METHODS IN MOLECULAR BIOLOGY™

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

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

Fungal Genomics

Methods and Protocols

Edited by

Jin-Rong Xu

Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN, USA

Burton H. Bluhm

Department of Plant Pathology, Division of Agriculture, University of Arkansas, Fayetteville, AR, USA

Humana Press

Editors
Jin-Rong Xu
Department of Botany
and Plant Pathology
Purdue University
West Lafayette, IN
USA
jinrong@purdue.edu

Burton H. Bluhm
Department of Plant Pathology
Division of Agriculture
University of Arkansas
Fayetteville, AR
USA
bbluhm@uark.edu

ISSN 1064-3745 e-ISSN 1940-6029
ISBN 978-1-61779-039-3 e-ISBN 978-1-61779-040-9
DOI 10.1007/978-1-61779-040-9
Springer New York Dordrecht Heidelberg London

Library of Congress Control Number: 2011923225

Printed on acid-free paper

Humana Press is part of Springer Science+Business Media (www.springer.com)

Preface

Fungal genomics has experienced unprecedented growth since the turn of the millennium. Starting with the completion of the first fungal genomes nearly 10 years ago, the genomes of over 60 species spanning major taxonomic groups and ecological niches have been sequenced. The rate at which fungal genomes are being sequenced has increased dramatically with the refinement of next-generation sequencing technologies, making genomics-based approaches feasible for a broad range of fungi. This dramatic expansion of resources and techniques is poised to fundamentally redefine the study of fungal biology.

In upcoming years, fungal genomics is likely to advance on three fronts. First, more and more genomes will be sequenced. To assist readers in this ongoing process, we present chapters describing techniques for genome sequencing and assembly, including a discussion of next-generation sequencing technologies. Second, sequenced genomes will be mined extensively for useful information. To this end, we have included chapters that describe protocols and programs to identify and analyze telomeres and repetitive sequences in the fungal genomes. Third, genomic sequences will provide a foundation for powerful techniques to explain biological processes, and much of this book is dedicated to explaining established and emerging genomics-based technologies in filamentous fungi. Four chapters describe gene expression profiling techniques, including expressed sequence tags (ESTs) and microarrays. Three chapters describe techniques for fungal proteomics, including how to identify proteins in a given biological sample, affinity purification of proteins based on protein–protein interaction, and how ChIP-chip can be used to study promoter elements and other functions at the chromatin or DNA–protein interaction levels. Other chapters provide case studies that could be adapted to a wide range of fungi, including procedures to generate, characterize, and manage a large number of knockout mutants in *Neurospora crassa*, the study of mycoviruses and hypovirulence in the chestnut blight fungus, metabolic fingerprinting in *Fusarium verticillioides* to determine gene function, and large-scale insertional mutagenesis in *Magnaporthe oryzae* to identify novel virulence or pathogenicity factors.

Contributors to this book were urged to emphasize unpublished tips, potential pitfalls, common mistakes, and special considerations based on their unique experiences. Our goal was to provide fungal biologists at any stage of their careers a user-friendly resource for fungal genomics, especially as readers branch out into unfamiliar but exciting new areas of study.

Particular thanks to all of the contributing authors as well as to Dr. John Walker and the entire Humana Press editorial staff.

West Lafayette, IN — ***Jin-Rong Xu***
Fayetteville, AR — ***Burton H. Bluhm***

Contents

Preface . . . . . *v*
Contributors . . . . . *ix*

1 Genome Sequencing and Assembly 1
Manfred G. Grabherr, Evan Mauceli, and Li-Jun Ma

2 Targeted Cloning of Fungal Telomeres 11
Mark L. Farman

3 Identification and Annotation of Repetitive Sequences in Fungal Genomes 33
Braham Dhillon and Stephen B. Goodwin

4 Next-Generation Sequencing and Potential Applications in Fungal Genomics 51
Phillip SanMiguel

5 Getting the Most Out of Your Fungal Microarray Data: Two Cost- and Time-Effective Methods 61
Sandra M. Mathioni, André Beló, Jeffrey P. Townsend, and Nicole M. Donofrio

6 *Fusarium graminearum* from Expression Analysis to Functional Assays 79
Heather E. Hallen-Adams, Brad L. Cavinder, and Frances Trail

7 EST Analysis Pipeline: Use of Distributed Computing Resources 103
Francisco Javier González and Juan Antonio Vizcaíno

8 The Application of ChIP-chip Analysis in the Rice Blast Pathogen 121
Soonok Kim and Thomas K. Mitchell

9 Proteome Studies of Filamentous Fungi 133
Scott E. Baker and Ellen A. Panisko

10 A Bioinformatics Pipeline for Sequence-Based Analyses of Fungal Biodiversity 141
D. Lee Taylor and Shawn Houston

11 Identifying Protein Complexes by Affinity Purification and Mass Spectrometry Analysis in the Rice Blast Fungus 157
Wende Liu, Anton Iliuk, Andy Tao, and Shengli Ding

12 Large Scale Identification of Genes Involved in Plant–Fungal Interactions Using Illumina's Sequencing-by-Synthesis Technology 167
R.C. Venu, Yuan Zhang, Brian Weaver, Peter Carswell, Thomas K. Mitchell, Blake C. Meyers, Michael J. Boehm, and Guo-Liang Wang

13 High-Throughput Production of Gene Replacement Mutants in *Neurospora crassa* 179
Gyungsoon Park, Hildur V. Colot, Patrick D. Collopy, Svetlana Krystofova, Christopher Crew, Carol Ringelberg, Liubov Litvinkova, Lorena Altamirano, Liande Li, Susan Curilla, Wei Wang, Norma Gorrochotegui-Escalante, Jay C. Dunlap, and Katherine A. Borkovich

14 Phenotypic Analysis of *Neurospora crassa* Gene Deletion Strains 191
Gloria E. Turner

15 Efficient Approaches for Generating GFP Fusion and Epitope-Tagging Constructs in Filamentous Fungi 199
Xiaoying Zhou, Guotian Li, and Jin-Rong Xu

16 Large-Scale Insertional Mutagenesis in *Magnaporthe oryzae* by *Agrobacterium tumefaciens*-Mediated Transformation 213
Xiao-Lin Chen, Jun Yang, and You-Liang Peng

17 Molecular Methods for Studying the *Cryphonectria parasitica* – Hypovirus Experimental System. 225
Angus L. Dawe, Rong Mu, Gloricelys Rivera, and Joanna A. Salamon

18 Metabolic Fingerprinting in *Fusarium verticillioides* to Determine Gene Function 237
Jonathon E. Smith and Burton H. Bluhm

19 Tapping Genomics to Unravel Ectomycorrhizal Symbiosis 249
Jonathan M. Plett, Barbara Montanini, Annegret Kohler, Simone Ottonello, and Francis Martin

Index *283*

Contributors

Lorena Altamirano • *Department of Plant Pathology and Microbiology, University of California, Riverside, CA, USA*

Scott E. Baker • *Pacific Northwest National Laboratory, Richland, WA, USA*

André Beló • *Department of Plant and Soil Sciences, University of Delaware, Newark, DE, USA*

Burton H. Bluhm • *Department of Plant Pathology, Division of Agriculture, University of Arkansas, Fayetteville, AR, USA*

Michael J. Boehm • *Department of Plant Pathology, The Ohio State University, Columbus, OH, USA*

Katherine A. Borkovich • *Department of Plant Pathology and Microbiology, University of California, Riverside, CA, USA*

Peter Carswell • *Department of Plant Pathology, The Ohio State University, Columbus, OH, USA*

Brad L. Cavinder • *Department of Plant Biology and Plant Pathology, Michigan State University, East Lansing, MI, USA*

Xiao-Lin Chen • *State Key Laboratory of Agrobiotechnology and Department of Plant Pathology, China Agricultural University, Beijing, China*

Patrick D. Collopy • *Department of Plant Pathology and Microbiology, University of California, Riverside, CA, USA*

Hildur V. Colot • *Department of Genetics, Dartmouth Medical School, Hanover, NH, USA*

Christopher Crew • *Department of Plant Pathology and Microbiology, University of California, Riverside, CA, USA*

Susan Curilla • *Department of Plant Pathology and Microbiology, University of California, Riverside, CA, USA*

Angus L. Dawe • *Department of Biology and Molecular Biology Program, New Mexico State University, Las Cruces, NM, USA*

Braham Dhillon • *Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN, USA*

Shengli Ding • *Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN, USA*

Nicole M. Donofrio • *Department of Plant and Soil Sciences, University of Delaware, Newark, DE, USA*

Jay C. Dunlap • *Department of Genetics, Dartmouth Medical School, Hanover, NH, USA*

Mark L. Farman • *Department of Plant Pathology, University of Kentucky, Lexington, KY, USA*

Francisco Javier González • *Oncostem Pharma, Salamanca, Spain*

Stephen B. Goodwin • *USDA-ARS, Crop Production and Pest Control Research Unit, Purdue University, West Lafayette, IN, USA*

Norma Gorrochotegui-Escalante • *Department of Genetics, Dartmouth Medical School, Hanover, NH, USA*
Manfred G. Grabherr • *The Broad Institute Cambridge Center, Cambridge, MA, USA*
Heather E. Hallen-Adams • *Department of Plant Biology, Michigan State University, East Lansing, MI, USA*
Shawn Houston • *Institute of Arctic Biology, University of Alaska, Fairbanks, AK, USA*
Anton Iliuk • *Department of Biochemistry, Purdue University, West Lafayette, IN, USA*
Soonok Kim • *Department of Plant Pathology, The Ohio State University, Columbus, OH, USA*
Annegret Kohler • *INRA-Nancy, Champenoux, France*
Svetlana Krystofova • *Department of Plant Pathology and Microbiology, University of California, Riverside, CA, USA*
Guotian Li • *Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN, USA*
Liande Li • *Department of Plant Pathology and Microbiology, University of California, Riverside, CA, USA*
Liubov Litvinkova • *Department of Plant Pathology and Microbiology, University of California, Riverside, CA, USA*
Wende Liu • *Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN, USA*
Li-Jun Ma • *University of Massachusetts, Amkebt, MA, USA*
Francis Martin • *INRA-Nancy, Champenoux, France*
Sandra M. Mathioni • *Department of Plant and Soil Sciences, University of Delaware, Newark, DE, USA*
Evan Mauceli • *The Broad Institute Cambridge Center, Cambridge, MA, USA*
Blake C. Meyers • *Department of Plant and Soil Sciences, Delaware Biotechnology Institute, Delaware Technology Park, University of Delaware, Newark, DE, USA*
Thomas K. Mitchell • *Department of Plant Pathology, The Ohio State University, Columbus, OH, USA*
Barbara Montanini • *INRA-Nancy, Champenoux, France*
Rong Mu • *Molecular Biology Program, New Mexico State University, Las Cruces, NM, USA*
Simone Ottonello • *INRA-Nancy, Champenoux, France*
Ellen A. Panisko • *Pacific Northwest National Laboratory, Richland, WA, USA*
Gyungsoon Park • *Department of Plant Pathology and Microbiology, University of California, Riverside, CA, USA*
You-Liang Peng • *State Key Laboratory of Agrobiotechnology and Department of Plant Pathology, China Agricultural University, Beijing, China*
Jonathan M. Plett • *INRA-Nancy, Champenoux, France*
Carol Ringelberg • *Department of Plant Pathology and Microbiology, University of California, Riverside, CA, USA*
Gloricelys Rivera • *Department of Biology, New Mexico State University, Las Cruces, NM, USA*

Joanna A. Salamon • *Department of Biology, New Mexico State University, Las Cruces, NM, USA*
Phillip SanMiguel • *Purdue Genomics Core Facility, Purdue University, West Lafayette, IN, USA*
Jonathon E. Smith • *Department of Plant Pathology, Division of Agriculture, University of Arkansas, Fayetteville, AR, USA*
Andy Tao • *Department of Biochemistry, Purdue University, West Lafayette, IN, USA*
D. Lee Taylor • *Institute of Arctic Biology, University of Alaska, Fairbanks, AK, USA*
Jeffrey P. Townsend • *Department of Ecology and Evolutionary Biology, Yale University, New Haven, CT, USA*
Frances Trail • *Departments of Plant Biology and Plant Pathology, Michigan State University, East Lansing, MI, USA*
Gloria E. Turner • *Department of Chemistry and Biochemistry, University of California, Los Angeles, CA, USA*
R.C. Venu • *Department of Plant Pathology, The Ohio State University, Columbus, OH, USA*
Juan Antonio Vizcaíno • *EMBL-EBI, Wellcome Trust Genome Campus, Hinxton, Cambridge, UK*
Guo-Liang Wang • *Department of Plant Pathology, The Ohio State University, Columbus, OH, USA*
Wei Wang • *Department of Genetics, Dartmouth Medical School, Hanover, NH, USA*
Brian Weaver • *Department of Plant Pathology, The Ohio State University, Columbus, OH, USA*
Jin-Rong Xu • *Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN, USA*
Jun Yang • *State Key Laboratory of Agrobiotechnology and Department of Plant Pathology, China Agricultural University, Beijing, China*
Yuan Zhang • *Department of Plant Pathology, The Ohio State University, Columbus, OH, USA*
Xiaoying Zhou • *Department of Botany and Plant Pathology, Purdue University, West Lafayette, IN, USA*

Chapter 1

Genome Sequencing and Assembly

Manfred G. Grabherr, Evan Mauceli, and Li-Jun Ma

Abstract

Decoding the genome sequence is becoming a fundamental tool for molecular, genetic, and genomic studies. This chapter reviews the history of DNA sequencing and technical principles of different sequencing platforms, and compares the strengths and weaknesses of different techniques for high-throughput genome sequencing applications are compared. It also covers brief descriptions on genome assembly and its validation.

Key words: Genome sequencing, Genome assembly, Next-generation sequencing, Single-molecule sequencing

1. Introduction

Deoxyribonucleic acid (DNA) molecules are the bearers of information needed for the development and functioning of all cells in a living organism. The DNA molecules of each organism are composed of chains of monomeric nucleotides (the four bases adenine, "A"; cytosine, "C"; guanine "G"; and thymine, "T") and are organized into chromosomes, called genome. To understand the mechanisms that govern the complex biological processes of an individual and to unveil the genetic differences among different organisms, it is necessary to know the composition of the DNA molecules through the process of genome sequencing.

To date, no technologies are available to read the entire sequence of a single molecule directly. Instead, various methods have been developed to decode small snippets of DNA, one piece at a time. In principle, these small sequence reads can be pieced together into longer pieces by bioinformatics analysis, ultimately to reach the completing of the genome, as long as there are sufficient amount of information. In reality, however, each

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722,
DOI 10.1007/978-1-61779-040-9_1,

genome project has to face multiple challenges, including to (a) obtain accurate sequence for each fragment, (b) get readings of sequences from the entire genome equally, (c) get readings that are long enough so that repetitive regions can be resolved, and (d) produce massive amounts of data quickly and inexpensively. In summary, the determining factors for each genome project are *accuracy*, *completeness*, *cost*, and *time*.

Earlier, most genomic projects were focused on the accuracy and completeness of issues. Since the early 2000s, the focus of genomic projects shifted, as a variety of sequencing technologies that emphasize on the cost and time issues have been established. Here, we briefly introduce the main sequence technologies, including the Sanger method and the emerging next-generation sequencing (NGS) technologies, and compare their strengths and weaknesses for genome sequencing applications.

2. Sanger Sequencing

In the 1970s, two pioneer sequencing technologies, the Maxam–Gilbert (1) and the Sanger method (2), were developed almost simultaneously, and the inventors of these technologies, Sanger, Maxam, and Gilbert were awarded the Nobel Prize. While the Maxam–Gilbert sequencing method was widely used initially, the development of automated high-throughput DNA sequence analyzers made Sanger sequencing the method of choice for all genome sequencing projects.

Frederick Sanger and colleagues introduced the chain termination method of sequencing in 1977 (3), which remained the standard method of DNA sequencing for the next 30 years. Chain termination sequencing starts with a preparation of identical single-stranded DNA (ssDNA) molecules (clonal templates). A short oligonucleotide is annealed to the same position in each ssDNA molecule as the primer for the synthesis of a new DNA strand complementary to the template strand. The DNA polymerase synthesizes the new strand by incorporating the appropriate deoxyribonucleotides (A, C, G, and T). A low concentration of dideoxyribonucleotides present in the reaction terminates the nascent strand when a dideoxyribonucleotide is incorporated. Strands terminated by each of the four bases (A, C, G, and T) are size-fractioned by electrophoresis. The smallest bit of DNA travels the farthest through the polyacrylamide gel and corresponds to the first base of the template DNA. By reading the banding pattern in the gel, one can reconstruct the sequence of the template DNA strand.

The ascendency of Sanger sequencing as the choice for genome sequencing applications was due to a number of key

laboratory and computational innovations. Many of these technical improvements were spurred by the sequencing of the human genome (4, 5). A pair of key laboratory innovations (although by no means a complete list) were the replacement of the original radioactive labels used to band detection in the gel with fluorescent labels (6, 7) and the replacement of the slab polyacrylamide gel with capillary gels (8, 9). Development of the PHRED (10, 11) software package introduced the concept of a "base-quality score" to each base, which is related to the probability of a base being sequenced incorrectly and allows for a quick assessment of the quality of raw data generated by the sequencing machines.

3. Genome Assembly

Chain termination sequencing can identify a stretch of several hundred nucleotides in a single "read" of genomic sequence. The goal of whole-genome sequencing is to reconstruct the master sequence of an organism's chromosomal DNA from these reads. This process is called "assembly." In early genome sequencing projects for organisms with small genomes, assembling the sequencing reads was done by hand. For more complex organisms, assembly is a computational task.

There are two major components to ensure the success of a whole-genome assembly (WGA): sequence accuracy and long-range continuity. In addition to the base-quality score, sequence accuracy can be achieved through increasing redundancy. Usually, each base in the genome is captured five to ten times in the sequence reads. Long-range continuity is accomplished by the construction of libraries of various sizes. These libraries are produced by cloning size-selected fragments of genomic DNA into a specific vector transforming them into the host *Escherichia coli*. Both ends of the insert DNA are then sequenced to generate two related reads (read "pairs") with a known genomic distance. This pairing information provides the long-range continuity needed. Most genome projects combine libraries of 2–10 kilobases (kb; plasmids), ~40 kb (fosmids), and ~150 kb (bacterial artificial chromosomes or BACs).

Given the read sequences and their pairing information, the assembly process begins with an alignment process, where overlapping sequences that originate from the same genomic location are identified. These are stitched together into larger continuous stretches of genome called "contigs." Using the pairing information, these contigs are ordered and oriented to reflect their proximity in the genome. This process is complicated by a number of factors: (a) wrongly called or missing sequence in the reads, or reads that actually come from two distinct regions of the genome

("chimerism") instead of one contiguous region; (b) cloning bias, where some regions of the genome are not represented in the reads, usually because they are lethal to the host *E. coli* cells; (c) repetitive sequences, such as transposable elements, tandem repeats, and segmental duplications that exceed the length of a read, making it difficult to map a read back to its correct copy of the repeat in the genome; (d) polymorphism in diploid (or polyploidy) genomes that cause the read set to contain a mixture of sequences from two sister chromosomes; (e) large data sets, especially for mammalian-sized genomes, which are on the order of two to three billion nucleotides, requiring a great deal of attention to ensure algorithms are efficient in both run-time and memory usage.

Over the years, a number of WGA software program packages have been developed. An incomplete list includes: SEQAID (12), CAP (13), PHRAP (14), TIGR assembler (15), AMASS (16), the Celera assembler (17), EULER (18), Jazz (19), Phusion (20), PCAP (21), Arachne versions 1 and 2 (22, 23), and Arachne 3 (http://www.broadinstitute.org/science/programs/genome-biology/crd/).

Even though WGA is by no means a solved problem with a clear solution that can be applied to all genomes, the efficacy of the Sanger sequencing and assembly method is demonstrated by the over 10,000 sequenced genomes across every kingdom of life present in the National Center for Biotechnology Information's (NCBI) data repository (http://www.ncbi.nlm.nih.gov/sites/entrez?db=genome). All fungal genomes sequenced at the Broad Institute are assembled with the Arachne package (22, 23). Most of these fungal genome assemblies have gaps. Some of these gaps may be resulted from DNA sequences that are not clonable in *E. coli* or difficult for Sanger sequencing. It is also true that most fungal genome sequencing projects have excluded sequence reads. Many of these unassembled reads are repetitive sequences.

4. Validation of Sequence Assembly

Genome assemblies are often validated by comparison with genetic and physical maps. Ideally, they should match perfectly with each other. For some organisms, such as asexual fungi, genetic maps are not available. Techniques, such as HAPPY mapping and optical mapping, can be used as alternative approaches to validate genome assemblies. In HAPPY mapping, PCR assays with native genomic DNA are used to determine the order and spacing of DNA markers. This method can be used to construct regional or genome-wide physical maps. For optical mapping, a set of restriction enzymes are used to digest high-molecular weight DNA molecules bound under tension to a glass surface.

The restriction patterns of individual DNA molecules are visualized by fluorescence microscopy. For validation, the order and distance between restriction sites obtained by optical mapping are compared with in silico digests of genome assemblies.

5. Genome Sequencing with NGS Technologies

While Sanger sequencing produces long, high-quality reads, it is relatively expensive and time-consuming to complete large genome sequencing projects. Even for sequencing small fungal genomes of 40 Mb in size, it can cost up to $500,000. In the past few years, the so-called NGS technologies have been developed to generate data at considerably lower cost by massive parallel sequencing, albeit at the expense of sequence quality and/or read length. Below, we briefly discuss their advantages and disadvantages in genome sequencing projects. The performance of different platform was collected around summer 2009, when we were developing this manuscript.

5.1. 454 Pyrosequencing

Pyrosequencing, the first NGS platform ready for practical use, was developed in 1996 (24) and brought to the market by *454 Life Sciences* around 2005. The platform creates clonally amplified DNA fragments through emulsion PCR (ePCR) that amplify individual single-stranded, bead-bound small DNA fragment in a water-in-oil mixture. Millions of such clonal DNA fragments are attached to a picotiter plate and sequenced simultaneously. For each sequencing reaction, only one type of nucleotide (T, A, C, or G) is added. Optical brightness of newly added nucleotides is detected with a charge coupled device (CCD) camera. Such synthesis process is repeated for all four bases in a fixed order. The signal strength of each cycle is correlated with the number of nucleotides incorporated. For example, a homopolymer stretch (AAAA) generates a stronger signal than a single nucleotide (A). Because the light signal detected in each cycle is not directly proportional to the number of bases in a homopolymer, the prevalent error patterns of 454 sequencing are insertions and deletions when the number of bases in a run was estimated incorrectly. Its nucleotide substitution error rate is similar to that of Sanger sequencing, making this platform suitable for single nucleotide polymorphism (SNP) detection even at low sequence coverage (25). The latest system of 454 pyrosequencing (Roche GS-FLX) produces about one billion bases per day at the cost of less than 10 cents per kilobase. Its read length has reached about 400 bp and may get longer through further refinements of the technology. Because of its read lengths and low substitution errors, *454* sequencing is suitable for de novo genome sequencing and

assembly. This process can be performed in flow space (i.e., using intensity signals rather than base calls), which greatly alleviates the homopolymer problem after the consensus is built from individual reads. Newbler, an assembly program specifically developed for data generated in this platform, is distributed with 454 sequencing machines (http://www.454.com/).

5.2. Illumina Sequencing

For the *Illumina* platform, each sequence run takes 2–3 days and yields over 1 Gb of sequence in up to 100 bp/read. The cost, at less than a cent per kilobase, makes it one of the most inexpensive options among NGS technologies. Small DNA fragments (100–300 bp) are attached to the surface of a flow cell through adapters to both ends of the single-stranded fragments. Each attached DNA fragment is locally amplified using the adapters as primers to form a small clonal DNA cluster. Each flow cell contains hundreds of millions of such clusters. These templates are sequenced in cycles base-by-base, by adding four labeled reversible terminators with removable fluorescent dyes. The base-by-base sequencing reactions used in this platform eliminate sequence context-specific errors and enable sequencing through homopolymers and repetitive sequences. *Illumina* has been mostly used for re-sequencing. However, the relatively high base error rates (more than 1 in 100 bp) require high redundancy (30-fold oversampling) to accurately call polymorphisms. While it is challenging to create de novo genome assemblies with sequence data of such high base error rates, various assemblers have developed, such as Velvet (26), ALLPATHS (27), and SOAPdenovo (http://soap.genomics.org.cn/).

5.3. ABI SOLiD Sequencing

The *ABI SOLiD* system uses hybridization–ligation methodologies for massively parallel sequencing. Currently, the system produces 20–40 Gb (25 bp reads) in an 8–10-day sequencing run, at a cost even lower than *Illumina*. Although short, the *SOLiD* reads have the lowest error rates among NGS technologies, making it particularly attractive for re-sequencing. The initial ePCR step to generate clonal DNA fragments is similar to that of the *454* platform. Amplified products are then covalently linked to a glass surface. Sequencing is carried out using random 8-mer probes with the first and second position containing dinucleotides, which are semi-degeneratively labeled with a fluorescent dye. Each nucleotide position is ascertained using a four-dye encoding schema, and each position is interrogated twice. As a result, every base is read in two different dinucleotide frames allowing for error correction. This scheme delivers raw read accuracy in excess of 99.9% (with some variation depending on the base position in the read), and provides even more power in distinguishing sequencing errors from SNPs, as a single sequencing error manifests itself as a single wrong color, whereas

an SNP appears as two consecutive mismatched colors with a given pattern.

5.4. Single Molecule Sequencing

Sequencing technologies described above require the amplification of individual DNA fragments in bacteria or in vitro before sequencing. A number of sequencing technologies that eliminate this amplification step are actively under development. The Helicose Single Molecule Sequencer is the first commercially available single-molecule sequencing platform (28). The sequencing process begins with fragmentation of DNA followed by poly-A tailing of the fragments. These template libraries are hybridized to an array of poly-T oligomers that are tethered to a planar surface. For each sequencing cycle, a single fluorescently labeled nucleotide is added and incorporated to the template strand. The array is imaged, the fluorescent label is cleaved off, and the next round of extension and imaging takes place. After numerous cycles, the process yields read lengths of about 30 bases. The dominant error pattern of this platform is deletions (<5% raw error rate). It also had a substantial rate of substitution errors (0.5%). This sequencing technology, with a throughput of approximately 150 Mb per hour and >20 Gb in a single run (28), is new and remains to be tested its potential applications.

The Pacific Biosciences' single-molecule real-time sequencing (SMRT) platform has been under development since 2004. The sequencing process begins with an SMRT chip, which contains thousands of zero-mode waveguides (ZMWs). A ZMW is a hole in a semiconductor that creates an illuminated observation volume that is small enough to observe a single nucleotide being incorporated by DNA polymerase (29). Each of the four bases is attached to one of four different fluorescent dyes. A detector captures the fluorescent signal of the nucleotide incorporation. The SMRT platform is not yet commercially available, but Pacific Biosystems has published sequencing results, where the dominant error pattern was found to be deletions (<8% raw error rate), with a similar rate (<5%) for mismatches (30). This sequencing platform promises read lengths in the kilobase range with minutes of running time and has potential applications in de novo whole-genome sequencing and re-sequencing projects.

Nanopore-based devices work by driving a single DNA molecule through a nanoscale pore. As individual nucleotides passing through the pore, the ionic current in the nanopore is modulated in a nucleotide-specific manner, enabling a direct readout of the DNA sequence. A number of nanopore designs are being investigated for sequencing, including a pore-forming protein (a "biopore") and a man-made solid-state pore (29, 31). A major defect of these designs is that the current change due to a single nucleotide passing through the pore is masked by field-effects from nearby nucleotides. Although this problem may

be addressed by using a solid-state pore articulated with tunneling probes (32), there are a number of other technological challenges that need to be addressed. However, the promise of nanopore-based sequencing technologies is immense. Read lengths can be in the tens of kilobases. Sequencing of a mammalian genome can take roughly a day and cost under $1,000.

6. Conclusive Remarks

With the development of various high-throughput sequencing technologies, which hold the promises to lower the cost of DNA sequencing with much increased speed. Ultimately, the replacement of the Sanger sequencing technology is inevitable for genome sequencing. When major challenges, such as lower error rates and longer reads, are solved, sequencing of entire human genomes may become routine and play a vital role in medicine and health care of the future. The goal set by the US National Institutes of Health to sequence a mammalian genome in a day at the cost of under $1,000 is certainly reachable.

References

1. Maxam AM, and Gilbert W. (1977) A new method for sequencing DNA. *Proc. Natl. Acad. Sci. USA.* **74**, 560–4.
2. Sanger F, and Coulson AR. (1975) A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *J. Mol. Biol.* **94**, 441–8.
3. Sanger F, Air GM, Barrell BG, et al. (1977) Nucleotide sequence of bacteriophage phi X174 DNA. *Nature* **265**, 687–95.
4. Lander ES, Linton LM, Birren B, et al. (2001) Initial sequencing and analysis of the human genome. *Nature* **409**, 860–921.
5. Venter JC, Adams MD, Myers EW, et al. (2001) The sequence of the human genome. *Science* **291**, 1304–51.
6. Prober JM, Trainor GL, Dam RJ, et al. (1987) A system for rapid DNA sequencing with fluorescent chain-terminating dideoxynucleotides. *Science* **238**, 336–41.
7. Smith LM, Sanders JZ, Kaiser RJ, et al. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–9.
8. Luckey JA, Drossman H, Kostichka AJ, et al. (1990) High speed DNA sequencing by capillary electrophoresis. *Nucleic Acids Res.* **18**, 4417–21.
9. Swerdlow H, and Gesteland R. (1990) Capillary gel electrophoresis for rapid, high resolution DNA sequencing. *Nucleic Acids Res.* **18**, 1415–9.
10. Ewing B, and Green P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res.* **8**, 186–94.
11. Ewing B, Hillier L, Wendl MC, and Green P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res.* **8**, 175–85.
12. Peltola H, Soderlund H, and Ukkonen E. (1984) SEQAID: a DNA sequence assembling program based on a mathematical model. *Nucleic Acids Res.* **12**, 307–21.
13. Huang X, and Madan A. (1999) CAP3: A DNA sequence assembly program. Genome Res. **9**, 868–77.
14. Green P. PHRAP documentation. http://wwwphraporg 1994.
15. Sutton G, White O, Adams M, and Kerlavage A. (1995) TIGR assembler: A new tool for assembling large shotgun sequencing projects. *Genome Sci Technol.* **1**, 9–19.
16. Kim S, Segre AM. (1999) AMASS: a structured pattern matching approach to shotgun sequence assembly. *J. Comput Biol.* **6**, 163–86.
17. Myers EW, Sutton GG, Delcher AL, et al. (2000) A whole-genome assembly of Drosophila. *Science* **287**, 2196–204.

18. Pevzner PA, Tang H, and Waterman MS. (2001) An Eulerian path approach to DNA fragment assembly. *Proc. Natl. Acad. Sci. USA*. **98**, 9748–53.
19. Dehal P, Satou Y, Campbell RK, et al. (2002) The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. *Science* **298**, 2157–67.
20. Mullikin JC, and Ning Z. (2003) The phusion assembler. *Genome Res.* **13**, 81–90.
21. Huang X, Wang J, Aluru S, Yang SP, and Hillier L. (2003) PCAP: a whole-genome assembly program. *Genome Res.* **13**: 2164–70.
22. Batzoglou S, Jaffe DB, Stanley K, et al. (2002) ARACHNE: a whole-genome shotgun assembler. *Genome Res.* **12**, 177–89.
23. Jaffe DB, Butler J, Gnerre S, et al. (2003) Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13**, 91–6.
24. Ronaghi M, Karamohamed S, Pettersson B, Uhlen M, and Nyren P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal Biochem.* **242**, 84–9.
25. Brockman W, Alvarez P, Young S, et al. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res.* **18**, 763–70.
26. Zerbino DR, and Birney E. (2008) Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Res.* **18**, 821–9.
27. Butler J, MacCallum I, Kleber M, et al. (2008) ALLPATHS: *de novo* assembly of whole-genome shotgun microreads. *Genome Res.* **18**, 810–20.
28. Harris TD, Buzby PR, Babcock H, et al. (2008) Single-molecule DNA sequencing of a viral genome. *Science* **320**, 106–9.
29. Levene MJ, Korlach J, Turner SW, Foquet M, Craighead HG, and Webb WW. (2003) Zero-mode waveguides for single-molecule analysis at high concentrations. *Science* **299**, 682–6.
30. Eid J, Fehr A, Gray J, et al. (2009) Real-time DNA sequencing from single polymerase molecules. *Science* **323**, 133–8.
31. Li J, Stein D, McMullan C, Branton D, Aziz MJ, and Golovchenko JA. (2001) Ion-beam sculpting at nanometre length scales. *Nature* **412**, 166–9.
32. Zwolak M, Di and Ventra M. (2005) Electronic signature of DNA nucleotides via transverse transport. *Nano Lett.* **5**, 421–4.

Chapter 2

Targeted Cloning of Fungal Telomeres

Mark L. Farman

Abstract

Telomeres are the sequences that form the ends of eukaryotic chromosomes and are essential structures that confer genome stability and guide chromosome behavior. In addition, the terminal regions of the chromosomes tend to house genes with predicted roles in ecological adaptation. Unfortunately, however, most fungal genome assemblies contain very few telomeres and, therefore, the identities of genes residing near the chromosome ends are often unknown. In an effort to develop a complete understanding of the organization and gene content of chromosome ends in a number of fungi, we developed efficient methods for the identification and targeted cloning of telomeres. This chapter describes the basic steps and shows exemplary results from the targeted cloning of *Epichloë festucae* telomeres.

Key words: Genome sequencing, Subcloning, Southern blotting, Colony blotting

1. Introduction

Telomeres protect chromosome ends from degradation caused by normal DNA replication processes and enzymatic activity. They also play important roles in chromosome biology by initiating chromosome pairing (1, 2) and directing chromosome movement (3, 4). The chromosome regions near the telomeres tend to be highly dynamic (5–7), show increased genetic variation (8, 9) and often house genes that enhance an organism's adaptive capabilities (10). Finally, there is evidence that fungi possess specialized mechanisms for regulating the expression of telomere-linked genes (11, 12). However, despite their obvious importance, telomeres are frequently missing from fungal genome sequences (13–15). This is disadvantageous for two reasons. First, the telomeres can be important landmarks for guiding the genome assembly process and verifying the final product. Second, if the telomeres are absent, the sequences that reside at the chromosome ends are unknown and, therefore, information on fungal

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722,
DOI 10.1007/978-1-61779-040-9_2,

genes with potential evolutionary and ecological significance remains elusive.

Analysis of raw sequence data has shown that telomere sequences are frequently captured in genome sequencing efforts but simply escape assembly. Such sequences can be identified and incorporated into genome assemblies through the use of the bioinformatic tool TERMINUS (13). However, even after exhaustive mining of raw sequence data, we find that most genome assemblies still lack sequence information for a number of telomeres. Therefore, in most cases, it is necessary to clone the missing chromosome ends using a targeted approach.

Native telomeres are refractory to cloning because their 3′ ends protrude as single stranded tails (16). Therefore, the overhanging nucleotides must be removed before the telomeres can be cloned. Once this has been achieved, however, it is possible to take advantage of the telomere's terminal position to enrich for telomeric restriction fragments. This can be accomplished by using a directional cloning strategy to select for fragments that are blunt at one end and sticky at the other. In this manner, only fragments that are at the ends of DNA molecules are recovered.

2. Materials

2.1. Extraction of Genomic DNA

1. Lysis buffer: 0.5 M NaCl, 1% sodium dodecyl sulfate (SDS), 10 mM Tris–HCl, pH 8, 1 mM EDTA, pH 8. Store at room temperature. Heat at 65°C before using to dissolve the SDS.
2. Phenol:chloroform:isoamylalcohol. 25:24:1 (PCI): 25 ml phenol equilibrated with Tris–HCl, pH 8; 24 ml chloroform (equilibrated with Tris–HCl, pH 8), 1 ml isoamylalcohol. Store under 0.1 M Tris–HCl, pH 8 in a tightly capped, dark glass bottle at 4°C. Before using, check that the PCI is colorless. Discard and make up a fresh batch if any hint of color is detected.
3. Chloroform:isoamylalcohol 24:1 (CI) 24 ml chloroform (equilibrated with Tris–HCl, pH 8), 1 ml isoamylalcohol. Store in a tightly capped, dark glass bottle at 4°C.
4. T0.1E buffer: 10 mM Tris–HCl, pH 8, 0.1 mM EDTA, pH 8. Store at room temperature.

2.2. Agarose Gel Electrophoresis

1. TBE (10× stock): Add the following to 800 ml of H_2O: 108 g Tris base, 55 g boric acid and 9.3 g EDTA. Adjust volume to 1 l with additional H_2O. Make a 0.5× working solution by diluting 20-fold in H_2O.
2. A 0.7% agarose gel solution (200 ml): add 1.4 g agarose to 200 ml 0.5× TBE. Heat in a microwave at high setting for 4–5 min. Swirl to ensure that all the agarose has dissolved

fully. The agarose solution can be stored molten in a 55°C oven for up to 2 days until needed.

3. Loading Dye (6× stock): 10 mM Tris–HCl, pH 8, 60 mM EDTA, 0.03% bromophenol blue.
4. Kilobase plus DNA size marker (Invitrogen, Carlsbad, CA).
5. Parafilm® (Alcan, Inc., Montreal, QE).
6. Ethidium bromide: Make a stock solution containing 5 mg/ml ethidium bromide.

2.3. Electroblotting

1. Denaturation solution: 0.5 M NaOH. Store at room temperature.
2. 20× SSC: 3 M NaCl, 0.3 M Na citrate, pH 7. Store at room temperature. Make working solutions by diluting in H_2O.
3. Jumbo Genie blotting apparatus (Idea Scientific, Minneapolis, MN).
4. Pall Biodyne B Hybridization membrane (Pall Corp., Pensacola, FL).
5. Whatman 3M paper (Whatman, Florham Park, NJ).

2.4. Southern Hybridization Analysis

1. Telomere oligonucleotides: TEL1, 5′-TTAGGGTTAGGGTTAGGGTTAGGG-3′ and TEL2, 5′-CCCTAACCCTAACCCTAACCCTAA-3′.
2. ExTaq PCR reagents (Takara, Shiga, Japan).
3. Labeling kit (Promega Corp., Madison, WI).
4. Dye Stop solution: 10 mM Tris–HCl, pH 8, 10 mM EDTA, pH 8, 0.8% dextran blue, 0.04% orange G.
5. Illustra MicroSpin™ G50 columns (GE Healthcare, Piscataway, NJ).
6. Hybridization solution: 0.125 M $NaHPO_4$ (from a 4× stock consisting of 0.5 M Na_2HPO_4 that has been adjusted to pH 7.5 with phosphoric acid), 7% SDS, 1 mM EDTA (use 0.5 M EDTA, pH 8 stock).
7. Wash solutions: low stringency, 2× SSC; high stringency, 0.1× SSC, 0.1% SDS.

2.5. Plasmid Vector

1. The vector we use for telomere cloning, pBS-TEL1, is based on the pBluescript KS II+ (Stratagene, La Jolla, CA) and contains a ~2 kb *Eco*RI "stuffer" fragment inserted into the *Eco*RI site (Fig. 1). The advantage of using this particular plasmid is explained in Note 1.

2.6. Molecular Biology Reagents

1. Restriction enzymes: *Hin*dIII and *Sma*I (New England Biolabs, Ipswich, MA).
2. Bovine serum albumin (BSA): make a 10× stock (1 mg/ml) by diluting the 100× stock provided by the manufacturer (NEB) tenfold with H_2O.

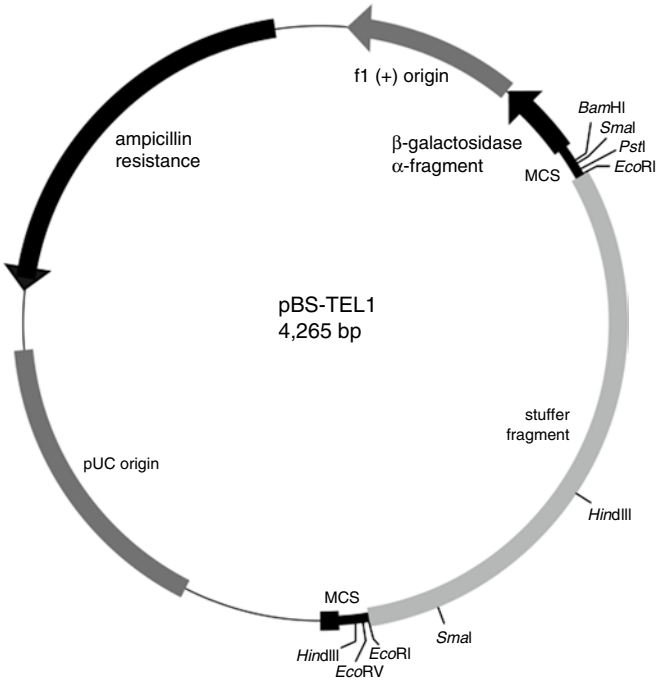

Fig. 1. Plasmid pTEL1. The pBLUESCRIPT KS II + (pBS) vector backbone and 2 kb stuffer fragment are labeled. Ap, ampicillin resistance gene; ORI, origin of replication; f1 ori, F1 origin of replication. Relevant restriction sites in the pBS polylinker and the stuffer fragment are shown.

3. Prime-A-Gene® DNA labeling kit (Promega Corp., Madison, WI).
4. End Repair: End-It™ kit (Epicentre® Biotech., Madison, WI).
5. Calf Intestinal Alkaline Phosphatase (CIAP, Promega).
6. LigaFast™ Rapid DNA Ligation System (Promega).

2.7. Escherichia coli *Transformation*

1. Commercially prepared electrocompetent cells: EPI300™ (Epicentre) or Ecloni® 10G (Lucigen Corp., Madison, WI) (see Note 2).
2. Electroporation cuvettes with 2 mm gap.
3. Luria-Bertani (LB) medium: One liter of medium contains 10 g tryptone, 5 g yeast extract and 10 g NaCl. For solid media, add agar (15 g/l). Sterilize by autoclaving at 121°C, 15 p.s.i. for 20 min.
4. Petri plates containing LB agar supplemented with 100 μg/ml ampicillin. Store plates for up to 1 month in the dark at 4°C.
5. Rattler Plating Beads: Zymo Research Corporation (Orange, CA).
6. Slide-A-Lyzer Mini Dialysis tubes (3,500 MWCO): Pierce (Rockford, IL).

2.8. Colony Hybridization

1. Whatman 541 paper (Whatman, Florham Park, NJ): cut into 82 mm circles, wrap in foil and sterilize by autoclaving.
2. Colony lysis buffer: 0.5 M NaOH. Store at room temperature.
3. Neutralization solution: 1 M Tris–HCl, pH 7.5. Store at room temperature.
4. 2× SSC: Add 100 ml of 20× SSC (see above) to 900 ml H_2O.
5. 95% EtOH.
6. Colony blot, prehybridization solution: 5× SSC, 0.1% SDS.

3. Methods

3.1. Preparation of High Molecular Weight Genomic DNA (see Note 3)

"Cut-off" pipette tips and slow pipetting should be used throughout the following procedures to minimize shearing of the DNA.

1. Place ~200 mg of freeze-dried mycelium in a 15 ml Falcon tube and use a glass rod to grind it against the side of the tube, forming a powder. Add 1.5 ml of lysis buffer that has been preheated to 65°C and mix gently using the glass rod. Place the cap on the tube and incubate in a 65°C water bath for 10 min. Add 1 ml of PCI and mix by gentle inversion. Screw the cap on tightly and return the tube to the water bath. Incubate for 30 min and gently invert the tube several times every 10 min to remix the PCI layer with the aqueous phase.
2. Pellet the cell debris by centrifuging at 3,000 × *g* for 30 min. Use a 1 ml pipette with a cut-off tip to recover 1 ml of supernatant and transfer it to a microfuge tube. Precipitate the DNA by adding 0.54 ml of room temperature isopropanol. If there is a large mass of DNA, spool it onto a glass rod (or sealed Pasteur pipette). Otherwise, pellet the DNA by centrifuging at 18,000 × *g* (see Note 4). Wash the pellet with 70% ethanol and air dry. Redissolve in 100 µl of T0.1E buffer.
3. Quantify the DNA solution using a fluorometer (see Note 5) and adjust DNA concentration to 100 ng/µl.
4. If the DNA solution contains too much polysaccharide, this could interfere with future manipulations. Therefore, if it is very milky in appearance, a differential precipitation procedure (17) should be used to reduce the level of polysaccharide contamination.

3.2. Restriction Digestion of Genomic DNA

1. Pipette 500 ng of genomic DNA (100 ng/µl) into a microcentrifuge tube. Add 5 µl of 10× restriction buffer and 5 µl of 10× BSA. Bring the volume up to 49 µl with sterile dH_2O. Then, add 1 µl (10–20 U) of restriction enzyme and mix well by gently flicking the tube. Incubate at 37°C (or other appropriate temperature) overnight.

2. Remove a 10 µl aliquot of the digest and pipette onto a small sheet of Parafilm. Add 2 µl of 6× loading dye solution.
3. Store the remaining digestion reaction at –20°C so that if electrophoresis shows the digestion to be not quite complete, additional enzyme can be added and the tube incubated for an additional overnight period.

3.3. Agarose Gel Electrophoresis

1. Prepare a 0.7% agarose gel solution in 0.5× TBE buffer and pour into a gel unit that is at least 20 cm long. Use gel combs with teeth that are ≤1 cm wide but which occupy a volume of at least 50 µl.
2. Place the gel in an electrophoresis unit filled with 0.5× TBE buffer. Load each restriction digestion reaction into a separate well. Include a DNA size ladder in at least one well. Run the gel overnight using a low voltage (e.g. 30 V, 20 h).
3. Stain the gel by placing it in ethidium bromide staining solution (made by adding 10 µl EtBr stock solution to 200 ml 0.5× TBE) and incubate for 30 min with gentle shaking.
4. Transfer the gel to a transilluminator and place a ruler alongside it. Switch on the UV lamp and take a photographic/digital image of the gel. Make sure that the gradations on the ruler are visible in the image.

3.4. Electroblotting

1. Presoak two electroblotter pads (supplied with the unit) in a large tray containing 0.5× TBE. Use gloved hands to press out air bubbles.
2. Place the cathode into the Genie blotter tray, followed by a plastic grid. Fill the tray to half full with 0.5× TBE.
3. Put a single electroblotter pad on top of the plastic grid and press again to expel air.
4. Cut two pieces of Whatman 3M paper and a single sheet of Pall Biodyne B hybridization membrane to the size of the agarose gel. Soak one piece of 3 M paper in 0.5× TBE and place it on top of the 3 M blotting pad.
5. Place the gel – open ends of the wells facing downward – on top of the 3 M sheet. Make sure that no air bubbles are trapped underneath the gel.
6. Wet the membrane in 0.5× TBE and place on top of the gel, being careful to avoid trapping air bubbles.
7. Wet the second sheet of 3 M paper in 0.5× TBE and place on top of the membrane, again being aware of air bubbles. Then, to expel any remaining bubbles, take a 20 mm glass test tube (or similar object), press down firmly and roll from one end the gel to the other.

8. Place the second blotting pad over the top sheet of 3 M paper and add sufficient 0.5× TBE to just cover it. Cover the pad with a second plastic grid, and then insert the anode, followed by the plexiglass top cover.
9. Slide the blotting setup into the blotter housing, stand the whole unit upright and, if necessary, top up with sufficient 0.5× TBE to submerge the gel fully.
10. Apply a voltage of 12 V with constant current of 10 A for 2 h.
11. After the transfer is complete, disassemble the unit and, using a pencil, mark the positions of the wells on the membrane. Cut a notch out of the bottom left of the membrane (which corresponds to the bottom right of the gel, which currently is upside-down). Then, peel the membrane off the gel and place it on a paper towel to wick off excess moisture.
12. Denature the immobilized DNA by floating the membrane on a solution of 0.5 M NaOH for 10 min.
13. Rinse the membrane with 2× SSC and then soak it in fresh 2× SSC for 10 min.
14. Blot the membrane dry with paper towels and then use a cross-linker to fix the DNA. Label the membrane on the top right-hand corner.

3.5. Preparation of Telomere Probe (See Note 6)

1. Pipette the following into a 200 µl thin walled PCR tube: 5 µl 10× ExTaq buffer, 4 µl dNTP mix (2.5 nM each nucleotide), 20 pmol TEL1 primer, 20 pmol TEL2 primer, PCR grade water to 49.8 and 0.2 µl ExTaq enzyme.
2. Mix well by flicking and centrifuge the tube briefly to collect the reagent mix at the bottom.
3. Place in the polymerase chain reaction machine and run the following program: 94°C for 1 min, followed by 35 cycles of 94°, 30 s; 55°C, 30 s; 72°C, 2 min. A final extension step at 72°C for 5 min is provided to complete the synthesis of any incompletely extended molecules.
4. Add 10 µl of 6× loading dye solution, mix by pipetting and load into three wells of a 0.7% agarose mini-gel (10 cm × 7 cm). Load a kilobase plus size marker into an adjacent well, and then run the gel at 20 V for 9.5 h.
5. Stain the gel in an aqueous ethidium bromide solution for 30 min. Then, visualize the DNA on a transilluminator emitting long wavelength UV light (312 nm). The PCR products will appear as a smear. Use a scalpel to excise DNA in the size range from 1.6 to 2 kb and recover the DNA using a commercial gel extraction kit.
6. Quantify the telomere probe using a spectrophotometer.

7. Pipette ~50 ng of telomere probe solution into a microcentrifuge tube and adjust the volume to 16 µl with H_2O. Denature the DNA by placing in a boiling water bath for 10 min. Then, remove the tube from the waterbath and snap cool on ice.
8. To the denatured DNA, add 5 µl of 5× Prime-A-Gene labeling buffer, 1 µl of dNTP mix containing dATP, dGTP and dTTP, 1 µl of BSA (100 µg/µl), and 1 µl of Klenow polymerase. Mix the reagents by flicking the tube and briefly spin in the microcentrifuge. Then, add 2 µl of ^{32}P-dCTP (20 µCi; 3,000 Ci/mmol) and incubate at room temperature for 4 h to overnight.
9. Stop the labeling reaction by adding 75 µl of dye stop solution and mix by pipetting up and down.
10. Use a MicroSpin™ G50 column to remove unincorporated nucleotides: First, give the column a "prespin" to compact the sieving matrix. Remove the filter column and discard the buffer that collects in the collection tube. Then, return the filter column to the collection tube and then pipette the labeling reaction on to the top of the sieving matrix. Spin for 15 s at full speed. Retain the collection tube and its contents and discard the spin column in a suitable radioactive waste container.
11. Check ^{32}P incorporation by pulling 1 µl of the column flow up into a pipette tip and holding the tip to a Geiger counter. An adequately labeled probe should emit ≥10,000 counts/min. Return the test sample to the collection tube.

3.6. Detection of Telomeric Restriction Fragments by Hybridization

1. Place the membrane containing the immobilized DNA in a hybridization bottle with the DNA side facing the inside of the tube. Add 20 ml of hybridization buffer and place in the hybridization oven. Incubate with rolling at 65°C for 10 min.
2. While the membrane is prehybridizing, transfer 50 µl of ^{32}P-labeled telomere probe into a fresh microcentrifuge tube and denature it by adding 5 µl of a freshly prepared 2 N NaOH solution. Incubate for 8 min at room temperature.
3. After the denaturation step, neutralize the probe solution by adding 5 µl of 1 M Tris–HCl, pH 7.4.
4. Decant off the buffer used for prehybridization and replace it with 5 ml of fresh hybridization solution. Add the probe directly to the hybridization buffer. Be careful not to let any probe touch the membrane directly.
5. Replace the hybridization bottle in the oven and incubate with rolling for 16–24 h.

6. Decant hybridization buffer into a container approved for the disposal of ^{32}P-dCTP waste. Rinse the membrane by pouring 50 ml 2× SSC into the hybridization tube and returning it to the chamber for 5 min.
7. Decant the rinse solution into the radioactive waste container and replace it with 50 ml low stringency wash buffer. Return the tube to the chamber and incubate for 20 min.
8. Decant the supernatant into the waste container and replace with 50 ml of high stringency wash buffer. Return the tube to the chamber and incubate for another 20 min.
9. Using forceps, remove the membrane from the hybridization tube and place on a paper towel to wick off excess liquid. *Do not allow the membrane to dry out.*
10. Place the membrane, right-(DNA-side)side up, on a sheet of plastic wrap and cover with a second layer of wrap. Expose to an autoradiographic film or phosphorimage screen overnight.
11. Develop the photographic film or scan the phosphorimage.
12. If necessary, print the phosphorimage at a scale of 1:1. Measure the positions of each telomere-hybridizing band relative to the well. Then, determine the molecular sizes of each band by cross-referencing with the size marker in the image of the original ethidium bromide stained gel (this is where the image of the ruler alongside the gel comes in handy). Alternatively, one can juxtapose the gel image of the size marker with the image of the blot (as shown in Fig. 2).
13. If desired, the membrane can be reused after stripping off the telomere probe. Probe removal is performed by performing two 30 min washes in 0.4 N NaOH at 50°C. The membrane is then washed with 2× SSC and allowed to dry before storage.

3.7. End Repair of Genomic DNA for Telomere Cloning

"Cut-off" pipette tips and slow pipetting should be used throughout the following procedures to minimize shearing of the DNA.

1. Place in a microcentrifuge tube 2–4 µg of genomic DNA in a total volume of ≤34 µl of T0.1E. Add 5 µl of End-It buffer, 5 µl of nucleotide mix, 5 µl of 10 mM ATP, and then 1 µl of the T4 polynucleotide kinase/T4 polymerase enzyme mix. Allow the reaction to proceed for 45 min at room temperature.
2. Inactivate the enzymes by heating at 70°C for 10 min.
3. The following PCI/CI extraction steps are critical. Add 50 µl of T0.1E followed by 50 µl of PCI. Vortex briefly and centrifuge at 18,000×*g* for 2 min. Recover the aqueous phase (top layer) and transfer to a fresh microcentrifuge tube. Add 50 µl of PCI to this tube and repeat the process (do not add any more T0.1E).

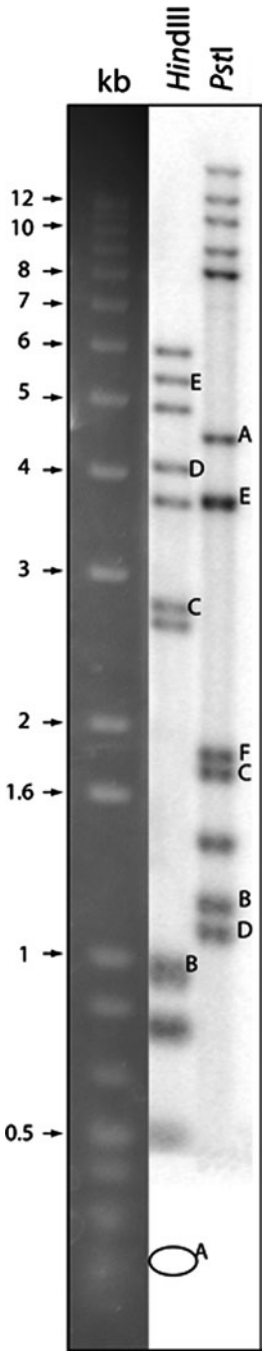

Fig. 2. Identification of telomeric restriction fragments in *Epichloë festucae* isolate E2368. Genomic DNA samples from *E. festucae* were digested separately with *Hin*dIII and *Pst*I and fractionated by agarose gel electrophoresis alongside a lane containing a 1 kb plus DNA ladder. The gel was then stained with ethidium bromide, imaged and electroblotted to a nylon membrane. After hybridization with the telomere probe, the membrane was exposed to a phosphorimage screen. The figure shows the resulting phosphorimage adjacent to an image of the gel lane that contained the DNA ladder. Counting doubly intense signals as two telomeres, 14 telomeric fragments are visible in the *Pst*I digest and 12 with *Hin*dIII (two fragments ran off the gel, one – a 0.2 kb fragment was present in the genome sequence and is represented with an *oval*). Fragments that were present in the *E. festucae* genome assembly are labeled with letters. The unlabeled fragments in the *Hin*dIII lane were all targets for cloning. Molecular sizes are listed on the *left.*

4. Recover the aqueous phase from the second PCI extraction and then add 50 µl of CI. Vortex briefly and centrifuge at 18,000 × *g* for 2 min. Recover the aqueous phase and add 0.1× vol. 3 M Na acetate, pH 5.2 and 2 vol. room temperature 100% EtOH. Mix gently and precipitate the DNA by centrifugation (18,000 × *g*, 10 min).

3.8. Restriction Digestion of End-Repaired Genomic DNA

1. Add 24 µl of 1× restriction buffer (+100 µg/ml BSA) to the pellet of end-repaired DNA (*see* Subheading 3.7) and leave on the bench for 30 min to dissolve. Gently flick the tube to disperse the solution and then add 20 U of restriction enzyme. Incubate at 37°C (or other appropriate temperature) overnight. Add 5 µl of 6× loading dye solution.
2. Prepare a 0.7% agarose gel solution in 0.5× TBE buffer and pour into a gel unit that is at least 20 cm long. Use gel combs with teeth that are ≤1 cm width but which occupy a volume of at least 50 µl.
3. Place the gel in the electrophoresis unit and submerge with 0.5× TBE buffer. Load each restriction digest(s) into a single well. Load a DNA size ladder into at least one well in the gel.
4. Run the gel overnight using a low voltage (e.g. 30 V, 20 h). Place the gel in ethidium bromide solution for 30 min.
5. UV light causes DNA damage which, in turn, prevents the recovery of subclones. Therefore, *do not photograph the gel before cutting out the DNA fractions.* Also, while cutting out bands, make sure that the DNA is exposed only to LONG wavelength (i.e. low energy) UV light (312 nm).
6. Use a scalpel to cut out gel slices containing DNA fragments of the desired sizes (see Fig. 3a). A good rule of thumb is to recover fragments ±0.2 kb for fragments up to ~4 kb, ±0.5 kb for fragments of 4–8 kb and then ±1 kb for larger ones (telomeres of similar sizes can be excised in a single gel slice and resolved after they have been cloned and characterized). Extract the DNA from the gel slices using a commercial kit (e.g. the QiaQuick extraction kit, Qiagen, Valencia, CA). Place the gel in fresh water and put in a refrigerator until it has been confirmed that the correct-sized fragments were excised.
7. Check that the desired telomeres were successfully recovered by running the purified fractions on an agarose gel, alongside a lane of total genomic DNA digested with the same enzyme that was used to produce the fragments. Blot the gel to a membrane and then incubate with the telomere probe (see Subheadings 3.3–3.6). Successful recovery of the desired telomeres is revealed by strong hybridization signals in the lanes containing the different size fractions (see Fig. 3b and c).

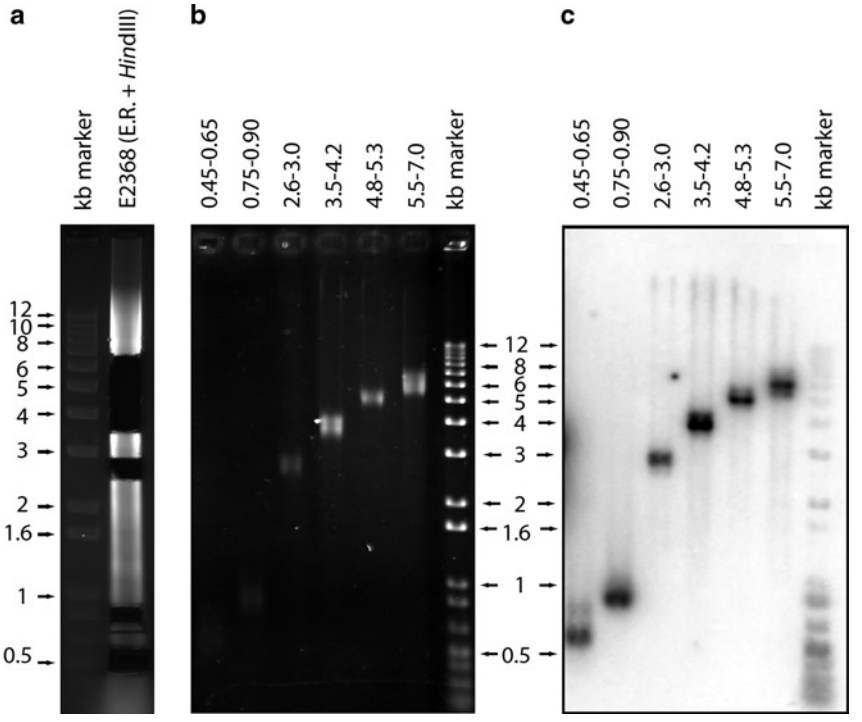

Fig. 3. Purification and verification of DNA fractions containing target telomeres. Approximately 2 µg of E2368 DNA was end-repaired and subsequently digested with *Hin*dIII. After fractionation by electrophoresis, the desired DNA fragments were excised from the gel (**a**). The fragments were extracted with a Qiagen kit and were then fractionated on a mini-gel to check DNA (and telomere) recovery (**b**). The mini-gel was electroblotted to a membrane and the immobilized DNAs were then hybridized with the telomere probe. The resulting phosphorimage is shown in (**c**).

3.9. Preparation of Linearized Plasmid Vector

1. Pipette ~2 µg of plasmid DNA (100 ng/µl) into a microcentrifuge tube. To this add 10 µl of 10× reaction buffer, 10 µl of 100 µg/ml BSA and 59 µl of H_2O. Finally, add 0.5 µl (10 U) of each enzyme (see Note 7), mix well and centrifuge briefly to collect the reaction mix at the bottom of the tube. Incubate the tube at 37°C for 2 h.
2. Add 10 µl of 10× CIAP buffer followed by 0.1 U of CIAP. Mix well and then incubate at 37°C for 1 h. Add another 0.1 U of CIAP and incubate this time at 50°C.
3. Stop the reactions by adding 10 µl of 100 mM EDTA, pH 8 and inactivate the enzymes by heating at 70°C for 10 min.
4. *This step is critical*: Extract the reaction two times with PCI followed by one extraction with CI (follow the extraction method described in Subheading 3.7 but omit the addition of T0.1E).
5. Adjust the Na^+ salt concentration of the reaction mix to 100 mM with 5 M NaCl. Then, precipitate the DNA by adding 200 µl of room temperature 100% EtOH. Centrifuge immediately for 10 min at 18,000×*g*.

6. Discard the supernatant and then rinse the DNA pellet with 500 µl of 70% EtOH. Remove the 70% EtOH and allow the pellet to air dry.
7. Dissolve the pellet in 25 µl of T0.1E buffer and add 5 µl of 6× loading dye solution.
8. Prepare an agarose gel (at least 15–20 cm long and made with 0.5× TBE).
9. Immerse the gel in running buffer in the electrophoresis unit. Load the DNA + dye solution into a well that is at least 2 cm in width. Apply a voltage of 30 V for 20 h.
10. Stain the gel by immersing in an ethidium bromide solution for 20 min.
11. *Do not photograph the gel before cutting out the band of interest.* Visualize the gel under LONG wavelength UV light (312 nm) and use a scalpel to cut out a gel slice containing the band that corresponds to the completely cut vector.
12. Extract the vector DNA from the gel slice using a commercial kit and quantify by spectrophotometry.

3.10. Ligation of Telomeric Fragments to the Linearized Vector

1. For the negative control: Pipette the following into a microfuge tube: 1 µl of plasmid DNA (~50 ng), 2 µl of H_2O, 3.5 µl of 2× ligation buffer and 0.5 µl T4 DNA ligase. Mix by gently flicking the tube and then incubate overnight at 12°C.
2. For the standard reaction: Pipette the following into a microfuge tube: 1 µl of plasmid DNA, 2 µl of size-fractionated genomic DNA, 3.5 µl of 2× ligation buffer and 0.5 µl T4 DNA ligase. Mix by gently flicking the tube and then incubate overnight at 12°C.
3. Inactivate the ligase by heating at 70°C for 10 min, and then add 13 µl of T0.1E buffer to dilute the salt in the ligation buffer.

3.11. Electrotransformation of E. coli (See Note 2)

1. Prechill the microcentrifuge tubes (one per ligation) and electroporation cuvettes (one per ligation) on ice.
2. Set the electroporator to 2.5 kV with a resistance of 200 Ω and a capacitance of 100 µF.
3. Quickly thaw the competent cells by rolling the tube between your fingers. As soon as the suspension begins to thaw, transfer as many 10 µl aliquots as are needed to the chilled tubes. Remaining cells can be refrozen for later use simply by returning them to the –80°C freezer.
4. Add 1 µl of the diluted ligation reaction to the thawed competent cell suspension. Mix thoroughly by pipetting and transfer the entire mix to an electroporation cuvette. Place the cuvette in the chamber and apply a single shock. *As soon*

as possible, add 400 µl of recovery medium (supplied with the competent cells) and incubate at 37°C for 30 min.

5. While the cells are recovering from the electroshock treatment, place four to five sterile glass beads into the Petri dishes that contain LB + ampicillin selection medium (two plates per transformation). Then, place the plates in a laminar flow hood with their lids off to allow any surface moisture to evaporate.
6. Pipette 200 µl of transformation mix into each Petri dish, cover with the lids and then rock the plates back and forth to distribute the mix evenly across the agar surface. Place the dishes on the bench until all moisture is absorbed into the agar surface, dump the glass beads into a waste receptacle and then place the plates upside down in a 37°C incubator. Allow to incubate for 24 h.
7. Count the number of colonies obtained for each transformation. The negative control should produce fewer than ten colonies (usually, we obtain just one or two). In contrast, the "vector plus insert" ligation should yield a total of >500. In this case, screening usually results in the recovery of at least one telomere-containing clone. Conversely, it is usually not worth screening for telomeres unless at least 200 colonies are recovered.

3.12. Dialysis of Ligation Reactions

1. If the transformation frequency is too low (see above), it can be increased up to tenfold by using dialysis to remove salts from the ligation reaction. This can be done very conveniently through the use of microdialysis chambers.
2. Fill a glass beaker with 500 ml T0.1E buffer and drop in a stir bar. Prewet the dialysis membrane by floating the empty chamber in the T0.1E for 10 min with stirring. Be sure to keep the level of the membrane barely below the surface of the buffer to prevent hydrostatic pressure from forcing too much liquid into the chamber.
3. Pipette the ligation mix into the chamber and, again, maintain the chamber's buoyancy. Incubate with stirring for 2 h.
4. Collect the dialyzed DNA solution with a pipette and transfer into a fresh microfuge tube. It is ready for use immediately.

3.13. Colony Lifts (See Note 8)

1. Use a permanent marker to place orientation marks on Whatman 541 paper disks, as shown in Fig. 4a. In addition, give each disk a label that corresponds to the Petri plate containing the colonies that are to be lifted.
2. With the label side up, use gloved hands to bend a paper disk upward into a U-shape. Touch the bent surface to the middle of the agar and, working outward from the initial point of contact, immediately smooth the sides of the disk evenly over the agar surface. Gently poke any bubbles to ensure good paper-agar contact across the whole plate.

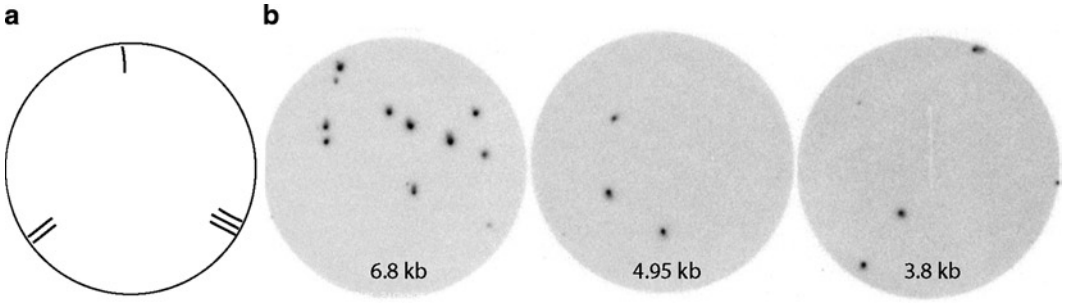

Fig. 4. Colony hybridization. (**a**) Marking pattern used to record the orientations of Whatman 541 paper disks. (**b**) Example of colony hybridization results obtained for three different *E. festucae* telomeres (the figure shows only one out of the two disks that were used for each telomere). The sizes of the telomere fragments that were cloned are shown on the respective images.

3. *Before removing the disk*, use a permanent marker to mark the bottom of the Petri dish with lines that correspond with the orientation marks on the disk. This facilitates the eventual alignment of hybridization spots with specific colonies on the agar surface.
4. Use forceps to grasp the edge of the disk. Lift it off the agar surface and place colony-side-up on a paper towel. This should have resulted in transfer of the colonies onto the disk, with only faint spots remaining on the agar. The disk can be processed immediately, or left overnight to air dry.
5. Place the agar plates in a safe location on the bench and allow the colonies to regrow overnight (do not incubate at 37°C, or the colonies will grow too large).
6. Prepare a positive control membrane by spotting 1 µl of a 100-fold dilution of a plasmid containing a known telomere. If a positive control plasmid is not available, dilute the (unlabeled) telomere probe 100-fold with T0.1E and spot 1 µl on a small square (~1.5 cm × 1.5 cm) of hybridization membrane. Allow the membrane to dry and then process along with the Whatman paper disks.
7. Pipette 1 ml aliquots of 0.5 M NaOH onto the surface of a clean plexiglass sheet. Place each disk colony-side-up onto an NaOH "puddle." Allow the solution to soak across the whole disk and incubate for 10 min.
8. Transfer the disks to a tray containing 1 M Tris–HCl, pH 7.5 and incubate with shaking for 10 min.
9. Transfer the disks to a tray containing 2× SSC and incubate with shaking for 10 min.
10. Transfer the disks to a tray containing 95% EtOH and incubate with shaking for 10 min.
11. Remove the paper disks from the EtOH and allow to air dry completely.

3.14. Colony Hybridization

1. Place the paper disks and the positive control membrane in a hybridization tube – colony side facing inward. Add colony blot prehybridization solution (~5 ml/disk) and incubate, with rolling, in the hybridization chamber at 65°C for 1 h.
2. Remove the prehybridization solution and replace with a total of 5 ml hybridization buffer. Incubate with rolling for 10 min at 65°C.
3. Pipette 50 µl of denatured telomere probe (~15,000 counts/µl) directly into the hybridization solution, making sure that it does not touch the disks directly. Incubate with rolling for 2–18 h at 65°C.
4. Decant the hybridization buffer into a container approved for the storage/disposal of radioactive waste. Rinse the disks by pouring 50 ml of 2× SSC into the tube and returning it to the chamber for 5 min.
5. Decant rinse solution into approved container and replace with 50 ml 2× SSC. Return tube to the chamber and incubate for 20 min.
6. Decant supernatant into approved container. Using forceps, remove disks from the hybridization tube and place on a paper towel to wick off excess liquid. *Do not allow the disks to dry out.*
7. Place the control membrane and the Whatman paper disks, labels-side-up, on a sheet of plastic wrap and cover with a second layer of wrap. If the disks are sufficiently moist, they should remain firmly in place (this is important for subsequent steps). If they do not stay in place, fix them to the bottom sheet with tape.
8. Expose the disks to autoradiographic film or phosphorimage screens (label-side-up) for 2 h to overnight.
9. Develop the film or scan the phosphorimage. Allow sufficient exposure that the outlines of the paper disks are clearly visible. If telomere-containing clones are present, they will produce spots of very strong hybridization (see Fig. 4b). In this case, *do not remove disks from plastic wrap after exposure.* If no signals are present on the paper disks but the hybridization did work, as indicated by a strong signal on the control membrane, then there will be no telomere-containing clones on the plates (see Note 9).

3.15. Identification of Colonies Containing Telomeric Clones (See Note 10)

1. If the marks on the paper disks faded during hybridization, use the permanent marker to re-mark their positions on the plastic wrap (do not remove the disks).
2. If phosphorimaging was used, print the digital image to paper (or a clear plastic sheet) at a scale of 1:1.
3. Lay the autoradiographic film or printed image over the plastic wrap and use the outlines of the paper disks to register the

photographic/digital disk images on top of the disks (use of a light box allows the outlines of the paper disks and the orientation marks to be seen through plain white paper). Copy the positions of the orientation markers onto the disk images with positive hybridization spots.

4. Once the orientation marks have been copied onto the images, place the corresponding Petri dish on top of the disk image and align the orientation marks. If the disks were accidentally removed from the Petri dishes before the positions of the orientation marks were copied onto them, it will be necessary to stain the paper disks to expose the lysed colonies (see Note 11).
5. Identify the *E. coli* colonies that gave rise to hybridization signals and pick them to liquid LB medium supplemented with 50 μg/ml ampicillin.
6. Grow cultures overnight with shaking at 37°C.
7. Prepare DNA using a commercial plasmid extraction kit and sequence to confirm that a bona fide telomere has been cloned.

3.16. Second Round Screening

If there are several colonies in the region that gave rise to the hybridization signal, it may be necessary to perform second round screening. This can be accomplished in two ways.

1. If the colonies in the region are separated from one another, these can be individually picked (using a dissecting scope, if necessary) and spotted on a fresh LB + ampicillin plate. After overnight growth at 37°C, these can be screened using the colony lift procedure described above.
2. If there are too many colonies in the region of hybridization, it will be difficult, or impossible, to identify the correct one. In this case, a scalpel blade can be used to cut out an agar plug that contains the relevant colonies. Place the plug in a microcentrifuge tube containing 500 μl LB medium, cap the tube and vortex vigorously for a few seconds. Perform serial dilutions of the resulting suspension, plate the suspensions on fresh LB + ampicillin plates and incubate overnight at 37°C. Select plates containing 100–500 colonies and perform colony lifts to screen for telomere-containing clones.

4. Notes

1. The success of the directional cloning procedure is highly dependent on the quality of the linearized plasmid preparation. The vast majority of fragments in the digested genomic DNA sample have two sticky ends, and it is essential that the vector precludes the cloning of such fragments while allowing

efficient cloning of the telomeric fragments that have one blunt and one sticky end. In addition, because it is necessary to screen for true telomeres among large numbers of clones derived from false ends generated by DNA breakage, it is important that only recombinant plasmids are recovered following transformation of *E. coli*. For this reason, the vector should not be capable of self-ligation.

It is very difficult to generate satisfactory linearized plasmid preparations from standard cloning vectors due to an inability to obtain complete cleavage of restriction sites. To address this issue, we utilize a recombinant pBluescript plasmid that contains a 2 kb stuffer fragment inserted at the *Eco*RI restriction site. To prepare vector samples for cloning telomeres, pBS-TEL1 is cut with either *Sma*I or *Eco*RV and with a second enzyme on the other side of the stuffer fragment. This way, a preparative electrophoretic gel can be used to separate the fully digested plasmid from uncut and singly cut molecules, as well as from the stuffer fragment.

2. Successful recovery of the rare telomeric restriction fragments is highly dependent on obtaining high transformation efficiencies ($>10^9$ colony forming units μg^{-1} DNA). For this reason, the use of commercially prepared competent cells is strongly recommended.
3. In an ideal world, only true chromosome ends would be recovered. However, the breakage of molecules during DNA isolation produces "false" ends that are also amenable to cloning by the methods described here. It follows that the more intact the starting DNA, the greater the proportion of clones that contain true telomeres. For this reason, it is important to avoid DNA isolation methods that are overly disruptive, such as the use of bead beaters and excessive pipetting. With careful execution, the following method can produce DNA with an average size of ~200 kb. For a fungus, with a genome size of ~40 Mb and eight chromosomes, this would mean that ~1 in 25 DNA ends should be true telomeres.
4. It is not necessary to use cold temperatures or extended incubation times to obtain essentially quantitative recovery of DNA following precipitation. In fact, the use of cold temperatures simply increases the recovery of undesirable contaminants, such as polysaccharides. Therefore, all precipitations should be performed using room temperature ethanol/isopropanol and centrifugation should be performed immediately after the alcohol has been mixed in.
5. Genomic DNA should be quantified using a fluorometer because the polysaccharides that are usually copurified cause the DNA concentration to be vastly overestimated.

6. Many telomere hybridization studies are performed using end-labeled oligonucleotides as probes. These produce very weak hybridization signals because there is only one labeled nucleotide per probe molecule (~5% of residues). The method for probe production described here generates probes in which 25% of residues are labeled. Consequently, the hybridization signals are at least five times stronger.
7. Depending on the enzymes used for telomere cloning, it may be necessary to perform separate digestion reactions on the plasmid DNA, due to the need for different reaction buffers. In such cases, the DNA should be precipitated between digests.
8. Successful colony lifts require that the agar plates contain just the right amount of moisture. To achieve the desired moisture content, it is important to leave the lids off the Petri plates while the molten agar is setting. Keep the lids off for ~30 min while the plates dry in a laminar flow hood. Colonies should be lifted immediately after removing the plates from the incubator.
9. Under normal circumstances, screening of 500–1,000 colonies should result in the recovery of at least one telomere-containing clone. Failure to do so indicates a problem with the cloning procedure. The most likely cause of failure is inefficient end-repair of the genomic DNA due to the presence of polysaccharides or other impurities. Use of a differential precipitation procedure should help to address these issues (17). However, if the fungus under study frequently yields DNA that is recalcitrant to restriction digestion regardless of what types of "clean-up" methods are used, then it may be necessary to use a DNA isolation kit that uses a column-binding procedure (e.g. Illustra™ Tissue and Cells GenomicPrep Midi Flow, GE Healthcare). Alternatively, growing the fungus under different cultural conditions may reduce the inhibitors to acceptable levels.
10. The method described for aligning hybridization images with the original paper disks requires at least two disks to be imaged on the same film/screen; or one disk and a control membrane. Otherwise, it is difficult to orient the image correctly.
11. If one forgets to mark the bottom of an agar plate before lifting off the Whatman paper disks, this makes it more difficult to match up a positive hybridization spot with a specific colony on the agar plate. In such instances, follow Subheading 3.15 up to the end of step 3. Once the digital image has been labeled with the disk orientation marks, remove the disks from the plastic and incubate them for 2 h

in a dilute solution of gel loading dye (add 100 µl 6× dye to 100 ml TE). The debris from the immobilized colonies stain a faint blue color. The hybridization image can then be placed over the stained disks to identify the specific colony(ies) that produced the hybridization signal(s). The desired clone can then be found on the original agar plate by visual comparison of colony positions.

Acknowledgments

I would like to acknowledge the technical support of David Thornbury. The development of these methods was supported by grants MCB 0653930 and EF 0523661 from the National Science Foundation.

