. *Nucleic Acids Res* 28(1):235–242
28. Sebastian A, Contreras-Moreira B (2014) footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics* 30(2):258–265. doi:[10.1093/bioinformatics/btt663](https://doi.org/10.1093/bioinformatics/btt663)
29. Mathelier A, Zhao X, Zhang AW, Parcy F, Worsley-Hunt R, Arenillas DJ, Buchman S, Chen CY, Chou A, Ienasescu H, Lim J, Shyr C, Tan G, Zhou M, Lenhard B, Sandelin A, Wasserman WW (2014) JASPAR 2014: an extensively expanded and updated open-access database of transcription factor binding profiles. *Nucleic Acids Res* 42(Database issue):D142–D147. doi:[10.1093/nar/gkt997](https://doi.org/10.1093/nar/gkt997)
30. Hume MA, Barrera LA, Gisselbrecht SS, Bulyk ML (2015) UniPROBE, update 2015: new tools and content for the online database of protein-binding microarray data on protein-DNA interactions. *Nucleic Acids Res* 43(Database issue):D117–D122. doi:[10.1093/nar/gku1045](https://doi.org/10.1093/nar/gku1045)
31. Jolma A, Yan J, Whittington T, Toivonen J, Nitta Kazuhiro R, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas Juan M, Vincentelli R, Luscombe Nicholas M, Hughes Timothy R, Lemaire P, Ukkonen E, Kivioja T, Taipale J (2013) DNA-binding specificities of human transcription factors. *Cell* 152(1):327–339

32. Kulakovskiy IV, Medvedeva YA, Schaefer U, Kasianov AS, Vorontsov IE, Bajic VB, Makeev VJ (2013) HOCOMOCO: a comprehensive collection of human transcription factor binding sites models. *Nucleic Acids Res* 41(Database issue):D195–D202. doi:[10.1093/nar/gks1089](https://doi.org/10.1093/nar/gks1089)
33. Salgado H, Peralta-Gil M, Gama-Castro S, Santos-Zavaleta A, Muniz-Rascado L, Garcia-Sotelo JS, Weiss V, Solano-Lira H, Martinez-Flores I, Medina-Rivera A, Salgado-Osorio G, Alquicira-Hernandez S, Alquicira-Hernandez K, Lopez-Fuentes A, Porron-Sotelo L, Huerta AM, Bonavides-Martinez C, Balderas-Martinez YI, Pannier L, Olvera M, Labastida A, Jimenez-Jacinto V, Vega-Alvarado L, Del Moral-Chavez V, Hernandez-Alvarez A, Morett E, Collado-Vides J (2013) RegulonDB v80: omics data sets, evolutionary conservation, regulatory phrases, cross-validated gold standards and more. *Nucleic Acids Res* 41(Database issue):D203–D213. doi:[10.1093/nar/gks1201](https://doi.org/10.1093/nar/gks1201)
34. Siervo N, Makita Y, de Hoon M, Nakai K (2008) DBTBS: a database of transcriptional regulation in *Bacillus subtilis* containing upstream intergenic conservation information. *Nucleic Acids Res* 36(Database issue):D93–D96. doi:[10.1093/nar/gkm910](https://doi.org/10.1093/nar/gkm910)
35. Bülow L, Engelmann S, Schindler M, Hehl R (2009) AthaMap, integrating transcriptional and post-transcriptional data. *Nucleic Acids Res* 37(Database issue):D983–D986. doi:[10.1093/nar/gkn709](https://doi.org/10.1093/nar/gkn709)
36. Franco-Zorrilla JM, Lopez-Vidriero I, Carrasco JL, Godoy M, Vera P, Solano R (2014) DNA-binding specificities of plant transcription factors and their potential to define target genes. *Proc Natl Acad Sci U S A* 111(6):2367–2372. doi:[10.1073/pnas.1316278111](https://doi.org/10.1073/pnas.1316278111)
37. Down TA, Bergman CM, Su J, Hubbard TJ (2007) Large-scale discovery of promoter motifs in *Drosophila melanogaster*. *PLoS Comput Biol* 3(1), e7. doi:[10.1371/journal.pcbi.0030007](https://doi.org/10.1371/journal.pcbi.0030007)
38. Enuameh MS, Asriyan Y, Richards A, Christensen RG, Hall VL, Kazemian M, Zhu C, Pham H, Cheng Q, Blatti C, Brasefield JA, Basciotta MD, Ou J, McNulty JC, Zhu LJ, Celniker SE, Sinha S, Stormo GD, Brodsky MH, Wolfe SA (2013) Global analysis of *Drosophila* Cys(2)-His(2) zinc finger proteins reveals a multitude of novel recognition motifs and binding determinants. *Genome Res* 23(6):928–940. doi:[10.1101/gr.151472.112](https://doi.org/10.1101/gr.151472.112)
39. Matys V, Kel-Margoulis OV, Fricke E, Liebich I, Land S, Barre-Dirrie A, Reuter I, Chekmenev D, Krull M, Hornischer K, Voss N, Stegmaier P, Lewicki-Potapov B, Saxel H, Kel AE, Wingender E (2006) TRANSFAC and its module TRANSCOMP: transcriptional gene regulation in eukaryotes. *Nucleic Acids Res* 34(Database issue):D108–D110. doi:[10.1093/nar/gkj143](https://doi.org/10.1093/nar/gkj143)
40. Higo K, Ugawa Y, Iwamoto M, Korenaga T (1999) Plant cis-acting regulatory DNA elements (PLACE) database: 1999. *Nucleic Acids Res* 27(1):297–300. doi:[10.1093/nar/27.1.297](https://doi.org/10.1093/nar/27.1.297)
41. Davuluri RV, Sun H, Palaniswamy SK, Matthews N, Molina C, Kurtz M, Grotewold E (2003) AGRIS: Arabidopsis gene regulatory information server, an information resource of Arabidopsis cis-regulatory elements and transcription factors. *BMC Bioinformatics* 4:25. doi:[10.1186/1471-2105-4-25](https://doi.org/10.1186/1471-2105-4-25)
42. Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon J, Delerce J, Spinelli L, Jaeger S, Blanchet C, Vincens P, Caron C, Staines D, Contreras-Moreira B, Artufel M, Charbonnier L, Hernandez C, Thieffry D, Thomas-Chollier M, van Helden J (2015) RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res* 43:W50–W56
43. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS (2009) MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res* 37(Web Server issue):W202–W208. doi:[10.1093/nar/gkp335](https://doi.org/10.1093/nar/gkp335)
44. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ (1990) Basic local alignment search tool. *J Mol Biol* 215(3):403–410. doi:[10.1006/jmbi.1990.9999](https://doi.org/10.1006/jmbi.1990.9999)
45. Mahony S, Benos PV (2007) STAMP: a web tool for exploring DNA-binding motif similarities. *Nucleic Acids Res* 35(Web Server issue):W253–W258
46. Finn RD, Bateman A, Clements J, Coghill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J, Punta M (2014) Pfam: the protein families database. *Nucleic Acids Res* 42(Database issue):D222–D230. doi:[10.1093/nar/gkt1223](https://doi.org/10.1093/nar/gkt1223)
47. Martinez-Garcia JF, Moyano E, Alcocer MJ, Martin C (1998) Two bZIP proteins from *Antirrhinum* flowers preferentially bind a hybrid C-box/G-box motif and help to define a new sub-family of bZIP transcription factors. *Plant J* 13(4):489–505
48. Dubos C, Kelemen Z, Sebastian A, Bülow L, Huep G, Xu W, Grain D, Salsac F, Brousse C, Lepiniec L, Weisshaar B, Contreras-Moreira B, Hehl R (2014) Integrating bioinformatic resources to predict transcription factors interacting with cis-sequences conserved in co-

- regulated genes. *BMC Genomics* 15(1):317. doi:[10.1186/1471-2164-15-317](https://doi.org/10.1186/1471-2164-15-317)
49. Che D, Jensen S, Cai L, Liu JS (2005) BEST: binding-site estimation suite of tools. *Bioinformatics* 21(12):2909–2911
50. Serra TS, Figueiredo DD, Cordeiro AM, Almeida DM, Lourenco T, Abreu IA, Sebastian A, Fernandes L, Contreras-Moreira B, Oliveira MM, Saibo NJ (2013) OsRMC, a negative regulator of salt stress response in rice, is regulated by two AP2/ERF transcription factors. *Plant Mol Biol* 82(4–5):439–455. doi:[10.1007/s11103-013-0073-9](https://doi.org/10.1007/s11103-013-0073-9)
51. Bryne JC, Valen E, Tang MH, Marstrand T, Winther O, da Piedade I, Krogh A, Lenhard B, Sandelin A (2008) JASPAR, the open access database of transcription factor-binding profiles: new content and tools in the 2008 update. *Nucleic Acids Res* 36(Database issue):D102–D106. doi:[10.1093/nar/gkm955](https://doi.org/10.1093/nar/gkm955)
52. Wingender E, Karas H, Knuppel R (1997) TRANSFAC database as a bridge between sequence data libraries and biological function. *Pac Symp Biocomput*:477–485
53. Chang WC, Lee TY, Huang HD, Huang HY, Pan RL (2008) PlantPAN: plant promoter analysis navigator, for identifying combinatorial cis-regulatory elements with distance constraint in plant gene groups. *BMC Genomics* 9:561. doi:[10.1186/1471-2164-9-561](https://doi.org/10.1186/1471-2164-9-561)
54. Paz-Ares J, Regia C (2002) REGIA, an EU project on functional genomics of transcription factors from *Arabidopsis thaliana*. *Comp Funct Genomics* 3(2):102–108. doi:[10.1002/cfg.146](https://doi.org/10.1002/cfg.146)

Chapter 18

RSAT::Plants: Motif Discovery Within Clusters of Upstream Sequences in Plant Genomes

Bruno Contreras-Moreira, Jaime A. Castro-Mondragon, Claire Rioualen, Carlos P. Cantalapiedra, and Jacques van Helden

Abstract

The plant-dedicated mirror of the Regulatory Sequence Analysis Tools (RSAT, <http://plants.rsat.eu>) offers specialized options for researchers dealing with plant transcriptional regulation. The website contains whole-sequenced genomes from species regularly updated from Ensembl Plants and other sources (currently 40), and supports an array of tasks frequently required for the analysis of regulatory sequences, such as retrieving upstream sequences, motif discovery, motif comparison, and pattern matching. RSAT::Plants also integrates the footprintDB collection of DNA motifs. This protocol explains step-by-step how to discover DNA motifs in regulatory regions of clusters of co-expressed genes in plants. It also explains how to empirically control the significance of the result, and how to associate the discovered motifs with putative binding factors.

Key words Co-expression, DNA motif, Position-weight matrix, Upstream sequence, Cluster

1 Introduction

Transcriptome data (microarrays, RNA-seq) have been extensively used as a proxy for genetic regulation in many organisms, as the analysis of genome-wide profiles of gene transcription under different treatments uncovers clusters of genes with correlated behaviors, which may result from direct or indirect co-regulation. A classical application of this approach was done by Beer and co-workers [1] with yeast microarray data sets obtained in a variety of experimental conditions. In that experiment, expression data-mining was demonstrated to be an effective strategy for finding regulons, groups of genes that share regulatory mechanisms and functional annotations.

Other studies have unveiled that the outcome of these approaches largely depends on the genomic background of the species under study. For instance, Sand and others [2] reported that the significance of DNA motifs discovered in *Saccharomyces*

cerevisiae promoters is much higher for regulons than for random gene sets of the same sizes, but for human promoters the signal-to-noise ratio is almost null, because random gene sets give highly significant motifs due to heterogeneities in promoter compositions and biases due to repetitive elements. For metazoans, it is thus a real challenge to distinguish *bona fide* motifs from noise [2]. These observations suggest that motif discovery on sequence clusters faces intrinsic properties of the genomes under study, regardless of the software used for the task.

Among plants, these strategies have so far been tested on the model *Arabidopsis thaliana*, and they have been successfully applied to the identification of novel *cis*-regulatory elements validated with synthetic promoters [3]. Yet, with the exception of this model, these sorts of experiments have not been possible in plants until recently. In spite of this, the growing list of available plant genomes encourages these analyses in combination with expression profiles obtained from either microarray or RNA-seq data sets, as in the recent work of Yu and collaborators [4], provided that the following factors are considered:

- Plant genomes are rich in repetitive elements (RE) distributed along the genome [5], which pose particular problems for motif discovery statistics (violation of the independence assumption).
- Current genome assemblies range from 119.7 Mbp (*A. thaliana*) to 6.48 Gbp (*Triticum aestivum*). *Brachypodium distachyon*, a model species for grasses, is 271.9 Mbp. The quality of these assemblies and their RE content is also quite variable, as shown in Fig. 1 and Table 1.
- Upstream regions, defined by annotated gene coordinates, are also of variable length, going from 1,123 bp on average in *A. thaliana* to 1,856 bp in *Aegilops tauschii* (see Table 1).

This chapter presents a step-by-step protocol for the task of discovering and annotating DNA motifs in clusters of upstream sequences for species supported by RSAT::Plants, which have been obtained mostly from Ensembl Plants (<http://plants.ensembl.org>) [6], but also include data from the JGI Genome Portal (<http://genome.jgi.doe.gov>) [7], and the National Institute of Agrobiological Sciences in Japan (<http://barleyflc.dna.affrc.go.jp/bexdb>) [8]. In addition, RSAT::Plants integrates footprintDB (<http://floresta.eead.csic.es/footprintdb>) [9], a collection of position-specific scoring matrices (PSSM) representing transcription factor binding motifs (TFBM), as well as their cognate binding proteins, which can be used to annotate discovered motifs and to predict potentially binding transcription factors, as illustrated in Chapter 17.

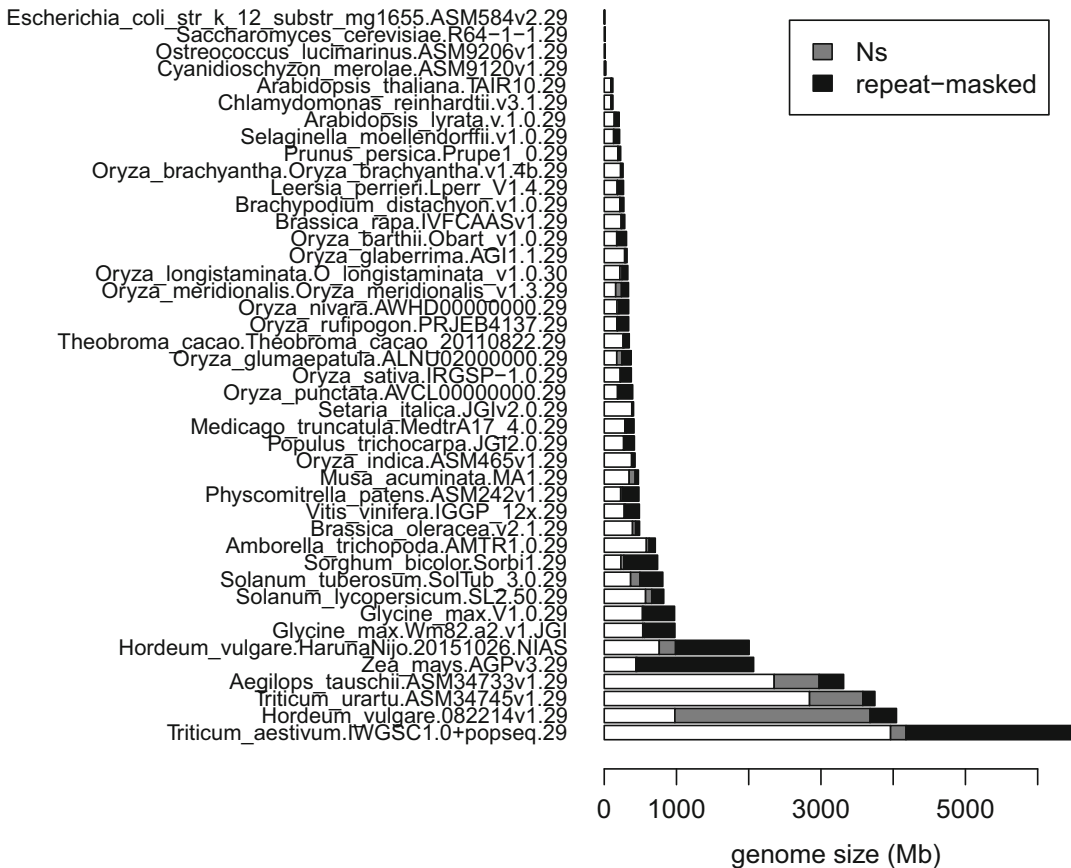


Fig. 1 Genome size of some plant species annotated in RSAT::Plants, showing the fraction of Ns and repeat-masked segments. “Ns” are stretches of uncharacterized nucleotides which often connect assembled sequence contigs. “Repeat-masked” segments are sequences with significant similarity to plant repetitive DNA sequences, which are masked in order to calculate background oligonucleotide frequencies. The full dataset is available at <http://plants.rsat.eu/data/stats>. Most genomes have been downloaded from Ensembl Plants [6]. The yeast genome (*S. cerevisiae*) is plotted as a reference model organism

Discovering regulatory elements within natural genomic sequences is certainly an important scientific goal on its own, but can also be part of the design and validation of synthetic promoters. We envisage at least two applications in this context:

1. The characterization of promoters of genes with known expression properties, which can then be used to engineer the expression of genes of interest.
2. The validation of engineered promoters in order to make sure that they contain the expected regulatory elements, which might be natural or engineered, depending on the application.