References

1. Bass, H.W., Reiera-Lizarazu, O., Ananiev, E.V., Bordoli, S.J., Rines, H.W., Phillips, R.L., Sedat, J.W., Agard, D.A. and Cande, W.Z. (2000) Evidence for the coincident initiation of homolog pairing and synapsis during the telomere-clustering stage of meiotic prophase. *J. Cell Sci.*, **113**, 1033–1042.
2. Niwa, O., Shimanuki, M. and Miki, F. (2000) Telomere-led bouquet formation facilitates homologous chromosome pairing and restricts ectopic recombination in fission yeast. *EMBO J.*, **19**, 3831–3840.
3. Tomita, K. and Cooper, J.P. (2007) The telomere bouquet controls the meiotic spindle. *Cell*, **130**, 113–126.
4. Chikashige, Y., Ding, D.Q., Funabiki, H., Haraguchi, T., Mashiko, S., Yanagida, M. and Hiraoka, Y. (1994) Telomere-led pre-meiotic chromosome movement in fission yeast. *Science*, **264**, 270–273.
5. Farman, M.L. and Kim, Y.-S. (2005) Telomere hypervariability in *Magnaporthe oryzae*. *Mol. Plant Pathol*, **6**, 287–298.
6. Horowitz, H., Thorburn, P. and Haber, J.E. (1985) Rearrangements of highly polymorphic regions near telomeres of *Saccharomyces cerevisiae*. *Mol. Cell. Biol.*, **4**, 2509–2517.
7. Wada, M. and Nakamura, Y. (1996) Antigenic variation by telomeric recombination of major-surface-glycoprotein genes of *Pneumocystis carinii*. *J. Euk. Microbiol.*, **43**, 8S.
8. Cuomo, C.A., Guldener, U., Xu, J.R., Trail, F., Turgeon, B.G., Di Pietro, A., Walton, J.D., Ma, L.J., Baker, S.E., Rep, M. *et al.* (2007) The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science*, **317**, 1400–1402.
9. Kasuga, T., Mannhaupt, G. and Glass, N.L. (2009) Relationship between phylogenetic distribution and genomic features in *Neurospora crassa*. *PLoS One*, **4**, e5286.
10. Barry, J.D., Ginger, M.L., Burton, P. and McCulloch, R. (2003) Why are parasite contingency genes often associated with telomeres? *Int. J. Parasitol.*, **33**, 29–45.
11. Wada, M. and Nakamura, Y. (1996) Unique telomeric expression site of major-surface-glycoprotein genes of *Pneumocystis carinii*. *DNA Research*, **3**, 55–64.
12. Schaffzin, J.K. and Stringer, J.R. (2004) Expression of the Pneumocystis carinii major surface glycoprotein epitope is correlated with linkage of the cognate gene to the upstream conserved sequence locus. *Microbiology*, **150**, 677–686.
13. Li, W., Rehmeyer, C.J., Staben, C. and Farman, M.L. (2005) TERMINUS: Telomeric-end read mining in unassembled sequences. *Bioinformatics*, **21**, 1695–1698.
14. Rehmeyer, C., Li, W., Kusaba, M., Kim, Y.-S., Brown, D., Staben, C., Dean, R. and Farman, M.

(2006) Organization of chromosome ends in the rice blast fungus, *Magnaporthe oryzae*. *Nucleic Acids Res.*, **34**, 4685–4701.

15. Wu, C., Kim, Y.-S., Smith, K.M., Li, W., Hood, H.M., C., S., Selker, E.U., Sachs, M. and Farman, M.L. (2009) Characterization of chromosome ends in the filamentous fungus *Neurospora crassa*. *Genetics*, **181**, 1129–1145.
16. Henderson, E.R. and Blackburn, E.H. (1989) An overhanging 3' terminus is a conserved feature of telomeres. *Mol. Cell. Biol.*, **9**, 345–348.
17. Michaels, S.D., John, M.J. and Amasion, R.M. (1994) Removal of polysaccharides from plant DNA by ethanol precipitation. *Biotechniques*, **17**, 274–276.

Chapter 3

Identification and Annotation of Repetitive Sequences in Fungal Genomes

Braham Dhillon and Stephen B. Goodwin

Abstract

Advances in sequencing technologies have fundamentally changed the pace of genome sequencing projects and have contributed to the ever-increasing volume of genomic data. This has been paralleled by an increase in computational power and resources to process and translate raw sequence data into meaningful information. In addition to protein coding regions, an integral part of all the genomes studied so far has been the presence of repetitive sequences. Previously considered as "junk," numerous studies have implicated repetitive sequences in important biological and structural roles in the genome. Therefore, the identification and characterization of these repetitive sequences has become an indispensable part of genome sequencing projects. Numerous similarity-based and de novo methods have been developed to search for and annotate repeats in the genome, many of which have been discussed in this chapter.

Key words: Repetitive sequence identification, Transposable elements, Repeat annotation, Similarity-based methods, De novo methods, *k*-mer methods

1. Introduction

Recent years have seen an exponential increase in the number of genomes that have been sequenced. This has been facilitated by two factors, the rate at which DNA can be sequenced and the reduced cost of sequencing. Over the course of a decade, with the improvement in sequencing technologies, the cost to sequence a 1-kb region has decreased from US $1 to ~7 cents (1). With the development of next-generation sequencing methods, we are getting closer to the Holy Grail of sequencing, the thousand-dollar genome (2). Advances in sequencing technologies have been closely rivaled by improvements in computational biology. Faster and efficient computing methodologies and infrastructure have paved the way for scientists to ask better and more complex biological questions.

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722,
DOI 10.1007/978-1-61779-040-9_3,

It was observed fairly early on that genome size is not correlated to the phenotypic complexity of the organism. This is known as the C-value paradox (3). A further validation for C-value paradox comes from the plethora of sequenced eukaryotic genomes, which reveal that although genome size may vary by at least five orders of magnitude (4), this discrepancy in genome size is not due to a variation in gene numbers (5). One factor that contributes to this disparity in genome size is an increase in length and number of repetitive sequences, especially transposable elements (TEs) (4).

Repetitive sequences can be described simply as any DNA sequence that occurs in two more copies in the genome. In one form or another, repetitive sequences are a common feature of all the genomes studied so far. They can be further classified into five broad categories (Fig. 1): interspersed repeat; tandem repeats, segmental duplications (SD), multicopy gene families, and pseudogenes. These categories are defined by several criteria, including distribution in the genome (interspersed versus tandem repeats), presence or absence of protein domains (tandem repeats versus gene families, TEs), and type of protein domain (gene families versus TEs).

1. Interspersed repeats/mobile genetic elements/TEs are DNA sequences that can move about in the genome, often with an increase in copy number.
2. Tandem repeats can be broadly classified into two categories, based on the length of the repeat unit. Microsatellites consist of very short sequences of 1–9 bp in length, repeating multiple times at a locus. These are dispersed throughout the chromosomes but often are found in and around genes (6).

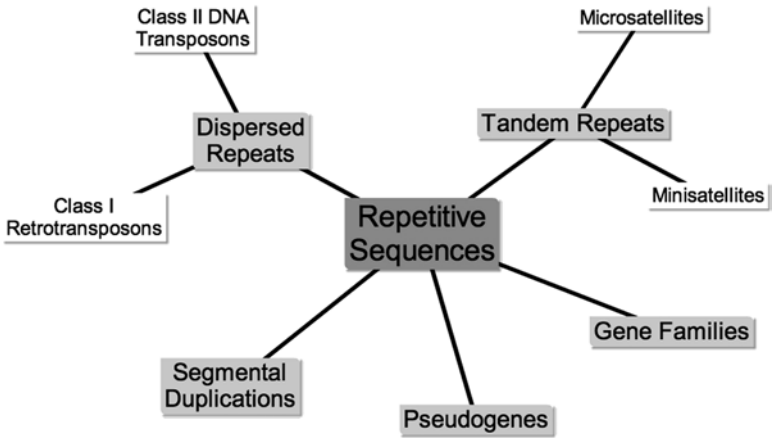

Fig. 1. *Classification of repetitive sequences.* A mind-mapping diagram showing the five main classes and additional subclasses of repetitive sequences.

Minisatellites have repeat units that vary from 10 to 150 bp in length. These repeat units can be present 2–100 times at a given locus.

3. SD are defined as duplicated genomic regions that are longer than 1 kb and share ≥90% sequence identity (7). SDs account for ~5.2% of the human genome (8) and are widespread in the primate lineage (9). In yeasts, spontaneous generation of SDs occurs at a frequency of 10^{-9}/cell/division (10).
4. Duplicated genes are present in all the organisms studied so far (11). Members of a gene family are often referred to as paralogs. Gene family size varies among species and gene families (12). The biggest family in *Drosophila melanogaster* has 111 members (13), whereas in mammals, the biggest family has ~1,000 members (14). In fungi, one of the largest gene families reported to date is the protein kinase family-2 from *Laccaria biclor*, which has 150 members (15).
5. One of the by-products of gene duplication is the generation of pseudogenes. A pseudogene can be described as a DNA sequence that has been derived from a functional gene, but has been rendered nonfunctional by mutations.

In this chapter, we focus mainly on the identification and annotation of TEs. In the latter part of the chapter, we briefly mention the remaining categories of repetitive sequences.

2. Classification of TEs and Their Structural Features

TEs can be defined as DNA fragments that can move into new locations in the host genome by excision or replication of an existing copy. Based on their mode of transposition, TEs may be further classified as:

1. Class I TEs or retrotransposons (copy and paste).
   Class I elements transpose via an RNA intermediate. They encode for a reverse transcriptase (RT) protein that acts on the RNA transcript to generate a double-stranded cDNA molecule, which is then inserted back into the genome (16). Therefore, after each replication cycle, a new copy of the retroelement is produced. This "copy–paste" mechanism allows for an exponential increase of retroelements in the genome. Copy numbers of Class I TEs often are very high. Retrotransposons are either flanked by long terminal repeats (LTR retrotransposons) or they have a poly-adenylated sequence at their 3′ end (non-LTR retrotransposons) (Fig. 2).

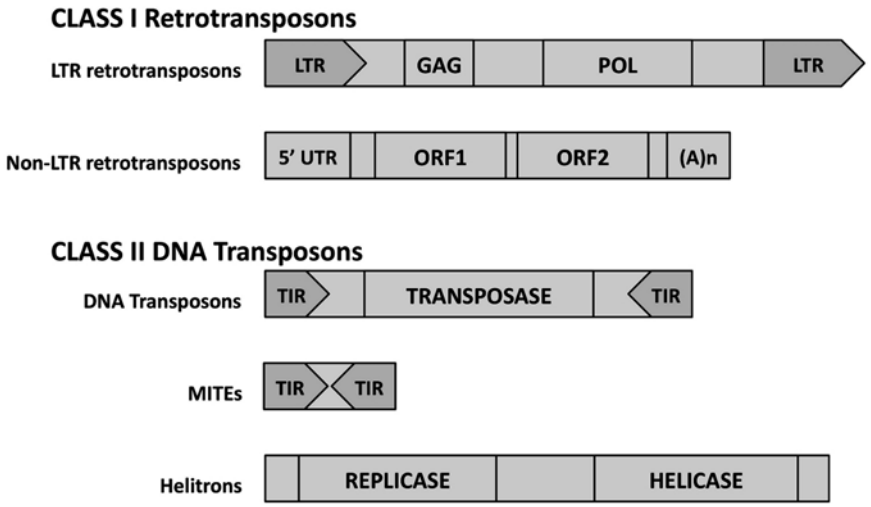

Fig. 2. *Structural features of transposable elements*. Characteristic features typical of Class I and Class II transposable elements. LTR, long terminal repeat; Gag, Pol-protein domains in LTR retrotransposons; 5′ UTR, 5′untranslated region; ORF1 and ORF2, open reading frames in non-LTR retrotransposons; (A)n, poly-A tail; TIR, terminal inverted repeat; transposase, protein domain in DNA transposons; replicase and helicase, protein domains in helitrons.

2. Class II TEs or DNA transposons (cut and paste).
   Class II elements carry a transposase domain that recognizes terminal inverted repeats (TIRs) at the ends of the DNA transposon (Fig. 2). The DNA transposon is cut at each end and excised from the original site before it can be inserted at a new site. Therefore, DNA transposons follow a "cut–paste" mechanism. As opposed to retrotransposons, DNA transposons follow a nonreplicative strategy to proliferate in the genome, and their copy numbers usually are low. Two modes have been documented for DNA transposons: transposition event concomitant with chromosome replication (17) and gap repair via homologous recombination (18).

   One unique class of DNA transposons that replicates via a "copy–paste" mechanism is the helitrons. Helitrons encode for replication initiator (Rep) and DNA helicase (Hel) domains and replicate via a rolling-circle mechanism (Fig. 2). Helitrons can be structurally distinguished by the presence of dinucleotides TC at their 5′ end and tetranucleotides CTRR at their 3′ termini (R stands for A or G). Another feature is the presence of a 15–20 bp palindrome, 10–12 bp from the 3′ end (19). Helitrons have been best characterized in maize, where they have been shown to capture host gene fragments (20).

Based on the presence/absence of protein domains, TEs can be classified as autonomous or nonautonomous elements, respectively. Autonomous TEs encode for proteins responsible for their own movement, whereas nonautonomous elements are usually

derived from deletions of internal protein-coding domains of autonomous elements. Therefore, nonautonomous elements are incapable of self-activation and require autonomous elements to be mobilized. One class of nonautonomous elements is MITEs, miniature inverted-repeat TEs. The relationship between MITEs and their related autonomous DNA transposons was first shown in rice (21).

Generally, during the process of insertion of a TE at a new site, short fragments of genomic DNA flanking the insertion site get duplicated. These are called target site duplications (TSD), and are a result of staggered endonucleolytic cleavage of target-site DNA by the element-encoded transposase or integrase (22). TSD length usually is unique to a particular TE family. However, helitrons do not generate TSDs.

3. Historical Perspective

Since the initial discovery of TEs by Barbara McClintock in the late 1940s, we have come a long way in understanding the role they play in the evolution of genome structure and function. Barbara McClintock was able to correlate and attribute the variegated color of corn kernels to activator/dissociator (Ac/Ds) activity in the maize plant. Later, as more molecular data emerged, it became clear that TEs could transpose and replicate themselves in the genome. This inherent ability of TEs to proliferate in the genome was presumed to mean that they were genomic parasites and therefore often described as "junk" DNA (23). Advances in sequencing technologies led to an exponential increase in the availability of genomic data, which aided in the transformation of thinking from TEs being regarded as "junk" to essential genomic elements.

In fungi, the first TE to be studied was the Ty element, an LTR retrotransposon from *Saccharomyces cerevisiae* (24). The first TE to be cloned from filamentous fungi was the "Tad" element from *Neurospora crassa* (25). The first eukaryotic genome to be completely sequenced was of the yeast, *S. cerevisiae* (26). In 2000, the Fungal Genomics Initiative was initiated at the Broad Institute (http://www.broadinstitute.org/) to sequence candidate fungal species based on their phylogenetic relationships, and relevance to medical, industrial, and research purposes. As more fungal genomes were sequenced, a number of interesting features were found in their genomes, e.g. the lack of active TEs in the *N. crassa* genome sequence (27) could be correlated to the presence of a genome defense mechanism in fungi, known as repeat induced point (RIP) mutation (28).

4. Why Do We Need to Look for Repeats?

Identification of repetitive sequences is important for a number of reasons. First, TEs could be used as genetic tools for gene tagging and phylogenetic studies. Repetitive sequences are also essential in the field of evolutionary biology because of their long-standing relationship with the host genomes and their contribution to genome structure, evolution, and function. From a structural genomics standpoint, repetitive sequences are a major hurdle in the assembly of the final sequenced regions and could lead to spurious matches. For automated annotation, they have to be masked before gene-prediction programs can work effectively. Besides being a hindrance in assembly and annotation, TE constitution and distribution can reveal much about the evolutionary history of the genome.

5. Methods for Identification and Analysis of Repetitive Sequences

Even before complete genome sequences were available, the proportion of repetitive content in a genome can be estimated by $C_o t$-curve analysis (29). This method is based on the reannealing kinetics of denatured DNA strands. Traditionally, the most common and efficient way to check for copy-number variants of a particular region in the genome was to use the sequence as a probe in Southern hybridization (30). Another nucleic acid hybridization-based method is fluorescence in situ hybridization (FISH) (31). Fluorochrome-labeled DNA/RNA probes can be used to visualize the hybridization directly in specific cells or tissues to estimate the copy number of a sequence of interest. However, hybridization-based methods are tedious and time consuming. To overcome these limitations, quantitative real-time PCR (qRT-PCR)-based assays have been developed to estimate copy numbers of specific sequences in a genome (32).

As more and more genomes are sequenced, the field of computational biology is expanding rapidly. Bioinformatic tools with improved algorithms for mining and analysis of information from large data structures are being developed. Here, we briefly describe the characteristics of some of the common bioinformatic tools that are available for whole-genome analysis and that are used commonly to search for repetitive sequences. These can be broadly grouped into three classes depending on the approach followed to identify repetitive sequences.

5.1. Similarity-Based Searches for Repetitive Sequences

Repetitive sequences can be identified based on similarity to elements in a preexisting repeat library. The success rate of similarity-based programs depends greatly on the content of the

precompiled library. Precompiled libraries can be obtained either from RepBase Update or by creating a custom repeat library. RepBase Update is a reference database of repetitive sequences. It contains over 3,600 annotated repeat sequences from a diverse array of eukaryotic organisms (33) and is available from the Genetic Information Research Institute (GIRI, http://www.girinst.org/). However, when working with a recently sequenced genome, one has to keep in mind that repetitive sequences from this nonmodel organism may not be represented in RepBase Update and a custom repeat library may be required.

1. Censor was the first RepBase tool for repetitive sequence detection and masking (34), but its implementation of the search algorithm was inefficient (35). However, recent versions of Censor use the faster WU-BLAST search engine.
2. RepeatMasker (RM) (36) is presently the most widely used tool for masking repeats in a genome (37). The output from RM contains a detailed annotation of the repeats present in the query sequence. It also generates a copy of the query sequence, where the repeat regions have been masked. A common concern is whether RM masks coding regions, although it has been documented that false matches in coding regions are extremely rare ((36); http://repeatmasker.org).
3. MaskerAid was an improvement on the existing RM (38). Without making any changes to RM itself, the default CrossMatch search engine of RM ((39); http://www.phrap.org/phredphrap/phrap.html) was replaced by the faster search engine, WU-BLAST. Swapping the search engine increased the masking speed by more than 30-folds (38).

5.2. Alignment-Based Methods for Repeat Identification

When little to no information is available about the repeat components of a genome, repetitive elements can be identified using a de novo approach. Unlike similarity-based programs, de novo repeat prediction methods do not require any a priori information about repetitive sequences in the genome. As these approaches identify novel repeat families that might be missed otherwise, they should be considered for all genome sequences. Majority of the de novo prediction programs use self-alignment to identify repeats. Several programs with different strengths and weaknesses are available.

1. RECON uses whole-genome self-comparison information to extract repeat elements. Boundaries of individual repeat elements are defined by using multiple alignments. RECON is meant for first-pass automatic classification of repeats in newly sequenced genomes (40). However, it has been used much more widely and has become the major tool for repeat sequence analysis in newly sequenced genomes (41).

2. REPuter consists of three subroutines: Repfind, Repselect, and Repvis (42). Repfind finds exact repeats and uses them as seeds to find other significant degenerate repeats. Repfind output is sorted by significance scores (*E*-values). Repselect is used to parse the Repfind output using different criteria, such as length or degeneracy of the repeats. Finally, Repvis is used for the visualization of Repfind output, producing a color-coded output based on significance scores.
3. Vmatch subsumes REPuter, but it has a very flexible user interface (43). A set of sequences is preprocessed into an index structure, called the persistent index and stored as a collection of several files. During the matching process, only the required part of the index is accessed, considerably minimizing the time constraints. The persistent index can be used for a number of different matching tasks characterized by the kind of sequences to be matched, the kind of matches sought, additional constraints on the matches, and the type of post-processing to be done with the matches.
4. RepeatFinder identifies exact repeats by building a suffix tree data structure and groups these repeat elements into separate repeat classes (44). RepeatFinder needs the support of the REPuter program mentioned earlier. The output from RepeatFinder lists the entire set of repeat classes identified, a multi-FASTA file of the actual repeat sequences, and reports a simple statistical analysis of the results.
5. RepeatMatch can rapidly align very large sequences of either DNA or amino acids and is available as a part of the MUMmer package (45). Like MUMmer, RepeatMatch also utilizes a suffix tree data structure and is specifically designed to find maximal exact repeats. However, it can analyze only one sequence at a time.
6. PILER uses a novel approach to de novo identify subsets of hits forming characteristic patterns of local alignments typical of certain repeats classes (46). There are variants of PILER available for identifying different classes of repeats: PILER-DF (for dispersed families), PILER-PS (for pseudosatellites), PILER-TA (for tandem arrays), and PILER-TR (for terminal repeats). PILER is used in conjunction with two other programs-pairwise alignment of long sequences (PALS) (46) and MUSCLE (47).

5.3. k-mer-Based Methods

Strings of nucleotides or amino acids of a finite length are defined as *k*-mers, where *k* is the length of the sequence. Identifying the frequency of particular *k*-mers in the genome is the focus of additional methods of repetitive sequence annotation.

1. RepeatScout proceeds in four phases (41). First, a file containing the frequency of all *k*-mers in the sequence to be analyzed is created. Using the frequency table and the sequence, an FASTA file containing all the repetitive elements in the sequence is created. Low-complexity and tandem repeats are then filtered out from the output with the help of two external programs, Nseg (48) and tandem repeats finder (49), respectively. Also removed are repeat elements that do not occur at a certain number of times (by default, 10). Finally, SD or exons are removed by comparing the RepeatScout output to the repeats found by RM.
2. TALLYMER computes occurrence counts by making use of enhanced suffix arrays and constructs a *k*-mer frequency index (50). Using this index, the frequency of each *k*-mer can be retrieved efficiently. This strategy enables the processing of very large genomes for a broad range of values of *k* with exceptional speeds.
3. Repeat analysis program (RAP) can identify both exact as well as inexact repeats as it uses a direct indexing of gapped words (51). At any genomic position, several different overlapping gapped words can be produced and counted, which considerably improves the specificity of the signal. Moreover, the same index is used for a given word and its reverse complement.
4. Factor ORacle Repeats (FORRepeats) is named for a novel data structure called factor oracle (52), which is time and space economical and alphabet independent (i.e. it can accommodate degenerate nucleotides as well as amino acid sequences). FORRepeats starts by detecting exact repeats in large sequences, followed by identification of approximate repeats and computing pair-wise comparisons (53). It is a very fast algorithm and can be used for finding and comparing repeats within and between species rapidly.
5. Recovery of ancestral sequences (ReAS) works on unassembled reads of a whole genome shotgun sequence to reconstruct TEs (54). This program is based on the assumption that these TEs exist in high copies across the genome and have not diverged significantly. Reads that contain a particular high-copy *k*-mer are retrieved and assembled into an initial consensus sequence. Then, a search for new *k*-mers at the ends of the consensus is done and iteratively extended until no further extensions are possible. The end result is an ReAS TE.

The detection of repetitive sequences generated by the above-mentioned programs is not the final output, but serves as a starting point for a number of downstream analyses. Once identified, these repetitive sequences need to be annotated and validated. Except for the similarity-based repeat-finding programs that use

sequence similarity to a precompiled repeat library to annotate the repeats, most repeat-finding programs only function to identify repetitive sequences in the genome. They cannot be used to determine the repeat classes or identify the particular structural features that characterize a repeat class. Therefore, to annotate the existing repeat sequences or to look for specific classes of repeats in the whole genome or in a smaller subset of the data, further analyses are needed.

Characterization of repeat families to classes depends on similarity searches to existing sequence databases. To help standardize the different genome annotation projects, an 80–80–80 rule was proposed (55). This rule states that for a repetitive sequence to be assigned to a particular repeat family, it should be at least 80 bp long with 80% or more sequence similarity, across at least 80% of the aligned sequence. Aligned sequence here refers to the repeat-encoded protein domains or terminal repeat regions, or both. An equal consideration should be given to both the coding and non-coding regions, as it helps in the classification of nonautonomous as well as truncated elements.

Different versions of the Basic Local Alignment Search Tool (BLAST) can be used for both nucleotide alignments as well as protein domain searches in repetitive sequences (56). The National Center for Biotechnology Information (NCBI) provides a BLAST Uniform Resource Locator (URL) Application Program Interface (API), which facilitates the automation of the BLAST step. Using this BLAST URL API, a simple script can be used to pass sequences to the NCBI databases and return results locally. This circumvents the need to download huge database files from NCBI to your local machine. Initially, sequence similarity at the nucleotide level can be determined using BLASTN, followed by BLASTX to search for protein domains.

In addition to the identification of protein domains in the repeat families, repetitive sequences also can be characterized based on their associated structural features, such as LTR, TIRs, palindromic sequences, and many more. A number of different programs are available that specifically search for such explicit patterns. This step is necessary for the classification of nonautonomous elements, as they lack a protein domain. Some of the programs that can be used to find specific types of transposons or their associated structural features include:

1. LTRs: LTR_STRUC (57), LTR_FINDER (58), LTRharvest (59), LTR_MINER (60), LTR_par (61), and find_ltr (62).
2. MITEs: FINDMITE (63), MAK (64), and MUST (65).
3. Helitrons: HelSearch1.0 (66) and HelitronFinder (67).

Running different repeat-finding programs and comparing their output from a single genome sequence is a slow, labor-intensive

process. To combine evidence from different programs more efficiently, an automated annotation system is required. Several labs have attempted to merge the processes of TE identification and annotation into a single pipeline. Within each pipeline, several different programs can be used for parallel analyses. Results of these analyses may be stored in a central database, which makes subsequent queries and data retrieval easier. When several different programs are used in a pipeline for the same task, deficiencies in the results of one program can be compensated for by the output from another program. However, the choice of programs that can be used is limited to those that have been put together into a pipeline. Depending on how the pipeline was assembled, it may not be easy to plug in a favorite repeat-finding program. Here, we briefly mention two of the pipelines developed for analyzing repetitive sequences.

1. REPET is a TE annotation pipeline that combines results from multiple de novo and similarity-based TE identification programs (68). In its current form, REPET utilizes results from several programs, including RM, RECON, BLASTER (69), and TE-HMM (70). The REPET pipeline has been further divided into two components, TEdenovo (for the detection of repeats) and TEannot (for repeat annotation). REPET results can be visualized and curated manually with the Apollo genome annotation tool (71). REPET can be used for both detection and annotation of repeats in genomic sequences. One advantage of using REPET is that by thorough annotation of complex TE models, it allows for the detection of highly truncated and/or nested elements in the genome. However, the downside is that it has only been setup to run on a specific cluster, although its source code is available. Furthermore, as all the programs required by REPET are bundled together as one package, no new programs can be added to the existing pipeline.
2. Distributed Annotation Working Group Pipeline to Annotate Wheat Sequences (DAWG-PAWS) is a conglomeration of a number of sequence annotation programs that are held together by a collection of PERL scripts designed to facilitate high-throughput generation of computational results (72). Scripts are available to convert the native output of individual programs to the standard Generic Feature Format (GFF). As with REPET, the Apollo genome annotation tool (71) can be used for the visualization and manual curation of the results. DAWG-PAWS can be used for annotating both genes and TEs. One disadvantage is that all of the required programs need to be downloaded and installed separately. On the other hand, DAWG-PAWS allow the users to use additional or newer programs with the existing pipeline.

6. General Workflow for Identifying and Annotating Repeats

Below is an outline of steps that may be followed to identify and annotate TEs in a particular genome.

1. As most of the commonly used programs have been developed for Unix-based systems and lack a graphical user interface, familiarity with command line interface and basic Unix commands make the job a lot easier. Manual files for all Unix commands can be accessed using the "man" command, followed by the command. The program "readme" file is a good source to determine the order in which the subprograms should be called and the options or parameters to be used with the program.
2. When dealing with the genome sequence of a novel organism, a precompiled repeat dataset most likely is not available. The best approach is to use one of the above-mentioned de novo or *k*-mer-based repeat finding programs. The default output is usually parsed according to a user-defined criterion. Even though sometimes the parsing scripts are included with the program, these may still have to be modified or written from scratch to fit a particular criterion.
3. Once the repeat families have been identified, the next step is to do multiple sequence alignment for all of the elements in each family. Any commonly used sequence alignment program can be used, such as ClustalX (73), MUSCLE (47), etc. This step takes care of the discrepancy arising from the improper merging of two separate families into a single family.
4. When the elements in each family have been verified to be similar, the next step entails similarity searches to a nonredundant repeat database and a protein database to categorize individual repeat families to different classes of repeats.
5. Once the repeat class is known, it becomes easier to search for structural features that represent that particular repeat class. This step helps to verify the findings of the previous step.

The decision to use a repeat-finding program or a pipeline depends on a number of factors. Using different repeat-finding programs and compiling their output is a tedious process, whereas using a preassembled pipeline allows for automation of the above steps. However, as mentioned earlier, it may impose a limit on the choice of repeat-finding programs as it may not be easy to plug in a new program into the existing pipeline.

7. Tandem Repeats

Occurrence of at least two nucleotides repeated head-to-tail in two or more adjacent copies are referred to as tandem repeats. Different functions ranging from gene regulation (74) to a role in recombination (75) have been associated with tandem repeats in the genome. In humans, at least ten genetic disorders have been attributed to a dramatic increase in the copy number of certain tandem repeats (76). Due to frequent variation in copy numbers, tandem repeats have been used for DNA fingerprinting (77).

Many programs have been written to specifically search for tandem repeats. Tandem repeat finder (TRF) (49) is one of the most flexible programs available for the identification of both perfect and degenerate tandem repeats (78). A downstream application, Tandem Repeats Analysis Program (TRAP) (78), processes TRF output to give a list of nonredundant tandem repeats. RepeatMasker, by default, also has an option to search for simple and low-complexity repeats. However, it only scans for di- to hexameric repeats. EQUICKTANDEM, in the EMBOSS (79) package, identifies tandem repeats for each pattern size up to a specified length. EQUICKTANDEM is normally followed by ETANDEM to calculate a consensus for the repeat region. There are other tandem repeat-finding programs, including mreps (80), tandem repeat occurrence locator (81), approximate tandem repeats (82), search for tandem approximate repeats (83) and E-TRA (84). This list is not comprehensive and other programs may be available.

Tandem repeat datasets from different organisms have been assembled into databases. One such database is the tandem repeat database (http://tandem.biomath.mssm.edu/cgi-bin/trdb/trdb.exe; (85)), which is a Web-based repository of tandem repeats from different genomes. It also supports tools to search, store, and manipulate tandem repeats. A more specific database is the short tandem repeat DNA internet database (STRBase, http://www.cstl.nist.gov/biotech/strbase/), maintained by the National Institute of Standards and Technology (NIST) since 1997. It specializes in short tandem repeats used for genetic mapping, forensic DNA typing, and identity testing in humans.

8. Segmental Duplications

Large blocks of genome sequence with high sequence similarity dispersed across the genome are labeled as SD. These may arise from polyploidy events, tandem duplication through unequal crossing over (86), and duplicative transposition (intra- and

interchromosomal exchange). The human genome has a tenfold increase in SD as compared to other sequenced vertebrate genomes (87).

One tool to study SD is DupMasker (88), which uses a library of nonredundant consensus sequences of human SD to query human and nonhuman primate sequences. Based on sequence similarity, the origin and degree of sequence identity of each duplicon can be obtained. The order and orientation of duplicons within complex duplication blocks and differences between sequenced human haplotypes also can be determined.

9. Gene Families and Pseudogenes

Paralogs originate from the duplication of an ancestral gene within the same genome, whereas orthologs are derived from a single ancestral gene after a speciation event. Multicopy gene families may result from a number of mechanisms, including unequal crossing over (89), retroposition (90), SD (91), and whole-genome duplication. ORTHOMCL uses reciprocal best hits within and between genomes followed by Markov clustering algorithm (92) to determine both paralogs and orthologs, respectively (93). FAST_PAN was developed to identify novel members of any well-characterized gene family (94). Protein query sequences are used to search for distantly related EST sequences of a gene family. The final output contains the statistical significance, alignment coverage, percent identity, and phylogenetic position. The most thorough method to determine orthologs or paralogs, involves comparing the gene tree to the species tree considered as a reference (tree reconciliation). This is a very tedious approach. For large-scale studies, an algorithm, RAP, was developed to infer speciation and duplication events by comparing gene and species trees (95).

10. Summary

Many tools and approaches are available for identification and analysis of repeated sequences within genomes. The choice of method used may depend on a number of factors, including technical aspects such as, familiarity with command line interface, availability of source code or binaries for a given operating system, ease of installation, or could be determined by kind of analysis or type of output required. There remains a great need for an automated pipeline for identification, classification, and nomenclature of repetitive elements within sequenced genomes. A basic

outline of one approach for detection and annotation of repeats was provided. The specific programs and methods needed for analysis with a particular fungal genome depend on resources and expertise available and the level of annotation desired.

References

1. Hugenholtz, P., and Tyson, G. W. (2008) Metagenomics *Nature* **455**, 481–3.
2. Wolinsky, H. (2007) The thousand-dollar genome *EMBO reports* **8**, 900–3.
3. Thomas, C. A. (1971) The genetic organization of chromosomes *Annu Rev Genet* **5**, 237–56.
4. Lynch, M., and Conery, J. S. (2003) The origins of genome complexity *Science* **302**, 1401–4.
5. Walbot, V., and Petrov, D. A. (2001) Gene galaxies in the maize genome *Proc Natl Acad Sci USA* **98**, 8163–4.
6. Thomas, E. E. (2005) Short, local duplication in eukaryotic genomes *Curr Op Genet Dev* **15**, 640–4.
7. Kim et al (2008) Analysis of copy number variants and segmental duplications in the human genome: Evidence for a change in the process of formation in recent evolutionary history *Genome Res* **18**, 1865–74
8. Bailey, J. A., Gu, Z., Clark, R. A., Reinert, K., Samonte, R. V., Schwartz, S., Adams, M. D., Myers, E. W., Li, P. W., and Eichler, E. E. (2002) Recent segmental duplications in the human genome *Science* **297**, 1003–7.
9. Cheng, Z., Ventura, M., She, X., Khaitovich, P., Graves, T., Osoegawa, K., Church, D., DeJong, P., Wilson, R. K., Paabo, S., Rocchi, M. and Eichler E. E. (2005) A genome-wide comparison of recent chimpanzee and human segmental duplications *Nature* **437**, 88–93.
10. Koszul, R. S., Caburet, B. D., and Fischer, G. (2004) Eucaryotic genome evolution through the spontaneous duplication of large chromosomal segments *EMBO J* **23**, 234–43.
11. Zhang, J. (2003) Evolution by gene duplication: an update *Trends Ecol Evol* **18**, 292–8.
12. Lespinet, O., Wolf, Y. I., Koonin, E. V., and Aravind, L. (2002) The role of lineage-specific gene family expansion in the evolution of eukaryotes *Genome Res* **12**, 1048–59.
13. Gu, Z., Cavalcanti, A., Chen, F. C., Bouman, P., and Li, W. H. (2002) Extent of gene duplication in the genomes of Drosophila, nematode, and yeast *Mol Biol Evol* **19**, 256–62.
14. Zhang, X., and Firestein, S. (2002) The olfactory receptor gene superfamily of the mouse *Nat Neurosci* **5**, 124–33.
15. Rajashekar, B., Kohler, A., Johansson, T., Martin, F., Tunlid, A., and Ahrén, D. (2009) Expansion of signal pathways in the ectomycorrhizal fungus Laccaria bicolor- evolution of nucleotide sequences and expression patterns in families of protein kinases and RAS small GTPases *New Phytol* **183**, 365–79.
16. Wilhelm, M., and Wilhelm, F. X. (2001) Reverse transcription of retroviruses and LTR retrotransposons *Cell Mol Life Sci* **58**, 1246–62.
17. Kunze, R., and Weil, C.F. (2002) The hAT and CACTA superfamilies of plant transposons In *Mobile DNA II* (eds. Craig, N., Craigie, R., Gellert, M., and Lambowitz, A.) ASM Press, Washington, DC, 565–610.
18. Engels, W. R., Johnson-Schlitz, D. M., Eggleston, W. B., and Sved, J. (1990) High-frequency P-element loss in *Drosophila* is homolog dependent. *Cell* **62**, 515–25.
19. Kapitonov, V. V., and Jurka, J. (2001) Rolling-circle transposons in eukaryotes *Proc Natl Acad Sci USA* **98**, 8714–9.
20. Morgante, M., Brunner, S., Pea, G., Fengler, K., Zuccolo, A., and Rafalski, A. (2005) Gene duplication and exon shuffling by Helitron-like transposons generate intraspecies diversity in maize *Nat Genet* **37**, 997–1002.
21. Jiang, N., Feschotte, C., Zhang, X., and Wessler, S. R. (2004) Using rice to understand the origin and amplification of miniature inverted repeat transposable elements (MITEs) *Curr Op Plt Biol* **7**, 115–9.
22. Feschotte, C., and Wessler, S. R. (2001) Treasures in the attic: Rolling circle transposons discovered in eukaryotic genomes *Proc Natl Acad Sci USA* **98**, 8923–4.
23. Orgel, L. E., and Crick, F. H. (1980) Selfish DNA: the ultimate parasite *Nature* **284**, 604–7.
24. Cameron, J. R., Loh, E. Y., and Davis, R. W. (1979) Evidence for transposition of dispersed repetitive DNA families in yeast *Cell* **16**, 739–51.
25. Kinsey, J. A., and Helber, J. (1989) Isolation of a transposable element from *Neurospora crassa Proc Natl Acad Sci USA* **86**, 1929–33.
26. Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J.,

Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996) Life with 6000 genes *Science* **274**, 563–7.

27. Galagan, J. E., Calvo, S. E., Borkovich, K. A., Selker, E. U., Read, N. D., Jaffe, D., FitzHugh, W., Ma, L. J., Smirnov, S., Purcell, S., Rehman, B., Elkins, T., Engels, R., Wang, S., Nielsen, C. B., Butler, J., Endrizzi, M., Qui, D., Ianakiev, P., Bell-Pedersen, D., Nelson, M. A., Werner-Washburne, M., Selitrennikoff, C. P., Kinsey, J. A., Braun, E. L., Zelter, A., Schulte, U., Kothe, G. O., Jedd, G., Mewes, W., Staben, C., Marcotte, E., Greenberg, D., Roy, A., Foley, K., Naylor, J., Stange-Thomann, N., Barrett, R., Gnerre, S., Kamal, M., Kamvysselis, M., Mauceli, E., Bielke, C., Rudd, S., Frishman, D., Krystofova, S., Rasmussen, C., Metzenberg, RL., Perkins, DD., Kroken, S., Cogoni, C., Macino, G., Catcheside, D., Li, W., Pratt, R. J., Osmani, S. A., DeSouza, C. P., Glass, L., Orbach, M. J., Berglund, J. A., Voelker, R., Yarden, O., Plamann, M., Seiler, S., Dunlap, J., Radford, A., Aramayo, R., Natvig, D. O., Alex, L. A., Mannhaupt, G., Ebbole, D. J., Freitag, M., Paulsen, I., Sachs, M. S., Lander, E. S., Nusbaum, C., and Birren, B. (2003) The genome sequence of filamentous fungus *Neurospora crassa Nature* **422**, 859–68.
28. Selker, E. U., Cambareri, E. B., Jensen, B. C., and Haack, K. R. (1987) Rearrangement of duplicated DNA in specialized cells of *Neurospora Cell* **51**, 741–52.
29. Britten, R. J., Graham, D. E., and Neufeld, B. R. (1974) Analysis of repeating DNA by reassociation *Methods Enzymol* **29**, 363–418.
30. Sambrook, J., and Russell, D. W. (2001) Molecular Cloning: A Laboratory Manual CSHL Press, Cold Spring Harbor, NY.
31. Zhong, X. B., Fransz, P. F., Wennekes-van, E. J., Zabel, P., van Kammen, A., and de Jong, J. H. (1996) High resolution mapping by fluorescence in situ hybridisation to pachytene chromosomes and extended DNA fibres *Plant Mol Biol Rep* **14**, 232–42.
32. Yuan, J. S., Burris, J., Stewart, N. R., Mentewab, A., and Stewart, C. N. (2007) Statistical tools for transgene copy number estimation based on real-time PCR *BMC Bioinformatics* **8**, S6.
33. Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O., and Walichiewicz, J. (2005) Repbase Update, a database of eukaryotic repetitive elements *Cytogenet Genome Res* **110**, 462–7.
34. Jurka, J., Klonowski, P., Dagman, V., and Pelton, P. (1996) CENSOR-a program for identification and elimination of repetitive elements from DNA sequences *Comput Chem* **20**, 119–21.
35. Kohany, O., Gentles, A. J., Hankus, L., and Jurka, J. (2006) Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor *BMC Bioinform* **7**, 474.
36. Smit, A. F. A., Hubley, R., and Green, P. (1996–2004) RepeatMasker Open-3.0. http://repeatmasker.org.
37. Morgulis, A., Gertz, E. M., Schäffer, A. A., and Agarwala, R. (2006) WindowMasker: window-based masker for sequenced genomes *Bioinformatics* **22**, 134–41.
38. Bedell, J. A., Korf, I., and Gish, W. (2000) MaskerAid: a performance enhancement to RepeatMasker *Bioinformatics* **16**, 1040–1.
39. Green, P. (1994–1999) http://www.phrap.org/phredphrap/phrap.html.
40. Bao, Z., and Eddy, S. R. (2002) Automated *de novo* identification of repeat sequence families in sequenced genomes *Genome Res* **12**, 1269–76.
41. Price, A. L., Jones, N. C., and Pevzner, P. A. (2005) *De novo* identification of repeat families in large genomes *Bioinformatics* **21**, Suppl 1, i351–8.
42. Kurtz, S., Choudhuri, J. V., Ohlebusch, E., Schleiermacher, C., Stoye, J., and Giegerich, R. (2001) REPuter: The manifold applications of repeat analysis on a genomic scale *Nucleic Acids Res* **29**, 4633–42.
43. Kurtz, S., and Schleiermacher, C. (1999) REPuter: fast computation of maximal repeats in complete genomes *Bioinformatics* **15**, 426–7.
44. Volfovsky, N., Haas, B. J., and Salzberg, S. L. (2001) A clustering method for repeat analysis in DNA sequences *Genome Biol* **2**, research0027.1–0027.11.
45. Delcher, A. L., Kasif, S., Fleischmann, R. D., Peterson, J., White, O., and Salzberg, S. L. (1999) Alignment of whole genomes. *Nucleic Acids Res* **27**, 2369–76.
46. Edgar, R. C., and Myers, E.W. (2005) PILER: identification and classification of genomic repeats. *Bioinformatics* **21**, Suppl 1, i152–8.
47. Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput *Nucleic Acids Res* **32**, 1792–7.
48. Wootton, J. C., and Federhen, S. (1993) Statistics of local complexity in amino acid sequences and sequence databases *Comput Chem* **17**, 149–63.
49. Benson, G. (1999) Tandem repeats finder: a program to analyze DNA sequences *Nucleic Acids Res* **27**, 573–80.

50. Kurtz, S., Narechania, A., Stein, J. C., and Ware, D. (2008) A new method to compute K-mer frequencies and its application to annotate large repetitive plant genomes *BMC Genomics* **9**, 517.
51. Campagna, D., Romualdi, C., Vitulo, N., Favero, M. D., Lexa, M., Cannata, N., and Valle, G. (2005) RAP: a new computer program for de novo identification of repeated sequences in whole genomes *Bioinformatics* **21**, 582–8.
52. Allauzen, C., Crochemore, M., and Raffinot, M. (1999) Factor oracle: a new structure for pattern matching In Pavelka, J., Tel, G., and Bartosek, M. (eds), SOFSEM '99, Theory and Practice of Informatics, Lecture Notes in Computer Science, 1725, Springer, Milovy, Czech Republic, Berlin, pp. 291–306.
53. Lefebvre, A., Lecroq, T., Dauchel, H., and Alexandre, J. (2003) FORRepeats: detects repeats on entire chromosomes and between genomes *Bioinformatics* **19**, 319–26.
54. Li, R., Ye, J., Li, S., Wang, J., Han, Y., Ye, C., Wang, J., Yang, H., Yu, J., Wong, G. K., and Wang, J. (2005) ReAS: Recovery of Ancestral Sequences for transposable elements from the unassembled reads of a whole genome shotgun *PLoS Comput Biol* **1**, e43.
55. Wicker, T., Sabot, F., Hua-Van, A., Bennetzen, J. L., Capy, P., Chalhoub, B., Flavell, A., Leroy, P., Morgante, M., Panaud, O., Paux, E., SanMiguel, P. and Schulman, A. H. (2007) A unified classification system for eukaryotic transposable elements *Nat Rev Genet* **8**, 973–82.
56. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool *J Mol Biol* **215**, 403–10.
57. McCarthy, E. M., and McDonald, J. F. (2003) LTR_STRUC: a novel search and identification program for LTR retrotransposons *Bioinformatics* **19**, 362–7.
58. Xu, Z., and Wang, H. (2007) LTR_FINDER: an efficient tool for the prediction of full-length LTR retrotransposons *Nucleic Acids Res* **35**, W265–8.
59. Ellinghaus, D., Kurtz, S., and Willhoeft, U. (2008) LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics* **9**, 18.
60. Pereira, V. (2004) Insertion bias and purifying selection of retrotransposons in the *Arabidopsis thaliana* genome *Genome Biol* **5**, R79.
61. Kalyanaraman, A., and Aluru, S. (2006) Efficient algorithms and software for detection of full-length LTR retrotransposons. *J Bioinform Comput Bio* **4**, 197–216.
62. Rho, M., Choi, J. H., Kim, S., Lynch. M., and Tang, H. (2007) *De novo* identification of LTR retrotransposons in eukaryotic genomes *BMC Genomics* **8**, 90.
63. Tu, Z. (2001) Eight novel families of miniature inverted repeat transposable elements in the African malaria mosquito, *Anopheles gambiae Proc Natl Acad Sci USA* **98**, 1699–704.
64. Yang, G., and Hall, T. C. (2003) MAK, a computational tool kit for automated MITE analysis *Nucleic Acids Res* **31**, 3659–65.
65. Chen, Y., Zhou, F., Li, G., and Xu, Y. (2009) MUST: A system for identification of miniature inverted-repeat transposable elements and applications to *Anabaena variabilis* and *Haloquadratum walsbyi Gene* **436**, 1–7s.
66. Yang, L., and Bennetzen, J. (2009) Structure-based discovery and description of plant and animal Helitrons *Proc Natl Acad Sci USA* **106**, 12832–7.
67. Du, C., Caronna, J., He. L., and Dooner, H. K. (2008) Computational prediction and molecular confirmation of Helitron transposons in the maize genome *BMC Genomics* **9**, 51.
68. Quesneville, H., Bergman, C. M., Andrieu, O., Autard, D., Nouaud, D., Ashburner, M., and Anxolabehere, D. (2005) Combined evidence annotation of transposable elements in genome sequences *PLoS Comput Biol* **1**, 166–75.
69. Quesneville, H., Nouaud, D., and Anxolabehere, D. (2003) Detection of new transposable element families in *Drosophila melanogaster* and *Anopheles gambiae* genomes *J Mol Evol* **57**, S50–9.
70. Andrieu, O., Fiston, A. S., Anxolabehere, D., and Quesneville, H. (2004) Detection of transposable elements by their compositional bias *BMC Bioinformatics* **5**, 94.
71. Lewis, S. E., Searle, S. M., Harris, N., Gibson, M., Iyer, V., Richter, J., Wiel, C., Bayraktaroglu, L., Birney, E., Crosby, M. A., Kaminker, J. S., Matthews, B., Prochnik, S. E., Smith, C. D., Tupy, J. L., Rubin, G. M., Misra, S., Mungall, C. J., and Clamp, M. E. (2002) Apollo: A sequence annotation editor *Genome Biol* **3**, Research0082.
72. Estill, J. C., and Bennetzen, J. L. (2009) The DAWGPAWS pipeline for the annotation of genes and transposable elements in plant genomes *Plant Methods* **5**, 8.
73. Larkin, M. A., Blackshields, G., Brown, N. P., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R., Thompson, J. D., Gibson, T. J., and Higgins, D. G. (2007) Clustal W and Clustal X version 2.0 *Bioinformatics*, **23**, 2947–8.

74. Richards, R., Holman, K., Yu, S., and Southerland, G. (1993) Fragile X syndrome unstable element, p(CCG)n, and other simple tandem repeat sequences are binding sites for specific nuclear proteins *Hum Mol Genet* **2**, 1429–35.
75. Majewski, J., and Ott, J. (2000) GT repeats are associated with recombination on human chromosome 22 *Genome Res* **10**, 1108–14.
76. Wells, R. D. (1996) Molecular basis of genetic instability of triplet repeats *J Biol Chem* **271**, 2875–8.
77. Edwards, A., Hammond, H. A., Jin, L., Caskey, C. T., and Chakraborty, R. (1992) Genetic variation at five trimeric and tetrameric tandem repeat loci in four human population groups *Genomics* **12**, 241–53.
78. Sobreira, T. J., Durham, A. M., and Gruber, A. (2006) TRAP: automated classification, quantification and annotation of tandemly repeated sequences *Bioinformatics* **22**, 361–2.
79. Rice, P., Longden, I., and Bleasby, A. (2000) EMBOSS: The european molecular biology open software suite *Trends Genet* **16**, 276–7.
80. Kolpakov, R., Bana. G., and Kucherov, G. (2003) mreps: Efficient and flexible detection of tandem repeats in DNA *Nucleic Acids Res* **31**, 3672–8.
81. Castelo, A. T., Martins, W., and Gao, G. R. (2002) TROLL-tandem repeat occurrence locator *Bioinformatics* **18**, 634–6.
82. Krishnan, A., and Tang, F. (2004) Exhaustive whole-genome tandem repeats search *Bioinformatics* **20**, 2702–10.
83. Delgrange, O., and Rivals, E. (2004) STAR: an algorithm to Search for Tandem Approximate Repeats *Bioinformatics* **20**, 2812–20.
84. Karaca, M., Bilgen, M., Onus, A. N., Ince, A. G., and Elmasulu, S. Y. (2005) Exact tandem repeats analyzer (E-TRA): a new program for DNA sequence mining *J Genet* **84**, 49–54.
85. Gelfand, Y., Rodriguez, A., and Benson, G. TRDB-the Tandem Repeats Database (2007) *Nucleic Acids Res* **35**, D80–7.
86. Lupski, J. R. (1998) Genomic disorders: structural features of the genome can lead to DNA rearrangements and human disease traits *Trends Genet* **14**, 417–22.
87. Eichler, E. E. (2001) Recent duplication, domain accretion and the dynamic mutation of the human genome *Trends Genet* **17**, 661–9.
88. Jiang, Z., Hubley, R., Smit, A., and Eichler, E. E. (2008) DupMasker: A tool for annotating primate segmental duplications *Genome Res* **18**, 1362–8.
89. Leh-Louis, V., Wirth, B., Potier, S., Souciet, J. L. and Despons, L. (2004) Expansion and contraction of the DUP240 multigene family in *Saccharomyces cerevisiae* populations *Genetics* **167**, 1611–9.
90. Schacherer, J., Tourrette, Y., Souciet, J. L., Potier, S. and De Montigny, J. (2004) Recovery of a function involving gene duplication by retroposition in *Saccharomyces cerevisiae Genome Res* **14**, 1291–7.
91. Korbel, J. O., Kim, P. M., Chen, X., Urban, A. E., Weissman, S., Snyder, M., and Gerstein, M. B. (2008) The current excitement about copy-number variation: How it relates to gene duplications and protein families *Curr Op Struct Biol* **18**, 366–74.
92. Van Dongen, S. (2000) Graph clustering by flow simulation *PhD Thesis* University of Utrecht The Netherlands.
93. Li, L., Stoeckert Jr., C. J., and Roos, D. S. (2003) OrthoMCL: Identification of Ortholog Groups for Eukaryotic Genomes *Genome Res* **13**, 2178–89.
94. Retief, J. D., Lynch, K. R., and Pearson, W. R. (1999) Panning for genes-A visual strategy for identifying novel gene orthologs and paralogs *Genome Res* **9**, 373–82.
95. Dufayard, J. F., Duret, L., Penel, S., Gouy, M., Rechenmann, F., and Perrière G. (2005) Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases *Bioinformatics* **21**, 2596–603.

Chapter 4

Next-Generation Sequencing and Potential Applications in Fungal Genomics

Phillip SanMiguel

Abstract

Since the first fungal genome was sequenced in 1996, sequencing technologies have advanced dramatically. In recent years, it has become possible to cost-effectively generate vast amounts of DNA sequence data using a number of cell- and electrophoresis-free sequencing technologies, commonly known as "next" or "second" generation. In this chapter, we present a brief overview of next-generation sequencers that are commercially available now. Their potential applications in fungal genomics studies are discussed.

Key words: Second-generation sequencer, GS-FLX, 454, SOLiD, Solexa, Transcriptome analysis, Comparative genomics

1. Introduction

Determining DNA and RNA sequence is central to many biological studies. Automated Sanger or "first generation" sequencing (1) was, until recently, the dominant technology for large sequencing projects. In the past few years, a number of second-generation, highly parallel sequencing technologies have been rapidly developing and likely impact many aspects of molecular and genetic studies. For Sanger sequencing (2), a set of nascent strands are synthesized from the same primer site on template DNA fragments typically originating from plasmid clones maintained and amplified in *Escherichia coli*. The polymerization reaction contains dideoxy nucleotides that, once incorporated, both halt the further extension of that strand and label that terminating base with one of four base-specific dyes. These products are fractionated by size via electrophoresis and the resulting "peaks" or "bands" are detected and used to decode the DNA sequence.

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722,
DOI 10.1007/978-1-61779-040-9_4,

In general, next-generation sequencing technologies use cell-free systems for template preparation (3). Because some DNA sequences are not clonable in *E. coli*, using specialized PCR reactions for template amplification in vitro potentially avoids some of the gaps in sequence coverage that occur in Sanger sequencing. Generally, PCR "amplicons" are constructed as follows. Double-stranded oligonucleotide adaptors are attached to fragmented DNA via intermolecular ligation. For some applications, segments from both ends of size-selected DNA fragments are combined into single amplicons as "mate paired" libraries. Removing the use of bacteria to store and amplify the amounts of DNA to be sequenced removes one bottleneck present in most first-generation sequencing methodologies. Another major improvement is that next-generation sequencers use electrophoresis-free systems to decipher DNA sequences by monitoring each sequencing step in situ. By removing the requirement for electrophoretic fractionation of reaction products, all next-generation sequencing methodologies can produce large amounts of sequence data in a single instrument run. Below is a brief summary of next-generation sequencing technologies that are now commercially available. Different techniques are used to amplify template DNA and decode nucleotide sequences in different next-generation sequencing platforms. Their unique features and potential applications in fungal genomics studies are discussed.

2. Sequencers

2.1. The Roche GS-FLX Genome Sequencer (Also Known as "454")

The 454 sequencer, in its initial incarnation, the GS-20, was the first commercially available next-generation sequencing instrument (4). For sequencing template preparation, DNA fragments are ligated to flanking adapters to form amplicons, bound to 20 μm polystyrene beads, during amplification by emulsion PCR (emPCR). These beads, each carrying thousands of copies of a different and likely unique DNA amplicon, are each deposited into one of 3.6 million hexagonal wells of a PicoTiterPlate (PTP) by centrifugation. The PTP is divided into two, four, eight, or sixteen "regions" with a gasket. Successful nucleotide sequence reads of DNA bound to one to two million of these beads result from pyrosequencing (5), that uses, like Sanger sequencing, a DNA polymerase-driven method. However, instead of detecting chain termination events, pyrosequencing deploys a series of reactions ending in luciferase to detect the release of inorganic pyrophosphate upon nucleotide incorporation. Extension terminators are not employed, nor is color-encoded data collected. Instead data is collected in real time upon the addition of sequential "flows" of single nucleotide types. The number of bases incorporated in a single flow is detected by the amount of light generated

Table1
Comparison of commercially available next-generation sequencers

Sequencers	Sequence generated (billions of bases)		Flowcells/run	Regions (lanes)/flowcell	
	Per run	Per day		Max	Min
Life Technologies (Applied Biosystems) 3730XL	0.0001	0.001	n/a	n/a	n/a
Roche GS-FLX (454)	0.5	0.5	1	16	2
Illumina Genome Sequencer (GA-2x)	25	2	1	8	8
Life Technologies (Applied Biosystems) SOLiD	60	4	2	8	1
Helicos HeliScope	20	2.5	2	25	25

by luciferase. Hence, depending on the sequence composition of the templates, 200 nucleotide flows tend to collect a median of 500 bases or more of sequence over the course of roughly 9 h. However, difficulties in distinguishing the true lengths of long runs of the same nucleotide in a DNA template via the intensity of light generated via pyrosequencing are endemic to this methodology but may be largely overcome by increased sequencing depth.

Currently, the Roche GS-FLX can produce roughly 0.4 billion bases of sequence per run in the form of approximately one million reads of mean length 400 bases using the GS-FLX Titanium chemistry and protocols (Table 1). Unlike all other next-generation sequencers, a GS FLX run completes in less than 1 day of instrument time. Cost per raw base of sequence obtained from a GS-FLX is one to two orders of magnitude less expensive than Sanger sequencing. This makes the instrument ideal for projects benefiting from long read lengths, such as full genome shotgun and de novo transcriptome sequencing (6–8).

Seven to eight hundred base reads should be possible with the next revision of the GS-FLX chemistry, nearly doubling the sequence generated per run of the instrument and bringing read lengths that achieve parity with first-generation sequencers. This new methodology utilizes 350 nucleotide flows rather than 200 and requires modification of the GS-FLX instrument to allow storage space of the extra reagent volumes required.

2.2. The Illumina Genome Sequencer (Also Known as "Solexa")

The Solexa sequencer was the first of two major commercial instruments to offer vastly increased read numbers but, initially, at the cost of short read lengths and somewhat lower accuracy. In comparison with GS-FLX/454, Illumina/Solexa uses different methods for template preparation, sequencing, and detection.

Genomic fragment or mate-pair templates of the size range of 150–200 bp are ligated with two unique adapters, denatured, and bound to the flow cell coated with oligonucleotides corresponding to the adapter sequences. Solid-phase amplification (bridge amplification) occurs when the free end of a single-stranded DNA anneals to a complementary oligonucleotide on the flow cell surface. Localized amplification of individual DNA molecules results in the formation of millions of molecular clusters across the flow cell surface.

Similarly to first-generation sequencing, Illumina/Solexa sequencing uses dye-labeled termination chemistry and a DNA polymerase to enable sequence determination. However, the dye-terminators are reversible (9). After the addition of a single dye-terminated base to each clonally amplified DNA cluster, the flow cell surface is scanned, the dye and extension-blocking moiety of terminal base is removed, allowing another extension cycle to begin. Current Illumina instruments are capable of producing upward of 12 million reads in each of eight channels of a flow cell. The large number of reads generated in a single Illumina/Solexa run results in raw sequence costs an order of magnitude less than those of the GS-FLX. Because of the relatively small average size of fungal genomes, each lane of a flow cell can be used to sequence different fungal species or strains. A variety of read lengths, typically between 30 and 150 bases can be collected, with longer reads consuming more reagents and machine time than shorter ones. Further, amplicon clusters can be strand-converted and resequenced from the opposite end to produce paired end sequences that may overlap. Long read length, paired end runs require upward of 2 weeks of instrument time to perform, but this nevertheless represents a data generation rate of nearly two billion bases per day.

2.3. Life Technologies/ Applied Biosystems SOLiD System

The Sequencing by Oligonucleotide Ligation and Detection (SOLiD) system was last to market of the three major commercial next-generation instruments available today. It is based on the sequence-by-ligation chemistry (10). One micrometer diameter beads binding clonal DNA populations amplified from adapter-flanked DNA fragments (amplicons) by emPCR are deposited and covalently linked onto a glass slide. Each slide comprises one, four, or eight gasket-separated sections or "spots," and two slides can be independently run simultaneously. A universal primer is annealed to the template amplicons and a library of dye-labeled nine base oligonucleotide probes then are pumped into to the flow cell containing the slide. The design of these probes is complex, but briefly, they contain two 5′ interrogating bases followed by seven random or nonspecific bases and a dye ambiguously specifying the identity of the first two bases. These oligos are allowed to hybridize to the template, followed by ligation, washing away of nonligated oligos

and four-color imaging. Subsequent deblocking of the ligated nonamers removes their terminal three bases along with the dye and repeated cycles of ligation and detection ensue out to the length of the sequence read. Because a single primer only detects sequence every five bases, five cycles of ligation extending from a different initiating base position must be undertaken to obtain unbroken sequence across the read. The SOLiD is unique in collecting "dual base" encoded data, where each base of sequence produced is the result of two independent measurement events. In principle, this is a great advantage because it allows miscall error detection when a read is mapped against a reference sequence.

The latest SOLiD 4 system can generate over 1 billion high quality sequence reads per run. Short sequence reads, 35 or 50 bases, generated by SOLiD make it useful almost exclusively for resequencing projects or sequence tag-based experiments (such as digital gene expression). The SOLiD can consecutively sequenced from two priming sites within the same amplicon, which allows the construction of mate pair templates such that generation of sequence tags derived from either end of a DNA fragment from roughly 0.6 to 6 kb in length are produced. In general, the SOLiD 4 system collects roughly 50 billion bases of sequence over the course of a 1 week run – in excess of seven billion bases per day. If the average genome size for filamentous fungi is estimated to be 40 Mb, one SOLiD run could accommodate resequencing of multiple fungal genomes.

2.4. Helicos HeliScope Single Molecule Sequencer

Unlike the three common or major next-generation sequencing technologies, the HeliScope single molecular sequencer analyzes a single DNA template molecule. As the first commercially available sequencer to eliminate template amplification, Heliscope uses its so-called tSMS chemistry to detect sequencing-by-synthesis reactions for billions of individual DNA molecules in parallel. Sequencing a single DNA molecule directly can eliminate the need of template amplification, which is a bottleneck of the above described next-generation (second-generation) sequencing technologies. Therefore, the HeliScope sequencing platform is kind of a third-generation technology (Fig. 1).

The Heliscope offers 50 channels in two flow cells capable of producing >20 billion bases over the course of an 8-day run. This places its throughput at nearly the same as the SOLiD. However, with very few HeliScope single molecular sequencers currently available in the field, it is difficult to assess the real-world performance of this sequencing platform. That said, the elimination of template amplification prior to sequencing certainly would reduce the often extensive and laborious presequencing sample preparation required by second-generation sequencers. It may also reduce sequence bias be eliminating any requirement for PCR amplification.

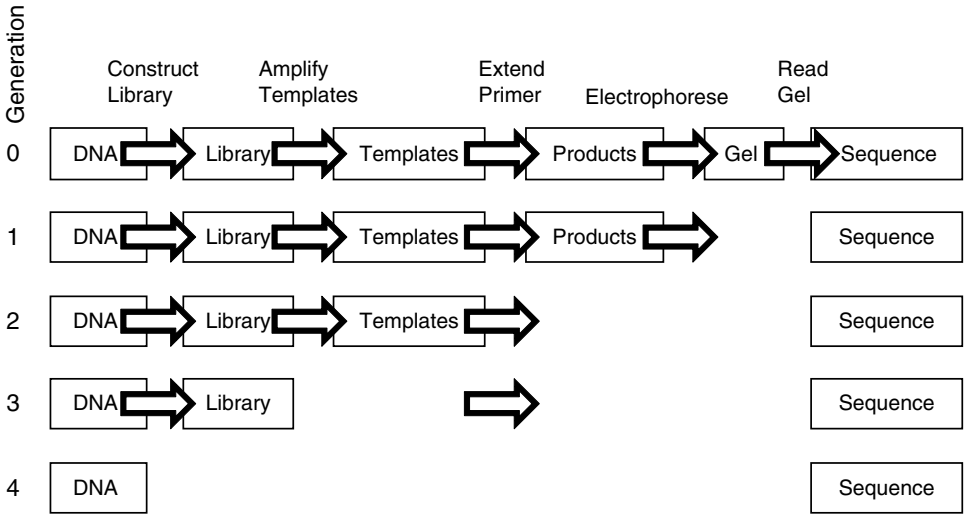

Fig. 1. *Schematic drawing of five sequencing paradigm variants.* The standard paradigm for DNA sequencing depicted as a series of data (*boxes*) converted by processes (*arrows*). The bottlenecks in DNA sequencing are successively removed by different generations of sequencing technologies. Generation zero uses gel electrophoresis to separate radioisotope-labeled nascent DNA fragments and detect signals on autoradiographs to infer the sequence of the template strand. Automated Sanger sequencing is considered the first-generation sequencing paradigm in which the dye-labeled terminating base of each product strand is detected during electrophoresis. Next- or second-generation sequencing uses electrophoresis-free detection systems to decode sequences of template DNA (amplicons) amplified in vitro. Further development allows the template DNA being sequenced without amplification. Sequencing platforms using single molecule detection systems can be described as third-generation sequencing. However, the ultimate sequencing paradigm likely transcends the use of enzymes to read DNA sequence directly with little manipulation required of the user. No commercial sequencers of this type currently exist, but this is an area of research with some promising candidates, such as nanopore technologies that detect the sequence of a strand of DNA as it passes through a small pore (11).

Another single molecular sequencing technique is being developed by *Pacific Biosciences* (PacBio, http://www.pacificbiosciences.com). It uses individual polymerases tethered to the base of small "zero wave guide" reactor cells that each processes a single template. Optical properties of these small reactors allow successful nucleotide incorporation events to be detected by the duration of the dwell time of the prospective dNTP in the active site of the polymerase as the base is added to the nascent strand. This real time, single molecular sequencing technology, when becomes commercially available, also represents a third-generation sequencing technology (Fig. 1). Further, this technology is said to be capable of producing long reads – in excess of even first-generation read lengths.

2.5. Dover Systems Polonator G.007

The Polonator uses a similar template preparation and sequence-by-ligation strategy to the SOLiD. During each run, a series of anchor primers are annealed to the adapter sequences added to the 17–18 bp proximal or distal (5′ or 3′) genomic DNA tags (DNA fragments or mate-pairs) during template preparation. Fully degenerate nonanucleotides ("nonamers") with each

component labeled with one of four fluorophores are then ligated on the anchor primer. The four colors of newly added bases at the query position of genomic DNA tags are detected to identify the base type. Each sequence cycle takes about 3 h. This sequence cycle is repeated for both the 5′ and 3′ ends of the proximal or distal tags. A read length of 26 bases (13 from each of the paired tags) is generated for each DNA template (with four to five bases in the middle remaining unread). In total, about 10 Gb of sequence can be generated in each instrument run, which takes about 80 h.

In contrast to the SOLiD, only one base is queried in each reaction cycle and each base is only interpreted once by the Polonator G.007. Therefore, the accuracy of the Polonator G.007 is likely lower than that of SOLiD. However, the Polonator G.007 is an open platform and its low upfront and recurring costs likely are very attractive to researchers with limited budgets or do-it-yourself mentality. Users can use the standard protocols, reagents, and software but all aspects of this Polonator G.007 system are open and programmable, providing the flexibility that users may find helpful for different sequencing projects.

3. Applications of Next-Generation Sequencing in Fungal Genomics Studies

Next-generation sequencers can rapidly generate tremendous numbers of DNA sequences unbiased by maintenance in *E. coli* host strains and with ever-decreasing cost per base. Below are some of their major applications in fungal genomics studies. Based on specific scopes of different projects, one next-generation sequencing platform may be more suitable than other ones. And it is also true that a number of new sequencing technologies are under development.

3.1. De Novo Sequencing of Fungal Genomes

The difficulty of a de novo sequencing project depends mainly upon the total size of the genome and the nature of the repetitive fraction of that genome. De novo sequencing large, highly repetitive genomes, such as mammals and many plant species can easily be hundreds-fold more resource intensive than resequencing the same genome. However, fungal genomes often are compact and contain little repetitive DNA. Therefore, next-generation sequencers, such as the 454 and the Solexa are suitable for de novo sequencing of fungal genomes (usually ranging from 20 to 60 Mb). Various techniques have been designed to overcome the difficulties of reassembling the sequence of a full genome from the snippets of sequences represented by sequence reads. The most important one is the design of assembly programs ("assembly engines") and the algorithms they deploy to overlap

short reads into contigs. Sequencing both ends of DNA fragments as "paired end" and "mate pair" also is important for extending the length of sequence contigs as well as linking adjacent contigs lacking sequence bridges into scaffolds. In addition, a host of other methodologies, including restriction enzyme mapping, optical mapping, and even fluorescent microscopy can be used to assist genome assembly.

3.2. Resequencing for Comparative Analysis

Given a reference sequence to map reads to, the need for long sequence reads is greatly diminished in resequencing projects. Considering the size and simplicity of fungal genomes, all next-generation sequencing technologies are suitable for resequencing. To date, over 70 fungi have been sequenced. Many of them are important plant or animal pathogens. Some of them are representative or model species of major taxonomic groups. However, except for a few of them (12), most of these sequenced fungal species are not close enough for any meaningful comparative genomics analysis. With the amount of sequence data that can be generated, multiple strains can be sequenced in a single instrument run of the SOLiD or Solexa sequencers. For example, with the SOLiD 3.5 chemistry, a conservative calculation is that $1,600 is sufficient to generate 320 million bases of raw sequence for each fungal strain if eight strains are sequenced at a single run at, for example, Purdue Genome Core Facility. That is equivalent to 30× coverage for each strain, assuming the genome size of sequenced species is 40 Mb.

Resequencing of closely related species or different strains of one species (such as subspecies or different races) is useful to determine genetic mechanisms underlying the differentiation of these species or strains. For example, genes involved the production of specific myco- or phytotoxins, race specificity determination, regulation of sexual or asexual reproduction, and colonization of a particular host or environmental niche may be identified by comparative analysis of genome sequences of these closely related fungi. However, one practical concern for fungal biologists is that data generated by next-generation sequencers could be overwhelming. While sequencing cost continues to decrease, it is likely that fungal biologists have to spend more resources on managing and analyzing actual sequence data.

3.3. Transcriptome Sequencing

Transcriptome de novo sequencing differs in two major ways from genome sequencing. First, RNA transcripts are orders of magnitude shorter than the chromosomes from which they are transcribed. This greatly simplifies assembly as the sequence reads approach the lengths of short transcripts. Second, as transcript abundance implies gene activity, determining the relative abundance of various transcripts at a given time, tissue or set of environmental condition is often an early step in decoding

biological mechanisms. Deep sequencing of transcripts present in a specific fungal tissue by next-generation sequencers can be used to determine the sequences of transcribed genes (de novo transcriptome determination) and the relative abundance of each transcript type (digital gene expression analysis). The former may require the more expensive long read methodologies, whereas the latter may not. In comparison with microarray analysis, next-generation sequencing is more suitable for identifying and profiling low abundance transcripts (3). With the decrease in sequencing cost and the development of software suitable for analyzing next-generation sequences, it is foreseeable that most applications of microarray analysis and other conventional gene expression profiling methods can be replaced by next-generation sequencing. For example, using Roche GS-FLX to assay ten total RNA samples in a single run at Purdue costs approximately $14,000. For each sample, 50–100 thousand reads and 20–40 million bases of sequence are obtained.

Although it is not necessary for most of fungi, methods such as "high Cot" (13) can be deployed to remove highly repetitive sequences in the genome before sequencing. Similar methods can be used to deplete cDNAs of highly expressed transcripts by normalization (7). There are also methods to target specific chromosomal regions for sequencing. They normally involve detection and capture of DNA fragments from chromosomal regions by hybridization to oligonucleotides immobilized on the surface of beads or glass substrates (14). In addition, next-generation sequencers can be efficiently used to determine the sequences of DNAs bound to or recognized by a specific protein or protein complex, such as ChIP-seq (15) and DNase-seq (16).

References

1. Smith LM, Sanders JZ, Kaiser RJ, Hughes P, Dodd C, et al. (1986) Fluorescence detection in automated DNA sequence analysis. *Nature* **321**, 674–679.
2. Sanger F, Nicklen S, Coulson AR (1977) DNA sequencing with chain-terminating inhibitors. *Proc. Nat. Acad. Sci. USA.* **74**, 5463–5467.
3. Metzker ML (2010) Sequencing technologies - the next generation. *Nature Rev* **11**, 31–46.
4. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380.
5. Ronaghi M, Karamohamed S, Pettersson B, Uhlén M, Nyrén P (1996) Real-Time DNA Sequencing Using Detection of Pyrophosphate Release. *Anal. Biochem.* **242**, 84–89.
6. Cheung F, Win J, Lang JM, Hamilton J, Vuong H, et al. (2008) Analysis of the Pythium ultimum transcriptome using Sanger and Pyrosequencing approaches. *BMC Genomics* **9**, 10.
7. Meyer E, Aglyamova G, Wang S, Buchanan-Carter J, Abrego D, et al. (2009) Sequencing and de novo analysis of a coral larval transcriptome using 454 GSFlx. *BMC Genomics* **10**, 219.
8. Barakat A, DiLoreto DS, Zhang Y, Smith C, Baier K, et al. (2009) Comparison of the transcriptomes of American chestnut (Castanea dentata) and Chinese chestnut (Castanea mollissima) in response to the chestnut blight infection. *BMC Plant Biol.* **9**, 11.
9. Guo J, Xu N, Li Z, Zhang S, Wu J, et al. (2008) Four-color DNA sequencing with 3'-O-modified nucleotide reversible terminators and chemically cleavable fluorescent dideoxynucleotides. *Proc. Nat. Acad. Sci. USA.* **105**, 9145–9150.

10. Shendure J, Porreca GJ, Reppas NB, Lin X, McCutcheon JP, et al. (2005) Accurate multiplex Polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732.
11. Fologea D, Gershow M, Ledden B, McNabb DS, Golovchenko JA, et al. (2005) Detecting single stranded DNA with a solid state nanopore. *Nano Lett.* **5**, 1905–1909.
12. Kellis M, Patterson N, Endrizzi M, Birren B, Lander ES (2004) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature* **423**, 241–254.
13. Yuan YN, SanMiguel PJ, Bennetzen JL (2003) High-Cot sequence analysis of the maize genome. *Plant J* **34**, 249–255.
14. Okou DT, Steinberg KM, Middle C, Cutler DJ, Albert TJ, et al. (2007) Microarray-based genomic selection for high-throughput resequencing. *Nature Meth* **4**, 907–909.
15. Mardis ER (2007) ChIP-seq: welcome to the new frontier. *Nature Meth* **4**, 613–614.
16. Song L, Crawford GE (2010) DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb Protoc doi:10.1101/pdb.prot5384.

Chapter 5

Getting the Most Out of Your Fungal Microarray Data: Two Cost- and Time-Effective Methods

Sandra M. Mathioni, André Beló, Jeffrey P. Townsend, and Nicole M. Donofrio

Abstract

Advances in genome sequencing technologies have facilitated production of a wealth of fungal data; within the last 5 years, experimental costs and labor have diminished, shifting the production bottleneck from genomic data generation to data analysis. Genome sequences and microarrays now exist for many fungi, and transcriptional profiling has been shown to be an efficient way to examine how the entire genome changes in response to many different environments or treatments. Multiple platforms, programs, and protocols exist for analyzing such data, making this task daunting for the bench-based scientist. Furthermore, many existing programs are expensive and require license renewals on a yearly basis for each user in the laboratory. Costs may be prohibitively high for bench-based scientists in academia. Our combined experiences with this kind of analysis have favored two programs, depending upon whether the scientist is working with single- or dual-channel hybridization data. Our protocols are aimed toward helping the bench-based PI get the most possible information from their data, without the need for expensive software or an experienced bioinformaticist.

Key words: Microarrays, Limma, BAGEL, Data analysis, Fungi

1. Introduction

1.1. Analyzing Microarray Data: An (Over?) Abundance of Options

Microarray technology has been widely used over the past decade to measure large-scale gene expression in many organisms, including fungi (1, 2). In addition to microarrays fabricated by academics, platforms featuring both single- and dual-channel functionality are commercially available, such as Agilent, Affymetrix, and Nimblegen. Data analysis of the different microarray platforms can be summarized by four main steps: (1) experimental design, (2) normalization and scaling, (3) hypothesis tests, and (4) validation.

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722, DOI 10.1007/978-1-61779-040-9_5,

Experimental design involves the formulation of hypotheses that will be tested and requires determination of the number of replicates required for assaying each treatment, or in the case of two-color arrays, replicates for assaying each pair of treatments. Deciding on an experimental design requires matching the level of detail of desired gene expression measurement against the cost of extra replicates. More replicates with either platform will reveal ever more subtle changes in expression among treatments (3). For experiments using dual-color hybridizations, most well-regarded designs test several treatments against one reference sample or use a loop design, in which each sample is hybridized against the next until many or all samples have been directly compared. A good pair of rules of thumb is that treatments whose differences are directly of interest should be compared directly, and that the degree of transitive inference (number of treatments through which other treatments are compared) should be minimized. Direct comparisons and cross-loop hybridizations will improve experimental acuity, allowing statistically significant inference of smaller differences in gene expression, as measured by post hoc analyses such as the GEL50 (gene expression level at which there is an empirical 50% probability of detecting statistical significance (3)).

After experimental hybridizations are complete, data normalization and scaling aim to correct for systematic sources of variation in the data. Such variation could arise from the use of different dyes, between samples hybridized to different arrays; reagents from different batches; slides not processed in the same day; etc. Accounting for all these effects may appear daunting, as there are many microarray data analysis methods available, and no consensus as to which method is the most robust (4). However, there are numerous software packages available for microarray data analysis that will provide normalizations along with various statistical analyses, including Rosetta Resolver, Genespring, and Limma. There is also a free normalization software for two-color arrays implemented as a web tool in the active experiment Filamentous Fungal Gene Expression Database (FFGED, (5)). For several software options, a yearly license is usually required (Rosetta Resolver and Genespring), which can be expensive for small laboratories. We have typically used Limma for background correction and normalization within and between slides (see below).

Statistical tests are recommended to identify the differentially expressed genes between treatments. The most used tests are *t*-test, ANOVA, and regression analysis. These tests will compare the signal intensity of each sample and their replicates, and output some statistics such as the value of the test and a *P*-value. The *P*-value corresponds to the probability that the observed difference in intensity is caused by chance rather than by an actual difference in expression. For this calculation, more replicates performed in the experiment will provide more statistical power to detect differentially expressed genes. There is no consensus about

which statistical test is the most appropriate, and different softwares implement different tests. Considering a transcript as differentially expressed between treatments should be based on a statistical test; looking only at the fold change between treatments is insufficient (4). We generally use linear models (implemented by Limma) and Bayesian statistics (implemented in BAGEL) to identify differentially expressed genes.

The final step in an experiment is the validation of the microarrays, which is usually done at two levels – in the laboratory and in silico. Again, there is no consensus about the best method to choose or when the researcher can feel confident that their data have, indeed, been validated (4). The criteria for microarray validation are dependent upon the objectives of the specific experiments. Usually, laboratory validation is performed with lower throughput on ten or more genes whose expression changes (or stasis) are key to the conclusions reached. A percentage of validation may be reported. The technique that has been most used for microarray validation is real-time quantitative reverse transcriptase PCR (real-time qRT-PCR). Validation in silico is really the use of several checkpoints during the data analysis to ensure that there are no errors in the data and that the steps performed are having the desired effect. A careful look at the data and the output of the data analysis can often reveal the presence of systematic errors that have to be accounted for. The use of certain types of graphs proves helpful in these tasks; perhaps the best check on data with microarray experiments is biological consistency across gene sets or pathways known to be co-regulated. Sets of related genes or pathways are unlikely to change expression in co-ordinated ways based on measurement errors.

Although there is no consensus among the options that are available for microarray data analysis, there is a consensus about the need for biological replication, normalization, and use of statistical tests to define the significance of the differentially expressed genes (4).

1.2. Limma: Linear Models for Microarray Data Analysis

Limma is an open-source and free software available as an R package (6) in the Bioconductor project (7) for analyzing gene expression data generated from microarray experiments (8–12). Limma uses linear models to detect variation in gene expression and provides the ability to analyze many samples simultaneously. Using Limma for two-color spotted arrays, we were able to assess the quality of array hybridizations, perform normalizations, and explore the data analysis output. The linear models and differential expression functions used in Limma and shown below may be applied to all microarray platforms, including Affymetrix, along with dual- or single-channel microarray experiments (some manipulations in R are necessary when analyzing single channel arrays). Limma runs within the R computing environment, which is an open-source command-driven statistical and data analysis software and language.