Table 1

Features of some plant genomes in RSAT:Plants, taken from <http://plants.rsat.eu/data/stats>. Each ID concatenates the organism, the assembly and the source. Most genome IDs add to the end the Ensembl Plants release number. For instance, *Arabidopsis thaliana.TAIR10.29* corresponds to *A. thaliana* assembly 10 from TAIR (<https://www.arabidopsis.org>) [12], annotated in release 29 of Ensembl Plants. The yeast genome (*S. cerevisiae*) is listed as a reference. “%Ns” are stretches of uncharacterized nucleotides which often connect assembled sequence contigs. “%repeat-masked” segments are sequences with significant similarity to plant repetitive DNA sequences, which are masked

Organism/assembly ID	Genome size (Mb)	Contigs	%Ns	% repeat-masked	Gene models	Mean upstream length
<i>Aegilops tauschii</i> .ASM34733v1.29	3,314	429,892	18.8	10.2	37,035	1,856
<i>Amborella trichopoda</i> .AMTR1.0.29	706	5,745	5.4	12.0	28,721	1,832
<i>Arabidopsis lyrata</i> .v1.0.29	207	695	11.1	21.8	32,667	1,411
<i>Arabidopsis thaliana</i> .TAIR10.29	120	7	0.2	19.3	33,602	1,123
<i>Brachypodium distachyon</i> .v1.0.29	272	83	0.4	20.1	26,552	1,723
<i>Brassica oleracea</i> .v2.1.29	489	32,928	8.8	11.0	59,225	1,628
<i>Brassica rapa</i> .IVFCAASv1.29	284	40,367	3.8	13.9	42,846	1,622
<i>Chlamydomonas reinhardtii</i> .v3.1.29	120	1,558	12.5	11.1	14,487	1,148
<i>Cyanidioschyzon merolae</i> .ASM9120v1.29	17	22	0.0	2.2	5,106	804
<i>Escherichia coli</i> _str_k_12_substr_mgl655.ASM584v2.29	5	1	0.0	0.6	4,497	129
<i>Glycine max</i> .Wm82.a2.v1.JGI	978	1,190	2.4	43.1	56,044	1,806
<i>Hordeum vulgare</i> .082214v1.29	4,045	19,705	66.8	9.0	26,066	1,769
<i>Hordeum vulgare</i> .HarunaNijo.20151026.NIAS	2,006	1,712,261	11.3	50.7	51,249	804
<i>Leersia perrieri</i> .Lperr_V1.4.29	267	12	0.4	31.3	30,615	1,629

Medicago_truncatula.MedtrA17_4.0.29	413	2,186	5.5	25.3	54,073	1,678
Musa_acuminata.MA1.29	473	12	17.4	9.6	37,579	1,469
Oryza_indica.ASM465v1.29	427	10,490	3.8	7.5	88,438	1,512
Oryza_longistaminata.O_longistaminata_v1.0.30	326	60,198	9.8	24.1	31,686	1,566
Oryza_sativa.IRGSP-1.0.29	374	61	0.0	40.5	91,080	1,444
Ostreococcus_lucimarinus.ASM9206v1.29	13	21	0.0	14.7	7,640	510
Physcomitrella_patens.ASM242v1.29	480	2,106	5.4	47.1	32,273	1,607
Populus_trichocarpa.JGI2.0.29	417	2,518	3.2	32.5	41,377	1,794
Prunus_persica.Prupe1_0.29	227	202	1.2	16.1	29,499	1,635
Saccharomyces_cerevisiae.R64-1-1.29	12	17	0.0	6.4	7,126	423
Selaginella_moellendorffii.v1.0.29	213	759	1.9	36.9	34,888	1,168
Setaria_italica.JGIv2.0.29	406	336	1.2	4.8	35,471	1,673
Solanum_lycopersicum.SL2.50.29	824	3,144	10.4	20.1	38,735	1,724
Solanum_tuberosum.SolTub_3.0.29	811	13	15.8	39.1	42,974	1,763
Sorghum_bicolor.Sorbi1.29	738	3,304	5.5	63.2	34,567	1,773
Theobroma_cacao.Theobroma_cacao_20110822.29	346	711	4.4	20.9	29,188	1,253
Triticum_aestivum.IWGSC1.0+popseq.29	6,483	317,977	3.3	35.6	112,496	1,391
Triticum_urartu.ASM34745v1.29	3,747	499,222	19.7	4.4	37,604	1,806
Vitis_vinifera.IGGP_12x.29	486	33	3.3	39.9	29,971	1,728
Zea_mays.AGPv3.29	2,068	523	0.6	78.2	39,625	1,829

2 Materials

This protocol requires disposing of:

1. A computer with any Web browser installed.
2. A set of gene clusters from any of the species currently supported at RSAT::Plants (<http://plants.rsat.eu>, *see Note 1*). Here, we will use three example clusters of co-expressed maize genes, shown in Table 2 (*see Note 2*). More generally, expression data can be obtained from Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo>) [10] and used to produce gene clusters of plant genes (*see Table 3*).

3 Methods

The following protocol enumerates the steps required to discover DNA motifs, based on the over-representation of k-mers (oligonucleotides) and dyads (spaced pairs of oligonucleotides), in clusters of upstream sequences. The protocol comprises two stages, analyzing first co-expressed genes and then random clusters as a negative control (*see Note 3*). Only after both stages have been completed, it is possible to objectively estimate the relevance of the results.

The time required for carrying out the following steps is approximately 1 h.

3.1 *Collecting the Full Set of Promoters for the Genome of Interest*

Before the proper analysis of the gene cluster, we will retrieve the promoter sequences of all the genes of the organism of interest, which will serve below to estimate the background model.

1. Open a connection to the RSAT::Plants server. It can be reached at <http://plants.rsat.eu> and also at <http://floresta.cead.csic.es/rsat>. On the left-side menu, select “Sequence tools → retrieve sequence”.
2. Choose “Single organism → *Zea_mays*.AGPv3.29” for the examples of this protocol (*see Note 1*). At the time of publication this corresponds to Ensembl Plants release 29, but that might change over time.
3. Choose “Genes → all”; this will retrieve all upstream sequences of the maize genome.
4. Set appropriate upstream bounds. Default values are -2000, -1. To replicate the work of Yu et al. [4] these should be set to “From” -1000 “To” +200, with position 0 corresponding to transcriptional start sites (TSS). Beware that TSS positions in plant genomes often correspond to start codons, probably due to incomplete annotations.

Table 2

Clusters of maize (*Zea mays*) genes used along the protocol, extracted from the published work of Yu et al. [4]. Experimentally verified regulatory motifs of these clusters are shown

Cluster name	Confirmed motif	Number of sequences	Gene IDs
ABI4	GCGCRSGCGGSC	16	GRMZM2G025062 GRMZM2G053503 GRMZM2G069082 GRMZM2G069126 GRMZM2G069146 GRMZM2G076896 GRMZM2G081892 GRMZM2G124011 GRMZM2G129674 GRMZM2G142179 GRMZM2G169654 GRMZM2G172936 GRMZM2G173771 GRMZM2G174347 GRMZM2G175525 GRMZM2G421033
E2F	TTCCCGCCA	18	AC197146.3_FG001 GRMZM2G017081 GRMZM2G021069 GRMZM2G037700 GRMZM2G057571 GRMZM2G062333 GRMZM2G065205 GRMZM2G066101 GRMZM2G075978 GRMZM2G100639 GRMZM2G112074 GRMZM2G117238 GRMZM2G130351 GRMZM2G139894 GRMZM2G154267 GRMZM2G162445 GRMZM2G327032 GRMZM2G450055
WRI1	CGGCGGCGS	56	AC210013.4_FG019 GRMZM2G008430 GRMZM2G009968 GRMZM2G010435 GRMZM2G010599 GRMZM2G014444 GRMZM2G015097 GRMZM2G017966 GRMZM2G022019 GRMZM2G027232 GRMZM2G028110 GRMZM2G035017 GRMZM2G041238 GRMZM2G045818 GRMZM2G047727 GRMZM2G048703 GRMZM2G064807 GRMZM2G068745 GRMZM2G074300 GRMZM2G076435 GRMZM2G078779 GRMZM2G078985 GRMZM2G080608 GRMZM2G092663 GRMZM2G096165 GRMZM2G098957 GRMZM2G107336 GRMZM2G108348 GRMZM2G111987 GRMZM2G115265 GRMZM2G119865 GRMZM2G122871 GRMZM2G126603 GRMZM2G126928 GRMZM2G132095 GRMZM2G140799 GRMZM2G148744 GRMZM2G150434 GRMZM2G151252 GRMZM2G152599 GRMZM2G170262 GRMZM2G181336 GRMZM2G311914 GRMZM2G312521 GRMZM2G322413 GRMZM2G325606 GRMZM2G343543 GRMZM2G353785 GRMZM2G409407 GRMZM2G439201 GRMZM5G823135 GRMZM5G827266 GRMZM5G831142 GRMZM5G835323 GRMZM5G870606 GRMZM5G882378

Table 3
Number of high-throughput sequencing expression data sets available at Gene Expression Omnibus (<http://www.ncbi.nlm.nih.gov/geo>) as of January, 2016

Taxon	GEO RNA-seq series
Metazoa	4,869
<i>Homo sapiens</i>	1,911
Fungi	398
<i>Saccharomyces cerevisiae</i>	167
Viridiplantae	649
<i>Arabidopsis thaliana</i>	235
<i>Zea mays</i>	62
<i>Oryza sativa</i>	51
Bacteria	415
Archaea	12
Total	6,378

5. We recommend to tick the option “Mask repeats”, as plant genomes are frequently repeat-rich (*see* Fig. 1 and Table 1; the maize genome contains 78 % of REs). This option should not be used if you suspect that the transcription factors of interest bind to repeated sequences.
6. Press “GO” and wait until the retrieve-seq result page is displayed (*see* Note 4). The results include the executed command and a URL to the “sequences” file, which must be saved. We will refer to this URL as “**all.fasta.URL**”. This FASTA-format file can also be stored as a local file on your computer, but note it can be rather large (52 Mb in this example).

3.2 Analyzing Upstream Sequences of Co-expressed Genes

We will now retrieve the upstream sequences of a cluster of co-expressed genes, and use *peak-motifs* to discover exceptional motifs in their promoters. The tool *peak-motifs* was initially conceived to discover motifs in ChIP-seq peaks, but it can also be used to analyze other sequence types, as illustrated here.

1. Choose cluster E2F from Table 2, copy the corresponding gene IDs (last column) and paste them in a new text file that you will store on your computer. Insert newline characters between genes (*see* Note 5).

2. In the left menu of the RSAT server, click on “retrieve sequence” to get a fresh form. Make sure that the option “Genes →selection” is activated and that the right organism, in this case “*Zea_mays.AGPv3.29*”, is selected. Tick “Mask repeats”, and set the same size limits as for the whole collection of promoters: from -1000 to +200. Paste the list of IDs of your gene cluster (one gene ID per row).
3. Press “GO” and wait a few seconds until the result page is displayed. Inspection of these sequences might reveal N-masked sequence stretches, which correspond to annotated repeats. Save both “query genes” and “sequences” files to local files on your computer, we will refer to them as “**cluster.genes**” and “**cluster.fasta**” later on this protocol.
4. Press the “peak-motifs” button. The **peak sequences** section is automatically filled with a link to the selected cluster sequences.
5. Add a title for this job, such as “E2F cluster”.
6. On the right side of “Peak sequences”, under **Control sequences**, paste the “**all.fasta.URL**” on the “URL of a sequence file available on a Web server” entry.
7. Click on “Reduce peak sequences” and leave both fields blank (“number of top sequences to retain” and “cut peak sequences”) to avoid having the sequences clipped.
8. Click on “Motif discovery parameters”. Select two algorithms: “Discover over-represented words” (**oligo-analysis**) and “Discover over-represented spaced word pairs” (**dyad-analysis**). Uncheck the program **position-analysis** (*see Note 6*).
9. Click on “Compare discovered motifs with databases” and select appropriate databases which will be used to annotate any found motifs. For plant promoters, we recommend to check “*footprintDB-plants*”, but you can also check other databases such as “*AthaMap*”, “*ArabidopsisPBM*”, and “*JASPAR plants*” (*see Note 7*). You can also upload your own collection of DNA motifs in TRANSFAC format.
10. Click on “Reporting Options”. Set “Origin” to “end” and “Offset” to -200 (*see Note 8*).
11. Select output type (display or email) and press “GO”.
12. After few seconds the server should have uploaded the sequences and display a page with the URL of the future result page. You can already click on this link: the result page will be periodically updated to show the progress of the analysis. At the end of the processing, a box will appear at the top of the result page, with a short summary of the discovered motifs, and links to different sections of the results. Once the job is complete click on the link [**Download all results (peak-motifs_archive.zip)**] to **save the results** on your computer.

You will later be able to uncompress this archive in order to check the result after its removal from the server (results are only available on the server for 7 days after job completion). We also recommend downloading the full set of discovered motifs, by clicking on the link [**Download all matrices (transfac format)**] and saving a local file named “**cluster.motifs.tf**”. This file contains all motifs in the form of position-weight matrices (PWMs) in TRANSFAC format.

On the result page, the section entitled “*Discovered motifs (with motif comparison)*” lists the discovered motifs, displays their sequence logos and their distribution along clustered sequences, in addition to top matches with the motif databases selected in **step 9**. The top motifs found by *oligo-analysis* and *dyad-analysis* are reported in Table 4.

3.3 Negative Control: Random Groups of Genes

In this section, we propose a procedure to obtain an empirical estimation of the rate of false positives, by discovering motifs in the promoters of genes picked up at random.

1. On the left-side menu of RSAT::Plants select “Build control sets → random gene selection”.
2. Choose “Organism → Zea_mays.AGPv3.29” for the examples of this protocol.

Table 4
Top hexamers and dyads enriched on the E2F cluster of maize upstream sequences and a random cluster of the same size

Cluster	Type	Motif	exp_freq	occ	exp_occ	occ_P	occ_E	occ_sig
E2F	Hexamer	gcggga	0.00046	37	6.65	3.1e-16	6.5e-13	12.19
E2F	Hexamer	cgggaa	0.00031	28	4.55	1.1e-13	2.2e-10	9.66
E2F	Hexamer	cccgcc	0.00072	36	10.49	5.7e-10	1.2e-06	5.93
Random	Hexamer	cttega	0.00032	15	4.78	0.00014	2.9e-01	0.53
Random	Hexamer	ccaaaa	0.00083	27	12.16	0.00016	3.4e-01	0.47
Random	Hexamer	aacacc	0.00046	18	6.78	0.00025	5.2e-01	0.28
E2F	Dyad	gcgn{1}gaa	0.00036	31	5.21	1.3e-14	2.6e-10	9.58
E2F	Dyad	ggcn{1}gga	0.00062	40	8.79	1.3e-14	2.7e-10	9.57
E2F	Dyad	ggcn{2}gaa	0.00042	27	6.00	2.9e-10	6.1e-06	5.22
Random	Dyad	accn{8}aaa	0.00055	23	7.66	5.7e-06	1.2e-01	0.91
Random	Dyad	aatn{3}aaa	0.00126	39	17.95	1.1e-05	2.4e-01	0.62
Random	Dyad	cttn{2}gac	0.00027	15	3.87	1.4e-05	2.9e-01	0.53

Abbreviations: *exp_freq* expected relative frequency, *occ* observed occurrences, *exp_occ* expected occurrences, *occ_P* occurrence probability (binomial), *occ_E* E-value for occurrences, *occ_sig* occurrence significance

3. Set “Number of genes” to the size of one of the sample clusters on Table 2. For instance, the size of the negative control sets would be 18 for cluster E2F, 16 for cluster ABI4, and 56 for cluster WR11. For convenience, in this tutorial only one random group is generated (the default), but this utility can generate several random groups in one go (*see Note 9*).
4. Press “GO” and click the “Next step” button “retrieve sequences” at the bottom of the result page. In the retrieve-seq form, set the other parameters as above: from -1000 to +200, check the “Mask repeats” option and press “GO”.
5. Save “query genes” and “sequences” files to local “**random.genes**” and “**random.fasta**” files and repeat **steps 4–11** of Subheading 3.2. The top motifs found by oligo-analysis and dyad-analysis on such a random cluster are reported in Table 4.

3.4 Validating Motifs by Scanning Promoter Sequences

This part of the protocol is devoted to validating sequence motifs discovered by their over-representation, which are scanned against the original sequences from which they were discovered, plus, optionally, orthologous sequences from a related species (*see Note 10*). The first goal of this section is to check whether the discovered motifs show patterns of occurrence along promoter sequences, and to see how many cluster sequences actually harbor them. This can be done empirically by comparing the results of expression-based motifs with those of shuffled motifs, with columns permuted, which play the role of negative controls. A second goal is to investigate whether these regulatory motifs are conserved on orthologous promoters of a related plant, *Sorghum bicolor* in this case study.

1. On the left-side menu select “Comparative genomics →get orthologs-compara”.
2. Choose “Reference organism →*Sorghum bicolor*” for the maize example.
3. Upload file “**cluster.genes**” generated in **step 3** of Subheading 3.2. Press “GO” and finally press “retrieve sequences” on the next screen.
4. Repeat **steps 4–6** of Subheading 3.1 but now select *Sorghum bicolor* as organism. Save “sequences” to local file “**cluster_orths.fasta**”.
5. On the left-side menu select “Build control sets →permute-matrix”.
6. Upload “**cluster.motifs.tf**” (obtained in **step 12** of Subheading 3.2) and press “GO”. Save the results file as “**cluster.motifs.perm1.tf**” (*see Note 11*).
7. Select “Pattern matching →matrix scan (full options)”.
8. In the sequence box paste the contents of “**cluster.fasta**” and, optionally, “**cluster_orths.fasta**”, if you wish to assess motif

conservation. Alternatively, **steps 7–12** can be performed separately with maize and *S. bicolor* sequences.

9. Upload file “**cluster.motifs.tf**” and select “TRANSFAC” format.
10. In the “Background model” section select Markov order 2 and choose “Organism-specific →Zea_mays.AGPv3.29”. Press “GO”.
11. Save the “Scan result” file as “**cluster.scan.ft**” and press the “feature map” button to draw a map of the matched motif instances.
12. Repeat **steps 6–11** using the set of permuted PWMs “**cluster.motifs.perm1.tf**” and save the results as “**cluster.perm1.scan.tf**”.

3.5 Interpretation of Results

The last stage of the protocol is the interpretation of results, which requires having at hand results of both clusters of co-expressed genes and random clusters, which play the role of negative controls. Figure 2 summarizes the results of clusters in Table 2 compared to 50 random clusters of the same size. There are three types of evidence to look at, which will be discussed with the examples in this figure.

- The **distributions of motif significance** yielded by *oligo-analysis* (A, E, I) and *dyad-analysis* (B, F, J). Motifs discovered in random clusters (grey bars) typically have significances below 4. The motifs found in ABI4 and WRI1 clusters (black bars) are not more significant than those of random gene sets of the same sizes. The reason for having significant motifs in the random gene sets may result from the occasional presence of low complexity motifs, which should not be considered as reliable predictions. In contrast, the most significant oligomer found within E2F upstream sequences clearly supersedes those of random clusters, and a very similar motif is reported by *dyad-analysis*, with a lower but still strong significance. For this reasons, E2F motifs can be considered as promising predictions.
- Panels A, E and I also show the comparisons between some motifs returned by *peak-motifs* and those reported by the authors of the reference experimental study [4]. They used MEME as motif discovery tool. For E2F and WRI1 the different motif discovery tools return similar motifs (logos) with some differences in the matrix width and in the conservation at some positions. Note that this protocol did not produce any motifs matching the binding sequence reported by Yu et al. [4].
- The **distributions of scanning scores** (C, G, K) show to which extent motif matches in upstream sequences of both maize genes and their *S. bicolor* orthologues (dark boxes) depart from matches of permuted matrices (lighter boxes, *see Note 11*), used here as negative controls. On these boxplots, the horizontal bars indicate the median score of all the predicted sites in a

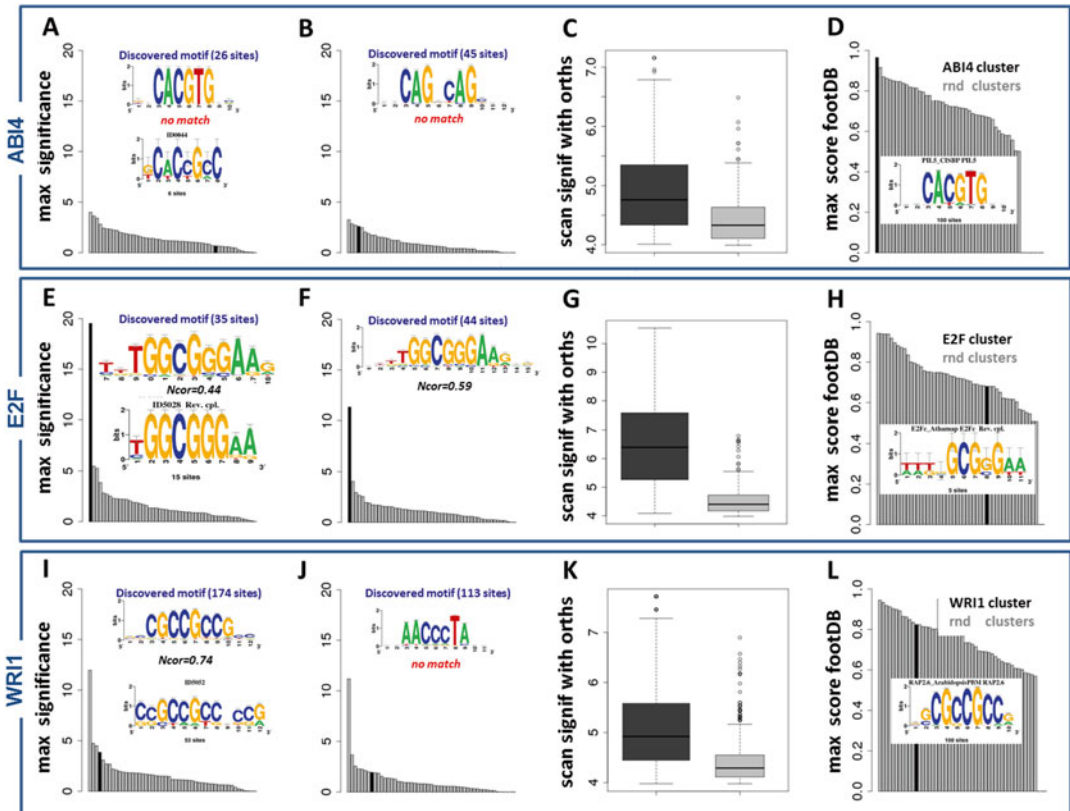


Fig. 2 Summary of motif discovery results with three clusters of maize genes (ABI4, *top*; E2F, *middle*; WRI1, *bottom*) used along the protocol (Table 2). *Dark bars* correspond to clusters of co-expressed genes, grey bars to 50 random clusters of genes drawn from the maize genome. Maximum significance of *oligo-analysis* (A, E, I) and *dyad-analysis* (B, F, J) motifs. The sequence logo of motifs reported by each algorithm is shown on *top*, indicating the number of sites used to compute it and the Ncor score of the comparison to the expected motif (*bottom*) (see Note 12). Note that the *oligo-analysis* sequence logo of E2F was trimmed to fit in the panel, the original has width=20. Panels C, G, K show the scores of discovered motifs when scanned back to the original maize upstream sequences and sequences from orthologous genes in *Sorghum bicolor*. Here *dark bars* are the reported PWMs, while the *grey bars* correspond to permuted PWMs. Panels D, H, L show the Ncor scores of discovered motifs when compared to annotated PWMs in footprintDB. A full report including cluster MYB59 can be browsed at http://plants.rsat.eu/data/chapter_expression_clusters

given set of promoter sequences, and the shaded rectangles show the interquartile range, i.e. the extent between the 25th and 75th percentiles. In the example, the results for E2F motifs confirm their relevance (Fig. 2g): the interquartile range of the E2F cluster (dark rectangle) is clearly separated from the corresponding rectangle of the random selections (gray box). For the ABI4 cluster (Fig. 2c), there is a noticeable overlap between the interquartile boxes of the cluster and the random gene selections. Besides, the random selections show several “outliers” (circles) indicating sites predicted with high matching

scores. Even though the mean scores are clearly higher for the actual cluster, the results may thus not be considered very significant. WRI1 results show a somewhat intermediate situation, where the interquartile boxes show a moderate overlap, but the random gene selections frequently bear relatively high-scoring sites (circles) for the discovered motifs.

- The **distributions of scores in footprintDB** (D, H, L) describe how similar the discovered motifs are when compared to motifs (PWMs) annotated in footprintDB. Similarities are measured by the normalized correlation score (*Ncor*, see **Note 12**). In each example 50 random sets of promoters were analyzed with *peak-motifs*, and the discovered motifs compared to footprintDB. The black bar indicates the best matching score for the original, expression-based gene clusters, and the corresponding logo is overlaid on the histogram. For E2F and WIR1, the best matching motifs correspond to the motifs experimentally confirmed by Yu et al. [4]. However, in both cases motifs discovered from random gene selections present even better matching scores with some motif database. This result indicates that the matching score between a discovered motif and a repository, while essential for annotation purposes (identifying putative factors for a given gene cluster), is not particularly helpful in order to distinguish relevant expression-supported motifs from PWMs constructed from random sequence clusters. For ABI4, the best-scoring matches correspond to phytochrome interacting factors. These proteins belong to the bHLH family of transcription factors and there are many annotated motifs for them in databases such as footprintDB.

In summary, motifs discovered in promoters of co-expressed genes should always be evaluated based on a combination of complementary criteria:

1. The primary key of interpretation is the significance reported by the motif discovery algorithms. This significance has to be interpreted by comparison with the results obtained in random promoter sets of the same size as the gene cluster of interest (negative controls).
2. Sequence scanning permits to predict putative binding sites, but the matching scores should be evaluated relative to randomized motifs (column-permuted).
3. Comparison between discovered motifs and databases of known TF-binding motifs suggests candidate transcription factors which could intervene in the co-regulation of the co-expressed cluster.

4 Notes

1. As gene models can change from one assembly to another it is important to use the right assembly version, which is indicated for each genome in Table 1. If the assembly of interest is not available on the RSAT::Plant server, please contact the first author.
2. Twelve clusters of maize genes, found to be co-expressed in 22 transcriptomes and enriched on Gene Ontology terms (<http://geneontology.org>) [11], were analyzed in detail by Yu et al. [4]. First, they discovered potential regulatory motifs within their upstream sequences, and then they performed electrophoretic mobility shift assays (EMSA) to confirm them. Table 2 shows three of those clusters which are used in this protocol. For each cluster a list of gene identifiers is given next to the EMSA-confirmed motifs. The remaining clusters were left out for being too small, as the statistical approaches in this protocol require at least ~10–15 genes. Cluster MYB59 was left out due to space restrictions but its results can be browsed at http://plants.rsat.eu/data/chapter_expression_clusters/
3. A crucial parameter to evaluate the results of motif discovery is to estimate the rate of false positives (FP). RSAT programs compute a significance score, which is the minus log of the expected number of false positives (e-value = $10^{-\text{signif}}$). For example, a motif associated with a significance of 1 should be considered as poorly significant, since on average we would expect $10^{-1} = 0.1$ false positives, i.e. one FP every ten random trials. In contrast, a significance of, e.g. 16 is very promising, since on average such a result would be expected every 10^{-16} random trials. However, the theoretical significance relies on the correctness of the background model (computed here as k-mer and dyad frequencies in the whole set of promoters). In some cases, sets of plant promoters can deviate from the theoretical model, due to heterogeneity of the input (e.g. inclusion of repetitive sequences). The negative control consists in measuring the significance obtained by submitting a random selection of promoters from the organism of interest (maize in the example). Although each of these genes is likely to be regulated by one or more transcription factors (and its promoter should contain corresponding binding sites), in principle the random set as a whole should not be co-regulated, so that the elements would differ from gene to gene, and there should thus be no over-represented motif in their promoters.
4. Should the connection to the server interrupt it might be safer to go back and choose “email” as delivery option. The mail message provides a link to the data, which is actually stored at the server.

5. It is crucial to have one gene ID per row for submitting queries to retrieve-seq, because only the first word of each row is considered as a query.
6. This program is generally relevant when analyzing sets containing a large number of sequences such as ChIP-seq peaks or genome-wide promoter sets.
7. Plant transcription databases are unfortunately still very fragmentary, so one might be tempted to check more complete collections such as *footprintDB* or *JASPAR core all*. However, the results should be interpreted with caution, because there is no conservation of *cis*-regulation between plants and other kingdoms of the tree of life.
8. The option “*Origin*” indicates the reference position relative to each sequence (start, center, or end). When this option is set to “end”, the coordinates are computed relative to the end of the sequence, with negative values indicating upstream location. The option “*Offset*” enables to shift the reference point by a given number. For the current example, setting the offset to -200 will give coordinates from -1000 to +200, the 0 corresponding to the TSS.
9. Clearly, more than one random cluster should be evaluated, as suggested in Fig. 2, where the results of up to 50 random groups are displayed next to the clusters of [4].
10. Orthologues reported are annotated in Ensembl Compara, generated by a pipeline where maximum likelihood phylogenetic gene trees play a central role. These gene trees, reconciled with their species tree, have their internal nodes annotated to distinguish duplication or speciation events, and thus support the annotation of orthologous and paralogous genes, which can be part of complex one-to-many and many-to-many relations. Adapted from: http://www.ensembl.org/info/genome/compara/homology_method.html.
11. This will permute the columns of input PWMs producing matrices with different consensus. Column-permuted matrices are used as negative controls because they conserve the information content and nucleotide frequencies of the original motifs, but at the same time alter the sequence of nucleotides captured by the original motif, which is not recognized anymore.
12. “Ncor” is the relative width-normalized Pearson correlation of two PWMs aligned with *matrix-scan*. This normalized score prevents spurious matches that would cover only a subset of the aligned matrices (e.g. matches between the last column of the query matrix and the first column of the reference matrix, or matches of a very small motif against a large one).

Acknowledgements

This work was funded in part by Fundación ARAID and by the Enseignants-Chercheurs invités program of Aix-Marseille Université (to BCM). CR was supported by the France Génomique National infrastructure, funded as part of the Investissements d'Avenir, program managed by the Agence Nationale pour la Recherche (contract ANR-10-INBS-09).