The advantages of Limma are twofold: (1) Limma is open-source software, available for Windows, Mac, and Linux users;

and (2) the complete analysis of microarrays can be done using Limma and R, from background correction to graphical representations. The only disadvantage that we have been able to discern is that the R computing environment is command driven, requiring some time and training to learn for those who are not already familiar with it, although graphical user interfaces for Affymetrix single-channel and others dual-channel platforms are also available for Limma (13, 14).

In order to use Limma, one must install the R program and the Bioconductor packages. Detailed instructions about installation are available in their respective websites (http://www.r-project.org; http://www.bioconductor.org). A basic knowledge about how to use R is also required, and although it is relatively easy to use, any inexperienced user will likely have to find out what the most relevant functions are and how to use them. We note that the R documentation, especially "An introduction to R" (http://www.r-project.org, in Documentation, Manuals), is a good overview of R and how to get started with it. After reading it, the user should be able to read specific documentation for each function or package.

1.3. BAGEL: Bayesian Analysis of Gene Expression Levels

BAGEL software is freely available for academic use (http://www.yale.edu/townsend/software.html, (15)). It may be used to analyze two-color arrays, and runs in an interactive text mode requiring very little expertise on the part of the inexperienced user. Within-array normalization of the image-analysis output file must be performed prior to BAGEL analysis, but can be accomplished with any appropriate method, including easy normalization by the normalization web tool provided by the FFGED (5). BAGEL requires a tab-delimited text-based input file that is straightforward to assemble. The input file may also be assembled by an automated web tool provided by the FFGED.

A major feature of BAGEL is its Bayesian Markov chain Monte Carlo (MCMC) algorithm, which provides seamless analysis of diverse well-replicated experimental designs, requiring only the naming of the treatments by the user above the normalized expression in the tab-delimited input file. The input file can be easily prepared in any spreadsheet program, and the output file containing expression levels, confidence intervals and *P*-values, is constructed to facilitate easy production of figures and tables via typical spreadsheet tools as well.

2. Methods

2.1. Using Limma to Analyze Your Data

We study global gene changes in the rice blast fungus, *Magnaporthe oryzae*. The example below is from a microarray dataset generated using the Agilent platform in order to gain

a better understanding of gene expression during nutrient starvation conditions. The fungus was grown in complete medium, which we refer as wild type (WT), and in minimal medium lacking either a carbon source (MMwoC) or a nitrogen source (MMwoN). Total RNA was extracted and used for whole genome expression profiling using microarrays (*M. oryzae* 4x44K version 2 slides from Agilent). The three samples were labeled with Cy3 and Cy5 dyes and used for two-color (channel) hybridizations. The samples grown on nutrient-limited media were hybridized together with the WT sample. No biological replicates per se were performed, though three biological replicates were pooled together for array hybridization. The dye-swaps (hybridizations with different dyes) were considered replicates to compare the expression profile of the two treatments with nutrient limitation to that of the WT. The protocol listed below was based on "LIMMA User's Guide.pdf" from October 22, 2008, with modifications specific for our fungal gene analysis. Each step of the protocol lists its rationale, as well as the command lines that can be copied and pasted directly. Several steps are optional, and these are denoted as such.

1. Download and install the R software (http://www.r-project.org). After opening R, issue the following commands to install Bioconductor, which includes Limma. Also install the package `statmod`, which will be used later for some statistical calculations:

```
source('http://www.bioconductor.org/biocLite.R')
biocLite()
install.packages('statmod', dependencies = TRUE)
```

2. Next, we load Limma and `statmod`, and create a file that lists the RNA targets hybridized to each channel of each array (Table 1). The default name for this file is "Targets.txt." It will contain a FileName column (the image analysis output

Table 1
Format and content of the hybridization file Targets.txt

SlideNumber	Name	FileName	Cy3	Cy5
1	MMwoC_WT	MMwoC_WT.txt	MMwoC	WT
2	WT_MMwoC	WT_MMwoC.txt	WT	MMwoC
3	MMwoN_WT	MMwoN_WT.txt	MMwoN	WT
4	WT_MMwoN	WT_MMwoN.txt	WT	MMwoN

file name), a Cy3 column (RNA labeled with Cy3 dye for that array), and a Cy5 column (RNA labeled with Cy5 dye for that array). It can be created using a spreadsheet or any text editor provided as it is saved as tab-delimited file format.

```
library(limma)
library(statmod)
targets <- readTargets('Targets.txt')
```

3. Load the microarray data files into a Limma's `RGlist` object using a filter to remove spots with high background (e.g., background intensity > 100). This filter is optional; however, it might be important to remove spots with high background. The same structure could be used to filter the data using different criteria, as explained in the Limma user's guide. The microarray data to be loaded should be in text files created by a scanner and contain the signal intensity and other information for each spot in the microarray slide. Limma should be able to read the most common formats including Agilent, Affymetrix, and Nimblegen:

```
bg_filter <- function(x) {
    bg_green_ok <- as.numeric(x$gProcessedBackground < 100)
    bg_red_red_ok <- as.numeric(x$rProcessedBackground < 100)
    as.numeric(bg_green_ok & bg_red_red_ok)
}
rg <- read.maimages(targets, source='agilent', wt.fun=bg_filter)
```

The statement `rg <- read.maimages(targets, source='agilent')` would do the same without using the above filter function. Here, we chose `rg` as the name of the Limma's `RGlist` object. The loaded data can be visualized by using different functions. `show(rg)`, `print(rg)`, or simply `rg` will print the top five rows of the object `rg`. Its class and all other R object's classes can be identified by using the function `class()` (e.g., `class(rg)` will return `RGList` and show that this is an object of the package Limma). The function `dim(rg)` can be used to obtain the size of the data. Individual components or data subsets of the object can be identified with the function `attributes(rg)`, and visualized or used by accessing the RGlist attributes with the `$` operand:

`rg$source` will print the type of slide
`rg$G` will print the intensities of the green (Cy3) channel
`rg$R` will print the intensities of the green (Cy5) channel
`rg$Gb` will print the intensities of the green (Cy3) channel background
`rg$Rb` will print the intensities of the green (Cy5) channel background
`rg$genes` will print the gene IDs and descriptions

4. (Optional) The correlation between intensities of the same treatment in different slides in either the same or distinct channels can be used as a quality control to check if all samples were properly prepared and hybridized to the microarrays,

Table 2
Correlation between signal intensities of samples labeled with different dyes and hybridized to different slides

Slides	^{g}MMwoC_WT	^{g}WT_MMwoC	^{g}MMwoN_WT	^{g}WT_MMwoN
MMwoC_WT^{r}	0.716	0.968	0.587	0.947
WT_MMwoC^{r}	0.950	0.664	0.744	0.692
MMwoN_WT^{r}	0.734	0.933	0.646	0.943
WT_MMwoN^{r}	0.803	0.664	0.901	0.680

Light-gray cells show correlation between WT treatments; *dark-gray cells* show correlations between the two MMwoC and between the two MMwoN treatments; ^{r}: sample in the red channel labeled with Cy5; ^{g}: sample in the green channel labeled with Cy3

and no errors were performed during these processes. The following code calculates the correlations between the two channels of each sample against the two channels of every other sample, and exports a text file with these correlations (Table 2):

```
sample_green <- character()
sample_red <- character()
correlations <- numeric()
for (g in 1:(dim(rg$G)[2])) {
    sample_green <- c(sample_green, paste('[G]',
        colnames(rg$G)[g], sep=''))
    sample_red <- c(sample_red, paste(colnames(rg$G)[g],
        '[R]', sep=''))
    for (r in 1:(dim(rg$R)[2])) {
    correlations <- c(correlations, cor(rg$G[, g],
        rg$R[, r], use='pairwise.complete.obs'))
    }
}
correlation_matrix <- matrix(correlations, nrow=dim(rg$G)[2],
    ncol=dim(rg$R)[2])
colnames(correlation_matrix) <- sample_green
rownames(correlation_matrix) <- sample_red
write.table(correlation_matrix, file='correlations.txt',
    sep='\t', quote=FALSE, row.names=TRUE, col.names=NA)
```

5. (Optional) This step is another quality check on the data. It plots the background of each channel for inspection and a boxplot of the backgrounds in all slides for comparison. The first block of code is needed for Agilent arrays because the number of spots does not correspond to a rectangular format. A solution is to fill in the spots needed to create a rectangle and then plot them as missing data:

```
# create dummy spots for Agilent arrays
Block <- 1
colnames(rg$genes)[colnames(rg$genes) == 'Col'] <- 'Column'
rg$genes <- cbind(rg$genes, Block)
```

```
rg$printer <- getLayout(rg$genes)
number_of_rows <- max(rg$genes$Row)
number_of_columns <- max(rg$genes$Col)
# plot background

png(file='bg_%02d.png')
for (i in 1:dim(rg)[2]) {
    array_name <- colnames(rg$R)[i]
    array_data <- rep(NA, number_of_rows * number_of_columns)
    index <- (rg$genes$Row - 1) * number_of_columns + rg$genes$Col
    array_data[index] <- log2(rg$Rb[,i])
    imageplot(array_data, rg$printer, low='white', high='red',
        main=paste(array_name, '_Cy5', sep=''))
array_data[index] <- log2(rg$Gb[,i])
imageplot(array_data, rg$printer, low='white', high='green',
        main=paste(array_name, '_Cy3', sep=''))
}
dev.off()

# plot boxplot of backgrounds
png(file='boxplot.png', height=1000, width=600, res=96)
oldpar <- par()
par(mfrow=c(2,1), mar=c(3,3,1,1))
boxplot(data.frame(log2(rg$Rb)), main='Cy5 background')
boxplot(data.frame(log2(rg$Gb)), main='Cy3 background')
par(oldpar)
dev.off()
```

6. (Optional) Load the information about the spots and set the spot types (Table 3). This operation is useful in slides that contain probes from different organisms and/or probes corresponding to hybridization controls, such as the 4x44K slide of *Magnaporthe* from Agilent, which contain probes that hybridize to *Magnaporthe* and probes that hybridize to rice. The subsets of probes could, for example, be useful to analyze only probes for one of the species present in the microarray when infected or inoculated samples are compared to noninfected or non-inoculated samples:

```
spottypes <- readSpotTypes('SpotTypes.txt')
rg$genes$Status <- controlStatus(spottypes, rg)
    or:
spottypes_mg_rice <- readSpotTypes('SpotTypes_mg_rice.txt')
rg$genes$Status <- controlStatus(spottypes_mg_rice, rg)
```

7. (Optional). MA plots and density plots provide a good way to check the relationship between ratio and intensity for each probe, and the distribution of the intensities in different channels of different slides, respectively. These graphs can be created from `MAlist` and `RGList` objects. For example, they could be created following each step of the normalization to see what the changes in the data are:

```
plotMA3by2(rg, prefix='rg_Magnaporthe_rice', cex.main=2,
       cex.lab=2, cex.axis=2)
png(file='densities_rg.png', res=96)
plotDensities(rg)
dev.off()
```

Table 3
Different probe annotation is used to classify the spots present in the slide

SpotType	ControlType	ProbeName	Color
file "SpotTypes.txt"			
Samples	0	*	Black
Positive	1	*	Green
Negative	−1	*	Red
file "SpotTypes_mg_rice.txt"			
Mg_probes	0	A_98*	Black
Rice_probes	0	A_97*	Green

8. The background correction can prove useful in improving the quality of the data as background can be detrimental by adding variation. This will mostly depend on the type of scanner used to read the intensity of each spot in the microarrays and the type of preprocessing analysis performed by the scanner program or any other program that handles the data. For Agilent slides, background corrections are usually automatically performed by a program called Feature Extractor, which reads the image from the scanner, identifies the spots, and exports a file with the intensities and background intensities for each spot.

```
rg_bg <- backgroundCorrect(rg, method='subtract', printer=rg$printer)
```

   Any positive or negative probes/spots might be removed as well as spots from other species, when present and not of interest: `rg_bg <- rg_bg[grep("A_98.*", rg_bg$genes$ProbeName),]`.

9. Signal normalization. Normalization is usually performed within (effect of different fluorophores) and between arrays (systematic variations between combinations of samples and arrays). Some options about normalization can be set as arguments to the functions `normalizeWithinArrays()` and `normalizeBetweenArrays()`. Detailed options can be obtained in R with `?function` or in "LIMMA User's Guide.pdf."

```
ma <- normalizeWithinArrays(rg_bg, method='loess')
ma_n <- normalizeBetweenArrays(ma, method="Aquantile")
```

10. (Optional). Combining duplicated spots with the same probe sequence. When duplicated spots contain the same probe sequence in a microarray, it can be used by Limma during the estimation of the linear models. To achieve this in the Agilent array, in which the identical probes are printed in different

positions along the slide, first, the probes have to be sorted by name, and then, the ratios averaged:

```
ma_ns <- ma_n[orde.r(ma_n$genes$ProbeName),]
```

11. (Optional). After the normalization, the ratios for each probe in each array can be exported as an input to BAGEL:

```
ma_ns_a <- avedups(ma_ns, ndups=2, spacing=1, weights=ma_ns$weights)
ma_ns_ar <- cbind(ma_ns_a$genes$ProbeName, ma_ns_a$genes$GeneName,
      2 ^ ma_ns_a$M)
ma_ns_ar <- rbind(c(NA, 'Cy5', targets$Cy5),
      c(NA, 'Cy3', targets$Cy3), ma_ns_ar)
write.table(ma_ns_ar, file='input_bagel.txt', sep='\t',
      row.names=FALSE, quote=FALSE)
```

12. Hypothesis to be tested between treatments. The following commands will create an experimental design for the microarrays loaded at the beginning and will allow the test of different hypotheses about the gene expression of the treatments. Detailed discussions of model designs can be found in Limma's documentation. Creation of experimental design:

```
design <- modelMatrix(targets, ref='WT')
print(design)
```

The following commands will estimate the correlation between intensities of different spots that contain the same probe sequence and estimate the linear models:

```
corFit <- duplicateCorrelation(ma_ns, design, ndups=2, spacing=1)
print(corFit$consensus)
```

```
fit <- lmFit(ma_ns, design, ref='WT', ndups=2, spacing=1,
cor=corFit$consensus)
```

13. Now we can test different contrasts (comparisons) of treatments in relation to their gene expression. For this, we need to update the linear model with the desired contrasts. In this case, we are testing the comparisons of the two treatments with nutrient limitation in relation to the complete media, that is, MMwoC and MMwoN in relation to the reference treatment as assigned in the previous statement.

```
contrast_matrix <- makeContrasts(MMwoC, MMwoN, levels=design)
print(contrast_matrix)
fit_contrast <- contrasts.fit(fit, contrast_matrix)
fit_contrast <- eBayes(fit_contrast)
```

At this point, we are ready to test a specific hypothesis. Examples of hypotheses would be "what probes (genes) are differentially expressed in each treatment in relation to the reference treatment?" or "what are the probes (genes) differentially expressed in at least one of the comparisons to the reference treatment?" The following code tests the previous hypotheses using a *P*-value of 0.01 and a fold change higher than two (`lcf=1`):

```
# testing the first hypothesis
hyp1 <- decideTests(fit_contrasts, method='separate',
       adjust.method='BH', p.value=0.01, lfc=1)
write.table(summary(hyp1), file='hyp1_summary.txt',
       quote=FALSE, sep='\t', col.names=TRUE, row.names=TRUE)
write.fit(fit1, results=hyp1, file='hyp1.txt', digits=3,
       adjust='BH', method='separate')

# testing the second hypothesis
hyp2 <- decideTests(fit_contrasts, method='global',
       adjust.method='BH', p.value=0.01, lfc=1)
write.table(summary(hyp2), file='hyp2_summary.txt',
       quote=FALSE, sep='\t', col.names=TRUE, row.names=TRUE)
write.fit(fit1, results=hyp2, file='hyp2.txt', digits=3,
       adjust='BH', method='global')
```

14. (Optional). With the previous commands, we have tables with the results in the files "hyp1.txt" and "hyp2.txt." It might be useful to transform the log2 of the fold changes to fold changes in the output file. This can be done for both hypotheses with the following code:

```
temp <- fit_contrasts$coefficients
temp <- 2 ^ temp
for (i in 1:dim(temp)[2]) {
   temp[temp[, i] < 1 & !is.na(temp[, i]), i] <-
       -1/(temp[temp[, i] < 1 & !is.na(temp[, i]), i])
}
fit_contrasts_fc <- fit_contrasts
fit_contrasts_fc$coefficients <- temp
write.fit(fit_contrasts_fc, results=hyp1, file='hyp1_fc.txt',
       digits=3, adjust='BH', method='separate')
write.fit(fit_contrasts_fc, results=hyp2, file='hyp2_fc.txt',
       digits=3, adjust='BH', method='global')
```

The Limma output, as we processed it, corresponds to the files "hyp#_summary.txt", "hyp#.txt", and "hyp#_fc.txt". The file hyp#_summary.txt shows the number of genes upregulated (row 1), with no difference in expression (row 0), and downregulated (row –1) for all the contrasts selected. The file hyp#.txt contains several columns representing the averaged log2 of intensities (A); the log2 of fold changes (columns starting with "Coef"); *t*-test statistics (columns starting with "t"); *P*-values for the *t*-tests (columns starting with "p.value"); adjusted *P*-values for the *t*-test (columns starting with "p-value.adj"); *F*-test of the model; *P*-value for the *F*-test; the result of the contrast comparison based on the *F*-test specifying if the probe is upregulated (1), downregulated (–1), or not differentially expressed (columns starting with "Res"); and several columns showing additional information about the probes. The columns showing whether the probes are differentially expressed according to the *F*-test can be used to create subsets of genes that are or are not differentially expressed in specific combinations of treatments. An

alternative option is to use the Limma function `TopTable()` to select the top differentially expressed genes for each treatment.

2.2. Using BAGEL to Analyze Your Data

1. Download the BAGEL software (http://bioinfo.towsend.yale.edu). Several practice datasets are included in the download so that you can see the input file format and immediately try running BAGEL without formatting your own. You may place the downloaded folder in your Applications directory, but be aware that you need not type directory information to load your input file if your input files reside in the same folder as the application.
2. Identify one of the practice datasets or create an acceptable input file for BAGEL. Acceptable files are tab-delimited text files with line feed (LF) characters, with three header rows (Table 4). The first row is for the use of the user in labeling the dataset and the arrays used (e.g., with array bar codes, or numbers – however, keep track of your hybridizations). The second and third rows must contain unique names for each experimental expression node and reference expression node, followed by any number of data rows for each gene of interest.

   By convention, Channel 1 (Experiment) is usually the Cy5 channel, and Channel 2 (Reference) is usually the Cy3 channel, but it is important to note that BAGEL makes no such distinction between experiment and reference; as long as Ratio1 is the normalized ratio of the experiment comparing Exp1/Ref1, where Exp1 and Ref1 are your preferred text names for the treatments you are using, Exp1 or Ref1 may be any treatment or "control" in your experimental design. Treatments often appear in the Channel1 row in one column and the Channel2 column in another row. In fact, they must appear in both rows if fluorophores are flipped in your experimental design or if any control "self-self hybridizations" are

Table 4
Format of the BAGEL input file

[Your Notes]	[Your Notes]	[Label1]	[Label2]	[Label3]	[…]
[Your Notes]	[Channel1]	Exp1	Exp2	Exp3	…
[Your Notes]	[Channel2]	Ref1	Ref2	Ref3	…
ORF1	CommonName1	Ratio1	Ratio2	Ratio3	…
…	…	…	…	…	…

Brackets indicate cells ignored by BAGEL

Table 5
BAGEL input file example based on the *Magnaporthe* experimental design

Donofrio et al.		MMwoC_WT	WT_MMwoC	MMwoN_WT	WT_MMwoC
Nutrient exp.	Cy5	WT	MMwoC	WT	MMwoN
	Cy3	MMwoC	WT	MMwoN	WT
Probe1	Gene1	0.72	1.4	4.2	0.24
…	…	…	…	…	…

performed. Table 5 presents an example of how this template might be filled in for the *Magnaporthe* nitrogen starvation experiment detailed in Subheading 2.1.

3. Execute BAGEL. How you tell your computer to do this varies by platform; for instance, on a Macintosh OS 10.0 or later, you must open the "Terminal" application, type in "cd_" where the underscore is actually a space character, use the mouse to drag the folder containing the BAGEL application and your data files into the new terminal window, press [RETURN], and then type in "./UBAGEL4.1" – or the current version name – [RETURN] for the interactive text mode. For the bioinformatically savvy, it is also possible to use unix flags to pipeline many analyses. See http://www.yale.edu/townsend/Software/BAGELTutorial.html for details.
4. In the interactive text mode, BAGEL will prompt you:

```
Please type the exact name of a text file of microarray
ratio results to analyze:
```

At this prompt, type the name of the input file, including the directory path. For example, in UNIX, directory pathnames look something like

```
/Users/jeff/DOCUMENTS/RESEARCH/Software/BAGEL/Datafilename.txt
```

If your data files are in the same folder as the BAGEL application, then you only need to type the data file name (with any extension such as .txt). Helpful hint: Keep your data file name and your experimental node names short. If the data file name is too long, BAGEL has to truncate it and use a far less intuitive name for your output file.

5. Now you are on easy street. From this point on, you may just press return to verify and accept the default settings, and BAGEL will analyze the data, and if your input file is correctly composed, those settings will work beautifully 99.9% of the time. However, you may want to follow along with the following narration as you might want to use the verification

functions and some more advanced functions of BAGEL. BAGEL will ask you to verify its count of the number of hybridizations and the names of the expression nodes (treatments including controls or references) in your experimental design, based on its reading of your input file. Press RETURN to verify, or press q to quit and correct your input file. You are then presented with a menu of options:

```
Current MCMC settings:
(E)rror Model: Additive errors, estimating/constraining
    Coefficient of Variation terms
(C)onstrained Coefficient of Variation: True
(I)nitial values:
Mu[M1-2] := 1.00    Coefficient of Variation[M1-2] := 0.2000
Mu[M2-8] := 1.00    Coefficient of Variation[M2-8] := 0.2000
Mu[M5-7] := 1.00    Coefficient of Variation[M5-7] := 0.2000
Mu[M7-8] := 1.00    Coefficient of Variation[M7-8] := 0.2000
(M)u step size: 0.50
(V)ariance/CV step size: 0.500
(B)urn in, # generations: 20000
(P)eriod of sampling from the Markov chain: 20
(G)enerations to be sampled: 10000
(F)ull output of the chain: False
(T)uning depth maximum: 8
```

If you have some knowledge of MCMC methods and are interested in modifying the default BAGEL execution, here are brief descriptions of the significance of these options:

`(E)rror Model`: Please read the papers for a description of these models. The model in the original BAGEL paper (15) is Additive errors, estimating/constraining Variance terms. The default model is Additive errors, estimating/constraining Coefficients of Variance (15). All the models that BAGEL implements work well and results are usually quite similar to results with other well-established methods, so just use the default unless you are really into the details.

`(C)onstrained Variance/Coefficient of Variation:` If true, variances/coefficients of variation for all expression nodes are assumed to be the same. With an uninformative prior, you must have as many measurements as expression nodes. If false, variances for all expression nodes are separately estimated. With an uninformative prior, you must have at least twice as many measurements as expression nodes (minus one). Unless your data are very highly replicated, using constrained variance is recommended. With under-replicated data, estimation of variances for each sample is very imprecise and can lead to misleading results. When a design is well replicated, to the extent it has been tested so far, it seems that estimating each variance independently changes accuracy of BAGEL estimates of the expression

levels very little, but costs a great deal in computation time and in precision of estimation.

(I)nitial values: The starting Mu and Sigma Squared parameter values for the Markov Chain. In some applications of the MCMC method, it is very important to try many different initial starting values to ensure that the chain does not get stuck in one region of the state space. This is not an issue with the BAGEL models. With any moderately decent microarray data, BAGEL does not get stuck in local peaks.

(M)u step size: The step size is a very important parameter in terms of how long BAGEL needs to run to achieve good estimates of gene expression. BAGEL automatically tunes the step size for you as long as the (T)uning Depth maximum, below, is greater than one. Note that BAGEL uses information from genes previously analyzed in your dataset to help it guess the right step size.

(V)ariance/CV step size: See (M) above.

(B)urn in, # generations: The default burn-in (20,000 iterations) is rather excessive for most datasets. However, it is nice to feel confident that stationarity in the chain has been reached. Decreasing this parameter will make BAGEL run slightly faster. From a formal Bayesian perspective, it is up to you, the user, to ensure that BAGEL is reaching stationarity, but with the default burn-in, I have never heard of there being any problem in this regard. One way to verify stationarity is to run BAGEL multiple times and check that they are converging on the same results.

(P)eriod of sampling from the Markov chain: This is how many iterations are performed until the current state is sampled from the chain to construct posterior distributions. When the period is greater than one, this sampling is referred to as thinning the Markov chain. Thinning the chain subdues correlations that are present between subsequent states of the chain. Decreasing this period substantially decreases computation time, but compromises the independence of the samples and thereby the adequate mixing of the chain.

(G)enerations to be sampled: Ten thousand generations yield accuracy to about three digits. Increasing the number of generations (iterations) increases the number of digits of accuracy. The product P * G largely determines the amount of time necessary for a BAGEL run for a gene.

(F)ull output of the chain: FALSE. Keep it that way, unless your input file has only one or two genes and you really like BAGEL to talk about everything it is doing.

If you run with Full Output, BAGEL saves the posterior distributions for Mu for every sample. On a genome-wide dataset with lots of nodes, it could fill a small hard drive.

`(T)uning depth maximum:` How hard (in MCMC runs) you wish BAGEL to try to find an optimal step size for a gene. Usually, an optimal step size is discovered in a few tries, and almost always in six or seven. A gene for which no optimal step size is found is marked FALSE under "Acceptable?" in the output file, but the tuning algorithm has been refined to a degree such that this virtually never happens.

6. When you are done changing settings, press RETURN and BAGEL will begin to work on your dataset.

   Depending on your dataset and the speed of your computer, BAGEL can take a long time to run, say, a minute per gene or more. Thus, it is frequently convenient to set it going on a computer you will not need for the night, and leave it alone.

   BAGEL output is a tab-delimited text file with estimates for each expression node, additions for 95% upper-bounds, and subtractions for 95% lower-bounds. These are formatted such that creation of an Excel column or bar graph should be very easy.

   Other columns let you know of the Mu and Variance/CV step acceptance rate as well as an acceptable column, which discloses whether BAGEL has found acceptable acceptance rates (between 0.15 and 0.5) for both parameters. Lastly, *P*-values for whether expression level is greater in one sample than another occupy a number of columns. You should be aware that you should have the appropriate scientific skepticism and look carefully for corroborating biologically consistent evidence. The results text file is on the same drive in the same folder as your input file. The output file name will be the same as the original Data file name, but will have the characters .BAR appended.

3. Conclusions

We have provided two methods routinely used by our laboratories for handling transcriptomic data from fungal species. Limma and BAGEL are freely available, downloadable software packages; Limma is an open-source, command-based software that runs under the R package and, while requiring a bit of a learning curve, can be utilized with success by an inexperienced user following our protocols listed above. Limma provides robust data

analysis dual- or single-channel microarray experiments. BAGEL also provides robust datasets from dual-channel array experiments and runs on a web-based interface that is easily utilizable for bench-based scientists with limited to no experience in bioinformatics.

References

1. Breakspear A., and Momany M. (2007) The first fifty microarray studies in filamentous fungi *Microbiology* **153**, 7–15.
2. Kasuga T., and Glass N. L. (2008) Dissecting colony development of *Neurospora crassa* using mRNA profiling and comparative genomics approaches. *Eukaryotic Cell* **7**, 1549–1564.
3. Townsend J. P. (2004) Resolution of large and small differences in gene expression using models for the Bayesian analysis of gene expression levels and spotted DNA microarrays *BMC Bioinformatics* **5**, 54.
4. Allison D. B., Cui X., Page G. P., and Sabripour M. (2006) Microarray data analysis: from disarray to consolidation and consensus. *Nat Rev Genet* **7**, 55–65.
5. Zhang Z., and Townsend J. P. (2010) The filamentous fungal gene expression database (FFGED). *Fungal Gen Biol* **47**, 199–204.
6. R: A Language and Environment for Statistical Computing; R Development Core Team; 2009. http://cran.r-project.org/doc/manuals/refman.pdf.
7. Gentleman R., Carey V., Bates D., Bolstad B., Dettling M. Dudoit S., Ellis B,. Gautier L., Ge Y., and Gentry J. (2004) Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol* **5**, R80.
8. Ritchie M. E., Diyagama D., Neilson J., van Laar R., Dobrovic A., Holloway A., and Smyth G. K. (2006) Empirical array quality weights in the analysis of microarray data *BMC Bioinformatics* 7, 261.
9. Ritchie M. E., Silver J., Oshlack A., Holmes M., Diyagama D., Holloway A., and Smyth G. K. (2007) A comparison of background correction methods for two-colour microarrays *Bioinformatics* **23**, 2700–2707.
10. Smyth G. K. (2004) Linear models and empirical bayes methods for assessing differential expression in microarray experiments *Stat Appl Genet Mol Biol* **3**, Article3.
11. Smyth G. K., Michaud J., and Scott H. S. (2005) Use of within-array replicate spots for assessing differential expression in microarray experiments *Bioinformatics* **21**, 2067–2075.
12. Smyth G. K., and Speed T (2003) Normalization of cDNA microarray data *Methods* **31**, 265–273.
13. Wettenhall J. M., Simpson K. M., Satterley K., and Smyth G. K. (2006) affylmGUI: a graphical user interface for linear modeling of single channel microarray data *Bioinformatics* **22**, 897–899.
14. Wettenhall J. M., and Smyth G. K. (2004) limmaGUI: a graphical user interface for linear modeling of microarray data *Bioinformatics* **20**, 3705–3706.
15. Townsend J. P., and Hartl D. L. (2002) Bayesian analysis of gene expression levels: statistical quantification of relative mRNA level across multiple strains or treatments. *Genome Biol* **3**, 12.

Chapter 6

Fusarium graminearum from Expression Analysis to Functional Assays

Heather E. Hallen-Adams, Brad L. Cavinder, and Frances Trail

Abstract

Fusarium graminearum, the causal agent of head blight of wheat, was the third filamentous fungus to have a completed genome sequence. Since the release of the genome sequence in 2003, *F. graminearum* has become a model for studies of genomics and transcriptomics, mycotoxins, fungal population genetics, gene function, and sexual development. Herein we present the techniques we have used in our laboratory to perform expression analyses on life cycle stages of *F. graminearum* and techniques to functionally characterize those genes identified as potentially interesting.

Key words: Affymetrix, *Gibberella zeae*, Microarrays, RNA extraction, Protoplasts, Transcriptomics, Transformation

1. Introduction

Fusarium graminearum (sexual state *Gibberella zeae*) is a serious pathogen of small grains, causing head blight of wheat and barley, and stalk rot of maize; it is also capable of infecting several other cereal crops. In addition to preharvest losses, *F. graminearum* produces mycotoxins (notably the trichothecene toxin deoxynivalenol [DON] and its derivatives, and the estrogenic mycotoxin zearalenone) in stored grains that may render a sizeable portion of the harvested crop unfit for consumption. *F. graminearum* was among the first filamentous fungi to have its genome sequenced, with 36.45 MB encoding more than 13,000 predicted genes. To date (December 2010), it arguably possesses the most thorough annotation of any filamentous fungal genome (1). The sequenced strain, Michigan field isolate PH-1 (2), can be induced easily to produce synchronous sexual development, making developmental time courses straightforward (3). Transformation is efficient,

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722, DOI 10.1007/978-1-61779-040-9_6,

and the availability of nitrate reductase (NIT) mutants provides an endogenous marker for controlled crosses (4). Protocols for inoculation of wheat and barley under controlled conditions are established (5, 6). Arabidopsis can also be infected successfully (7), although the progress of the disease differs from that of infected grains. Recently, work has begun on *Brachypodium* as a model system for host–pathogen interactions (8). While this work is still in its infancy, the small genome of *Brachypodium* and its small size and relative genetic tractability (compared to the hexaploid wheat) make it potentially a very valuable system.

An Affymetrix GeneChip is available based on the sequenced strain (9). Published studies using the Fusarium GeneChip include infection time courses (6, 9, 10), developmental studies (3, 11, 12), and mutant vs. wild-type comparisons (13). The availability of Affymetrix GeneChips for the major host plants wheat, barley (14), maize, and rice makes possible the dissection of plant–pathogen interactions within the same platform.

2. Materials

For all protocols, it will be assumed that the reader possesses standard laboratory materials and equipment, including a microcentrifuge, centrifuge capable of handling 50- and 250-ml tubes, vortex mixer, spectrophotometer, balance, heat blocks and/or water baths capable of being heated to different temperatures, Eppendorf tubes, glassware, etc. Our autoclave parameters are 121°C, >20 psi.

2.1. Cultures and Growth Media

2.1.1. Fungal Strains

Our procedures have been optimized for the sequenced strain of *Fusarium graminearum*, Michigan field isolate PH-1 (FGSG 9075; NRRL 31084). Some techniques, such as RNA extraction and statistical analysis of microarrays, are applicable across a wide variety of strains and species. Others, such as the sexual development time course, may require further optimization for specific strains.

2.1.2. Sporulation Media

1. Bilays (15): dissolve 1 g KH_2PO_4, 1 g KNO_3, 0.5 g $MgSO_4 \cdot 7H_2O$, 0.5 g KCl, 0.2 g starch powder, 0.2 g glucose, 0.2 g sucrose, 15 g agar in 1 l water. Autoclave for 20 min. Pour in 60 mm diameter petri plates.
2. Carboxymethyl cellulose medium (CMC) (16): dissolve 15 g of carboxymethyl cellulose sodium salt (Sigma-Aldrich, St. Louis, MO), 1 g NH_4NO_3, 1 g KH_2PO_4, 0.5 g $MgSO_4 \cdot 7H_2O$, and 1 g yeast extract in 1 l water (carboxymethyl cellulose dissolves slowly, and will need to be heated and stirred). Aliquot 100 ml apiece into 250-ml flasks; autoclave for 20 min.

3. Miracloth (Calbiochem, La Jolla, CA).
4. Sterile 250-ml centrifuge tubes (one per sample).
5. 35% Glycerol, autoclaved.
6. Hemocytometer.

2.1.3. Sexual Development Media

1. Carrot agar: Chop 350 g carrots into 2 cm pieces. Add 400 ml of water. Autoclave for 30 min in a 2-l beaker. Decant into blender. Blend at low, medium, and finally high speeds. Transfer to flasks, rinsing blender into flasks and bringing total volume to 1 l. Add 20 g of agar. Autoclave for 35 min. Pour into 60 mm diameter petri plates.
2. *Neurospora* Synthetic Crossing medium (SC; (17)) Dissolve 3 g KNO_3, 2.1 g K_2HPO_4 (anhydrous), 1.5 g KH_2PO_4 (anhydrous), 0.3 g NaCl, 0.3 g $CaCl_2 \cdot 2H_2O$ (dissolved separately and added in), 0.3 ml trace element stock solution in 1 l water. The pH is about 6.5. Add 10 g of sucrose and 20 g of agar. Autoclave and pour plates.
3. Trace element stock solution: Dissolve 5 g of citric acid $\cdot 1H_2O$, 5 g $ZnSO_4 \cdot 7H_2O$, 1 g $Fe(NH_4)_2(SO_4)_2 \cdot 6H_2O$, 0.25 g $CuSO_4 \cdot 5H_2O$, 0.05 g $MnSO_4 \cdot 1H_2O$, 0.05 g anhydrous H_3BO_3, and 0.05 g $Na_2MoO_4 \cdot 2H_2O$ successively in 95 ml of distilled water. Add 1 ml of chloroform for storage (18).
4. 2.5% Tween 60: dilute Tween 60 in water. Autoclave.

2.1.4. DNA Extraction

1. YES medium: Dissolve 20 g of Bacto yeast extract (Difco Laboratories, Detroit, MI) and 60 g sucrose in 1 l water. Bring pH to 5.8 with concentrated HCl. Dispense into 100 ml aliquots in 250-ml flasks. Autoclave for 20 min.
2. CTAB extraction buffer: 0.1 M Tris–HCl, pH 7.5; 1% hexadecyltrimethylammonium bromide (CTAB); 0.7 M NaCl; 10 mM EDTA; water to 99 ml. Autoclave and add 1 ml 2-mercaptoethanol once cooled. The CTAB will not dissolve completely until the NaCl is added and the solution is heated to 65°C and stirred.
3. Tris-saturated phenol.
4. Chloroform:isoamyl alcohol 24:1 v:v (chloroform:IAA).
5. Absolute ethanol, chilled to –80°C.
6. Ice cold 70–80% ethanol.
7. RNase, DNase free.
8. Proteinase K, 20 mg/ml.
9. 3 M sodium acetate (NaOAc).
10. Heat blocks at 65 and 37°C.
11. Glass beads, ≤0.5 mm diameter (for tissue disruption).
12. 1.5 ml polypropylene pellet pestle (Kimble/Kontes).

2.1.5. Protoplast Production and Transformation

1. CMC (see Subheading 2.1.2, above).
2. YEPD: 3 g yeast extract (Difco Laboratories, Detroit, MI); 10 g Bacto peptone (Difco Laboratories, Detroit, MI); 20 g dextrose (anhydrous; = D-glucose); dissolve in 1 ml water, aliquot 100 ml apiece into 250-ml flasks and autoclave.
3. STC buffer: 1.2 M sorbitol; 10 mM Tris–HCl, pH 8.0; 50 mM $CaCl_2$; autoclave. Use 4.1 ml per transformation reaction.
4. Sterilized (autoclaved): miracloth disks cut to fit inner diameter of Büchner funnel, glass funnel, Büchner funnel, 250-ml Erlenmeyer, beaker (≥50 ml, depending on how much protoplasting buffer (see below) is prepared), 250 ml flat bottomed centrifuge bottle with screw cap, 50-ml round bottom centrifuge tube with screw cap, 30 µm diameter Nitex membrane (Sefar America, Inc. Kansas City, MO; 03-30/18).
5. Protoplasting buffer: To 20 ml of 1.2 M KCl, add 500 mg Driselase from Basidiomycetes (Sigma Chemical Co., St. Louis; D8037); 1 mg Chitinase from *Streptomyces griseus* (Sigma Chemical Co., St. Louis; C6137); and 100 mg lysing enzyme from *Trichoderma harzianum* (Sigma Chemical Co., St. Louis; L1412); stir for 30 min and filter sterilize through a 0.45-µm Millex-HA filter (Millipore, Bedford, MA).
6. 30% PEG solution: 30% polyethylene glycol (PEG) 8000 (Sigma-Aldrich, St. Louis, MO; P2139); 10 mM Tris–HCl, pH 8.0; 50 mM $CaCl_2$; filter sterilize using a 0.45-µm Millex-HA filter (Millipore, Bedford, MA). Always prepare fresh on the day of use. 2.05 ml is used per reaction.
7. Regeneration medium (RM): 135.5 g sucrose in 500 ml water; heat to dissolve, then add 0.5 g yeast extract; 0.5 g N-Z-Amine AS (Sigma-Aldrich, St. Louis, MO; N4517); and 3.72 g agar; autoclave. Makes approximately twenty 100 mm diameter petri plates – sufficient for two transformations. Keep regeneration medium warm (55–65°C) and liquid until use.
8. RM + Hyg: regeneration medium (recipe as above) amended with 150 µg/ml hygromycin B (Calbiochem-Novabiochem Corp., San Diego, CA) after RM has cooled to 50°C following autoclaving.
9. V8 medium + Hyg: 163 ml V8 Juice; 1 g $CaCO_3$; 15 g agar in 1 l water. Autoclave; add 150 µg/ml hygromycin B (Calbiochem-Novabiochem Corp., San Diego, CA) once medium has cooled to 50°C following autoclaving. Pour into 60 mm diameter petri plates.
10. Hemocytometer.

2.1.6. Plant Materials

1. Wheat: varieties susceptible to Fusarium head blight (FHB) include spring wheat cultivars Norm, Wheaten, and Bobwhite (19). Spring wheat variety Alsen is moderately resistant (20). These varieties are commonly used in wheat-*Fusarium* interaction studies. The sequenced strain (cultivar Chinese Spring) is also moderately resistant.
2. Barley: FHB-susceptible cultivars include Morex and Stander (6, 21). Chevron is resistant (21). Morex has been sequenced.
3. 0.04% Tween 20: dilute Tween 20 in water. Autoclave.
4. Air brush or atomizer.

2.1.7. Long-Term Storage and Culture Maintenance

1. 35% Glycerol.
2. Sterile soil: Fill 10-ml glass screw-top vials halfway with moderately fine-grain soil. Add 200 µl of water. Autoclave for 45 min, cool overnight, and autoclave again.

2.2. RNA Extraction from Culture and Wheat Stems

1. Starting materials: fungal biomass scraped from culture plates during time course or harvested from liquid culture. Stems from infected wheat plants. For culture materials, good results have been consistently obtained from sufficient materials to fill a 2.0-ml Eppendorf tube at least halfway (~ the scrapings of five to ten 60-mm petri plates, depending on developmental stage). For wheat stems, stems at the appropriate infection stage should be sectioned into 1 cm pieces, and ≥20 pieces of the same stage should be pooled and used for one extraction. Flash-freeze samples on harvest (liquid nitrogen or dry ice bath), lyophilize, and store at −80°C until use.
2. TRIzol reagent (Invitrogen, Carlsbad, CA).
3. Chloroform.
4. RNA CTAB: 2% hexadecyltrimethylammonium bromide (CTAB), 2% polyvinylpyrrolidone PVP K30, 100 mM Tris–HCl (pH 8.0), 25 mM EDTA, 2 M NaCl, and 0.5 mg/ml spermidine.
5. Solution of chloroform: isoamyl alcohol (24:1 v/v) (chloroform:IAA).
6. Ethanol at 100% (absolute) and 70% concentrations.
7. 3 M NaOAc.
8. 2-Propanol.
9. Mortars and pestles (one each per sample) baked at ≥180°C for ≥8 h.
10. 30-ml Corex tubes (four per sample) baked at ≥180°C for ≥8 h.
11. Glass beads, ≤0.5 mm diameter (for tissue disruption).

2.3. RNA Extraction from Wheat Kernels

1. Infected wheat kernels. 20 or more kernels (in a single sample) will consistently yield good results. RNA can sometimes be successfully extracted from fewer than ten kernels, but not always in sufficient quantities for further analyses. Flash-freeze kernels at harvest; lyophilize and store at –80°C until use.
2. Extraction buffer: 0.1 M Tris–HCl (pH 8.0), 0.1 M LiCl, 10 mM EDTA, and 1% (v/v) SDS.
3. Tris-saturated phenol.
4. LiCl at 10 and 2 M concentrations.
5. Ethanol at 100% (absolute) and 70% concentrations.
6. 3 M NaOAc.
7. Mortars and pestles (one each per sample) baked at ≥180°C for ≥8 h.
8. 30- and 15-ml Corex tubes (one each per sample) baked at ≥180°C for ≥8 h.
9. Glass beads, ≤0.5 mm diameter (for tissue disruption).

2.4. RNA Cleanup

1. DNase (RNase-free) and 10× DNase buffer (Roche, Mannheim, Germany; 04 716 728 001).
2. 0.2 M EDTA.
3. RNeasy Mini Kit (Qiagen), containing spin columns, Buffer RLT, Buffer RW1 and Buffer RPE (add ethanol to Buffer RPE prior to first use, following manufacturer's instructions).
4. Absolute ethanol.

2.5. Affymetrix Microarray Analyses

1. Fusarium GeneChip (Affymetrix).
2. GeneChip 3′ IVT Express Kit (Affymetrix).
3. Absolute ethanol.
4. Programmable thermocycler.
5. 96-Well plate.
6. Magnetic stand for 96-well plates (Ambion; #AM 10050 or #AM 10027).
7. Orbital shaker for 96-well plates.
8. Spectrophotometer.
9. GeneChip Instrument System and GeneChip Operating Software (Affymetrix) – this step is usually hired out, and not done directly by the researcher.

2.6. Statistical Analyses with R and Bioconductor

1. R – computer software environment available for free download through The R Project for Statistical Computing, <http://www.r-project.org>.

2. Bioconductor – bioinformatics programs that run in and require R (above). Available for free download at <http://www.bioconductor.org>.
3. CEL files – generated by the Affymetrix GeneChip Operating Software, one file per microarray experiment.
4. CDF file – this file allows the software to connect the position of a spot on your array with the identity of the relevant probe. It is generated by the GeneChip Operating Software, and may require some additional manipulation in R before the first use.

3. Methods

3.1. Growing and Maintaining Fusarium graminearum

Wild-type and some mutant strains can be obtained through the Fungal Genetics Stock Center <http://www.fgsc.net/>; the sequenced strain is FGSC 9075. Mutants of interest can be generated through homologous recombination (see Subheading 3.1.5, below).

3.1.1. Use of CMC and Bilays for Spore Stocks

1. Inoculate 100 ml of CMC with a colonized agar plug or 10 µl of conidia or a small portion of a soil stock.
2. Shake at 200 rpm for 3–4 days at RT.
3. Filter through sterile Miracloth into a sterile 250-ml centrifuge tube.
4. Centrifuge at 4,000 × *g* for 5 min.
5. Discard the supernatant.
6. Resuspend the pellet in 1 ml of dH_2O. Pipette into 1.5-ml Eppendorf tubes.
7. Centrifuge at 4,500 × *g* for 5 min.
8. Discard the supernatant. Resuspend in 1 ml of dH_2O and repeat centrifugation. Discard the supernatant.
9. Resuspend in 1 ml of 35% glycerol.
10. Use hemocytometer to quantify. Dilute in 35% glycerol if necessary for a final concentration of 5×10^5 conidia/ml.
11. Store at –80°C.

Optional – If it is difficult to obtain conidia in CMC, or if conidia are needed for single spore isolation, center inoculate one plate of Bilays for each strain. Allow to grow for 2–3 days at RT. Conidia may be isolated as single spores or used to inoculate CMC.

3.1.2. Sexual Development Time Course with Carrot Agar Cultures

1. Inoculate carrot agar with 10 µl of conidia, or agar plug, or a small portion of a soil stock. (If inoculating with conidial stock, conidia can be spread across the plate with a flame-sterilized glass rod to facilitate synchronous development).
2. Grow at RT under continuous white fluorescent lighting 4 days or until the mycelium reaches the edge of the plate, if center inoculated (see Note 1).
3. Use a sterile toothpick or flame sterilized scalpel to remove surface hyphae, scraping down to the agar without disturbing the agar surface. If comparing gene expression over sexual development, transfer aerial hyphae (0 h time point) to a 2-ml Eppendorf tube and store at −80°C (see Note 2).
4. Add 800 µl of 2.5% Tween 60 to the center of the plate. Spread with a flame-sterilized glass rod.
5. Return plates to incubation under lights. Plates should be facing upwards, and incubated in a single layer (i.e., not stacked).

3.1.3. Sexual Development Time Course with Neurospora SC Medium Cultures

1. Inoculate SC plates with 10 µl of conidia, or agar plug, or a small portion of a soil stock.
2. Grow at RT under continuous white fluorescent lighting for 4 days or until the mycelium reaches the edge of the plate, if center inoculated.
3. Use a sterile toothpick or flame sterilized scalpel to remove surface hyphae, scraping down to the agar without disturbing the agar surface (see Note 3).

3.1.4. DNA Extraction

1. Label a 1.5-ml tube for each sample. Add a small amount of glass beads for tissue disruption.
2. Tare each tube and weigh out 15 mg of lyophilized tissue into the tube.
3. Close tube and place in −80°C freezer for 10 min or drop tubes in liquid nitrogen for a minute.
4. Retrieve tube; open and grind mycelia well with pellet pestle.
5. Add 700 ml of CTAB buffer; continue grinding. Mix well by inverting and shaking to ensure that all the mycelia are suspended (turn the tube upside down and slam top on counter if needed). Do not vortex at any step.
6. Incubate at 65°C for 20 min. Invert tubes several times after the first 10 min. Cool on ice (1–2 min).
7. Add 300 ml of phenol, 300 ml of chloroform:IAA (24:1). Mix well by inversion. Spin at high speed for 5 min.
8. Transfer (upper) aqueous phase to new tube.
9. Add 300 ml phenol, 300 ml chloroform:IAA. Mix well by inversion. Incubate for 20 min at room temperature. Spin at high speed for 5 min.

10. Transfer aqueous phase to new tube.
11. Add 500 ml of chloroform. Mix well by inversion. Spin at high speed for 5 min.
12. Transfer the aqueous phase to new tube.
13. Add cold 100% ethanol from a –80°C freezer until tube is full (800–900 µl); invert several times and place in –80°C freezer for 15–30 min (until the solution has thickened but not frozen).
14. Spin down at max speed for 10–15 min in 4°C centrifuge. Pour off supernatant.
15. Wash pellet in 600 µl of ice cold 70–80% ethanol. Spin down at high speed for 1 min.
16. Pour off supernatant. Leave the pellet to dry or place on 65°C heating block for 2 min.
17. Resuspend in 50–100 ml water. (For multiple replicates of the same sample, pool them here and resuspend in smallest amount of water possible) Place on a 65°C heating block with open cap for several minutes to ensure that ethanol has completely evaporated. Also effective with a closed cap to help dissolve the pellet too – although sometimes pellets resuspend better at 37°C.
18. Add 1 µl Rnase, Dnase free enzyme solution per 100 µl of DNA suspension. Incubate at 37°C for at least 30 min.
19. Add 20 µl of proteinase K solution (20 mg/ml) and digest for at least 1 h up to but *not over* 65°C (over 65°C inactivates proteinase K).
20. Add ½ volume of 3 M NaOAC, then add CTAB extraction buffer to reach a volume of 700 µl, invert several times, and follow steps 7–17 as mentioned above. You now have high quality DNA.

3.1.5. Protoplasting and Transformation

1. All experiments should be performed in a Biosafety/sterile hood. Inoculate 100 ml of CMC solution in a 250-ml Erlenmeyer flask with 0.3 g of soil suspension or a pea sized piece of mycelia. Incubate for 72–96 h (up to 6 days) on a rotary shaker table at 25°C at 250 rpm (16). Start two CMC flasks per YEPD flasks in step 3.
2. Filter the culture from the two CMC flasks through a sterile Miracloth in a glass funnel into a 250-ml flat bottom centrifuge bottle. Rinse with sterile dH_2O. Spin at room temperature (RT) at 4,000 × *g* in an appropriate rotor for 10 min.
3. Discard all but 2–4 ml of the supernatant and resuspend conidia. Place conidia in 100 ml of YEPD broth in a 250-ml Erlenmeyer flask, and grow in a rotary shaker for 10–14 h at 25°C at 175 rpm. Timing is critical here as older cultures do not digest well into protoplasts.

4. Filter culture from each YEPD flask, through sterile Miracloth, in a Büchner funnel under vacuum (can use regular glass funnels without vacuum) and collect the mycelial mat. Rinse the mat with sterile dH_2O and allow the water to drain. Place the mat back into the flask from where it was obtained or into a new sterile 250-ml Erlenmeyer flask if needed. Add 30 ml of Protoplasting Buffer to each flask (should have already been prepared and filter sterilized).
5. Digest for 1–2 h on a rotary shaker table at 30°C at 80 rpm. Check for protoplasts after the first 45 min and then after every 15–20 min (see Note 4).
6. Filter the digestion mixture through a 30 μm Nitex nylon membrane (see Note 5) into 50-ml round bottom centrifuge tubes. Filtrate should be turbulent due to the presence of protoplasts. Centrifuge at RT at 3,000 × *g* for 5 min in an appropriate rotor. Protoplasts are very fragile. Treat them gently.
7. Discard supernatant and gently resuspend protoplasts in 10 ml of STC Buffer using wide orifice glass pipettes. Spin the solution at 3,000 × *g* for 5 min.
8. Discard the supernatant and gently resuspend protoplasts in 1 ml of STC Buffer using wide orifice pipette tips. Transfer to a 2-ml tube. Spin in a microcentrifuge at RT at 3,500 × *g* for 5 min. Repeat once.
9. Resuspend protoplasts in a final volume of 300–600 μl. Quantify using a hemocytometer; a good preparation can be expected to yield 10^6–10^8 protoplasts/ml. This is your protoplast suspension. Make the following mixture: 100 μl – protoplast suspension, 100 μl – STC Buffer, 50 μl of 30% PEG Solution, and 10 μl of plasmid (20–50 μg) or PCR product (0.25–0.5 μg). For split-marker vectors (22, 23), mix 6 μl of each partial vector first and then add 10 μl (0.25–0.5 μg each fragment) (see Note 6). It is useful to perform two reactions for each set of vectors (target) and ½ reactions for controls. Extra protoplasts can be frozen for later use: add DMSO to 7% volume, aliquot and freeze at −80°C. When using frozen stocks, spin to collect and resuspend in STC at least twice before use. Start at step 9.
10. Add 2 ml of 30% PEG Solution and incubate for 5 min.
11. Add 4 ml of STC Buffer and gently mix by inversion.
12. Pour reactions into cooled Regeneration Medium (RM): 250 ml RM per transformation reaction. *RM must be cool enough to touch and hold to the inside of your arm else protoplasts will be killed!* If the medium feels hot, allow it to cool more. The agar should be close to solidifying. Mix and pour into plates.

13. Allow protoplasts to regenerate for 12–15 h and then overlay with RM amended with 150 μg/ml hygromycin B.
14. When transformants emerge (usually within 4–7 days), screen putative transformants on V8 medium containing 450 μg/ml hygromycin.

3.1.6. Wheat Infection

1. Plant wheat, four kernels per four inch pot in BACCTO Professional Planting mix (Michigan Peat Company, Houston, TX). Grow in 16 h light (20°C) and 8 h dark (18°C). Water every 2 days.
2. Fourteen days after planting, fertilize with ten pellets of Osmocote Plus Multipurpose Plant Food (Scotts, Marysvile, OH) per pot.
3. Select wheat plants just prior to anthesis (~7 weeks after planting). Stamens should be present within the glumes, but ideally should not have emerged yet.
4. Gently pull back the glume and slowly inject 10 μl of conidia (5×10^5 conidia/ml) with a pipette into the floret (see Fig. 1).
5. Mark inoculated spikelets with a Sharpie on the outside of the glume.

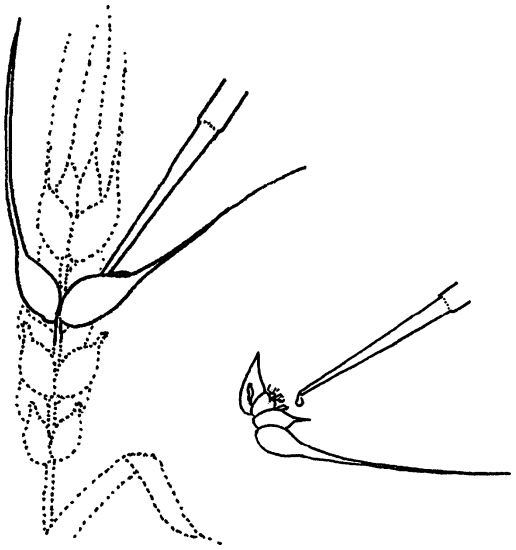

Fig. 1. Inoculation of wheat. Select a spikelet near the middle of the head for inoculation. *Inset*: From *left* to *right*: palea; coreopsis with cottony style at top and anthers to the sides (ideally, anthers have not yet emerged at time of inoculation); lemma; and lower glume. Gently peel back the lemma and lower glume and insert the pipette tip beside the coreopsis. Inject ten microliters of conidial suspension and remove the pipette; gently return the glume and lemma to position. Mark the glume with a Sharpie to denote the inoculation point.

6. Incubate plants for 3 days in a mist chamber.
7. Remove plants from the mist chamber. Plants will not need watering for 1–2 days; afterwards, resume normal watering schedule.

3.1.7. Barley Infection

1. Plant barley (6), four seeds per six inch pot in Scotts MetroMix 200. Grow in 16 h light (20°C) and 8 h dark (18°C). Water every 2–3 days until 14 days, then daily until sampling.
2. Seven days after planting, apply 5 ml of Osmocote 14/14/14 (Marysville, OH).
3. Select plants 2–3 days following emergence from the boot. Anthesis has been completed.
4. Spray inoculate in the afternoon with a suspension of 2×10^6 conidia/ml in 0.04% Tween 20 using an airbrush at a pressure of 82.8 kPa.
5. Bag inoculated heads in clear plastic for 3 days.
6. Remove plastic bags. Continue normal watering schedule throughout inoculation and follow-up.

3.1.8. Long-Term Storage and Culture Maintenance

1. Soil stocks – inoculate a vial of sterile soil with 100 µl of conidia (5×10^5 conidia/ml stock). Maintain at room temperature long enough for the conidia to germinate and the fungus to colonize the soil (~1 week – 10 days). Store at –20°C.
2. Agar plugs – cut plugs from the young, growing edge of a colony on agar. Immerse in 35% glycerol and store at –80°C.
3. See Note 7.

3.2. Trizol-Based RNA Extraction from Cultures and Wheat Stems (3, 13)

1. Grind samples in mortar and pestle in liquid nitrogen.
2. Add 5 ml of TRIzol solution. Continue grinding.
3. Decant into 30-ml baked Corex tubes. Incubate at RT for 5 min.
4. Add 1 ml of chloroform, vortex, and incubate at RT for 2–3 min.
5. Centrifuge for 15 min at 12,000 × *g* at RT.
6. Transfer the upper aqueous layer to a new 30-ml Corex tube (discard the lower layer).
7. Add 1 volume of RNA CTAB, vortex, and incubate for 25 min – 2 h at 65°C.
8. Add 1 volume of chloroform:IAA. Vortex briefly.
9. Centrifuge for 10 min at 12,000 × *g* at RT.
10. Transfer the upper aqueous layer to a new 30-ml Corex tube.
11. Repeat once the chloroform:IAA extraction (steps 8–10).

12. Add 0.25 volume of 3 M NaOAc and 1 volume of 2-propanol.
13. Incubate for 10 min (to overnight) at −20°C.
14. Centrifuge for 10 min at 12,000 × *g* at RT.
15. Discard the supernatant. Wash the pellet with 1 ml of 70% ethanol.
16. Centrifuge for 5 min at 7,500 × *g* at RT.
17. Discard the supernatant. Resuspend the pellet in 100 µl of RNase-free water and transfer to a 1.5-ml Eppendorf tube. Leave the tube open in a 65°C heatblock for 5 min to evaporate remaining ethanol (see Note 8).
18. Quantify.
19. Proceed to RNA cleanup (Subheading 3.4).

This method results in poor yields from wheat kernels.

3.3. Lithium Chloride-Based RNA Extraction from Wheat Kernels (24)

1. Heat a 1:1 mixture of extraction buffer:phenol to 80°C.
2. Grind sample (ideally 20+ kernels, see above, Subheading 2.3) in mortar and pestle in liquid nitrogen.
3. Add 5 ml of the extract buffer:phenol mixture. Continue grinding.
4. Decant into a 30-ml baked Corex tube.
5. Add 2.5 ml of chloroform. Vortex to mix.
6. Centrifuge for 30 min at 2,500 × *g* at RT.
7. Transfer the upper aqueous layer to a baked 15-ml Corex tube (discard the lower layer).
8. Add 0.2 volume of 10 M LiCl.
9. Incubate on ice for at least 2 h or overnight at −20°C.
10. Centrifuge for 30 min at 12,000 × *g* at 4°C.
11. Discard the supernatant. Wash the pellet with 2 M LiCl.
12. Centrifuge for 5 min at 12,000 × *g* at RT.
13. Discard the supernatant. Wash the pellet with 70% ethanol.
14. Centrifuge for 5 min at 12,000 × *g* at RT.
15. Remove the supernatant and allow the sample to air dry.
16. Resuspend the pellet in 2 ml of RNase-free water. Add 200 µl of 3 M NaOAc and 5.5 ml absolute ethanol.
17. Incubate for 15 min at −80°C.
18. Centrifuge for 5 min at 12,000 × *g* at RT.
19. Discard the supernatant. Wash the pellet with 70% ethanol.
20. Centrifuge for 5 min at 12,000 × *g* at RT.

21. Discard the supernatant. Resuspend the pellet in 100 µl of RNase-free water and transfer to a 1.5-ml Eppendorf tube. Leave the tube open in a 65°C heatblock for 5 min to evaporate remaining ethanol (see Note 8).
22. Quantify.
23. Proceed to RNA cleanup (Subheading 3.4).

This method is not efficient in extracting RNA from culture materials or infected wheat stems.

3.4. RNA Cleanup Using the RNeasy Mini Kit (Qiagen)

1. Bring no more than 100 µg RNA to 88 µl in RNase-free water.
2. Add 10 µl of 10× incubation buffer and 2 µl (=20 U) of DNase I.
3. Incubate for 15 min at 37°C.
4. Add 4 µl of 0.2 M EDTA.
5. Incubate for 10 min at 75°C to halt the reaction.
6. Add 350 µl of Buffer RLT (from the RNeasy Mini Kit) and 225 µl of absolute ethanol. Apply sample to column (from the kit).
7. Centrifuge for 15 s at 8,000 × *g*.
8. Discard the flow through. Add 650 µl of Buffer RW1 (see Note 9).
9. Centrifuge for 15 s at 8,000 × *g*.
10. Transfer column to a new 2-ml Eppendorf tube; add 500 µl of Buffer RPE (from the kit).
11. Centrifuge for 15 s at 8,000 × *g*.
12. Discard the flow through. Add 500 µl of Buffer RPE.
13. Centrifuge for 2 min at 8,000 × *g*.
14. Transfer the column to a new 1.5-ml Eppendorf tube. Add 15 µl of RNase-free water.
15. Centrifuge for 1 min at 8,000 × *g*.
16. *Optional* – Add an additional 15 µl of RNase-free water. Repeat centrifugation. (Recommended if the concentration of starting RNA is 0.5 µg/µl or greater.)
17. Quantify and obtain 260/280 ratio (see Note 10).

3.5. Preparation of aRNA and Affymetrix GeneChip Hybridization (see Note 11)

1. Dilute RNA to 200 ng in 3 µl.
2. Prepare a serial dilution of Poly-A RNA stock control: 2 µl in 38 µl Poly-A Control Dilution Buffer; 2 µl of first dilution in 98 µl of buffer; 2 µl of second dilution in 98 µl of buffer; 10 µl of third dilution in 90 µl of buffer. This final dilution is the poly-A spike (see Note 12).

3. Add RNA to 2 µl of poly-A spike.
4. Prepare First Strand Master Mix of 4 µl First-Strand Buffer Mix and 1 µl First-Strand Enzyme Mix per reaction.
5. Add 5 µl of First Strand Master Mix to 5 µl RNA + poly-A spike.
6. Incubate for 2 h at 42°C. Place on ice.
7. Prepare Second Strand Master Mix: 13 µl of water, 5 µl of Second-Strand Buffer Mix, 2 µl of Second Strand Enzyme Mix (per reaction).
8. Add 20 µl of Second Strand Master Mix to each completed first strand reaction.
9. Incubate for 1 h at 16°C and 10 min at 65°C in a thermal cycler (keep mix on ice until thermal cycler reaches 16°C.
10. Prepare IVT Master Mix: 4 µl of IVT Biotin Label, 20 µl of IVT Labeling Buffer, 6 µl of IVT Enzyme Mix per sample.
11. Add 30 µl of IVT Master Mix to each completed second strand reaction.
12. Incubate for 4 h at 40°C. The product is aRNA (=amplified RNA; cRNA = complimentary RNA in previous Affymetrix protocols).
13. Place on ice or freeze at −20°C.
14. Preheat 50 µl of aRNA Elution Solution per sample at 50–60°C for 10 min.
15. Prepare aRNA Binding Mix of 10 µl RNA Binding Beads and 50 µl aRNA Binding Buffer Concentrate per reaction.
16. Add 60 µl of aRNA Binding Mix to each aRNA sample, mix, and transfer to a 96-well plate.
17. Add 120 µl of absolute ethanol and mix by pipetting.
18. Shake slowly on shaker for ≥2 min to allow aRNA to bind to beads.
19. Place 96-well plate on a magnetic stand for approximately 5 min, or until sample clears and beads are captured by the magnets.
20. Remove the supernatant with pipette without disrupting magnetic pellet (leave plate on magnetic stand during this step).
21. Remove plate from the magnetic stand and wash the sample by adding 100 µl of aRNA Wash Solution (to which ethanol has been added) and shaking at moderate speed for 1 min.
22. Repeat capture of beads with magnetic stand and removal of supernatant, as above (Subheading 3.5, steps 19 and 20).
23. Repeat wash (Subheading 3.5, steps 21 and 22).

24. Remove plate from magnetic stand and dry by shaking vigorously for 1 min to evaporate ethanol.
25. Add 50 µl of preheated aRNA Elution Solution to each sample.
26. Shake vigorously until the beads are fully dispersed (≥3 min).
27. Capture beads on a magnetic stand.
28. Transfer the supernatant to a new 1.5-ml Eppendorf; this contains the cleaned aRNA. Samples may be stored at −80°C at this point.
29. Measure OD and 260/280 ratio with a spectrophotometer (see Note 13).

3.6. Basic Statistical Analyses

3.6.1. Normalization (RMA) (25)

1. Place the .CEL file(s) in the working directory. By default, the working directory is the R folder. If any other directory is desirable, do so as follows: >setwd("c:\\My Documents\MyCelFiles").
2. Open R.
3. Load the "affy" program: >library(affy) (see Note 14).
4. If using Windows/PC, set the memory as high as possible: >memory.limit(size=3000) (see Note 15).
5. Load the .CEL file data into R: >rawdata <- ReadAffy() (see Note 16).
6. Normalize the data:>eset <- rma(rawdata) (see Note 17).
7. Export normalized data to Excel (optional, but useful for visualizing results): >write.exprs(eset, file="NormalizedDate.txt").
8. Open Excel. Open the normalized file and go through the steps to turn a text file into a proper Excel file (see Note 18).
9. If doing further analyses within R, proceed to do so. If not, quit R: >q().
10. When asked whether to save the workspace, select yes (y) to preserve the normalized file for future analyses.

3.6.2. Comparison of Different Conditions (Limma) (27, 28)

1. Compile a list of genes of choice, as: >Fats <- eset[c(4661, 9881,8332...), c(1:5,12:26,6,7...)] where "4661, 9881, 8332..." are the numbers corresponding to the genes you are interested in (as determined from viewing the RMA file in Excel; see above) and "1:5, 12:26, 6, 7..." correspond to the input CEL files (see Note 19).
2. Open the limma library: >library (limma).
3. Design a matrix in which you tell R which treatments are which. For the example of time course experiment: >design <- model.matrix(~ 1+factor(c(1,1,1,1,1,2,2,2,3,3,3))) (see Note 20).
4. Designate names for your treatments: >colnames(design) <- c("group1", "group2", "group3") (see Note 21).

5. Apply the matrix and treatment names to the list of genes and/or CEL files you designated at the beginning: >fit <-lmFit(Fats, design).
6. State which comparisons to make: > contrast.matrix <-make Contrasts(group2-group1, group3-group1, levels=design) (see Note 22).
7. The next few commands run the statistical analyses on the comparisons: > fit2 <- contrasts.fit (fit, contrast.matrix); > fit2 <- eBayes(fit2).
8. To generate a list of the top ten differentially expressed genes: > topTable(fit2, coef=1, adjust="BH") (see Note 23).

3.6.3. Heatmaps (gplots and heatmap.2)

1. Compile a list of genes of choice: >Fats <- eset[c(4661, 9881,8332...), c(1:5,12:26,6,7...)] (see Note 24).
2. Obtain the programs needed to generate the heatmap: >install.packages("gplots"); >install.packages("gtools"); >install.packages("gdata") (see Note 25).
3. Open the gplots library: > library(gplots).
4. Run heatmap.2 on the dataset. Almost every parameter in the command line can be varied. Please see heatmap.2 documentation (within the program) for more details. The following parameters are useful for *Fusarium* GeneChip data: > heatmap.2(exprs(Fats), col=redblue(75), key=TRUE, symkey=FALSE, density.info="none", trace="none", cexRow=0.5) (see Note 26).

4. Notes

1. Do not overgrow mycelia, or perithecium production will be poor. Note that continuous light suppresses the circadian rhythm. Some long wave UV light is essential for perithecium production. Older white fluorescent bulbs may lose their ability to induce sexual development. A second, "blacklight" or "sunlight balanced" bulb can be used to assist, if light appears to be the trouble.
2. *F. graminearum* PH-1 produces the following discrete stages during sexual development: wide binucleate hyphae (~18–22 h post-induction [hpi]), protoperithecia (48 hpi), immature perithecia with paraphyses (72 hpi), immature perithecia with developing asci (96 hpi), mature perithecia discharging ascospores (144 hpi). Times are approximate. If seeking a particular developmental stage, always check a representative sample under the microscope (2, 3).

Strain PH-1 was selected from a wide variety of field isolates on the basis on its readiness to consistently and reliably undergo sexual differentiation in the laboratory, its tendency to conidiate prolifically, and the ease with which it produces protoplasts (29). Not all *F. graminearum* strains can be as readily induced to undergo sexual development.

3. The protocol for SC cultures is similar to that for carrot agar cultures. However, growth on SC is characterized by very little surface mycelium and perithecium initials begin to develop very soon after inoculation, but not as synchronously as on carrot agar. The benefit of using SC medium is that it is defined, so one can test nutrients and additives that may affect early stages of sexual development in a defined medium. In addition, SC is transparent, allowing for easier viewing of the stages under the microscope. SC does not have various carrot cellular structures that can interfere with fluorescence microscopy. Finally, there is no need for addition of Tween 60 (as there is for carrot agar), which in some work can be problematic (12). Later stages of sexual development are not as prolific on SC medium as they are on carrot agar.
4. Protoplasts are spherical, while intact *Fusarium* cells occur in a variety of shapes, but are not spherical. Under the microscope, many round protoplasts should be present in the field of view when observed with a 40× objective; if only a few protoplasts are present, continue the reaction. After 2½ h, further incubation will not be of benefit.
5. Nitex or some similar nylon membrane with 30 μm mesh must be used; Miracloth is not an acceptable substitute. The pore size in Miracloth is larger and is too variable to be effective.
6. Note that too much DNA can be inhibitory, too little can be ineffective. Incubate at room temperature for 20 min.
7. *Fusarium* species are notoriously unstable in culture or in serial transfer conditions. Soil stocks and agar plugs (stored in glycerol at −80°C) provide effective long-term storage. Conidial stocks may remain usable for ≥3 years, but will degrade over time, particularly if subjected to repeated freeze–thaw cycles. Keep a master stock in the freezer and several substocks to culture from. For each experiment, start a fresh culture from the −80°C freezer stock.
8. Do not use a waterbath, and do not put water in the heatbath; dry heat is best for evaporation.
9. The use of Buffer RW1 was adopted following discussions with Qiagen technical support. Some developmental stages, particularly those in which perithecia are developing and maturing, are rich in polysaccharides and can form an insoluble gel during RNA extraction and cleanup. Buffer RW1 helps to minimize this problem.

10. QRT-PCR requires a 260/280 ≥ 2.0 at this stage. Affymetrix is less stringent, but should still have 260/280 ≥ 1.9. Lower 260/280 ratios can sometimes be improved by repeating the RNeasy steps (Section 3.4, steps 6–17).
11. As of September, 2009, Affymetrix has discontinued production and sale of their One Step Labeling Kit, which has been superceded by the GeneChip 3′ IVT Express Kit. Please note if you have used the One Step Kit, the GeneChip 3′ IVT Express Kit produces comparable results with significantly smaller quantities of input RNA.
12. We find 200 ng starting material to be sufficient in most cases (the previous kit required 5 μg). First dilution can be stored at −20°C to 6 weeks, and subjected to eight freeze–thaw cycles; second through fourth dilutions must be made fresh).
13. Hybridization of cleaned, labeled aRNA is generally performed off-site, as it is not economical for the average molecular biology lab to own and maintain their own GeneChip Instrument System and Operating Software. 20–80 μg total yield is common. 260/280 ratios greater than 1.8 are desirable, but good results can be obtained with lower ratios if the RNA is not badly degraded (midpoint of smear ~1 kb when visualized on an RNA gel or Bioanalyzer [Agilent]).
14. This opens the "affy" library in Bioconductor, which enables you to process Affymetrix data.
15. 3,000 MB (3 GB) is the most Windows is capable of allocating to R. R will probably respond with the word "NULL." Mac and Linux systems do not have the same memory issues as Windows.
16. This will choose a name for an object ("rawdata" is sensible, but you could call it anything you like), and defining "rawdata" as the temporary file that holds the output of the "ReadAffy()" command. "ReadAffy()" reads all of the .CEL files in the R folder into "rawdata." If you only wanted to work with 24 h replicates, the command will be: >rawdata <- ReadAffy("24H_rep1.CEL", "24H_rep2.CEL", "24H_rep3.CEL").
17. This will create an object (named "eset" by convention; again, it could be whatever you like it to be), and defining it as the outcome of running "rma" on your previously-named "rawdata". "rma" does three things: "convolution background correction, quantile normalization, and a summarization based on a multi-array model fit robustly using the median polish algorithm" (26).
18. Insert one cell in the upper left-hand corner of the Excel file, as all of the column headings are shifted one to the left of where they should be. In general, probe sets are arranged by R in the alphabetical order, but that can vary a bit between

versions of R/Bioconductor. You may or may not be given a column with the numbers 1-x, x being the total number of probe sets on your GeneChip. If you do not have such a column, make one; your first probe (as ordered in the normalized file) is number 1 (and corresponds with the second row on your Excel file); your xth probe is number x (and corresponds with row x+1). This is important, because this is how R thinks of the probe sets. Probe names (fg00505_at) are not meaningful to R and unsuitable for more sophisticated analyses.

19. R orders things in an alphanumerical fashion that places 144 between 0 and 24, so, for a time course, it is necessary to reorder the genes. Reordering may or may not be necessary with your dataset.

    For the mutant vs. wild-type comparison, imagine your CEL files 19-23 correspond to wildtype at 96 h, and 31-32 correspond to the mutant at 96 h. You are interested in all genes, so you would enter >Mutant <- eset[,c(19:23,31,32)].

20. It will direct the software to put the first five CEL files it has read into one category (0H), the second three in another (24H) and the third group of three in a third (48H).

21. Syntax is important. R will not start with a number (e.g., the first treatment cannot be named "0H").

22. This compares 24H to 0H, and 48H to 0H.

23. The "coef=1" refers to the first comparison, 24H vs. 0H. To see 48H vs. 0H, enter "coef=2" instead.

    The output table will be in the format:

ID	M	A	t	P.Value	adj.P.Val	B
17 fgd233-470_at	–3.46	6.3	–9.9	3.74E-09	3.88E-07	11.2
20 fgd93-380_at	3.03	6.8	9.8	4.36E-09	3.88E-07	11.1
129 fgd160-950_at	–4.31	4.8	–6.9	9.82E-07	4.37E-05	5.8

    The first number is the order in which that gene was found in the exprSet. "ID" numbers are the names with which the genes were labeled on the GenChip. "M" is the $\log_2$-normalized fold change; 3.03 corresponds to a fold change of $2^{3.03}$, or 8.17-fold upregulation at 24H vs. 0H. The negatives correspond to down-regulation; –3.46 is $2^{3.46}$ lower expression (11-fold) in 24H vs. 0H. Any number in this column exceeding 1 represents a twofold increase or greater. "t" is the t-statistic. "P. value" is obvious; "adj.P.Val," the adjusted P-value, is somewhat more conservative measure. "B" is the log-odds.

    In many cases, you will be interested in more than the top ten genes. In the command >FatComparison <-topTable(fit2, coef=1, number=x, adjust="BH", "x" can be set to as many genes as desired. In many cases, you will be interested in a

lot more genes than you can easily scan or analyze in R's command window. You can export the file to Excel as: >write.table(FatComparison, file= "FatComparison.txt", sep= "\t")

24. (See above, Subheading 3.6.2, for more detail). Unlike the statistical comparisons of limma, you will not be able to examine all of the genes on the GeneChip (as in a mutant vs. wild-type comparison). For a heatmap to be readable, do not exceed 100 genes.
25. At the first installation, a window will open up asking which CRAN mirror you wish to download from (R is hosted at – and downloadable from – many sites). Generally, the one nearest you will be the quickest.
26. This produces a graphic (which can be exported as jpeg or bitmap) showing your heatmap with all of the columns (CEL files) and rows (the individual genes) labeled. This is useful for your own reference, but almost certainly you will want to relabel for publication purposes. To remove all column and row labels>heatmap.2(exprs(FATS), Colv=FALSE, Rowv=FALSE, col=redblue(75), key=TRUE, symkey=FALSE, density.info= "none", trace= "none", cexRow=0.5).

    R will generate a warning message if the column and/or row labels are removed, but will still generate the graphic. The above commands generate heatmap scaled to the $\log_2$-normalized expression values (red being lowest and blue being highest). To normalize values by gene, add the command scale=row within the parentheses.

Acknowledgments

This work was supported in part by a grant from the National Science Foundation (0923794 to FT) and the Michigan Agricultural Experiment Station. We thank the many Fusarium workers past and present who contributed in all ways to the evolution of these protocols.

References

1. Güldener, U., Mannhaput, G., Münsterkötter, M., Haase, D., Oesterheld, M., Stümpflen, V., Mewes, H.-W., and Adam, G. (2006) FGDG: a comprehensive fungal genome resuorce on the plant pathogen *Fusarium graminearum*. *Nucleic Acids Res.* **34**, D456–D458.
2. Trail, F. and Common, R. (2000) Perithecial development by *Gibberella zeae*: a light microscopy study. *Mycologia*. **92**, 130–138.
3. Hallen, H. E., Huebner, M., Shiu, S.-H., Güldener, U., and Trail, F. (2007) Gene expression shifts during perithecium development in *Gibberella zeae* (anamorph *Fusarium graminearum*), with particular emphasis on ion transport proteins. *Fungal Genet. Biol.* **44**, 1146–1156.
4. Bowden, R. L. and Leslie, J. F. (1999) Sexual recombination in *Gibberella zeae*. *Phytopathology*. **89**, 182–188.