References

1. Beer MA, Tavazoie S (2004) Predicting gene expression from sequence. *Cell* 117:185–198
2. Sand O, Turatsinze TJ, van Helden J (2008) Evaluating the prediction of cis-acting regulatory elements in genome sequences. In: Valencia A, Frishman D (eds) *Modern genome annotation*. Springer Verlag, Wien, pp 55–89
3. Koschmann J, Machens F, Becker M, Niemeyer J, Schulze J, Bülow L, Stahl DJ, Hehl R (2012) Integration of bioinformatics and synthetic promoters leads to the discovery of novel elicitor-responsive cis-regulatory sequences in *Arabidopsis*. *Plant Physiol* 160:178–191
4. Yu CP, Chen SC, Chang YM, Liu WY, Lin HH, Lin JJ, Chen HJ, Lu YJ, Wu YH, Lu MY, Lu CH, Shih AC, Ku MS, Shiu SH, Wu SH, Li WH (2015) Transcriptome dynamics of developing maize leaves and genomewide prediction of cis elements and their cognate transcription factors. *Proc Natl Acad Sci U S A* 112:E2477–E2486
5. Schmidt T, Heslop-Harrison J (1998) Genomes, genes and junk: the large-scale organization of plant chromosomes. *Trends Plant Sci* 3:195–199
6. Kersey PJ, Allen JE, Armean I et al (2016) *Ensembl Genomes 2016: more genomes, more complexity*. *Nucleic Acids Res* 44:D574–D580
7. Nordberg H, Cantor M, Dusheyko S, Hua S, Poliakov A, Shabalov I, Smirnova T, Grigoriev IV, Dubchak I (2014) The genome portal of the Department of Energy Joint Genome Institute: 2014 updates. *Nucleic Acids Res* 42(1):D26–D31
8. Tanaka T, Sakai H, Fujii N, Kobayashi F, Itoh T, Matsumoto T, Wu J (2013) *bexdb: bioinformatics workbench for comprehensive analysis of barley-expressed genes*. *Breed Sci* 63:430–434
9. Sebastian A, Contreras-Moreira B (2014) *footprintDB: a database of transcription factors with annotated cis elements and binding interfaces*. *Bioinformatics* 30:258–265
10. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A (2013) *NCBI GEO: archive for functional genomics data sets—update*. *Nucleic Acids Res* 41:D991–D995
11. The Gene Ontology Consortium (2015) *Gene Ontology Consortium: going forward*. *Nucleic Acids Res* 43:D1049–D1056
12. Lamesch P, Berardini TZ, Li D, Swarbreck D, Wilks C, Sasidharan R, Muller R, Dreher K, Alexander DL, Garcia-Hernandez M, Karthikeyan AS, Lee CH, Nelson WD, Ploetz L, Singh S, Wensel A, Huala E (2012) *The Arabidopsis Information Resource (TAIR): improved gene annotation and new tools*. *Nucleic Acids Res* 40:D1202–D1210

RSAT::Plants: Motif Discovery in ChIP-Seq Peaks of Plant Genomes

Jaime A. Castro-Mondragon*, Claire Rioualen*,
Bruno Contreras-Moreira, and Jacques van Helden

Abstract

In this protocol, we explain how to run ab initio motif discovery in order to gather putative transcription factor binding motifs (*TFBMs*) from sets of genomic regions returned by ChIP-seq experiments. The protocol starts from a set of peak coordinates (genomic regions) which can be either downloaded from ChIP-seq databases, or produced by a peak-calling software tool. We provide a concise description of the successive steps to discover motifs, cluster the motifs returned by different motif discovery algorithms, and compare them with reference motif databases. The protocol is documented with detailed notes explaining the rationale underlying the choice of options. The interpretation of the results is illustrated with an example from the model plant *Arabidopsis thaliana*.

Key words Chromatin immunoprecipitation DNA-sequencing (ChIP-seq), Transcription factor (TF), Transcription factor binding motifs (TFBM), Transcription factor binding site (TFBS), Gene ontology (GO), Functional enrichment

1 Introduction

1.1 The ChIP-seq Technology

The ChIP-seq method [1, 2], which enables one to characterize transcription factor binding sites (*TFBS*) or chromatin marks in a whole genome, has gained a tremendous popularity to study genetic and epigenetic regulation. Although the main field of application so far has been Human and model organisms (Table 1), the ChIP-seq technology opens wide perspectives for the analysis of plant regulation.

Chromatin immunoprecipitation, followed by high-throughput sequencing and mapping on a reference genome, shows regions with high enrichment in reads. These regions, so-called *ChIP-seq peaks*, can be detected by using *peak-calling* algorithms. They typically encompass a few hundreds base pairs, and

*The authors of this chapter contributed equally to the work.

Table 1

ChIP-seq samples per taxa. Number of ChIP-seq samples available in the Gene Expression Omnibus database [13] (Dec 18, 2015) per taxonomic group (see Note 14)

Taxon	GEO ChIP-seq series
<i>No taxon specified (any taxon)</i>	4722
Metazoa	4255
<i>Homo sapiens</i>	1542
<i>Mus musculus</i>	1793
<i>Caenorhabditis elegans</i>	410
<i>Drosophila melanogaster</i>	542
Fungi	238
<i>Saccharomyces cerevisiae</i>	163
Viridiplantae	157
Bacteria	64
<i>Escherichia coli</i>	24
Alveolata	14
Archaea	1

are centered on a binding site for the immunoprecipitated transcription factor (*TF*). They thus need to be further processed in order to discover transcription factor binding motifs (*TFBM*) and define the precise locations of the binding sites.

The characterization of *TFBM* from ChIP-seq experiments presents several advantages:

1. ChIP-seq peaks provide a relatively precise information about TF binding locations (~200 bp precision). This makes a drastic difference with the approaches based on co-expression clusters (transcriptome arrays, RNA-seq), in particular for multicellular organisms (Metazoa, Plants), where regulatory regions can be found not only in the upstream promoter, but also in introns, downstream, and dispersed over wide distances.
2. The transition from ChIP-chip to ChIP-seq yet increased the precision of genome-wide location analyses.
3. Motifs discovered in ChIP-seq peaks are typically built from several hundreds or thousands of binding sites, and are thus much more robust than the previous-generation motifs built from a handful of sites that had been gathered one by one with Electrophoretic Mobility Shift Assays (EMSA) or footprint (low throughput) experiments.

4. Peak collections better reflect the in vivo diversity of binding sites for the TF of interest than in vitro methods such as Systematic Evolution of Ligands by EXponential Enrichment (SELEX).
5. Since peaks encompass a few hundred base pairs, they contain binding sites not only for the immunoprecipitated factor, but also for other interacting factors. Ab initio motif discovery thus enables us to detect additional motifs, and infer putative partners of the studied factor.

The knowledge gained from analyzing motifs and sites in ChIP-seq peaks may be used to enforce the design of synthetic promoters by predicting potentially important interactions between multiple TF (i.e. co-occurring motifs), synthetic promoters, and native promoters of the target species.

Since ChIP-seq peaks typically encompass several megabases or tens of megabases, specialized bioinformatics tools have been developed to discover motifs ab initio and scan the peaks for putative binding sites [3–6]. In this chapter, we explain how to combine the motif discovery workflow *peak-motifs* [5, 6] and some other tools of the Regulatory Sequence Analysis Tools (RSAT, <http://rsat.eu/>) [7] to discover and interpret TFBMs from plant ChIP-seq peaks.

1.2 Principle of the ChIP-seq Technology

The principle of Chromatin Immunoprecipitation sequencing (ChIP-seq) technology [1, 2] is to cross-link a DNA-binding protein (TF, histone) with its bound DNA, shear the DNA by ultrasonication, immunoprecipitate the protein of interest, release the cross-link, select DNA fragments of reasonable size (~300 bp), and sequence their extremities (NGS sequencing is typically restricted to sequences smaller than the fragments). The primary result of a ChIP-seq experiment is a file with *raw short reads* (typically 36–75 bp), which can be mapped onto a reference genome.

Figure 1a shows the density profile of ChIP-seq reads for the transcription factor MYB3R3, mapped onto chromosomes 1 and 2 of the genome of *A. thaliana* (TAIR10 assembly version). This primary view of the data reveals a first difficulty for the interpretation of ChIP-seq data: some genomic regions are covered by a huge number of reads. These regions correspond to repetitive elements in centromeric and telomeric regions of the chromosomes. For the sake of comparison, Fig. 1b shows the density profile of a control experiment where the ChIP-seq protocol was run with an anti-GFP antibody, supposed to give an unspecific signal. This mock experiment reveals the same hyper-mapped regions, and can serve to estimate background and discard unspecific reads for the *peak-calling*. Note that mock experiments generally give reduced libraries. An alternative way to estimate unspecific background is to sequence genomic DNA without applying the immunoprecipitation procedure (*genomic input*).

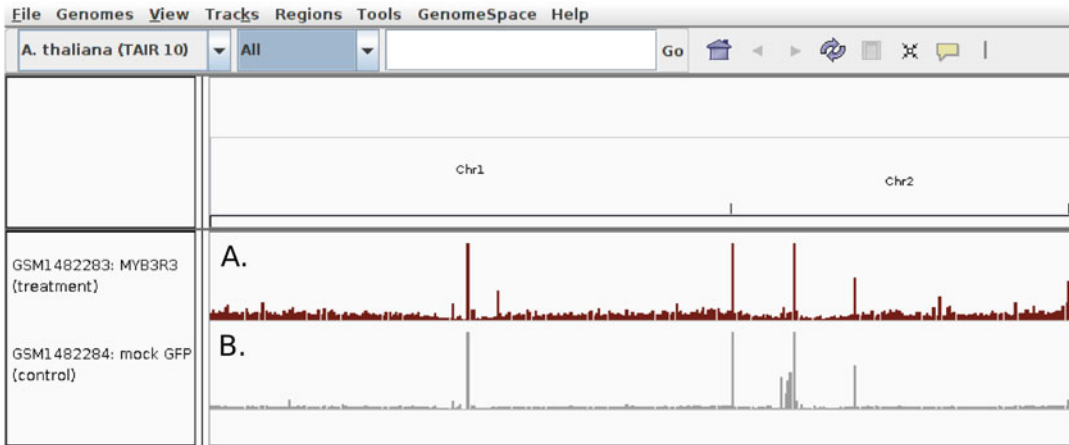


Fig. 1 Density profiles of reads mapped on chromosomes 1 and 2. (a) MYB3R3-bound immunoprecipitated chromatin. Reads were mapped on the TAIR10 assembly of *Arabidopsis thaliana* genome. (b) Control experiment (mock with anti-GFP antibody). Reads from the control experiment are used as “input” for the peak-calling, which enables peak-callers to avoid reporting peaks in the repetitive regions. In both ChIP and control tracks, note the striking concentration of reads in particular genomic locations, corresponding to repetitive regions. The map was generated with the Interactive Genome Viewer [12]

1.3 Choice of a Peak-Caller and Tuning of Its Parameters

One of the most crucial steps of the ChIP-seq analysis is the choice of a peak-calling program and the tuning of its parameters.

The *peak-calling* procedure consists in identifying genomic regions presenting a significant enrichment in reads in the ChIP-seq data, compared to some control set. The control set can either be a mock experiment, as in Fig. 1b, or a full-genome sequencing. A large number of different programs exist for peak-calling [8, 9].

Figure 2 shows a detailed view of the peaks identified by some popular peak-callers on an arbitrary genomic region of the MYB3R3. Note the difference between the numbers and widths of the peaks, depending on the peak-calling tool. One of the most popular peak-calling programs, MACS, comes in two releases [10]. The first version, MACS14, tends to return wide regions encompassing several topological peaks (compare the peaks with the MYB3R3 density profiles). MACS2, an upgraded version of MACS14, allows to specify parameters to obtain narrower peaks. Homer [11], based on the findPeaks algorithm, outputs very sharp peaks. The series of SWEMBL [12] peaks illustrates the impact of the parameters. This peak-caller proposes a “gradient” option (*-R*), which strongly affects the number of peaks and their width. SPP [13], using the FDR as a main parameter, is also to be carefully configured.

Most publications rely on the prior choice of a popular peak-caller, which is run with default parameters. Table 2 shows the wide range of peaks that can be found in a single dataset depending on the peak-calling algorithm and its configuration. However, the most appropriate algorithm and, even more, the fine-tuning of its

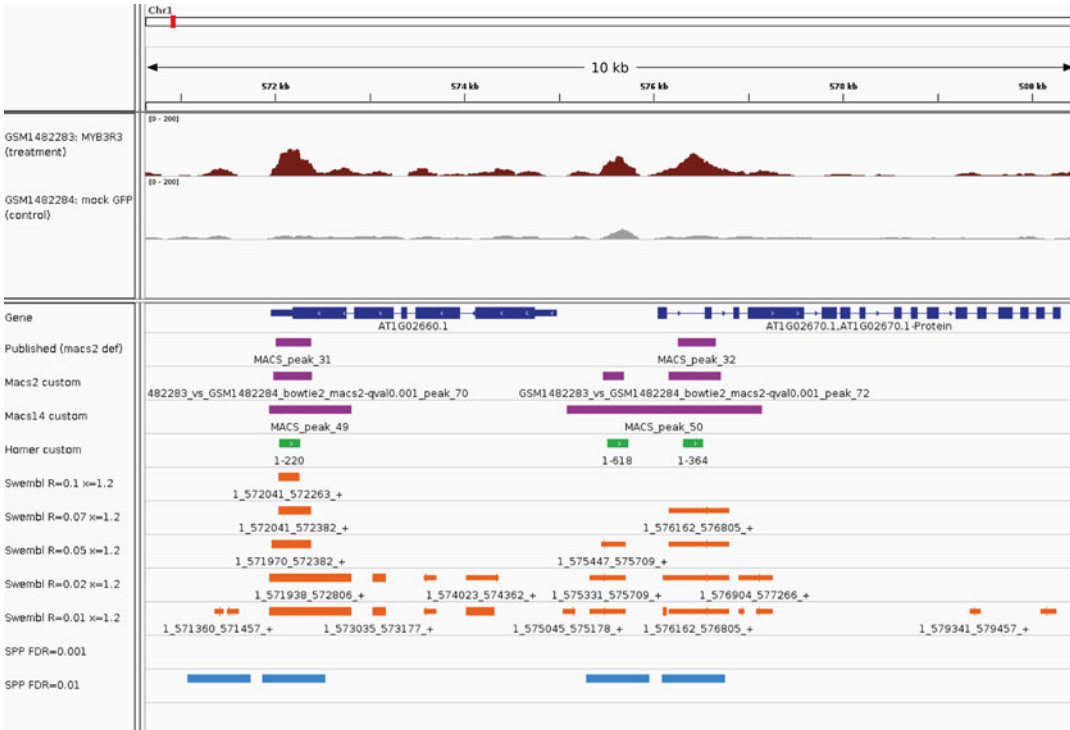


Fig. 2 Peak profiles obtained with a variety of peak-calling algorithms and parameters. Zoom of the reads and peaks in an illustrative region of chromosome 1 (coordinates 570,625–580,625). Each peak-caller is denoted by a specific color: MACS (*pink*) [10], Homer (*green*) [11], SWEMBL (*orange*) [12], and SPP (*cyan*) [13]. Two peaks are detected by most peak-callers, although with different widths. One of these peaks is located in a gene promoter (at 572 kb), and another one within an intron (576 kb). The sensitivity of each peak-caller can be tuned with some specific parameters, as illustrated with the SWEMBL series (sensitivity increases from *top* to *bottom*) or SPP (false discovery rate set to 0.001 or 0.01, resp.). Relatively stringent settings are recommended to obtain a good trade-off between sensitivity and relevance of the peaks

parameters depend on the organism, data type, and even the purpose of the analysis (gathering high-confidence binding locations, identifying likely target genes, building a transcription factor binding motif, etc.) [9]. There is unfortunately no gold standard that would permit to assess the relative merits of peak-callers, and define their optimal parameters.

However, a variety of criteria can be used to evaluate the relevance of the returned peaks by various indirect indications, some of which will be illustrated in this protocol:

- Enrichment of the reference motif (annotated motif for the immunoprecipitated factor) in the peak sequences (RSAT *matrix-quality*);
- Concentration of the reference motif at peak centers;
- Significance of the motifs discovered by ab initio approaches (RSAT *peak-motifs*);

Table 2

Contingency tables comparing peak-caller results

Peak-caller	Macs2 (qval=0.05)	Macs2 (qval=0.001)	Macs14 (pval=0.000001)	Homer (fdr=0.01)	SPP (fdr=0.01)	SPP (fdr=0.001)	SWEMBL (R=0.1)	SWEMBL (R=0.07)	SWEMBL (R=0.05)	SWEMBL (R=0.02)	SWEMBL (R=0.01)
Macs2 (qval=0.05)	2931	3335	2699	2854	3408	532	1298	2224	2704	2848	3263
Macs2 (qval=0.001)	2930	9711	5494	8136	9576	544	1340	2767	5020	8840	11,973
Macs14 (pval=0.00001)	2895	7360	6242	7359	9767	535	1325	2659	4510	9518	17,225
Homer (fdr=0.01)	2851	8884	6114	18,812	16,898	534	1328	2743	5091	17,125	24,503
SPP (fdr=0.01)	2920	9291	6166	15,364	24,781	544	1333	2751	5104	20,399	39,018
SPP (fdr=0.001)	534	640	533	534	680	544	532	533	536	557	654
SWEMBL (R=0.1)	1374	1764	1294	1377	1786	534	1352	1343	1340	1368	1502
SWEMBL (R=0.07)	2355	3606	2561	2861	3679	535	1352	2788	2765	2864	3424
SWEMBL (R=0.05)	2852	6302	4224	5316	6697	538	1352	2787	5256	5541	7277
SWEMBL (R=0.02)	2931	9692	6236	16,734	20,518	544	1352	2788	5256	31,867	41,904
SWEMBL (R=0.01)	2931	9710	6242	18,611	24,365	544	1352	2788	5256	31,864	92,695

Each cell indicates the number of peaks of one peak-calling result covered by peaks of another peak-calling result. The diagonal (in bold) indicates the number of peaks detected by each one of them. *See* also Fig. 9 for a heatmap of the relative frequencies

		Macs2_qval0.05	Macs2_qval0.001	Macs14_pval0.00001	Homer_fdr0.01	SPP_fdr0.01	SPP_fdr0.001	SWEMBL_R0.1	SWEMBL_R0.07	SWEMBL_R0.05	SWEMBL_R0.02	SWEMBL_R0.01
Macs2_qval0.05	2931	1	0.34	0.43	0.15	0.14	0.98	0.96	0.8	0.51	0.09	0.04
Macs2_qval0.001	9711	1	1	0.88	0.43	0.39	1	0.99	0.99	0.96	0.28	0.13
Macs14_pval0.00001	6242	0.99	0.76	1	0.39	0.39	0.98	0.98	0.95	0.86	0.3	0.19
Homer_fdr0.01	18812	0.97	0.91	0.98	1	0.68	0.98	0.98	0.98	0.97	0.54	0.26
SPP_fdr0.01	24781	1	0.96	0.99	0.82	1	1	0.99	0.99	0.97	0.64	0.42
SPP_fdr0.001	544	0.18	0.07	0.09	0.03	0.03	1	0.39	0.19	0.1	0.02	0.01
SWEMBL_R0.1	1352	0.47	0.18	0.21	0.07	0.07	0.98	1	0.48	0.25	0.04	0.02
SWEMBL_R0.07	2788	0.8	0.37	0.41	0.15	0.15	0.98	1	1	0.53	0.09	0.04
SWEMBL_R0.05	5256	0.97	0.65	0.68	0.28	0.27	0.99	1	1	1	0.17	0.08
SWEMBL_R0.02	31867	1	1	1	0.89	0.83	1	1	1	1	1	0.45
SWEMBL_R0.01	92695	1	1	1	0.99	0.98	1	1	1	1	1	1

Fig. 9 Heatmap of mutual coverage of peak-calling results. The *second column* indicates the number of peaks depending on the peak-calling program and the main parameters affecting the stringency of the result. *Further columns* indicate the proportion of peaks of one peak-calling result (*row*) covered by peaks of another peak-calling result (*columns*)

- Biological relevance of the transcription factors putatively bound to the discovered motifs (*FootprintDB* search);
- Functional enrichment of the genes linked to the peaks (Gene ontology);
- Concentration of the discovered motifs at the peak centers (RSAT *position-analysis*);

1.4 The Plant Regulatory Sequence Analysis Tools

Regulatory Sequence Analysis Tools (RSAT, <http://rsat.eu/>) is a specialized software suite for the analysis of cis-regulatory elements in genomic sequences [7]. Since 2015, the services have been distributed on taxon-specific servers, including a Plant RSAT (<http://plants.rsat.eu/>). This address will redirect you to the host server <http://floresta.cead.csic.es/rsat>, which will be used for this protocol.

1.5 Functional Interpretation of ChIP-Seq Peaks

RSAT supports several approaches to interpret the peaks in functional terms:

1. *Motif enrichment.* In some cases, the immunoprecipitated factor is already known, and a reference motif exists in some database.

It is generally a good practice to start by measuring the enrichment of the peak set for this reference motif, in order to check that the procedure went fine (from the wet lab to the bioinformatics workflow that produced the peaks).

2. *Motif discovery.* Several ab initio methods can be used to detect exceptional motifs in the peak sequences, based on different criteria: over-representation, biased positional distribution relative to the peak centers, etc.

1.6 Transcription Factor Binding Motifs

Transcription Factor Binding Motifs (*TFBMs*) are generally represented as Position-Specific Scoring Matrices (*PSSMs*). They are built from an alignment of TF binding sites. Each cell of the matrix indicates the frequency of a given nucleotide (matrix rows) in a given column of the aligned sites (matrix columns). They can be depicted as sequence logos [14].

The widespread use of high-throughput technologies, for example ChIP-seq, allows to discover novel TFBMs or improve the quality of those existing (i.e. by increasing the number of sites to build the TFBMs). As more TFBMs are available, repertoires are required to give an easy access to these motifs. Currently there are many public and private motif databases, some of them specialized on few organisms (Athamap for *Arabidopsis thaliana*; Hocomoco for Human and Mouse, etc.) and others have taxon-wide collections of TFBMs (Jaspar, TRANSFAC, CisBP) for plants, vertebrates, fungi, insects, etc. However, as these databases are growing, and since a single new study could produce an entire collection of motifs [15], efforts to collect, integrate and update many motif databases must be done. One option is FootprintDB [16] which is a meta-database encompassing 14 up-to-date motif databases (*see* Chapter 17).

In this protocol, we show how to run ab initio discovery on a set of ChIP-seq peak sequences, compare discovered motifs with a reference motif database, and cluster the discovered motifs to obtain a non-redundant collection.

2 Materials

2.1 Required Software

This protocol requires to dispose of

- a computer with any Web browser installed;
- a set of peak coordinates from a ChIP-seq or related experiment.

For visualization purposes (Figures 1, 2), we also recommend to install the Integrative Genome Viewer [17].

2.2 Data Sources

Peaks can be obtained either from NGS databases [18, 19] or by running a peak-calling software tool on genome-mapped reads. This protocol starts from pre-computed peak coordinates, and does not cover the read mapping and peak calling procedures.

2.3 Data Formats

Peak coordinates should be provided in *bed* format (*see* the description of NGS file formats at the UCSC genome browser (*see* **Note 1**)). Alternatively, this protocol can be run with peak sequences in *fasta* format (in which case the sequence retrieval steps can be skipped).

2.4 Study Case

As a study case we take a recent MYB3R3 study [20]. We will use a BED file available at the Gene Expression Omnibus Database (GEO), under accession GSE60554 (<http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE60554>), which contains the results of a ChIP-seq experiment with the MYB3R3 transcription factor of *Arabidopsis thaliana*. The peaks can be found at the bottom of the GEO Web page for the MYB3R3-ChIP-ped sample (GSM1482283, <http://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSM1482283>, peak file *GSM1482283_MYB3R3-GFP_ChIP_peaks.bed.txt.gz*) (*see* **Note 2**).

The reference motif for this case is that of c-Myb in tobacco [21], likely to be similar to MYB3R3 in *Arabidopsis thaliana*.

3 Methods

3.1 Retrieval of Peak Sequences from the Peak Coordinates

1. Obtain a bed-formatted list of peak coordinates (*see* **Note 3**).
2. Open a connection to the Plant Regulatory Sequence Analysis Tools server (<http://plants.rsat.eu/>).
3. On the left-side panel, open the toolbox “**Sequence tools**” and click “**sequences from bed/gff/vcf**”.
4. Choose the appropriate genome in the **Organism** pop-up menu (*see* **Note 4**). For the study case, the reference organism is *Arabidopsis thaliana.TAIR10.29*, where the suffix *TAIR10* indicates the assembly, and the number *29* the EnsemblGenome version.
5. Enter the **Genomic coordinates** of your peaks (*see* **Note 5**). Coordinates can be entered in different ways: (1) directly pasted in the text area; (2) large files can be uploaded from your computer to the server (option **Choose file**); (3) enter the *URL of a coordinates file available on a Web server* (e.g. BED file on your account of a Galaxy server). For the study case you can enter the downloaded file *GSM1482283_MYB3R3-GFP_ChIP_peaks.bed.txt.gz*.
6. Verify that the option **Mask repeats** is checked, as plant genomes are often repeat-rich (*see* **Note 6**).

7. For the **Output** option, choose *server*, and click **GO** to submit the job.
8. After a few seconds, a result page (shown in Fig. 3) should appear with the links to the FASTA file containing the peak sequences, plus some additional links to the input BED file and a log file. Note that the results are kept on the server for a restricted duration (72 h). If you want to keep track of the results, you can right-click on the fasta sequence file and download it to your computer.

At this stage of the protocol, you should have at your disposal a file containing peak sequences in fasta format. Typical peak sets include a few hundreds to tens of thousands of peaks, with lengths varying from tens to hundreds of base pairs each.

Note that the results page contains links to other RSAT tools. These enable you to transfer the obtained fasta file directly to the next step of the analysis.

3.2 *Ab Initio Motif Discovery in ChIP-Seq Peak Sequences*

We will now describe the way to discover motifs from ChIP-seq peak sequences. We obtained these sequences in the previous section, in the form of a fasta file, but it is also possible to upload your own fasta file from your computer directly in the *peak-motifs* section. We assume here that the sequences are transferred from the previous step.

1. At the bottom of the sequence retrieval result page, the **Next step** box presents a series of buttons to transfer the fasta sequences to another tool for further analyses (Fig. 3). Click on the *peak-motifs* button. This will display a new Web form shown in Fig. 4, pre-loaded with the URL of the peak sequences.

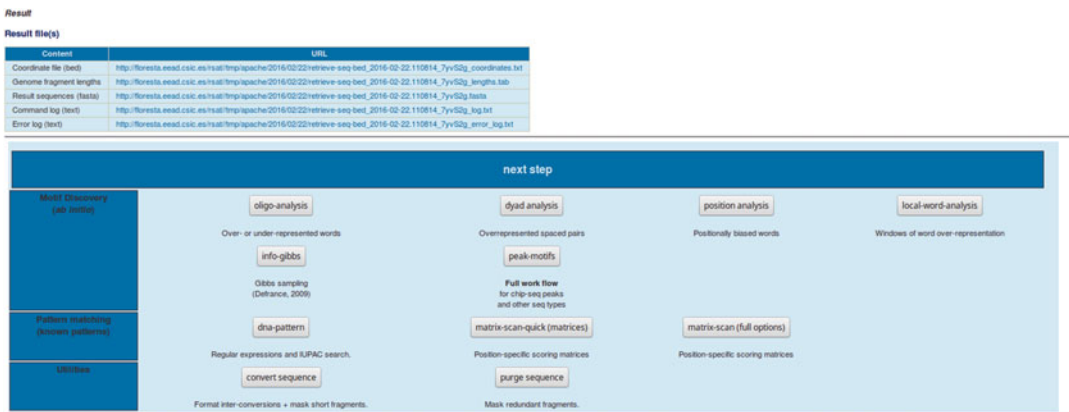


Fig. 3 Results of the sequences retrieval procedure. View of the result page from the sequence retrieval step, made using a BED file [15] and the tool “sequences from bed/gff/vcf”. Next analysis steps can be processed with by simply clicking the corresponding buttons

RSAT - peak-motifs

Pipeline for discovering motifs in massive ChIP-seq peak sequences.

References

1. Thomas-Chollier, M., Herrmann, C., Defrance, M., Sand, O., Thieffry, D. and van Helden, J. (2011). RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets Nucleic Acids Research doi:10.1093/nar/gkr1104, 9. [Open access]
2. Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J. (2012). A complete workflow for the analysis of full-size ChIP-seq (and similar) data sets using peak-motifs. Nat Protoc 7(8): 1551-1568. [PMID 22836136]

► Information on the methods used in peak-motifs

► Reduce peak sequences

► Motif discovery parameters

► Compare discovered motifs with databases (e.g. against Jaspur) or custom reference motifs

► Locate motifs and export predicted sites as custom UCSC tracks

► Reporting options

Output display email

Note: email output is preferred for very large datasets or many comparisons with motifs collections

Fig. 4 View of the *peak-motifs* form. Two fields are required in order to proceed with the analysis: “title” and “peak sequence”. Here, the sequence file was automatically uploaded from the previous step

2. Before running *peak-motifs*, you are requested to type a **Title** for the job. For the study case, we can for example type “*A. thaliana* MYB3R3 versus GFP - GSM1482283”.
3. The **Reduce peak sequences** frame allows you to trim the number and length of the peaks. By default all peaks are retained but those longer than 1 Kb (500 bp on either side of the peak center) are shortened, because they are suspected to result from peak-calling artifacts rather than to represent trustworthy binding sites.
4. The **Motifs discovery** frame permits to choose the discovery algorithms and tune their parameters. By default only *oligo-analysis* and *position-analysis* are activated (see **Note 7**).
5. Under **Motifs discovery** activate **oligomer lengths** 6 and 7 (see **Note 8**).
6. Check that the **Markov order** is set to *automatic (adapted to sequence length)* (see **Note 9**).

7. Check that the **Number of motifs per algorithm** is set to 5 (*see Note 10*).
8. Under **Compare discovered motifs with databases**, you can select one or more motif collections in order to annotate any discovered motifs. For plant sequences we recommend *footprintDB-plants*, which integrates motifs from diverse public databases (*see Chapter 17*).
9. Optionally, the button below **Add your own motif database** allows you to upload a custom database of transcription factor binding motifs in a TRANSFAC-formatted file.
10. If there is a known motif for the immunoprecipitated factor, you can upload it with option **Add known reference motifs for this experiment** (*see Note 11*).
11. Click on the title **Locate motifs and export predicted sites**, check the option **Search putative binding sites in the peak sequences**, and activate the option **Peak coordinates specified in fasta headers in bedtools getfasta format (also for retrieve-seq-bedoutput)**. Here, we assume that the sequences were obtained from RSAT *retrieve-seq-bed* as indicated above (*see Note 12*) but some alternative formats are also supported.
12. You can type in your **email address** to be notified of the job submission and completion, or you can choose **display**, and click **GO**.

After a few seconds, the server displays a confirmation of the job submission, with a link to the result Web page. Clicking on this link will open the result page on a separate tab of your Web browser. This page will be progressively updated to show the results of the analysis. A typical analysis should take from a few minutes to 1 h, depending on the sequence size and the selected options (motif discovery algorithms, motif databases, sequence scanning).

13. Results will progressively be displayed on this page. Once the job is completed, a summary of all results will appear in a box at the top of the results page. After completion of the *peak-motifs* workflow, we recommend to **download the results** on your computer for further analyses, since they are kept on the server for a restricted time.
 - (a) Clicking on the link **Download all results**, in the header box of the result Web page, will allow you to save a zipped file containing the whole HTML report. You will thus be able to visualize these pages locally on your computer.
 - (b) Right-clicking on the link **Download all matrices (TRANSFAC format)** and saving it as *peak-motifs_motifs_discovered.tf* will allow you to keep a file containing all the motifs matrices. This file contains all discovered motifs, in

the flat-file motif description format designed for the TRANSFAC database (this format is convenient because it allows to associate annotations to each motif). We will use it below in the section about matrix clustering (Subheading 3.3).

- (c) In the **Sequence composition (test sequences)** section, right-click on the link “[coordinates: UCSC BED track]” (right panel) and save the BED file as *peak-motifs_test_seqcoord.bed*. This file contains the peaks used for the peak-motifs analysis.
- (d) At the bottom of the Web page, look for section **Motif locations (sites)**, then **Predicted sites on test peaks (all motifs)**. Right-click on the “[bed]” link to download the corresponding file *peak-motifs_all_motifs_seqcoord.bed*. This file can be loaded in a genome browser such as IGV [17].