5. Guenther, J. C. and Trail, F. (2005) The development and differentiation of *Gibberella zeae* (anamorph: *Fusarium graminearum*) during colonization of wheat. *Mycologia* **97**, 229–237.
6. Boddu, J., Cho, S., Kruger, W. M., and Muehlbauer, G. J. (2006) Transcriptome analysis of the barley-*Fusarium graminearum* interaction. *Mol. Plant Microbe Interact.* **19**, 407–417.
7. Urban, M., Daniels, S., Mott, E., and Hammond-Kosack, K. (2002) *Arabidopsis* is susceptible to the cereal ear blight fungal pathogens *Fusarium graminearum* and *Fusarium culmorum*. *Plant J.* **32**, 961–973.
8. Parker, D., Beckmann, M., Ent, D. P., Overy, D. P., Rios, Z. C., Gilbert, M., Talbot, N, and Draper, J. (2008) Rice blast infection of *Brachypodium distachyon* as a model system to study dynamic host/pathogen interactions. *Nat. Proc.* **3**, 435–445.
9. Güldener, U., Seong, K.-Y., Boddu, J., Cho, S., Trail, F., Xu, J.-R., Adam, G., Mewes, H.-W., Muehlbauer, G. J., and Kistler, H. C. (2006) Development of a *Fusarium graminearum* Affymetrix GeneChip for profiling fungal gene expression in vitro and in planta. *Fungal Genet. Biol.* **43**, 316–325.
10. Stephens, A. E., Gardiner, D. M., White, R. G., Munn, A. L., and Manners, J. M. (2008) Phases of infection and gene expression of *Fusarium graminearum* during crown rot disease of wheat. *Mol. Plant Microbe Interact.* **21**, 1571–1581.
11. Seong, K. Y., Zhao, X., Xu, J.-R., Güldener, U., and Kistler, H. C. (2008) Conidial germination in the filamentous fungus *Fusarium graminearum*. *Fungal Genet. Biol.* **45:**389–399.
12. Guenther, J. C., Hallen-Adams, H. E., Bücking, H., Shachar-Hill, Y., and Trail, F. Triacylglyceride metabolism by *Fusarium graminearum* during colonization and sexual development on wheat. *Mol. Plant Microbe Interact. in press*
13. Hallen, H. E. and Trail, F. (2008) The L-type calcium ion channel cch1 affects ascospore discharge and mycelial growth in the filamentous fungus *Gibberella zeae* (anamorph *Fusarium graminearum*). *Eukaryot. Cell.* **7**, 415–424.
14. Schreiber, A. W., Sutton, T., Caldo, R. A., Kalashyan, E., Lovell, B., Mayo, G., Muehlbauer, G. J., Druka, A., Waugh, R., Wise, R. P., Langrisge, P., and Baumann, U. (2009) Comparative transcriptomics in the Triticeae. *BMC Genomics.* **10**, 285.
15. Booth, C. (1971) The Genus *Fusarium*. Commonwealth Mycological Institute, Kew, Surrey, England.
16. Cappellini, R. A. and Peterson, J. L. (1965) Macroconidium formation in submerged cultures by a non-sporulating strain of *Gibberella zeae*. *Mycologia* **57**, 962–966.
17. Westergaard, M. and Mitchell, H. K. (1947) Neurospora. V. A synthetic medium favoring sexual reproduction. *Am. J. Bot.* **34**, 573–577.
18. http://www.fgsc.net/neurosporaprotocols/How%20to%20choose%20and%20prepare%20media.pdf.
19. Shin, S., Mackintosh, C. A., Lewis, J., Heinen, S. J., Radmer, L., Dill-Macky, R., Baldridge, G. D., Zeyen, R. J., and Muehlbauer, G. J. (2008) Transgenic wheat expressing a barley class II chitinase gene has enhanced resistance against *Fusarium graminearum*. *J. Exp. Bot.* **59**, 2371–2378.
20. Oelke, L. M. and Kolmer, J. A. (2005) Genetics of leaf rust resistance in spring wheat cultivars alsen and norm. *Phytopathology.* **95**, 773–778.
21. Ma, Z., Steffenson, B. J., Prom, L. K., and Lapitan, N. L. V. (2000) Mapping of quantitative trait loci for Fusarium head blight resistance in barley. *Phytopathology.* **90**, 1079–1088.
22. Fairhead, C., Llorente, B., Denis, F., Soler, M., and Dujon, B. (1996) New vectors for combinatorial deletions in yeast chromosomes and for gap-repair cloning using "split-marker" recombination. *Yeast.* **12**, 1439–1457.
23. Catlett, N. L., Lee, B., Yoder, O. C., and Turgeon, B. G. (2003) Split-marker recombination for efficient targeted deletion of fungal genes. *Fungal Genet. Newsletter.* **50**, 9–11.
24. Goswami, R. S., Xu, J. R., Trail, F., Hilburn, K., and Kistler, H. C. (2006) Genomics analysis of host-pathogen interaction between *Fusarium graminearum* and wheat during early stages of disease development. *Microbiol.* **152**, 1877–1890.
25. Irizarry, R. A., Hobbs, B., Collin, F., Beazer-Barclay, F., Antonellis, K. J., Scherf, U., and Speed, T. P. (2003) Exploration, normalization and summaries of high density oligonucleotide array probe level data. *Biostatistics.* **4**, 249–264.
26. Bolstad, B. M., Irizarry, R. A., Gautier, L., and Wu, Z. (2005) "Preprocessing high-density oligonucleotide arrays", pp. 13-32 in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Gentleman, *et al.*, eds. New York, Springer.
27. Smyth, G. K. (2004) Linear models and empirical Bayes methods for assessing differential expression in microarray experiments. *Stat. Appl. Genet. Mol. Biol.* **3**, 3.

28. Smyth, G. K. (2005) "Limma: linear models for microarray data", pp. 397–420 in *Bioinformatics and Computational Biology Solutions Using R and Bioconductor*, Gentleman, *et al.*, eds. New York, Springer.

29. Trail, F., Gaffoor, I., Guenther, J. C., and Hallen, H. E. (2005) Using genomics to understand the disease cycle of the Fusarium head blight fungus, *Gibberella zeae* (anamorph *Fusarium graminearum*). *Can. J. Plant Pathol.* **27**, 486–498.

Chapter 7

EST Analysis Pipeline: Use of Distributed Computing Resources

Francisco Javier González and Juan Antonio Vizcaíno

Abstract

This chapter describes how a pipeline for the analysis of expressed sequence tag (EST) data can be implemented, based on our previous experience generating ESTs from *Trichoderma* spp. We focus on key steps in the workflow, such as the processing of raw data from the sequencers, the clustering of ESTs, and the functional annotation of the sequences using BLAST, InterProScan, and BLAST2GO. Some of the steps require the use of intensive computing power. Since these resources are not available for small research groups or institutes without bioinformatics support, an alternative will be described: the use of distributed computing resources (local grids and Amazon EC2).

Key words: Expressed sequence tag (EST), Workflow, Distributed computing, Grid, Functional annotation

1. Introduction

Although fungal genomes are somewhat larger than bacterial genomes, many important fungal genomes are relatively modest in size (7–40 Mb). However, when it is not possible to obtain the complete genome sequence from a given fungus, a common alternative is to generate expressed sequence tags (ESTs). Many EST-based projects have been completed in recent years; an extensive list describing both genomic and EST-based resources can be found at the Genome OnLine Database, or GOLD (1). Another list, which only contains fungal projects, can be obtained from http://fungalgenomes.org/.

Extensive computational strategies have been developed to organize and analyze data from small- and large-scale EST

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722,
DOI 10.1007/978-1-61779-040-9_7,

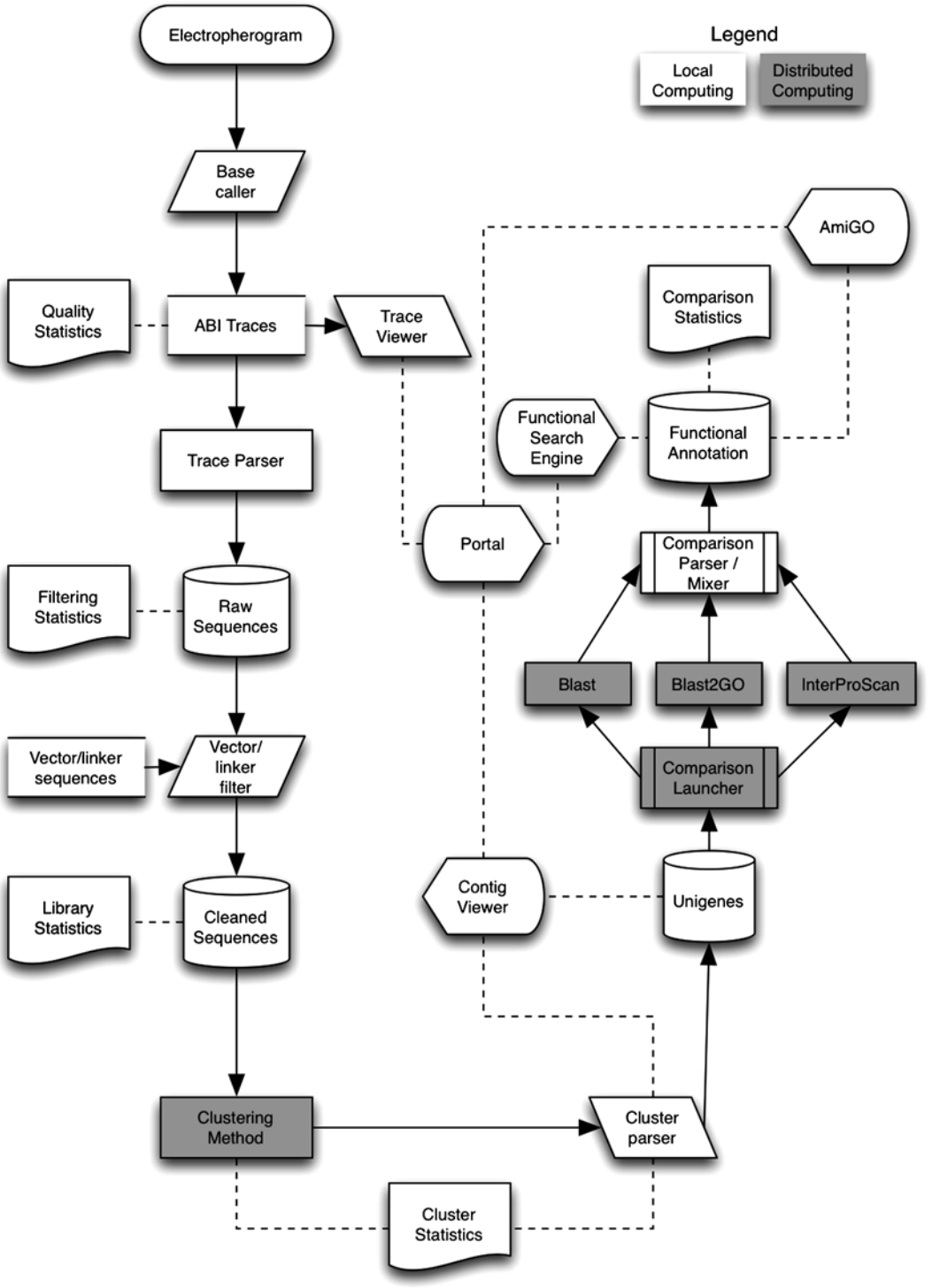

Fig. 1. General scheme of the project pipeline proposed in this chapter.

projects. Many sequencing projects, especially those generating EST data, follow a similar data management workflow (Fig. 1). Multiple steps require access to high-throughput computing resources, since it is frequently necessary to analyze large numbers of sequences with algorithms that require a considerable amount of time to process. Additionally, data storage requires a substantial amount of memory. Finally, stored data must be easily accessible, usually through a web-based portal replete with a variety of graphical and textual data analysis tools.

When planning an EST sequencing project, it is almost certain that no perfect software solution will exist for all the required steps of your particular workflow. Therefore, in-house tailored solutions need to be developed, an endeavor that represents a considerable amount of the total workload. As an additional complication, hardware infrastructure sufficient to perform some of the required tasks is not always available. Fortunately, alternative strategies based on distributed computing approaches allow researchers to circumvent this potential obstacle.

The workflow represented in Fig. 1 was the foundation of a EU-funded project called "TrichoEST," in which almost 35,000 EST sequences from different *Trichoderma* species were generated (2, 3). In addition to the tools that were used in that project, updates and/or new tools that could be incorporated into an EST analysis pipeline will also be presented. We will focus on some of the key steps of the workflow: processing raw data from the sequencers, clustering ESTs, and annotating the resulting sequences using distributed computing resources.

This chapter is aimed for bioinformaticians and not for pure wet lab biologists, although we have tried to keep explanations as simple as possible. A last remark is that we will not discuss data visualization. Viewers in this workflow need to be implemented by specialized programmers with basic knowledge in the programming languages such as Perl, Python, PHP, and HTML or indeed customizing projects like AmiGO (http://amigo.geneontology.org/).

2. Materials

2.1. Electropherograms (Chromatograms or Flowgrams)

Currently, there are two main sources of sequencing data. The traditional technology involves the generation of electropherograms based on dye-terminator sequencing, mainly coming from capillary electrophoresis on Beckman Coulter CEQ or Applied Biosystems 31xx sequencers. However, in recent years, new sequencing technologies collectively known as next-generation

sequencing have emerged (4). Some next-generation technologies are based on pyrosequencing, e.g., the Roche 454. Pyrosequencing follows the "sequencing by synthesis" principle, which relies on detection of pyrophosphate release on nucleotide incorporation. Additionally, two other platforms based on solid phase amplification (Solexa, from Illumina) and sequencing by ligation (SOLiD by Applied Biosystems), complete this breakthrough group of massively parallel sequencing methods. In these new technologies, the outputs are "runs" representing multiplexed "reads" of short sequences.

2.2. Raw Data Processing Algorithms

Once the trace has been registered for each run by the sequencer (see Note 1), a base caller algorithm determines the quality of the trace and provides a sequence in plain text format as output. Depending on the sequencing method, different base-calling algorithms are applied. The chromatogram files from each sequencing platform have proprietary extensions (SCF for Beckman Coulter, ABI, AB1 or AB for Applied Biosystems, and SFF for 454). These extensions refer to standard formats registered at NCBI (http://www.ncbi.nlm.nih.gov/Traces/), e.g., SFF (Standard Flowgram Format) or SCF (Standard Chromatogram Format).

For ultra-high throughput sequencing platforms like 454, there are three different output files: a .fna file, which is a fasta file containing the sequence for each read, a .qual file that stores its corresponding base quality score, and a .sff file that stores the information on the signal strength for each flow (the equivalent to SCF or AB1 files in the Sanger sequencing methods). However, Sanger sequencing results have only a fasta file with a .txt or .fas extension if the management software has been set to process the trace directly with its proprietary base caller algorithm. Alternatively, raw traces can also be base called using a program called Phred (5, 6) (http://bozeman.genome.washington.edu/phredphrapconsed.html#block_phred, see Note 2). Additionally, other commercial applications (http://www.dnabaser.com/features.html) deploy their own proprietary algorithms for base calling (see Notes 3 and 4).

2.3. Clustering Method

Clustering analyses group EST sequences into sequences (or clusters) that have stringent similarity and are likely derived from the same template sequence. The TGI Clustering tools (TGICL) (7) automates clustering and assembly of large EST/mRNA datasets and can be used for fast clustering. The clustering process is performed by a slightly modified version of the Mega BLAST program from the NCBI, and the resulting clusters are then assembled by another program called CAP3 (8). TGICL can be obtained from http://compbio.dfci.harvard.edu/tgi/software/.

2.4. Distributed Computing (Local Grid and Amazon EC2)

Several components of the EST analysis workflow require high throughput computing capacity. Different internet-based options are available that use either open-source parallelized software or pay-per-use services running on remote infrastructures. The easiest option is to gather the required computer power using grids. A grid is the combination of computer resources applied to a common task that requires a great number of computer processing cycles or the need to process large amounts of data. There is a grid solution, namely the community-oriented open-source version of the Fura middleware family (http://www.gridsystems.com/?p=products/products.php&s=8). The Fura project is a self-contained grid middleware that allows the grid enablement and distribution of applications on heterogeneous computational resources. Its architecture is based on plugins that allow grid services to be extended or replaced, and new ones developed, reusing existing components. Fura features a web-based graphical user interface (GUI), wizard-guided installation and configuration, and web service compliance. The company GridSystems launched the Fura Project in 2007, adopting an open-source business model in order to lower the barriers, ease the adoption of Grid technology, and create a community. Fura has currently a GPL2 type license.

2.5. InterProScan

Functional annotation of the generated sequences is the ultimate goal of most EST-based projects. InterPro is a resource that integrates a number of signature protein databases and can be a powerful tool to annotate protein sequences (9). Signatures from different databases that describe the same domain, family, repeat, active site, binding site, or post-translational modification are grouped into single InterPro entries. Protein matches are calculated using the program InterProScan (10), a tool that is also available for user query sequence searches (http://www.ebi.ac.uk/InterProScan/).

2.6. BLAST2GO

Blast2GO is a bioinformatics tool that can be used for the functional annotation and analysis of gene or protein sequences (11, 12). It is a platform-independent Java application made available via Java Web Start (JWS). According to the authors, the main concept behind the development of this tool was that it should be easy for biological researchers to use. However, more advanced functionalities are available for researchers with a higher computational background. The basic functionality of BLAST2GO is that the program uses BLAST searches in order to find similar sequences to one or several input sequences. Blast2GO then extracts associated gene ontology (GO) terms and returns an evaluated GO annotation for each query sequence. Additionally, the program can also retrieve enzyme codes (EC numbers), InterPro (9) protein domains, and KEGG pathway maps. Blast2GO can be downloaded from http://www.blast2go.de.

3. Methods

3.1. Raw Data Processing Algorithms: Phred

As mentioned before, Phred (5, 6) can be used instead of proprietary software from the sequencers to process raw data. To determine quality scores, Phred calculates several parameters related to peak shape and peak resolution at each base. Then, it uses these parameters to assign a quality score based on hard-coded lookup tables generated from sequence traces for which the correct sequence was known. Different lookup tables are used for different sequencing chemistries and machines. An evaluation of Phred quality scores for various sequencing chemistries and instruments showed that the scores are highly accurate (13).

Once the raw data have been base called, the sequences must be trimmed to eliminate low-quality bases. Phred quality scores are logarithmically related to error probabilities. For example, a Phred quality score of 30 indicates that the base in question is called incorrectly 1 time in 1,000. Generally, bases with a quality score of 20 and above are counted. Depending on the sequence coverage (number of reads for the same DNA fragment), the researcher can decide to use a Phred score cut-off value of either 20 or 30. In some cases, trimming sequences using these cut-off values can eliminate raw sequences of 300–500 bp from future analyses.

On the command-line, sequences are trimmed with various processing options of the *phred* executable:

- The command *-trim_alt* performs sequence trimming on the current sequence. Bases are trimmed from the start and end of the sequence on the basis of trace quality.
- The command *-trim_cutoff* sets trimming error probability for the *-trim_alt* option and the trimming points written in the .phd files. The default value is 0.05.
- The command *-trim_fasta* trims sequences written to sequence and quality value FASTA files. It also sets trimming information in the FASTA headers to reflect the high quality of the sequence, and append the string "trimmed" to the header (see Notes 5 and 6).

For instance, the command *phred -id* <chromatogram files directory> *-trim_alt 0.01 -trim_cutoff -st fasta -sa seqs_fasta* will trim bases with quality scores below 20 and will write fasta files containing all the chromatograms processed in the directory. Additionally, each sequence's fasta description will have a format similar to:

>*chromat_name 1323 15 548 ABI*, where:

- the sequence name immediately follows the header delimiter ">",
- the first integer is the number of bases called by Phred,

- the second integer is the number of bases "trimmed off" the beginning of the sequence,
- and the third integer is the number of bases "remaining following trimming," and the string describes the type of input file.

3.2. EST Clustering

TGICL requires input files containing multiple sequences in fasta format (and optionally, peer quality values file) and outputs assembly files as produced by CAP3. Both the clustering and assembly phases can be parallelized by distributing the searches and the assembly jobs across multiple CPUs, since TGICL can take advantage of either SMP (Symmetric Multi Processing) machines or PVM (Parallel Virtual Machine) clusters. However, the clustering phase does not perform any multiple alignment based approach, but only fast pairwise alignments (using Mega BLAST), which are then filtered and used to build subsets of sequences by a transitive closure approach. In the assembly phase, each cluster is then sent to the assembly program (CAP3), which attempts the multiple alignment of the sequences in the cluster and creates one or more contigs (the resulting consensus sequences containing one or more of the initial ESTs used as inputs).

After installing the package by decompressing the distribution (http://compbio.dfci.harvard.edu/tgi/software/), a Perl script can be found at the root directory (tgicl). To cluster and assemble with the default options, type: *tgicl* <name of fasta file>. This will run both clustering and assembly procedures. The resulting ACE format assembly files will be in a directory named: *./asm_*/ACE files*. An "asm_X" subdirectory is created for each CPU in a parallel processing setting (where X is a number from 1 to the number of CPUs you specified, e.g. *asm_1*, *asm_2*, etc.). Standard execution will only create an *asm_1* directory, as by default the program is using one CPU, but if you have a dual-CPU machine you can instruct the program to use both CPUs by adding "-c 2" to the command line specified above. The list of singletons will be included in the file called <fasta_filename>.*singletons*. During the clustering phase, *cluster_** temporary directories will be created for distributed searching of the databases. However, if only one CPU is used, only the cluster_1 directory will be created. After the clustering process is finished, these directories are removed and subsequently, one or more files named *hitsort_NNN.Z* will be created instead (where NNN can be 001,002, etc). This file contains the compressed output of the Mega BLAST pairwise searches, a special tab delimited format.

The clustering programs take the overlap data contained in the *hitsort_*.Z* files as an input. The resulting clusters are written in a **_clusters* file <fasta_filename> *_clusters*). A cluster file has a pseudo fasta format: each record is actually a cluster definition and consists of a header line having this format: >*cluster_name number_of_components*, followed by the list of sequence names

contained in the cluster. In the assembly phase, a fasta file containing the actual sequence data is built for each cluster, and this file is passed to the assembler. If only the actual fasta file with the singleton sequences is needed, these can be easily obtained with the following command: *cdbyank fastadb.cidx <fastadb.singletons> singleton.seqs*

The TIGCL program also creates a few log files for the main process, as well as for each of the sub-processes launched for each CPU/node in a parallel processing environment. These log files should be inspected after TIGCL terminates to ensure that no error messages or unnoticed errors appeared during the process (see Note 7). More information on using TGICL, can be found on the DFCI-Gene Index Project website (http://compbio.dfci.harvard.edu/tgi/software/) and in the documentation available within the distribution.

3.3. Distributed Computing (Local Grid and Amazon EC2)

If the infrastructure needed to deploy your own grid platform is not available, you can try "Fura In The Cloud" (http://www.grid-systems.com/?p=products/products.php&s=6). A cloud is a pool of highly scalable, virtualized infrastructures hosting an application that is billed by consumption such as Amazon Elastic Compute Cloud (Amazon EC2, http://aws.amazon.com/ec2/). Cloud computing capacity can be used for grid-enabled applications and users only pay for the computing capacity utilized. "Fura In The Cloud" acts as middleware that simplifies the way in which Amazon EC2 resources can be used. Included in both Fura open-source and "Fura In The Cloud" distributions, extensive documentation is provided with clear instructions about how to have a correct installation running and launching Blast, InterProScan, or other algorithms as if they were executed in the command line. However, the skills of an experienced developer are probably required to deploy your project applications with this middleware (see Note 8).

Details about running BLAST comparisons on a local installation of the Fura open-source distribution are provided in Note 9, and Note 10 describes a protocol to launch a Fedora 64-bit Amazon EC2 extra-large (8 CPUs) image, to serve as a basis to install, as an example, TGICL. Here, we outline the procedure to launch a virtual machine in Amazon EC2 that can run TGICL and other software:

1. Create an Amazon EC2 account at http://aws.amazon.com/
2. Sign up for Amazon EC2 service. You will be prompted for a suitable payment method, since this is a pay-per-use service.
3. Create an X.509 certificate.
4. Sign into the Amazon EC2 web console (http://aws.amazon.com/console/).
5. Launch a Quick Start AMI selecting a 64 bit Fedora extra-large image.

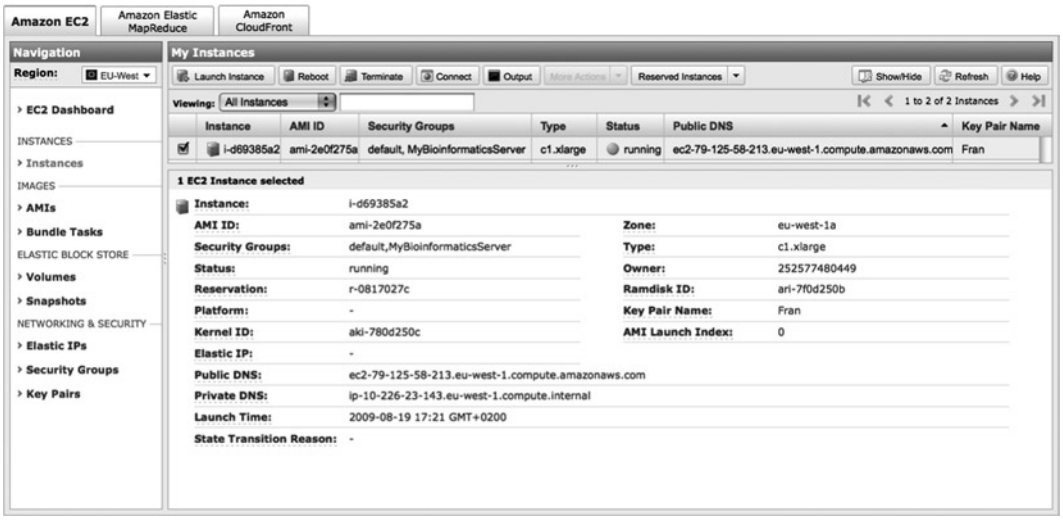

Fig. 2. *Screenshot that shows the* Amazon EC2 web console once the instance has been launched.

6. Create and download a key pair that will be used to enter in your Amazon EC2 instances (through SSH and SFTP protocols), once launched.
7. Open port 22 (via SSH) on your "security group" to enable access to this instance from the Internet.
8. Launch the extra-large instance.
   In the Amazon web console (http://console.aws.amazon.com) you could have something similar to Fig. 2. More documentation can be found at http://docs.amazonwebservices.com/AWSEC2/latest/GettingStartedGuide/.
9. Wait for 2–4 min and check that your instance is running and that you can access via SSH and SFTP.
10. Upload your files (for instance, the TGICL distribution and a zip file containing all the fasta sequences) using a remote client such asWinSCP (for Windows machines) or CyberDuck (for Mac OS X machines).

3.3.1. Run BLAST Searches

Figure 3 outlines how to build a grid module and launch a task that corresponds to the command line. To run BLAST searches, run the following command:

blastall -p tblastx -d /NCBI/nt -i <fasta_filename> *-o* <result_filename>*.xml -T T -m 7*, in which several fasta files (one for each EST) are swept in the comparison against the formatted database available in /NCBI/nt (in this example). The screenshots in Fig. 3 are from the Fura open-source graphical user interface. These are the steps that have to be configured through the portal of this middleware:

1. In *GridTools* > *File Viewer* navigate to */home/<your Fura user>* and create a <fasta directory> folder including all the sequences

Fig. 3. Key parameters needed to configure a distributed calculation of NCBI Blast using either Fura open source or "Fura In The Cloud" platforms.

in independent fasta files. All files can be uploaded in one zip file and decompressed automatically in the server.

2. In the same File Viewer, navigate to */modules* and create a *NCBIblast* directory, and click on *New platform* as many times as different operating systems you have in your grid. Inside each folder, upload the corresponding *blastall* executable. In the commonFiles folder, upload the selected substitution matrix you want to use for your BLAST analyses (for instance, *BLOSUM62)*.
3. In *GridStudio* > *Modules* > *New,* select the recently created *NCBIblast* directory and "*Add Platform*" for each of the operating systems you have, selecting the correct executable in the *General* tab. Name this module *NCBI_tblastx.* Fill in the form as indicated on the next picture.
4. In the *Ranges & Parameters* tab, "*Add range*" fasta_file in the iteration ranges section and fill the parameters section as detailed in the next picture.
5. In the *Execution control* tab, add a *.xml filter in the filtering of results section to capture the outputs of each individual execution. Up to this point, the module defines all the parameters

needed to run blast executions in the command line of each grid node. Fill in the form as indicated on the next picture.

6. Launch the grid-distributed execution. In *GridTask* > *Tasks* > *New*, select the recently created module *NCBI_tblastx* and give this task a name, for instance *NCBI_tblastx_execution*. Automatically, the parameters needed to run this calculation are displayed in a form. In this form, select the *Folder Path* in which the fasta files were saved in the first step and give a **fasta File Filter* to tell Fura that only those files with a fasta extension are to be distributed. Fill in the form as indicated on the next picture.

   This protocol assumes that all the nodes have access to a preexistent *NCBI* folder containing the *nt* database formatted with NCBI blast *formatdb* executable. This folder can be a shared folder or a replicated folder in each node (see Notes 11–13).

3.3.2. Run InterProScan Searches

InterProScan is available for running via the web site (http://www.ebi.ac.uk/InterProScan/). Alternatively, InterProScan and the underlying applications are freely available under the GNU license agreement from the EBI FTP server (ftp://ftp.ebi.ac.uk/pub/databases/interpro/iprscan/). An important feature of InterProScan is the possibility for distributed execution of individual jobs. The main script of the stand-alone version is *iprscan*, located in the *bin/* directory. The iprscan script starts jobs by calling another script (*iprscan_wrapper.pl*), which in turn launches and tracks jobs for each application included in the program. Results are parsed. When running the command line, the results are written by default to *stdout*, unless the *-o* option is used to redirect the output to a specified file. A normal way to launch *iprscan* in the command-line would be:

iprscan -cli -i <fasta file name>*.fasta -o* <output>.xml *-iprlookup -goterms -seqtype (n|p)*, where

- *cli* specifies to the script to be used in command line mode.
- *iprlookup* switches on look up of corresponding InterPro annotation.
- *goterms* switches on look up of corresponding Gene Ontology annotation.

It is relatively easy to configure a new module and task to distribute the *iprscan* command-line execution using the Fura middleware as detailed above for BLAST. An important difference is that you have to check *Pre-installed Software* in the General tab of the New Module description. To run *iprscan* in distributed computing mode, preinstall the software in all the nodes of the system, maintaining the following directories in the same root folder of *iprscan*:

- the *data* directory contains all databases and required indices.
- the *tmp* directory is used to store temporary user sessions and temporary jobs outputs.
- the *bin* directory contains some Perl scripts and platform specific binaries of scanning programs (in the binaries/ subdirectory).
- the *lib* directory contains all Perl modules necessary for *iprscan* to work properly.
- the *conf* directory contains configuration files for each database/application used.

3.4. Blast2GO

The Blast2GO application (12, 14), which can be downloaded from http://www.blast2go.de, has the following operational requirements: an Internet connection, Java 1.5 or higher, and an open network port 3306 for a direct MySQL connection to the Blast2GO database. Throughout this protocol, screenshots were obtained from Blast2GO v.2.3.6, and following the described protocol using a set of test DNA sequences ("*sequence_data_example.fasta*"), which can be downloaded from the "Downloads" section of the Blast2GO web page ("Example files to try and test Blast2GO: *b2g_example_file.zip*").

1. Open the Blast2GO application. Go to "File" -> "Load FASTA File" and select your file containing your set of EST sequences in fasta format. They will appear in the "Main Sequence Table."
2. BLAST the loaded sequences. Go to "Blast" -> "Make BLAST." The "Blast Configuration" dialog will pop up (Fig. 4). Different parameters can be configured here. The default options include: the nonredundant (nr) database as "Blast DB," the BLASTX (for DNA sequences) as "Blast Program" and 1.0E-3 as "Blast ExpectValue" (Fig. 4) (see Note 14).
   In our original pipeline we decided to use a "Blast ExpectValue" of 1.0E-5. Lower thresholds are more stringent, leading to fewer chance matches being reported.
3. Click on the "arrow" icon in the top left part of the "BLAST Configuration" dialog to start the BLAST search. Then, choose the file format to save the BLAST results. This can be done by selecting the corresponding check boxes. By default, the results will be stored in .xml format. However, results can also be saved as .txt and/or .html files. Once your analysis is complete, visualize your results at "Statistics"->"BLAST Statistics." As the BLAST search progresses, sequences with successful results will change color on the "Main Sequence Table" from white to light-red, and the corresponding BLAST result-related col-

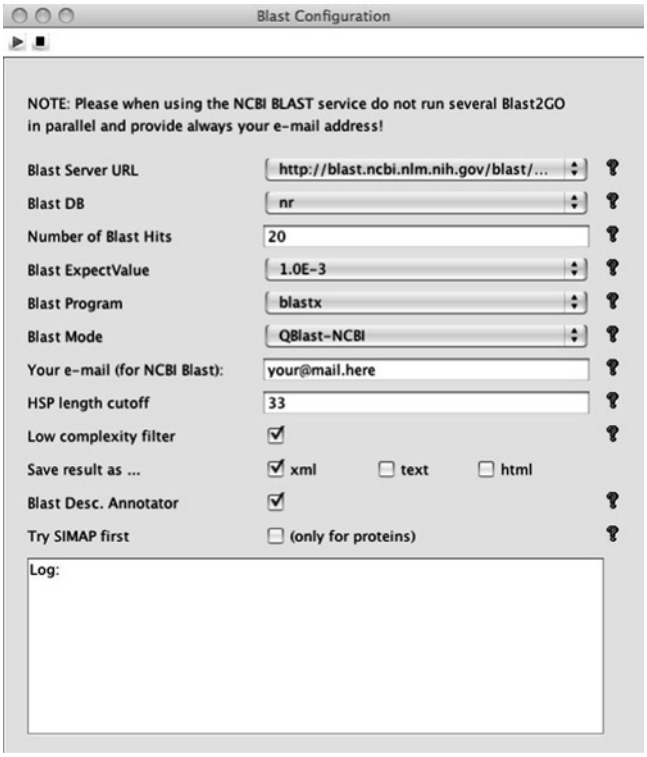

Fig. 4. Screenshot that shows the BLAST2GO "Blast Configuration" dialog.

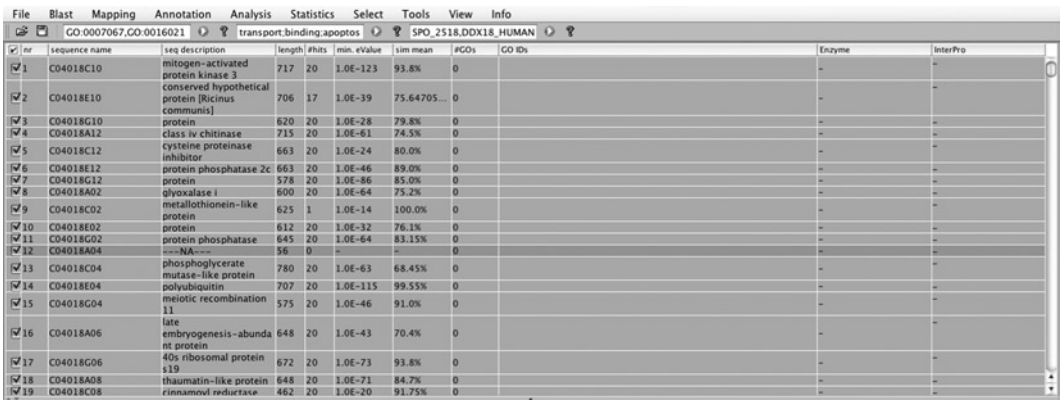

Fig. 5. Screenshot from the BLAST2GO "Main Sequence Table" after the BLAST step has been performed.

umns will be filled. If no results are found for a given sequence, the corresponding row will turn dark-red (Fig. 5).

4. Perform the mapping step. Mapping is the process of retrieving GO terms associated to the hits after a BLAST search. Go to "Mapping"->"Run GO-Mapping Step." The "Mapping" annotation dialog will appear with an explanation of the four different types of mappings that Blast2GO performs. To start

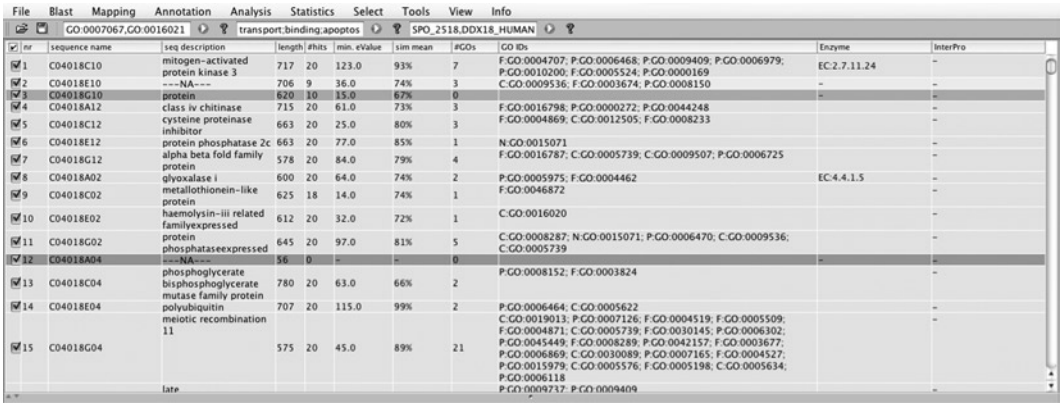

Fig. 6. Screenshot from the BLAST2GO "Main Sequence Table" after the annotation step has been performed.

mapping, click on the "arrow" icon on the top left part of the window. Once your analysis is finished, you can visualize your results at "Statistics"->"Mapping Statistics." When at least one GO term is successfully mapped to a BLAST result, the sequence row position will turn light green.

5. Perform the annotation step. This is the process of selecting GO terms from the GO pool obtained by the mapping step and assigning them to the query sequences. Go to "Annotation"->"Run Annotation Step." The "Annotation Configuration" dialog will pop up. To start the annotation, click again on the "arrow" icon (see Note 15). Once the annotation is complete, visualize the results at "Statistics"->"Annotation Statistics." Successfully annotated sequences will turn blue, and only the annotated GOs will remain in the GO IDs column.
6. Additional annotations can be performed as optional steps, such as InterPro enzyme code annotation or KEGG mapping. Go to "Annotation"->"InterProScan" or "Annotation"->"Enzyme Code and KEGG," respectively (see Note 16). The corresponding columns will be filled in the "Main Sequence Table."
7. The last step is to export the annotations. This function is available at "File"->"Export"->"Export Annotations." The default option is a *.annot* file, in which annotations are provided in a three column, tab-delimited file. The first column contains the sequence name, the second the annotation code, and the third, the sequence description (Fig. 6).

4. Notes

1. After a 454 instrument completes a run, output data is deposited into a run folder. This folder will contain a subdirectory

named "SFF," which will contain one or more SFF (Standard Flowgram Format) files. Initially, one SFF file will be generated for each 454 region, and they can be freely concatenated together to represent an entire run in one SFF file if desired. High-quality base-calls will be extracted from these SFF files with determination criteria incorporated in the 454 software. Note that prior to version 1.0.52 of the 454 software, SFF files were not created automatically during the run, but instead were created from other data files deposited in the analysis directory using the script *sffcall* after the run was completed. An SFF file can hold one or more sequences (typically it will consist of all the reads of a single 454 run, which may number in the hundreds of thousands). Base calls and quality values (which use a Phred-like scoring system) can be extracted on a per-read basis from an SFF file analogous to how similar data would be extracted from a standard SCF trace file. Once the SFF file is transferred to a working directory, high-quality base calls (as determined by the 454 base calling software) can be extracted for each sequence with the program *sffinfo*, which is included in the 454 software package and is easy to use. To extract the base calls for the high-quality region of all traces described by a given SFF file run this command: *sffinfo -s* <SFF file filename> > <fasta output file>, which will deposit fasta nucleotide sequence into the <fasta output file> representing what the 454 software believes to be the high-quality region of each individual read. Additionally, if using the option *-q* <QUAL file filename> the output will be a QUAL format using Phred quality scores.

2. If using Phred for previous sequencing platforms, a fasta file can be extracted from .abi or .scf files with the command: *phred* <name of abi file> *-trim_alt ″″ -trim_out -c* <name of output for scf file> *-qa* <name of output for qual file> *-sa* <name of output for ph1 file> *-p* <name for ouput of the phfile>. Phred can produce a variety of different output files: SCF files containing Phred base calls and quality values, sequence files in fasta (or XBAP) format, quality files in fasta (or XPAB) format, or PHD files - text files which contain base call and quality information, which can later be used during the contig editing by Consed and similar programs (15, 16). Phred makes 40–50% fewer errors than the ABI software (5).
3. If the 454 software cannot be used, there is a free alternative: use *sff_extract* in the third party scripts from the MIRA package (http://www.chevreux.org/mira_downloads.html).
4. Alternatively, in a local computer, one of these two software packages can be used for the initial steps of parsing the raw chromatograms and set-up the contigs: MIRA (http://www.chevreux.org/projects_mira.html) or CodonCode Aligner (http://www.codoncode.com/aligner/). MIRA is a whole-

genome shotgun and EST sequence assembler for the Sanger, 454 and Solexa /Illumina platforms. MIRA is free and runs on Linux but it can also be compiled for other platforms. On the other hand, CodonCode Aligner is a commercial program for sequence assembly, contig editing, and mutation detection, available for Windows and Mac OS X.

5. Phred uses data from a chemistry parameter file called *phredpar.dat* in order to identify dye primer data. Set the environment variable "PHRED_PARAMETER_FILE" to point to the full path name of the file. For example, if you copy the file *phredpar.dat* to*/usr/local/etc/PhredPar*, and you are using the C shell, then issue the command *setenv PHRED_PARAMETER_FILE /usr/local/etc/PhredPar/phredpar.dat*. It is most convenient to set the environment variable in the system-wide shell startup (*cshrc* or equivalent) file. ename The Phred parameter file can be renamed, but the PHRED_PARAMETER_FILE environment variable must reflect the new name. With Windows NT, give the command *set PHRED_PARAMETER_FILE=\usr\local\etc\PhredPar\phredpar.dat* in the DOS command window in which you will run Phred.
6. After obtaining the cleaned sequences, trim the sequences to eliminate vector/linker contamination with one of several binaries from the EMBOSS package (for instance, *vectorstrip*, *trimseq,* and *trimest*) (17).
7. This procedure can be executed on a machine with several CPUs. If these resources are not available, it is possible to pay for the use of a Linux machine using Amazon EC2 Machine images.
8. The open-source version of Fura can be downloaded from http://fura.sourceforge.net. The project documentation describes how to build the grid middleware. If you prefer not to build the middleware, you can use "Fura In The Cloud" at Amazon EC2.
9. To set-up a distributed calculation of any particular algorithm in the Fura middleware, take into account that a Fura module has to be written that converts each argument in the normal command line into a parameter in the Fura module, using GridStudio (the graphical user interface of Fura).
10. The TGICL clustering method can be executed directly with the software installed on an Amazon EC2 Machine Image provisioning a base Linux installation of Fedora or Ubuntu. It is very useful to be skilled at using Amazon EC2 developer tools and have other tools such as ElasticFoxa installed on your browser.
11. An alternative is to use AutoFACT (Automatic Functional Annotation/Classification Tool) (18). AutoFACT is a Perl script that reads a FASTA sequence file and the corresponding BLAST output files, and then performs automatic functional

annotation. AutoFACT is very useful to generate outputs in HTML, text, and GFF formats, depending on the user's preference. BLAST calculations in AutoFACT can be distributed in several grid computing resources editing the script.

12. Once you have all the XML results files coming from the TBLASTX comparison, you can use a free blast viewer like Korilog Blast Viewer (http://www.korilog.com/index.php/BlastViewer.html).
13. A very good parser in C# ("BlastXML2Database – bxml2db," http://frenesssi.wordpress.com/2008/06/21/blastxml2database-bxml2db-v01/), that can be run using the mono platform (http://www.mono-project.com/) has been written by J. Cervantes. This application directly parses BLAST results on XML into MySQL or other database format tables. If you would like to use this code, contact him at jacobnix@gmail.com.
14. More parameters in the BLAST search can be configured in the "Blast Configuration" dialog. To change some options, you will need to edit the Blast2GO properties file that can be found in the local user profile directory (e.g., under Linux, */home/yourname/blast2go/blast2go.properties*).
    - In "Number of Blast hits" choose the number of alignments that will be stored (1–100, default is 20).
    - In "Blast Server URL" specify your own BLAST server direction (you will need to edit the Blast2GO properties file). To make this possible, change the default "Blast Mode" from "QBlast-NCBI" to "WWW-Blast."
    - In "HSP length cutoff," the cut-off value for the minimal length of the first High-scoring Segment Pair (HSP) of a BLAST hit. This can be used to exclude hits with only small local alignments from the BLAST result. The default value is 33 for nucleotides.
15. Evidence Code Weights can be modified at "Annotation"-> "Evidence Code Weights." Note that IEA (Inferred from Electronic Annotation) is usually overwhelmed in the mapping results. The contribution of the type of annotation can be modulated in the annotation step. If you want to exclude GO annotations of a certain EC (for instance IEA), you can set this EC weight at 0. Alternatively, if no influence by evidence codes is desired, all ECs must be set to 1.
16. At any point during the different BLAST2GO annotation steps (BLAST search, Mapping process or Annotation), the "Single Sequence Menu" can be accessed by right-clicking on a given sequence row in the Main Sequence Table. Some of the functionalities that can be found there include: "Show Blast Result," "Show GO Descriptions," and "Change Annotation and Description" (this function allows to edit manually annotations and sequence descriptions).

Acknowledgments

The authors would like to thank everyone involved in the "TrichoEST" project, especially Prof. E. Monte, Prof. A. Llobell and Dr. M. Rey.

References

1. Liolios, K., Mavromatis, K., Tavernarakis, N., and Kyrpides, N. C. (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata, *Nucleic Acids Res* ***36***, D475–479.
2. Vizcaino, J. A., Gonzalez, F. J., Suarez, M. B., Redondo, J., Heinrich, J., Delgado-Jarana, J., Hermosa, R., Gutierrez, S., Monte, E., Llobell, A., and Rey, M. (2006) Generation, annotation and analysis of ESTs from *Trichoderma harzianum* CECT 2413, *BMC Genomics 7*, 193.
3. Vizcaino, J. A., Redondo, J., Suarez, M. B., Cardoza, R. E., Hermosa, R., Gonzalez, F. J., Rey, M., and Monte, E. (2007) Generation, annotation, and analysis of ESTs from four different *Trichoderma* strains grown under conditions related to biocontrol, *Appl Microbiol Biotechnol 75*, 853–862.
4. Ansorge, W. J. (2009) Next-generation DNA sequencing techniques, *N Biotechnol* ***25***, 195–203.
5. Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment, *Genome Res* ***8***, 175–185.
6. Ewing, B., and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities, *Genome Res* ***8***, 186–194.
7. Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., and Quackenbush, J. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets, *Bioinformatics* ***19***, 651–652.
8. Huang, X., and Madan, A. (1999) CAP3: A DNA sequence assembly program, *Genome Res* ***9***, 868–877.
9. Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., Finn, R. D., Gough, J., Haft, D., Hulo, N., Kahn, D., Kelly, E., Laugraud, A., Letunic, I., Lonsdale, D., Lopez, R., Madera, M., Maslen, J., McAnulla, C., McDowall, J., Mistry, J., Mitchell, A., Mulder, N., Natale, D., Orengo, C., Quinn, A. F., Selengut, J. D., Sigrist, C. J., Thimma, M., Thomas, P. D., Valentin, F., Wilson, D., Wu, C. H., and Yeats, C. (2009) InterPro: the integrative protein signature database, *Nucleic Acids Res 37*, D211–D215.
10. Mulder, N., and Apweiler, R. (2007) InterPro and InterProScan: tools for protein sequence classification and comparison, *Methods Mol Biol* ***396***, 59–70.
11. Conesa, A., Gotz, S., Garcia-Gomez, J. M., Terol, J., Talon, M., and Robles, M. (2005) Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research, *Bioinformatics* ***21***, 3674–3676.
12. Gotz, S., Garcia-Gomez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talon, M., Dopazo, J., and Conesa, A. (2008) High-throughput functional annotation and data mining with the Blast2GO suite, *Nucleic Acids Res* ***36***, 3420–3435.
13. Richterich, P. (1998) Estimation of errors in "raw" DNA sequences: a validation study, *Genome Res* ***8***, 251–259.
14. Aparicio, G., Gotz, S., Conesa, A., Segrelles, D., Blanquer, I., Garcia, J. M., Hernandez, V., Robles, M., and Talon, M. (2006) Blast2GO goes grid: developing a grid-enabled prototype for functional genomics analysis, *Stud Health Technol Inform* ***120***, 194–204.
15. de la Bastide, M., and McCombie, W. R. (2007) Assembling genomic DNA sequences with PHRAP, *Curr Protoc Bioinformatics Chapter 11*, Unit11.4.
16. Gordon, D. (2003) Viewing and editing assembled sequences using Consed, *Curr Protoc Bioinformatics Chapter 11*, Unit11.2.
17. Rice, P., Longden, I., and Bleasby, A. (2000) EMBOSS: the European Molecular Biology Open Software Suite, *Trends Genet* ***16***, 276–277.
18. Koski, L. B., Gray, M. W., Lang, B. F., and Burger, G. (2005) AutoFACT: an automatic functional annotation and classification tool, *BMC Bioinformatics* ***6***, 151.

Chapter 8

The Application of ChIP-chip Analysis in the Rice Blast Pathogen

Soonok Kim and Thomas K. Mitchell

Abstract

To attempt to gain an understanding of the molecular underpinnings of disease, many researchers have turned to expression profiling of genes during various stages of host recognition, entry, invasive growth, and host responses. While these studies have proven valuable, a deeper level of knowledge of the control circuitry affecting observed gene expression profiles can lead to a better understanding of the host pathogen interaction. Transcription factors are key switches in signal transduction circuits regulating gene expression. One powerful method to define target sequence specificity for this important group of transcription regulators is chromatin immunoprecipitation (ChIP) with microarray chips (chip), commonly called ChIP-chip. A more recent variation of this technique is ChIP-seq where DNA sequencing replaces the microarray chip. Here, we describe how we elucidated the binding sites for the *Magnaporthe oryzae* Ca^{2+}/calcineurin-dependent transcription factor *MoCRZ1* with the ChIP-chip approach.

Key words: Chromatin immunoprecipitation, ChIP-chip, *Magnaporthe oryzae*, MoCRZ1

1. Introduction

All organisms have the innate ability to perceive their environment and respond to it largely through controlling gene expression. The movement of information through the cell into the nucleus is the work of signal transduction circuits common among most organisms. Specificity of a response is primarily achieved through unique receptors, downstream transcription factors, and the genes these transcription factors regulate. For fungal plant pathogens, conserved signal transduction cascades are involved in

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722, DOI 10.1007/978-1-61779-040-9_8,

perception of hosts, transgression of physical barriers, suppression or elicitation of host defenses, *in planta* nutrient acquisition, and completion of their life cycle. To date, signaling networks common to all eukaryotic organisms, such as cAMP, MAP kinase, and Ca^{2+}, have been described (1–6). We know that the Ca^{2+}/calcineurin signaling pathway is a central conduit regulating aspects of growth, development, and infection for many fungal plant pathogens (4, 7–9). We also know that much of the yeast Ca^{2+}/calcineurin signaling machinery are conserved in filamentous fungi (6). This case study will focus on the rice blast pathogen *Magnaporthe oryzae*. Like the other more recognized model systems, the rice blast pathogen *M. oryzae* has all the trappings of an ideal model complete with extensive genomic resources, a sequenced genome, and sophisticated bioinformatic tools (10–12). Most importantly, *M. oryzae* has the advantage in that its importance in worldwide nutrition, economy, and social rest make new discoveries immediately impacting and relevant (13). Moreover, much of what we have learned from *M. oryzae* (specifically as it relates to signaling) has had direct bearing on other fungal pathogens of plants and animals that are less biologically tractable. As such, the impact of understanding the molecular circuitry underpinning fungal disease and virulence for *M. oryzae* has a broad and far reaching impact.

We applied ChIP-chip technology to focus on a central node in the Ca^{2+} signaling network, specifically calcineurin and the transcription factor it regulates, MoCRZ1. Briefly, MoCRZ1 is activated in a Ca^{2+}/calcineurin-dependent manner and regulates calcium homeostasis, host penetration, and cell wall maintenance (14). Using ChIP-chip combined with comprehensive expression profiling, we have revealed several new finding on the control of Ca^{2+} signaling and virulence.

The ChIP-chip approach allows for the comprehensive identification of target sequences and downstream genes of transcription factors (15, 16). Cells are briefly fixed to crosslink transcription factor proteins to the DNA fragments that they bind in vivo. The cells are lysed, the chromatin is sheared, and the transcription factor with its associated DNA is immunoprecipitated. In the protocol below, we tagged MoCRZ1 with GFP; however, other tags such as Tap or His can be used. The bound DNA fragments are then recovered and fluorescently labeled and hybridized to a DNA chip harboring all the regulatory (intergenic) regions of the genome investigated (17, 18). Alternatively, fragments can be amplified and sequenced (19). On average, each transcription factor in the yeast genome interacts with 38 target genes (range 0–181) (20). In *M. oryzae*, we identified 346 potential MoCRZ1 binding sites.

2. Materials

2.1. Sample Preparation

1. *M. oryzae* strain: KJ201 over-expressing the *MoCRZ1*-GFP construct (see Note 1).
2. 1 M $CaCl_2$: Filter sterilized with 0.22-μm bottle top filter.
3. 5 μg/μl FK506: Resuspended in DMSO. Store at –20°C.
4. Buffer A: 0.4 M sucrose, 10 mM Tris–HCl, pH 8.0, 1 mM EDTA, 1 mM PMSF, 1% formaldehyde. Store at room temperature (PMSF and formaldehyde should be added just before use).
5. 100 mM PMSF: Dissolve 0.1742 g PMSF (Sigma–Aldrich, St. Louis, MO) in 10 ml isopropanol. Wrap the tube in foil and store at room temperature. PMSF inhibits serine proteases.
6. 2 M Glycine.
7. Miracloth (Calbiochem, La Jolla, CA).

2.2. Nuclei Isolation

1. CelLytic™ PN Plant Nuclear Isolation Kit (Sigma–Aldrich).
2. 1× Nuclear isolation buffer (NIB): 500 μl 4× NIB from the CelLytic™ PN Plant Nuclear Isolation Kit, 20 μl 100 mM DTT, 1,480 μl dH_2O.
3. 100 mM DTT.
4. NIBA: Add 15 μl of proteinase inhibitor cocktail into 1,500 μl of NIB.
5. Proteinase inhibitor cocktail (Sigma–Aldrich).
6. 10% Triton X-100: Autoclave and store at room temperature.
7. Lysis buffer (nuclear membrane lysis buffer): 50 mM HEPES pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% deoxycholate (Sigma D6750), 0.1% SDS (10%), 10 mM sodium butyrate (Sigma–Aldrich), 1 mM PMSF, 1% (v/v) proteinase inhibitor cocktail (Sigma–Aldrich) (see Note 2).

2.3. DNA Fragmentation and Chromatin Immunoprecipitation

1. Biorupter Sonicator (Cosmo Bio, Tokyo, Japan).
2. Salmon sperm/protein A agarose (Upstate Biotechnologies, Lake Placid, NY).
3. GyroMini Nutating Mixer (Labnet, Edison, NJ).
4. Anti-GFP antibody (Abcam, Cambridge, MA).
5. Rabbit anti-IgG antibody (Abcam).
6. Input DNA: Keep at 4°C.

7. LNDET: 0.25 M LiCl, 1% NP40 (Nonidet® P40, Fluka 74385), 1% deoxycholate, 1 mM EDTA.
8. Elution buffer: 1% SDS, 0.1 M $NaHCO_3$ (sodium bicarbonate), 0.25 mg/ml proteinase K, 1 mM DTT.
9. TE: 10 mM Tris–HCl (pH 8.0), 1 mM EDTA (pH 8.0).

2.4. Purification of Chromatin Immunoprecipitated (ChIPed) DNA

1. 25 mg/ml RNaseA: Store at –20°C.
2. Qiaquick PCR purification kit (Qiagen).
3. EB (Elution buffer) from the Qiaquick PCR Purification Kit.
4. ND-1000 UV/Vis Spectrophotometer (Nanodrop Technologies).

2.5. Amplification of ChIPed DNA

1. Primers A and B: Primers designed according to sequences at the promoter region of known direct control targets.
2. Platinum Taq (Invitrogen).

2.6. Whole Genome Amplification Hybridization to Microarray Chips

1. GenomePlex Whole Genome Amplification Kit (WGA2, Sigma).
2. Library preparation buffer.
3. Library stabilization solution.
4. Library preparation enzyme.
5. Qiaquick PCR purification kit (Qiagen).

3. Methods

This protocol is optimized for small-scale chromatin immunoprecipitation. To minimize the variation between biological replicates, it is important to obtain ample fungal biomass for each replicate although only a fraction of the sample will be used. The overall procedure of ChIP-chip experiments is illustrated in Fig. 1.

3.1. Sample Preparation

1. Prepare mycelia with test and control treatments. We treated mycelia grown in liquid culture with 200 mM $CaCl_2$ for 1 h with shaking with or without 10 μg/ml FK506 calcineurin inhibitor (see Note 3).
2. Prepare buffer for crosslinking (Buffer A).
3. Collect mycelia with one layer of Miracloth and wash with excess water.
4. Crosslink by resuspending mycelia in Buffer A for 20 min on a shaker. 1% formaldehyde is enough to crosslink, however, up to 3% has been reported.

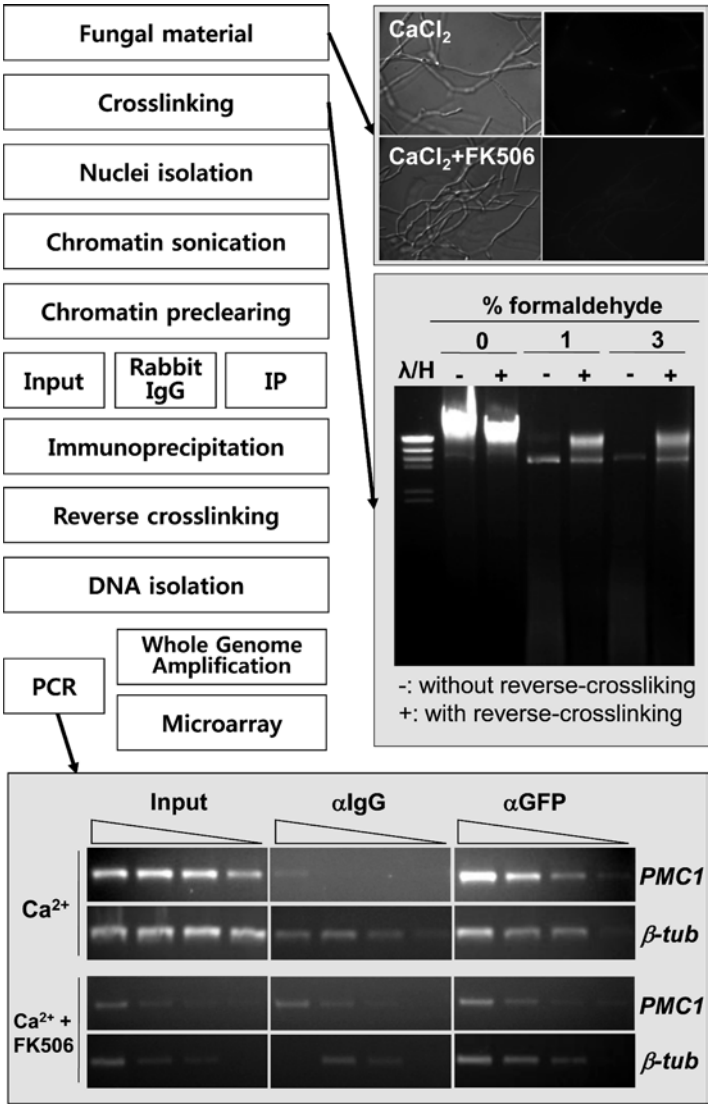

Fig. 1. ChIP-chip experimental work flow and validation steps.

5. Stop crosslinking by adding 1/20 volume of 2 M glycine (final concentration 0.1 M), continue incubation for 5–10 min.
6. Collect mycelia with two layers of Miracloth and rinse with excess distilled water. Squeeze to remove as much water as possible.
7. Freeze immediately with liquid nitrogen. Grind with pre-chilled mortar and pestle, and store at −80°C until used.

3.2. Nuclei Isolation

1. Prepare NIB (nuclear isolation buffer, included in kit):
   - 1.5 ml/sample is required.
   - For 2 ml NIB (/sample) – 500 µl 4× NIB, 20 µl 100 mM DTT, 1,480 µl dH_2O.

     *After using 500 µl for resuspending mycelia, add 15 µl (1:100 (v/v)) proteinase inhibitor cocktail to make NIBA.
2. Weigh the frozen mycelia and resuspend 50–100 mg mycelia in 500 µl NIB by gentle tapping.
3. Filter through two layers of Miracloth.
4. Collect the nuclei by centrifuging at 1,200×*g* for 10 min at 4°C.
5. Discard the supernatant and resuspend the pellet into 500 µl of NIBA by brief vortex at half power.
6. Lyse cell membrane by adding 15 µl of 10% Triton X-100 to a final concentration of 0.3%. Mix by gentle tapping or inverting.
7. Centrifuge at 10,000×*g* for 10 min at 4°C.
8. Discard the supernatant and resuspend the pellet into 500 µl NIBA by vortexing 5 s.
9. Centrifuge at 10,000×*g* for 10 min at 4°C.
10. Discard the supernatant and resuspend the pellet into 300 µl nuclear membrane lysis buffer by sonication until resuspended.

3.3. DNA Fragmentation and Chromatin Immunoprecipitation

1. Sonicate with Biorupter (Cosmo Bio, Tokyo, Japan) in the following condition: 30 s ON and 30 s OFF at power level H for total 30 min. Change ice every 10 min (see Note 4).
2. Centrifuge at 10,000×*g* for 5 min at 4°C.
3. Preclear supernatant with 30 µl of salmon sperm/protein A agarose for ~4 h (minimum 1 h) with rotation on GyroMini Nutating Mixer at 4°C.
4. Centrifuge at 3,000 rpm for 1 min at 4°C.
5. Transfer 100 µl each supernatant into three tubes. Add 0.5 µl of anti-GFP antibody in one tube (αGFP), 1 µl anti-IgG antibody into another tube (αIgG). The third tube is for input DNA (Input).
6. Incubate overnight with rotation on GyroMini Nutating Mixer at 4°C. Keep the input DNA at 4°C.
7. Add 30 µl of Salmon sperm/protein A agarose, continue incubation for ~4 h (minimum 2 h).
8. Wash in cold room:
   (a) Add 0.5 ml of Lysis buffer, inverse six times, centrifuge at 3,000 rpm for 1 min at 4°C, and discard the supernatant.

(b) Add 0.5 ml of Lysis buffer, rotate for 5 min, centrifuge, and discard the supernatant.

(c) Add 0.5 ml of LNDET, invert six times, centrifuge, and discard the supernatant.

(d) Add 0.5 ml of LNDET, rotate for 5 min, centrifuge, and discard the supernatant.

(e) Add 0.5 ml of TE, invert six times, centrifuge, and discard the supernatant.

(f) Add 0.5 ml of TE, rotate for 5 min, centrifuge, and discard the supernatant.

9. Elute with 40 µl of Elution buffer and incubate at 65°C for 10 min. Centrifuge and transfer the supernatant to a new tube.
10. Elute again with 40 µl of Elution buffer. Final elution volume is 80 µl. In parallel, add 50 µl of Elution buffer into 30 µl of input fraction for the 30% input control.
11. Incubate overnight at 65°C for reverse crosslinking.

3.4. Purification of ChIPed DNA (see Note 5)

1. Add 1 µl of RNase A and incubate for 30 min at room temperature.
2. Extract DNA by using the Qiaquick PCR purification kit (Qiagen). Elute in 30 µl of EB.
3. Quantify input DNA using Nanodrop.

3.5. Amplification of ChIPed DNA

1. Set up PCR reactions with 0.1–4 µl of ChIPed DNA to test the efficacy of ChIP.
   - Perform PCR on all samples: Input, αIgG, and αGFP
   - Reaction components (20 µl): Make master-mix for all samples for each primer.

Component	Per Rxn (20 µl)	Per µl	For 400 µl
10× PCR buffer w/o $MgCl_2$	2.0 µl	0.1 µl	40 µl
25 mM $MgCl_2$	0.6 µl	0.03 µl	12 µl
dNTPs 10 mM	0.4 µl	0.02 µl	8 µl
Primer A (10 µM)	0.5 µl	0.025 µl	10 µl
Primer B (10 µM)	0.5 µl	0.025 µl	10 µl
dH_2O	15.92 µl	0.796 µl	318.4 µl
Platinum Taq	0.08 µl	0.004 µl	1.6 µl

Use 4× dilution series for this PCR reaction.

Transfer 19 + 15 + 15 + 15 µl of master-mix w/o template into each PCR tube.

For each treatment, the following amount of Rxn mix was required:

– Input DNA: 19 + 15 + 15 + 15

αIgG: 19 + 15 + 15 + 15

– αGFP: 19 + 15 + 15 + 15

- For treatment and control samples $(19 \times 3 \times 2) + (15 \times 9 \times 2) = 384$ µl of master-mix is required.
  - After adding 1 µl of template into 19 µl, mix it well by pipetting and transfer 5 µl into second tube. Mix well and transfer 5 µl into third tube. From the final tube, discard 5 µl to make 15 µl.

2. Run PCR with following conditions:
   - 94°C 3 min for initial denaturation.
   - 94°C 30 s – 55–60°C 30 s – 72°C 30 s for 35 cycles.
   - 72°C 5 min for final extension.
3. Check by electrophoresis, and reduce cycles to get the optimum result for input/GFP ratio.

3.6. Whole Genome Amplification and Hybridization to Microarray Chip

Amplify using the GenomePlex Whole Genome Amplification Kit (Sigma) following manufacturer's instruction with slight modification (see Note 6).

1. Library preparation.
   (a) Take 10 µl of 10 ng of ChIPed DNA (αGFP).
   (b) Take 10 ng of input DNA, and make to 10 µl by adding ultrapure water.
   (c) Add 2 µl of 1× Library preparation buffer to each sample.
   (d) Add 1 µl of Library stabilization solution.
   (e) Mix well by vortexing, spin down, and incubate at 95°C for 2 min.

   After cooling the sample on ice, add 1 µl of Library preparation enzyme, vortex thoroughly, and centrifuge briefly.
   (f) Place the sample in a thermal cycler and incubate as follows:

   16°C for 20 min (precool cycler to this temperature)

   24°C for 20 min

   37°C for 20 min

   75°C for 5 min

   4°C hold

(g) Remove samples from thermal cycler and spin down. Samples may be amplified immediately or stored at –20°C for 3 days.

2. Amplification
   (a) Prepare master-mix with the followings for each sample and add 60 µl:

   7.5 µl of 10× Amplification master-mix.

   47.5 µl of nuclease-free water.

   5 µl of WGA DNA polymerase.

   For multiple samples, multiply by the number of samples, and add 1/10 volume extra of each component.

   (b) Add 60 µl master-mix to 15 µl sample.

   (c) Mix thoroughly by vortexing, spin down, and PCR in the following condition:

   95°C for 3 min (initial denaturation)

   14 cycles of

   94°C for 15 s (denature)

   65°C for 5 min (anneal/extend)

   4°C hold

   (d) Purify amplified DNA with Qiaquick PCR purification kit (Qiagen). Elute with water.

   (e) Measure the DNA concentration using Nanodrop.
3. Re-amplification (see Note 7)
   (a) Use 10 ng each of the amplified DNA, and add ultrapure water to make 10 µl.

   (b) Repeat amplification step as described above.
4. Check the efficacy of ChIPed DNA by PCR described in Subheading 3.5.
5. Follow specifications provided by microarray chip producer/processing unit (see Note 8).

4. Notes

1. In this case, we over-expressed a *MoCRZ1*-GFP construct in the wild-type background. An alternated approach can be to complement a null mutant with the tagged transcription factor gene.
2. PMSF and proteinase inhibitor should be added just before use. For 1 ml, add 10 µl 100 mM PMSF + 10 µl proteinase inhibitor cocktail to 980 µl of the lysis buffer mixture.

3. It is highly recommended to observe the nuclear localization of the transcription factor in the test and not in the negative control condition prior to ChIP. We used GFP-tagged MoCRZ1 to visually validate our test and negative control condition, i.e., Ca^{2+} treatment resulting in nuclear translocalization of MoCRZ1 and Ca^{2+}/FK506 treatment blocking nuclear localization, respectively.
4. Sonication is the most important step for ChIP. Conditions depend on the type of sonicator. If other sonicators are used (other than Biorupter), conditions need to be optimized prior to ChIP to break nuclear DNA to an average size of 500 bp ranging from 200 to 1,000 bp. Maximum volume for Biorupter is 300 µl. If the sample volume exceeds 300 µl, split into two tubes
5. ChIPed DNA should be confirmed for the enrichment of transcription factor bound fragments by PCR with serial dilutions or by real-time PCR using primer sets for putative target gene(s). For this ChIP-chip case study, we compared the amount of DNA in the ChIPed sample (αGFP) with that of the input sample (input is cross-linked material that has not been immunoprecipitated) and in. Alternative approaches include comparing a ChIPed deletion mutant with that of wild type. A second negative control condition is needed to ascertain the specificity of enriched fragments in the test sample. In this case, it was the FK506 treatment.
6. The first DNA fragmentation step was omitted because the ChIPed DNA was already sheared.
7. Total recovery for immunoprecipitated samples will be in the 1–4 µg range. Re-amplification with amplified DNA can be applied to amass the DNA amount needed for ChIP-chip.
8. There are several data analysis tools and over 13 different algorithms available to analyze ChIP-chip results. For a comprehensive review of the most used tools, refer to Johnson et.al. 2008 (21). Additionally, most microarray producers have custom platforms to view and analyze ChIP-chip data.

References

1. Bruno, K.S., Tenjo, F., Li, L., Hamer, J.E., and Xu, J.R. (2004) Cellular localization and role of kinase activity of *PMK1* in *Magnaporthe grisea*. Eukaryot Cell. *3*: 1525–32.
2. Dean, R.A. (1997) Siganl pathways and appressorium morphogenesis. Annual Rev. Phytopathology. *35*: 211–234.
3. Dixon, K.P., Xu, J.R., Smirnoff, N., and Talbot, N.J. (1999) Independent signaling pathways regulate cellular turgor during hyperosmotic stress and appressorium-mediated plant infection by *Magnaporthe grisea*. Plant Cell. *11*: 2045–58.
4. Lee, S.C. and Lee, Y.H. (1998) Calcium/calmodulin-dependent signaling for appressorium formation in the plant pathogenic fungus *Magnaporthe grisea*. Mol Cells. *8*: 698–704.
5. Mitchell, T.K. and Dean, R.A. (1995) The cAMP-dependent protein kinase catalytic

subunit is required for appressorium formation and pathogenesis by the rice blast pathogen *Magnaporthe grisea*. Plant Cell. *7*: 1869–78.

6. Zelter, A., Bencina, M., Bowman, B.J., Yarden, O., and Read, N.D. (2004) A comparative genomic analysis of the calcium signaling machinery in *Neurospora crassa, Magnaporthe grisea,* and *Saccharomyces cerevisiae*. Fungal Genet Biol. *41*: 827–41.
7. Ahn, I.P. and Suh, S.C. (2007) Calcium restores prepenetration morphogenesis abolished by methylglyoxal-bis-guanyl hydrazone in *Cochliobolus miyabeanus* infecting rice. Phytopathology. *97*: 331–7.
8. Steinbach, W.J., Cramer, R.A., Jr., Perfect, B.Z., Asfaw, Y.G., Sauer, T.C., Najvar, L.K., Kirkpatrick, W.R., Patterson, T.F., Benjamin, D.K., Jr., Heitman, J., and Perfect, J.R. (2006) Calcineurin controls growth, morphology, and pathogenicity in *Aspergillus fumigatus*. Eukaryot Cell. *5*: 1091–103.
9. Yang, Y., Cheng, P., Zhi, G., and Liu, Y. (2001) Identification of a calcium/calmodulin-dependent protein kinase that phosphorylates the Neurospora circadian clock protein FREQUENCY. J Biol Chem. *276*: 41064–72.
10. Dean, R.A., Talbot, N.J., Ebbole, D.J., Farman, M.L., Mitchell, T.K., Orbach, M.J., Thon, M., Kulkarni, R., Xu, J.R., Pan, H., Read, N.D., Lee, Y.H., Carbone, I., Brown, D., Oh, Y.Y., Donofrio, N., Jeong, J.S., Soanes, D.M., Djonovic, S., Kolomiets, E., Rehmeyer, C., Li, W., Harding, M., Kim, S., Lebrun, M.H., Bohnert, H., Coughlan, S., Butler, J., Calvo, S., Ma, L.J., Nicol, R., Purcell, S., Nusbaum, C., Galagan, J.E., and Birren, B.W. (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature. *434*: 980–6.
11. Mitchell, T.K., Thon, M., Jeong, J.S., Brown, D., Deng, J., and Dean, R.A. (2003) The rice blast pathosystem as a case study for the development of new tools and raw materials for genome analysis of fungal plant pathogens. New Phytologist. *159*: 53–61.
12. Xu, J.R. and Xue, C. (2002) Time for a blast: genomics of *Magnaporthe grisea*. Mol Plant Pathol. *3*: 173–176.
13. Talbot, N.J. (2003) On the trail of a cereal killer: Exploring the biology of *Magnaporthe grisea*. Annual Rev. Microbiol. *57*: 177–202.
14. Choi, J., Kim, Y., Kim, S., Park, J., and Lee, Y.H. (2009) *MoCRZ1*, a gene encoding a calcineurin-responsive transcription factor, regulates fungal growth and pathogenicity of *Magnaporthe oryzae*. Fungal Genet Biol. *46*: 243–54.
15. Buck, M.J. and Lieb, J.D. (2004) ChIP-chip: considerations for the design, analysis, and application of genome-wide chromatin immunoprecipitation experiments. Genomics. *83*: 349–60.
16. Hanlon, S.E. and Lieb, J.D. (2004) Progress and challenges in profiling the dynamics of chromatin and transcription factor binding with DNA microarrays. Curr Opin Genet Dev. *14*: 697–705.
17. Lieb, J.D., Liu, X., Botstein, D., and Brown, P.O. (2001) Promoter-specific binding of Rap1 revealed by genome-wide maps of protein-DNA association. Nat Genet. *28*: 327–34.
18. Iyer, V.R., Horak, C.E., Scafe, C.S., Botstein, D., Snyder, M., and Brown, P.O. (2001) Genomic binding sites of the yeast cell-cycle transcription factors SBF and MBF. Nature. *409*: 533–8.
19. Mardis, E.R. (2007) ChIP-seq: welcome to the new frontier. Nat Methods. *4*: 613–4.
20. Borneman, A.R., Gianoulis, T.A., Zhang, Z.D., Yu, H., Rozowsky, J., Seringhaus, M.R., Wang, L.Y., Gerstein, M., and Snyder, M. (2007) Divergence of transcription factor binding sites across related yeast species. Science. *317*: 815–9.
21. Johnson, D.S., Li, W., Gordon, D.B., Bhattacharjee, A., Curry, B., Ghosh, J., Brizuela, L., Carroll, J.S., Brown, M., Flicek, P., Koch, C.M., Dunham, I., Bieda, M., Xu, X., Farnham, P.J., Kapranov, P., Nix, D.A., Gingeras, T.R., Zhang, X., Holster, H., Jiang, N., Green, R.D., Song, J.S., McCuine, S.A., Anton, E., Nguyen, L., Trinklein, N.D., Ye, Z., Ching, K., Hawkins, D., Ren, B., Scacheri, P.C., Rozowsky, J., Karpikov, A., Euskirchen, G., Weissman, S., Gerstein, M., Snyder, M., Yang, A., Moqtaderi, Z., Hirsch, H., Shulha, H.P., Fu, Y., Weng, Z., Struhl, K., Myers, R.M., Lieb, J.D., and Liu, X.S. (2008) Systematic evaluation of variability in ChIP-chip experiments using predefined DNA targets. Genome Res. *18*: 393–403.

Chapter 9

Proteome Studies of Filamentous Fungi

Scott E. Baker and Ellen A. Panisko

Abstract

The continued fast pace of fungal genome sequence generation has enabled proteomic analysis of a wide variety of organisms that span the breadth of the Kingdom Fungi. There is some phylogenetic bias to the current catalog of fungi with reasonable DNA sequence databases (genomic or EST) that could be analyzed at a global proteomic level. However, the rapid development of next generation sequencing platforms has lowered the cost of genome sequencing such that in the near future, having a genome sequence will no longer be a time or cost bottleneck for downstream proteomic (and transcriptomic) analyses. High throughput, nongel-based proteomics offers a snapshot of proteins present in a given sample at a single point in time. There are a number of variations on the general methods and technologies for identifying peptides in a given sample. We present a method that can serve as a "baseline" for proteomic studies of fungi.

Key words: Proteomics, Fungi, Pigments, Gene models

1. Introduction

Members of the fungal kingdom influence life in many ways. Fungi are key players in global carbon and nitrogen cycling, economically important plant and animal pathogens, and are the source of many pharmaceuticals, industrial enzymes, and commodity chemicals. As the number of fungi whose genome sequences have been elucidated continues to rapidly grow, so too does the potential to perform high throughput proteomic analysis on these organisms. While genome sequencing illuminates the genetic potential of an organism, proteomic analysis can indicate the relative amounts of proteins produced by an organism at a given point in time. Proteomic analysis is therefore an important method for deriving biological knowledge from genome sequences.

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722,
DOI 10.1007/978-1-61779-040-9_9,

There are technical issues that set proteomic analysis of fungi apart from other organisms. These issues can be associated with the fungal cell wall, highly abundant pigments, or secondary metabolites. A "typical" proteomics experiment is difficult to describe since there are different objectives designed by individual investigators. Here, we present a "baseline" method that can be modified as needed to suit particular fungi. An excellent overview of techniques is reviewed elsewhere (1). In addition, since proteomics is such an active area of research HPLC separations (2) and methods for quantitation of peptides (3) are performed under varied conditions between research groups.

2. Materials

2.1. Isolation of Cellular Proteins from Filamentous Fungi

1. Fungal biomass of interest. Any fresh fungal biomass free of substrates or media is suitable. The example given in this chapter is about vegetative hyphae harvested from liquid cultures.
2. Miracloth (EMD Chemicals, Darmstadt, Germany).
3. Lyophilizer (Fisher Scientific, Pittsburgh, PA).
4. Glass beads: 0.5 mm zirconia/silica glass beads (Biospec Products, Bartlesville, OK). Autoclave for 40 min. Store at room temperature.
5. Bead beating tubes with screw caps (Sarstedt Inc., Newton, NC).
6. Mini beadbeater (Biospec Products).

2.2. Isolation of Secreted Proteins from Filamentous Fungi

1. Spent media from fungal fermentation of interest.
2. Amicon Ultra-15 centrifugal filter units (Millipore, Billerica, MA).
3. Acetone.

2.3. Denaturation, Alkylation, and Tryptic Digestion of Proteins

1. Silonized Eppendorf tubes 1.7 ml (Fisher Scientific).
2. 8 M Guanidine hydrochloride (Sigma–Aldrich, St. Louis, MO).
3. Tris(2-carboxyethyl)phosphine (Bond-Breaker TCEP, Pierce Protein Research Products, Rockford, IL).
4. Iodoacetamide (Sigma–Aldrich, reagent is light sensitive and extremely toxic to fish – follow appropriate disposal procedures).
5. 500 and 50 mM Ammonium bicarbonate: dissolve ammonium bicarbonate (Sigma–Aldrich) in water. Make fresh.
6. Trypsin, mass spectrometry grade (Promega Corp., Madison, WI).
7. 50 mM Acetic acid: dilute glacial acetic acid (Sigma–Aldrich) in water.

2.4. Solid Phase Extraction of Peptides

1. C18 Solid phase extraction 100 mg, 1 ml tubes (Supelco, St. Louis, MO).
2. Vacuum manifold for solid phase extraction (Supelco).
3. 100% HPLC grade methanol (Sigma–Aldrich).
4. 80% HPLC grade acetonitrile (Sigma–Aldrich).
5. Silonized Eppendorf tubes 2.0 ml (Fisher Scientific).
6. Centrifugal concentrator (Thermo Scientific, Asheville, NC).

2.5. HPLC/MS Analysis

1. 1100 series HPLC system (Agilent, Santa Clara, CA).
2. 1100 series HPLC vials, caps, and glass inserts (Agilent, Santa Clara, CA).
3. LTQ mass spectrometer (Thermo Scientific, Asheville, NC).
4. 40 cm, 150 μm (ID) C18 column (Jupiter 5 μm resin, Phenomenex).
5. Fused silica emitter (4).
6. Formic acid ultrapure diluted to 0.1% (Fluka).
7. 90% HPLC grade acetonitrile, 0.1% formic acid ultrapure.