3.2.1 Interpretation of the Peak-Motifs Results

The *peak-motifs* results are displayed in a Web form giving access to all the files generated during the analysis.

Figure 5 shows a partial snapshot of the *peak-motifs* results with the study case. Since the workflow covers many types of analyses and results, here we attempted to present a human-readable report, organized according to the successive steps of the workflow: sequence composition (Fig. 5a), motif discovery (Fig. 5b), and comparison of discovered motifs with known motifs (Fig. 5c).

Sequence Composition

This section, described in Fig. 5a, shows some properties of the peak sequences.

- The top panel of the synthetic table shows the distribution of sequence lengths. In this study case, we can observe that most sequences have a length around 200 bp, which is a good indication for transcription factor ChIP-seq peaks (histone peaks are generally longer).
- The second panel shows the nucleotide composition of the sequences, with a heatmap indicating the frequencies of each nucleotide, and a plot displaying the profile of frequencies for each nucleotide along the peaks. In this example, we can see that *G* and *C* are less frequent than *A* and *T* over the whole peak width. Interestingly, we also notice a nucleotidic skew, with an enrichment of *As* and *Gs* upstream peak centers, and a symmetrical enrichment of *Ts* and *Cs* downstream.
- The third panel shows the dinucleotide composition of the sequences. The *transition table* indicates the probabilities of each nucleotide (column) depending on the preceding nucleotide (“prefix”, rows). Gray shades denote the relative frequencies, and highlight dependencies between adjacent nucleotides. For example, in the study case, we observe that the frequency of *As* varies from 0.36 after another *A* (*AA* dinucleotide) to

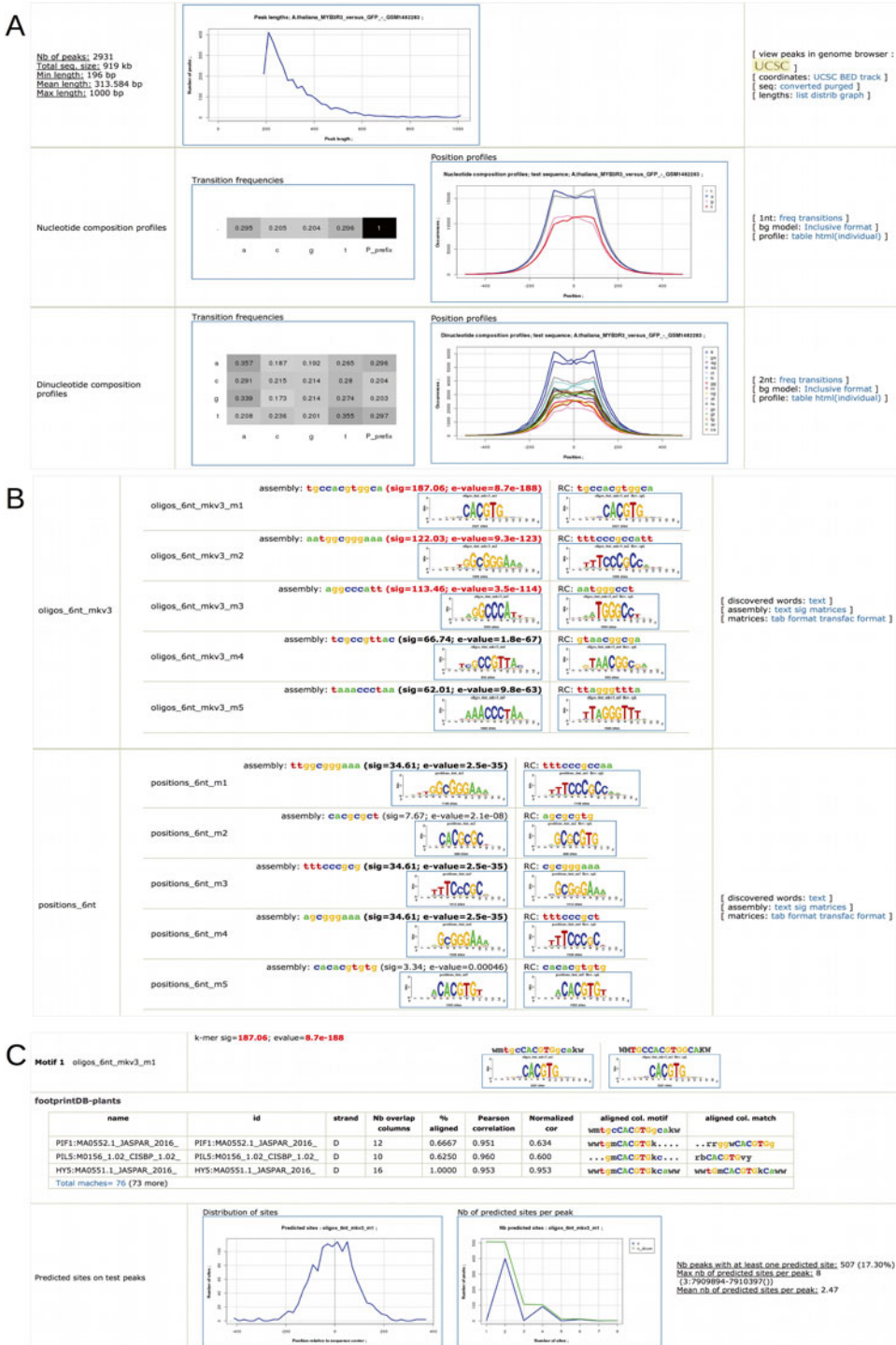


Fig. 5 Peak-motifs results. (a) General information about the peak sequences of our study case, including their composition in nucleotides and dinucleotides, and the corresponding profiles. (b) Discovered motifs (by algorithm). This example shows the 6-nucleotide motifs found with the *oligo-analysis* algorithm, using a Markov model of order 3. (c) Discovered motifs (with motif comparison). Shows the comparison of the discovered motifs versus a collection of TFBMs databases, and the distribution profile of the motifs in the peaks

0.21 after a T (TA dinucleotide). The *dinucleotide profiles* provide a visual representation of the positional distribution for each dinucleotide. On the study case we note an upstream–downstream skew for AA, TT, CC and GG, and a local depletion of TA and AT in the peak centers.

Discovered Motifs (by Algorithm)

This section (Fig. 5b) shows the full list of discovered motifs, organized by motif discovery algorithm (*oligo-analysis*, *position-analysis*) and by k-mer size.

The name of each motif (e.g. oligos_6nt_mkv3_m1) indicates:

- The algorithm used (oligos for *oligo-analysis*, positions for *position-analysis*).
- The k-mer length used to build the motif (6 nt, 7 nt).
- The order of the Markov model (mkv).
- The rank of the motif (m1 to m5).

In addition, the motif logo is displayed in both orientations.

In this section, an important information is that each discovered motif is associated with an e-value and a derived significance score: $sig = -\log_{10}(E\text{-value})$. The e-value indicates the expected number of false positives. E-values much lower than 1 (corresponding to highly positive *sig* scores) indicate a very significant over-representation (*oligo-analysis*) or positional bias (*position-analysis*) of the motif. The most significant motifs are highlighted in red and bold. In our study case, the motif CACGTG is over-represented with a significance of 187, which corresponds to an e-value (expected number of false positives) of $\sim 10^{-187}$. The same motif is found by *position-analysis*, yet with a much lower significance ($s=3.34$, e-value 0.00046). It is thus the most significant motif in terms of over-representation, but other motifs are much more significant in terms of positional bias, in particular wttG-GCGGGAaaat (positions_6nt_m1), which achieves a significance of 34.61. This example shows the interest of combining two independent criteria to discover exceptional motifs.

Discovered Motifs (with Motif Comparison)

Illustrated in Fig. 5c, this section displays each motif individually with matches found in collections of known TFBMs (e.g. FootprintDB plants, Jaspar plants, etc.).

Additionally, for each motif, two other plots are shown:

- The positional distribution of predicted sites relative to peak centers (e.g. showing that most matches are located around the center of the peaks).
- The distribution of the number of binding sites per sequence. For the CACGTG motif, occurrences per peak show a particular teeth-shaped distribution due to the reverse complementary palindromic nature of the motif (occurrences are systematically found on both strands).

Note that the algorithms produce redundant motifs. For example a motif with the core GGCGGG is found by both *oligo-analysis* and *position-analysis*, with different k-mer lengths; thus, the next step in the analysis is to reduce the redundancy of the motifs.

3.3 Motif Clustering

Using different motif discovery algorithms to analyze the same sequences is useful and recommended to increase the sensitivity (some algorithms discover motifs that others do not) or to corroborate the results (e.g. gain confidence by observing that the same motif is both over-represented and concentrated on the peak centers). However in some cases the redundancy between motifs returned by different algorithms and with different parameters makes it difficult to interpret the results as a whole.

The RSAT website includes a new specialized tool called *matrix-clustering*, which identifies groups of similar motifs, generates consensus matrices, and provides a dynamical visual interface to browse and inspect the relationships between multiple motifs. We will use this tool to obtain a non-redundant collection of motifs from the motifs discovered with *peak-motifs*.

- 1 Open a connection to the Plants Regulatory Sequence Analysis Tools server (<http://plants.rsat.eu/>).
- 2 On the left-side panel, open the toolbox “**Matrix tools**” and click “**matrix-clustering**”.
- 3 On the *title* box you can give a title to the analysis for example *Myb3R3 discovered motifs*.
- 4 **Upload** the motif file obtained from *peak-motifs* and select the **TRANSFAC format**.
- 5 In the **Motif comparison options** section, you can fine-tune the thresholds that will be used to split the tree with all the motifs in a collection of trees (forest). The default cutoffs are relatively lenient, but for this application more conservative values can be chosen. In the column *lower threshold*, set **w** to 5, **cor** to 0.75, and **Ncor** to 0.55.
- 6 In the **Clustering options** section, select *Ncor* (Normalized Pearson Correlation) as a **Metric to build the trees** and *average* as the **Agglomeration rule**.
- 7 You can either select **email** output and fill up your address, or **display**, and click **GO**.

After a few seconds, the website displays a link to the result page. You can already open this page as soon as the link appears. Even though the program may take a few minutes to accomplish the clustering, the result page will be updated periodically.

3.3.1 Interpretation of the Matrix-Clustering Results

The *matrix-clustering* results are organized in different sections (Fig. 6). You can display/hide each one by clicking on the buttons.

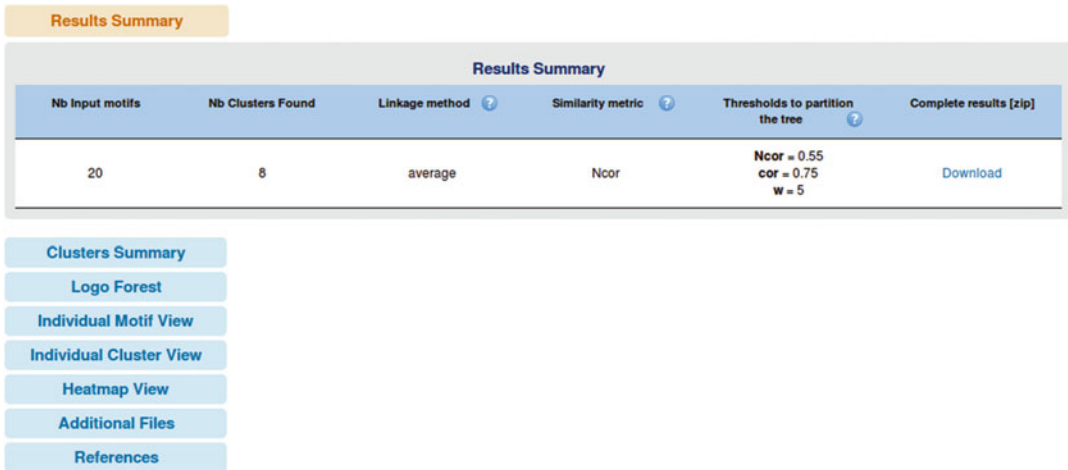


Fig. 6 Matrix-clustering results. General view of matrix-clustering report. Each button can be clicked to show/hide details

- The *Results summary* section shows a table indicating the number of input motifs, the number of clusters and the parameters used to cluster the motifs, additionally a link to download all the results in zip. In this case the 20 motifs discovered with *peak-motifs* were regrouped in eight distinct clusters.
- The *Clusters summary* section shows a table with the motifs belonging to each cluster and the logos in both orientations representing the *root motifs* of each cluster (i.e. a motif formed by summing or averaging the counts of all the motifs belonging to the cluster).
- The *Logo Forest* section points to a link where the clusters are displayed as a set of trees, each corresponding to a cluster. In this link you can dynamically expand/collapse the tree, each time a branch is collapsed, it shows the *branch-motif* which represents all the descendant motifs of the collapsed branch. Figure 7 shows the first three clusters of the logo forest produced by *matrix-clustering* from the motifs discovered by *peak-motifs* in MYB3R3 peaks.
- The *Individual Motif View* section shows a table with all the input motifs and some of their attributes (assigned cluster, aligned and colored consensus, small logos).
- The *Individual Cluster View* section shows some properties of each cluster individually. You can select a specific numbered node of tree to select its corresponding *branch-motif*.
- The *Heatmap view* section shows a heatmap of the motifs grouped in clusters.
- The *Additional Files* section shows a table with additional files (motif comparison results, the motifs associated to each cluster,

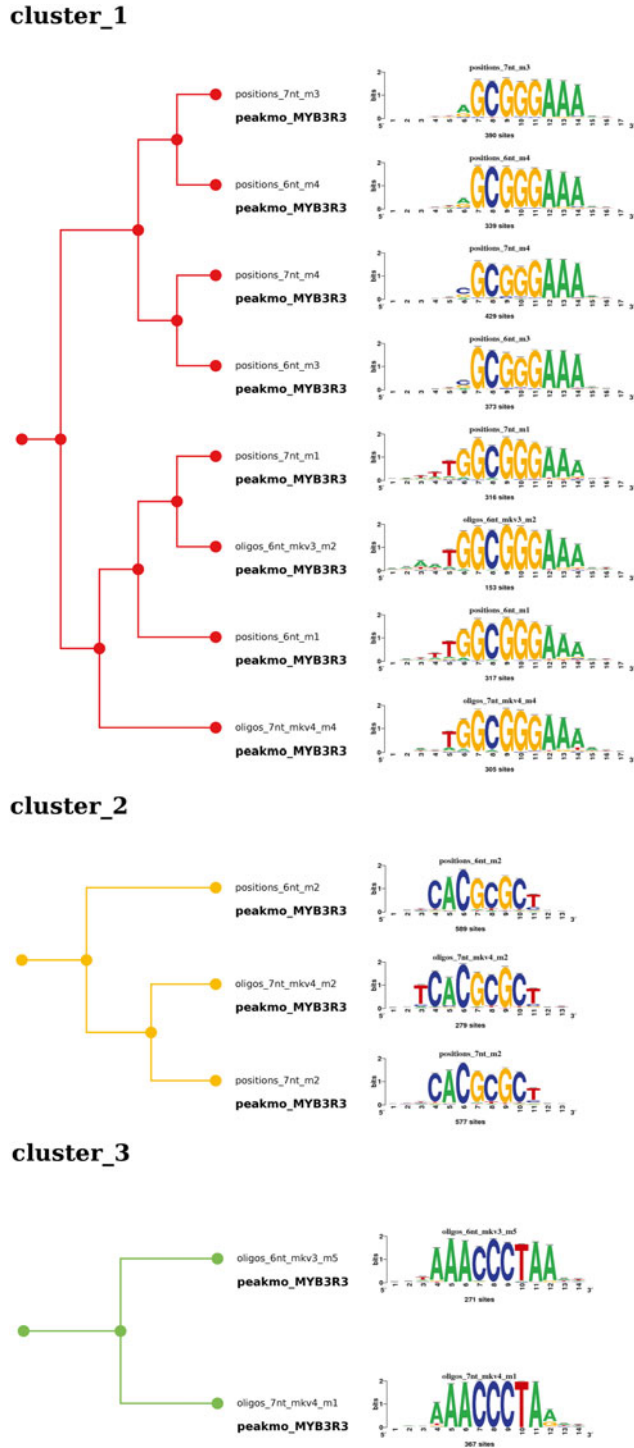


Fig. 7 Matrix-clustering results. The 20 motifs discovered by peak-motifs in the MYR3R3 ChIP-seq peaks were separated in eight clusters. Each tree shows the alignments of a cluster of similar motifs. The *leaves* indicate the motif discovery algorithm with which each motif was found. Note that the similar motifs are discovered independently by different algorithms (*oligo-analysis*, *position-analysis*), or are found with different parameters (e.g. k-mer length) of the same algorithm

etc.) including the *Root motifs* file, which contains the collection of non-redundant motifs. This file will be used for the following part of the analysis.