3. Methods

As mentioned above, sometime filamentous fungal samples present unique problems for proteome analysis. If possible, it can often be more expedient to prepare several samples and analyze them by LC/MS as a trial run in order to determine if any pigments present are problematic before adding additional sample manipulations to remove them. As a reminder for researchers who are unfamiliar with analytical HPLC/MS techniques, impurities introduced can obliterate any chance of quality data collection (see Note 1). Detergents in particular, once introduced into a sample are often unable to be removed. It is advised that when researchers are utilizing a new proteomics sample preparation method, they consult with the laboratory that is performing the HPLC/MS analysis to ensure compatibility. In terms of the amount of protein necessary for HPLC/MS analysis researchers should also consult with the laboratory that is performing the analysis, but typically for fungal experiments protein is not limiting. For example, it is fairly common to perform 1 μl injections for HPLC/MS on peptide concentrations in the 0.2 μg/μl (microliter) range. Depending on the HPLC system utilized 10 μl may be required in order to provide enough volume for the autosampler needle. So analyses of samples well under 100 μg of protein are possible.

3.1. Isolation of Soluble Cellular Proteins from Filamentous Fungi

1. Isolate fungal biomass by filtering from fresh cultures through sterile Miracloth. Rinse with sterile distilled water.
2. Remove excess water by pressing hyphal mat with paper towels.
3. Transfer fungal biomass to a 50 ml centrifuge tube and freeze the sample by placing the tube into liquid nitrogen.
4. Lyophilize fungal hyphae in a lyophilizer. Dried samples can be stored at −80°C until ready to complete processing (see Note 2).
5. Break or cut a piece of fungal biomass approximately the size of 15 mm sphere and place into a bead beating tube with a screw cap.
6. Add siliconized glass beads to the tube until it is approximately half full.
7. Replace cap and bead beat in a mini bead beater on highest power for 30 sec.
8. Cool samples on ice for 1 min.
9. Repeat steps 7 and 8 for five additional times.

The sample is ready for resuspension as outlined in step 1 of Subheading 3.3.

3.2. Isolation of Secreted Proteins from Filamentous Fungi

1. Isolate spent media by filtering out fungal biomass with two layers of Miracloth.
2. Reduce the sample volume by concentrating with an Amicon Ultra-15 centrifugal filter unit.
3. Transfer the concentrated filtrate to a centrifuge tube. Add 4 volumes of ice cold acetone and mix well.
4. Incubate at −20°C for 2 hr to overnight.
5. Centrifuge for 15 min at 15,000 × *g* and 4°C.
6. Decant the supernatant and reserve the pellet.

Samples are ready for resuspension as outlined in step 1 of Subheading 3.3.

3.3. Denaturation, Alkylation, and Tryptic Digestion of Proteins

1. Resuspend the protein as described in Subheading 3.1, or pellet as described in Subheading 3.2, in 1–2 ml 8 M guanidine hydrochloride (see Note 3). Vigorous shaking, sonication, and heating may be utilized to aid protein solubilization.
2. Remove any debris or unsolubilized proteins by centrifugation at 10,000 × *g* for 5 min.
3. Transfer the supernatant to a new silonized eppendorf tube (see Note 4).
4. Determine protein concentration with the BCA assay kit by using the microplate procedure (see Note 5).

5. Reduce disulfide bonds by adding TCEP to a final concentration of 2.5 mM.
6. Heat the sample for 30 min at 60°C.
7. Cool down to room temperature.
8. Calculate the amount of iodoacetamide to have a tenfold excess over cysteine residues. Assume every milligram of protein is equivalent to 30 nmol and each mole protein contains 6 mol of cysteine residues.
9. Dissolve iodoacetamide in 100–200 µl of 500 mM ammonium bicarbonate.
10. Add iodoacetamide/ammonium bicarbonate solution to the sample. Wrap in foil and incubate for 1 h on a rocker.
11. Dilute the sample to a final concentration of 0.9 M guanidine hydrochloride with 50 mM ammonium bicarbonate.
12. Calculate the amount of required trypsin needed for protein digestion. Typically 1 µg trypsin is used per 100 µg protein.
13. Dissolve trypsin in the buffer provided by the supplier or 50 mM acetic acid to a volume of 0.5 µg/µl (microliter).
14. Add appropriate volume of trypsin solution to the sample and digest 4 hr to overnight in a 37°C water bath.

3.4. Solid Phase Extraction of Peptides

1. Set up the vacuum manifold with a 15 ml disposable conical tube to capture flow through.
2. Place a 1 ml C18 tube onto the manifold (see Note 6).
3. Draw a total of 2 ml methanol through the tube to activate the resin. Leave a small volume of liquid on top of the frit. Once activated with methanol the C18 resin should not be allowed to dry until the final elution step.
4. Wash the C18 resin by drawing a total of 4 ml of water through, again reserving a small volume of liquid above the frit.
5. Add the sample to the tube and draw through. Reserve some liquid above the frit.
6. Wash the sample tube with 1 ml of 50 mM ammonium bicarbonate and add wash to C18 tube. Again leave a small volume of liquid above the frit.
7. Rinse the C18 resin by drawing 8 ml of 50 mM ammonium bicarbonate through the vacuum apparatus.
8. Replace the 15 ml conical tube in vacuum apparatus with a new tube and place a 2 ml silonized eppendorf tube on top of the 15 ml tube to collect the peptide eluate.
9. Elute peptides with 2 volumes of 900 µl of 80% acetonitrile.
10. Dry the sample in a centrifugal concentrator and store at –20°C until ready for HPLC/MS analysis.

3.5. HPLC/MS Analysis

1. Resuspend peptide samples in 0.1% formic acid at a concentration between 0.1 and 1 µg/µl.
2. Centrifuge for 5 min at top speed in a microfuge to remove any insolubles.
3. Transfer peptide sample to HPLC vial with glass insert, cap and place into HPLC autosampler.
4. Equilibrate HPLC column in solvent A (0.1% formic acid solution) at 2 µl/min for 1 h.
5. Inject 1 µl of peptide sample onto column and elute into mass spectrometer by electrospray ionization utilizing the fused silica emitter. The following linear gradients can be used: 15 min solvent A from 100% solvent A to 20% solvent B (80% acetonitrile, 0.1% formic acid) over 5 min, from 20% solvent B to 50% solvent B over 55 min, from 50% solvent B to 95% solvent B over 5 min and hold for 5 min, from 5% solvent A to 100% solvent A over 5 min and hold for 40 min, at a flow rate of 2 µl/min.
6. Spectra are collected by the mass spectrometer using the data dependent mode with the five most intense ions from the survey scan being selected for dissociation.
7. Peptide identifications are performed by Sequest (5) analysis of the raw data against an appropriate protein database. Data for subsequent analysis is filtered so that only total tryptic peptides with accepted scoring criteria remain (6).
8. Peptide abundances are calculated by the MASIC program (7) and protein abundances are performed by mixed-effects statistical modeling (8).

4. Notes

1. All solutions used in the protocol should be made utilizing 18 M ohm water. Plastic disposable pipets are not recommended for use with HPLC grade organic solvents as contaminants leech from the plastic into the solvent over time. Often, laboratories separate solvents used for proteomics protocols from the general laboratory supply and utilize glass pipets exclusively.
2. For comparative proteomics where multiple experimental replicates are performed storing lyophilized biomass at –80°C until all samples are ready to be digested is recommended. Processing samples as to not introduce confounding variables is discussed in detail elsewhere (9).
3. Here, the volume that the protein is resuspended in is not critical. This only affects the size of tubes you need for further

processing as well as the time it takes for the solid phase extraction procedure. For those samples that contain zirconia/silica beads, the resuspended sample can be transferred to a silonized eppendorf tube with a glass Pasteur pipet before spinning to remove cell debris.

4. For cell extracts that contain pigments that interfere with HPLC/MS analysis acetone precipitation would be performed as outlined in Subheading 3.2, steps 3–7.
5. The BCA is recommended for its high tolerance to salts (guanidine hydrochloride in particular). However, some pigments produced by certain fungi interfere with protein determination assays. In several cases, we have utilized the Bio-Rad stain free gel system in concert with densitometry analysis to determine protein concentration. After protein determination, sample preparation replicates could be performed by splitting the sample equivalently between several eppendorf tubes and completing the remainder of the procedure in parallel.
6. The 1 ml C18 tubes contain 100 mg of resin. The capacity of these tubes is 5 mg of digested proteins. Larger tubes (and subsequently increased washing volumes) should be utilized for greater amounts of protein.

References

1. Nesvizhskii, A.I., VItek, O., and Aebersold, R. (2007) Analysis and validation of proteomic datasets generated by tandem mass spectrometry. *Nature Methods*, **4**, 787–797.
2. Sandra, K., Moshir, M., D'hondt, F., Verleysen, K., Kas, K., and Sandra, P. (2008) Highly efficient peptide separations in proteomics Part 1. Unidimensional high performance liquid chromatography. *Journal of Chromatography B, Analytical Technologies in the Biomedical and Life Sciences* **866**, 48–63.
3. Bantscheff, M., Schirle, M., Sweetman, G., Rick, J., and Kuster, B. (2007) Quantitative mass spectrometry in proteomics: a critical review, *Analytical and Bioanalytical Chemistry* **389**, 1017–1031.
4. Kelly, R.T, Page, J.S., Luo, Q., Moore, R.J., Orton, D.J., Tang, K., and Smith, R.D. (2006). Chemically Etched Open Tubular and Monolithic Emitters for Nanoelectrospray Ionization Mass Spectrometry. *Analytical Chemistry* **78**, 7796–7801.
5. Eng, J.K., McCormack, A.L., and Yates III, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *Journal of the American Society for Mass Spectrometry*, **5**, 976–989.
6. Wolters, D. A., Washburn, M. P., and Yates III, J. R. (2001) An automated multidimensional protein identification technology for shotgun proteomics. *Analytical Chemistry*, **73**, 563–569.
7. Monroe, M.E., Daly, D.S., Adkins, J.N., and Smith, R.D. (2008). MASIC: a software program for fast quantitation and flexible visualization of chromatographic profiles form detected LC-MS(/MS) features. *Computational Biology and Chemistry* **32**, 215–217.
8. Daly, D.S., Anderson, K.K., Panisko, E.P., Purvine, S.P., Fang, R., Monroe, M.E., and Baker, S.E. (2008) Mixed-effects statistical model for comparative LC-MS proteomics studies, *Journal of Proteome Research*, **7**, 1209–1217.
9. Oberg, A.L., and Vitek, O. (2009) Statistical design of quantitative mass spectrometry-based proteomic experiments. *Journal of Proteome Research* **8**, 2144–2156.

Chapter 10

A Bioinformatics Pipeline for Sequence-Based Analyses of Fungal Biodiversity

D. Lee Taylor and Shawn Houston

Abstract

The internal transcribed spacer (ITS) is the locus of choice with which to characterize fungal diversity in environmental samples. However, methods to analyze ITS datasets have lagged behind the capacity to generate large amounts of sequence information. Here, we describe our bioinformatics pipeline to process large fungal ITS sequence datasets, from raw chromatograms to a spreadsheet of operational taxonomic unit (OTU) abundances across samples. Steps include assembling of reads originating from one clone, identifying primer "barcodes" or "tags," trimming vectors and primers, marking low-quality base calls and removing low-quality sequences, orienting sequences, extracting the ITS region from longer amplicons, and grouping sequences into OTUs. We expect that the principles and tools presented here are relevant to datasets arising from ever-evolving new technologies.

Key words: Fungi, Next-generation sequencing, Biodiversity, ITS, Ribosomal, Automated, Pipeline

1. Introduction

Studies of microbial diversity in complex communities present in natural environments have exploded in the last 20 years, in large part due to the advent of the polymerase chain reaction. Analyses of prokaryotic diversity have led the way, starting with the pioneering studies of Woese et al. (1–3) while parallel studies of microbial eukaryotes, including fungi, have blossomed more recently (4–6). Methods for prokaryotic microbes have matured to the point that streamlined, Web-accessible software pipelines are available (7, 8), although these tools continue to advance rapidly (9). Unfortunately, many methods developed for prokaryotes are not appropriate for fungi, due largely to the fact that the ribosomal small subunit (SSU), the linchpin of prokaryotic diversity

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722,
DOI 10.1007/978-1-61779-040-9_10,

analyses, is insufficiently variable to precisely discriminate fungal taxa. Instead, the highly variable internal transcribed spacer (ITS) region of the nuclear ribosomal operon is more appropriate for fungal diversity analyses (10, 11). The ITS region is well suited for species discrimination, but is too variable to be aligned across fungal genera. As a result, many methods to analyze prokaryotic diversity that depend upon global alignments of SSU sequences cannot be applied to fungi. Although efforts to develop standardized protocols and tools to analyze fungal diversity are underway, at present, most research groups either develop their own sets of tools or are confounded by the lack of tools.

Sequence-based studies of microbial biodiversity face two diametrically opposed technical challenges. The first challenge is posed by the overestimation of diversity that arises from artifactual DNA sequences. Sequence artifacts arise through a number of mechanisms, the most common being mistaken base calls derived from low-quality sequence reads (12) and sequences derived from two different organisms as a result of chimera formation during PCR (13, 14). Taq error and pseudogenes can also contribute to false diversity (15). Thus, strict quality controls and careful analyses are critical for accurate assessments of diversity. The tremendous diversity of microbes poses the second challenge: massive sequencing efforts are required to enumerate a significant fraction of total diversity (16). Depth of sequencing is perhaps most pressing when analyzing prokaryotic diversity, but also applies to fungi and other microbial eukaryotes (4, 6, 17). Adequately large datasets are beyond the reach of manual quality control and analysis. In short, the opposing imperatives of large sequence datasets and strict quality control evoke the need for carefully conceived and implemented bioinformatics pipelines.

Here, we describe steps developed by our lab to process large fungal ITS sequence datasets. We emphasize quality control and the underlying biological principles. The steps are not linked in an end-to-end pipeline, so user intervention is required at multiple steps. However, our approach relies on readily available programs and Web sites, most of which are freeware; unique scripts have been made publicly available through a Web portal. This text assumes no background in bioinformatics; our pipeline obviates command line programs and scripting entirely. Our target audience is those new to sequence-based analysis of fungal diversity rather than the experienced practitioner.

Examples are given from our experiences with high-throughput Sanger sequencing of clone libraries generated from PCR of soil DNA extracts. Although Sanger sequencing is a somewhat dated technology given the current explosion of next-generation sequencing (18), we hope that the principles and tools discussed here will be useful regardless of sequencing methodology. Our bench methods have been presented elsewhere (19–22).

In brief, we amplify a fungal nuclear ribosomal region spanning ITS1, the 5.8S, ITS2, and roughly 700 bp of the large subunit (LSU) using the primers ITS1-FL and tagged versions of TW13 (20, 23). By obtaining LSU data, we are able to infer the broad relationships of novel fungi that lack closely related, known fungi in public databases, which is difficult using the ITS alone.

2. Methods

2.1. Chromatogram Processing

Quality scores ("Q scores") for each base call are a critical element of any automated sequence cleanup and analysis pipeline. In the past, sequence reads from the same template, clone, or amplicon were assembled, and the chromatograms inspected manually to insure sequence quality, to correct base calls, and to remove regions of low confidence. This approach is clearly untenable with large datasets, and automated approaches that utilize aspects of chromatogram peak height and shape to determine base confidence values were elaborated in the early 1990s to support the growing needs of genome sequencing efforts (24, 25). Researchers carrying out phylogenetic and biodiversity analyses have adopted these automated quality control approaches slowly, but today's massive datasets are forcing the transition. Due to the statistical nature of the derived base call confidence values, there is less need for multiple coverage of a given sequence region (e.g., the traditional bidirectional and completely overlapping reads often used for phylogenetic studies). For example, if the probability that a base call is incorrect is below 1/100th of 1% (phred score of 40), there is little need to cover the same base from an additional read. On the other hand, our lab routinely generates two reads per clone in order to obtain adequate coverage of the 1,200–1,500 bp region. In this case, an assembly step is still required in order to join these overlapping reads.

We use the program CodonCode Aligner (http://www.codoncode.com/aligner/) to carry out initial sequence assembly and processing. Aligner provides a Mac OSX graphical user interface (GUI) for the popular freeware package phred/phrap/consed that is written in C (http://www.phrap.org/phredphrapconsed.html). Phred is a basecaller that includes the key feature of determining a confidence score ("phred score") for each base call (24, 25). A phred score of ten implies that there is a one in ten chance that the base call is incorrect. A phred score of 20 implies a one in one hundred chance of an incorrect base call, and so on. Analogous scores can now be generated with pyrosequencing platforms, despite the rather different patterns of basecall errors in comparison to Sanger sequencing (26). Phrap carries out assembly of reads into "contigs" utilizing phred scores to derive the

best supported consensus sequence from overlapping reads (27). Usually, assembly refers to the joining of numerous reads from random shotgun sequencing of genomes into consensus sequences spanning contiguous regions of the genome. However, in our procedure, we simply wish to assemble the reads derived from a single clone. Usually, these are two paired-end reads obtained using the vector primers flanking the inserted PCR product. With small numbers of clones, the reads to be assembled together can be selected manually in Aligner. However, this is not practical for large datasets. Fortunately, phrap and Aligner offer a variety of options for automatically selecting sets of reads to be joined based on shared elements within the read names. Note that pyrosequencing circumvents the cloning step and creates reads in only a single direction from a particular template (paired-end read technologies are being developed, but are not currently available for the sequencing of 400–800 bp PCR products), meaning that assembly is not required in the analysis of pyrosequence datasets. This is also a good point at which to eliminate short reads, most of which likely represent either (1) poor quality sequences, or (2) sequences of undesired clone inserts, particularly primer-dimers. An appropriate cutoff for sequence length after assembly depends on the locus. We typically use 200 bp as the cutoff after trimming the ends.

1. It is often necessary to rename your sequence reads before importing the chromatograms to a base caller/assembler in order to allow automated joining of reads arising from the same clone (see Note 1).
2. CodonCode aligner is able to instruct phrap to assemble reads that have a shared component of the name that is set off by a specified delimiter (often dot "." or underscore "_"). This component that unites your target reads must not be found in any other reads within the Codoncode Project; consult the Codoncode Help for more detail.
3. Once your reads are named properly, import all of the reads you wish to assemble into Aligner using the File → Import → Add Folder command.
4. Save the project with an appropriate file name and remember to save after every step along the way, especially before and after sequence assembly.
5. Trim the dirty ends of the reads using Sample → Clip Ends. The default clipping settings are usually adequate, as we will further trim ends and deal with low-quality base calls in subsequent steps.
6. Remove sequences with low numbers of high-quality base pairs. Sort sequences by clicking on the "quality" column heading, then visually inspect chromatograms for sequences with low-quality scores. Choose your threshold (<300 quality

bases is often used), select all the sequences you wish to remove, then navigate to Edit → Move to Trash.

7. Remove vector sequences using Sample → Trim Vector. Note that you must specify which cloning vector you are using in the Vector Trimming settings under CodonCode Aligner → Preferences. As with end trimming, there are later opportunities to remove vector regions that were missed by Aligner.
8. Save your project, and then assemble sets of reads derived from a single clone using the Contig → Assemble with Options command. Click the radio button next to "assemble in groups." Note that you need to click the button "Define Groups" to instruct Aligner how to recognize reads that should be joined. Aligner parses components of the read names between delimiters from left to right. Click the "Preview" button to confirm that your definitions accurately specify the sets of reads that should be assembled.
9. It is likely that some reads will not assemble into contigs due to poor sequence quality, insufficient read overlap, or other problems. We generally keep these reads because the phred scores allow us to screen out poor quality sequence and keep high-quality sequence, even for regions lacking bidirectional coverage.
10. Save your project again, and then export both your sequences and phred scores. Select all the contigs created at the assembly step, then use File → Export → Consensus Sequences as Single File command, making sure to check the box for "export quality scores" (see Note 2).
11. You need to separately export any unassembled reads by selecting them and using File → Export → Sequences.
12. Assuming that you had some unassembled reads, you should now have four text files: (1) a .contig file of the contig sequences, (2) a matching .qual file of the contig quality scores, (3) a file of the unassembled sequences, and (4) a file of the unassembled sequence quality scores.
13. For simplicity in downstream steps, combine the contig and unassembled .fasta files into one file. Also combine the corresponding .qual files.
14. From this point forward, a series of text files is used instead of the chromatogram data. It is important to keep these files in a simple "txt" format – do not open them in Word and save them in a default Word format, as the various bioinformatics programs are not able to open and read Word documents. On a MacIntosh platform, we use TextEdit and TextWrangler to work with these text files. WordPad can be used on the PC platform.

2.2. Masking

The next goal is to permanently mark low-quality base calls so that questionable bases are not evaluated as reliable sequence data. At this point, the phred base call quality scores are essential. We use an "in-house" Perl script to change consensus bases with phred scores below a certain threshold, usually 20, to Ns or to lowercase letters.

1. Direct your browser to our Fungal Metagenomics Web portal at http://www.borealfungi.uaf.edu/. Under Search ITS, click on the "Mask" link.
2. Submit your sequence file and corresponding file of phred scores using the upload links on the Web page. Set your desired phred score threshold and decide whether to convert low-quality bases to Ns or to lowercase letters (see Note 3).
3. Download the resulting "masked" files from the portal. The process is repeated for each pair of text files containing sequences and corresponding phred scores (see Note 3).

2.3. Finding Tags

For automated, high-throughput sequencing of clone libraries, we pool many samples into single-clone libraries. To attribute clone sequences to source samples after sequencing, we add 10-bp "tags" to the PCR primers that uniquely identify each sample in a pooled clone library (20). To identify these tags after sequencing, we utilize an approach that allows some sequence error within the tagged primer region, yet accurately assigns sequences to the correct source sample.

1. Direct your browser to our Fungal Metagenomics Web portal at http://www.borealfungi.uaf.edu/. Under "Search ITS," click on the "Tag Finder" link.
2. Submit each of the masked files together with a text file listing all of the primer sequence tags to the appropriate upload boxes. Additional instructions are provided on the Web site.
3. Open the output files in a text editor, and rename reads to include the information from the tags. This can be accomplished easily in Textedit by searching for ">" with the Edit → Find command, then placing "*>sample_name_or_tag_*" in the "Replace with" box and clicking "Replace All." The greater than sign occurs at the beginning of each sequence name when the sequences are in fasta format, and provides a convenient key for adding or changing components of sequence names. Similar commands are available in TextWrangler and Wordpad.

2.4. Sequence Orientation

The Invitrogen TOPO TA pcr4 cloning vector is not directional, meaning that the PCR product can be cloned in either of the two possible orientations. To obtain high-quality sequence reads through the primers, which contain the critical tag region, we

sequence with vector primers. A given vector primer may read into either the ITS or the LSU end of a particular amplicon, meaning that we must analyze the reverse complement of any contigs that are in the wrong orientation with respect to the ribosomal operon. We accomplish this by comparing our sequences to a series of short motifs corresponding to conserved regions of the SSU, 5.8S and LSU. If a strong match is found to a particular motif, the input sequence is returned unaltered. If a match to the reverse complement of the motif is found, the input sequence is returned as the reverse complement. If no strong match is found, the input sequence is returned in a separate file. The motifs are used consecutively until nearly all sequences are in the proper orientation. In our experience, the few sequences for which no strong matches to our motifs are found are usually nonfungal, and should be discarded from the dataset.

1. Submit fasta files of sequences to the Orient page of our Fungal Metagenomics Web site, http://www.borealfungi.uaf.edu/.
2. Download the resulting "oriented.contigs" and "orientation_unknown" files.
3. Analyze any remaining unoriented sequences for the presence of nonfungal sequences and check their orientation by performing a standard nucleotide BLAST search on the NCBI Web site (http://blast.ncbi.nlm.nih.gov/Blast.cgi, click nucleotide blast, then check "others" under Database).

2.5. Splitting ITS from LSU

Our amplicons span the entire ITS plus about 700 bp of the LSU. However, species-level discrimination of fungi using percent identity thresholds is currently based on the highly variable ITS regions, rather than the LSU. Hence, we must split ITS from LSU prior to carrying out clustering of our clones into operational taxonomic units (OTUs).

1. Align your cleaned sequences using a fast multiple sequence alignment program, such as Clustal (28) or Muscle (29); there are numerous Web servers for these programs, e.g., http://www.ebi.ac.uk/Tools/clustalw2/index.html.
2. For Clustalw, we use the following nondefault settings: -gapext = 1; transweight = 0.2; -pwgapext = 1; apply –kimura.
3. Convert the alignment from clustal format to fasta format using the public Web tool Format Converter at http://hcv.lanl.gov/content/sequence/FORMAT_CONVERSION/form.html (see Note 4).
4. Open the alignment in an editor. For Mac OS, we recommend SeAl (http://tree.bio.ed.ac.uk/software/seal/) and for Windows, we recommend BioEdit (30) http://www.mbio.ncsu.edu/BioEdit/bioedit.html).

5. In SeAl, use Alignment → Find to locate the conserved motif of the ITS4 primer site (reverse complement): GCATATCAATAAGCGGAGG. Under the menu for Treat Gaps, select "as if removed from sequence." If using Bioedit, the analogous menus are Edit → Search → Find.
6. When properly aligned, large blocks of downstream sequence can be selected and deleted (see Note 5).
7. The 5′ end of the sequences should be well aligned in the forward primer ITS1-F region: CTTGGTCATTTAGAGG AAGTAA (see Note 6).
8. This is also a convenient opportunity to remove the forward primer sequence, in our case ITS1-F. Since primers do not necessarily agree with the template sequence of the organism, base calls from primer regions should not be included in sequence submissions to GenBank. This step also ensures that any remaining vector sequence is removed.
9. Select blocks of sequence from the 5′ end to remove everything up to the 3′ end of ITS1-F (or the primer used in your studies).

2.6. Additional Sequence Cleanup

At this point, there are likely to be a number of relatively low-quality sequences remaining in the dataset. This is a good point at which to identify and remove these sequences.

1. If you used lowercase, rather than N, to replace low-quality bases at the masking step (2.2), change the lowercase bases to Ns. Open the file with Textedit, open the Find dialog and unselect the Ignore Case box. Search consecutively for "a," "c," "t," and "g," replacing with "N" in each case using the Replace All command. Beware of any a, c, t, or gs in your read names, as these are also replaced.
2. With Sanger sequencing, most of the poorer base calls are located at the beginning and ends of the sequence reads. It is helpful to trim the sequences from the beginning and end to remove these low-quality regions. We use the program TrimSeq, which is available as a Web tool at many locations, including http://imed.med.ucm.es/cgi-bin/emboss.pl?_action=input&_app=trimseq. Set the window size to 40 and the percent ambiguity to 5% (or lower), and select "yes" for "trim at the start?" and "trim at the end?".
3. Now that the ends are trimmed, sequences that still have a number of Ns should be deleted from the dataset. We use BioEdit to accomplish this. Open the fasta file of sequences in BioEdit, then use Edit → Search → Find/Replace → Replace With command to replace Ns with gaps (dash "-"). Then use Sequences → Filter Out Sequences Containing → Greater than X% Gaps to delete sequences with greater than 2% gaps.

A more stringent cutoff is acceptable, less stringent cutoffs are not recommended, since a 3% difference would cause two sequences to be placed in different OTUs (see Subheading 2.7). Finally, the gaps must be converted back to Ns using the Edit → Search → Find/Replace command in reverse.

2.7. Grouping ITS Sequences into OTUs

In a large sequence dataset, there are many identical and nearly identical sequences that likely represent the same fungal species. With large datasets, the most practical approach is to group or cluster sequences based upon a percent identity threshold (i.e., the percentage of bases that are identical throughout the entire overlapping region in a pairwise sequence alignment) with a sequence assembly program. For smaller datasets, this can be accomplished directly in sequence editors such as Sequencher, CodonCode Aligner, or Bioedit. However, for large datasets and to achieve greater control over the grouping behavior, we use the program TGICL (31) which carries out a first-pass grouping using BLAST and a second, finer grouping via the genome assembler Cap3 (32). A wide range of percent identity values for the ITS region have been used in various fungal studies, ranging from 90 to 99%. A balance must be struck between lumping discrete species when using a low percent threshold and splitting a single species due to base call errors, polymerase error, and intraspecific variation when applying a stringent threshold. A consensus of 95–98% seems to be emerging in the recent literature (6, 17, 33).

The next task is to reformat the output from Cap3 in such a way that it is useful for subsequent analyses of diversity. Cap3 creates a number of files for each run. The '.singletons' and '.contigs' files contain DNA sequences that are rarely used, since the contig sequences are a consensus of all reads that were grouped in that OTU, and thus represent an artificial sequence, not one that necessarily occurs in nature. The key file is the Cap3.out file, which includes both a list of reads that were grouped into specific contigs and the multiple alignments comprising those contigs. It is worth visually checking several of the multiple alignments to ensure satisfaction with the way Cap3 grouped the input sequences. The list of reads by contig is the most important data for downstream ecological analyses, but must be 'parsed' in order to clean up the file format. We use a freeware program called TextWrangler to accomplish this initial parsing.

1. Submit fasta file of ITS sequences to the OTU_Grouping page of our Fungal Metagenomics Web site, http://www.borealfungi.uaf.edu/.
2. Open the resulting Cap.out files in TextWrangler.
3. Copy the top part of the file containing the list of contigs (not the multiple sequence alignments) into a new text file.

4. In the new text file, replace all instances of "**** Contig" by clicking the "grep" check box and inserting the following text in the Find box under the Search menu: *+\s*Contig*\s, then typing \t*SampleName* in the Replace With box (where SampleName is something of your choosing to indicate the source of the sequence), then selecting "replace all" (see Note 7).
5. Remove unneeded spaces by unclicking the "grep" box and simply placing a space in the Find box and deleting all contents from the Replace With box, then selecting "replace all."
6. Remove unneeded asterisks the same way spaces were removed, but using * in the Find box.
7. Remove unneeded read name after "is in" using by clicking "grep" and inserting the following text in the Find box: *sis*\ *sin*\s[0-9A-Za-z_+-?]+\r, then inserting \r in the Replace With box and clicking "replace all."
8. Remove the unneeded "+" and "–" symbols by checking the grep box and inserting the text [+-] in the Find box, emptying any contents from the Replace With box, and selecting "replace all."
9. Save this 'parsed' text file.

Now that you have simplified the Cap3 output, you are ready to import the data into Excel (or a spreadsheet program of your choice).

1. In Excel, select File > Open, and select your parsed Cap3 text file.
2. When the dialog box comes up asking you how to interpret the data, select "Delimited" then click "Next."
3. In the choice of delimiters given in the dialog box, select Tab, and then click OK.
4. At this point, Excel should present you with a spreadsheet in which your read names (clones, sequences) are in the first column, and the OTU groupings or contigs are in the second column. The only remaining task is to fill out the contig names for every row, since they are now only at the top of each sequence group.
5. To fill out the contig information, select a contig name and pull down the column to select up to the the next contig name, then select "fill down" under the Edit menu; repeat for all contigs.
6. We routinely incorporate information about the source site or sample into each read name. Then, in Excel, this information can easily be propagated to a new column using the "text to columns" command under the Data menu. This allows one to sort sequences by sample or OTU.

7. Lastly, the resulting long list of sequences, OTU grouping and site or sample can be summarized by selecting the relevant columns then navigating to Data → Pivot Table Report. Click Next twice, then click Layout, then drag OTU to the "rows" area, site or sample to the "columns" area and OTU to the central "data" area in the Format dialog box to create a useful summary (see Note 8).

3. Discussion

The steps outlined above describe our analysis pipeline from sequence chromatograms to a spreadsheet of OTU abundances across samples with high confidence in the quality of the sequences and groupings. There are many possible directions for downstream ecological analyses, such as diversity analyses with programs like EstimateS (34), community ordination with programs, including PC Ord (35) or Vegan (36), or phylogenetic community analysis with Unifrac (37) or Phylocom (38).

Several important steps are not covered here. One is the identification of the fungal taxa represented by the OTUs. The traditional approach to this is a simple BLAST search of GenBank. However, the large numbers of unidentified and misidentified sequences in the international databases limit the utility of these searches. Pipelines and curated databases that help overcome these issues have been described elsewhere (39, 40). We also provide tools for fungal identification at our Fungal Metagenomics Web site.

Another issue not addressed in this chapter is the detection of chimeric sequences within the dataset. This issue is not trivial, as estimated proportions of chimeric sequences in fungal ITS clone libraries have been as high as 30% (41). In our datasets, we commonly find roughly 3% chimeras. As each chimera is typically a "singleton" with regard to OTU grouping in Cap3, these chimeras obviously lead to a considerable inflation of estimated diversity. Unfortunately, methods developed for prokaryotic 16S sequences are not appropriate for fungal ITS datasets because the prokaryotic methods depend on broad multiple sequence alignments (14, 42). We have developed a multistep BLAST approach to identify chimeras that is not amenable to automation. Due to the complexity of the method, it is beyond the scope of this chapter. We hope that automated detection methods appropriate for fungal ITS datasets will become available in the near future.

With regard to areas for future development, it is clear that there are shortcomings of the "one size fits all" percent sequence identity approach (21, 22). Different clades of fungi have ITS sequences that evolve at different rates. Thus, any arbitrary identity threshold may lump discrete species in some clades while

splitting a single species in other clades. Eventually, we expect that phylogenetic approaches in which well-supported clades are used to distinguish OTUs will supplant current percent identity approaches. However, at present, automated pipelines for fungal phylogenetic analysis of very large datasets are not publicly available.

4. Notes

1. Renaming of your reads can be accomplished using BioEdit (http://www.mbio.ncsu.edu/BioEdit/bioedit.html) on a PC or the "Replace text in names" script in the Finder Scripts folder for the application Applescript on a Mac. If you have control over the read naming convention, you should not need to rename reads – just use an appropriate convention at the beginning.
2. While the assignment of base call phred scores to bases within a single read follows a strict algorithm developed by Phil Green, there are several ways that these scores can be combined to approximate a phred score for a consensus base call derived from assembled reads. These options can be set in Aligner Preferences. We use the subtract scores for conflicting bases option.
3. If you have incorporated "tags" or "barcodes" in your primers, as described in the next section, it will be helpful to convert low-quality base calls to lowercase rather than N so that the tags can be identified in a higher proportion of your sequences.
4. If this Web tool is unavailable, most versions of the freeware Readseq should be able to perform the desired conversion.
5. It may be convenient at this point to paste this LSU components into a new alignment so that they can be used for subsequent phylogenetic analyses.
6. There are a number of Ascomycetes that have a large and highly variable intron just downstream of the ITS1-F primer (43); if such taxa are present in your dataset, they may cause the ITS1-F region to be located erratically in the alignment. As long as the ITS1-F primer region can be located and removed, these introns should not cause major problems. Because they can be present or absent among isolates of the same putative species, it is reasonable to remove them prior to OTU grouping.
7. If you would like to better understand and perhaps modify these commands, look up Grep in the online TextWrangler Help menu.

8. These steps are for Excel 2000–2004; the menus and commands are organized differently in Excel 2009, but pivot table reports are still available.

Acknowledgments

We thank James Long for writing several of the original pipeline scripts and Dan Cardin for writing the tag-finder script. Niall Lennon and Chad Nusbaum of the Broad Institute, MA, spearheaded high-throughput Sanger sequencing of our fungal clone libraries. Lab members Michael Booth, Robert Burgess, Ian Herriott, Jack McFarland, and Ina Timling have assisted with testing and improving our pipeline and also provided valuable comments on earlier drafts of the chapter. This work was supported in part by the National Science Foundation under grant numbers EF-0333308 and ARC-0632332. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the National Science Foundation. This publication was also made possible by grant number 2P20RR016466 from the National Center for Research Resources (NCRR), a component of the National Institutes of Health (NIH).

References

1. Fox, G. E., Stackebrandt, E., Hespell, R. B., Gibson, J., Maniloff, J., Dyer, T. A., Wolfe, R. S., Balch, W. E., Tanner, R. S., Magrum, L. J., Zablen, L. B., Blakemore, R., Gupta, R., Bonen, L., Lewis, B. J., Stahl, D. A., Luehrsen, K. R., Chen, K. N., and Woese, C. R. (1980) The phylogeny of prokaryotes, *Science* ***209***, 457–463.
2. Pace, N. R., Stahl, D. A., Lane, D. J., and Olsen G. J. (1985) Analyzing natural microbial populations by rRNA sequences, *ASM American Society for Microbiology News* ***51***, 4–12.
3. Giovannoni, S. J., Britschgi, T. B., Moyer, C. L., and Field, K. G. (1990) Genetic diversity in Sargasso Sea bacterioplankton, *Nature* ***345***, 60–63.
4. Vandenkoornhuyse, P., Baldauf, S. L., Leyval, C., Straczek, J., and Young, J. P. W. (2002) Evolution – extensive fungal diversity in plant roots, *Science* ***295***, 2051–2051.
5. Schadt, C. W., Martin, A. P., Lipson, D. A., and Schmidt, S. K. (2003) Seasonal dynamics of previously unknown fungal lineages in tundra soils, *Science* ***301***, 1359–1361.
6. O´Brien, H. E., Parrent, J. L., Jackson, J. A., Moncalvo, J. M., and Vilgalys, R. (2005) Fungal community analysis by large-scale sequencing of environmental samples, *Appl Environ Microb* ***71***, 5544–5550.
7. Maidak, B. L., Cole, J. R., Lilburn, T. G., Parker, C. T., Saxman, P. R., Farris, R. J., Garrity, G. M., Olsen, G. J., Schmidt, T. M., and Tiedje, J. M. (2001) The RDP-II (Ribosomal Database Project), *Nucleic Acids Res* ***29***, 173–174.
8. DeSantis, T. Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E. L., Keller, K., Huber, T., Dalevi, D., Hu, P., and Andersen, G. L. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB, *Appl Environ Microb* ***72***, 5069–5072.
9. Schloss, P. D., Westcott, S. L., Ryabin, T., Hall, J. R., Hartmann, M., Hollister, E. B., Lesniewski, R. A., Oakley, B. B., Parks, D. H., Robinson, C. J., Sahl, J. W., Stres, B., Thallinger, G. G., Van Horn, D. J., and Weber, C. F. (2009) Introducing mothur: open-source, platform-independent, community-supported

software for describing and comparing microbial communities, *Appl Environ Microb* **75**, 7537–7541.

10. Gardes, M., and Bruns, T. D. (1993) ITS primers with enhanced specificity for basidiomycetes – application to the identification of mycorrhizae and rusts, *Mol Ecol* **2**, 113–118.
11. Seifert, K. A. (2009) Progress towards DNA barcoding of fungi, *Mol Ecol Resour* **9**, 83–89.
12. Kunin, V., Engelbrektson, A., Ochman, H., and Hugenholtz, P. (2010) Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates, *Environ Microbiol* **12**, 118–23.
13. Meyerhans, A., Vartanian, J. P., and Wainhobson, S. (1990) DNA recombination during Pcr, *Nucleic Acids Res* **18**, 1687–1691.
14. Ashelford, K. E., Chuzhanova, N. A., Fry, J. C., Jones, A. J., and Weightman, A. J. (2006) New screening software shows that most recent large 16S rRNA gene clone libraries contain chimeras, *Appl Environ Microb* **72**, 5734–5741.
15. Valentini, A., Miquel, C., Nawaz, M. A., Bellemain, E., Coissac, E., Pompanon, F., Gielly, L., Cruaud, C., Nascetti, G., Wincker, P., Swenson, J. E., and Taberlet, P. (2009) New perspectives in diet analysis based on DNA barcoding and parallel pyrosequencing: the trnL approach, *Mol Ecol Resour* **9**, 51–60.
16. Sogin, M. L., Morrison, H. G., Huber, J. A., Mark Welch, D., Huse, S. M., Neal, P. R., Arrieta, J. M., and Herndl, G. J. (2006) Microbial diversity in the deep sea and the underexplored "rare biosphere", *P Natl Acad Sci USA* **103**, 12115–12120.
17. Buee, M., Reich, M., Murat, C., Morin, E., Nilsson, R. H., Uroz, S., and Martin, F. (2009) 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity, *New Phytol* **184**, 449–456.
18. Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., Berka, J., Braverman, M. S., Chen, Y. J., Chen, Z. T., Dewell, S. B., Du, L., Fierro, J. M., Gomes, X. V., Godwin, B. C., He, W., Helgesen, S., Ho, C. H., Irzyk, G. P., Jando, S. C., Alenquer, M. L. I., Jarvie, T. P., Jirage, K. B., Kim, J. B., Knight, J. R., Lanza, J. R., Leamon, J. H., Lefkowitz, S. M., Lei, M., Li, J., Lohman, K. L., Lu, H., Makhijani, V. B., McDade, K. E., McKenna, M. P., Myers, E. W., Nickerson, E., Nobile, J. R., Plant, R., Puc, B. P., Ronan, M. T., Roth, G. T., Sarkis, G. J., Simons, J. F., Simpson, J. W., Srinivasan, M., Tartaro, K. R., Tomasz, A., Vogt, K. A., Volkmer, G. A., Wang, S. H., Wang, Y., Weiner, M. P., Yu, P. G., Begley, R. F., and Rothberg, J. M. (2005) Genome sequencing in microfabricated high-density picolitre reactors, *Nature* **437**, 376–380.
19. Taylor, D. L., Herriott, I. C., Long, J., and O'Neill, K. (2007) TOPO TA is A-OK: a test of phylogenetic bias in fungal environmental clone library construction, *Environ Microbiol* **9**, 1329–1334.
20. Taylor, D. L., Booth, M. G., Mcfarland, J. W., Herriott, I. C., Lennon, N. J., Nusbaum, C., and Marr, T. G. (2008) Increasing ecological inference from high throughput sequencing of fungi in the environment through a tagging approach, *Mol Ecol Resour* **8**, 742–752.
21. Geml, J., Laursen, G. A., and Taylor, D. L. (2008) Molecular diversity assessment of arctic and boreal Agaricus taxa, *Mycologia* **100**, 577–589.
22. Geml, J., Laursen, G. A., Timling, I., Mcfarland, J. M., Booth, M. G., Lennon, N., Nusbaum, C., and Taylor, D. L. (2009) Molecular phylogenetic biodiversity assessment of arctic and boreal ectomycorrhizal Lactarius Pers. (Russulales; Basidiomycota) in Alaska, based on soil and sporocarp DNA, *Mol Ecol* **18**, 2213–2227.
23. White, T. J., Bruns, T., Lee. S., Taylor, J. (1990) Amplification and direct sequencing of fungal ribosomal RNA Genes for phylogenetics, *PCR protocols: a guide to methods and applications* **42**, 315–322.
24. Ewing, B., Hillier, L., Wendl, M. C., and Green, P. (1998) Base-calling of automated sequencer traces using phred. I. Accuracy assessment, *Genome Res* **8**, 175–185.
25. Ewing, B., and Green, P. (1998) Base-calling of automated sequencer traces using phred. II. Error probabilities, *Genome Res* **8**, 186–194.
26. Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W. L., Russ, C., Lander, E. S., Nusbaum, C., and Jaffe, D. B. (2008) Quality scores and SNP detection in sequencing-by-synthesis systems, *Genome Res* **18**, 763–770.
27. Gordon, D., Abajian, C., and Green, P. (1998) Consed: A graphical tool for sequence finishing, *Genome Res* **8**, 195–202.
28. Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994) Clustal-W – improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice, *Nucleic Acids Res* **22**, 4673–4680.
29. Edgar, R. C. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res* **32**, 1792–1797.

30. Hall, T. A. (1999) BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT, *In*: Nucleic acids symposium series. p. 95–98.
31. Pertea, G., Huang, X. Q., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., Tsai, J., and Quackenbush, J. (2003) TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets, *Bioinformatics **19***, 651–652.
32. Huang, X. Q., and Madan, A. (1999) Cap3: A DNA sequence assembly program, *Genome Res **9***, 868–877.
33. Higgins, K. L., Arnold, A. E., Miadlikowska, J., Sarvate, S. D., and Lutzoni, F. (2007) Phylogenetic relationships, host affinity, and geographic structure of boreal and arctic endophytes from three major plant lineages, *Mol Phylogenet Evol **42***, 543–555.
34. Colwell, R. K., and Coddington, J. A. (1994) Estimating terrestrial biodiversity through extrapolation, *Philos T Roy Soc B **345***, 101–118.
35. McCune, B., Mefford, M. J. (1999) PC-ord. Multivariate analysis of ecological data, version 4(0).
36. Oksanen, J., Kindt, R., Legendre, P., O'Hara, B., Stevens, M. H. (2007) vegan: Community Ecology Package. R package version 1.8-8. Online at: http://r-forge.r-project.org/projects/vegan.
37. Lozupone, C., and Knight, R. (2005) UniFrac: a new phylogenetic method for comparing microbial communities, *Appl Environ Microb **71***, 8228–8235.
38. Webb, C. O., Ackerly, D. D., and Kembel, S. W. (2008) Phylocom: software for the analysis of phylogenetic community structure and trait evolution, *Bioinformatics **24***, 2098–2100.
39. Koljalg, U., Larsson, K. H., Abarenkov, K., Nilsson, R. H., Alexander, I. J., Eberhardt, U., Erland, S., Hoiland, K., Kjoller, R., Larsson, E., Pennanen, T., Sen, R., Taylor, A. F. S., Tedersoo, L., Vralstad, T., and Ursing, B. M. (2005) UNITE: a database providing web-based methods for the molecular identification of ectomycorrhizal fungi, *New Phytol **166***, 1063–1068.
40. Nilsson, R., Bok, G., Ryberg, M., Kristiansson, E., Hallenberg, N. (2009) A software pipeline for processing and identification of fungal ITS sequences. *Source Code Biol Med* **4**, 1.
41. Jumpponen, A. (2003) Soil fungal community assembly in a primary successional glacier forefront ecosystem as inferred from rDNA sequence analyses, *New Phytol **158***, 569–578.
42. Huber, T., Faulkner, G., and Hugenholtz, P. (2004) Bellerophon: a program to detect chimeric sequences in multiple sequence alignments, *Bioinformatics **20***, 2317–2319.
43. Perotto, S., Nepote-Fus, P., Saletta, L., Bandi, C., and Young, J. P. W. (2000) A diverse population of introns in the nuclear ribosomal genes off ericoid mycorrhizal fungi includes elements with sequence similarity to endonuclease-coding genes, *Mol Biol Evol **17***, 44–59.

Chapter 11

Identifying Protein Complexes by Affinity Purification and Mass Spectrometry Analysis in the Rice Blast Fungus

Wende Liu, Anton Iliuk, Andy Tao, and Shengli Ding

Abstract

Affinity purification and mass spectrometry analyses have been used in various organisms to identify protein complexes and determine protein–protein interactions in vivo. In comparison with the TAP (tandem affinity purification) tag, the 3× FLAG is a relatively small epitope tag. It has been used to systematically identify protein–protein interactions in the budding yeast. We have used the 3× FLAG tag to isolate proteins co-purified with a number of genes in the rice blast fungus, including *TIG1*, *MST50*, *PMK1*, and *MST12*. For the example given in the text, five genes homologous to components of the yeast Set3C complex were identified by mass spectrometry analysis.

Key words: *Magnaporthe oryzae*, Protein interaction, Interactome, Epitope tagging, Proteomics

1. Introduction

Rice blast disease caused by the heterothallic ascomycete *Magnaporthe oryzae* is one of the most severe fungal diseases of rice throughout the world (1). In the past decade, it has been developed as a model system to study fungal–plant interactions (2–4). Like in many other filamentous fungi, random insertional mutagenesis and targeted gene deletion are two approaches commonly used to determine the function of individual genes in *M. oryzae* (5–7). One limitation of these two approaches is that only one specific gene is genetically characterized in a disruption or deletion mutant. Because most proteins do not act alone and often interact with other proteins to carry out specific biological functions, it is critical to determine protein–protein interactions and characterize their relationship during fungal development and pathogenesis.

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722,
DOI 10.1007/978-1-61779-040-9_11,

In the budding yeast, both the TAP (tandem affinity purification)-tag and FLAG tag affinity purification approaches have been used to systematically study protein–protein interactions (8, 9). Similar approaches have been used in identifying protein complexes from other eukaryotes. TAP-tag purification has less background and is more sensitive. However, the advantages of the FLAG-tag approach include the small size of the tag (less impact on the target protein) and efficient one-step protein purification (10). In genome-wide systematic studies, FLAG-tag purification was as efficient as TAP-tag purification in establishing protein–protein interaction networks in *Saccharomyces cerevisiae* (8, 9). In filamentous fungi, there are only limited studies on isolating protein complexes by affinity purification (11). In *M. oryzae*, we have used the FLAG-tag to purify the *TIG1* complex and proteins interacting with various components of the *PMK1* pathway. Below is the procedure that was used in our labs to identify the *TIG1* histone deacetylase (HDAC) complex in the rice blast fungus.

2. Materials

2.1. 3× FLAG Fusion Construct

1. Advantage 2 Polymerase Mix: Clontech Laboratories, Inc., Mountain View, CA.
2. QIAquick Gel Extraction kit: Qiagen, Valencia, CA.
3. Alkali Cation Yeast transformation kit: BIO 101, Inc., Vista, CA.
4. SD-trp medium: 6.7 g yeast nitrogen base without amino acids (Difco Laboratories, Detroit, MI), 182.2 g D-sorbitol, 15 g agar, 0.74 g Trp DO supplement (Clontech Laboratories). Add double deionized H_2O (DDW) to a final volume of 960 mL. Adjust pH to 5.8. Autoclave for 15 min. Cool down to 55°C and then add 40 mL of 50% glucose (sterilized by filtration or autoclave for 15 min). Pour into petri plates.
5. The Pierce ECL Supersignal System: Thermo Fisher Scientific Inc., Rockford, IL.
6. Plasmid DNA of pDL2 (5).
7. Yeast strain XK1-25 (*MATa*, *trp1*) (12).
8. The *M. oryzae* wild-type strain 70-15 (13).

2.2. Fungal Cultures and Protein Extraction

1. 5× YEG liquid media (l L): 5 g yeast extract and 10 g glucose. Autoclave for 20 min.
2. Complete media (CM): 50 mL 20× nitrate salts (see below), 10 g D-glucose, 2 g peptone, 1 g yeast extract, 1 g casamino acids, 1 mL vitamin solution (see below), pH 6.5 (w/NaOH). Add DDW to 1 L. Autoclave for 20 min.

3. 20× nitrate salts (1 L): 120 g $NaNO_3$, 10.4 g KCl, 10.4 g $MgSO_4 \cdot 7H_2O$, and 30.4 g KH_2PO_4. Autoclave for 15 min. Store at 4°C.
4. Vitamin solution (100 mL): 0.01 g biotin, 0.01 g pyridoxine, 0.01 g thiamine, 0.01 g riboflavin, 0.01 g PABA (*p*-aminobenzoic acid), and 0.01 g nicotinic acid.
5. Miracloth: EMD Biosciences, Inc., La Jolla, CA.
6. Lysis buffer: 50 mM Tris–HCl (pH 7.4), 150 mM NaCl, 1 mM EDTA, and 1% Triton X-100. Autoclave for 20 min. Store at 4°C.
7. Protease inhibitor cocktail: Sigma-Aldrich, St. Louis, MO.
8. 0.1 M phenylmethylsulfonyl fluoride (PMSF): Dissolve 17.4 mg PMSF (Sigma-Aldrich) in 1 mL methanol. Store at –20°C.
9. Acid-washed glass beads (0.5 mm): Biospec Products, Inc. Bartlesville, OK. Autoclave for 40 min.
10. A mini bead beater: Biospec Products, Inc.

2.3. Immunoprecipitation of 3× FLAG Fusion Proteins

1. EZview Red ANTI-FLAG M2 Affinity Gel: Sigma-Aldrich.
2. Tris-buffered saline (TBS): 50 mM Tris–HCl (pH 7.4) and 150 mM NaCl.
3. 50 mM trimethylammonium bicarbonate buffer (TMAB): Dilute 1 M TMAB buffer (Sigma-Aldrich) with DDW.
4. 0.1% RapiGest (14): RapiGest (Waters Corporation, Milford, MA) dissolved in 50 mM TMAB.
5. A GyroMini mini-rocker shaker: Labnet International, Inc., Woodbridge, NJ.

2.4. Protein Denaturalization and Digestion

1. 5 mM dithiothreitol (DTT): dissolve in water and store in single-use aliquots at –20°C.
2. 15 mM iodoacetamide: freshly prepared in water and used immediately.
3. Trypsin (1 μg): dissolve in 5 μL of trypsin resuspension buffer (or 50 mM acetic acid) and stored at –20°C.
4. Hydrochloric acid is diluted to 1 M with water.
5. MS analysis solution: 0.1% formic acid in water.
6. Savant Speed Vac concentrator.

2.5. Liquid Chromatography Tandem Mass Spectrometry

1. Agilent HPCL autosampler vials: Agilent Technologies, Wilmington, DE.
2. Agilent 1100 nanoflow HPLC system: Agilent Technologies.
3. Thermo LTQ-Orbitrap XL mass spectrometer: ThermoFisher, San Jose, CA.
4. Proteome Discoverer software: Thermo Fisher.

3. Methods

3.1. Generating Transformants Expressing the –3× FLAG Fusion Construct

The 3× FLAG tag contains three tandem FLAG epitope tags, which enhances the detection of fusion proteins by several commercially available, highly specific anti-FLAG monoclonal antibodies (M1, M2, and M5 from Sigma-Aldrich). It is suitable for constructing both N- or C-terminal fusion proteins.

1. The *TIG1* coding region was amplified with primers RPF (tttcgtaggaacccaatcttcaaaatgaaggaatttctcgactcgg) and RPR (gaacagctcctcgcccttgc*tcactta*cttgtcatcgtcatccttgtaatcgatatcatgatctttataatcaccgtcatggtctttgtagtccct*aaagttgatcacagcg*) with the Advantage 2 Polymerase Mix. Primer RPR contains sequences encoding three copies of the FLAG epitope sequence (underlined) followed by stop codons in two frames (in italic). PCR conditions: 95°C 2 min, 30 cycles of 94°C 30 s and 68°C 60 s, followed by 68°C 10 min.
2. The resulting PCR products were purified with the QIAquick Gel Extraction kit (Qiagen) and co-transformed with *Xho*I-digested pDL2 in equal molecular ratio into competent cells of yeast strain XK1-25 prepared with the Alkali Cation Yeast transformation kit (BIO 101). The transformation mixture was plated out on SD-Trp plates and incubated at 30°C for 2 days.
3. The *TIG1*–3× FLAG fusion vector pSD37 was recovered from the TRP^+ yeast transformants and confirmed by sequence analysis (see Note 1).
4. Plasmid DNA of pSD37 was transformed into protoplasts of *M. oryzae* strain 70-15. The resulting hygromycin-resistant transformant TFG3 was confirmed to contain the transforming pSD37 in the genome by Southern blot analysis. The expression of the *TIG1*–3× FLAG fusion construct was confirmed by western blot analysis (Fig. 1) with a monoclonal anti-FLAG antibody (M2, Sigma-Aldrich).

3.2. Isolation of Total Proteins from Fungal Hyphae

1. Transfer one square inch of 10-day-old oatmeal agar cultures into a sterile blender cup. Add 50 mL of 5× YEG and blend for 1 min. Transfer the culture mixture to a 250-mL flask and incubate for 24 h at 25°C with shaking (150 rpm).
2. Transfer the overnight culture to a sterile blender cup and blend for 1 min. Pour the blended culture back to the 250-mL flask and add 50 mL of complete medium (CM). Incubate for 24 h at 25°C with shaking (150 rpm).
3. Vegetative hyphae were harvested from 2-day-old cultures by filtering through one layer of Miracloth and rinsed with 100 mL of sterile distilled water.

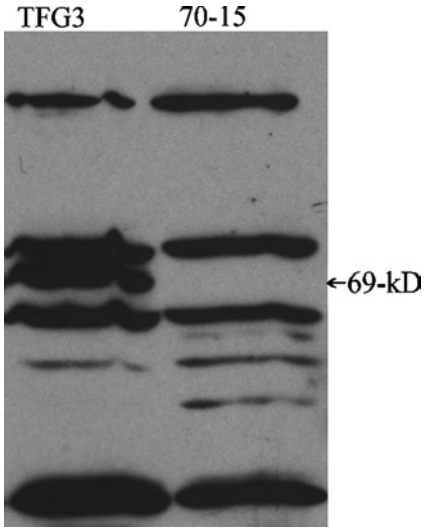

Fig. 1. Total proteins were isolated from the wild-type strain 70-15 and transformant TFG3 and separated on a 12% SDS-PAGE gel. After blotting on to a nitrocellulose membrane, the presence of FLAG-fusion proteins of expected 69 kD was detected with the anti-FLAG antibody (Sigma) using the Pierce ECL Supersignal System.

4. Remove excessive water by pressing hyphal pads between layers of paper towels.
5. Resuspend 250 mg of semi-dry vegetative mycelia in 2 mL of protein lysis buffer (4°C) with 10 µL of protease inhibitor cocktail and 10 µL of 0.1 M PMSF. Add 0.35 g of sterile, acid-washed glass beads.
6. Hyphae were homogenized with a Biospec mini bead beater for five times of 40 s beating with 2 min interval on ice (see Note 2).
7. The lysate was centrifuged at 25,000 × *g* in a HERMLE Z383K centrifuge for 20 min at 4°C.
8. The supernatant containing soluble proteins was transferred to a new sterile Eppendorf tube that was prechilled on ice (see Note 3). Store 50 µL of the supernatant at –80°C for using as the total protein control in western blot analyses.

3.3. Binding of 3× FLAG Fusion Proteins to the Anti-FLAG Beads

1. Gently shake the tube containing the EZview Red anti-FLAG M2 Affinity Gel until the beads are completely suspended. Transfer an aliquot containing 40 µL of the 50% slurry into a clean 1.5-mL microcentrifuge tube that has been prechilled on ice (see Note 4).
2. Add 500 µL of prechilled TBS. Vortex briefly and centrifuge for 30 s at 18,000 × *g* at 4°C.
3. Carefully remove the supernatant with a pipette or by aspiration. Set the tube with the anti-FLAG bead pellet on ice.

4. Repeat washing the anti-FLAG beads with 500 µL of prechilled TBS as described above (see Note 5).
5. Add 200–1,000 µL of total proteins extracted from fungal hyphae (the supernatant from Subheading 3.2, step 7 (see above) to the washed beads) (see Note 6).
6. Incubate the protein–bead mixture at 4°C for 2 h with gentle shaking (40 rpm) (see Note 7).

3.4. Elution of 3× FLAG Fusion Proteins

1. Centrifuge the protein–bead mixture at 4°C in a microcentrifuge for 30 s at 8,200 × *g*. Carefully remove the supernatant with a sterile pipette tip. Place the tube with the beads on ice (see Note 8).
2. Add 500 µL of prechilled lysis buffer to the pellet. Gently resuspend the beads.
3. Incubate at 4°C for 5 min with gentle shaking. Centrifuge at 4°C for 30 s at 8,200 × *g*. Remove the supernatant.
4. Repeat washing the beads (steps 2 and 3) with 500 µL of prechilled lysis buffer two more times.
5. Wash the beads three times with 500 µL each of prechilled 50 mM trimethylammonium bicarbonate (TMAB) as outlined in steps 2 and 3.
6. Wash the beads three times with 500 µL each of prechilled sterile distilled water using similar approaches described in steps 2 and 3.
7. Add 100 µL of prechilled 0.1% RapiGest (14) to the pellet. Incubate at 25°C for 5 min with gentle shaking.
8. Centrifuge the sample for 30 s at 8,200 × *g* then incubate it in a boiling water bath for 5 min.
9. Place the sample on ice for 2 min. Centrifuge at 8,200 × *g* for 30 s at 4°C.
10. Transfer the supernatant containing proteins eluted from the anti-FLAG beads into a new Eppendoff tube prechilled on ice (see Note 9).
11. Take 10 µL of the eluate for western blot analysis (Fig. 2). The rest will be used for mass spectrometry analysis (see Note 10).

3.5. Protein Denaturalization and Digestion

1. The proteins eluted with 50 mM TMAB containing 0.1% RapiGest are reduced with 5 mM dithiothreitol (DTT) for 30 min at 50°C to remove disulfide bridges, cooled down, and then alkylated with 15 mM iodoacetamide for 1 h at room temperature in the dark (15) (see Note 11).
2. The pH is adjusted with 2–5 µL TMAB to ~8.0.
3. Proteins are digested with 1 µg trypsin overnight (or 12–16 h) at 37°C (see Note 12).

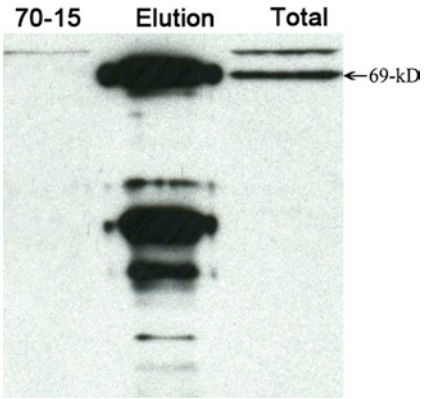

Fig. 2. Western blot analysis with the anti-FLAG antibody. The left lane was loaded with total proteins from the wild-type strain (70-15) as the negative control. The right and middle lanes were loaded with total proteins (total) and proteins eluted from anti-FLAG beads (elution) of the *TIG1*–3× FLAG transformant TFG3.

4. The resulting peptides are mixed with 1 M hydrochloric acid to a final concentration of 50 mM HCl and a pH <3.0.
5. Incubate the sample in a 37°C water bath for 45 min.
6. Centrifuge at 16,100 × *g* for 10 min.
7. Carefully transfer the supernatant into a clean microcentrifuge tube without disturbing the pellet.
8. The soluble peptide sample is dried completely using a Savant SpeedVac concentrator and stored at −20°C until MS analysis (16).

3.6. Mass Spectrometry-Based Peptide Analysis

1. The dried peptide sample is resuspended in 9 µL of 0.1% formic acid and transferred into an Agilent HPLC autosampler vial.
2. Load 8 µL of the sample by autosampler injection onto a C18 precolumn using the Agilent 1100 nanoflow HPLC system.
3. The peptides are eluted by hydrophobicity at 300 nL/min using a 90 min increasing acetonitrile gradient and analyzed subsequently by the in-line Thermo LTQ-Orbitrap XL mass spectrometer.
4. The instrument is operated at the data-dependant mode with one MS scan followed by four MS/MS scans.
5. Search the resulting MS data with the Proteome Discoverer software using the SEQUEST algorithm against corresponding protein database with activated reverse database search to estimate false discovery rate (FDR).
6. Static modification is put into search parameters to account for alkylated cysteine residues (+57.0214).

Table 1
Proteins copurified with the *TIG1* gene in vegetative hyphae

Gene ID	Yeast homolog	Predicted function
MGG_01633	*HOS2*	Histone deacetylase
MGG_09174	*SNT1*	DNA binding protein
MGG_02488	*HST1*	Histone deacetylase
MGG_10447	*CPR1*	Catalyzes the cis–trans isomerization of peptide bonds N-terminal to proline residues
MGG_05727	YIL112w (*HOS4*)	Four ankyrin repeats protein
MGG_03741	YGR266W	Protein of unknown function

7. Variable modification of +15.9949 is put into search parameters to account for possible oxidation of methionine residues.
8. The conditions should be set where only peptides with the FDR <5% (or <1% for more stringent results) are accepted as correct identifications.

In our studies with the *TIG1* gene, putative Tig1-interacting genes were identified by mass spectrometry analysis with three independent biological samples of proteins co-immunoprecipitated with the *TIG1*–3× FLAG fusion. Proteins that bind unspecifically to the anti-FLAG antibody were removed by comparing with background (abundant) proteins that also were copurified with various components of the Pmk1 MAP kinase pathway. Table 1 lists predicted *M. oryzae* genes that appeared to be specifically copurified with *TIG1*–3× FLAG. *TIG1* is an ortholog of yeast *SIF2*, which is a component of the Set3 complex (17). The top five genes in the *Tig1* pull down list are homologous to yeast, *HOS2*, *SNT1*, *HST1*, *CPR1*, and *HOS4* genes (Table 1). All of them are components of the yeast Set3 complex involved in the regulation of late stage sporulation genes (17).

4. Notes

1. In this example, the resulting construct is a *TIG1*–3× FLAG C-terminal fusion. To generate the N-terminal fusion constructs, the oligo nucleotides encoding the 3× FLAG epitope sequence should be placed right behind the start cordon of the target gene with similar yeast GAP repair approach.
2. Continuous grinding may lead to elevated temperature and result in protein degradation. If necessary, the interval on ice

could be increased to 5 min. Do not use liquid nitrogen for grinding because some protein complexes or protein–protein interactions may be sensitive to freezing temperatures.

3. Do not disturb the interface. Although 2 mL lysis buffer is added, normally only about 1 mL of the supernatant could be recovered.
4. To avoid damaging the beads, use tips with wider openings to transfer the beads. Cutting off about 1 mm from the tip of regular 200 μL Pipetman tips will work well for this purpose.
5. If there are multiple samples for immunoprecipitation, the anti-FLAG resin needed for all the samples can be prepared together. For each wash, at least 20 times of the bed volume of TBS should be used. After washing three times with TBS (at least 20 times of the bed volume for each wash), the anti-FLAG resin can divided into the desired number of aliquots.
6. The volume of protein extract to be used depends on the expression level of 3× FLAG fusion protein in vegetative hyphae of the transformant. If necessary, bring the final volume to 1 mL with lysis buffer.
7. A mini-rocker shaker (such as the GyroMini from the Labnet) is better than a regular orbital bench top shaker for keeping anti-FLAG beads in suspension.
8. If desired, the supernatant could be saved as the washing through control.
9. If necessary, the anti-FLAG beads can be eluted again with 50 μL of prechilled 0.1% RapiGest as described in steps 6–9 to increase the recovery efficiency. The first and second eluates can be combined.
10. The volume could be varied from 10 to 20 μL, depending on the concentration of proteins eluted from the beads.
11. Check the pH before adding the iodoacetamide. If pH is below 7.0, add 2–5 μL of 1 M TMAB to increase the pH to above neutral.
12. If more than 100 μg of total protein is present in the sample, add trypsin at 1:100 ratio.

Acknowledgments

The authors would like to thank Xiaoying Zhou at Purdue University for technical assistance. This work was supported by a USDA-NRI grant 2005-35319-16073.

References

1. Ford, T. L., Cooley, J. T., and Christou, P. (1994) Current status for gene transfer into rice utilizing variety-independent delivery systems. In *Rice Blast Disease*. Ziegler, R.S., Leong, S.A., and Teng, P.S., (eds). Wallingford, UK: CAB International, pp. 195–208.
2. Talbot, N. J. (2003) On the trail of a cereal killer: Exploring the biology of *Magnaporthe grisea*. *Annu. Rev. Microbiol.* **57**, 177–202.
3. Xu, J. R., Zhao, X., and Dean, R. A. (2007) From genes to genomes: A new paradigm for studying fungal pathogenesis in *Magnaporthe oryzae*. *Adv. Genet.* **57**, 175–218.
4. Ebbole, D. J. (2007) *Magnaporthe* as a model for understanding host-pathogen interactions. *Annu. Rev. Phytopathol.* **45**, 437–456.
5. Li, L., Ding, S. L., Sharon, A., Orbach, M., and Xu, J. R. (2007). Mir1 is Highly Upregulated and Localized to Nuclei During Infectious Hyphal Growth in the Rice Blast Fungus. *Mol. Plant-Microbe Interact.* **20**, 448–458.
6. Betts, M. F., Tucker, S. L., Galadima, N., Meng, Y., Patel, G., Li, L., Donofrio, N., Floyd, A., Nolin, S., Brown, D., Mandel, M. A., Mitchell, T. K., Xu, J. R., Dean, R. A., Farman, M. L., and Orbach, M. J. (2007). Development of a high throughput transformation system for insertional mutagenesis in *Magnaporthe oryzae*. *Fungal Genet. Biol.* **44**, 1035–1049.
7. Valent, B., and Chumley, F. G. (1991). Molecular Genetic Analysis of the Rice Blast Fungus, *Magnaporthe grisea*. *Annu. Rev. Phytopathol.* **29**, 443–467.
8. Gavin, A. C., Aloy, P., Grandi, P., Krause, R., Boesche, M., et al., (2006). Proteome survey reveals modularity of the yeast cell machinery. *Nature* **440**, 631–636.
9. Krogan, N. J., Cagney, G., Yu, H., Zhong, G., Guo, X., et al., (2006). Global landscape of protein complexes in the yeast *Saccharomyces cerevisiae*. *Nature* **440**, 637–643.
10. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., Moore, L., et al., (2002). Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry. *Nature* **415**, 180–183.
11. Liu, W. D., Ding, S. L., and Xu, J. R. The Set3 complex regulates penetration and infectious growth in *Magnaporthe grisea*. Fungal Genet. Newsl. **58S**, 557.
12. Bruno, K. S., Tenjo, F., Li, L., Hamer, J. E., and Xu, J. R. (2004). Cellular localization and role of kinase activity of *PMK1* in *Magnaporthe grisea*. *Eukaryotic Cell* **3**, 1525–1532.
13. Dean, R. A., Talbot, N. J., Ebbole, D. J., Farman, M. L., Mitchell, T. K., et al., (2005). The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature* **434**, 980–986.
14. Yu, Y. Q., Gilar, M., Lee, P. J., Bouvier, E. S. P., and Gebler, J. C. (2003) Enzyme-friendly, mass spectrometry-compatible surfactant for in-solution enzymatic digestion of proteins. *Anal. Chem.* **75**, 6023–6028.
15. Tao, W. A., Wollscheid, B., O'Brien, R., Eng, J., Li, X., Bodenmiller, B., Watts, J., Hood, L., and Aebersold, R. (2005) Quantitative phosphoproteome analysis using a dendrimer conjugation chemistry and mass spectrometry. *Nature Methods* **2**, 591–598.
16. Benjamini, Y., and Hochberg, Y. (1995) Controlling the false discovery rate – a practical and powerful approach to multiple testing. *J. Royal Stat. So. Ser. B-Method.* **57**, 289–300.
17. Pijnappel, W. W, Schaft, D., Roguev, A., Shevchenko, A., Tekotte, H., Wilm, M., Rigaut, G., Séraphin, B., Aasland, R., and Stewart, A. F. (2001). The *Saccharomyces cerevisiae SET3* complex includes two histone deacetylases, Hos2 and Hst1, and is a meiotic-specific repressor of the sporulation gene program. *Genes Dev.* **22**, 2991–3004.

Chapter 12

Large Scale Identification of Genes Involved in Plant–Fungal Interactions Using Illumina's Sequencing-by-Synthesis Technology

R.C. Venu, Yuan Zhang, Brian Weaver, Peter Carswell, Thomas K. Mitchell, Blake C. Meyers, Michael J. Boehm, and Guo-Liang Wang

Abstract

Deep transcriptome profiling of pathogen-infected tissues enhances the understanding of molecular mechanisms underlying host–pathogen interactions. Illumina's next generation sequencing technology sequencing-by-synthesis (SBS) is a powerful tool to rapidly sequence genomes and transcriptomes at an affordable rate. We modified the procedure for SBS library construction to significantly increase the efficiency of library construction. Using our improved method, two *Sclerotinia homoeocarpa* libraries were constructed from mycelia grown in potato dextrose broth (PDB) or potato dextrose agar (PDA) for 96 h, respectively, and two creeping bentgrass libraries were constructed from leaves 96 h after inoculation with *S. homoeocarpa* or water sprayed, respectively. About 4–7 million mRNA signatures were sequenced from each library. Sequence analysis using BLAST was performed against sequenced fungal genomes and rice genomic sequence to identify the expressed genes in both *S. homoeocarpa* mycelia and creeping bentgrass. Bioinformatic analysis identified many expressed genes in the pathogen and host. A public database to access the sequence data was developed at http://www.dstidb.org. Our results demonstrate how SBS technology can unravel transcriptome complexity during the creeping bentgrass–*S. homoeocarpa* interaction.

Key words: Sequencing-by-synthesis, Transcriptome analysis, Creeping bentgrass, *Sclerotinia homoeocarpa*, Bioinformatics

1. Introduction

Dollar spot, caused by *Sclerotinia homoeocarpa* F.T. Bennett, is one of the most devastating diseases of creeping bentgrass (turf) on golf courses throughout the world (1–3). *S. homoeocarpa*

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722, DOI 10.1007/978-1-61779-040-9_12,

produces oxalic acid and many cell wall degrading enzymes *in vitro* similar to other members of the genus *Sclerotinia* (4), and infects a wide range of monocots and dicots (5). Its broad host range underscores the significance of its secretome in pathogenesis. Identification of important host and fungal genes expressed specifically during the infection process is important to advance the current understanding of pathogenesis in *S. homoeocarpa* and defense mechanisms in creeping bentgrass.

Over the past decade, microarrays and tag-based approaches such as serial analysis of gene expression (SAGE) or massively parallel signature sequencing (MPSS) were widely used to understand host-pathogen interactions in many crop plants (6–9). However, with the recent development of next generation sequencing (NGS) technologies, it is now possible to produce millions of DNA sequences in a single run in few days. NGS technologies include 454 Life Sciences's pyrosequencing (10), Illumina's SBS (11), and ABI's sequencing by oligo ligation and detection system (SOLiD) (12, 13). The major advantages of these NGS technologies are deep sequencing coverage to identify lowly expressed genes, simplified library construction, and low sequencing cost, all of which make these technologies attractive for a wide range of applications in genomic studies such as host–pathogen interactions.

Dollar spot management in intensively cultivated turfgrass depends heavily on timely fungicide applications. With integrated pest management strategies, dollar spot can be managed effectively (1–3). However, with changes in population dynamics of the dollar spot fungus and concerns regarding the effects of applied fungicides on human health and the environment, alternative strategies/fungicides targets to control dollar spot are needed. Understanding the creeping bentgrass–*S. homoeocarpa* interaction at the molecular level will accelerate the development of alternative disease management strategies. Specifically, identifying turf and *S. homoeocarpa* genes expressed during the host–pathogen interactions will reveal molecular events underlying pathogen recognition and disease development of dollar spot. Additionally, identifying and characterizing genes in *S. homoeocarpa* encoding secretory proteins will provide information about the interaction with turf grasses. Therefore, we initiated a project to understand creeping bentgrass–dollar spot interactions at the molecular level. We made several modifications in the Illumina's sample preparation protocol that can double the number of libraries to be made with the reagents provided in the kit. Also, we developed a bioinformatics pipeline to analyze the millions of signatures generated by Illumina's genome analyzer. Our methods described here can be used for SBS library construction and data analysis in any other fungal or plant species.

2. Materials

2.1. S. homoeocarpa and Creeping Bentgrass Growth Conditions

1. *S. homoeocarpa* isolate MB01.
2. Potato dextrose agar.
3. Potato dextrose broth.
4. Incubator-shaker.
5. Creeping bentgrass seeds (creeping bentgrass variety – Crenshaw).
6. Conviron growth chamber.

2.2. S. homoeocarpa Inoculation

1. 50-mL centrifuge tubes.
2. Sterile distilled water.
3. Mortar and pestle.
4. Hemocytometer.

2.3. Total RNA Isolation and SBS Library Construction

Illumina's digital gene expression-tag profiling *Dpn*II sample preparation kit was used to construct the libraries for SBS (Catalog FC-102-1007(1000910)). The original protocol can be downloaded at http://www.genomics.ucr.edu/facility/genomics/DGE_DpnII_Sample_Prep.pdf.

2.3.1. RNA Isolation

1. Liquid nitrogen.
2. Trizol solution.
3. Chloroform.
4. Isopropanol.
5. 95 and 75% ethanol.
6. Diethyl pyrocarbonate (DEPC)-treated H_2O.

2.3.2. Single Strand and Second Strand Synthesis (All Reagents Are from Illumina's Kit Unless Specified)

1. Silicon.
2. GEX Sera-Mag magnetic oligo(dT) beads.
3. GEX binding buffer.
4. GEX washing buffer.
5. 5× first strand buffer.
6. 10 mM 5mC–dNTP mix.
7. RNase H enzyme.
8. SuperScript II reverse transcriptase with 100 mM DTT (Invitrogen, Carlsbad, CA).
9. GEX 2nd strand buffer.
10. DNA Polymerase I.
11. GEX cleaning solution additive.

12. 10× *Dpn*II buffer.
13. GEX buffer C.
14. GEX buffer D.
15. Magnetic stand (Invitrogen, Carlsbad, CA).
16. RNase- and DNase-free water.
17. Thermomixer R (Eppendorf).

2.3.3. Restriction Digestion with DpnII and Ligation with GEX DpnII Adapter 1 (All Reagents Are from Illumina's Kit Unless Specified)

1. 10× *Dpn*II buffer.
2. *Dpn*II enzyme.
3. 5× T4 DNA ligase buffer.
4. T4 DNA ligase.
5. GEX *Dpn*II adapter 1.
6. 10× restriction enzyme buffer.

2.3.4. Restriction Digestion with MmeI and Ligation with GEX Adapter 2 (All Reagents Are from Illumina's Kit Unless Specified)

1. 10× restriction buffer.
2. GEX 32 mM *S*-adenosylmethionine.
3. *Mme*I.
4. CIAP.
5. Glycogen.
6. Phenol/chloroform/isoamyl alcohol (25:24:1) (Invitrogen, Carlsbad, CA).
7. 3 M sodium acetate, pH 5.2.
8. –20°C 100% ethanol.
9. Room temperature 70% ethanol.
10. GEX adapter 2.
11. Savant Speed Vac.

2.3.5. PCR and Gel Purification of the cDNA Tags (All Reagents Are from Illumina's Kit Unless Specified)

1. 5× Phusion HF buffer (Finnzymes Oy, Espoo, Finland).
2. Phusion Polymerase (Finnzymes Oy, Espoo, Finland).
3. Primer GX2.
4. 25 mM dNTP mix.
5. Primer GX1.
6. 25 bp ladder.
7. 6× DNA loading Dye.
8. PAGE electrophoresis system (Bio-Rad, Hercules, CA).
9. Acrylamide bisacrylamide mix.
10. TEMED.
11. Electrophoresis power supply (Bio-Rad, Hercules, CA).

12. 10× gel elution buffer.
13. Resuspension buffer.
14. Spin-X cellulose acetate filter.
15. Glycogen.

3. Methods

3.1. S. homoeocarpa Isolation from Leaf Tissue

S. homoeocarpa isolate MB01 was originally isolated from symptomatic leaves of creeping bentgrass (*Agrostis stolonifera* L.; syn = *A. palustris* Huds.) growing at the Ohio Turf Foundation research center, Columbus, OH. Diseased leaf tissue was surface disinfected for 1 min in a 3% sodium hypochlorite solution, rinsed twice in sterile water, and placed on acidified potato dextrose agar prepared by adding 0.75 mL of 85% lactic acid (Fisher Scientific, Fair Lawn, NJ) per 1 L of PDA (Difco, Becton Dickinson and Company, MD) after autoclaving. Cultures were incubated at 25°C and *S. homoeocarpa* was purified by repeatedly transferring hyphal tips to new plates as required. *S. homoeocarpa* was identified by cultural morphology and microscopic observation as originally described by Bennett (1937) (14). Isolate MB01 is sensitive to benzimidazole, DMI and dicarboximide fungicides.

3.2. S. homoeocarpa Inoculum Preparation

1. Transfer one 5-mm diameter plug of actively growing *S. homoeocarpa* to 100 mL potato dextrose broth (24 g/L; pH 6.0). Incubate at 25°C for 48 h with shaking at 150 rpm.
2. Fragment the mycelial balls with a mortar and pestle and adjust the concentration of inoculum to 1×10^5 mycelial fragments/mL.

3.3. S. homoeocarpa Mycelium Tissue from In Vitro: PDB Culture

1. Grow *S. homoeocarpa* isolate MB01 in PDB. Adjust each flask of PDB (24 g/L) to pH 6.0, inoculate with one 5-mm diameter plug of actively growing *S. homoeocarpa* in 100 mL PDB, and incubate at 25°C for 96 h with shaking at 150 rpm.
2. Centrifuge the medium to collect the mycelium and freeze it in liquid nitrogen for RNA isolation.

3.4. S. homoeocarpa Mycelium Tissue from In Vitro: PDA Culture

1. Grow *S. homoeocarpa* isolate MB01 on PDA (39 g/L) adjusted to pH 6.0 with 1 M potassium hydroxide (Fisher Scientific, Pittsburgh, PA). Agar plugs (5 mm in diameter) containing actively growing *S. homoeocarpa* mycelium were transferred to PDA plates and the plates were incubated for 5 days at 26°C with no illumination.
2. Collect the mycelium with a clean scalpel, and freeze it in liquid nitrogen for RNA isolation.

3.5. Leaf Tissue from S. homoeocarpa-Infected Crenshaw Creeping Bentgrass Plants

1. Grow Crenshaw seedlings for 21 days in a Conviron growth chamber at 26°C, 80% relative humidity, and 12 h light at 200 μmol photons m^{-2}/s during the day, and 22°C, 60% relative humidity at night.
2. Spray *S. homoeocarpa* inoculum (concentration – 1×10^5 mycelia fragments/mL) evenly on Crenshaw plants. For control plants, spray water only.
3. Keep the plants in a sealed plastic container in the dark for 24 h with 100% humidity. Remove the bags and transfer the plants to a Conviron growth chamber for 96 h to promote disease development.
4. Harvest infected leaves at 96 h after inoculation for RNA isolation.

3.6. SBS Library Construction

The diagrammatic representation of the SBS procedure is shown in Fig. 1. Detailed protocol of SBS library construction is available at http://www.illumina.com/products/dge_tag_profiling_sample_prep_kits.ilmn. Please see Illumina's kit instructions for reagent and protocol details. Below are the major steps and some modifications made to the original SBS library construction protocol (indicated in **BOLD**). **To double the number of samples prepared per kit, you can use all reagents at half of the volume recommended in Illumina's protocol** (see Notes 1 and 2).

1. Grind approximately 2 g of leaf or fungal mycelium into a fine powder using liquid nitrogen. Immediately transfer ground tissue into 15 mL of Trizol solution. Mix well and incubate at room temperature for 10 min.
2. Add 4 mL of chloroform, incubate at room temperature for 5 min, and then centrifuge for 20 min (7,000 × *g*) at 4°C. Transfer supernatant into 25-mL tube containing 10 mL of ice-cold isopropanol, mix well, and then incubate on ice for 10 min. Centrifuge for 15 min (7,000 × *g*) at 4°C.
3. Wash the RNA pellet with 15 mL of 75% ethanol, centrifuge for 10 min (7,000× *g*), and then discard the alcohol. Dry the RNA pellet at room temperature for 10–15 min.
4. Dissolve the RNA pellet in 700 μL of DEPC-treated H_2O at 65°C for 10 min. Quantify the amount of total RNA with a spectrophotometer by measuring the OD at 260/280 nm. For SBS library construction, dilute **1–6** μg **of total RNA** (see Note 3) with RNase-free water to 50 μL, then heat at 65°C in a thermal cycler for 5 min and immediately place on ice.
5. Add 50 μL oligo(dT) beads in GEX binding buffer and then mix 50 μL of diluted total RNA. Remove the supernatant and wash beads by resuspending them twice in 200 μL of GEX washing buffer.

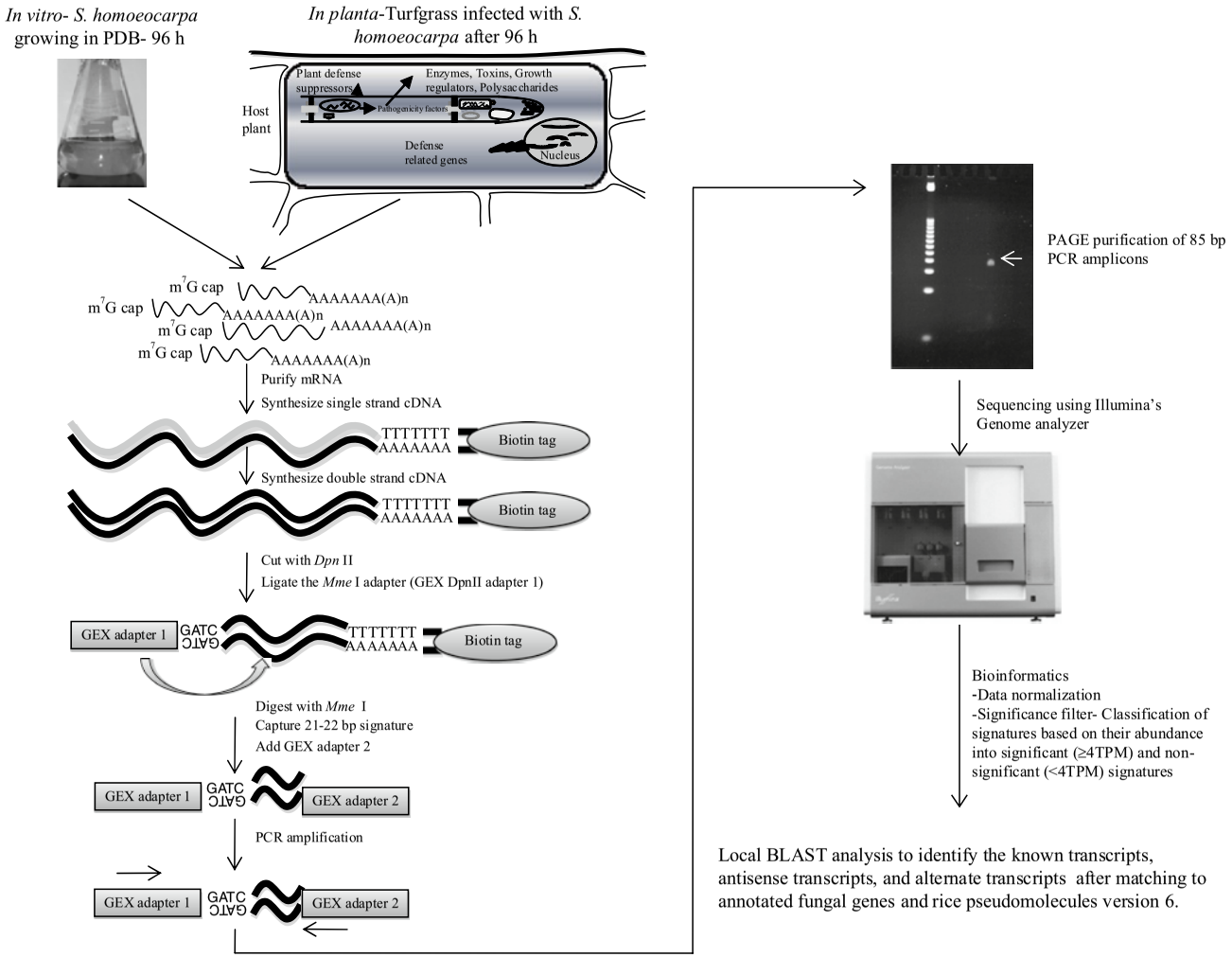

Fig. 1. Diagrammatic representation of sequencing-by-synthesis (SBS) methodology and data analysis pipeline. Total RNA from *Sclerotinia homoeocarpa* grown for 96 h in potato dextrose medium and turfgrass infected with *S. homoeocarpa* after 96 h is isolated. Library construction was done according to the Illumina's kit instructions with few modifications. The 85 bp PCR amplicons are purified using poly acrylamide gel electrophoresis (PAGE) and sequenced by Illumina's genome analyzer. The data was normalized across the libraries and used for BLAST analysis.

6. For single strand cDNA synthesis, wash the beads by resuspending them in 100 μL freshly prepared 1× first strand buffer. Resuspend the beads in 48 μL of the first strand cDNA synthesis premix and SuperScript II Reverse Transcriptase and incubate at 42°C in thermomixer for 1 h.
7. For second strand synthesis, add GEX second strand buffer and 5mC–dNTP mix in the mRNA/cDNA hybrid. Then, add DNA Polymerase I and RNase H and incubate for 2.5 h. at 16°C Wash the beads by resuspending them in GEX buffer C. Again, wash and resuspend the beads in GEX buffer D. Finally, resuspend the beads in 100 μL of 1× *Dpn*II buffer.
8. Resuspend the beads in *Dpn*II digestion premix and *Dpn*II enzyme and incubate at 37°C in a thermomixer for 1 h (see Note 4). Wash the beads by resuspending them in GEX buffer C. Again, wash and then resuspend the beads in 100 μL fresh working cleaning solution for 15 min at 37°C. Finally, resuspend the beads in GEX buffer D.
9. For ligation of GEX *Dpn*II Adapter 1, resuspend the beads in 100 μL of 1× T4 DNA ligase buffer. Add ultra pure water, GEX *Dpn*II adapter 1, 5× T4 DNA ligase buffer and T4 DNA ligase. **Incubate at 20°C in a thermomixer for 4 h.** Wash the beads by resuspending them in GEX buffer C. Then, wash and resuspend the beads in 100 μL of fresh working cleaning solution. Then, wash with GEX buffer D and resuspend the beads in 100 μL of 1× restriction buffer (see Note 5).
10. Resuspend the beads in 100 μL of *Mme*I restriction digest premix and **incubate at 37°C in a thermomixer for 2 h**. Collect the supernatant and add CIAP for dephosphorylation (see Note 6). Then, extract with phenol:chloroform:isoamyl alcohol (25:24:1) and precipitate the DNA by adding 1 μL of glycogen, 10 μL 3 M sodium acetate, and 325 μL –20°C 100% ethanol. Wash the DNA pellet with room temperature 70% ethanol. Dry the pellet using the speed vac. Resuspend the pellet in 6 μL ultra pure water and add GEX Adapter 2, 5× T4 DNA ligase buffer and T4 DNA ligase. Incubate the mix at 20°C for 4 h in a thermomixer (see Note 5).
11. Prepare the PCR reaction mix following the instructions described in Illumina's kit. **Instead of using 2.5 GEX adapter 2-ligated cDNA as recommended in the Illumina's kit, take 1.5 μL of the ligation mix for PCR amplification so that the ligation mix is enough for seven PCR reactions. PCR is set for 20 cycles.** PCR amplicons are separated on a **6% TBE polyacrylamide gel at 50 V for 5 h** (see Note 7).
12. For cDNA tag purification, excise the 85 bp band from the acrylamide gel. Extract the DNA with 1× gel elution buffer (see Note 8). Precipitate the DNA with 1 μL glycogen, 10 μL

3 M sodium acetate, and 325 μL cold ethanol (−20°C). Wash the DNA pellet with 500 μL room temperature 70% ethanol. Resuspend the pellet in 10 μL resuspension buffer and use it for sequencing with Illumina's genome analyzer.

3.7. SBS Data Analysis (see Note 9)

1. Normalize expressed signatures in each library to Transcripts Per Million (TPM) (15, 16). The signatures should be classified into significant (≥4 TPM) and nonsignificant (≤4 TPM) based on their abundance in the library.
2. Convert the SBS experimental signatures into FASTA format.
3. Isolate virtual sense (from sense strand or Watson strand of the DNA) and antisense (or Crick strand of DNA) signatures from annotated genes/ESTs. In this protocol, to identify the expressed genes in creeping bentgrass, isolate virtual signatures from rice annotated genes (http://rice.plantbiology.msu.edu/pseudomolecules/info.shtml). Similarly, to identify the *S. homeocarpa* expressed genes, isolate the virtual signatures from sequenced fungal genomes deposited at the Broad institute (http://www.broad.mit.edu/node/568).
4. Convert all the virtual SBS signatures into FASTA format. Use NCBI local BLAST software version 2.2.17 to match the experimental SBS signatures (derived from the libraries of *S. homoeocarpa* and creeping bentgrass) with the virtual SBS signatures to identify the expressed genes and to calculate the matching rate of the experimental tags. The Local BLAST analysis can be parallelized using OSC's HPC clusters so the time of the analysis is shortened to 1/20–1/50 of the original analysis.
5. Extract output files in FASTA format containing the list of signatures, their corresponding abundances, top five matched target sequences (based on E-value), and also the list of signatures with no hit in the database. Perl scripts can be used to parse these results for import into Microsoft Excel.
6. Identify known transcripts, novel transcripts, antisense transcripts, and alternatively spliced transcripts.

3.8. Digital Northern Analysis of SBS Signatures at DSTI Database

The DSTI database (http://www.dstidb.org) hosts the transcriptome data from both *S. homoeocarpa* and creeping bentgrass (see Note 10). The database provides information about the repertoire of expressed genes and their expression levels in different conditions in both creeping bentgrass and *S. homoeocarpa*. We have deposited the SBS data from: (1) *S. homoeocarpa* grown on PDA plates for 96 h, (2) *S. homoeocarpa* grown on PDB for 96 h, (3) Crenshaw plants infected with *S. homoeocarpa* for 96 h, and (4) noninoculated Crenshaw plants. Using keyword or SBS signature or gene sequence, the database can be searched for the expressed genes in a given library. The abundances show the level

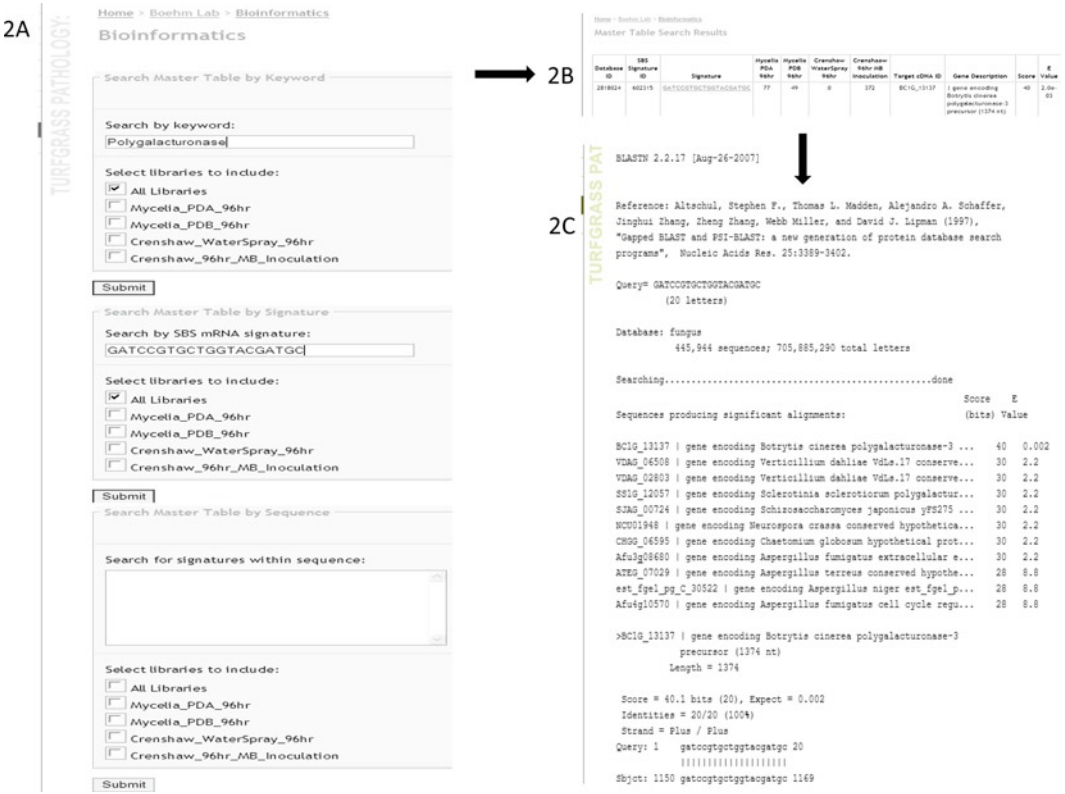

Fig. 2. The Dollar Spot–Turfgrass Interactome (DSTI) database. (2a) Features of two *Sclerotinia homoeocarpa* and two turfgrass SBS libraries are summarized. In each search engine, check the box to select the library and put the keyword, mRNA SBS signature, gene sequence, respectively. (2b) Search result for the selected libraries. This consists of signatures, frequency of signatures, signature hit annotated gene (fungus or plant), score and *e* value. Clicking on the signature sequence directs to next page showing the BLAST output for the selected signature. (2c) The BLAST output page showing the signature alignment on the annotated genes, score and *e* value. Scrolling one's computer mouse down gives the result of all the matched genes in the given database.

of expression in each library. The SBS data are normalized, thus allowing the comparison of signatures across the libraries. Results are displayed on a new page with hyperlinked signature sequences. Each hyperlink provides the BLAST summary result for a given signature (Fig. 2).

4. Notes

1. Our modified SBS protocol is simple and easy to follow. We can construct 4–8 libraries per week. The protocol described here can be used to identify expressed genes from both plants and fungi.

2. The Illumina's sample preparation kit provides reagents for construction of eight libraries. We made 16 libraries by reducing the amount of each reagent to half in each reaction.
3. This method can be used with as little as 1.0 μg total RNA, which is especially advantageous when the amount of RNA is limited.
4. Similar to MPSS and SAGE protocols, SBS also uses a single anchoring enzyme, *Dpn*II (GATC), which may occur in every 256 bp. Therefore, some transcripts in the genome that have no GATC sites might be missed during transcriptome analysis. To overcome this limitation, another library could be made with a different anchoring enzyme, such as *Nla*III.
5. Ligation reactions are carried out for longer periods (4 h) to increase the efficiency of ligation reactions.
6. Restriction digestion with *Mme*I releases the cDNAs from beads. Be careful when transferring the supernatant to another tube.
7. This is helpful when the initial input RNA is low, as more cDNA can be obtained from the increased PCR reactions for DNA sequencing.
8. For efficient elution of DNA from the acrylamide gel debris (fine sliced gel), add up to 500 μl of 1× gel elution buffer and incubate on the thermomixer at 600 rpm at room temperature for 1 h.
9. The bioinformatics pipeline developed can be useful for analyzing the SBS data from any other fungus or plant.
10. We constructed the first genomic database of the dollar spot disease hosting transcriptome data from both creeping bentgrass and *S. homoeocarpa*. The web based search options like keyword, signature search, and gene sequence searches are especially useful for people working in fungal biology and host-pathogen interactions to search for their genes of interest.

Acknowledgments

This work was supported by grants from the NSF Plant Genome Program (#0321437 and #0701745; Dr. Guo-Liang Wang and Dr. Blake Meyers), and state and federal funds appropriated to the Ohio Agricultural Research and Development Center (Dr. Michael Boehm). We thank Mr. Gaurav Jain for his help in extraction of experimental signatures from raw SBS data.

References

1. Couch, H.B. (1995) *Diseases of turfgrasses*, 3rd ed. Krieger publishing co., Malabar, FL.
2. Smiley, R.W., Dernoeden, P.H. and Clarke, B.B. (2005) *Compendium of Turfgrass Diseases*. Third Edition, APS Press.
3. Vargas, J.M. (2005) *Management of turfgrass diseases*. 3rd edition, John Wiley and Sons, Inc., New Jersey.
4. Venu, R. C., Beaulieu, R.A., Graham, T.L., Medina, A.M. and Boehm, M.J. (2009) Dollar spot fungus *Sclerotinia homoeocarpa* produces oxalic acid. *Inter. Turfgrass Soc. Res. J.* **11**, 263–270.
5. Walsh, B., Ikeda, S.D. and Boland, G.J. (1999) Biology and management of dollar spot (*Sclerotinia homoeocarpa*): an important disease of turfgrass. *Hort. Sci.* **34**, 13–21.
6. Gowda, M., Jantasuriyarat, C., Dean R., Wang, G.L. (2004) A robust-longSAGE method for large-scale gene discovery and transcriptome profiling. *Plant Physiol.* **134**, 890–897.
7. Jantasuriyarat, C., Gowda, M., Haller, K., Hatfield, J., Lu, G., Stahlberg, E., et al. (2005) Large-scale identification of ESTs involved in rice and rice blast (*Magnaporthe grisea*) interaction. *Plant Physiol.* **138**, 105–115.
8. Venu, R.C., Jia, Y., Gowda, M., Jia, M.H., Jantasuriyarat, C., Stahlberg, E., et al. (2007) RL-SAGE and microarray analysis of the rice transcriptome after *Rhizoctonia solani* infection. *Mol. Genet. Genomics* **278**, 421–431.
9. Wise, P.R., Moscou, J, M., Bogdanove, J.A., and Whitham, A.S. (2007) Transcript profiling in host-pathogen interactions. *Annu. Rev. Phytopathol.* **45**, 329–369.
10. Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., Bemben, L.A., et al. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature* **437**, 376–380.
11. Bentley, D.R. (2006) Whole-genome resequencing. *Curr. Opin. Genet. Dev.* **16**,545–552.
12. Shendure, J., Porreca, G.J., Reppas, N.B., Lin, X., McCutcheon, J.P., Rosenbaum, A.M., et al. (2005) Accurate multiplex polony sequencing of an evolved bacterial genome. *Science* **309**, 1728–1732.
13. Simon, S.A., Zhai, J., Nandety, R.S., McCormick, Zeng, J., Mejia, D., et al. (2009) Short-read sequencing technologies for transcriptional analyses. *Annu. Rev. Plant Biol.* **60**, 305–333.
14. Bennett, F.T. (1937) Dollar spot disease of turf and its causal organism. *Sclerotinia homoeocarpa* n. sp. *Ann. Appl. Biol.* **24**, 236–257.
15. Meyers, B.C., Tej, S.S., Vu, T.H., Haudenschild, C.D., Agrawal, V., Edberg, S.B., et al. (2004) The use of MPSS for whole-genome transcriptional analysis in Arabidopsis. *Genome Res.* **14**, 1641–1653.
16. Nobuta, K., Venu, R.C., Lu, C., Belo, A., Vemaraju K., Kulkarni K., et al. (2007) An expression atlas of rice mRNA and small RNA. *Nat. Biotechnol.* **25**, 473–477.

Chapter 13

High-Throughput Production of Gene Replacement Mutants in *Neurospora crassa*

Gyungsoon Park, Hildur V. Colot, Patrick D. Collopy, Svetlana Krystofova, Christopher Crew, Carol Ringelberg, Liubov Litvinkova, Lorena Altamirano, Liande Li, Susan Curilla, Wei Wang, Norma Gorrochotegui-Escalante, Jay C. Dunlap, and Katherine A. Borkovich

Abstract

The model filamentous fungus *Neurospora crassa* has been the focus of functional genomics studies for the past several years. A high-throughput gene knockout procedure has been developed and used to generate mutants for more than two-thirds of the ~10,000 annotated *N. crassa* genes. Yeast recombinational cloning was incorporated as an efficient procedure to produce all knockout cassettes. *N. crassa* strains with the Δ*mus-51* or Δ*mus-52* deletion mutations were used as transformation recipients in order to reduce the incidence of ectopic integration and increase homologous recombination of knockout cassettes into the genome. A 96-well format was used for many steps of the procedure, including fungal transformation, isolation of homokaryons, and verification of mutants. In addition, development of software programs for primer design and restriction enzyme selection facilitated the high-throughput aspects of the overall protocol.

Key words: Functional genomics, Gene deletion, High-throughput gene knockout, Large scale mutagenesis, Yeast recombinational cloning

1. Introduction

Genome sequences are now available for an increasing number of filamentous fungi. Recent advances in molecular-genetic approaches have accelerated identification of genes and genetic mechanisms governing pathogenesis and development in these organisms (1, 2). A collection of deletion mutants for all genes in the genome is a very useful resource for functional studies in all

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722, DOI 10.1007/978-1-61779-040-9_13,

organisms, including filamentous fungi. In the yeast *Saccharomyces cerevisiae*, high-throughput production of gene deletion (also known as gene replacement or knockout) mutants was facilitated by the high rate of homologous recombination of DNA constructs (3). However, the low rate of homologous recombination observed in wild-type strains of most filamentous fungi limits the feasibility of such an approach.

Neurospora crassa is a filamentous fungus with ~10,000 predicted genes (4) and serves as a model for species that are important pathogens of plants and animals (5). There are several factors that support efficient production of gene knockouts in *N. crassa*: (1) the transformation efficiency is very high, requiring only a few micrograms of DNA (6), (2) transformation is easily accomplished by electroporation of conidia (7), and (3) the organism has a fast growth rate and is easy to culture (8). However, similar to other filamentous fungi, wild-type *N. crassa* strains have a very low rate of homologous recombination (<10%; (9)). This negative factor has been mitigated by a recent study demonstrating that deletion of *mus-51* or *mus-52*, genes required for nonhomologous end-joining DNA repair, greatly reduces ectopic integration and improves the incidence of homologous recombination (>90%; (10)).

The discovery that mutation of either *mus-51* or *mus-52* could produce a strain with levels of homologous recombination in excess of 90% provided a foundation for a high-throughput gene knockout method in *N. crassa*. In addition, the high recombination rate observed in *S. cerevisiae* was exploited to produce gene knockout cassettes for all *N. crassa* genes (11). Fragments corresponding to 1-kb fragments of DNA 5′ and 3′ to each open reading frame and the selectable marker hygromycin B phosphotransferase (*hph*) were amplified using the polymerase chain reaction, with primers designed to produce fragments with complementary ends. The fragments were then joined in vivo using yeast recombinational cloning (11) to produce the final gene replacement/knockout cassettes, with the 5′ and 3′ regions flanking the selectable marker *hph* (12).

In this chapter, we present a high-throughput procedure for production of gene knockout mutants in *N. crassa*. We first describe the protocol used for generation of *mus* deletion strains. The general method used to produce the Δ*mus* knockout cassettes is presented; for details regarding yeast strains, yeast media, etc., the reader is directed to another volume (20). It should also be noted that long flanking sequences (3 kb) were necessary to achieve homologous recombination of the Δ*mus* constructs in the wild-type background. In the last part of the chapter, we present the method for high-throughput production of knockout mutants for every gene in the genome, using the Δ*mus* mutants as transformation recipients. Again, details regarding construction of the actual knockout cassettes are presented in another volume (20).

2. Materials

Note: All plates are 100 mm in diameter and contain 30 ml agar medium.

2.1. Generation of *N. crassa* Δ*mus* Mutants

1. *S. cerevisiae* strain FY834 (20).
2. YPD medium: in 100 ml, 1 g yeast extract, 2 g peptone, and 2 g dextrose. Autoclave.
3. SD-Ura medium: in 1 L, 6.8 g yeast nitrogen base, 20 g dextrose, and 2 g drop-out mix minus uracil (w/o yeast nitrogen base). Adjust pH to 5.8, add 15 g agar (for plates) and autoclave.
4. *N. crassa* strains FGSC 4200 (wild type, mat a) and FGSC 6103 (*his-3*, mat A).
5. Eppendorf electroporator 2510 (Eppendorf Scientific, Westbury, NY).
6. Electroporation cuvette with 2-mm gap width.
7. QiaexII gel extraction kit (Qiagen, Valencia, CA).
8. Phosphinothricin (PPT) purified as previously described (13).
9. 50× Vogel's minimal medium salts for use with PPT (50× VM-PPT salts; 1 L): 126.8 g $Na_3C_6H_5O_7 \cdot 2H_2O$, 250 g KH_2PO_4, 10 g $MgSO_4 \cdot 7H_2O$, 5 g $CaCl_2 \cdot 2H_2O$, 5 ml biotin solution, 5 ml trace elements solution, and 5 ml chloroform as preservative (8).
10. Biotin solution (filter sterilize): 5 mg biotin/100 ml 50% (v/v) ethanol.
11. Trace elements (filter sterilize): in 100 ml, 5 g $C_6H_8O_7 \cdot H_2O$, 5 g $ZnSO_4 \cdot 7H_2O$, 1 g $Fe(NH_4)_2(SO_4) \cdot 6H_2O$, 0.25 g $CuSO_4 \cdot 5H_2O$, 0.05 g $MnSO_4 \cdot H_2O$, 0.05 g H_3BO_3, and 0.05 g $NaMoO_4 \cdot 2H_2O$.
12. 10× FGS additive (filter sterilize): in 1 L, 5 g fructose, 5 g glucose, and 200 g sorbose (6).
13. VM-PPT medium: 1× VM-PPT Salts, 1.5% sucrose, 0.5% L-proline, and 1% agar (for solid medium) (14). Add PPT to 400 μg/ml after autoclaving.
14. VM-PPT-His medium: 1× VM-PPT salts, 1.5% sucrose, 0.5% L-proline, and 1% agar (for solid medium). Add PPT to 400 μg/ml and L-histidine to 100 μg/ml after autoclaving.
15. FGS-PPT plates: 1× VM-PPT salts, 0.5% L-proline, and 1% agar. After autoclaving, add 10× FGS additive to 1× final concentration and PPT to 400 μg/ml.
16. PPT regeneration agar: 1× VM-PPT Salts, 0.5% L-proline, 1 M sorbitol, and 1% agar. Add 10× FGS additive to 1× final concentration after autoclaving.

17. FGS-PPT-His plates: 1× VM-PPT Salts, 0.5% L-proline, and 1% agar. After autoclaving, add 10× FGS additive to 1× final concentration, PPT to 400 μg/ml, and L-histidine to 100 μg/ml.

2.2. High-Throughput Transformation of N. crassa

1. *N. crassa* strains FGSC 9718 (Δ*mus-51*::*bar*, *mat a*) and FGSC 9719 (Δ*mus-52*::*bar*, *mat a*).
2. 96-well 2 mm electroporation plate (BTX, Holliston, MA).
3. High-throughput electroporation plate handler (BTX).
4. Electro cell manipulator, model ECM 630 (BTX).
5. Sterile 96-deep-well plate (VWR, West Chester, VA).
6. Multichannel pipet (Rainin, Oakland, CA).
7. 15 ml EZ clip polyethylene tubes (Thermo Scientific, Waltham, MA).
8. Hygromycin B (Calbiochem, San Diego, CA).
9. 50× Vogel's salts (50× VM Salts; 1 L): 126.8 g $Na_3C_6H_5O_7 \cdot 2H_2O$, 250 g KH_2PO_4, 100 g NH_4NO_3, 10 g $MgSO_4 \cdot 7H_2O$, 5 g $CaCl_2 \cdot 2H_2O$, 5 ml Biotin solution, 5 ml Trace elements solution, and 5 ml chloroform as preservative (8).
10. Biotin solution (see above).
11. Trace elements (see above).
12. 10× FGS additive (see above).
13. VM: 1× VM salts, 1.5% sucrose, and 1% agar (for agar media) (8).
14. FGS-YE-His plates: 1× VM salts, 2% yeast extract, and 1% agar. After autoclaving, add FGS additive to 1×, L-histidine to 100 μg/ml, and hygromycin to 300 μg/ml.
15. His-YE Regeneration agar: 1× 50× VM Salts, 1 M sorbitol, 2% yeast extract, and 1% agar. Add FGS additive to 1× and L-histidine to 100 μg/ml after autoclaving.
16. Recovery medium: 1× VM salts, 2% yeast extract. Add L-histidine to 100 μg/ml after autoclaving.

2.3. Isolation of Homokaryotic Mutants

1. *N. crassa* strain FGSC 2489 (mat A), *fluffy* (*fl*) mutants FGSC 4317 (*mat A*) and FGSC4347 (*mat a*).
2. Synthetic crossing medium (SCM) agar: in 1 L, 1.0 g KNO_3, 0.7 g K_2HPO_4. 0.5 g KH_2PO_4, 0.5 g $MgSO_4 \cdot 7H_2O$, 0.1 g $CaCl_2$, 0.1 g NaCl, 0.1 ml Biotin solution, 0.1 ml trace elements, 15 g sucrose, and 10 g agar. Autoclave to sterilize.
3. FGS-Hyg plates: 1× VM salts, 1% agar. Add FGS additive to 1× and hygromycin to 200 μg/ml after autoclaving.
4. VM-Hyg agar slants: 1× VM salts, 1% agar. Add hygromycin to 200 μg/ml final concentration after autoclaving.
5. VM-PPT agar (see above) in 12-tube strips: each tube contains 250 μl of VM-PPT agar medium.

2.4. Confirmation of Knockout Mutants by Southern Blot Analysis

1. 12-Well vacuum manifold (Millipore, Billerica, MA).
2. Whatman 2.5 cm grade 1 filter paper circles (Whatman, Piscataway, NJ).
3. 96-Deep-well plates (E&K Scientific, Santa Clara, CA).
4. Caps for 96-deep-well plates (Qiagen, Valencia, CA).
5. 5-mm stainless steel beads (Qiagen).
6. TissueLyser (Qiagen).
7. Puregene DNA extraction kit (Qiagen).
8. Qiagen MagAttract 96 DNA plant core kit (Qiagen).
9. PCR DIG probe synthesis kit (Roche, Basel, Switzerland).
10. DIG wash and block buffer set (Roche).
11. DIG Easy Hyb (Roche).
12. Anti-digoxigenin–AP conjugate (Roche).
13. DIG-labeled DNA molecular weight marker VII (Roche).
14. CDP-star (Roche).

3. Methods

3.1. Generation of *N. crassa* Δ*mus* Mutants

3.1.1. mus Knockout Cassette Construction

1. PCR primers were designed to amplify the *bar* gene (15) and 3-kb fragments corresponding to the 5′ and 3′ regions of *mus-51* and *mus-52*. The 3-kb flanking regions are necessary to achieve a high rate of homologous recombination in a wild-type genetic background. The *bar* gene confers resistance to the chemical phosphinothricin, also known as Ignite or Finale.
2. Amplify the 3-kb 5′ and 3′ flanking regions of *mus-51* or *mus-52* and the *bar* gene using primers and genomic DNA (for *mus-51* and *mus-52* fragments) or plasmid pTJK1 (for *bar*; (16) as template.
3. Digest vector pRS426 with *Eco*RI and *Xho*I (see Note 1).
4. Purify all PCR products and linearized vector using the QiaexII gel extraction kit following the manufacturer's instructions.
5. Inoculate 50 ml of YPD with 0.3 ml of a saturated culture of yeast strain FY834 and incubate overnight at 30°C with shaking.
6. Pellet yeast cells in a 50-ml conical tube by centrifugation at 2,500 rpm for 5 min at room temperature.
7. Discard supernatant and resuspend cells in 0.4 ml 100 mM lithium acetate.
8. Transfer 50 μl of cells to a 1.5 ml microcentrifuge tube and spin down at maximum speed for 30 s at room temperature.

9. Completely remove supernatant and add the following ingredients in order: 240 µl 50% polyethylene glycol 3350, 36 µl 1 M lithium acetate, 50 µl boiled salmon sperm DNA (2 mg/ml), 27 µl sterile water, 1 µl linearized pRS426 (100 ng/µl), and 2 µl each of three PCR products.
10. After vortexing, incubate mixture at 42°C for 30 min.
11. Spin down cells as in step 8 and discard supernatant.
12. Resuspend cells in 1 ml YPD and recover by incubating at 30°C for 1 h.
13. Spin down cells at maximum speed for 30 s and discard supernatant.
14. Resuspend cells in 100 µl YPD and spread on SD-Ura agar plate.
15. After incubating at 30°C for 3–4 days, collect all yeast colonies by scraping the plate with a glass spreader and dipping in 1 ml of sterile water (see Note 2).
16. Spin down cells at maximum speed for 30 s and discard supernatant.
17. Extract yeast DNA using the "Smash and Grab" DNA extraction protocol (17).
18. *mus* knockout cassettes containing the *bar* gene flanked by 3-kb 5′ and 3′ regions of *mus-51* or *mus-52* are amplified using flanking primers with the extracted yeast DNA as template.

3.1.2. Transformation of N. crassa Wild Type with Δmus-51 or Δmus-52 Knockout Cassette DNA

1. Collect conidia of wild-type strain FGSC 4200 (*mat a*) from a 2-week-old culture and suspend in 50 ml of ice-cold sterile water.
2. Centrifuge the tube for 5 min at 2,500 rpm and discard the supernatant.
3. Resuspend conidia in 30 ml of sterile water by vortexing and repeat step 2.
4. Resuspend the conidial pellet in 30 ml of 1 M cold sorbitol.
5. Centrifuge the tube for 5 min at 2,500 rpm and discard the supernatant.
6. Repeat steps 4 and 5.
7. Resuspend the pelleted conidia in 1 M ice-cold sorbitol to give a concentration of 2.5×10^9 conidia/ml.
8. Place 40 µl conidial suspension in a 1.5-ml sterile microfuge tube and add 1 µg *mus* knockout cassette DNA.
9. Transfer the mixture to a 2 mm electroporation cuvette and electropulse using an Eppendorf electroporator 2510 set at

2000 V (or other appropriate electroporator and settings; see below). Add 1 ml 1 M ice-cold sorbitol and transfer mixture to a new tube.

10. Combine the electroporation mixture with 10 ml PPT regeneration agar and plate on FGS-PPT plates.
11. Pick transformant colonies onto VM-PPT slants after incubation at 30°C for 3–4 days.

3.1.3. Verification of mus Knockout Mutants Using Southern Analysis

1. Inoculate 3 ml VM-PPT liquid cultures in 18 × 150 mm glass tubes using hyphae from the transformant slant cultures. Incubate with shaking at 30°C for 24 h. Collect cultures by vacuum filtration and place in 2-ml plastic microcentrifuge tubes.
2. Grind cells in liquid nitrogen using glass rods.
3. Extract genomic DNAs using the Qiagen Puregene DNA isolation kit according to the manufacturer's directions.
4. Digest DNA with an enzyme that will show a different banding pattern for the wild-type and gene replacement *mus* gene. Subject the digests to Southern blot analysis (18, 19), using the entire knockout cassette as a probe.

3.1.4. Isolation of Homokaryotic mus Knockout Mutants

1. We crossed heterokaryotic *mus* deletion mutants as males to strain FGSC 6103 (*his-3*, *mat* A). The *his-3* background was chosen to facilitate future experiments targeting constructs to the *his-3* locus of knockout mutants. However, the heterokaryons could just as easily have been crossed to a wild-type strain of opposite mating type, as was done for all knockout mutants created using the *mus* mutants as recipients (see Section 3.2.2, below).
2. Collect ascospore progeny and suspend in 1 ml of sterile water.
3. Centrifuge tubes at 3,000 rpm for 5 min at room temperature. Discard the supernatant.
4. Resuspend ascospores in 1 ml of sterile water. Withdraw 100 µl of suspension into a 1.5-ml microcentrifuge tube.
5. Incubate the tube at 60°C for 45 min to activate the ascospores.
6. Plate activated ascospores on FGS-PPT-His plates.
7. After 24 h incubation at 30°C in the dark, transfer germinated ascospores to VM-PPT-His agar slants.
8. Isolate genomic DNA from progeny for each gene as described above (section 3.1.3).
9. Perform Southern blot analysis as described above (section 3.1.3) to confirm homokaryotic knockout mutants of

either *mus-51* or *mus-52*. The presence of the *his-3* mutation is determined by spot testing strains on VM, while mating type is determined using *fl* mating type tester strains (8). We selected Δ*mus-51 mat a* strain FGSC9718 and Δ*mus-52 mat a* strain FGSC9719 as the transformation recipients for all knockout cassettes described in Section 3.2, below.

3.2. High-Throughput Gene Replacements in Δmus-51 or Δmus-52 Strains

3.2.1. Transformation of Δmus Strains with Knockout Cassettes

1. The knockout cassette for each target gene consists of the hygromycin phosphotransferase (*hph*) gene flanked by 1 kb 5′ and 3′ regions of the gene (12). The *hph* gene confers resistance to the antibiotic hygromycin (see Note 3).
2. Determine the linkage group of target genes from the genome sequence (http://www.broad.mit.edu/annotation/fungi/neurospora_crassa_7). In cases where the target and *mus* gene are on the same chromosome, the other *mus* mutant strain should be the recipient for transformation (see Note 4).
3. Cultivate the *mus* deletion strain in VM agar flasks for 3 days at 30°C in the dark, followed by 11 more days at 25°C in the light.
4. Collect conidia from the 2-week-old cultures and suspend in 50 ml ice-cold sterile water.
5. Wash the conidia twice using sterile water and then twice with 1 M ice-cold sorbitol, as described in Section 3.1.2, steps 2–6.
6. Resuspend the conidial pellet in 1 M sorbitol to yield a concentration of 2.5×10^9/ml.
7. Transfer the conidial suspension (40 μl/well) into 96-well electroporation plates on ice.
8. Add approximately 1 μg (5–10 μl) of knockout cassette DNA to each well using a multichannel pipet.
9. Electropulse the plate using a BTX ECM 630 electroporator set at 1,500 V, 600 Ω, and 25 μF.
10. Immediately add cold 1 M sorbitol (60 μl) to each well using a multichannel pipet.
11. Transfer the electroporation mixture to 900 μl of chilled 1 M sorbitol in a 96-deep-well plate on ice.
12. Transfer approximately 500–1,000 μl of the electroporation mixture to 1 ml of recovery medium (with 2% yeast extract and 100 μg/ml histidine) in a 96-deep-well plate and incubate at 30°C in the dark with shaking for 2 h (see Note 5).
13. After the recovery step, combine the mixture (1.5–2 ml) with 10 ml of His-YE regeneration agar in a 15-ml EZ clip polyethylene tube and spread on a FGS-YE-His plate.
14. Incubate plates at 30°C in the dark. Transformant colonies begin to appear after 4–7 days.
15. Pick four colonies/gene onto VM-Hyg agar slants.

3.2.2. Generation of Homokaryotic Knockout Mutants with Wild-Type *mus* Alleles

1. Use two heterokaryotic transformants/gene as males to fertilize wild-type *mat A* strain FGSC 2489 cultured on SCM agar slants for 6 days at 25°C in constant light.
2. Incubate sexual crosses for 3 weeks at 25°C in constant light.
3. Collect and activate ascospores as described in Section 3.1.4, steps 2–5.
4. Plate activated ascospores on FGS-Hyg plates and incubate at 30°C for 1–2 days.
5. Pick 12 germinated ascospores from each plate onto VM-Hyg agar slants and incubate for 3 days at 30°C in the dark and then 2 days at 25°C in the light.
6. Check for the presence of the *mus* deletion by inoculating strains in 12-tube strips containing VM-PPT agar. Only those strains that do not grow on this medium (*mus*$^+$) are carried forward.
7. Determine the mating types of strains by crossing strains (male) to both mating types of *fl* tester strains (females; FGSC4317 *mat* A and FGSC4347 *mat* a).

3.2.3.Confirmation of Knockout Mutants by Southern Blot Analysis

1. Select two homokaryotic strains (one of each mating type) for each gene and use to inoculate 3 ml VM liquid cultures in 18 × 150 mm glass tubes.
2. Incubate tubes at 30°C in the dark with shaking for 2 days (longer for slow-growing mutants).
3. Collect tissues using a 12-well vacuum filter system. Transfer cell pads to 2-ml microcentrifuge tubes with screw caps or a 96-deep-well plate and store at −80°C.
4. Add liquid nitrogen to the tubes and pulverize the tissue using a glass rod. Alternatively, cells can be broken using a TissueLyser and 5-mm stainless steel beads.
5. Extract genomic DNA from ground tissues using the Qiagen MagAttract 96 DNA Plant Core Kit, following the manufacturer's recommendations.
6. Digest approximately 5 µg genomic DNA using an appropriate restriction enzyme. A program (http://borkovichlims.ucr.edu/southern/) was developed by John Jones for automatic identification of suitable restriction enzymes. Sequences of the wild-type (FGSC 4200) gene and the corresponding gene replaced with *hph* (as defined by the primers used to create the deletion cassette) were analyzed in silico with 17 selected restriction enzymes. The information was used to generate a list of usable enzymes and to report the sizes of the resulting bands that would hybridize to the probe (if the entire knockout cassette was used; see below) in both the wild-type and knockout strains.

7. Perform Southern blot analysis (18) using the entire knockout cassette or the *hph* gene as a probe. The DIG-labeled DNA molecular weight marker VII is run with samples on the agarose gel. The PCR DIG probe synthesis kit is used for probe amplification and labeling, while the DIG Easy Hyb, DIG wash and block buffer set, Anti-digoxigenin-AP conjugate, and CDP-Star are used for hybridization and detection.

4. Notes

1. Complete digestion is a critical step. Gel check on the digestion status is recommended.
2. Normally, each plate will have 50–100 colonies. If too many colonies are formed on each plate, it is likely that the vector pRS426 is not digested completely.
3. We recommend to use about 1 kb upstream and 1 kb downstream flanking sequences. Shorter flanking sequences may lead to lower efficiency of gene replacement.
4. This will allow generation of knockout mutant progeny with wild-type copies of the *mus* alleles later in the protocol (section 2B).
5. Addition of 2% yeast extract and 100 μg/ml histidine to recovery medium has previously been demonstrated to allow growth of various auxotrophic mutants (G.E. Turner, unpublished).

Acknowledgments

This work was supported by National Institute of General Medical Sciences grant P01 GM068087. We thank John Jones for software design and implementation of the LIMS and Gloria Turner for helpful comments on the manuscript.

References

1. Shafran, H., Miyara, I., Eshed, R., Prusky, D., and Sherman, A. (2008) Development of new tools for studying gene function in fungi based on the Gateway system, *Fungal Genet Biol* **45**, 1147–54.
2. Dunlap, J. C., Borkovich, K. A., Henn, M. R., Turner, G. E., Sachs, M. S., Glass, N. L., McCluskey, K., Plamann, M., Galagan, J. E., Birren, B. W., Weiss, R. L., Townsend, J. P., Loros, J. J., Nelson, M. A., Lambreghts, R., Colot, H. V., Park, G., Collopy, P., Ringelberg, C., Crew, C., Litvinkova, L., DeCaprio, D., Hood, H. M., Curilla, S., Shi, M., Crawford, M., Koerhsen, M., Montgomery, P., Larson, L., Pearson, M., Kasuga, T., Tian, C., Basturkmen, M., Altamirano, L., and Xu, J. (2007) Enabling a community to dissect an organism: overview of the *Neurospora* functional genomics project. *Adv Genet* **57**, 49–96.

3. Winzeler, E. A., Shoemaker, D. D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J. D., Bussey, H., Chu, A. M., Connelly, C., Davis, K., Dietrich, F., Dow, S. W., El Bakkoury, M., Foury, F., Friend, S. H., Gentalen, E., Giaever, G., Hegemann, J. H., Jones, T., Laub, M., Liao, H., Liebundguth, N., Lockhart, D. J., Lucau-Danila, A., Lussier, M., M'Rabet, N., Menard, P., Mittmann, M., Pai, C., Rebischung, C., Revuelta, J. L., Riles, L., Roberts, C. J., Ross-MacDonald, P., Scherens, B., Snyder, M., Sookhai-Mahadeo, S., Storms, R. K., Veronneau, S., Voet, M., Volckaert, G., Ward, T. R., Wysocki, R., Yen, G. S., Yu, K., Zimmermann, K., Philippsen, P., Johnston, M., and Davis, R. W. (1999) Functional characterization of the S. cerevisiae genome by gene deletion and parallel analysis. *Science* **285**, 901–6.
4. Galagan, J. E., Calvo, S. E., Borkovich, K. A., Selker, E. U., Read, N. D., Jaffe, D., FitzHugh, W., Ma, L. J., Smirnov, S., Purcell, S., Rehman, B., Elkins, T., Engels, R., Wang, S., Nielsen, C. B., Butler, J., Endrizzi, M., Qui, D., Ianakiev, P., Bell-Pedersen, D., Nelson, M. A., Werner-Washburne, M., Selitrennikoff, C. P., Kinsey, J. A., Braun, E. L., Zelter, A., Schulte, U., Kothe, G. O., Jedd, G., Mewes, W., Staben, C., Marcotte, E., Greenberg, D., Roy, A., Foley, K., Naylor, J., Stange-Thomann, N., Barrett, R., Gnerre, S., Kamal, M., Kamvysselis, M., Mauceli, E., Bielke, C., Rudd, S., Frishman, D., Krystofova, S., Rasmussen, C., Metzenberg, R. L., Perkins, D. D., Kroken, S., Cogoni, C., Macino, G., Catcheside, D., Li, W., Pratt, R. J., Osmani, S. A., DeSouza, C. P., Glass, L., Orbach, M. J., Berglund, J. A., Voelker, R., Yarden, O., Plamann, M., Seiler, S., Dunlap, J., Radford, A., Aramayo, R., Natvig, D. O., Alex, L. A., Mannhaupt, G., Ebbole, D. J., Freitag, M., Paulsen, I., Sachs, M. S., Lander, E. S., Nusbaum, C., and Birren, B. (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422**, 859–68.
5. Davis, R. H., and Perkins, D. D. (2002) Neurospora: A model of model microbes. *Nat Rev Genet* **3**, 397–403.
6. Case, M. E., Schweizer, M., Kushner, S. R., and Giles, N. H. (1979) Efficient transformation of *Neurospora crassa* by utilizing hybrid plasmid DNA. *Proc. Natl. Acad. Sci. USA* **76**, 5259–63.
7. Chakraborty, B. N., Patterson, N. A., and Kapoor, M. (1991) An electroporation-based system for high-efficiency transformation of germinated conidia of filamentous fungi. *Can J Microbiol* **37**, 858–63.
8. Davis, R. H., and deSerres, F. J. (1970) Genetic and microbiological research techniques for *Neurospora crassa*. *Methods Enzymol.* **71A**, 79–143.
9. Paietta, J. V., and Marzluf, G. A. (1985) Gene disruption by transformation in *Neurospora crassa*. *Mol Cell Biol* **5**, 1554–9.
10. Ninomiya, Y., Suzuki, K., Ishii, C., and Inoue, H. (2004) Highly efficient gene replacements in Neurospora strains deficient for nonhomologous end-joining. *Proc Natl Acad Sci USA* **101**, 12248–53.
11. Oldenburg, K. R., Vo, K. T., Michaelis, S., and Paddon, C. (1997) Recombination-mediated PCR- directed plasmid construction in vivo in yeast. *Nucleic Acids Res* **25**, 451–2.
12. Colot, H. V., Park, G., Turner, G. E., Ringelberg, C., Crew, C. M., Litvinkova, L., Weiss, R. L., Borkovich, K. A., and Dunlap, J. C. (2006) A high-throughput gene knockout procedure for Neurospora reveals functions for multiple transcription factors. *Proc Natl Acad Sci USA* **103**, 10352–7.
13. Hays, S., and Selker, E. (2000) Making the selectable marker *bar* tighter and more economical. *Fungal Genet. Newsl.* **47**, 107.
14. Pall, M. L. (1993) The use of Ignite (Basta; glufosinate;phosphinothricin) to select transformants of bar-containing plasmids in *Neurospora crassa*. *Fungal Genet. Newsl.* **40**, 58.
15. Avalos, J., Geever, R. F., and Case, M. E. (1989) Bialaphos resistance as a dominant selectable marker in *Neurospora crassa*. *Curr Genet* **16**, 369–72.
16. Jones, C. A., Greer-Phillips, S. E., and Borkovich, K. A. (2007) The response regulator RRG-1 functions upstream of a MAPK pathway impacting asexual development, female fertility, osmotic stress and fungicide resistance in *Neurospora crassa*. *Mol Biol Cell* **18**, 2123–36.
17. Hoffman, C. S. (2001) Preparation of yeast DNA. *Curr Protoc Mol Biol* Chapter 13, Unit13 11.
18. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) Molecular Cloning: A Laboratory Manual., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.
19. Ivey, F. D., Hodge, P. N., Turner, G. E., and Borkovich, K. A. (1996) The G alpha i homologue gna-1 controls multiple differentiation pathways in *Neurospora crassa*. *Mol. Biol. Cell* **7**, 1283–97.
20. Collopy, P. D., Colot, H. V., Park, G., Ringelberg, C., Crew, C. M., Borkovich, K. A. and Dunlap, J. C. (2010) High-throughput construction of gene deletion cassettes for generation of *Neurospora crassa* knockout strains. *Methods in Mol. Biol.* **638**, 33–40.

Chapter 14

Phenotypic Analysis of *Neurospora crassa* Gene Deletion Strains

Gloria E. Turner

Abstract

Phenotypic analysis of *Neurospora crassa* knockout (KO) mutants was used as a vehicle to introduce students to laboratory research. The availability of gene deletion strains was the impetus for the development of a program designed to introduce beginning science students to basic microbiology, genetics, microscopy and beginning bioinformatics. The goal was to provide a research experience, acquire laboratory skills and phenotype hundreds of KO mutants. The data provided by the students was used to build a phenotype database at the Broad Institute at Harvard/MIT for the fungal scientific community. Each mutant analysis consists of five assays that examine growth and morphology, asexual and sexual development using wild-type (parental) strains as a reference. This information indicates how loss of each gene impacts these basic and important processes.

Key words: *Neurospora crassa*, KO Phenotypic analysis, Aerial hyphae, Conidia, Protoperithecia, Perithecia, Ascospores

1. Introduction

Assigning a function to ten thousand *Neurospora crassa* genes, half of them novel, is a daunting task. The approach taken was a high throughput gene deletion schema that produced thousands of strains containing a single gene deletion (1). A method to systematically examine each of these strains for phenotypes was developed for undergraduate students. The assays allow us to determine if growth of developmental processes are impaired in knockout (KO) mutants. This information helps establish functions for the missing gene product in each mutant. The high throughput analysis was done in quadruplicate with data uploaded to the Broad Institute by student researchers. This information was curated, summarized, and published on the Broad Neurospora

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722,
DOI 10.1007/978-1-61779-040-9_14,

Web site. The data can be accessed by searching any NCU # and checking its KO 1 allele status. If the mutant has been phenotyped, a KO 1 will appear when a Feature Search is performed. A summary list can be found under Phenotypic Summaries (2).

Plate growth was examined on two types of media, minimal, and minimal supplemented with yeast extract and at two different temperatures (25°C and 37°C). Comparison of phenotypes under these conditions reveals mutants that are influenced by temperature and the presence of metabolites. Plate images were captured using a camera at 24 and 48 h of growth in order to document the presence or absence of gross morphological differences in each mutant. A microscope outfitted with a digital camera was used to capture images of the growing colony edge. Growth rates were measured on VM (minimal) medium at 25°C using race tubes (3). Aluminum racks were designed to hold ten race tubes. The racks fit on cafeteria trays and facilitated the sterilization, inoculation, incubation, and measurement processes.

Asexual development produces macro- and microconidia. Our assays address only macroconidia production which is monitored on VM agar slants at 25°C. Standing liquid VM and VM + yeast extract supplemented media were used to measure extension of aerial hyphae, a step in macroconidiation. Production of macroconidia is regulated by many factors and numerous *N. crassa* mutants have been isolated that are blocked at various stages. However, in most cases the missing or mutated gene products in these mutants are unknown (4). We do know that differentiation of macroconidia is an environmental stress response (5, 6) and is regulated in part by light-sensing and heterotrimeric G-protein signaling pathways (7–9).

The final assay examined the complex sexual cycle in *N. crassa* (10, 11). The analysis determined whether the gene deletion mutant was blocked in any of the major steps of this pathway, including production of protoperithecia (female sexual structures), development of perithecia after fertilization, and the ejection of ascopores (products of meiosis). Since matings are performed between opposite mating types, the mating type (*mat*) locus of the knockout strain determined which wild-type strain was used as the male. Mating types for all homokaryotic knockout progeny were determined at either UC Riverside or Dartmouth (12). UCLA received only *mat a* progeny except in instances, where only a *mat A* mutant was recovered from the knockout project. This occurred in less than 5% of the strains phenotyped. All observations were recorded on data entry forms by students and uploaded to the Broad Institute. All images were captured on laboratory computers sent via FTP to the Broad server for review and selection.

2. Materials

2.1. Growth and Morphology

1. Hygromycin B (Invitrogen Cat. No.: 10687-010).
2. Vogel's Minimal media (VM) is 1× Vogel's salts and 1.5% sucrose (13).
3. VM Supplemented media is VM + 2% Difco Yeast extract.
4. Vogel's 50× Minimal Medium Salts (1 L): 117.5 g $Na_3C_6H_5O_7 \cdot 2H_2O$, 250 g KH_2PO_4, 100 g NH_4NO_3, 10 g $MgSO_4 \cdot 7H_2O$, 5 g $CaCl_2 \cdot 2H_2O$, 5 mL Biotin solution, 5 mL Trace elements solution, 5 mL chloroform as preservative.
5. Biotin solution (filter-sterilize): 5 mg biotin/100 mL 50% (v/v) ethanol.
6. Trace elements (filter sterilize): in 100 mL, 5 g $C_6H_8O_7 \cdot H_2O$, 5 g $ZnSO_4 \cdot 7H_2O$, 1 g Fe $(NH_4)_2(SO_4) \cdot 6H_2O$, 0.25 g $CuSO_4 \cdot 5H_2O$, 0.05 g $MnSO_4 \cdot H_2O$, 0.05 g H_3BO_3, 0.05 g $NaMoO_4 \cdot 2H_2O$.
7. VM hyg: VM containing 200 μg/mL hygromycin B.
8. Avery Easy Peel Clear labels (#5667) were used for labeling all gene deletion strain tubes and assay tubes and plates.
9. Glass race tubes measuring 16.5 in. overall length with a 16 mm OD, 1.2 mm wall. The tube ends were bent at a 45° angle 2.6 in. from each end (Chemglass product #GC-4020). Race tubes were capped with size 16 disposable KimKap* closures (Kimax*).
10. Aluminum metal race tube racks.
11. Alliance pale crepe gold rubber bands.
12. S8 ApoStereo Zoom microscope mounted with a DFC 280 digital camera (Leica, Wetzlar, Germany).
13. Infinity camera with a Navitar Zoom 7000 lens (Lumenera Scientific, Ottawa, Canada).
14. Disposable 100 × 15 mm Petri plates (cat. 08-757-13).
15. Inoculating needles (LeStab) manufactured for Decon Labs, Inc. used for all inoculations.
16. Fisherbrand Fine point Marker pens.

2.2. Asexual Development

1. VM agar slants (3 mL) in 13 × 100 mm tubes (see Note 1).
2. VM and VM supplemented liquid media.
3. Liquid and agar tubes were capped with size 13 disposable KimKap* closures (Kimax*).
4. Inoculating needles (LeStab) manufactured for Decon Labs, Inc.
5. Fisherbrand Fine point Marker pens.

2.3. Sexual Development

1. Synthetic Crossing Medium (SCM) is a nitrogen poor medium that induces the sexual cycle (13). SCM: 1× Westergaard's salts, 1.0 g KNO_3, 0.7 g K_2HPO_4, 0.5 g KH_2PO_4, 0.5 g $MgSO_4 \cdot 7H_2O$, 0.1 g $CaCl_2$, 0.1 g NaCl, 0.1 mL Biotin solution, 0.1 mL Trace elements. 1% sucrose was used as the carbon source and 1.5% agar as a solidifying agent.
2. Disposable 18 × 150 mm tubes with 18 mm Bacti Capall closures. Each cross tube was filled with 7 mL of SCM agar medium. After sterilization, the tubes were slanted and left to solidify.
3. Inoculating needles (LeStab) manufactured for Decon Labs, Inc.
4. Conidia used as male were from parental reference strains FCSC 4200 (*mat a*) or FGSC 2489 (*mat A*). A conidial suspension of 7–10 day old culture was diluted 1:100 from a stock concentration of 1×10^8 conidia/mL.
5. P-1000 pipetman and sterile pipet tips used for dispensing conidial suspension to cross tubes.
6. Thermolyne Vortex shakers.
7. StereoZoom 7 microscope (Bausch & Lomb/Leica, Wetzlar, Germany).
8. S8 Apo StereoZoom microscope (Leica, Wetzlar, Germany).

3. Methods

Gene deletion strains were received from the University of California at Riverside (K. A. Borkovich/G. Park) and Dartmouth Medical School (J. C. Dunlap/ H. V. Colot) (see Note 2). All strains arrived on VM hyg in 13 × 100 mm slants and were stored at 5°C. Duplicate sets were made from the original stocks and maintained on VM hyg slants.

The Broad Institute has assigned every predicted open reading frame an NCU # (*N. Crassa U*nknown). All stocks were identified by NCU #'s. At the outset of the analysis, a sheet of 80 clear labels was generated for each NCU #. These labels were shared by four students phenotyping the same mutant and were used for all procedures. The phenotype assays were performed in the absence of hygromycin B. All *N. crassa* transfers were performed in biosafety hoods using inoculating needles sterilized by gas flame. Each student generated his or her own set of KO mutant VM hyg slants from the duplicated stocks. Students began phenotyping by performing all five assays with wild-type *N. crassa*. This introduction familiarized the students with the organism and the procedures. Wild-type was subsequently used as a control for all assays during the phenotypic analysis.

3.1. Growth and Morphology

1. Two VM and two VM Supplemented agar plates (100 × 15 mm) are labeled (agar side) and inoculated in the center with a KO mutant. A set of VM and VM Supplemented plates is incubated at both 25°C and 37°C for 48 h. Plate images for all four conditions are captured using an Infinity camera at both 24 and 48 h allowing documentation of growth and morphology. The hyphae at the edge of the colony are imaged at 4 × 20× magnification using an S8 Apo Stereo Zoom microscope mounted with a DFC 280 digital camera (Leica). Because the *N. crassa* control strains extend hyphae quickly and reach the edge of the plate before 48 h, 48 h edge photos are not available for mutant strains that grow as fast as the wild-type. The microscopy is performed for all plate growth conditions. Plate morphology is scored relative to the parental strains at 48 h for growth and pigmentation. All images are captured on laptop computers and saved in individual student folders. Student folders are organized by media (VM and VM sup), temperature (25°C and 37°C), and time (24 and 48 h) resulting in eight subfolders for each student folder. As images are captured, they are placed in the appropriate folder.
2. Growth rates of basal hyphae in one dimension are measured in glass race tubes (see Note 3) on VM agar medium. This assay takes 5 days to complete. Aluminum racks were designed and made to hold ten race tubes. The tubes are attached with quality rubber bands. The racks were placed on cafeteria trays allowing for convenient handling and incubation in 25°C walk-in constant temperature rooms. Each student inoculates a rack of ten tubes per week. Race tubes were labeled using NCU numbers and racks with student names. The tubes were inoculated at one end in the afternoon, and incubated overnight at 25°C. The growth front was marked with a marker pen early the next day ($t = 0$). The overnight incubation is considered the lag phase in this experiment. The race tubes are then returned to the 25°C warm room. Late in the day, the new growth front is marked. Growth fronts were marked twice per day, morning and afternoon, over a 72 h period. The distances between each front are measured and recorded on a data entry sheet. Using time and distance as coordinates, the data are graphed to verify linearity and only data with an R^2 value of 0.96 or better is accepted. The growth rate is expressed as mm/day.

3.2. Asexual Development

1. Asexual sporulation is assayed in slant tubes containing VM agar. Tubes are labeled and inoculated with a KO strain. The tubes are incubated at 25°C for 3 days and then put at room temperature (22–25°C) for 3–5 days. Conidiation, pigmentation, and aerial hyphae are scored using the parental strains as a reference. Data are recorded on data entry forms and uploaded to the student database at the Broad Institute.

2. Aerial hyphal extension is measured in VM and VM Supplemented standing liquid cultures. Test tubes (13 × 100 mm) containing 2 mL of liquid medium are inoculated using conidia from student mutant stocks (see Note 4). The racks of tubes are incubated statically at 25°C for 24 h with ambient light/dark cycle. A mycelial mat is formed at the top of the liquid media after 24 h. The first measurement is made at this time. Using a marker pen, a line is drawn on the outside of the tube to mark the top of the mycelial mat. In instances where no mat was formed, the tubes are marked at the top of the medium meniscus. This mark is taken as $t = 0$ for measuring aerial hyphal extension. If aerial hyphae have extended at this 24 h interval, a mark is made to indicate the height of extension. The difference in the two marks is the 24 h extension. The cultures are incubated statically for an additional 48 h and the total height is recorded in mm. It is important to mark the same section of the tube.

3.3. Female Sexual Development

1. SCM slants are labeled and inoculated with conidia from student KO strain stocks.
2. The slants are incubated in ambient light/dark conditions at room temperature (22–25°C) for 7 days. Formation of protoperithecia is observed at this time using a stereomicroscope (Bausch and Lomb or Leica).
3. Conidial suspensions are prepared from flasks of *N. crassa* grown for 7–10 days. Conidia were harvested by filtration and centrifugation and washed with sterile water (13).
4. SCM female cultures are fertilized using 400 µL of a 1:100 dilution of a 1×10^8 conidial stock suspension from FGSC 2489 (*mat A*) or FGSC 4200 (*mat a*). It is important to vortex the conidial suspension before removing an aliquot. Carefully pipet the suspension over the SCM agar surface, where protoperithecia have formed. Slant the tubes so that the agar surface is covered by the conidial suspension. Continue incubation of slanted tubes as in step 2.
5. The following day tubes are examined using a stereomicroscope (Bausch and Lomb or Leica) and scored for the formation of protoperithecia relative to parental controls. The data are recorded on data entry forms.
6. SCM tubes are incubated for an additional week as in step 2.
7. Perithecial formation is evaluated at this 2-week interval with a stereomicroscope using the wild-type cross as a reference. Observations are recorded and tubes are left for one remaining week at room temperature.
8. At the end of 3 weeks, tubes are checked for ascospore development and ejection. Ascospores are ejected with such force

they end up on the wall of the test tube opposite the agar surface. The wild-type cross is used as a reference for ascospore abundance and pigmentation (see Note 5).

4. Notes

1. 1.5% Difco Bacto agar was used for all agar-containing media.
2. See Chapter 13 authored by Park and colleagues for details.
3. For other fungi grow slower than *N. crassa*, growth rate can be measured with Petri plate cultures.
4. Once inoculated it is important not to tilt the tubes.
5. The wild-type ascospores of *N. crassa* are darkly pigmented. Ejected ascospores sticking to the side of the test tubes can be easily observed.

Acknowledgments

National Institute of General Medical Sciences grant P01 GM068087. Matthew Crawford for support and assistance with data organization at the Broad Institute and Katherine A. Borkovich for helpful comments on the manuscript.

References

1. Colot, H. V., Park, G., Turner, G. E., Ringelberg, C., Crew, C. M., Litvinkova, L., Weiss, R. L., Borkovich, K. A., and Dunlap, J. C. (2006) A high-thoughput gene knockout procedure for *Neurospora* reveals functions for multiple transcription factors, *Proc Natl Acad Sci USA* **103**, 10352–7.
2. Dunlap, J. C., Borkovich, K. A., Henn, M. R., Turner, G. E., Sachs, M. S., Glass, N. L., McCluskey, K., Plamann, M., Galagan, J. E., Birren, B. W., Weiss, R. L., Townsend, J. P., Loros, J. J., Nelson, M. A., Lambreghts, R., Colot, H. V., Park, G., Collopy, P., Ringelberg, C., Crew, C., Litvinkova, L., DeCaprio, D., Hood, H. M., Curilla, S., Shi, M., Crawford, M., Koerhsen, M., Montgomery, P., Larson, L., Pearson, M., Kasuga, T., Tian, C., Basturkmen, M., Altamirano, L., and Xu, J. (2007) Enabling a community to dissect an organism: overview of the *Neurospora* functional genomics project, *Adv Genet* **57**, 49–96.
3. Ryan, F. J., Beadle, G.W., and Tatum, E. L. (1943) The tube method of measuring the growth rate of *Neurospora*, *Amer. J. Bot.* **30,** 784–799.
4. Springer, M. L. and Yanofsky, C. (1989) A morphological and genetic analysis of conidiophore development in *Neurospora crassa*, *Genes and Devel.* **3**, 559–571.
5. Toledo, I., Aguirre, J., and Hansberg, W. (1994) Enzyme inactivation related to a hyperoxidant state during conidiation of *Neurospora crassa, Microbiology* **140**, 2391–2397.
6. Michán, S., Lledías, F., and Hansberg, W. (2003) Asexual development is increased in *Neurospora crassa cat-3*-null mutant strains, *Eukary. Cell* **2.4**, 798–808.
7. Ivey, F. D., Hodge, P. N., Turner, G. E., and Borkovich, K. A. (1996) The G alpha i homologue gna-1 controls multiple differentiation

pathways in Neurospora crassa, *Mol. Biol. Cell* **7**, 1283–97.

8. Yang, Q., Borkovich, K. A. (1999) Mutational activation of a Gai causes uncontrolled proliferation of aerial hyphae and increased sensitivity to heat and oxidative stress in *Neurospora crassa, Genetics* **151**, 107–117
9. Kays, A. M., Rowely, P. S., Baasiri, R. A., and Borkovich, K. A. (2000) Regulation of conidiation and adenylyl cyclase levels by the Ga protein GNA-3 in *Neurospora crassa*, *Mol Cell Biol* **20**, 7693–7705.
10. Westergaard, M., and Hirsch, H. M. (1954) Environmental and genetic control of differentiation in *Neurospora*, Proc. Symp. Colson Res. Soc. **7**, 171–183.
11. Springer, M. L. (1993) Genetic control of fungal differentiation: The thee sporulation pathways of *Neurospora crassa*, *BioEssays,* **15**, 365–374.
12. Park, G., Colot, H. V., Collopy, P. D., Krystofova, Crew, C., Ringelberg, C. Litvinkova, L., Altamirano, L., Li, L., Dunlap, J. C., and Borkovich, K. A. In press
13. Davis, R. H., and deSerres, F. J. (1970) Genetic and microbiological research techniques for *Neurospora crassa, Methods Enzymol.* **71A**, 79–143.

Chapter 15

Efficient Approaches for Generating GFP Fusion and Epitope-Tagging Constructs in Filamentous Fungi

Xiaoying Zhou, Guotian Li, and Jin-Rong Xu

Abstract

For functional characterization of predicted genes encoding hypothetical proteins in fungal genomes, it is complementary to genetic studies to determine their expression and subcellular localization patterns in different developmental or infection stages. It is also important to identify and characterize other proteins that are physically associated with or functionally related to these genes in vivo by co-immunoprecipitation or affinity purification analyses. In this chapter, we described a set of yeast shuttle vectors and protocols to generate fusion constructs by the yeast gap repair approach. Because of the simplicity and efficiency of yeast gap repair, these vectors and the general methods described in this chapter are suitable for functional genomics studies in filamentous fungi.

Key words: Fusion protein, Subcellular localization, Affinity purification, Gap repair

1. Introduction

Advances in sequencing technology have significantly reduced the cost of genome sequencing. To date, over 50 filamentous fungi have been sequenced. In general, about one-third of the genes in most of the sequenced fungal genomes have no distinct homologs in GenBank or have no known biological functions (1). Some of them are unique to certain fungal groups or specific species and have no recognizable protein domain or motif. Targeted gene deletion and disruption are two commonly used approaches to determine the function of these hypothetical genes (2, 3). One approach complementary to genetic studies with these genes encoding hypothetical proteins is to assay their expression and subcellular localization patterns in different fungal

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722,
DOI 10.1007/978-1-61779-040-9_15,

developmental or infection stages. Genes specifically expressed during certain stages or localized to particular organelles are likely to be involved in related biological functions or processes.

Like in many other organisms, GFP has been widely used as a molecular marker to determine the expression and localization of targeted genes in filamentous fungi. However, constructing GFP fusions by conventional restriction enzyme digestion and ligation with vectors suitable for fungal transformation is sometimes difficult or time consuming. The yeast gap repair approach is a recombination-based cloning method (4, 5), which takes advantage of the high efficiency of homologous recombination and extremely low transformation frequency with linearized plasmid DNA in yeast. Because there is no need for suitable restriction enzyme sites or ligation, gap repair is efficient and convenient for generating GFP constructs (6, 7).

For functional genomics studies, it is also important to identify and characterize other genes that are physically associated with or functionally related to the genes of interest in filamentous fungi. One common approach to determine protein–protein interactions in vivo is the co-immunoprecipitation assay. For this purpose, two genes of interest should be fused with different epitope tags, such as 13×Myc and 3×FLAg, that can be detected with antibodies. The other approach is to fuse the gene of interest with epitope tags and use the affinity purification and mass spectrometry analyses to identify its interacting proteins in fungal cells (8). For both co-immunoprecipitation and affinity purification assays, it is necessary to generate transformants expressing the fusion constructs with epitope tags. In previous studies, we have developed a set of yeast shuttle vectors that carry the hygromycin-, bleomycin-, or geneticin-resistance genes. These vectors are suitable for generating fusion constructs with GFP, 13×Myc, or 3×FLAG epitope tags with the gap repair approach. They can be directly used to transform *Magnaporthe oryzae*, *Fusarium graminearum*, and other filamentous fungi to generate transformants expressing various GFP fusion or epitope-tagging constructs (7, 9, 10). If necessary, the strong constitutive RP27 promoter (6) carried by these vectors can be used to overexpress the resulting fusion constructs.

Below is an example for generating the *MCM1*-GFP construct by homologous recombination and transformants of *M. oryzae* expressing the fusion proteins. Because of the simplicity and efficiency of yeast gap repair, the general methods and vectors described in this chapter are suitable for large scale studies of expression, localization, and affinity purification in *M. oryzae* and other fungi.

2. Materials

2.1. Amplification of Target Genes

1. Advantage 2 DNA polymerase mix: Clontech, Mountain View, CA.
2. TE: 10 mM Tris–HCl (pH 8.0), 0.1 mM EDTA. Autoclave for 20 min and store at room temperature (RT).
3. 50× dNTP mixture: 10 mM each for dATP, dGTP, dCTP, and dTTP. Store at –20°C.
4. PCR primers: all primers are dissolved in sterile distilled water to 50 μM.
5. 1× TAE buffer (1 L): 20 mL 50× TAE buffer and 980 mL distilled water.
6. 50× TAE buffer (1 L): 242 g Tris Base (Sigma-Aldrich, St Louis, MO), 57.1 mL glacial acetic acid, 100 mL 0.5 M EDTA (pH 8.0).

2.2. Amplification of Target Genes and Vector Treatment

1. QIAquick gel extraction kit: Qiagen, Valencia, CA.
2. *Xho*I restriction enzyme: 10 U/μl (Invitrogen, Carlsbad, CA).
3. 1× Reaction buffer 2 (Invitrogen) for *Xho*I digestion.

2.3. Preparation of Yeast Competent Cells

1. *Saccharomyces cerevisiae* strain XK1-25: *MATa trp1* (7).
2. YEPD liquid medium (1 L): 20 g glucose, 20 g peptone, 10 g yeast extract. Autoclave for 20 min and store at RT.
3. YEPD agar: add 15 g agar to 1 L YEPD liquid medium. Autoclave for 20 min. Pour plates.
4. TE (pH 7.5): from the Alkali-cation yeast transformation kit (BIO 101, La Jolla, CA).
5. Lithium/cesium acetate solution: from the Alkali-cation yeast transformation kit (BIO 101).
6. Dimethyl sulfoxide (DMSO).

2.4. Yeast Transformation

1. 10 mg/mL carrier DNA: from the Alkali-cation yeast transformation kit (BIO 101).
2. Histamine solution: from the Alkali-cation yeast transformation kit (BIO 101).
3. 50% PEG solution: from the Alkali-cation yeast transformation kit (BIO 101).
4. TE/Cation MIXX: from the Alkali-cation yeast transformation kit (BIO 101).
5. SOS solution: from the Alkali-cation yeast transformation kit (BIO 101) (11).

6. SD-Trp agar: dissolve 182.2 g sorbitol, 6.7 g yeast nitrogen base without amino acids (Difco Laboratories, Detroit, MI), and 0.74 g – Trp DO supplement (Clontech, Mountain View, CA) in water to the final volume of 960 mL. Add 15 g agar and autoclave for 20 min. When cooled down to 55°C, add 40 mL 50% glucose (sterilized by filtration). Pour plates.

2.5. Identification of Fusion Constructs

1. SCE (1 L): 1 M sorbitol, 50 mM EDTA, 100 mM sodium citrate. Adjust to pH 7.0. Autoclave for 20 min.
2. Lytic enzyme solution: dissolve approximately 50 μg lytic enzyme (Sigma-Aldrich, St Louis, MO) in 0.5 mL SCE buffer (~100 μg/mL).
3. Spheroplast lysis buffer: 50 mM Tris–HCl (pH 7.5), 25 mM EDTA, 0.5 M NaCl, 1% SDS, 0.1% β-2-mercaptoethanol.
4. 25:24:1 phenol/chloroform/isoamyl alcohol: USB, Cleveland, OH.
5. 100% and 70% ethanol (EtOH): store at –20°C.
6. Competent cells of *Escherichia coli* strain DH10B: Invitrogen.
7. LB (1 L): 10 g tryptone, 5 g yeast extract, 5 g NaCl, 18 g agar. Autoclave for 20 min.
8. 1,000× ampencillin: 50 mg/mL ampencillin in 70% EtOH. Store in –20°C.
9. LB + Amp plates: Melt LB agar medium. Cool down to 55°C. Add 1,000× ampencillin to the final concentration of 50 μg/mL before pouring into Petri plates.
10. Oatmeal agar plates (1 L): Heat 100 g Quaker oatmeal in 1 L of distilled water for 1 h at 70°C. Strain with a metal strainer. Add water to the filtrate to bring up the volume to 1 L. Add 14 g of agar. Autoclave for 40 min. Pour plates.

2.6. Generation of Fungal Transformation Expressing the Fusion Constructs

1. Miracloth: EMD Biosciences, Inc., La Jolla, CA.
2. 5× YEG liquid media (l L): 5 g yeast extract and 10 g glucose. Autoclave for 20 min.
3. 1 M sorbitol: dissolve 182.2 g sorbitol in 1 L of distilled water. Autoclave for 20 min.
4. Protoplasting solution: dissolve 5 mg/mL lysing enzyme (Sigma-Aldrich) and 1 mg/mL beta-glucanase (Sigma-Aldrich) in sterile distilled water.
5. STC: 20% sucrose, 50 mM Tris–HCl (pH 8.0), 50 mM $CaCl_2$. Autoclave for 20 min. Store at RT.
6. PTC: dissolve PEG 8000 (Sigma-Aldrich) in STC to the final concentration of 40% (w/v). Sterilize by filtration through 0.45 μm filter. Store at RT for up to 1 week.

7. TB3 (1 L): 3 g yeast extract, 3 g casamino acids, 20% sucrose. Autoclave for 20 min.
8. Bottom agar: add 0.65 g agar to 100 mL TB3. Autoclave for 20 min. When cooled down to 55°C, add 1 mL of 10 mg/mL caffeine, and then add bleomycin (Invitrogen) to the final concentration of 150 μg/mL.
9. Top agar: similar to the top agar except the final concentration of bleomycin is 250 μg/mL.
10. 10 mg/mL caffeine: dissolve 100 mg caffeine (Sigma-Aldrich) into 10 mL of sterile distilled water and store at 4°C for up to 1 month.
11. The *M. oryzae* wild-type strain 70-15: a strain derived from a genetic cross (12).

2.7. Yeast Shuttle Vectors

1. Vectors pDL2 (6), pYP1, and pFL1 carry the hygromycin (Hyg^R), bleomycin ($Bleo^R$), and geneticin (Gen^R) resistance genes, respectively (Fig. 1a). They are suitable for generating C-terminal GFP fusion constructs.
2. Vectors pLX1 (Hyg^R), pFL4 ($Bleo^R$), and pFL3 (Gen^R) are suitable for generating C-terminal 13×Myc tagged constructs (Fig. 1b).
3. Vectors pHZ126 (Hyg^R), pFL5 ($Bleo^R$), and pFL6 (Gen^R) are suitable for generating C-terminal 3×FLAG tagged constructs (Fig. 1c).
4. Vectors pFL6 ($Bleo^R$) and pFL8 (Gen^R) are suitable for generating N-terminal 3×FLAG tagged constructs (Fig. 1d).