1. Right-click the “**Root motifs**” link and save file as *matrix-clustering_cluster_root_motifs.tf* on your computer.

The 20 motifs discovered with *peak-motifs* were separated in eight clusters of variable size (Fig. 7). For example, cluster 1 contains 8 motifs corresponding to the EF2 family while the motif for cluster 6 (singleton) corresponds to the MSA motif reported in the published work selected as our case study [15].

3.4 Negative Controls with Random Genomic Regions

RSAT motif discovery tools compute the significance of the motifs based on theoretical models (Markov chains, which take into account the dependencies between adjacent nucleotides). However, it is not obvious a priori that these models perfectly suit the properties of biological sequences. A pragmatic way to check the correctness of the models is to measure the empirical rate of false positives with a *negative control set*, i.e. set of sequences supposedly not enriched for any particular TFBS. In principle, motif discovery programs should be able to return a negative answer (no result) when such datasets are submitted.

When analyzing genomic regions such as ChIP-seq peaks, the recommended negative control consists in analyzing regions of the same sizes as the peaks picked up at random in the reference genome.

- 1 Open a connection to the Plant Regulatory Sequence Analysis Tools server (<http://plants.rsat.eu/>).
- 2 In the left-side panel, open the toolbox “NGS ChIP-seq” and click “**random genome fragments**”.
- 3 Under **Random fragments**, click the “**Browse...**” button and locate the peak sequences file on your computer (the fasta file downloaded at **step 8** in Subheading 3.1).
- 4 Under **Organisms**, select the reference organism. For the study case, this is *Arabidopsis thaliana.TAIR10.29*.
- 5 In the **Output** section, select *Sequences in fasta format (only for RSAT organisms)* and check the *Mask repeats* option.
- 6 Select the *server* output and click **GO**. The selection of random genomic regions should take a few seconds.
- 7 On the result page, you can access the randomly picked up genomic sequences by clicking on the link to the fasta file (*Genomic fragments (fasta)*). You can optionally save this result to keep a copy of these random genomic fragments.
- 8 In the **Next Step** section of the result page, click on the **peak-motifs** button. This will display a *peak-motifs* form pre-filled with the URL of the random genomic sequences. Set the title to “A. thaliana random fragments”. Check that all the

other parameters have the same parameters as for the analysis of the actual ChIP-seq peaks in the previous sections, and click **GO** (see **Note 13**).

- 9 Once the job is completed, open the results page, and click the link **Download all matrices (TRANSFAC format)** in the summary, to store the matrices on your computer.
- 10 Repeat the **matrix-clustering** analysis (**steps 1–8** in Subheading **3.3**) using the matrices obtained with *Random fragments* (TRANSFAC file).

3.4.1 Interpretation of the Negative Control

The goal of this negative control is to obtain an empirical estimation of the rate of false positives. In some cases, these controls reveal that the actual rate of false positive exceeds the theoretical expectation (indicated by the e-value of the motif discovery programs).

When the sequences of interest are genomic regions such as ChIP-seq peaks, the most relevant negative control consists in selecting random genomic regions of the same sizes. For the study case, we analyzed a dataset made of 2,931 random regions from *Arabidopsis thaliana*. The sequence length distribution is, as expected, exactly the same as for the actual peaks analyzed above. However the mono- and di-nucleotide composition may differ, because they reflect a random sampling of any type of genomic regions rather than regulatory regions.

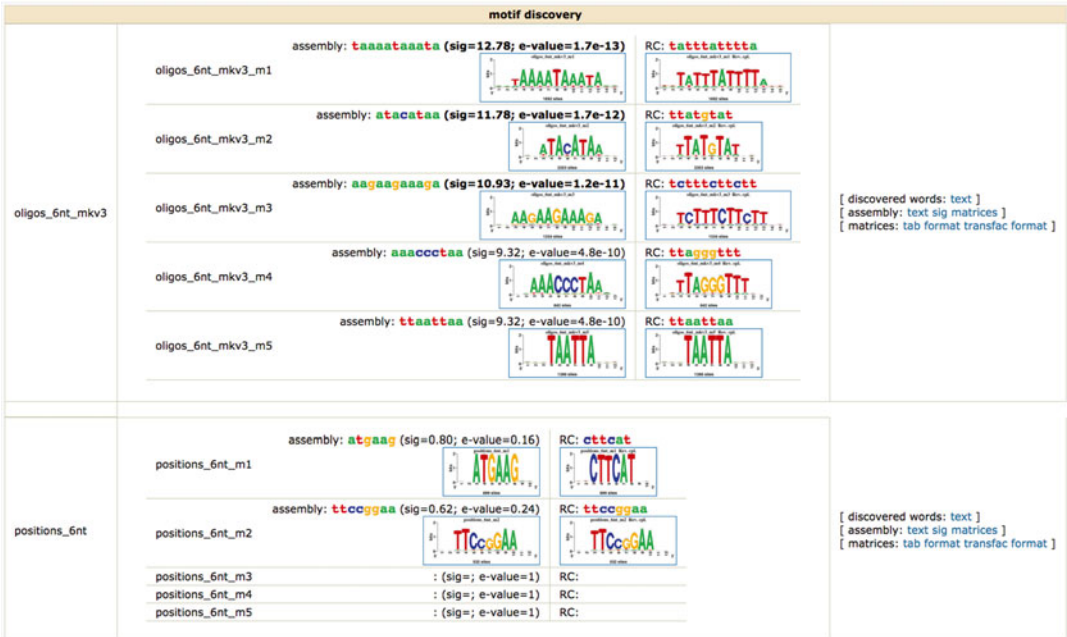
Peak-motifs

The analysis of random genomic regions returned 17 motifs (Fig. 8a), most of which are of low complexity (e.g. atAAaATAaata, aaaAACAAAA, or motifs showing repeated sequences, e.g. TATATATA). Some of these motifs show a high similarity with some reference motifs stored in FootprintDB, suggesting that they might correspond to some actual transcription factor.

The most important criterion in this control is to inspect the significance of the discovered motifs in the section **Discovered motifs (by algorithm)** (Fig. 8a). In our experience, programs based on a global over-representation (*oligo-analysis*, *dyad-analysis*) tend to return results even with random genomic regions, although with significance hopefully lower than with the real peaks: in the study case, *oligo-analysis* returns significance scores of 188 with the actual peaks, and 13.6 with random genomic regions. These motifs are actually correctly qualified of over-represented, but their over-representation is general in the genome rather than specific to the peaks. These motifs can correspond to low complexity regions or to functional elements found in abundance throughout the genome.

In contrast, programs relying on positional distributions (*position-analysis*, *local-word-analysis*) generally perform very well in negative controls (Fig. 8a), in the sense that they return motifs of poor significance (lower than 3) or no motif at all. This emphasizes once again the importance of evaluating multiple criteria before considering a motif as relevant.

a



b

Results Summary					
Nb Input motifs	Nb Clusters Found	Linkage method ?	Similarity metric ?	Thresholds to partition the tree ?	Complete results [zip]
17	14	average	Ncor	Ncor = 0.55 cor = 0.75 w = 5	Download

Fig. 8 Negative controls with random genomic regions. **(a)** Partial results of the *peak-motifs* motif discovery result in random genomic regions. Note that the most significant motifs are poor-complexity motifs corresponding to repetitive elements. **(b)** Overview of the *matrix-clustering* results for these motifs. Note the high number of clusters, indicating that most motifs are detected by only one motif discovery method. **(c)** Clustering of the motifs discovered in the random peaks

C

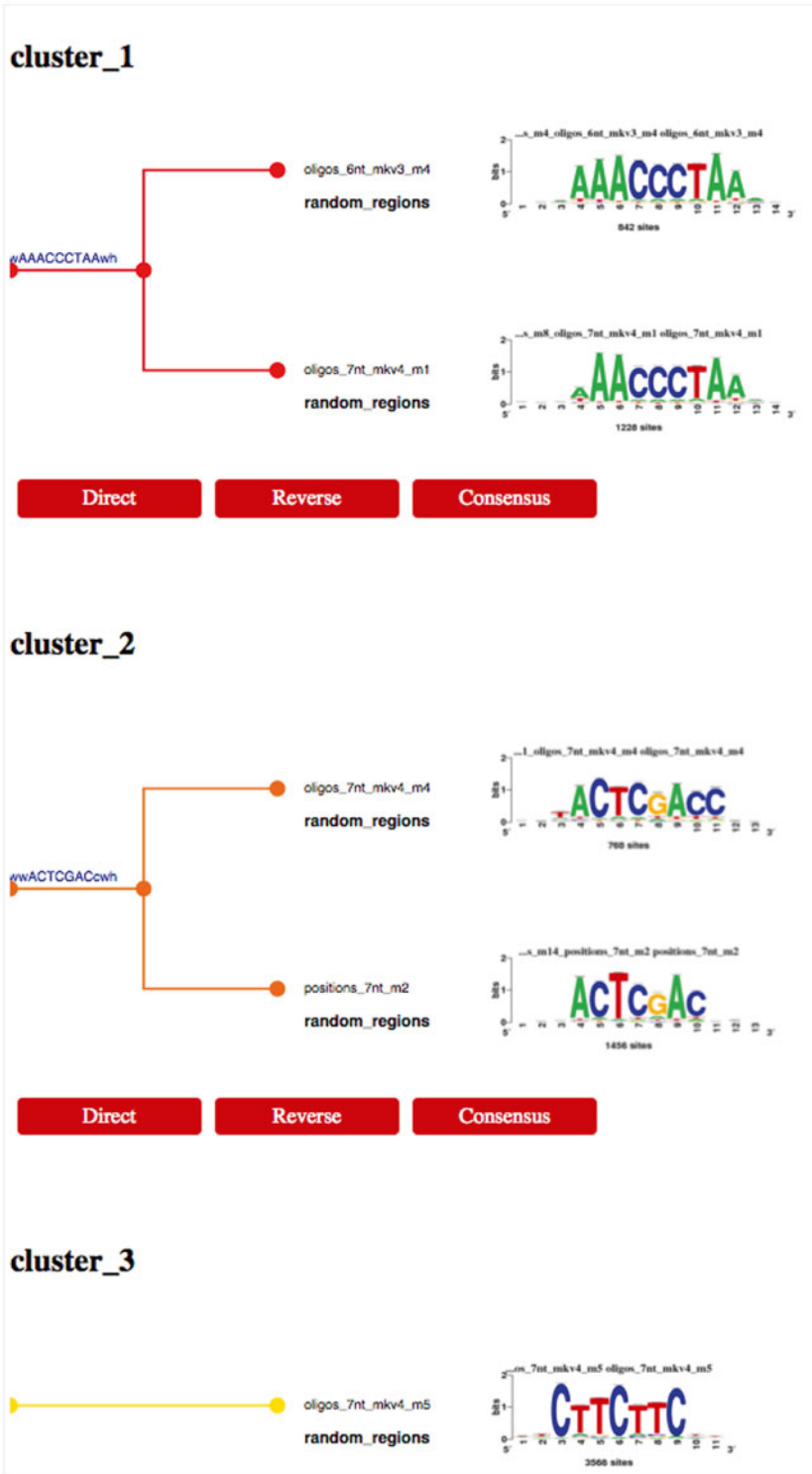


Fig. 8 (continued)

In the section **Discovered motifs (with motif comparison)**, the positional distribution of predicted sites is not as concentrated around the centers of random fragments as they were for actual MYB3R peaks (Fig. 5c). Also, the number of matches is generally lower than the real peaks.

Matrix-Clustering

With our random trial, the clustering separated the 17 significant motifs into 14 clusters (Fig. 8b, c), where only three clusters contain at least two motifs (the rest are singletons). This lack of consistency between the discovered motifs is also an indication of the poorer relevance of the motifs discovered in random regions, relative to those found in actual peaks.

4 Notes

1. Format descriptions at UCSC: <https://genome.ucsc.edu/FAQ/FAQformat.html>
2. Direct access to the peak coordinates of the study case: ftp://ftp.ncbi.nlm.nih.gov/geo/samples/GSM1482nnn/GSM1482283/suppl/GSM1482283_MYB3R3-GFP_ChIP_peaks.bed.txt.gz
3. When working with lab data, peaks are obtained by running peak-calling programs on the aligned reads. Alternatively they can be downloaded from specialized databases such as GEO ([18], <http://www.ncbi.nlm.nih.gov/geo/>) or ArrayExpress ([19], <https://www.ebi.ac.uk/arrayexpress/>).
4. It is very important to specify the same assembly as used for the read mapping, since otherwise the coordinates on the BED file might not match the correct genomic sequences. Please contact the administrator of the RSAT Plant site if the required assembly does not appear in the list.
5. A common difficulty with BED files is that the chromosome naming convention differs between genome databases. In particular, some databases systematically use a “chr” prefix (chr1, chr2, chr3, ..., chrMt, chrPt) whereas some others simply use the chromosome number (1, 2, 3, ...) or name (Mt, Pt). To circumvent this problem, the sequence retrieval tool automatically checks the consistency of chromosome names between the query BED file and the genome sequence file installed on RSAT, and prepends or removes the chr prefix if required.
6. In plant genomes, repeated elements may result from various sources: transposons, polyploidy, etc. (see Chapter 18). Repetitive elements cause particular problems for motif discovery, because the statistics of over-representation rely on an assumption of independence between the sequences. It is thus

recommended to mask repeated elements during the motif discovery step of a ChIP-seq analysis workflow. Note that in some other contexts (for example, scanning sequences with a TF binding motif), it might be relevant to keep the repetitive elements in order to detect all the putative binding sites.

7. Two other algorithms can be selected for finding motifs: *dyad-analysis* detects over-represented dyads (spaced pairs of trinucleotides), which are typically bound by dimeric transcription factors; *local-words* detects k-mers with local over-representation, i.e. having a higher number of occurrences in a particular positional window, relative to the rest of the peaks. Selecting more algorithms is sometimes helpful to gather a wider set of discovered motifs, as some algorithms can discover motifs that other would not. However, in many cases the different algorithms return very similar motifs, thus producing redundancy in the result. We thus activated by default the two algorithms offering a good trade-off between computing time and sensitivity, and which rely on two complementary criteria (over-representation and positional distribution relative to peak centers).
8. Beware, oligomer-length is not the same as motif length. Indeed, the significant k-mers and dyads are assembled and used as seeds to collect sites, which are in turn aligned to build the final motifs (position-specific scoring matrices). The resulting matrices are thus generally wider than the oligomer length. The default lengths were chosen because they generally provide a good trade-off between sensitivity and specificity, and were shown to return the most relevant motifs [22].
9. The program *oligo-analysis* relies on Markov models to compute the prior probability of each k-mer, i.e. its probability to be found at a given position in the sequence. In peak-motifs, the prior probability of each oligonucleotide (k-mer) is estimated on the basis of the frequencies of smaller k-mers in the sequence. The Markov order specifies the stringency of the background model. Increasing the order improves the specificity at the cost of sensitivity. This automatic option applies an ad-hoc rule to choose a Markov order ensuring a balance between sensitivity and specificity, depending on the total size of the peak set.
10. By default the program restricts the results to five motifs (assembled matrices) per algorithm. This number could be increased if you have some particular reason to think that the peak set contains a wider variety of motifs, with a proportional increase in the computing time. This can be useful for example for peaks from particular histone modification marks corresponding to enhancer regions supposedly bound by multiple factors.

11. Beware, there is a distinction between the options *reference motifs* and *custom database*. Reference motifs should be one or a few motifs expected to be found in the ChIP-seq peaks, whereas the custom database may be a large collection encompassing all the known motifs for the organism or taxon of interest.
12. By default, sequence scanning returns the putative binding site coordinates relative to the peak sequences. If appropriately formatted, the sequence headers of the peak file can indicate the coordinates of each peak relative to the chromosomes. The program can then convert each binding site coordinate from peak-relative to chromosome coordinates. The resulting files can then be loaded in a genome viewer (e.g. IGV).
13. The *peak-motifs* analysis will take approximately the same time as for the actual peaks, between a few minutes and several tens of minutes depending on the sequence size.
14. Example of structured query to gather ChIP-seq series (GSE) for a given taxon in GEO datasets (<http://www.ncbi.nlm.nih.gov/gds/>): (“gse”[Entry Type] AND “genome binding/occupancy profiling by high throughput sequencing”[DataSet Type] AND “Viridiplantae”[Organism]).

Acknowledgements

We thank C. Dubos for feedback on MYBR3 proteins. This work was funded in part by Fundación ARAID and by the Enseignants-Chercheurs invités program of Aix-Marseille Université (to B.C.M.). C.R. was supported by the *France Génomique* National infrastructure, funded as part of the *Investissements d’Avenir*, program managed by the *Agence Nationale pour la Recherche* (contract ANR-10-INBS-09). J.C-M PhD grant is funded by the Ecole Doctorale des Sciences de la Vie et de la Santé (EDSVS), Aix-Marseille Université.

References

1. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, Euskirchen G, Bernier B, Varhol R, Delaney A, Thiessen N, Griffith OL, He A, Marra M, Snyder M, Jones S (2007) Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. *Nat Methods* 4:651–657
2. Mardis ER (2007) ChIP-seq: welcome to the new frontier. *Nat Methods* 4:613–614
3. Kulakovskiy IV, Boeva VA, Favorov AV, Makeev VJ (2010) Deep and wide digging for binding motifs in ChIP-Seq data. *Bioinformatics* 26:2622–2623
4. Machanick P, Bailey TL (2011) MEME-ChIP: motif analysis of large DNA datasets. *Bioinformatics* 27:1696–1697
5. Thomas-Chollier M, Darbo E, Herrmann C, Defrance M, Thieffry D, van Helden J (2012) A complete workflow for the analysis of full-size

- ChIP-seq (and similar) data sets using peak-motifs. *Nat Protoc* 7:1551–1568
6. Thomas-Chollier M, Herrmann C, Defrance M, Sand O, Thieffry D, van Helden J (2012) RSAT peak-motifs: motif analysis in full-size ChIP-seq datasets. *Nucleic Acids Res* 40, e31
 7. Medina-Rivera A, Defrance M, Sand O, Herrmann C, Castro-Mondragon JA, Delerce J, Jaeger S, Blanchet C, Vincens P, Caron C, Staines DM, Contreras-Moreira B, Artufel M, Charbonnier-Khamvongsa L, Hernandez C, Thieffry D, Thomas-Chollier M, van Helden J (2015) RSAT 2015: Regulatory Sequence Analysis Tools. *Nucleic Acids Res* 43:W50–W56
 8. Pepke S, Wold B, Mortazavi A (2009) Computation for ChIP-seq and RNA-seq studies. *Nat Methods* 6:S22–S32
 9. Steinhauser S, Kurzawa N, Eils R, Herrmann C (2016) A comprehensive comparison of tools for differential ChIP-seq analysis. *Brief Bioinform*. doi:10.1093/bib/bbv110
 10. Zhang Y, Liu T, Meyer CA, Eeckhoutte J, Johnson DS, Bernstein BE, Nusbaum C, Myers RM, Brown M, Li W, Liu XS (2008) Model-based analysis of ChIP-Seq (MACS). *Genome Biol* 9(9):R137. doi:10.1186/gb-2008-9-9-r137
 11. Heinz S, Benner C, Spann N, Bertolino E, Lin YC, Laslo P, Cheng JX, Murre C, Singh H, Glass CK (2010) Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell* 38(4):576–589
 12. Wilder S (2009) SWEMBL: a generic peak-calling program. Unpublished. <http://www.ebi.ac.uk/~swilder/SWEMBL/>
 13. Kharchenko PV, Tolstorukov MY, Park PJ (2008) Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat Biotechnol* 26:1351–1359. doi:10.1038/nbt.1508
 14. Schneider TD, Stephens RM (1990) Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res* 18:6097–6100
 15. Jolma A, Yan J, Whittington T, Toivonen J, Nitta KR, Rastas P, Morgunova E, Enge M, Taipale M, Wei G, Palin K, Vaquerizas JM, Vincentelli R, Luscombe NM, Hughes TR, Lemaire P, Ukkonen E, Kivioja T, Taipale J (2013) DNA-binding specificities of human transcription factors. *Cell* 152:327–339
 16. Sebastian A, Contreras-Moreira B (2014) footprintDB: a database of transcription factors with annotated cis elements and binding interfaces. *Bioinformatics* 30:258–265
 17. Thorvaldsdóttir H, Robinson JT, Mesirov JP (2013) Integrative Genomics Viewer (IGV): high-performance genomics data visualization and exploration. *Brief Bioinform* 14:178–192
 18. Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M, Yefanov A, Lee H, Zhang N, Robertson CL, Serova N, Davis S, Soboleva A (2013) NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res* 41:D991–D995
 19. Kolesnikov N, Hastings E, Keays M, Melnichuk O, Tang YA, Williams E, Dylag M, Kurbatova N, Brandizi M, Burdett T, Megy K, Pilicheva E, Rustici G, Tikhonov A, Parkinson H, Petryszak R, Sarkans U, Brazma A (2015) ArrayExpress update—simplifying data submissions. *Nucleic Acids Res* 43:D1113–D1116
 20. Kobayashi K, Suzuki T, Iwata E et al (2015) Transcriptional repression by MYB3R proteins regulates plant organ growth. *EMBO J* 34:1992–2007
 21. Ito M, Araki S, Matsunaga S, Itoh T, Nishihama R, Machida Y, Doonan JH, Watanabe A (2001) G2/M-phase-specific transcription during the plant cell cycle is mediated by c-Myb-like transcription factors. *Plant Cell* 13:1891–1905
 22. van Helden J, André B, Collado-Vides J (1998) Extracting regulatory sites from the upstream region of yeast genes by computational analysis of oligonucleotide frequencies. *J Mol Biol* 281:827–842

INDEX

A

- ABI4..... 285, 289, 291, 292
Agrobacterium
A. tumefaciens.....20–22, 26, 27, 117, 127,
 129, 176–177, 181, 184, 186, 187, 189, 192, 205
 transformation 204–207, 215
 Agro-infiltration.....127
Arabidopsis thaliana.....140, 157, 178, 179,
 234–238, 242, 247, 248, 254
 ArabidopsisPBM.....263, 287
 At7 cell culture68
AtFer1..... 104, 107–109
 AthaMap.....242, 250, 263, 270, 287, 304

B

- β-glucuronidase68, 72, 76, 116–117,
 129–131, 157, 172, 177, 184, 191
 Binding-site Estimation Suite of Tools
 (BEST)..... 236–240, 243, 268
 Bioinformatics 50, 84, 94, 164, 241, 299, 304
 Bioluminescence.....23, 28, 106–107, 109
 BLAST..... 178, 179, 186

C

- Calmodulin-Binding Transcriptional Activator
 (CAMTA)16
Camelina sativa..... 191, 192
 Cauliflower mosaic virus 35S promoter
 (CaMV 35S)..... 6, 9, 12, 71, 78, 111, 116, 124, 194
Caulimovirus..... 72, 111, 154, 157, 159, 164, 177
 Cestrum yellow leaf curling virus (CmYLCV)..... 111–137
 Chromatin immunoprecipitation (ChIP).....260, 297, 300
 Chromatin immunoprecipitation DNA-sequencing
 (ChIP-seq).....140, 286, 294, 297–319
Cis-acting element..... 2–12, 67, 78
Cis-element78, 112, 140, 144–146, 148,
 163–165, 171, 219, 220, 236, 242, 251–256, 260–265,
 269, 270, 272, 273
Cis-regulatory element.....15–28, 68, 151, 219,
 242, 248, 256, 259–274, 280, 303
Cis-regulatory sequence31–46, 103, 139–148,
 233, 234, 241, 247
 Cluster.....3, 42, 79, 86, 91–92, 96, 128, 268,
 279–295, 304, 309, 312–315, 317

- CmYLCV. *See* Cestrum yellow leaf curling virus (CmYLCV)
 Co-expression..... 16, 43–44, 298
 Comparative sequence analysis..... 87, 96–97
 Coregulated genes 233–243, 247

D

- Database 179, 234, 241, 242, 248, 251
 Destination vector 190, 194, 197, 200–203, 208, 215
 DNA-binding domains (DBDs) 49–51, 58,
 67, 139, 140, 147, 151, 175, 180, 260, 261, 265, 267
 DNA-binding motif (DBM).....259, 261
 DNA binding sites (DBSs)259, 260, 263, 270, 272
 DNA methylation 83, 139
 DNA motif.....50, 58, 260, 262–267,
 269–271, 273, 279, 280, 284, 287
 DREME25
 Drought-Responsive Element (DRE).....248

E

- E2F.....285–292
 Electrophoretic mobility shift assay (EMSA)..... 50, 113,
 120–121, 132, 134–137, 260, 271, 293, 298
 Electroporation..... 115–116, 123–124
 Enhanced green fluorescent protein (EGFP)..... 191, 194,
 195, 213, 214
 Enhanced yellow fluorescent protein (EYFP)191, 195,
 213, 214
 Ensembl 280–284, 294
Escherichia coli.....54, 68, 74, 113, 119, 121–122,
 132, 136, 140, 141, 143–145, 154, 189–191, 198, 199,
 201–204, 215, 220, 273, 298
 Expression analysis 212–213, 247–256

F

- FASTA 19, 186, 223, 229, 266,
 268, 270, 306
 Ferritin104
 Floral dip..... 127, 206–208
 Flow cytometry..... 152, 155, 157, 160
 Fluorescence54–55, 62–64, 76, 104,
 124, 130, 152, 155, 157, 158, 171, 173, 184, 187, 193,
 208, 212–214
 FootprintDB.....259–274, 291, 292, 294,
 303, 304, 311, 316

G

Gateway..... 154, 159, 190–192, 194, 195, 197–202, 214, 215
 GenBank..... 121, 196, 229, 230
 GenoCAD..... 219–231
 Gene Expression Omnibus (GEO)..... 35, 284, 298, 305, 318, 319
 Gene ontology (GO)..... 35, 293, 303
 Green fluorescent protein (GFP) 104, 152, 155–159, 187

H

Histochemical GUS staining..... 117, 127, 129–131
 Hypomethylome..... 83–85, 96, 98, 99

I

Imaging 18, 23, 27, 103–109
 In silico validation 248
 Iron..... 106–109

J

JASPAR..... 25, 263, 265, 269, 270, 287, 294, 304, 311
 JGI Genome Portal 280

L

Labeling DNA probes..... 120–121
 Ligation-independent cloning (LIC) 140, 141, 143, 144, 147
 Luciferase (LUC) 69, 71, 72, 75–76, 78, 80, 104, 106–109, 117, 124, 170
 reporter 16, 20, 104, 164

M

mCherry 191, 193–195, 213, 229
 MEME 43, 237–239, 243, 263, 290
 Methylation-sensitive restriction enzyme sequencing (MRE-seq) 83–101
 Microarray..... 44, 140, 151, 234, 236, 248, 250, 255, 268, 279, 280
 Microbe-associated molecular patterns (MAMPs)..... 163, 164, 171–173
 MicroProtein (miP)..... 175–187
 Minimal promoter..... 3, 4, 6, 8–9, 16, 20, 74, 164, 165
 Monomeric cyan (teal) fluorescent protein (mTFP1) 191, 213
 Multi gene cloning 189
 MYB3R3..... 299, 300, 305, 307, 312, 313

N

NASCarrays 35, 234
 NCBI's Gene Expression Omnibus..... 234
 Next generation sequencing (NGS) 50, 84, 299, 305

Nicotiana benthamiana 18, 21–24, 26, 127, 177, 182, 184, 187, 229
 Nuclear extract preparation 119

O

Open-access..... 271
Oryza sativa..... 84, 88, 90, 265, 267, 271, 274, 286
 Overrepresented motifs 234, 236–241, 248, 293

P

Parsley protoplast..... 79, 163–173
 Parsley suspension culture..... 165, 167–168
 PathoPlant..... 234–236, 242, 248, 251, 252, 254
 pCAMBIA 228
 PC-GW-BAR..... 191, 194–198, 200–202, 208, 211
 Peak-caller 300–303
 Phylogenetic footprinting..... 32, 36
Physcomitrella patens..... 151–160
 Phytozome 185
Picea abies..... 84, 88, 90, 91
 Plain sequence 229
 Plant biotechnology..... 12
 Plant expression vector 113, 122, 127, 219–231
 Plant grammar 219–231
 Plant pathogen interaction 163
 Plant promoter 12, 67, 112, 189, 191, 220, 228, 240, 248, 263, 287, 293
 Polymerase chain reaction (PCR) 140, 141, 143, 148, 179–181, 190–192, 197–200, 203–206, 211–212, 214, 215
 Position-specific scoring matrices (PSSMs) 259–261, 263, 264, 271–273, 280, 304, 319
 Position-weight matrix (PWM) 288, 290–292, 294
 Preparation of high-molecular-weight genomic DNA..... 87, 88
 Promoter sequences 16, 18, 36, 39, 40, 112, 191, 220, 234, 236–238, 240, 242, 243, 247, 248, 262, 271, 284, 289–291
 Protein Data Bank (PDB)..... 261, 263, 265, 272
 Protein domain 12, 180, 267, 272
 Protein–protein interaction (PPI)..... 2, 175, 180, 220
 Proteome 265–269, 273
 Protoplast 67–80, 112, 113, 115, 121–124, 129, 136, 140, 152–160, 163–173
 isolation 71, 115–116, 123, 136, 152–154

Q

Quantitative DNA-protein-Interaction-ELISA (qDPI-ELISA)..... 49–64
 Quantitative real-time PCR (qRT-PCR)..... 112, 119, 131–132

R

RNA isolation 119, 131–132
 RNA-seq 35, 234, 279, 280, 298
 RSAT::Plants 263, 271, 274, 279–294, 297–319

S

Saccharomyces cerevisiae 142, 148, 279–283, 286, 298
 Single nucleotide polymorphism (SNP) 31–46, 100, 236
Solanum tuberosum 223, 227, 229
Sorghum bicolor 289–291
 Stable transformation 16–18, 25, 68, 204
 STAMP 241, 242, 264, 269, 272, 273
 Subcellular localization 152, 155
 SYBR–Green 55–56, 62, 119, 131
 Synthetic biology 12, 13, 50, 175, 220
 Synthetic promoter 140, 163–173, 177, 220, 233, 242, 250
 Systematic Evolution of Ligands by Exponential enrichment (SELEX) 50, 140, 151, 299

T

T-DNA 189, 190, 211
 Terminator 123, 190, 194, 195, 197, 203, 223, 228
 TGA1a 112, 119–120, 132–136
 The Arabidopsis Information Resource (TAIR) 24, 25, 34, 35, 234, 236

Tobacco cell suspension culture 115–116, 123
 Transcription factor binding motifs (TFBM) 280, 298, 299, 301, 304, 310, 311, 315
 Transcription factor binding sites (TFBS) 2, 5, 12, 32, 33, 35, 38, 40, 84, 96, 100, 238, 242, 250, 297
 Transcription factors (TFs) 139–148, 151, 163, 175, 177, 178, 180, 186, 233, 242, 247, 248
 Transcriptome 95, 234, 279, 293, 298
 TRANSFAC 263, 265, 266, 269, 271, 272, 287, 288, 290, 304, 308, 312, 316
 Transient expression 5, 22, 127, 140, 157
 Transient transformation 21, 26

U

uidA 68, 71, 78, 104, 116–117, 129–131, 164, 177
 Upstream sequence 39–41, 268, 279–294

W

WRI1 289–292

Y

Yeast one-hybrid (Y1H) 67, 139–148, 152, 268–270
 Yeast transformation 141–142, 145–146, 148

Z

Zea mays 285, 286