3. Methods

3.1. Amplification of Target Genes

The vectors depicted in Figs. 1 and 2 can be used to generate different fusion constructs that are under the control of the native promoter of the target gene (the gene of interest) or the RP27 promoter(6). The example described below is for generating the C-terminal GFP fusion construct of the *M. oryzae MCM1* gene (MGG_02773.6). Similar approaches can be used to generate different GFP fusion or epitope tag constructs with appropriate vectors (Figs. 1 and 2) and primers (Table 1).

1. Primers for constructing the *MCM1*-GFP fusion were MCM1-GF (Table 1, 5′-*TATAGGGCGAATTGGGTACT CAAATTGGTT*GTTGAGCTGTGCCGATGGTATC-3′) and MCM1-GR (Table 1, 5′-*CCCGGTGAACAGCTCCTCG CCCTTGCTCAC*TGACTGGTGAGAGCTGTGTTG-3′).

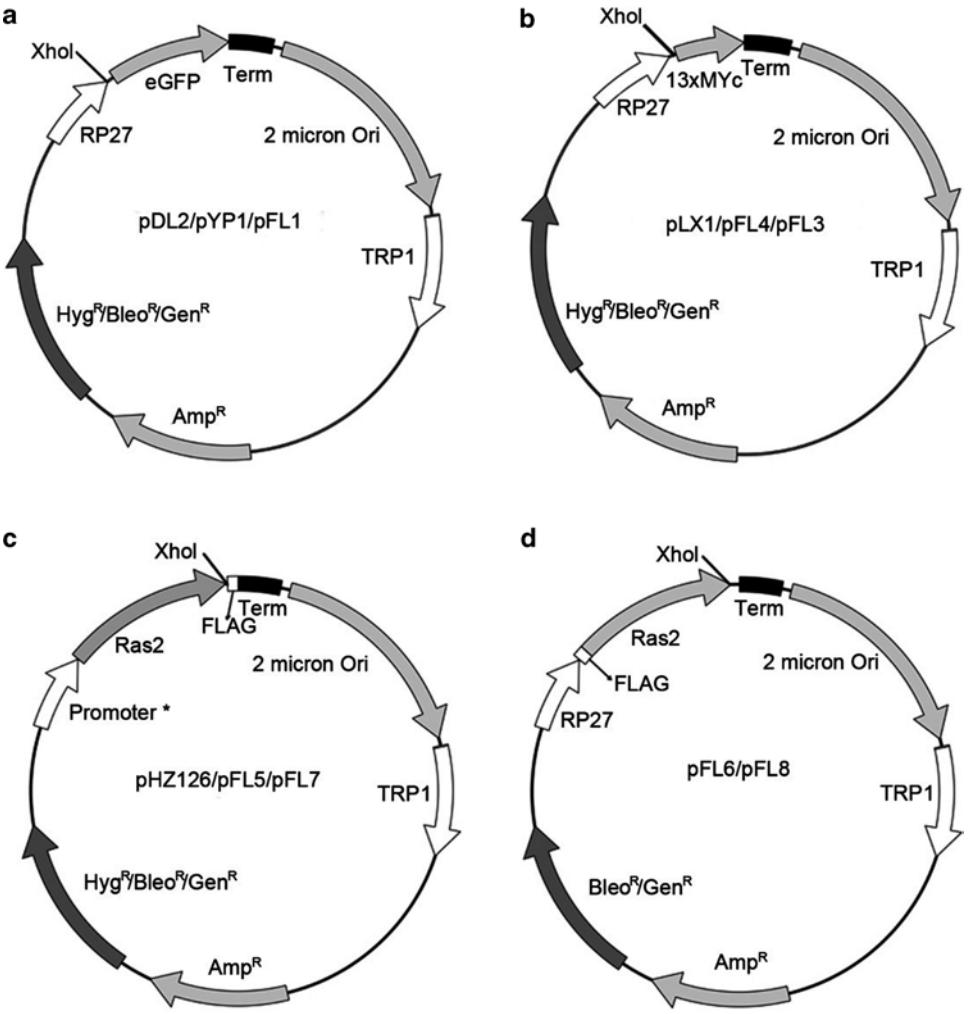

Fig. 1. Yeast shuttle vectors for generating C-terminal GFP or 13×Myc fusion constructs. (**a**) GFP fusion vectors pDL2, pYP1, and pFL1 carry the hygromycin (Hyg^R), bleomycin ($Bleo^R$), and geneticin (Gen^R) resistance genes, respectively. (**b**) Vectors pLX1 (Hyg^R), pFL4 ($Bleo^R$), and pFL3 (Gen^R) for constructing 13×Myc fusions. *Xho*I is the site for linearization. RP27 is the strong constitutive promoter from *M. oryzae* (6). The terminator (Term) sequence is from the *M. oryzae* beta-tubulin gene. The 2 μ origin of replication (2 micron *ori*) is for replication in yeast. *TRP1* and Amp^R are selectable markers for transformation of *S. cerevisiae* and *E. coli*, respectively. (**c**) Vectors pHZ126, pFL5, and pFL7 carry Hyg^R, $Bleo^R$, and Gen^R resistance genes, respectively, and are suitable for generating C-terminal 3×FLAG fusion constructs. (**d**) Vectors pFL6 ($Bleo^R$) and pFL8 (Gen^R) for constructing N-terminal fusions.

The underlined sequences are specific for the *MCM1* gene. The nucleotide sequences in italic are not gene-specific. They are homologous to the sequences flanking the *Xho*I site on the yeast shuttle vector pYP1 (Fig. 1) (see Note 1).

2. Set up the PCR reaction for amplifying the target gene as shown below (see Note 2):

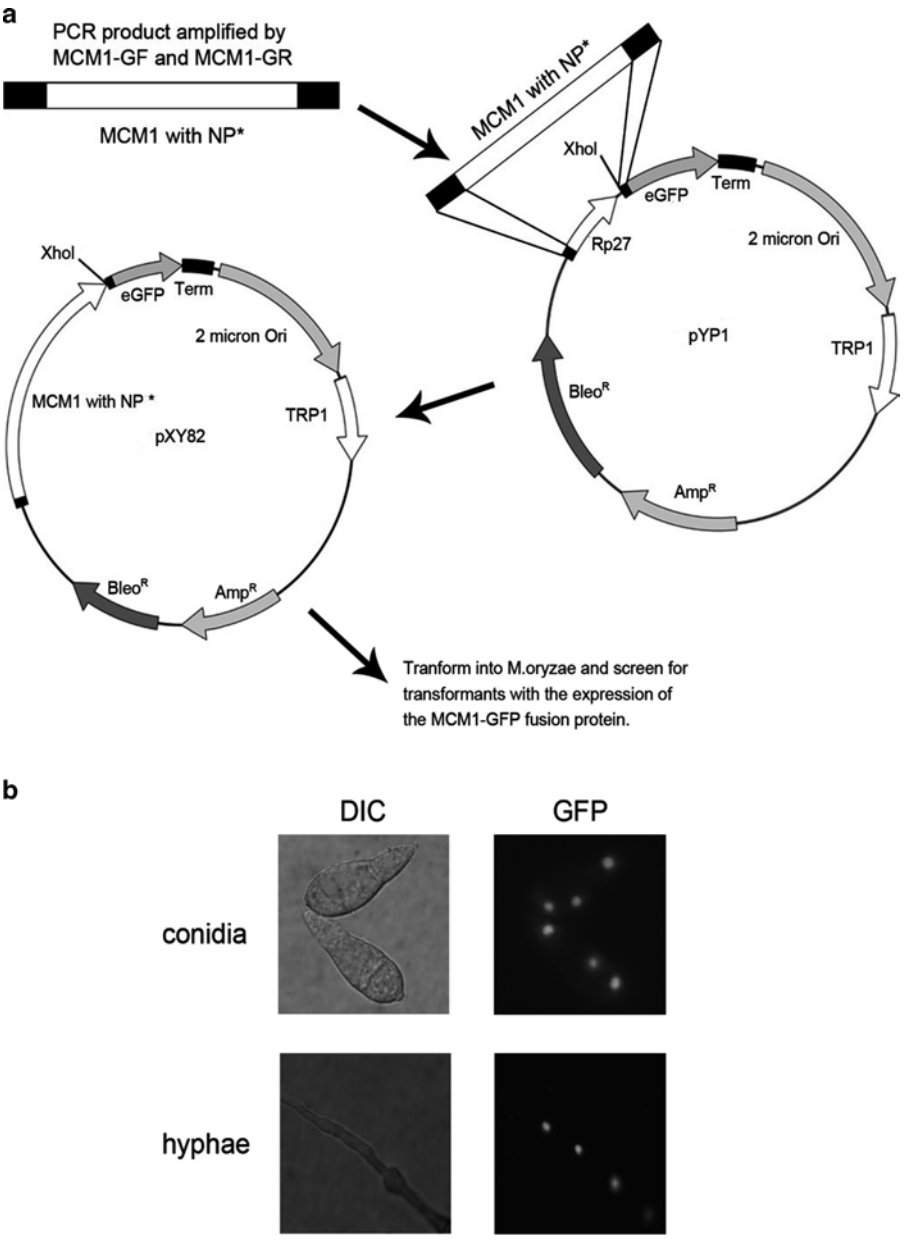

Fig. 2. The *MCM1*-GFP fusion construct and transformants. (**a**) The *MCM1* fragment amplified with primers MCM1-GF and MCM-GR was integrated into *Xho*I-digested pYP1 by double homologous recombinations (gap repair). The resulting construct was transformed into *M. oryzae*. (**b**) Expression and localization of Mcm1-GFP fusion proteins in the resulting transformants. GFP signals were observed in the nuclei in conidia and vegetative hyphae harvested from the *MCM1*-GFP transformant.

Table 1
Primer sequences

Primer	Sequence (5′–3′)
MCM1-GF	TATAGGGCGAATTGGGTACTCAAATTGGTT GTTGAGCTGTGCCGATGGTATC
MCM1-GR	CCCGGTGAACAGCTCCTCGCCCTTGCTCAC TGACTGGTGAGAGCTGTGTTGC
MCM-NF	CCAAGAAAACCGAGGCATTA
MCM-NR	GATGGGTCCATGCCTACATT
GFPR	GTGGTGCAGATGAACTTCA
Native-FP[a]	TATAGGGCGAATTGGGTACTCAAATTGGTT
RP27-FP	TTTCGTAGGAACCCAATCTTCAAA
GFP-RP[b]	CCCGGTGAACAGCTCCTCGCCCTTGCTCAC
13×Myc-RP	CCTGCAGCGTACGAAGCTTCAGCTG
3×FLAG-RP	TTTATAATCACCGTCATGGTCTTTGTAGTC

[a]For constructing C-terminal fusions (GFP, 13×Myc, or 3×FLAG), nucleotide sequences Native-C-FP and RP27-C-FP should be added to the forward primers that match to the 5′-end of the promoter region and open reading frame of the target gene, respectively. The resulting fusion constructs are under the control of the native promoter of the target gene or the RP27 promoter (6)
[b]For the reverse primers, nucleotide sequences GFP-C-RP, 13×Myc-C-RP, 3×FLAG-C-RP that match to the C-terminal sequences of GFP, 13×Myc, and 3×FLAG, respectively, should be added to the 3′-end of the open reading frame of the target gene

50× Advantage 2 polymerase mix	1 μL
50× dNTP mix	1 μL
Genomic DNA of strain 70-15	1 μL
5′-primer MCM1-GF (10 μM)	1 μL
3′-primer MCM1-GR (10 μM)	1 μL
10× Advantage 2 PCR buffer	5 μL
Add sterile distilled water to	50 μL

3. Run the PCR reaction as: 95°C 2 min, 30 cycles of 95°C 30 s and 68°C 1 min, followed with 68°C 5 min.
4. The resulting PCR products were separated on 1% agarose gel in 1× TAE.
5. Isolate the DNA band of the expected size from the agarose gel.
6. Purify the *MCM1* fragment with the QIAquick gel extraction kit (Qiagen). Elute PCR products in 30–50 μl TE (pH 8.0).

3.2. Vector Treatment

All the yeast shuttle vectors depicted in Figs. 1 and 2 have one single *Xho*I site that can be used for linearization.

1. Digest 80–100 μg vector DNA with 5 μl of *Xho*I in 500 μl volume with 1× reaction buffer 2 at 37°C for 10 h (see Note 3).
2. Heat at 65°C for 20 min to inactivate the restriction enzyme.
3. Run 3 μL of the digestion mixture on a 0.7% agarose gel. If the vector is completely digested by *Xho*I, there should be only one band of the linearized plasmid DNA. In case digestion is incomplete, leave the reaction mixture at 37°C for another 10 h (see Note 4).
4. Separate *Xho*I-linearized vector on a 0.7% agarose gel.
5. Purify the vector band with the QIAquick gel extraction kit (Qiagen).
6. Elute the *Xho*I-linearized vector DNA with TE. Adjust to the plasmid concentration to approximately 0.05–0.5 μg/μl with sterile distilled water. Store the treated vector at −20°C.

3.3. Preparation of Yeast Competent Cells

1. Streak the *S. cerevisiae* strain XK1-25 on a YEPD agar plate. Incubate at 30°C for 2 days.
2. Inoculate a single colony of XK1-25 into 100 mL YEPD broth in a 250-mL flask. Incubate at 30°C with vigorous shaking (250 rpm) until OD_{600} reaches 0.7 (see Note 5).
3. Transfer yeast cultures to sterile centrifuge tubes. Spin at 2,000 × *g* for 5 min at RT in a GSA rotor (Sorvall). Discard the supernatant.
4. Gently resuspend the pellet with 10 mL TE (pH 7.5). Centrifuge at 2,000 × *g* for 5 min. Discard the supernatant.
5. Gently resuspend yeast cells in 5 mL lithium/cesium acetate solution.
6. Incubate at 30°C for 30 min with gentle shaking at 80 rpm.
7. Spin at 2,000 × *g* for 5 min to pellet yeast cells. Discard the supernatant.
8. Gently resuspend yeast cells in 1 mL of prechilled TE (pH 7.5). Keep on ice.
9. Aliquot 40 μL of the resulting competent cells into sterile 1.5-mL Eppendorf tubes for transformation (see Note 6).

3.4. Yeast Transformation

The following yeast transformation procedure is modified from the instructions provided by the manufacturer of the Alkali-cation yeast transformation kit.

1. Gently mix 100 μL of yeast competent cells with (see Note 7):

   5 μL 10 mg/ml carrier DNA

   5 μL histamine solution

1 μL *Xho*I-digested vector DNA

9 μL PCR fragment of the target gene

2. Incubate at 25°C for 10–15 min without disturbance.
3. Add 800 μL of 50% PEG and 200 μL TE/cation MIXX into each transformation tube. Mix by pipetting gently.
4. Incubate the transformation mixture at 30°C for 10 min.
5. Place the reaction tube in a 42°C water bath. Incubate for exactly 10 min as the heat shock treatment.
6. Take the tubes out of water bath and cool down to RT.
7. Centrifuge in a microcentrifuge at 12,470 × *g* for 5 s, and then remove the supernatant.
8. Resuspend the yeast cells in 200 μL of SOS solution and plate out on an SD-Trp plate (see Note 8).
9. Incubate at 30°C for 3 days. The Trp + transformants are selected as colonies formed on SD-Trp plates.

3.5. Identification of Fusion Constructs

1. Scrape off all Trp + yeast transformants grown on SD-Trp plates with 1.5 mL of SCE buffer.
2. Transfer yeast cells into an Eppendorf tube and centrifuge at 4,722 × *g* for 5 min. Remove the supernatant.
3. Resuspend the yeast cells (from step 2) in 0.5 mL of freshly prepared lytic enzyme solution (see Note 9).
4. Incubate at 37°C for 2 h (invert the tubes every 20–30 min). Majority of yeast cells should become spheroplasts after lytic enzyme treatment (see Note 10).
5. Centrifuge at 8,000 rpm for 5 min. Remove the supernatant.
6. Resuspend the spheroplasts in 0.4 mL of spheroplast lysis buffer. Incubate at 70°C for 15 min.
7. Add an equal volume of phenol/chloroform/isoamyl alcohol (25:24:1). Mix thoroughly by inverting the tubes multiple times.
8. Centrifuge at 13,000 rpm for 10 min at RT. Transfer the supernatant to a new tube. Add 2× volume of 100% EtOH, mix briefly, then keep the tubes at –20°C for 30 min (see Note 11).
9. Centrifuge in a microcentrifuge for 15 min at 13,000 rpm. Wash the DNA pellet with 1 mL of 70% EtOH. Resuspend air-dried DNA in 50 μL of TE.
10. Transform competent cells of *E. coli* strain DH10B with the resulting yeast plasmid DNA (see Note 12). Screen ampicillin-resistant colonies grown on LB + Amp by colony PCR with primers MCM1-NF and MCM1-NR (Table 1) for clones containing the *MCM1*-GFP fusion construct (see Note 13).

11. Use the reverse primer (GFPR, Table 1) located at the N-terminal region of EGFP to sequence the *MCM1*-GFP fusion construct to verify the in-frame fusion and check for possible PCR errors.

3.6. Generation of Fungal Transformants Expressing the Fusion Constructs

Fusion constructs generated in vectors depicted in Figs. 1 and 2 have selectable markers in the vector backbone. They can be directly used for fungal transformation with general procedures as described (13). The *MCM1*-GFP construct was generated with the pYP1 vector. It contains the bleomycin-resistance marker.

1. Harvest vegetative hyphae of the *M. oryzae* wild-type strain 70-15 (see Note 14) from 100 ml 2-day-old liquid 5×YEG by filtration through one layer of Miracloth. Rinse with 100 mL of 1 M sorbitol. Digest the resulting hyphae with 20 ml of protoplasting solution for 2 h at 30°C. Filter through two layers of Miracloth. Centrifuge the filtrate at 4,000 × *g* for 5 min.
2. Wash the protoplast pellets twice with 50 mL of STC.
3. Resuspend protoplasts in STC to 1×10^7 protoplasts/mL.
4. Mix approximately 5 μg of the *MCM1*-GFP construct DNA with 200 μL of protoplasts. Incubate at RT for 20 min.
5. Transfer the mixture to a 15-mL conical tube. Add 1 mL of PTC. Mix well by inverting the tube, and then incubate at RT for 20 min.
6. Add 5 mL of TB3. Incubate at RT for 18 h with gentle shaking on a rocker.
7. Spin down the regenerated protoplasts at 4,000 × *g* for 5 min. Remove the supernatant.
8. Add 10 mL of bottom agar with 150 μg/mL bleomycin and 0.1 mg/mL caffeine. Mix briefly and pour into a Petri plate. Incubate at RT for 24 h.
9. Overlay with 10 mL of top agar containing 250 μg/mL bleomycin and 0.1 mg/mL caffeine. Incubate at RT for 4–6 days.
10. Transfer bleomycin-resistant transformants that become visible after 4 days to oatmeal agar plates.
11. Screen by PCR with primers MCM1-NF and MCM1-NR (Table 1) for *M. oryzae* transformants carrying the *MCM1*-GFP construct.
12. Examine for GFP signals in fungal hyphae or conidia of the *MCM1*-GFP transformants by epifluorescence microscopy (Fig. 2) (see Note 15).

Although the example described above is for the *MCM1*-GFP C-terminal fusion, suitable yeast shuttle vectors can be developed to construct N-terminal fusion with GFP. We have generated

vectors that are suitable for generating fusion constructs with 13×Myc and 3×FLAG tags (Figs. 1 and 2). The resulting constructs can be directly used for fungal transformation because hygromycin, bleomycin, or geneticin are three commonly used antibiotics in filamentous fungi. Therefore, approaches similar to what described in this chapter can be used to generate fungal transformants expressing various fusion or epitope-tagging constructs. In addition to PCR or Southern blot analysis, the resulting transformants can be further verified for the expression of fusion constructs by western blot analysis with commercially available anti-Myc or anti-FLAG antibodies.

4. Notes

1. Table 1 lists nucleotide sequences that should be added to the 5′-end of the PCR primers for generating different fusion constructs with vectors carrying desirable selectable markers under the control of the native or RP27 promoter.
2. In general, PCR products from reaction volume of 50–100 μl are sufficient to transform yeast competent cells.
3. It is a good practice to digest vector DNA with *Xho*I in large quantity. If any particular batch of *Xho*I-digested DNA has low background and high efficiency for gap repair, it could be saved for later use.
4. Complete digestion of the vector is critical for yeast gap repair because undigested plasmid causes background problems for yeast transformation and later screenings.
5. Yeast cultures with OD_{600} between 0.5 and 1 are suitable for preparing competent cells.
6. When DMSO is added to the final concentration of 7%, yeast competent cells could be stored for a short period time at −80°C. Do not freeze yeast cells with liquid nitrogen to avoid reduced transformation efficiency.
7. The molecular ratio between PCR fragment of the target gene and treated vector should be around 3:1.
8. Glucose should be autoclaved separately or filter sterilized and added to SD-Trp medium right before pouring into Petri plates.
9. If the amount of lytic enzyme cannot be weighed accurately, adding estimated amount will be fine.
10. Spheroplast formation can be monitored microscopically by adding SDS to yeast cells digested with lytic enzyme and observing for ghosts.

11. In generally, at least 300 µl of the supernatant could be obtained. It is optional to add 1/10 volume of 3 M NaAc to the supernatant for DNA precipitation.
12. Instead of using electroporation cells from Invitrogen, competent cells of *E. coli* strain DH10B can be prepared as described (14).
13. If the vector is digested completely, over 60% of *E. coli* colonies should be positive clones.
14. The *MCM1*-GFP fusion construct contains the bleomycin-resistance gene. It is suitable for transformation with both the wild-type and *mcm1* deletion mutant (hygromycin-resistant) strains. The fusion construct complemented the defects of the *mcm1* mutant, indicating that fusion of GFP to *MCM1* had no effect on its function.
15. In addition to examining for fluorescent signals, the expression of the *MCM1*-GFP fusion construct can be assayed by western blot analysis with anti-GFP antibodies that are commercially available.

Acknowledgments

The authors thank Dr. Flora Urmeev at Purdue University for assistance in constructing the yeast shuttle vectors. This work was supported by a USDA-NRI grant 2005-35319-16073 and grants from the USWBSI.

References

1. Xu, J. R., Peng, Y. L., Dickman, M. B., and Sharon, A. (2006) The dawn of fungal pathogen genomics, *Ann. Rev. Phytopathol.* ***44***, 337–366.
2. Ebbole, D. J. (2007) Magnaporthe as a model for understanding host-pathogen interactions, *Annu. Rev. Phytopathol.* ***45***, 437–456.
3. Xu, J. R., Zhao, X., and Dean, R. A. (2007) From genes to genomes; a new paradigm for studying fungal pathogenesis in *Magnaporthe oryzae*, *Adv. Genet.* ***57***, 175–218.
4. Raymond, C. K., Sims, E. H., and Olson, M. V. (2002) Linker-mediated recombinational subcloning of large DNA fragments using yeast, *Genome Res.* ***12***, 190–197.
5. Raymond, C. K., Mugford, V. R., and Sexson, S. L. (1999) Plasmid topologies that enhance the transformation efficiency of yeast, *Biotechniques 27*, 892–894, 896.
6. Bourett, T. M., Sweigard, J. A., Czymmek, K. J., Carroll, A., and Howard, R. J. (2002) Reef coral fluorescent proteins for visualizing fungal pathogens, *Fungal Genet. Biol.* ***37***, 211–220.
7. Bruno, K. S., Tenjo, F., Li, L., Hamer, J. E., and Xu, J. R. (2004) Cellular localization and role of kinase activity of *PMK1* in *Magnaporthe grisea*, *Eukaryot. Cell* ***3***, 1525–1532.
8. Ho, Y., Gruhler, A., Heilbut, A., Bader, G. D., et al. (2002) Systematic identification of protein complexes in *Saccharomyces cerevisiae* by mass spectrometry, *Nature* ***415***, 180–183.
9. Zhao, X. H., and Xu, J. R. (2007) A highly conserved MAPK-docking site in Mst7 is essential for Pmk1 activation in *Magnaporthe grisea*, *Mol. Microbiol.* ***63***, 881–894.
10. Seong, K., Li, L., Hou, Z. M., Tracy, M., Kistler, H. C., and Xu, J. R. (2006) Cryptic

promoter activity in the coding region of the HMG-CoA rediactase gene in *Fusarium graminearum*, *Fungal Genet. Biol.* **43**, 34–41.

11. Xu, G., Jansen, G., Thomas, D. Y., Hollenberg, C. P., and Rad, M. R. (1996) Ste50p sustains mating pheromone-induced signal transduction in the yeast *Saccharomyces cerevisiae*, *Mol. Microbiol.* **20**, 773–783.
12. Chao, C. C. T., and Ellingboe, A. H. (1991) Selection for mating competence in *Magnaporthe grisea* pathogenic to rice, *Can. J. Bot.* **69**, 2130–2134.
13. Sweigard, J. A., Chumley, F. G., and Valent, B. (1992) Disruption of a *Magnaporthe grisea* cutinase gene, *Mol. Gen. Genet.* **232**, 183–190.
14. Sambrook, J., and Russell, D. (2001) *Molecular Cloning – A Laboratory Manual*, 2nd ed., Cold Spring Harbor Laboratory Press, Cold Spring Harbor, NY.

Chapter 16

Large-Scale Insertional Mutagenesis in *Magnaporthe oryzae* by *Agrobacterium tumefaciens*-Mediated Transformation

Xiao-Lin Chen, Jun Yang, and You-Liang Peng

Abstract

With genome sequences of more and more fungi become available, high-throughput systematic mutagenesis is desirable for functional genomics studies. While a number of random insertional mutagenesis and targeted gene disruption approaches have been used in filamentous fungi, *Agrobacterium tumefaciens*-mediated Transformation (ATMT) remains one of the most effective methods for identifying genes required for specific fungal developmental or infection processes. Because of its simplicity, ATMT is suitable for large-scale insertion mutagenesis in fungi. *Magnaporthe oryzae*, the rice blast fungus is a model for studying host–pathogen interactions. Here, we describe protocols for generating a *M. oryzae* mutant library consisting of over 70,000 ATMT transformants and for identifying genes disrupted by T-DNA in the mutants by TAIL-PCR.

Key words: Rice blast, *Magnaporthe oryzae*, Pathogenicity genes, Functional genomics, Insertion mutants

1. Introduction

Filamentous fungi are groups of eukaryotes with compact genomes. Many fungi are cultivable in artificial media and are tractable for classic and molecular manipulations. As sequences of more and more fungal genomes become available, functional genomic studies of genes that are unique or common to certain fungal groups became necessary and possible (1–4). Several approaches have been used to generate disruption or deletion mutants in fungi (5). Insertional mutagenesis is an approach

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722,
DOI 10.1007/978-1-61779-040-9_16,

during which, exogenous or modified DNA vectors are randomly integrated into the recipient genome. Since the sequences of transforming vectors are known, it is relatively easy to identify genes that are disrupted or activated in mutants (5). There are several derivatives of insertion mutagenesis, including transposon-mediated mutation (6), restriction enzyme-mediated integration (REMI) (7), and *Agrobacterium tumefaciens*-mediated transformation (ATMT) (8, 9).

A. tumefaciens is a phytopathogenic bacterium that contains the Ti plasmid, which contains the T-DNA region (10) and the Vir region (11). The T-DNA region is a DNA segment that can be integrated into the genome of host cells. ATMT was initially used for plant transformation (12, 13). In 1995, Bundock et al. firstly reported ATMT of the budding yeast *Saccharomyces cerevisiae* (14). Since then, ATMT has been successfully applied to transform a number of filamentous fungi (9, 15). In 2001, Rho et al. first reported ATMT of the rice blast fungus *Magnaporthe oryzae* (16), which is the causal agent for the most destructive fungal disease of rice. Recently, we have established an efficient ATMT system for *M. oryzae* and constructed an insertional mutant library with over 70,000 independent transformants. This mutant library has been used for identifying genes required for fungal growth, differentiation, and plant infection processes. Here, we describe the detailed protocols of ATMT and TAIL-PCR used for mutagenesis and gene discovery in *M. oryzae*.

2. Materials

2.1. Construction of the A. tumefaciens Strain for Fungal Transformation

1. The binary vector pBI-G_3C (17) was kindly provided by Dr. Y. Kubo, Kyoto Prefectural University, Japan. It was constructed with pBIN19 (18) by cloning the hygromycin B phosphotransferase (*hph*) gene cassette, the ColE1 replication origin, and the chloramphenicol resistance gene between the left and right borders of the T-DNA (17).
2. *A. tumefaciens* strain: EHA105 (19).
3. *Escherichia coli* strain: JM109.
4. YM medium (1 l): 0.5 g KH_2PO_4, 10 g mannitol, 2 g l-glutamine, 0.2 g NaCl, 0.2 g $MgSO_4$, 0.3 g yeast extract, adjust pH to 7.0. Add 15 g/l agar for solid medium. Autoclave for 20 min.
5. LB (1 l): 10 g tryptone, 5 g yeast extract, 5 g NaCl, adjust pH to 7.0. Add 15 g/l agar for solid medium. Store at room temperature (RT) after autoclave.

6. SOB (1 l): 20 g tryptone, 5 g yeast extract, 0.25 g NaCl, 10 ml 0.25 M KCl, adjust pH to 7.0. Store at RT after autoclave.
7. 100 mM $CaCl_2$: Dissolve $CaCl_2$ in sterile distilled water (DW). Sterilize by filtration through a 0.45-μm filter.
8. Kanamycin stock solution 50 mg/ml: Dissolve kanamycin in DW and store at −20°C.
9. Streptomycin stock solution 50 mg/ml: dissolve streptomycin in ethanol. Store at −20°C.
10. Rifampicin stock solution 10 mg/ml: dissolve rifampicin in methanol. Store at −20°C.

2.2. Preparation of Fungal Conidia

1. *M. oryzae* strain: P131, a field isolate from rice.
2. Oatmeal tomato agar (OTA) medium (1 l) (20): 40 g boiled oatmeal filtrate, 150 ml tomato juice, 20 g agar. Boil the oatmeal in 800 ml water for 30 min, filtrated by double-deck pledgets. Extract tomato juice by extractor, then filtrated by double-deck pledgets. Mix 150 ml tomato juices with boiled oatmeal filtrate. Add water to 1 l. Add 20 g agar for solid media. Autoclave for 40 min.

2.3. Preparation of *A. tumefaciens* Cells for Transformation

1. Minimal Medium (MM) (1 l) (16)

   Mix the following reagents into 941.5 ml distilled water:

   10 ml K-phosphate buffer (pH 7.0) (200 g/l K_2HPO_4, 145 g/l KH_2PO_4).

   20 ml M-N solution (30 g/l $MgSO_4\ 7H_2O$, 15 g/l NaCl).

   1 ml 1% $CaCl_2\ 2H_2O$ (w/v).

   10 ml 20% glucose (w/v), sterilized by filtration.

   10 ml 0.01% $FeSO_4$ (w/v), sterilized by filtration.

   5 ml Spore Elements (100 mg/l $ZnSO_4\ 7H_2O$, 100 mg/l $CuSO_4\ 5H_2O$, 100 mg/l H_3BO_3, 100 mg/l $MnSO_4\ H_2O$, 100 mg/l $Na_2MoO_4\ 2H_2O$).

   2.5 ml 20% NH_4NO_3 (w/v).

   All these reagents except for $FeSO_4$ are stable at RT. They can be mixed and stored at RT. Add $FeSO_4$ before transformation.

2. Induction Medium (IM) (1 l) (16)

   Mix the following reagents into 898.7 ml distilled water:

   0.8 ml 1.25 M K-buffer (pH 4.9): 184 g/l K_2HPO_4 (pH adjusted with phosphoric acid).

   20 ml M-N solution (as described above).

   1 ml 1% $CaCl_2\ 2H_2O$.

   10 ml 0.01% $FeSO_4$ (w/v), sterilize by filtration.

5 ml Spore Elements (as described above).

2.5 ml 20% NH_4NO_3 (w/v).

10 ml 50% glycerol.

40 ml 1 M MES (2-(*N*-morpholino)ethanesulfonic acid), pH adjusted to 5.5 with NaOH, sterilize by filtration.

10 ml 20% glucose (w/v).

2 ml 100 mM acetosyringone (AS, dissolved in ethanol, store at –20°C).

All above regents can be stored at RT except AS.

Add $FeSO_4$, MES, and AS before transformation. AS should only be added when the medium temperature is lower than 55°C.

2.4. Media for Cocultivation and Transformant Selection

1. Cocultivation medium (plates) (1 l) (16): Same as IM (see above) except only 5 ml (not 10 ml) 20% glucose is added. Add 15 g agar for solidification before autoclave.
2. Complete medium (CM) (1 l): 6 g yeast extract, 3 g enzymatic casein hydrolysate, 3 g acidic casein hydrolysate, 10 g glucose. Add 15 g agar for solid medium.
3. Half CM (0.5× CM): Dilute CM in DW (1:1) before adding agar.
4. Hygromycin B: Dilute to 50 mg/ml in water. Store at 4°C.
5. Cefotaxime: Dissolve in DW to 200 mg/ml. Store at –20°C.
6. Filter paper (Φ9 cm disks, intermediate filtering velocity): Xinhua Filter Paper, Hangzhou, China.

2.5. Single Spore Isolation and Strain Preservation

1. Sterilized distilled water.
2. OTA plates.
3. Sterile filter paper pieces (3–4 mm^2).
4. Sterile waxed paper bags (Fengnong paper, Jinan, China).

2.6. Reagents for TAIl-PCR

1. CATB extract buffer: 10% CTAB (w/v), 50 mM Tri–HCl, 0.7 M NaCl, 100 mM EDTA.
2. TE: 10 mM Tris–HCl (pH 7.5), 1 mM EDTA.
3. dNTP Mixture: 2.5 mM each dATP, dCTP, dGTP, and dTTP (TaKaRa, Dalian, China), store at –20°C.
4. Taq DNA polymerase: TAKARA Taq (5 U/μl).
5. PCR primers: Table 1 lists the sequences of specific and degenerate TAIL primers. All primers are resuspended in DDW to 50 μM.
6. 1× TAE: 0.04 M Tris–acetate; 0.001 M EDTA (pH 8.0).
7. BioMed DNA purification kit: Biomed, Beijing, China.

Table 1
Primers used for TAIL-PCR with *M. oryzae* DNA

Primer name	Primer sequence (5′–3′)	Note
HS-1	GGCCGTGGTTGGCTTGTATGGAGCAGCAGA	Left border, specific
HS-2	TGGTCTTGACCAACTCTATCAGAGCTTGGT	Left border, specific
HS-3	TCTGGACCGATGGCTGTGTAGAAGTACTCG	Left border, specific
RHS-1	CTTGATTAGGGTGATGGTTCACGTAGT	Right border, specific
RHS-2	CAACACTCAACCCTATCTCGGTCTATTC	Right border, specific
RHS-3	GGTTTTCCCAGTCACGACGTTGT	Right border, specific
AD-1	NGTCGASWGANAWGAA	Degenerate
AD-2	GTNCGASWCANAWGTT	Degenerate
AD-3	WGTGNAGWANCANAGA	Degenerate

3. Methods

3.1. Transformation of A. tumefaciens with the Vector pBI-G_3C

1. For preparation of competent cells, inoculate a single colony of *A. tumefaciens* strain EHA105 into 5 ml YM, and shake for about 18 h at 220 rpm at 28°C to the log phase (OD_{600} about 0.5). Aliquot 0.2 ml of the culture into 1.5 ml sterile EP tubes and centrifuge for 5 min at 2,400 × *g* at RT. Remove the supernatant and add 0.2 ml prechilled 0.1 M $CaCl_2$ to the pellet. Mix gently by pipetting and place the cells on ice (see Note 1).
2. Add 1 μg pBI-G_3C DNA to the competent cells and mix thoroughly. Keep the tube on ice for 30 min followed by freezing at –70°C for 10 min.
3. Heat shock the transformation mixture at 42°C for 1 min. Add 800 μl YM liquid medium and shake at 28°C for 3 h at 175 rpm.
4. Spread 300–400 μl of the transformation mixture on a YM plate with 50 μg/ml kanamycin and incubate at 28°C for 48 h.
5. Kanamycin-resistant *A. tumefaciens* transformants were verified by PCR to contain pBI-G3C and preserved in 15% (v/v) glycerol at –80°C.

3.2. Preparation of M. oryzae Conidia

The approach described by Peng and Shishiyama (20) was used to prepare fresh conidia for transformation.

1. Inoculate OTA plates with *M. oryzae* strain P131 and incubate for 7 days in an illumination incubator at 28°C.

2. Add 2 ml sterilized distilled water to each plate. Gently scrape off aerial hyphae and conidia with a spreader. Spread 0.5 ml of the resulting suspension onto a fresh OTA plate (see Note 2).
3. Incubate at 28°C in an illumination incubator until new mycelia become visible by naked eyes (usually up to 36 h postinoculation).
4. Gently remove superficial mycelia with cotton swabs. Rinse with sterilized distilled water to remove debris. Remove excessive water from the surface with a sterile filter paper.
5. Cover the plates with two layers of gauze and incubate at 28°C for 48 h in an illumination incubator (see Note 3).
6. Add 15–20 ml sterile distilled water to each plate and scrape gently with a spreader to harvest conidia. Filter with two layers of sterile lens paper to remove hyphal fragments and resuspend conidia thoroughly by brief vortexing.
7. Determine the spore concentration with a hemocytometer and adjust the final concentration to 10^6 per ml in sterile distilled water (see Note 4).

3.3. Cocultivation of Agrobacterium Cells with Fungal Spores

1. Streak the pBI-G_3C transformant of *A. tumefaciens* EHA105 on an LB plate with 50 μg/ml kanamycin and 170 μg/ml rifampicin.
2. Incubate at 28°C for 2 days (see Note 5).
3. Scrape *A. tumefaciens* cells off the streak (2 mm width and 4 mm length) and inoculate into 10 ml MM with 50 μg/ml kanamycin and 170 μg/ml rifampicin in a 50-ml flask.
4. Shake the culture at 220 rpm at 28°C until OD_{600} reaches 1.2 (see Note 6).
5. Collect *A. tumefaciens* cells into two 50-ml tubes. Centrifuge at 2,400 × *g* for 1 min at RT.
6. Resuspend the pellet in IM with 50 μg/ml kanamycin and 200 μg/ml AS to $OD_{600} = 0.18$ (see Notes 7 and 8).
7. Transfer 15 ml of the cell suspension into a 50-ml flask. Shake at 220 rpm for 6 h at 28°C (see Note 9).
8. Mix 200 μl *A. tumefaciens* cells with 200 μl of the spore suspension (prepared in Subheading 3.2).
9. Spread 400 μl of the mixture evenly onto a sterile filter paper that is placed over the cocultivation medium in a petri plate. The sterile filter paper disk is sliced into 0.5 cm strips in advance but left to be connected at the edge (Fig. 1).
10. Incubate in the dark at 23°C for 3 days (see Note 10).

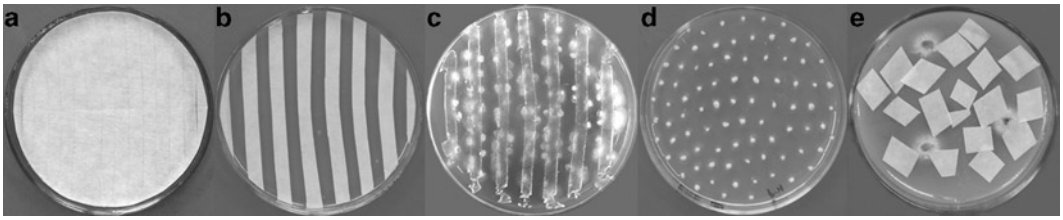

Fig. 1. Keep steps in the ATMT transformation of *M. oryzae*. (**a**) Cocultivation of *M. oryzae* spores with *A. tumefaciens* cells on a precut filter paper laid over the cocultivation medium. (**b**) Filter paper strips placed upside down in a half-CM plate for primary selection. (**c**) Transformants grown on a half-CM plate after removing the filter paper strips. (**d**) Putative hygromycin-resistant transformants from the primary transformation plate were picked and grown on a secondary selection plate. (**e**) Preparation of filter paper stocks for preservation.

3.4. Isolation of Hygromycin-Resistant Transformants

1. After cocultivation for 3 days, separate the filter strips and transfer them onto half CM plates with 200 μg/ml hygromycin B and 200 μg/ml cefotaxime (see Note 11). The filter strips should be placed upside down and 0.5 cm away from each other (Fig. 1).
2. Incubate at 28°C for 2 days in an illumination incubator.
3. Remove the filter paper strips. Incubate the half CM plates with the transformants for 2 days (see Note 12).
4. Transfer individual hygromycin-resistant transformants onto CM plates with 200 μg/ml hygromycin B using sterile toothpicks. Incubate at RT for 2 days (Fig. 1). We normally inoculate about 50 transformants onto one CM plate. This step serves as the secondary selection to reduce false positive transformants.
5. Incubate the original transformation plates for 2 more days at 25°C to allow transformants with reduced growth rate to grow. Pick these slow growing transformants onto CM plates with 200 μg/ml hygromycin B as described in step 4.
6. Individual hydromycin-resistant colonies on the secondary selection plates are then inoculated onto OTA plates and incubate for 2–3 days.

3.5. Single Spore Isolation and Preservation of Hygromycin-Resistant Transformants

1. Add 300 μl sterilized water onto individual colonies formed on the OTA plates. Pipette a few times to suspend conidia in water.
2. Spread the spore suspension onto a water agar plate. Transfer individual conidia under microscope to new OTA plates.
3. Incubated at 25°C for 3 days.
4. Place a dozen of sterilized filter paper pieces on the edges of the colony. Incubate at 25°C for another 10 days.

5. Transfer filter paper disks covered with fungal hyphae and conidia to a sterile waxed paper bag. Keep in a desiccator for 2 weeks. When dried, the filter paper cultures can be kept at –20°C.

3.6. TAIL-PCR for Recovering Flanking Sequences of the Integrated T-DNA

Various approaches, including inverse PCR, plasmid rescue, and TAIL-PCR, can be used to isolate genomic sequences flanking the integrated T-DNA. We found TAIL-PCR (21) is reliable and effective in *M. oryzae*. Table 1 lists the T-DNA specific primers and degenerate TAIL-PCR primers used in our lab.

1. Extracting genomic DNA from transformants with the CTAB protocol (22).
2. Set up the following primary TAIL-PCR.

10× Taq buffer	2.0 µl
Taq DAN polymerase	0.5 µl
dNTP mix (10 mM)	1.0 µl
Primer HS1	0.5 µl
Primer AD1/AD2/AD3	2.0 µl
Genomic DNA (about 30 ng/µl)	2.0 µl

Add ddH_2O up to 20 µl.

3. Run the primary PCR program as presented in Table 2.
4. Set up the following secondary TAIL-PCR.

10× Taq buffer	3.0 µl
Taq DAN polymerase	0.5 µl
dNTP mix (10 mM)	1.0 µl
Primer HS2	0.5 µl
Primer AD1/AD2/AD3	3.0 µl
Primary PCR product (diluted to 1/100)	2.0 µl

Add ddH_2O up to 30 µl.

5. Run the secondary PCR program as presented in Table 2.
6. If tertiary TAIL-PCR is necessary, set up the PCR similar to that of secondary PCR. Replace primer HS2 with primer HS3 and use diluted secondary TAIL-PCR product (1/100 dilution in DW) as the template.
7. Run the tertiary PCR program with the same conditions as the secondary TAIL-PCR (Table 2).
8. Separate TAIL-PCR products on 1% agarose gene in 1× TAE. Individual DNA bands are purified with the PCR purification kit (Biomed, Beijing, China) and sequenced with an ABI 3730 sequencer (Sunbio, Beijing, China).

Table 2
Programs of TAIL-PCR

Program	Number of cycles	PCR condition
Primary	1	94°C 2 min, 95°C 1 min
	5	94°C 30 s, 65°C 1 min, 72°C 2 min
	1	94°C 30 s, 30°C 1 min, 41°C 1 min, 52°C 2 min
		63°C 2 min, 72°C 2 min
	15	94°C 30 s, 68°C 1 min, 72°C 2 min, 94°C 30 s
		68°C 1 min, 72°C 2 min, 94°C 30 s, 44°C 1 min, 72°C 2 min
	1	72°C 5 min
Secondary	1	94°C 2 min, 95°C 1 min
	15	94°C 30 s, 68°C 1 min, 72°C 2 min, 94°C 30 s
		68°C 1 min, 72°C 2 min, 94°C 30 s, 44°C 1 min, 72°C 2 min
	1	72°C 5 min

9. Search the *M. oryzae* genome sequence (http://www.broadinstitute.org/annotation/genome/magnaporthe_grisea) with rescued flanking sequences to identify the integration site of T-DNA.
10. The same approach described above (steps 1–9) can be used to recover the right flanking sequences with specific primers RHS1, RHS2, and RHS3 (Table 1). When both left and right flanking sequences of the integrated T-DNA are identified, they should align to the same genomic region in *M. oryzae* if there is no deletion or unspecific PCR amplification.
11. If necessary, the integration event can be further proved by Southern blot analysis and complementation assays for individual transformants. Figure 2 shows the colony morphology ATMT transformant SX11404 (a slow growth mutant), gel electrophoresis of the TAIL-PCR product amplified from SX11404 with the left border primers, integration event in SX11404, and Southern blot analysis (23).

We have used the ATMT approach to generate over 70,000 transformants of P131. Based on preliminary analysis with random-selected transformants, more than half of the ATMT transformants has a single copy integration of T-DNA. Among these transformants, over 700 transformants that have hyphal

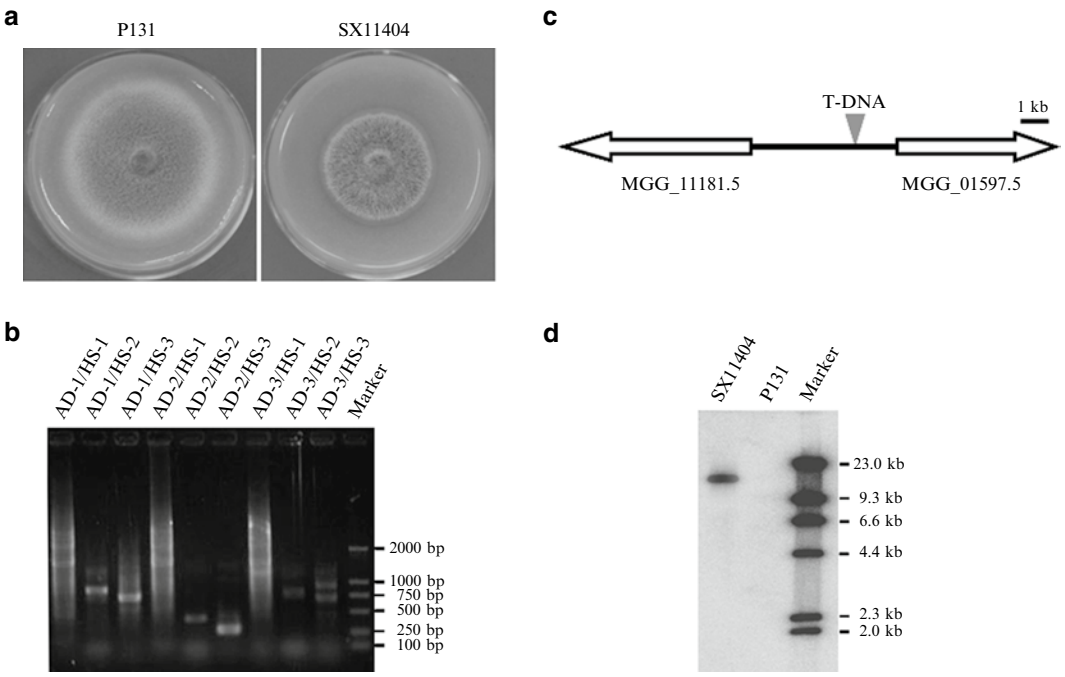

Fig. 2. ATMT mutant SX11404. (**a**) Colonies formed by the wild-type P131 and mutant SX11404 on OTA plates 5 days after inoculation. (**b**) Gel electrophoresis of the primary, secondary, and tertiary TAIL-PCR products amplified from genomic DNA of SX11404 with the left border primers. (**c**) Sequence analysis of the rescued flanking sequences revealed that T-DNA was inserted between two predicted genes in SX11404. (**d**) Southern blot of *Hin*dIII-digested SX11404 genomic DNA hybridized with the *hph* gene. Only SX11404 had one hybridization band.

growth defects were isolated. We have used the TAIL-PCR approach to recover the flanking border sequences from over 340 transformants. Figure 2 is an example to show the insertion event and related analysis of one ATMT transformant SX11404 that is disrupted between predicted genes MGG_01597.5 and MGG_11181.5.

4. Notes

1. The competent cells of *A. tumefaciens* can be frozen in liquid nitrogen and stored at −70°C. However, transformation efficiency is reduced when cells are stored longer than 1 month. We strongly recommend to use *A. tumefaciens* cells prepared within 2 weeks before transformation.
2. The OTA medium should be at least 6 mm thick in Φ9-cm petri plates.
3. Fresh spores (produced within 48 h) are necessary for high transformation efficiency. We noticed that transformation efficiency sharply declines when spores older than 4 days are used.

4. Normally, about 10^8 conidia can be harvested from each OTA plate.
5. We got high transformation efficiency with 2-day-old *A. tumefaciens* after streaking, but opposite effect appeared when prolonged to 3 days.
6. It usually takes about 24 h before OD_{600} reaching 1.2. OD_{600} up to 1.6 is acceptable. Older cultures ($OD_{600} > 1.6$) reduce transformation efficiency.
7. MES is a kind of acid organic buffers that functions between pH 5.0 and 6.0. The optimal pH for *M. oryzae* ATMT transformation is between pH 5.3 and 5.5. Because the cocultivation pH condition is mainly determined by MES, it is very important to exactly adjust pH of MES.
8. For induction, higher than 200 μg/ml AS concentration leads to multiple insertions of T-DNA.
9. OD_{600} of *A. tumefaciens* cells in IM before AS induction should be between 0.15 and 0.20. Higher OD_{600} results in excessive strain growth and cause problems for cefotaxime to kill bacterial cells after cocultivation. On the other hand, lower OD_{600} results in lower transformation efficiency. Under these conditions, OD_{600} normally reaches 0.35–0.45 after 6 h incubation.
10. We found that cocultivation at 23°C for 3 days is better for higher transformation efficiency.
11. We strongly recommend using half-CM medium for the primary selection to reduce background growth.
12. Using this protocol, we usually achieved high transformation efficiency. On average, 400–700 transformants were obtained from one cocultured filter paper (9 cm in diameter) (2,000–3,500 transformants per 10^6 spores). However, if the number of transformants in one plate is more than 200, it is necessary to decrease the mixture volume of *A. tumefaciens* cells and *M. oryzae* spores for cocultivation or cut the cocultivation filter paper into more stripes and overlay them in more plates. In order to avoid coalescence between the transformants, we usually overlay stripes of one cocultivation filter paper in two or three plates.

Acknowledgments

This work was supported by the 973 program (Grant No. 2006CB101901) and by the Innovative Project of State Key Laboratory for Agrobiotechnology (Grant No. 2008SKLAB0103) to Y. -L. Peng from the Ministry of Sciences and Technology, China.

References

1. Dean, R.A., Talbot, N.J., Ebbole, D.J., Farman, M.L., Mitchell, T.K., Orbach, M.J., et al. (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature* **434**, 980–986.
2. Galagan, J.E., Henn, M.R., Ma, L.J., Cuomo, C.A., and Birren, B. (2005) Genomics of the fungal kingdom: insights into eukaryotic biology. *Genome Res.* **15**, 1620–1631.
3. Xu, J.R., Peng, Y.L., Dickman, M.B., and Sharon, A. (2006) The dawn of fungal pathogen genomics. *Annu. Rev. Phytopathol.* **44**, 337–366.
4. Wilson, R.A., and Talbot, N.J. (2009) Under pressure: investigating the biology of plant infection by *Magnaporthe oryzae*. *Nature Rev. Microbiol.* **7**, 185–195.
5. Weld, R.J., Plummer, K.M., Carpenter, M.A., and Ridgway H.J. (2006) Approaches to functional genomics in filamentous fungi. *Cell Res.* **16**, 31–44.
6. Hamer, L., DeZwaan, T.M., Montenegro-Chamorro, M.V., Frank, S.A., and Hamer, J.E. (2001) Recent advances in large-scale transposon mutagenesis. *Curr. Opin. Chem. Biol.* **5**, 67–73.
7. Maier, F.J. and Schäfer, W. (1999) Mutagenesis via insertional- or restriction enzyme-mediated integration (REMI) as a tool to tag pathogenicity related genes in plant pathogenic fungi. *Biol. Chem.* **380**, 855–864.
8. Mullins, E.D. and Kang, S. (2001) Transformation: a tool for studying fungal pathogens of plants. *Cell Mol. Life Sci.* **58**, 2043–2052.
9. Caroline, B., Michielse, A.E., Paul, J.J., Hooykaas-Cees, A.M., van den Hondel, J.J., and Ram, F.J. (2005) *Agrobacterium*-mediated transformation as a tool for functional genomics in fungi. *Curr. Genet.* **48**, 1–17.
10. Zhu, J., Oger, P.M., Schrammeijer, B., Hooykaas, P.J., Farrand, S.K., and Winans, S.C. (2000) The bases of crown gall tumorigenesis. *J. Bacteriol.* **182**, 3885–3895.
11. Hoekema, A., Hirsch, P.R., Hooykaas, P.J., and Schilperoort, R.A. (1983) A binary vector strategy based on separation of Vir-region and T-region of the *Agrobacterium tumefaciens* Ti-plasmid. *Nature* **303**, 179–180.
12. Alonso, J.M., Stepanova, A.N., Leisse, T.J., Kim, C.J., Chen, H.M., Shinn, P., and *et al.* (2003) Genome-wide insertional mutagenesis of *Arabidopsis thaliana*. *Science* **301**, 653–657.
13. Jeon, J.S., Lee, S., Jung, K.H., Jun, S.H., Jeong, D.H., Lee, J., and *et al.* (2000) T-DNA insertional mutagenesis for functional genomics in rice. *Plant J.* **22**, 561–570.
14. Bundock, P., Dulk-Ras, A., Beijersbergen, A., and Hooykaas, P.J. (1995) Trans-kingdom T-DNA transfer from *Agrobacterium tumefaciens* to *Saccharomyces cerevisiae*. *EMBO J.* **14**, 3206–3214.
15. Groot, M.J., de Bundock, P., Hooykaas, P.J., and Beijersbergen, A.G. (1998) *Agrobacterium tumefaciens*-mediated transformation of filamentous fungi. *Nature Biotech.* **16**, 839–842.
16. Rho, H.S., Kang, S., and Lee, L.H. (2001) *Agrobacterium tumefaciens*-mediated transformation of the plant pathogenic fungus *Magnaporthe grisea*. *Mol. Cells* **12**, 407–411.
17. Tsuji, G., Fujii, S., Fujihara, N., Hirose, C., Tsuge, S., Shiraishi, T., and Kubo, Y. (2003) *Agrobacterium tumefaciens*-mediated transformation for random insertional mutagenesis in *Colletotrichum lagenarium*. *J. Gen. Plant Pathol.* **69**, 230–239.
18. Frisch, D.A., Harris-Hailer, L.W., Yokubaitis, N.T., Thomas, T.L., Hardin, S.H., and Hall, T.C. (1995) Complete sequence of the binary vector Bin 19. *Plant Mol. Biol.* **27**, 405–409.
19. Hellens, R., Mullineaux, P., and Klee, H. (2000) A guide to *Agrobacterium* binary Ti vectors. *Trends Plant Sci.* **5**, 446–451.
20. Peng, Y.L., and Shishiyama, J. (1988) Temporal sequence of cytological events in rice leaves infected with *Pyricularia oryzae*. *Can. J. Bot.* **66**, 730–735.
21. Liu, Y.G., Mitsukawa, N., Oosumi, T., and Whittier, R.F. (1995) Efficient isolation and mapping of *Arabidopsis thaliana* T-DNA insert junctions by thermal asymmetric interlaced PCR. *Plant J.* **8**, 457–462.
22. Doyle, J.J., and Doyle, J.L. (1987) A rapid DNA isolation procedure for small quantities of fresh leaf tissue. *Phytochem. Bull.* **19**, 11–15.
23. Sambrook, J., Fritsch, E.F., and Maniatis, T. (1988) Molecular cloning: a laboratory manual. Cold Spring Harbor Laboratory Press. Cold Spring Harbor, New York.

Chapter 17

Molecular Methods for Studying the *Cryphonectria parasitica* – Hypovirus Experimental System

Angus L. Dawe, Rong Mu, Gloricelys Rivera, and Joanna A. Salamon

Abstract

The interaction of the filamentous fungal plant pathogen *Cryphonectria parasitica* with its virulence-attenuating viruses provides a unique platform to explore the molecular biology and genetics of virus–host interactions. Following the development of transformation procedures for this fungus, subsequent advances include infectious cDNA clones of several members of the *Hypoviridae* and an imminently complete fungal genome project. Presented here are basic protocols for growth of the organism and the extraction of DNA, RNA, and protein. Additionally, two further protocols are provided for investigations of host protein phosphorylation and for viral genome secondary structure.

Key words: Hypovirulence, Protein phosphorylation, Phosphatase, RNA secondary structure, RNase digestion

1. Introduction

This chapter focuses on techniques used in the study of *Cryphonectria parasitica*, a filamentous fungal plant pathogen, and its associated virulence-attenuating mycovirus. *C. parasitica* is a member of the phylum Ascomycota and the causative agent of chestnut blight. First observed in the USA in the early part of the twentieth century (1), the fungus rapidly spread throughout the natural range of *Castanea dentata*, the American chestnut, resulting in the near-eradication of this species. The blight also appeared in Europe during the 1930s (2), affecting the European chestnut, *Castanea sativa*. However, the observation of healing trees (3) led to the isolation of hypovirulent strains of *C. parasitica* (4) that were subsequently shown to contain a double-stranded RNA

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722, DOI 10.1007/978-1-61779-040-9_17,

(dsRNA) species (5, 6) ultimately recognized as a new family of mycoviruses, the *Hypoviridae* (7).

The single most striking phenotype that these viruses impose upon the fungal host is a reduction in pathogenesis. Additional phenotypes are noted in laboratory cultures of hypovirulent strains, including reduced pigmentation, asexual sporulation and radial growth rate, altered colony morphology and female sterility (reviewed by ref. 8). The phenomenon of hypovirulence provides potential for biological control of the chestnut blight fungus (9). The hypovirus genome of the CHV1 species consists of a 12.7 kb dsRNA molecule (10) expressed in two open reading frames. The smaller ORF A encodes p69, which gives rise to two polypeptides, p29 and p40 via an autocatalytic event (11). The polyprotein of ORF B is less well characterized, with a single event known to liberate p48 from the N-terminus (10). Other mature protein products are as yet uncharacterized. Most closely related to positive-strand RNA viruses of the potyvirus group (12), the family *Hypoviridae* represent a unique experimental system that permits in-depth analysis using molecular tools to genetically modify both a host (*C. parasitica*) and its parasite (the mycovirus) and to observe further interactions with a third organism (the chestnut). It is then possible to examine the effects of these changes on both the host phenotype and virulence on the chestnut. A protoplast-based protocol for the transformation of *C. parasitica* was developed in 1990 (13) and is essentially unchanged. Presented here are basic methods to approach different aspects of the molecular biology of the mycovirus–host interaction (general procedures for isolation of protein, DNA and RNA) and two more detailed protocols we have used to examine host protein phosphorylation and viral genome secondary structure.

2. Materials

2.1. Growth and Harvesting of Fungal Strains (see Note 1)

1. Strains of *C. parasitica* including the most widely used "wild type" strain (EP155) and its isogenic counterpart infected with hypovirus CHV1-EP713 are available from the ATCC culture collection, #38755 and #52571, respectively.
2. Potato dextrose broth (PDB), 24 g/l.
3. Potato dextrose agar (PDA), 37 g/l.
4. A flat surface that can be illuminated on an approximately 12 h light/dark cycle and at an irradiance level of approximately 20–60 μmol/s/m^2. This can be verified with a radiometer (e.g., model HD 2302, Hotek Technologies). Room temperature is usually sufficient.
5. Hand-held tissue homogenizer (e.g., Polytron PT1600E, Kinematica Inc).

6. Cellophane circles, cut by hand to match the diameter of the Petri plates being used. These should be first submerged in water and then autoclaved. Store submerged at 4°C.
7. Miracloth (EMD Biosciences) cut to fit Buchner funnel. For procedures requiring additional culturing after filtration, Miracloth can be autoclaved if wrapped lightly in aluminum foil. Autoclaving is not generally necessary for procedures where cell lysis immediately follows harvesting.

2.2. Extraction of Proteins (see Notes 2–4)

1. Extraction Buffer: 100 mM Tris–HCl, pH 8.0, 1 mM EDTA, 150 mM NaCl store at room temperature.
2. For protein isolation, prior to use add DTT to 10 mM, CHAPS to 1% and yeast protease inhibitor cocktail (Sigma-Aldrich) at 50 µl/g of mycelium.
3. Liquid nitrogen.
4. Mortar and pestle.

2.3. Extraction of Total RNA

1. For total RNA isolation, use the Plant RNA Isolation Aid, RNAqueous and DNA*free* treatment kits (all from Applied Biosystems/Ambion; see Notes 5 and 6).

2.4. Extraction of Genomic DNA

1. For genomic DNA isolation, supplement Extraction Buffer in Subheading 2.2, item 1 with Triton X-100 or SDS at 2% in place of the components in Subheading 2.2, item 2 and prepare additional reagents:
   (a) Buffer-saturated phenol/chloroform/isoamyl alcohol (P/C/IAA), pH 7.9.
   (b) Chloroform.
   (c) 3 M sodium acetate, pH 5.2.
   (d) 100% ethanol; 70% (by volume) ethanol in water.
   (e) TE buffer (10 mM Tris–HCl, 1 mM EDTA, pH to 7.0 with HCl).
   (f) RNase Cocktail (Applied Biosystems/Ambion).

2.5. Dephosphorylation and Casein Kinase II-Mediated Rephosphorylation

1. Calf Intestinal Alkaline Phosphatase (CIAP) (Invitrogen).
2. Adenosine triphosphate.
3. Phosphatase inhibitor sodium orthovanadate (Sigma-Aldrich).
4. CK2 inhibitor DMAT, 2-dimethylamino-4, 5, 6, 7-tetrabromo-1H-benzimidazole (EMD Biosciences).
5. Whole protein extracts of strains to be tested.

2.6. In Vitro Transcription of Viral cDNA Clones, RNase Treatment and Sequencing

1. Plasmids bearing cDNA clones of the hypoviruses CHV1-EP713 (pLDST; (14)) or CHV1-Euro7 (pTE7; (15)) kindly provided by Don Nuss, University of Maryland Biotechnology Institute.

2. *Spe*I restriction endonuclease, BSA (10 mg/ml) and NEBuffer 2 (New England Biolabs).
3. AmpliCap™ T7 and SP6 high yield message maker kit components (Epicenter Biotechnologies).
4. RNeasy Mini Kit components (QIAGEN).
5. RNase T1, RNase A, RNase I, and RNase V1 (Applied Biosystems/Ambion).
6. Water baths at 50 and 65°C.
7. GlycoBlue coprecipitant (Applied Biosystems/Ambion) for visualization of the RNA pellet.
8. 80% ethanol (by volume) in RNase-free water.
9. Eppendorf Vacufuge concentrator.
10. IR-700 labeled primer, 5′-CCACTGTAGTAGGATCAAC-3′ (see Note 7; Li-Cor Biosciences).
11. SuperScript III Reverse Transcriptase, RNaseOUT RNase inhibitor, First-strand buffer, dithiothreitol (DTT; 0.1 M), dNTP mixture (10 mM each dATP, dGTP, dTTP, and dCTP), all from Invitrogen.
12. Sequitherm Excell II DNA sequencing kit (Epicenter Technologies).
13. Access to a Li-Cor 4200 DNA sequencer.

3. Methods

3.1. Growing and Harvesting Mycelium

1. Solid-medium cultures are grown by inoculating a standard Petri dish containing PDA medium with a small (~2 mm × 2 mm) plug from the actively growing edge of a previously grown culture. Plates are best grown under the conditions noted above regarding temperature and light.
2. For harvesting solid-grown mycelium, recovery is easiest if autoclaved cellophane is first placed on the surface of the medium after it has solidified. Once inoculated, the mycelium grows on the cellophane but does not penetrate into the agar. The fungal tissue can then be recovered easily by scraping the mycelium off of the cellophane.
3. For liquid cultures, several plugs of mycelium from a Petri plate are used to inoculate approximately 10 ml of potato dextrose broth. The cultures are left stationary at room temperature, but may be agitated daily with vigorous shaking or brief vortexing. After 3–4 days, the fungal mass should be homogenized using a handheld homogenizer to break up the mycelial clumps. An equal volume of fresh medium is then added and the culture incubated an additional 3–4 days.

4. Harvesting of the liquid cultures is best achieved by filtration. Place four layers of Miracloth into the Buchner funnel and slowly pour the culture onto the surface while applying a vacuum. The resulting mycelial pad can be washed with water and then compressed between paper towels to remove excess moisture.

3.2. Extraction of Proteins

1. Pulverize harvested mycelium under liquid nitrogen with a pestle and mortar, then carefully weigh. Transfer the ground mycelium to microfuge tube(s) and add 1.5 times the volume by weight (e.g., 100 mg ground mycelium requires 150 µl) of freshly prepared protein extraction buffer with DTT, CHAPS, and protease inhibitors.
2. Vortex vigorously and incubate on ice for 15 min. Repeat vortexing and incubate a further 15 min on ice.
3. Pellet the cellular debris for 5 min at 4°C and maximum speed in a microcentrifuge.
4. Carefully pipet the clear protein lysate and transfer to a clean microfuge tube. Avoid aspirating the pellet. If the lysates are not clear, repeat the centrifugation step.
5. Determine protein concentration using the Bradford Assay system from BioRad. This assay is not sensitive to the concentrations of CHAPS used in this extraction method.
6. The samples can be stored at –20°C for short-term use, but appear more stable at –80°C for longer term.

3.3. Extraction of Total RNA

1. For extraction of RNA, begin by harvesting and pulverizing the mycelium as described for the preparation of proteins. Resuspend the powdered tissue in the lysis/binding solution from the RNAqueous kit and add the Plant RNA Isolation Aid. Use a ratio of 0.2 g ground mycelium per 1.6 ml lysis/binding solution and 0.2 ml Plant RNA Isolation Aid.
2. Continue with the RNAqueous kit exactly as described in the manufacturer's literature. Elute the RNA in 60 µl elution solution preheated to 80°C. Remove any DNA from the recovered sample by following the DNA*free* rigorous treatment protocol. Perform this procedure twice to ensure the removal of all DNA.
3. After DNA digestion, validate quality of the preparation by checking on a spectrophotometer at 260/280 nm. Based on the values obtained, load approximately 1–2 µg on a standard 0.8–1% agarose gel. Be sure to first clean the gel box thoroughly with RNaseZap solution. When visualized by staining with Ethidium Bromide or SYBR Safe (Invitrogen), the two bands corresponding to the ribosomal RNA species should be clearly visible and any degradation products (seen as a

fast-migrating species at the bottom of the gel) should be minimal or absent. Use immediately or store at –80°C.

3.4. Extraction of Genomic DNA

1. For extraction of genomic DNA in volumes convenient for microfuge tubes, resuspend approximately 100 µg powdered mycelium in 1.5 volumes of cold Extraction Buffer with Triton X-100 and vortex vigorously.
2. Add an equal volume of alkaline Phenol/Chloroform/Isoamyl alcohol, vortex vigorously, and centrifuge in a microfuge at maximum speed for 5 min at 4°C. Recover the upper aqueous layer and repeat the extraction with P/C/IAA before performing a third extraction, but with an equal volume of chloroform alone.
3. Carefully aspirate the aqueous layer, add 1/10 volume of 3 M sodium acetate, pH 5.2, and 2 volumes of 100% ethanol. Precipitate for at least 30 min (but may be left overnight at –20°C). Pellet DNA in a microfuge for 20 min at maximum speed and 4°C. Aspirate the ethanol with vacuum or a pipettor and allow the pellet to air-dry. Carefully rinse the pellet with 70% ethanol to reduce the salt content, and repellet.
4. Resuspend the pellet in 100 µl TE buffer and add 5 µl of the RNase Cocktail. Incubate 30 min at room temperature. Verify the DNA yield by spectrophotometry.

3.5. Dephosphorylation and Rephosphorylation of Proteins

Phosphorylation by Protein Kinase 2 (CK2) is a common modification that influences a wide array of cellular signal transduction pathways. To confirm whether there are any physiologically relevant CK2 phosphorylation sites within a protein, the covalently bound phosphates can first be removed by in vitro treatment with Calf Intestinal Alkaline Phosphatase. This enzyme is the same routinely used for modification of DNA during certain cloning procedures and is purchased from the supplier (Invitrogen) prepared for that purpose. We have analyzed the phosphorylation state of the phosducin-like protein BDM-1 using antibodies raised specifically against this protein for the western blot. While this protocol was designed to analyze the modification of BDM-1 by CK2, if the protein of interest is suspected to be the target of a different kinase, other pharmacological agents are available that could be applied.

1. For protein dephosphorylation, dilute protein lysate to a concentration of 1 µg/µl with dilution buffer supplied by CIAP enzyme manufacturer. Aliquot 35 µl of protein lysate, 4 µl of 10× reaction buffer provided with the enzyme and 1 µl (20,000 U) of CIAP. Incubate the reaction mixture for 30 min at 37°C.
2. Dephosphorylated proteins are rephosphorylated by the addition of equal volume of protein extract in the presence of

1 mM ATP, 100 mM sodium orthovanadate (to inhibit residual CIAP activity), and 20 μM specific Casein Kinase II (CK2) inhibitor DMAT (16). The reaction is performed at room temperature in the dark for 22 h.

3. The presence or absence of the charged phosphate moiety on the protein of interest should be detectable as differential migration after conventional polyacrylamide electrophoresis and transfer to a nylon membrane for western blotting. For BDM-1, we have successfully used NuPAGE 10% Bis-Tris Gels (Invitrogen) in MOPS running buffer (50 mM MOPS, 50 mM Tris base, 0.1% SDS, 1 mM EDTA, pH 7.7) and a BioRad transblot semidry apparatus for transfer to Immobilon-P nylon membranes from Millipore. An affinity-purified polyclonal antiserum raised in rabbits against the entire protein and a horseradish peroxidase-conjugated anti-rabbit secondary antibody from BioRad completes the blotting procedure. Optimal separation parameters to view the subtle changes in migration may vary according to protein analyzed.

3.6. RNase Mapping of Structural Features in the Hypovirus Genome

Structural features of viral RNA genomes have been shown to be important for aspects of translation and viral genome replication (17, 18). In an effort to better understand the components of the hypovirus genome that are required for maintenance of the hypovirus-infected phenotype in the host mycelium, we have modified available protocols for RNA structure analysis that analyze the products of specific degradation by individual RNase activities. This approach has several distinct stages: the generation of an in vitro transcript from an available cDNA clone, RNase digestion, and reverse transcription of the resulting RNA fragments resulting in the incorporation of a labeled primer.

1. Linearize the cDNA clone by digesting 2.5 μg DNA with *Spe*I in a 10 μl reaction with 1 μl of NEBuffer 2 and 1 mg/ml BSA at 37°C for 90 min. Confirm the concentration of the product by spectrophotometer after linearization.
2. Using the materials provided in the AmpliCap™ T7 and SP6 high yield Message Maker kit, transcribe 1 μg of the linearized cDNA clone in a reaction volume that includes 2 μl Amplicap-Max transcription buffer, 2 μl 100 mM DTT, 8 μl Amplicap-Max Cap/NTP Premix, 2 μl Amplicap-Max T7 enzyme and water to 20 μl. Incubate at 37°C for 2 h.
3. Use the RNeasy mini kit to purify the transcripts, eluting in 30 μl RNase-free water. Check the concentration of the RNA by spectrophotometer, then store at –80°C.
4. Dilute the RNases in water to the following concentrations: 0.1 U/μl of RNase T1, 0.02 U/μl of RNase A, 4 U/μl of RNase I, and 0.01 U/μl of RNase V1 (see Note 8).

5. Add 4 µg RNA sample from step 3, 4 µl of 10× RNA structure buffer and 0.3 µl yeast RNA to a microfuge tube. Add RNase-free water to make the total volume 40 µl. Incubate at 65°C for 2 min, and then cool to room temperature.
6. Divide the 40 µl from step 5 into four microfuge tubes. Add 1 µl of a different diluted RNase to each tube and incubate at room temperature for 10 min. Stop the reaction by addition of inactivation/precipitation buffer to each tube.
7. Add 2 µl Glycoblue to each tube and incubate the tubes at –20°C for 15 min before pelleting in a microcentrifuge at maximum speed for 15 min.
8. Discard the supernatant. Add 200 µl 80% ethanol to each pellet, then centrifuge as in **step** 7. Carefully aspirate the ethanol and dry the pellets for 20 min under vacuum in a Vacufuge concentrator. Resuspend the dried pellets in 10 µl of RNase-free water.
9. Add reverse transcription reagents to each tube and make the total volume 20 µl: 1 µl RNase inhibitor, 5 µl 5× First-strand buffer, 1 µl DTT (0.1 M), 1 µl dNTP mixture (10 mM each), 1 µl Licor primer (5 pM), and 1 µl Superscript III Reverse Transcriptase. Cover tubes with foil and incubate at 50°C for 1 h. The reactions are stored at –20°C until analysis.
10. Control sequencing reactions are prepared by adding 1 µg cDNA clone (plasmid or linearized), 1 µl of Licor primer (2.5 pM), 7.2 µl of 3.5× buffer, and 1 µl of polymerase from the Sequitherm sequencing kit, with RNase-free water to a total volume 20 µl to a microfuge tube. Keeping the tube on ice, divide the 20 µl mixture into four thin-wall PCR tubes and add 2 µl of terminator mix A (or T or C or G) from the sequencing kit to each PCR tube.
11. Perform the PCR with the following parameters: 95°C for 3 min followed by 30 cycles of 95°C for 30 s, 53°C for 30 s, and 70°C for 1 min. Add 3 µl of the Stop solution provided and store the reactions at –20°C wrapped in foil.
12. The individual RNase-treated reactions, the control sequencing reaction and untreated negative controls must now be separated using the Li-Cor sequencer according to the protocols and guidelines of the facility you are working with.
13. From the resulting gel image, the location of the bands in the RNase treatment lanes represents a cleavage event. These are located in the sequence as a whole by reading the sequence control reaction from the bottom of the gel upward. Cleavage by RNase V1 occurs at double-stranded nucleotides, by RNase I at any single-stranded nucleotides, by RNase A at single-stranded C or U and by RNase T1 at single-stranded G. A small portion of a typical gel is shown in Fig. 1. By locating the sites at which there is a known

feature as defined by the RNase product, it is possible to constrain a model for the secondary structure of the RNA molecule using experimentally determined characteristics.

3.7. Analyzing Secondary Structural Motifs in the Hypovirus Genome: Mfold Prediction

1. Go to Web server http://mfold.bioinfo.rpi.edu/cgi-bin/rna-form1-2.3.cgi (19) to input the RNA sequence (see Note 9).
2. For *C. parasitica* hypoviruses, use the prediction temperature of 25°C, the closest approximation to the actual temperature at which the organism is grown in the laboratory. Other settings are left as defaults. Do not enter any constraint information at this time.
3. After selecting the "fold RNA" button at the bottom of the screen, the input sequence may generate many alternative secondary structure prediction results. Choose the lowest free energy for the most stable (likely) structure. This represents the baseline structure assuming no constraints imposed by the interaction of individual bases (see Fig. 1a).
4. Based on the RNase mapping results, repeat steps 1–3 above but include constraints that do not match the experimental observations concerning the locations of known single-stranded regions. Repeat the folding analysis to generate a secondary structure that includes experimentally validated constraints (see Fig. 1 for an example).

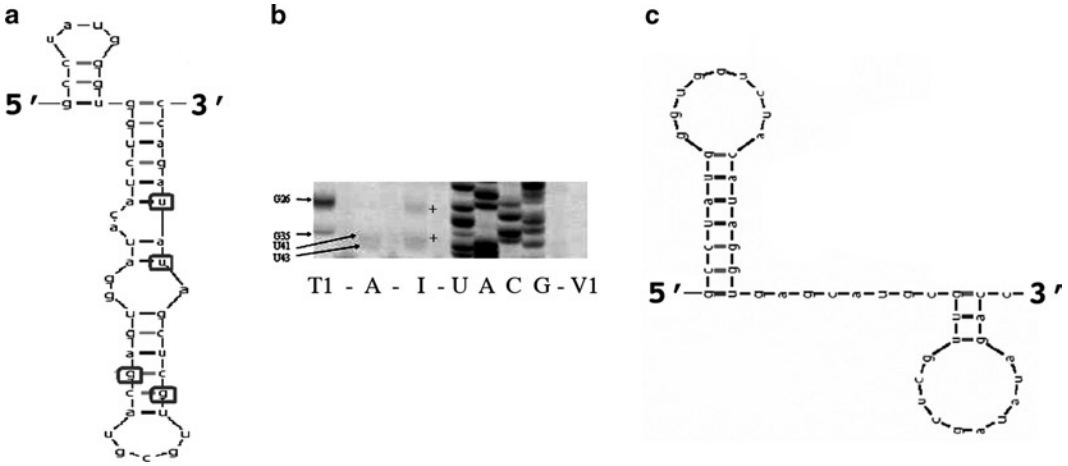

Fig. 1. An example of experimental constraints applied to a predicted RNA structure. (**a**) This represents the first 48 nucleotides of the 5′ and of the untranslated region of CHV1-Euro7, as predicted by mFold with no constraints. The *boxed* residues indicate, where the predicted structure contradicts the experimental analysis. (**b**) The RNase digest shows that the template was digested by RNase T1 at G26 and G35, and by RNase A at U41 and U43. These cleavage sites are also supported by the bands seen in the RNase I digest, indicated by (+). *Lanes T1*, *A*, *I*, and *V1* refer to the RNases used. *Lanes U*, *A*, *G*, and *C* refer to the control sequencing reaction. Intervening lanes (–) represent primer extension controls. (**c**) The resulting mFold prediction when the *marked* paired nucleotides from (**a**) are constrained to be unpaired.

4 Notes

1. Growth conditions can affect the phenotype of *C. parasitica*, particularly excessive light or heat. Generally, stable colony morphologies can be maintained at about 21–24°C. Under higher light intensity than indicated, sporulation and pigmentation may be increased. However, this can begin to ameliorate some effects of virus infection, as noted by Hillman et al. (20).
2. All solutions and media should be prepared with purified water (18.2 MΩ cm). Additionally, for critical procedures and particularly those involving RNA, RNase-free water should be used. This is easily obtained from many suppliers (e.g., Applied Biosystems, Sigma, Fisher, etc.) but can also be home-made according to protocols found in Sambrook et al. (21).
3. Traditionally, DNA, RNA and protein extracts of *Cryphonectria* have been generated by the method presented here – using liquid nitrogen to freeze the sample followed by manual grinding. However, the use of lyophilized mycelium has significant promise and may provide a viable alternative, especially when working with a large number of samples. After harvesting, mycelial samples (liquid or solid medium grown) can be lyophilized overnight following which the mycelium is easily powdered by agitation at room temperature using a pipet tip, microhomogenizer or by vortexing with acid-washed glass beads. The extraction procedure then continues as described above.
4. The method presented here is derived from an analysis of the suitability of different extraction methods for recovering G-protein signaling components described by Parsley et al. (22). Alternative extraction buffers may be found in that paper.
5. For all RNA isolation and handling procedures, be sure to clean the work area thoroughly with an RNase-decontaminating agent (e.g., RNase Zap, Applied Biosystems). Wear gloves and change regularly, especially if the gloves are removed during a pause in the process. Use microfuge tubes and pipet tips that are certified as "DNase and RNase free" by the manufacturer.
6. *Cryphonectria* mycelium, particularly in liquid culture, produces excessive carbohydrates that interfere with many column-based isolation protocols. The kit indicated is designed to isolate RNA from plant material and works well if all guidelines concerning biomass are observed. Other kits may also be equally effective.

7. The primer should be designed to anneal at least 40–50 nucleotides from the target sequence to be analyzed. The primer described above is suitable for examining the 5′ non-translated regions of the CHV1-EP713 and CHV1-Euro7 hypovirus genomes.
8. The exact concentration of RNase was determined by viewing results after sequencing. Too little RNase results in faint or absent bands, whereas too much precludes the resolution of bands that are close together. Unfortunately, for other viral templates, these values should be taken only as a guide and a certain amount of trial and error is likely required.
9. Alternate versions of the mFold software are available. This link directs the user to an older version of the software, but one where the temperature field is variable. In newer versions of the software, the temperature field is fixed at 37°C.

Acknowledgments

This work was supported by NSF award MCB-0718735 (to ALD). GR was supported by a fellowship from the NMSU-NIH MBRS-RISE program (R25GM061222).

References

1. Merkel, H. W. (1906) A deadly fungus on the American chestnut., *NY Zool. Soc. Annu. Rep.* **10**, 97–103.
2. Biraghi, A. (1946) Il cancro del castagno causato da *Endothia parasitica*., *Ital. Agric.* **7**, 1–9.
3. Biraghi, A. (1953) Possible active resistance to *Endothia parasitica* in *Castenea sativa*., in *11th Rep. Congr. Int. Union For. Res. Org.*, pp 149–157, Rome.
4. Grente, J. (1965) Les formes hypovirulentes d'*Endothia parasitica* et les espoires de lutte contre le chancre du chataigner., *C.R. Seances Acad. Agric. Fr.* **51**, 1033–1037.
5. Day, P. R., Dodds, J. A., Elliston, J. E., Jaynes, R. A., and Anagnostakis, S. L. (1977) Double-Stranded RNA in *Endothia parasitica* [from Europe and North America, Biological Control, Chestnut Blight], *Phytopathology* **67**, 1393–1396.
6. Anagnostakis, S. L., and Day, P. R. (1979) Hypovirulence conversion in *Endothia parasitica*., *Phytopathology* **69**, 1226–1229.
7. Hillman, B. I., Fulbright, D. W., Nuss, D. L., and Van Alfen, N. K. (1995) Hypoviridae, in *Virus taxonomy* (F.A. Murphy, C. M. F., D.H.L. Bishop, S.A. Ghabrial, A.W. Jarvis, G.P. Martelli, M.A. Mayo, and M.D. Summers (ed.), Ed.), pp 261–264, Springer-Verlag, New York, N.Y.
8. Dawe, A. L., and Nuss, D. L. (2001) Hypoviruses and chestnut blight: exploiting viruses to understand and modulate fungal pathogenesis, *Annu Rev Genet* **35**, 1–29.
9. Anagnostakis, S. L. (1982) Biological control of chestnut blight, *Science* **215**, 466–471.
10. Shapira, R., Choi, G. H., and Nuss, D. L. (1991) Virus-like genetic organization and expression strategy for a double- stranded RNA genetic element associated with biological control of chestnut blight, *EMBO J.* **10**, 731–739.
11. Choi, G. H., Pawlyk, D. M., and Nuss, D. L. (1991) The autocatalytic protease p29 encoded by a hypovirulence-associated virus of the chestnut blight fungus resembles the potyvirus-encoded protease HC-Pro, *Virology* **183**, 747–752.
12. Koonin, E. V., Choi, G. H., Nuss, D. L., Shapira, R., and Carrington, J. C. (1991)

Evidence for common ancestry of a chestnut blight hypovirulence- associated double-stranded RNA and a group of positive-strand RNA plant viruses, *Proc. Natl. Acad. Sci. U.S.A* **88**, 10647–10651.

13. Churchill, A. C. L., Ciufetti, L. M., Hansen, D. R., Van Etten, H. D., and Van Alfen, N. K. (1990) Transformation of the fungal pathogen *Cryphonectria parasitica* with a variety of heterologous plasmids., *Curr Genet* **17**, 25–31.
14. Choi, G. H., and Nuss, D. L. (1992) Hypovirulence of chestnut blight fungus conferred by an infectious viral cDNA, *Science* **257**, 800–803.
15. Chen, B., and Nuss, D. L. (1999) Infectious cDNA clone of hypovirus CHV1-Euro7: a comparative virology approach to investigate virus-mediated hypovirulence of the chestnut blight fungus *Cryphonectria parasitica*, *J. Virol.* **73**, 985–992.
16. Pagano, M. A., Meggio, F., Ruzzene, M., Andrzejewska, M., Kazimierczuk, Z., and Pinna, L. A. (2004) 2-Dimethylamino-4,5,6,7-tetrabromo-1H-benzimidazole: a novel powerful and selective inhibitor of protein kinase CK2, *Biochem Biophys Res Commun* **321**, 1040–1044.
17. Romero, T. A., Tumban, E., Jun, J., Lott, W. B., and Hanley, K. A. (2006) Secondary structure of dengue virus type 4 3′ untranslated region: impact of deletion and substitution mutations, *J Gen Virol* **87**, 3291–3296.
18. Van Den Born, E., Gultyaev, A. P., and Snijder, E. J. (2004) Secondary structure and function of the 5′-proximal region of the equine arteritis virus RNA genome, *RNA* **10**, 424–437.
19. Zuker, M. (2003) Mfold web server for nucleic acid folding and hybridization prediction, *Nucleic Acids Res* **31**, 3406–3415.
20. Hillman, B. I., Shapira, R., and Nuss, D. L. (1990) Hypovirulence-associated suppression of host functions in *Cryphonectria parasitica* can be partially relieved by high light intensity., *Phytopathology* **80**, 950–956.
21. Sambrook, J., Fritsch, E. F., and Maniatis, T. (1989) *Molecular cloning: a laboratory manual*, 2nd ed., Cold Spring Harbor Press, Cold Spring Harbor, NY.
22. Parsley, T. B., Segers, G. C., Nuss, D. L., and Dawe, A. L. (2003) Analysis of altered G-protein subunit accumulation in *Cryphonectria parasitica* reveals a third Gα homologue, *Curr Genet* **43**, 24–33.

Chapter 18

Metabolic Fingerprinting in *Fusarium verticillioides* to Determine Gene Function

Jonathon E. Smith and Burton H. Bluhm

Abstract

Fusarium verticillioides is a major pathogen of corn and poses a significant risk to human health by producing mycotoxins that accumulate in kernels. Considerable efforts have focused on identifying genes involved in secondary metabolism and pathogenesis. The availability of a sequenced genome accelerates gene discovery and characterization, but functional genomics approaches are hindered when disruption of a gene results in a phenotype that is not readily distinguishable from the wild type. To address this problem, we developed a metabolomics approach to characterize gene function. The technique involves culturing two fungal strains (wild type and a mutant) under identical conditions, extracting as wide a range of metabolites as possible, analyzing the metabolomes by gas chromatography/mass spectrometry, and comparing the unique metabolic fingerprint of each strain.

Key words: Fungal, Metabolic, Fingerprinting, *Fusarium verticillioides*, Gene function, Metabolomics

1. Introduction

Fungi produce a wide range of metabolites, broadly defined as low molecular weight products of cellular-level reactions. Primary metabolites are essential for growth and survival, such as the intermediates of glycolysis and structural components of the cell, and are generally conserved across fungal taxa. Secondary metabolites, however, are at least conditionally dispensable for survival and serve diverse roles in fungal growth and development (1). Fungal secondary metabolism is complex; the exact number of unique fungal secondary metabolites is impossible to determine, but is estimated to be well in excess of 200,000 (2). Examples of fungal secondary metabolites include toxins involved in host colonization (e.g., victorin, deoxynivalenol, cercosporin) (3), UV-protective

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722,
DOI 10.1007/978-1-61779-040-9_18,

pigments (e.g., mycosporines) (4), antibiotics (e.g., penicillin, cephalosporin) (5), and plant hormones (e.g., abscisic acid, gibberellic acid) (6, 7).

The availability of sequenced fungal genomes is constantly increasing, which greatly accelerates gene discovery and characterization. Gene function can be determined through functional genomics, a cornerstone of which is generating mutants by gene deletion, disruption, or silencing. However, disrupting a gene often does not result in an obvious morphological phenotype, in many cases because the gene regulates aspects of primary and/or secondary metabolism (8).

Metabolomics has emerged as a powerful tool to simultaneously detect and quantify numerous metabolites and thus determine gene function. Currently, two distinct approaches are employed to analyze metabolites: profiling and fingerprinting. Both approaches are based on the simultaneous examination of numerous metabolites through chromatography and mass spectrometry (MS). However, metabolic profiling involves the identification of compounds in order to elucidate a specific metabolic pathway, whereas metabolic fingerprinting aims to identify differences in metabolite production between treatments (9).

We developed a protocol for metabolic fingerprinting in *Fusarium verticillioides*, a mycotoxigenic pathogen of maize. *F. verticillioides* is well-suited for metabolomics because it is important from both an economic and food safety standpoint, it is easily cultured under a range of conditions, its genome has been sequenced, and secondary metabolism has been studied extensively among members of the genus *Fusarium* (10–14). Of these considerations, a sequenced genome is especially important because mutants generated through both forward and reverse genetic approaches can be fingerprinted to determine phenotypic effects.

The protocol we developed requires the creation of a metabolic library based on diverse cultures and known spectra. Then, a metabolic fingerprint consisting of chromatographic peaks and mass spectra of each metabolite is generated for the wild type in a defined set of culture conditions. Mutants (random or defined) are grown under identical culture conditions, fingerprinted, and compared to the wild type. Differences in the fingerprints are either quantitative (i.e., a compound is produced by the mutant and wild type but at significantly different levels) or qualitative (i.e., a compound is produced by one strain but not the other) (Fig. 1). Once components of the metabolic fingerprint are determined to be of interest, the database can be searched to provide information on identity (if known) and production by the wild type in a variety of environmental conditions. In the context of determining gene function, identifying differentially produced compounds provides unique insights.

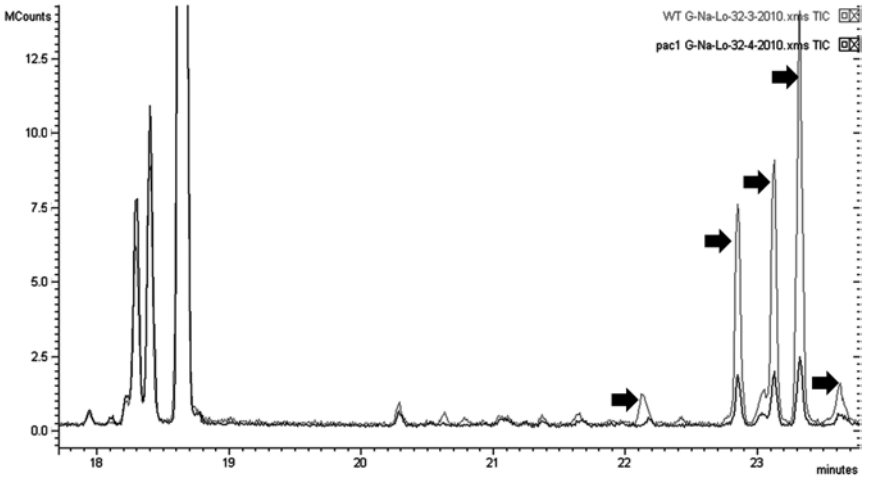

Fig. 1. Portion of an example total ion chromatogram comparing wild type *Fusarium verticillioides* (*light gray*) and the Δ*pac1* mutant (*black*). Differences in metabolite production are evident when comparing the compounds eluting between 22 and 24 min (*arrows*), whereas the compounds eluting between 17 and 19 min are present at nearly identical levels.

2. Materials

2.1. Fungal Strains

- *F. verticillioides* strain 7600 (FRC M3125 = NRRL 20956) (12–14).
- *F. verticillioides Δpac1* (12).

2.2. Culture Media

V8 agar (1 L):180 mL V8 vegetable juice, 2 g $CaCO_3$, 20 g Agar. Add H_2O to 1 L and stir. Autoclave for 40 min and pour into sterile petri dishes.

Liquid growth media (1 L): 10 g Glucose, 1.0 g NH_4NO_3, 1.0 g KH_2PO_4, 0.5 g $MgSO_4 + 7H_2O$, 0.5 g KCl, 0.5 mL 2% $FeSO_4 + 7H_2O$, and 0.2 mL Trace element solution. Add H_2O to 1 L and stir. Adjust to desired pH with either phosphoric acid or KOH and autoclave for 40 min (see Note 1).

Trace element solution (200 mL): 10 g Citric acid + H_2O, 10 g $ZnSO_4 + 7H_2O$, 0.5 g $CuSO_4 + 5H_2O$, 0.1 g $MnSO_4 + H_2O$, 0.1 g H_3BO_3, and 0.1 g $NaMoO_4 + 2H_2O$. Add H_2O to 200 mL and store at 4°C.

2.3. Metabolite Extraction

- Vacuum concentrator (e.g., Savant, Speedvac SVC100H).
- Methanol (HPLC grade or better).
- Isooctane (ReagentPlus grade or better).
- Type 1 H_2O.

2.4. Derivatization

- 1 M HCl in methanol.
- Compressed nitrogen.
- Nitrogen evaporator manifold.
- *TMSI:TMCS 100:1 (1.01 mL)* (make fresh) (see Note 2).
  - 1.0 mL Trimethylsilylimidazole (TMSI).
  - 10 µL Trimethylchlorosilane (TMCS).

2.5. GC–MS Analysis

- Gas chromatograph (e.g., Varian, 450-GC).
- GC column (low bleed 30 m 5% phenyl, e.g., Varian, VF-5 ms).
- Quadrupole mass spectrometer (e.g., Varian, 320-MS GC/MS).
- MS Workstation software (version 6.9.1, Varian).
- Retention index standards (7–40 carbon saturated n-Alkanes e.g., Supelco, 49452-U) (see Note 3).

2.6. Database Creation and Fingerprinting

1. Automatic Mass spectral Deconvoluion and Identification System (AMDIS version 2.6.5, http://chemdata.nist.gov/mass-spc/amdis/).
2. Golm Mass Spectral (MS) and Retention Time Index (RI) Library (Q_MSRI_ID version 2004-03-01, http://csbdb.mpimp-golm.mpg.de/csbdb/gmd/msri/gmd_msri.html).
3. ACD/MS Manager (version 12.0, http://www.ACDlabs.com).
4. ACD/IntelliXtract (version 12.0, http://www.ACDlabs.com).
5. Metabolite Detector (version 1.597, http://metabolitedetector.tu-bs.de/download.html).

3. Methods

3.1. Culture Conditions (See Note 4)

To generate inocula for liquid cultures, *F. verticillioides* wild type and mutants are initially grown on media that promotes the production of microconidia (such as V8 agar). The concentration of microconidia in each inoculum can be adjusted easily for optimal growth and reproducibility.

1. Inoculate V8 plates with a culture of *F. verticillioides*. Incubate at room temperature for 5–7 days.
2. Harvest microconidia with 1 mL of sterile water and a sterile "L" spreader. Adjust concentration to 10^7 microconidia/mL.
3. Inoculate liquid growth media with 100 µL of microconidial suspension (10^6 microconidia) and incubate at 25°C for 10 days (see Note 5).

3.2. Metabolite Extraction

Fungal tissue must be rinsed to minimize the detection of compounds derived from the growth medium. After rinsing, the tissue is ground in liquid nitrogen and stored at –80°C until it can be dried under vacuum. Storage at –80°C prevents breakdown of metabolites by enzymes released from ruptured cells during grinding, and drying improves reproducibility by standardizing tissue weights. Methanol is used to extract metabolites because it solubilizes compounds with varying chemical properties and allows these compounds to be separated from cellular debris.

1. Harvest fungal tissue by filtration with a filter funnel under vacuum. Rinse tissue thoroughly with Type I (ultra-filtered) water to remove residual growth media.
2. Transfer filtered tissue to 15-mL conical tubes resting in liquid nitrogen and store at –80°C until all samples are collected (see Note 6).
3. Grind tissue under liquid nitrogen in a mortar and pestle and transfer to 2-mL microcentrifuge tubes.
4. Place 2-mL tubes with lids open in the speedvac rotor and turn on the centrifuge. When the centrifuge attains top speed, apply vacuum and allow tissue to dry for at least 12 h.
5. Add 6.0 mg of ground tissue to 2 mL of methanol in a 4-mL autosampler vial (see Note 7) and vortex vigorously for 30 s. Sonicate until tissue is fully suspended, usually 30–60 s.
6. Incubate vial overnight at room temperature with gentle agitation on a flatbed shaker.

3.3. Derivatization

GC–MS analysis requires that analytes be in the gaseous phase. The following derivatization method replaces polar hydroxyl groups with nonpolar trimethylsilyl (TMS) groups, making metabolites more volatile and thus more amenable to separation by GC. TMS derivatives are not stable indefinitely and will degrade at room temperature. For this reason, reaction times should not be increased, derivatized samples should be stored in the dark at 4°C, and samples should be analyzed as soon as possible. Samples should not be exposed to moisture or humidity, as hydroxyl groups reverse the derivitization reaction.

1. Centrifuge autosampler vials containing tissue extracts in swinging bucket rotor at 4,000 rpm (~3,200 RCF) for 10 min or until pellet is solid (see Note 8).
2. Transfer 1 mL of the extract to a 2-mL autosampler vial and completely dry contents under nitrogen (see Note 9).
3. Add 100 μL of 1 M HCl, vortex vigorously for 30 s, and sonicate until extracts are resuspended. After incubating for

1 h at 50°C, completely dry contents under nitrogen (see Note 10).

4. Add 100 µL of TMSI:TMCS 100:1 (see Note 2) and vortex vigorously for 30 s. Verify that all interior surfaces of the vial are contacted. Incubate at 37°C for 1 h.
5. Add 100 µL of isooctane and vortex briefly to mix. Isooctane quenches the derivatization reaction and dilutes any remaining TMSI:TMCS so that water can be added without reacting violently or degrading derivatized products.
6. Add 200 µL of Type I water and vortex vigorously. Centrifuge at 3,200 × *g* for 5 min or until aqueous and organic layers are fully separated.
7. Transfer the top (organic) phase to a microvolume autosampler vial insert and discard the aqueous phase.
8. Place the microvolume insert in a 2-mL autosampler vial with cap and analyze by gas chromatograph mass spectrometry (GC–MS).

3.4. GC–MS Analysis (*See Note 11*)

GC–MS methods can be optimized to detect a wide range of fungal metabolites. A limitation of GC–MS is that compounds must be volatile and interact differentially with the analytical column in order to be separated. However, many metabolites can be volatilized by derivatization with TMS reagents, and a wide selection of GC columns is available to separate analytes based on their unique properties.

1. The autoinjector/autosampler should incorporate the following settings: three preinjection rinses with isooctane, one preinjection rinse with sample, three postinjection rinses with isooctane, a split ratio of 5:1, and an injector temperature of 260°C.
2. Set the column flow rate to 1 mL/min with helium as the carrier gas. The column oven temperature program starts at an initial temperature of 120°C with a 5-min hold. The temperature rises at a rate of 4°C/min until it reaches 300°C, where it is held for 15 min for a total run time of 65 min. When the run has ended, the column should be reequilibrated to 120°C and held there for at least 3 min before injecting another sample.
3. The mass spectrometer interface temperature should be 275°C and the ion source should be 200°C. Ensure that the filament is off and data acquisition is delayed for 3 min to prevent damage to the filament and overloading of the detector.

3.5. Database Creation and Fingerprinting

1. Open ACD/MS Manager and load a mass spectral trace. Select the IntelliXtract algorithm and click "run". ACD/IntelliXtract will smooth and pick peaks based on signal to

noise ratio and minimum peak height. These settings can be changed by clicking "options" while the IntelliXtract algorithm is selected, but the default values are appropriate for this application (see Note 12).

2. Create a .CDF file from the smoothed chromatogram by clicking "File" from the menubar and selecting "export". Choose the location name of file to be created then be sure to select "NetCDF" from the "Save as type" dropdown menu.
3. Repeat steps 1 and 2 above with all samples and retention index standards.
4. To find retention indices, a calibration file must be created from the retention index standards. Load the standard by opening AMDIS, clicking "File" and selecting "Open…" from the menubar. In the "Instrument:" dropdown menu, select "NetCDF". Next, find the location of the .CDF file, select it, and click "Open". Create the calibration file by selecting "Analyze", "Analyze GC/MS Data…" from the menubar at the top, and in the "Type of analysis:" dropdown menu, select "RI Calibration/Performance". Click the "RI Calib. Data…" button. In the "Analysis Settings" window, click the "Select New" button, choose the location and name for the new calibration file, and then click "open". In the "Analysis Settings" window, click "Save," and in the "AnalyzeGC/MS Data" window, click "Run". "T"s will appear over peaks that match spectra in the default database. Next, create a new database by clicking "Library" and selecting "Build One Library…" from the menubar. Click "Files…" then "Create New Library…". Choose the name and location for your retention index library and click "OK". Then, select the peak for each alkane by clicking the arrow above it on the chromatogram. In the library window, the retention time of the selected peak will appear on the "Add" button. Click this button, and then click "New Compound and Spectrum". Click "Edit…" then "Compound…" and enter the name under "Chemical ID:" and retention index (number of carbons × 100) in the "Index:" box. Click "Save", select the next alkane peak on the chromatogram, and repeat library entry steps.
5. To identify compounds and assign retention indices in AMDIS, load a mass spectral trace by clicking "File" and selecting "Open…" from the menubar. In the "Instrument:" dropdown menu, select "NetCDF". Next, find the location of the file you wish to analyze, select it, and click "Open". Ensure that the Golm metabolite database is loaded by selecting "Analyze", "Analyze GC/MS Data…" from the menubar at the top. Then, click "Target Library …", "Select New…", find and highlight the file "Q_MSRI_ID.msl", and click

"open". In the "Type of analysis:" dropdown menu, select "Use RI Calibration Data," and ensure your custom calibration data is loaded by clicking the "RI Calib. Data…" button. In the "Analysis Settings" window, click the "Select New" button, select your calibration file, and then click "open". Click "Save" in the "Analysis Settings" window and "Run" in the "AnalyzeGC/MS Data" window. "T"s will appear over peaks that match spectra in the database.

6. After analyzing with AMDIS, create a custom library by clicking "Library," followed by "Build One Library…". Next, create a new compound library by clicking "Files…" in the library window and "Create New Library…". Choose a name and location for the new library and click "OK". Select peaks on the chromatogram and add library entries as previously explained for alkanes. Click "Edit…" for each entry to add a compound ID if known (The "Index:" field should already have a number corresponding to the retention index of the compound) or any other designation.

3.6. Comparison of Mutant and Wild Type Fingerprints (see Note 13)

1. Once fingerprints are obtained and databases are created, open Metabolite Detector and import .CDF files by clicking "File", "Import", and "NetCDF Import…". Click the folder icon to bring up the "Open NetCDF Files" window and find the location of your files. Select all metabolic traces and the retention index standard by clicking and dragging a box around all files or by holding the "Ctrl" button on your keyboard and clicking each file. After selecting files, click "Open". The selected files should appear in the "NetCDF Import" box. Select the directory to save the converted files and click "OK".
2. Import the compound library and retention index standard library you created earlier by clicking "File", "Import", "Import MSL Library…". Only one library can be imported at a time, so select either library by clicking the folder icon next to the "Import file:" box and finding the .MSL file. Click "OK" and the number of compounds in the library will be displayed. Click the folder icon next to the "Library file:" box and choose a name and location for the converted library file. Click "Save" then "OK" and repeat import steps for the other library you created earlier. When "New Library" dialog box appears, click the "Yes" button.
3. Align the imported spectral traces by clicking "Tools", "RI-Calibration Wizard…" in the menubar. Click "Next", and then click the folder icon to select the retention index standard .bin file. Find and select this file and click "Open". Click "Next" and click the folder icon to select the .lbr file created from the library of retention index standards.

Find and select this file, and then click "Open". Select the compounds present in your standard mix from the left column and click the "≫" button, and then click "Next". The software will attempt to match retention index standard compounds to peaks in the chromatogram. If the assignments are not correct, double click the RT box associated with the compound in question and select the correct retention time from the dropdown menu (Do NOT type the correct retention time! If the software has not identified the exact time you enter as a component, the software will crash). After ensuring retention times are correct, click "Next," confirm that the "RI Calibration Table" is correct, and click "Next" again. Click the green "+" icon and find and select the mass spectral traces to be compared. Click "Open," followed by "Next". "Deconvolution Settings" can be adjusted if the default settings do not produce satisfactory results, however, the default settings should be used the first time an alignment is performed. Click "Next" followed by "Start".

4. Compare mass spectral traces by clicking "Tools" in the menubar and selecting "Batch quantification…". In the window that appears, click "Next >" then click the green "+" icon, locate and select the .bin files to be compared and click "Open". The selected files will appear in the "Chromatogram Selection" box. Click "Next >". Select "Targeted Analysis" in the "Analysis Type" window and click "Next >". In the "Settings" window, click the tool icon (which resembles a wrench and a screwdriver), select "Identification" in the box on the left, and click the folder icon next to the "Compound Lib:" box to select the compound library you created and imported earlier. Locate and highlight the .lbr file and click "Open". Ensure that "Combined score" is selected and click "OK" (Other settings can be changed from this window, but unless results are unsatisfactory, default settings should be used). In the "Integrated GCMS chromatogram analysis" window, click "Finish".
5. Click on the "Batch Quantification" tab at the bottom of the main "MetaboliteDetector" window to view the results. The table can be exported to a comma separated value file by clicking the "Export table to CSV" button (It looks like a page with a green arrow). In the "CSV Export" window click the "Sums" button and choose a name and location for the resulting file. These exported data can then be used with statistical analysis software of your choice. Alternatively, principal component analysis can be performed by clicking the green "PCA" button. This opens a new window with PCA data.

4. Notes

1. Liquid media composition (carbon source, nitrogen source, pH, etc.) can be adjusted depending on experimental conditions.
2. The derivatizing reagents used are extremely reactive with hydroxyl groups so great care should be taken to avoid contact with water and minimize contact with air. TMSI and TMCS should be stored under vacuum in a dessicator.
3. Retention index standards should be diluted to 50 μg/mL in hexane. At higher concentrations, higher carbon number alkanes are viscous to solid at room temperature and should be warmed slightly (~37°C). As these compounds are volatile, there is no need to derivatize this standard mix.
4. Growth parameters (incubation time, temperature, etc.) can be optimized for different fungi.
5. Mutants with inhibited growth will not produce enough tissue for analysis in 10 days. Incubation times should be increased to allow for more growth. For mutants that do not produce conidia, liquid media can be inoculated with hyphae collected from a V8 agar plate. However, to ensure reproducibility, special care must be taken to use equal amounts of inoculum.
6. The metabalome can change quite rapidly in response to stress or changing environments; therefore, to ensure reproducibility, tissue should be flash frozen in liquid nitrogen immediately after harvesting and kept frozen at all points between harvesting and extraction.
7. Plastics should be avoided when using solvents. For all steps after harvesting tissue, glass syringes and pipettes should be used for liquid transfer and centrifugation should be carried out in glass vials with PTFE lined caps.
8. Because the glass vials are somewhat fragile, special care should be taken during centrifugation. Any rotor can be used, however, removing vials is easiest from swinging buckets with inserts designed to fit 15-mL tubes.
9. Hydrolysis should be carried out with dried samples, so that the HCl is not diluted.
10. Hydrolysis products should be completely dried prior to the addition of derivatizing agents because TMSI and TMCS react strongly with hydroxyl groups. If the methanol is not evaporated, the efficiency of derivatization will be severely compromised due to an excess of hydroxyl groups.
11. Metabolic profiles will vary from species to species, so it may be necessary to modify GC parameters (oven program, column length, etc.) to improve resolution.

12. IntelliXtract options such as peak-picking, smoothing, and baseline correction will depend on the requirements of the experiment and the individual chromatogram, but it is generally better to assign peaks liberally at first then remove noise manually to avoid missing possible metabolites.
13. Metabolic fingerprints should be replicated no less than three times and comparisons should be made between the averages of peak areas.

Acknowledgments

The authors thank Dr. Jackson Lay of the Arkansas Statewide Mass Spectrometry Facility for expert advice and technical support. This work was supported by the Arkansas Agricultural Experiment Station, the Arkansas Biosciences Institute, and the Arkansas Center for Plant Powered Production (P^3 Center).

References

1. Drew, S., and Demain, A. (1977) Effect of primary metabolites on secondary metabolism *Annu Rev Microbiol* **31**, 343–356.
2. Sinha, K., and Bhatnagar, D. (1998) Mycotoxins in agriculture and food safety Marcel Dekker, inc. NY, 231.
3. Möbius, N., and Hertweck, C. (2009) Fungal phytotoxins as mediators of virulence *Curr Opin Plant Biol* **12**, 390–8.
4. Volkmann, M., Whitehead, K., Rütters, H., Rullkötter, J., Gorbushina, A. (2003) Mycosporine-glutamicol-glucoside: a natural UV-absorbing secondary metabolite of rock-inhabiting microcolonial fungi. *Rapid Commun Mass Spectrom* **17**, 897–902.
5. Aharonowitz, Y., Cohen, G., and Martin, J. (1992) Penicillin and cephalosporin biosynthetic genes: structure, organization, regulation, and evolution *Annu Rev Microbiol* **46**, 461–495.
6. Curtis, P., and Cross, B. (1954) Gibberellic acid – a new metabolite from the culture filtrates of *Gibberella fujikuroi Chem Ind* 1066.
7. Zeevaart, J., and Creelman, R. (1988) Metabolism and physiology of abscisic acid *Annu Rev Plant Physiol Plant Mol Biol* **39**, 439–73.
8. Weckwerth, W., Loureiro, M., Wenzel, K., and Fiehn, O. (2004) Differential metabolic networks unravel the effects of silent plant phenotypes *Proc Natl Acad Sci USA* **101**, 7809–14.
9. Fiehn, O. (2001) Combining genomics, metabolome analysis, and biochemical modeling to understand metabolic networks. *Comp Funct Genom* **2**, 155–68.
10. Calvo, A., Wilson, R., Bok, J., and Keller, N. (2002) Relationship between secondary metabolism and fungal development *Microbiol Mol Biol Rev* **66**, 447–459.
11. Yu, J., and Keller, N. (2005) Regulation of secondary metabolism in filamentous fungi, *Annu Rev Phytopathol* **43**, 437–458.
12. Flaherty, J., Pirttilä, A., Bluhm, B., and Woloshuk, C. (2003) PAC1, a pH-regulatory gene from *Fusarium verticillioides. Appl Environ Microbiol* **69**, 5222–7.
13. Sagaram, U., and Shim, W. (2007) *Fusarium verticillioides* GBB1, a gene encoding heterotrimeric G protein β subunit, is associated with fumonisin B_1 biosynthesis and hyphal development but not with fungal virulence. *Mol Plant Pathol* **8**, 375–84.
14. Bluhm, B., and Woloshuk, C. (2005) Amylopectin induces fumonisin B_1 production by *Fusarium verticillioides* during colonization of maize kernels. *Mol Plant Microbe Interact* **18**, 1333–9.

Chapter 19

Tapping Genomics to Unravel Ectomycorrhizal Symbiosis

Jonathan M. Plett, Barbara Montanini, Annegret Kohler, Simone Ottonello, and Francis Martin

Abstract

Given recent technological advances, we are in a golden era of cell and whole organism research. With the availability of so many sequenced genomes, and the data that has been mined there-in, it is easy to gain the impression that all our work as scientists is complete. Instead, such work and results have now provided oceans of data, but with minimal functional information. We also do not have a full grasp on the working relationships within a number of different plant developmental pathways. This is especially true in the study of the symbiotic interaction between ectomycorrhizal fungi and their plant hosts. One of the current interests in symbiotic and pathogenic interactions between plants and fungi is the role of small, secreted proteins. What makes fungal small secreted proteins so interesting is that only a few of them share sequence homology to any other known proteins, but some may act as effectors modulating plant metabolism and development. Therefore, it is difficult to make predictions as to the action of these proteins without functional analysis. For this reason, we created a pipeline to analyze the role and function of these proteins. Typically, this involves transcriptional analysis of genes followed by protein localization, identification of protein–protein interactions, and functional analysis of the protein through heterologous expression in yeast among many other different procedures. Due to the physiology of mycorrhizal root tips, there are a number of unique challenges that must be overcome to properly study a fungal effector. Here, we outline some of the methods, and hopefully helpful tips, that we are currently using to pursue the study of different effectors in the *Laccaria–Populus* interaction.

Key words: Functional genomics, Ectomycorrhizal fungi, *Laccaria bicolor*, *Populus trichocarpa*, Small secreted protein, *Saccharomyces cerevisiae*

1. Introduction

Trees are large, long lived, and stationary. As such, trees encounter problems when accessing nutrients and water while excluding pathogens, heavy metals, or other pollutants. To compensate against these potential problems, a great number of forest

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722, DOI 10.1007/978-1-61779-040-9_19,

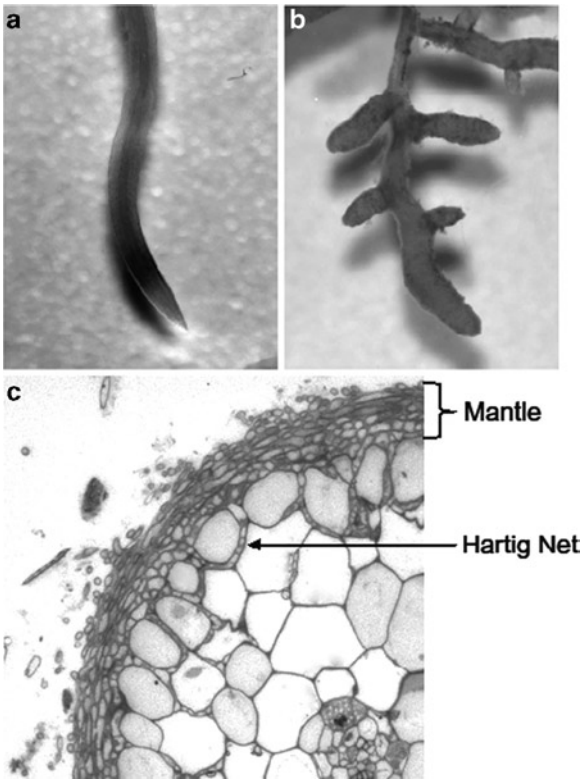

Fig. 1. *Mycorrhizal root tips of P. trichocarpa.* (**a**) Normal *P. trichocarpa* lateral root is slender and not bloated. (**b**) Mycorrhizal root tips of *P. trichocarpa* show expanded girth and a characteristic sheath of fungal hyphae surrounding the root. (**c**) Transverse cross section of *P. trichocarpa* mycorrhizal root tip showing the mantle surrounding the root epidermis as well as hyphae penetrating between the root epidermal cells to form a full Hartig net.

trees have established symbiotic relationships with soil-borne ectomycorrhizal (EM) fungi. Over 5,000 different EM fungi have been described (1) and several hundred can be detected in a few grams of forest soil (2). Mycelia surround the tree root in a sheath of fungal hyphae (a mantle) and penetrate between root cortical cells to form a hyphal network called the "Hartig net" (Fig. 1). In this symbiosis, exploratory hyphae, which can extend for several meters in the surrounding soil, transport water and nutrients to hyphae within the Hartig net, where exchange for plant photosynthate is enacted (3). Meanwhile, the dense mantle around the root protects the tree from fungal pathogenesis (e.g. *Fusarium oxysporum*) or from heavy metal uptake (4–7). The symbiotic tissues formed with these fungi are called ectomycorrhizal root tips.

One EM fungus that has received a lot of attention in the past few years is the basidiomycete *Laccaria bicolor* (Maire) P.D. Orton (Agaricomycotina, Agaricales; common name: bicolored

deceiver). With a broad host range, including both gymnosperms (e.g. Douglas fir) and angiosperms (e.g. Poplar), *L. bicolor* is a choice model organism when studying EM fungi. Additionally, the sequenced genome of *L. bicolor* was recently released (8), facilitating genomic and functional analysis. As the genome of poplar is also publicly available (9), and because there are a wide variety of poplar mutant populations (10, 11), we study the interaction between *L. bicolor* and *Populus,* as research in these model organisms is greatly facilitated to move beyond the pure genomic stage. As established gene models allow for the creation of microarrays, we frequently test expression patterns during the establishment of symbiosis between *L. bicolor* and poplar. Unfortunately, mycorrhizae are not leaves or stems – mycorrhizal root tips are small, contain little RNA, and have a large concentration of secondary metabolites that can degrade or sequester what little RNA is present. To help overcome these problems, we describe in Subheadings 3.1 and 3.2 a micro-RNA extraction protocol we devised from a combination of two commercially available RNA extraction kits followed by amplification of the cDNA before microarray analysis.

Once a fungal gene and transcript of interest has been identified through genomic and microarray expression analysis, it is important to identify where the actual protein is localized. Using antibodies specific to the protein of interest, it is possible to perform two levels of localization. First, by using immunofluorescence (Subheading 3.3), it is possible to ascertain on a tissue level where the protein is produced. Due to detection and resolution limitations, plus problems with autofluorescence, the final images that this technique provides may not be sufficient to localize proteins (i.e. fungal small secreted proteins) that are present at very low concentrations and in discrete compartments of the cell. To overcome these problems, it is often necessary to use electron microscopy and immunogold probes to locate the protein of interest (Subheading 3.4).

Gene validation and investigation of gene function can also be studied using the yeast *Saccharomyces cerevisiae* as a "test tube" for genome-wide heterologous functional analyzes. This is especially useful for postgenomic investigations of organisms, where efficient genetic transformation procedures are not yet available. We focus on four main yeast functional assays (outlined in Table 1) that are particularly well-suited for the large-scale interrogation of genes potentially involved in specific aspects of plant–fungus interaction and mycorrhiza development: (a) the signal sequence trap (SST) assay for the identification of secreted proteins, effector-like proteins and other proteins involved in intercellular communication, cell nutrition, and environmental sensing (Subheading 3.7); (b) the transcriptional activator trap (TAT) assay to identify positive regulators of gene expression

Table 1
Yeast assay targets and requirements

Yeast-based technique	Targeted proteins	Yeast fusion protein (position)	Heterologous cDNA library
Signal sequence trap	Secreted proteins	Invertase w/o signal peptide (N-term.)	5′-Enriched
Transcriptional activator trap	Transcriptional activators	DNA-binding domain (C-term.)	Full-length
Nuclear targeted trap	Nuclear proteins	Transcription factor w/o NLS (C-term.)	–
Two hybrid system	Interacting partners	Activation domain (C-term.)	Full-length
Heterologous functional complementation	Specific heterologous surrogates of a given yeast protein	None (unfused expression construct)	Full-length

(Subheading 3.8); (c) the nuclear targeted trap (NTT) to assess the ability of a protein to reach, and localize to, the nucleus (Subheading 3.9); and (d) the yeast two hybrid system (THS), an established method for gaining insight on the functional identity of a protein by identifying the protein partners with which it interacts (Subheading 3.10). All these screens are based on specific fusion libraries (Subheading 3.5), which are transformed into particular yeast strains (Subheading 3.6), followed by functional selection of clones expressing the proteins of interest, isolation of the corresponding heterologous cDNA inserts and sequence analysis (Subheadings 3.11 and 3.12). These methods can be applied to single genes, genome-wide or preselected ORF collections, or cDNA libraries. In general, total cDNA libraries, which can be prepared from particular life-cycle stages or tissues, are more practical and allow a higher throughput than ORF collections for various reasons: (a) cDNA library construction usually precedes the end of a genomic sequencing project and ensuing gene prediction, whereas assembly of a sequence information-supported ORF collection requires prior identification of all putative genes (indeed, cDNA libraries are usually constructed at the very beginning of a genome sequencing project as essential tools for training gene finding programs and for validating gene predictions); (b) cDNA library construction is generally less expensive and time-consuming than the assembly of an ORF collection, especially for organisms with a large number of ORFs, such as *L. bicolor*; (c) the use of cDNA libraries derived from a specialized tissue, such as the ectomycorrhiza, allows identification of all relevant mycorrhiza-expressed proteins; (d) due to construction artifacts, cDNA libraries may contain truncated

ORFs which, by encoding only a subset of protein domains, may improve the selection of specific (e.g. particularly large) proteins, especially in THS.

Immunoprecipitation is a time-tested method for confirming protein: protein interactions derived from THS analysis. While there are several published methods that have worked very well in the past, many of them have the disadvantage of needing large quantities of starting material, require the use of radioactive tracers, or take a number of days to complete. Given the small millimetric size of mycorrhizal root tips and the difficulty of amassing large quantities of tissue, traditional methods must be adapted to this particular case-study. Currently, there are new technologies on the market developed specifically to aid in speeding the process of immunoprecipitation while also reducing the amount of initial tissue that is needed. We currently use the Dynabead system available from Invitrogen. Unfortunately, the kit instructions and the suggested protein extraction buffers are best suited for bacterial or mammalian cell cultures. Given the presence of a cell wall, along with interfering secondary metabolites that may degrade the proteins during extractions, and due to the nonsterile conditions of samples found in soil, extraction of mycorrhizal root tips presents a number of problems that must be overcome. Subheadings 3.13 and 3.14 are adaptations that seek to deal with these problems while still maintaining protein–protein interactions.

One of the most powerful tools available to molecular biologists is the ability to genetically manipulate the model organisms under study. In this way, genes of interest can be differentially regulated to understand their roles in different developmental processes. In the study of the symbiosis between *L. bicolor* and poplar, it is possible to mutate both partners. Subheading 3.16 describes how genetic transformation of *L. bicolor* can be achieved while Subheading 3.17 deals with the transformation procedure employed for the poplar clone 717-1B4. Once a proven transgenic has been created, there are a number of ways to analyze the impact of that mutation on the formation of mycorrhizal root tips. While different in vitro methods to assess the so-called mycorrhization potential have been described (which have the advantage of being fast and to allow a strict control of nutrient supplementation), none of them is close to the natural growth conditions of either the tree or the fungus. When possible, we try to utilize both in vitro analysis of mycorrhizal root tips as well as mycorrhizae formed in pot culture in the greenhouse (described in Subheading 3.18). To control the nutrients that are being supplied to trees grown in pot culture, we use Terra-Green as an inert substrate in which to grow the poplar in the presence of the fungus.

Genomic studies of various systems have given us a wealth of information concerning all components encoded within the

genome, the structure and variation of genomes between different species, and the plasticity of these genomes. From these resources, different modeling and homology programs have emerged and that have advanced the study of gene function in new model organisms. It remains, however, that genomics is only able to predict the repertoire of putative proteins; their localization and function within an organism and the identification of the protein–protein interactions rely on high-throughput assays, such as those described in this chapter. Therefore, functional genomics is a prerequisite for shaping our understanding of the complex biotrophic interactions which play a key role in most terrestrial ecosystems.

2. Materials

2.1. Micro-extraction of RNA from Mycorrhizal Root Tips

1. RNeasy Plant Kit (Qiagen), microRNeasy kit (Qiagen).
2. PEG 8000 (Sigma), β-mercaptoethanol (Sigma).

2.2. Amplification of cDNA for Microarray Analysis

1. SMART™ PCR cDNA synthesis kit (Clontech).

2.3. Protein Immunofluo-Localization in Ectomycorrhizal Root Tips

1. Phosphate buffered saline (PBS): 135 mM NaCl, 25 mM KCl, 10 mM Na_2HPO_4, pH 7.4.
2. Digestion solution: 1% cellulase (Sigma), 0.01% pectolyase (Sigma), and 0.1% BSA (Sigma) in PBS buffer (pH 7.4).
3. Paraformaldehyde (Sigma) for use only under a fume hood to prevent inhalation of toxic fumes. Paraformaldehyde solutions are not stable for long-term storage at room temperature; therefore, make working solutions and store at –20°C until use. After long-term storage, check pH of solution before using.
4. 6% Agarose solution.

2.4. Protein Immunogold Localization in Ectomycorrhizal Root Tips

1. 10 mM sodium phosphate buffer (pH 7.4): Mix 19 mL of 0.02 M monobasic sodium phosphate with 81 mL of 0.02 M dibasic sodium phosphate and add 100 mL distilled water.
2. 25% Glutaraldehyde solution (Sigma) and paraformaldehyde (Sigma) stored at –20°C until use. Both of these fixatives should always be used under a fume hood to prevent inhalation of toxic fumes.
3. 4% Osmium tetroxide (Sigma) stored at 4°C until use. Osmium tetroxide is an extremely strong oxidizer and must be treated with great respect. Read manufacturer's MSDS

sheet and follow all safety protocols. Also, store remaining osmium tetroxide appropriately to avoid release into the environment.

4. Dilutions of ethanol for dehydration series (30, 50, 70, 90, and 100%) stored at room temperature.
5. LR white resin (Sigma).

2.5. cDNA Library Construction

1. SMART™ 5× first-strand buffer (Clontech), SuperScript™II reverse transcriptase 200 U/μL (Invitrogen), SUPERase In RNase inhibitor 20 U/μL (Ambion), Advantage® 2 PCR Kit (Clontech).
2. CloneMiner™ cDNA Library Construction Kit I (Invitrogen) and cDNA size fractionation columns (Invitrogen).
3. High-efficiency electrocompetent DH10T1^{R} *Escherichia coli*.
4. Gateway-modified pSUC2TM13ORI (12), pDEST™32 and pDEST™22 (ProQuest™ Two Hybrid System, Invitrogen) vectors.

2.6. Transformation and Other Yeast Procedures

1. 2× YPAD (2% yeast extract, 4% peptone, 4% glucose, 40 mg/L adenine).
2. YP solid medium: 1% yeast extract and 2% peptone in water adjusted to pH 6.5 with HCl. Add 2% agar. Sterilize by autoclaving.
3. YPD solid medium: add 2% glucose (from a 20% sterile stock solution) to melted YP.
4. 2 mg/mL ssDNA (in 10 mM Tris–HCl pH 8, 1 mM EDTA). Store at −20°C, boil for 5 min before use. Use ssDNA (e.g. Sigma D1626) for best efficiency.
5. 1.0 M lithium acetate stock solution (filter-sterilized).
6. 50% PEG MW 3350 (w/v) sterilized by autoclaving. Use PEG MW 3350 for best efficiency.
7. Drop-out mix (2 g adenine, 2 g L-Arg, 2 g L-Ile, 2 g L-Lys, 2 g L-Met, 3 g L-Phe, 2 g L-Ser, 2 g L-Thr, 2 g L-Tyr, 9 g L-Val; shake vigorously to mix).
8. SD-glucose medium: dissolve 6.7 g of yeast nitrogen base (YNB) without amino acids, 0.6 g of drop-out mix and 20 g of glucose in 600 mL of water, adjust to pH 5.6 with 2 M KOH, and bring to 1 L. If required, add 20 g of agar; sterilize by autoclaving. Using filter-sterilized stock solutions (2 mg/mL), add uracil, l-leucine, l-tryptophan, and/or l-histidine at a final concentration of 40 mg/L, as required by selection/screening media. For heterologous functional complementation assays, replace glucose with galactose to prepare modified SD-Ura + Gal medium.

2.7. SST Assay

1. Yeast strain YTK12.
2. pSUC-GW cDNA library.
3. YPR-A solid medium: add 2% raffinose and antimycin A (60 ng/mL) to melted YP-agar. Use freshly prepared filter-sterilized raffinose stock solutions (20%); add raffinose to melted YP when temperature is below 60°C.
4. SD-Trp media (solid and liquid).
5. Replica plating apparatus (150 mm), 10–20 velveteens.

2.8. TAT Assay

1. Yeast strain MaV103 harboring Gal4-dependent *LacZ*, *HIS3*, and *URA3* reporter genes.
2. pDEST™32 cDNA library.
3. 3AT stock solution (1 M): dissolve 840 mg of 3-amino-1,2,4-triazole (3AT) in 10 mL of deionized water.
4. SD-Leu-His + 25 mM 3AT, SD-Leu-His + 100 mM 3AT agar media.
5. SD-Leu and SD-Leu-Ura agar media.
6. SD-Leu liquid medium.
7. YPD solid medium.
8. Nitrocellulose or nylon membrane cut to the size of Petri dishes.
9. Z buffer solution: dissolve 16.1 g of $Na_2HPO_4 \cdot 7H_2O$ (8.52 g if anhydrous sodium phosphate is used), 5.5 g $NaH_2PO_4 \cdot H_2O$ (4.8 g of anhydrous salt), 0.75 g KCl, 0.246 g $MgSO_4 \cdot 7H_2O$ (0.12 g of anhydrous salt) in deionized water, sterilize by autoclaving, adjust pH to 7.0.
10. 2-Mercaptoethanol.
11. 4% X-Gal (5-bromo-4-chloro-3-indolyl-β-d-galactopyranoside, dissolved in dimethylformamide).

2.9. NTT Assay

1. Yeast strain L40.
2. pNIA-CEN-MBP vector.
3. SD-Leu-His + 50 mM 3AT agar medium.
4. SD-Leu and SD-Leu-Ura agar media.
5. YPD solid medium.
6. Nitrocellulose or nylon membranes cut to the size of Petri dishes.
7. Z buffer solution (see above).
8. 2-Mercaptoethanol.
9. 4% X-Gal (see above).

2.10. Two Hybrid Screen

1. ProQuest™ Two Hybrid System (Invitrogen).
2. pDEST™22 cDNA library.
3. Sterile cell scraper.
4. SD-Leu, SD-Trp, SD-Leu-Trp-His + 3AT (preoptimized 3AT concentration); SD-Leu-Trp, SD-Leu-Trp-Ura solid media; SD-Leu, SD-Trp, SD-Leu-Trp liquid media.
5. Yeast MaV103 strain (mating type "a").
6. YPD solid medium.
7. 0.45 μm Filters (47 mm Ø; one for each screen), 500 mL filter-holding funnel.
8. Nitrocellulose or nylon membranes cut to the size of Petri dishes.
9. Z buffer solution (see above).
10. 4% X-Gal (see above).
11. 2-Mercaptoethanol.
12. Sterile glycerol solution (65% glycerol, 0.1 M $MgSO_4$, 25 mM Tris–HCl, pH 8.0).
13. 1 M sorbitol (filter-sterilized).
14. 10 mg/mL cycloheximide.

2.11. DNA Insert Amplification by Yeast Colony PCR

1. 20 mM NaOH (freshly diluted from a 2 M stock solution).
2. 0.25 M Tris–HCl (pH 8.0).

2.12. Yeast Plasmid Miniprep

1. TE buffer (10 mM Tris–HCl pH 8.0, 1 mM EDTA).
2. RNaseA (20 mg/mL).
3. Lysis solution I: 25 mM Tris–HCl, 50 mM EDTA (pH 8.0), plus 10 μL of zymolyase (1.5 U/μL), and 5 μL of 2-mercaptoethanol in a final volume of 250 μL.
4. Lysis solution II: 0.2 M NaOH, 1% SDS.
5. Neutralization solution III: 3 M potassium acetate, pH 4.8.
6. Isopropanol (100%); ethanol (70%).
7. Electrocompetent *E. coli* ($DH10T1^R$) cells.

2.13. Protein Extraction for Immunoprecipitation

1. Extraction buffer: 0.01 M Tris–HCl buffer (pH 7.5) containing 150 mM NaCl, 1 mM EDTA, 1 mM EGTA (pH 8.0), 0.2 mM sodium vanadate, 0.2 mM PMSF, 50 mM sodium fluoride, 1 μg pepstatin A, 1 μg leupeptin, 1.5% triton-100, 1% (w/v) soluble polyvinylpyrrolidone (PVP) and 10% (w/ mass of sample ground) insoluble polyvinylpolypyrrolidone (PVPP). The protease inhibitors (sodium vanadate, pepstatin A, leupeptin, and PMSF) should be made up as stock solutions and diluted into the extraction medium immediately

before use. Pepstatin A should be made in DMSO or ethanol, PMSF in isopropanol and sodium vanadate in water. Store all inhibitors at –20°C except for sodium vanadate which can be stored at room temperature.

2.14. Immunoprecipitation

1. Dynabeads (Invitrogen).
2. PBS-T: 135 mM NaCl, 25 mM KCl, 10 mM Na_2HPO_4, pH 7.4, 0.02% Tween-20.
3. 5 mM suberic acid bis(3-sulfo-*N*-hydroxysuccinimide ester) sodium salt in 20 mM sodium phosphate, 0.15 M NaCl (pH 7.4). Make up this solution fresh every time as BS^3 is unstable over time in solution.
4. 1 M Tris–HCl (pH 7.5).
5. PBS: 135 mM NaCl, 25 mM KCl, 10 mM Na_2HPO_4, pH 7.4.
6. Denaturing elution buffer: 1 M Tris–HCl, 4% SDS, 20% glycerol, 9.8% B-mercaptoethanol, 0.04% bromophenol blue (pH 6.8).

2.15. Silver Staining Compatible with Mass Spectrometry Analysis

1. Gel fixative: 50% methanol, 5% acetic acid.
2. 50% Methanol.
3. 0.02% Sodium thiosulfate.
4. 0.1% Silver nitrate.
5. Developing solution: 0.04% formalin, 2% sodium carbonate.
6. 1% and 5% acetic acid solutions.

2.16. Transformation of L. bicolor

1. Malt extract medium: 1% malt extract, 2% agar.
2. Induction medium: 60 mM K_2HPO_4, 33 mM KH_2PO_4, 3.2 mM $(NH_4)_2SO_4$, 2 mM $Na_3C_3H_5O(CO_2)_3$, 0.8 mM $MgSO_4 \cdot 7H_2O$, 3 μM thiamine-HCl, 11 mM glucose, 40 mM MES, 0.5% glycerol, 2% agar, pH 5.3 supplemented with the appropriate antibiotic for selection and 200 μM acetosyringone.
3. Selection medium: 1% malt extract, 2% agar pH 7.5 supplemented with 100 μg/mL cefotaxime and the appropriate antibiotics for selection.

2.17. Populus Transformation

1. Medium M1: 21 mM ammonium nitrate, 19 mM potassium nitrate, 4 mM calcium chloride, 3 mM magnesium sulfate, 1.5 mM potassium dihydrogen phosphate, 100 μM boric acid, 112 μM manganese sulfate, 66 μM zinc sulfate, 5 μM potassium iodide, 1.2 μM sodium molybdate, 0.1 μM copper (II) sulfate pentahydrate, 0.15 μM cobalt chloride dehydrate, myo-inositol, 1.3 mM 2-(*N*-morpholino) ethanesulfonic acid, 110 μM ethylenediaminetetraacetic acid iron (III) sodium salt, 3% sucrose, 0.7% agar, pH 5.7–5.8.

Autoclave, cool to 55°C and then add filter-sterilized: 8 μM nicotinic acid, 5 μM pyridoxine, 3 μM thiamine, 2 μM calcium pantothenate, 4 μM l-cystein, 0.02 μM biotin, 10 μM (a-naphthalene) acetic acid, 1.5 mM l-glutamine, 5 μM (N6-(2-isopentenyl) adenine.

2. YEP medium: 10 g yeast extract, 10 g Bacto-peptone, 5 g NaCl, adjust to pH 7.0, and bring final volume to 1 L with deionized water.
3. M2 medium: same as M1 medium with the addition of filter-sterilized carbenicilline (500 mg/L) and cefotaxime (250 mg/L) after autoclaving.
4. M3 medium: M2 medium supplemented with the appropriate antibiotic.
5. Regeneration medium: M3 medium plus 0.1 μM TDZ.

2.18. Analysis of Mycorrhization Potential

1. Peat moss and vermiculite.
2. Pachlewski medium: 2.7 mM di-ammonium tartrate, 7.3 mM KH_2PO_4, 2.0 mM $MgSO_4 \cdot 7H_2O$, 13 mM maltose, 110 mM glucose, 2.9 μM thiamine-HCl, and 1 mL of a trace-element stock solution Kanieltra.
3. Terra-Green substrate (calcined attapulgite clay supplied by Turf-Pro, UK).
4. Hydroponic solution: 2.5 mM KNO_3, 0.8 mM KH_2PO_4, 1 mM $MgSO_4 \cdot 7H_2O$, 2.3 mM $Ca(NO_3)_2 \cdot 4H_2O$, 23 μM H_3BO_3, 4.6 μM $MnCl_2 \cdot 4H_2O$, 0.4 $ZnSO_4 \cdot 7H_2O$, 0.09 μM $H_8Mo_2NH_7$, 0.18 μM $CuSO_4 \cdot 5H_2O$, 20 μM FeNaEDTA, pH 5.8.
5. Nutrient solution: 0.8 mM KNO_3, 0.8 mM $Ca(NO_3)_2 \cdot 4H_2O$, 0.3 mM NaH_2PO_4, 0.3 mM $MgSO_4 \cdot 7H_2O$, 20 μL trace elements solution Kanieltra.

3. Methods

3.1. Micro-extraction of RNA from Mycorrhizal Root Tips

1. The extraction buffer used is the RLC buffer from the RNeasy Plant kit. This is a preferred buffer both for RNA extraction from fungal mycelium as well as from mycorrhizal root tips.
2. To the RLC buffer add 20 mg/mL of PEG 8000. If desired, heat this solution at 60°C for up to 30 min before the addition of the β-mercaptoethanol. The PEG 8000 does not fully dissolve but helps bind phenolics that would interfere with the extraction. Cool down the solution and add the β-mercaptoethanol.

3. Grind tissue in liquid nitrogen and then add the RLC extraction buffer as described in the manufacturer's instructions. If you have more than ten ectomycorrhizal root tips, add this solution to the RNeasy Plant Qiashredder columns (lilac columns). Centrifuge for 2 min at maximum speed and use the recovered supernatant for the subsequent steps in the microRNeasy kit (tissues protocol) starting at step 3 (page 27 of the instruction manual) "addition of ethanol to the supernatant" (see Note 1).
4. For the rest of the procedure, use the protocol as outlined by QIAGEN in their microRNeasy kit instruction booklet.

3.2. Amplification of cDNA for Microarray Analysis

A number of different and very efficient kits are currently available on the market for amplification of first-strand cDNA. We use the SMART™ PCR cDNA synthesis kit, available from Clontech, for our amplification as we have found that it is the most reliable. As we follow the kit instructions, we do not cover it in this section but rather refer the reader to the guidelines provided by the manufacturer. It is important to follow all instructions and the quality control guidelines as failure to do so will bias the results obtained during downstream analysis. The cDNA obtained with this method is of appropriate quality for either microarrays or, as is becoming popular, Illumina-Solexa RNA-Seq sequencing when working with a nonmodel organism. Once expression results have been obtained, it is possible to identify the genes that are most differentially regulated during the establishment of the symbiosis and, upon comparison with bioinformatics data, determine if there are any effector-like genes among them.

3.3. Protein Immunofluo-Localization in Ectomycorrhizal Root Tips

(Adapted from Ref. 8)

1. Immediately put freshly harvested tissue immediately into PBS supplemented with 4% paraformaldehyde. Leave for 2–16 h depending on the size of the tissue (see Note 2).
2. Following fixation, rinse the tissues twice with PBS (pH 7.4), 10 min each.
3. Embed the tissue sample in 6% agarose (see Note 3).
4. Leave the agarose block at 4°C for 24–48 h to fully harden.
5. Trim the agarose block and mount on the base of the vibratome using Roti Coll (Carl Roth) and cut 25–30 μm thick sections.
6. Once all the sections are prepared, it is optional to incubate them in digestion solution for 10 min at 35°C.
7. Wash five times in PBS buffer (pH 7.4).
8. Block in 1% BSA dissolved in PBS buffer (pH 7.4) for 1 h.

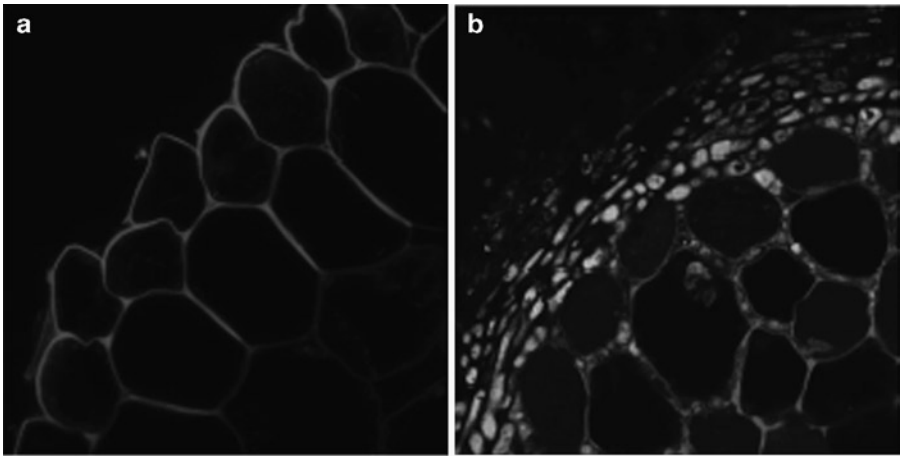

Fig. 2. *Immunofluorescent localization of a fungal protein in mycorrhizal root tips.* (**a**) Transverse section of a noncolonized root of *P. trichocarpa* with cell walls stained in *red* with propidium iodide. (**b**) Transverse section of mycorrhizal root tip with cell walls stained in *red* using propidium iodide. *Green signal* indicates the presence of a fungal protein induced by the mycorrhization process.

9. Wash five times with PBS buffer (pH 7.4).
10. Incubate vibratome cuts with the primary antibody overnight in 0.5% BSA at 4°C (see Notes 4–7).
11. The following day wash the samples five times with PBS buffer (pH 7.4).
12. Incubate samples for 2 h in the presence of the appropriate secondary antibody labeled with a fluorochrome compatible with the microscope visualization system available (e.g. anti-rabbit IgG-AlexaFluor 488; Molecular Probes). This antibody should be diluted in 0.5% BSA in PBS buffer (see Note 8).
13. Wash samples five times in PBS buffer (pH 7.4).
14. Mount the samples in 80% glycerol, 5% w/v propul gallate in PBS buffer (see Note 9).

An example of the results obtained with this method is shown in Fig. 2.

3.4. Protein Immunogold Localization in Ectomycorrhizal Root Tips

(Adapted from Ref. 13)

1. Immediately transfer freshly harvested ectomycorrhizal root tips into ice-cold 10 mM sodium phosphate buffer (pH 7.4) containing 2.5% glutaraldehyde (see Note 10 on fixatives; Fig. 3).
2. Put the sample under vacuum (400 mmHg) for 20–30 min to aid infiltration of the fixative. During this step, maintain samples on ice.

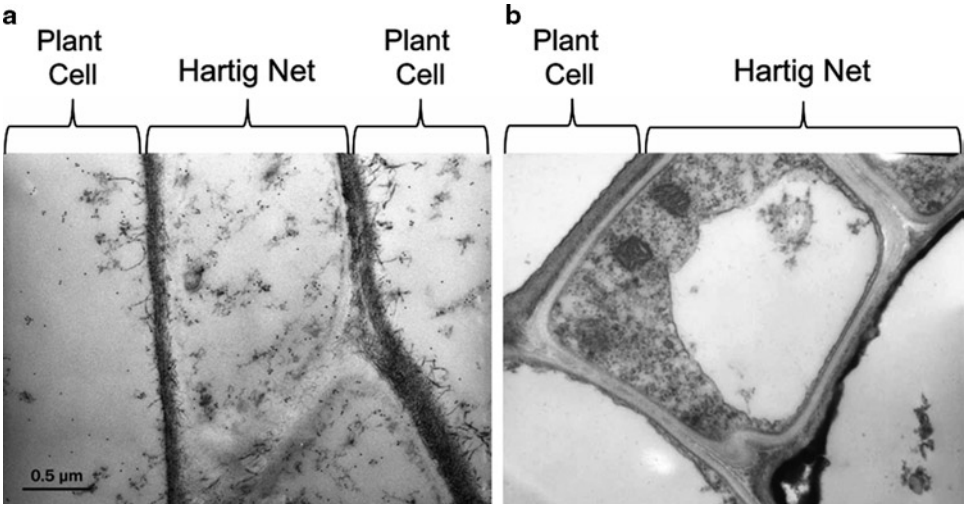

Fig. 3. *Different fixation techniques differentially preserve subcellular cell structure.* (**a**) Cross section of a mycorrhizal root tip preserved with 4% paraformaldehyde and 0.5% glutaraldehyde. Cells of both plant and fungal origin have not maintained subcellular structural integrity with loss of the main vacuole as well as most other organelles except for the nucleus. (**b**) Cross section of a mycorrhizal root tip preserved with 2.5% glutaraldehyde followed by a postfixation with 1% osmium tetroxide for 1 h. These samples maintain a high degree of structural integrity with vacuoles, mitochondria, vesicles, and the nucleus still visible.

3. Depending on the size of the tissue, leave in fixative solution for either 2 h at room temperature, or 16 h at 4°C (see Note 2 regarding length of incubation).
4. After fixation, rinse twice (10 min each) with 10 mM sodium phosphate buffer (pH 7.2).
5. Postfix the tissues with 1% osmium tetroxide diluted in deionized water for 1 h at room temperature in the dark (see Notes 11–13).
6. Wash the tissue three times with water, 10 min per wash.
7. Dehydrate the tissue in an ethanol series (30, 50, 70, 90, and 100% ethanol). The tissue should stay in each ethanol bath for 20 min (see Note 14).
8. Begin infiltration of the sample with LR white resin (Sigma), performing a series of different baths. First, use a 2:1 ethanol:LR white solution for 2 h followed by a 1:1 ethanol:LR white bath for 1 h, and then by 1:2 ethanol:LR white bath for 3 h. Replace the final bath with 100% LR white and place the samples at 4°C overnight (see Notes 15 and 16).
9. The following morning, replace the LR white with new 100% LR white and leave for 4 h at 4°C.
10. Transfer the samples to an embedding capsule and position in fresh LR white as desired. Polymerize for 24 h at 60°C without oxygen (see Notes 17–19).

3.5. cDNA Library Construction for Gene Validation and Functional Discovery Approaches in Yeast

1. Total RNA (usually 200–500 ng) is utilized as template for mRNA amplification as described in the mRNA amplification kit manual using modified, Gateway-compatible primers for amplification by RT-PCR (see Notes 20–22).
2. For first-strand cDNA synthesis, CDS Primer II A must be replaced by the AttB2-Smart I FL primer for full-length libraries, or the AttB2-Smart I 5R primer for 5′-enriched libraries; the SMART™ T7 primer must be replaced by the AttB1-Smart I primer (see Table 2 for oligonucleotide primer sequences).
3. First-strand cDNA is amplified as described in the SMART™ mRNA amplification kit manual (Clontech; Appendix B, "Analyzing First-Strand cDNA") using the AttB1-Smart II and the AttB2-Smart II primers instead of the PCR Primer II A and the T7 extension primer (Table 2; see Note 23).
4. PCR products are purified on size fractionation columns and cloned into the pDONR222 vector using the BP Recombination reaction (CloneMiner™ cDNA Library Construction; Invitrogen); a typical reaction mixture contains the pDONR222 vector (250 ng) and purified cDNA (100 ng) in a final volume of 10 μL. If BP clonase II is used, 2 μL are added to the reaction mixture.
5. The BP clonase reaction mixture is then transformed into high-efficiency electrocompetent DH10T1^{R} *E. coli* cells in order to amplify the entry library.
6. Entry libraries from different tissues and/or life-cycle stages can be used for cDNA sequence analysis as well as for the construction of secondary expression libraries. To this end, a given entry library (250 ng) is mixed with a particular destination vector (250 ng), plus 2 μL of LR clonase II, in a final volume of 10 μL. Gateway-modified pSUC2TM13ORI ((12); see Note 24), pDEST™32 and pDEST™22 (ProQuest™ Two Hybrid System Manual, Invitrogen) are used as destination vectors for SST, TAT, and THS, respectively. The galactose-inducible yeast expression vector pYES-DEST52 (Invitrogen) is used for functional complementation assays (Table 3).
7. Destination libraries are amplified as described above for the entry library.

3.6. Transformation and Other Yeast Procedures

1. Yeast transformation is carried out with a high-efficiency PEG/LiAc protocol (see, e.g. 14) with 20–50 μg of amplified library DNA in order to obtain ~5×10^5–2×10^6 colonies (see Note 25).
2. After transformation, bring the volume to 1 mL with sterile water, and plate 100 μL of 10×, 100×, and 1,000× dilutions on the appropriate selective media. Yeast strains, vectors, and media are listed in Table 3.

Table 2
Oligonucleotides utilized for library construction, PCR amplification and sequencing

Primer name	Primer sequence (5′–3′)[a]	Annealing temperature	Forward/ reverse	Application
AttB1-Smart I	GG**ACAACTTTGTACAAAAAAGTTGG**AGAGGGCGGG	Not applicable	F	**cDNA library construction**
AttB2-Smart I FL	**ACAACTTTGTACAAGAAAGTTGG**GTACT(30)VN	Not applicable	R	**cDNA library construction**(full-length libraries)
AttB2-Smart I 5R	**ACAACTTTGTACAAGAAAGTTGG**GTACTNBN(20)	Not applicable	R	**cDNA library construction**(5′-enriched libraries)
AttB1-Smart II	TCGTCGGGG**ACAACTTTGTACAAAAAAGTTGG**	Not applicable	F	**cDNA library construction**
AttB2-Smart II	GGCGGCCGC**ACAACTTTGTACAAGAAAGTTGGGT**	Not applicable	R	**cDNA library construction**
pSUC-GW-F	CTCGTTCCCTTTCTTCCTTGTTTC	47°C	F	**SST**: yeast colony PCR
pSUC-GW-R	GACCAAAGGTCTATCGCTAGTTTC	47°C	R	**SST**: yeast colony PCR and sequencing
pDEST32-F	GGAGACTGATATGCCTCTAAC	52°C	F	**TAT**: yeast colony PCR and sequencing
pDEST-R	AGCCGACAACCTTGATTGGAGAC	52°C	R	**TAT and THS**: yeast colony PCR
pDEST22-F	TATAACGCGTTTGGAATCACTAC	52°C	F	**THS**: yeast colony PCR and sequencing
pNIA-F	ATTCGATGATGAAGATACCC	52°C	F	**NTT**: PCR and sequencing
pNIA-R	GGACCTAGACTTCAGGTTGT	52°C	R	**NTT**: PCR
pYES-F	CCTCTATACTTTAACGTCAAGG	50°C	F	**Heterologous functional complementation**: yeast colony PCR and sequencing
pYES-R	GACATAACTAATTACATGATGCG	50°C	R	**Heterologous functional complementation**: yeast colony PCR and sequencing

[a]Gateway cassettes are in **bold**

Table 3
Strains, vectors, and selection media for yeast transformation and screening

Yeast-based technique	Yeast strain	Vector	Transformant selection medium	Phenotype screening medium
Signal sequence trap	YTK12	pSUC-GW	SD-Trp	YP raffinose + antimycin A (YPR-A)
Transcriptional activator trap	MaV103	pDEST32	SD-Leu	SD-Leu-His + 3AT
Nuclear targeted trap	L40	pNIA-CEN-MBP	SD-Leu	SD-Leu-His + 3AT
Two hybrid system	MaV203	pDEST22	SD-Trp	SD-Leu-Trp-His + 3AT
Heterologous functional complementation	Specific yeast mutant	pYES-DEST52	SD-Ura + Gal	Mutant strain-specific conditions

3. Incubate for 3–4 days at 30°C and count cells to determine transformation efficiency.
4. Use the optimal PEG concentration for whole library transformation, scaling-up the volumes of yeast culture and transformation mix in proportion to the amount of library DNA.
5. Positive colonies from the various assays (see below) are organized into 96- or 384-well microtiter plates and handled with a 96- or 384-multipinner device (V&P). Replica-plating can then be performed by pinning cells from one plate to another, dipping the cell-loaded multipinner (20 times) into a microtiter plate containing 40 µL of sterile water/well to dilute the cell input prior to transfer.

3.7. SST Assay

(Adapted from Ref. 12)

1. Transform yeast strain YTK12 as described in Subheading 3.6 with the entire library (5×10^5–2×10^6 colonies), plate about 50,000 library cDNA-transformed cells on SD-Trp medium, and culture them for 3–4 days at 30°C (see Notes 26 and 27).
2. Replica-plate on YPR-A medium with a velvet device.
3. Culture cells for 5–7 days at 30°C and then transfer single colonies to a 96-well microtiter plate (40 µL of SD-Trp medium/well). Culture overnight at 30°C with mild shaking, replica-plate individual colonies onto SD-Trp solid medium, and grow for 4–5 days (see Note 28). Replica-plate on YPR-A to confirm phenotype.

4. Use individual, confirmed colonies for PCR amplification of the cDNA insert or plasmid isolation, followed by sequence analysis (see Subheadings 3.11 and 3.12).

3.8. TAT Assay

(Adapted from Ref. 15)

1. Transform yeast strain MaV103 with the appropriate pDEST™32 expression library as described in Subheading 3.6 (see also Subheading 3.5) and plate about 50,000 cells on SD-Leu-His + 25 mM 3AT agar medium (see Notes 27, 29 and 30).
2. After 5–7 days at 30°C, transfer single colonies to a 96-well microtiter plate (40 µL of SD-Leu medium/well) and grow overnight at 30°C. Replica-plate on SD-Leu-His + 25 mM 3AT using a multipinner device (see Notes 28 and 31).
3. Replica-plate colonies grown on SD-Leu-His + 25 mM 3AT onto SD-Leu (three plates for each starting plate) until homogeneous growth are obtained. Using freshly grown colonies, perform separate gene reporter assays for each of the three reporters (see below) to assess the strength of heterologous putative activators (see Note 32).
4. For the *HIS3* reporter assay, replica-plate freshly grown colonies on SD-Leu (positive control), SD-Leu-His + 25 mM 3AT and SD-Leu-His + 100 mM 3AT (test); check growth after 2 and 4 days at 30°C for 25 mM 3AT and 100 mM 3AT, respectively. Clones yielding large or small colonies on 100 mM 3AT are classified as strong (+++) or medium (++) strength activators; clones that fail to grow on 100 mM 3AT, but grow on 25 mM 3AT are classified as weak activators (+), whereas clones that do not grow on either medium are considered as false positives (–). An example of this assay is shown in Fig. 4.
5. For the *URA3* reporter, replica-plate freshly grown colonies on SD-Leu (positive control) and SD-Leu-Ura (test) for 4–5 days. After culture, classify activator strength as none (–), weak (+), medium (++), or strong (+++), based on visual inspection of colony size using the positive control as a reference.
6. For the *LacZ* (β-Gal) reporter assay, pin candidate clones onto 150 mm-YPD plates overlaid by a nitrocellulose or nylon membrane (without intermediate pin washing and input cell dilution). Incubate overnight at 30°C with the cell-loaded surface upward, and then perform the β-Gal assay as described by Walhout and Vidal (16). After culture incubation at 37°C for 24 h, classify clones based on color development as no activation (–; white), weak (+; green), medium (++; light-blue), and strong (+++; dark-blue).
7. Cumulative scores from the three assays are used to evaluate activation strength: clones with score values ≥++ in at least

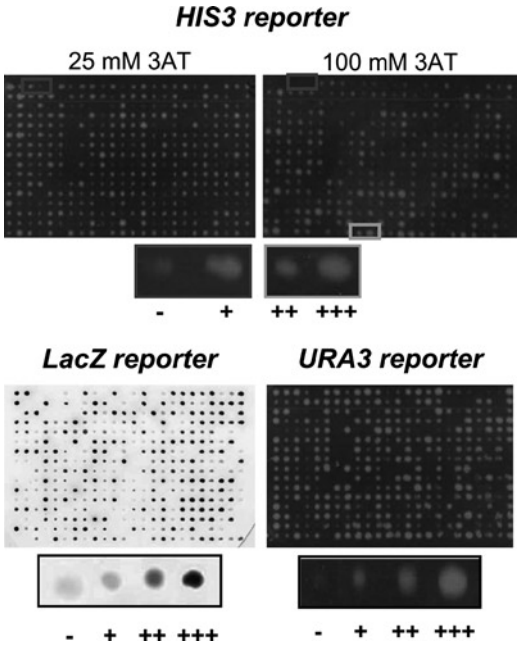

Fig. 4. *Different transcriptional activators from T. melanosporum functionally selected and visualized in yeast.* A representative subset of TAT-positive clones (plus a number of empty wells as technical negative controls) were grown on selective media (SD-Leu-His/+25 mM 3AT, SD-Leu-His/+100 mM for *HIS3*; SD-Leu-Ura for *URA3*; and SD-Leu followed by transfer to a nitrocellulose membrane for *LacZ*) and assayed for the indicated reporter gene activities. Activation strength was evaluated semiquantitatively based on colony size upon growth on different selective media (*HIS3*, *URA3*) or color intensity (*LacZ*) as indicated (see the text for details).

two assays and ≥+ in the remaining assay are considered to be putative strong activators (see Note 33).

3.9. NTT Assay

(Adapted from Ref. 17)

1. Subclone the coding sequence of a TAT-positive clone of interest into the pNIA-CEN-MBP vector in frame with the artificial transcription factor LexA-DBD/yGal4-AD, and transform the resulting construct into the L40 strain as described in Subheading 3.6 (see Notes 34 and 35).
2. Following transformant selection (SD-Leu), *HIS3* (+50 mM 3AT) and *LacZ*/β-Gal reporter assays are performed as described in Subheading 3.8, with slight modifications. For each clone, plus the unmodified pNIA-CEN-MBP vector (negative control), resuspend freshly grown transformants in sterile water. Starting from an OD_{600} of 1.0, serially dilute the resulting cell suspension 1,000× in tenfold increments. Spot an aliquot (2 μL) of each dilution onto SD-Leu and SD-Leu-His+50 mM 3AT plates and incubate at 30°C for 3–5 days.

3. Spot aliquots (2 μL) of each clone, pregrown to an $OD_{600} = 0.1$ on 150 mm YPD plates overlaid with a nitrocellulose or nylon membrane, incubate overnight at 30°C, and perform the β-Gal assay as described in Subheading 3.8. Colonies that grow on SD-His + 50 mM 3AT and score positively to the β-Gal assay express foreign proteins with a functional nuclear localization signal (NLS).

3.10. Two Hybrid Screen

(Adapted from Ref. 18)

1. Vectors from the ProQuest™ Two Hybrid System kit (Invitrogen) are first used to set up the necessary interaction controls. To this end, transform the MaV103 strain with pEXP32/Krev (SD-Leu selection), and then cotransform MaV103-pEXP32/Krev cells with pEXP22/RalGDS wt (strong interaction control), pEXP22/RalGDS m1 (weak interaction control), and pEXP22/RalGDS m2 (no interaction control). Select transformants on SD-Leu-Trp medium (see Note 36).
2. For "bait" construction, clone the gene of interest into the pDEST™32 plasmid in frame with the Gal4-DBD and transform the resulting plasmid into MaV103 cells (mating type "a"). Perform self-activation tests for the DBD "bait strain" (as in Ref. 16), using the above described interaction controls (see Note 37).
3. To prepare the "prey" strain, transform the MaV203 strain (mating type "α") with 20–50 μg of the amplified pDEST™22 library of interest to obtain ~5×10^5–2×10^6 colonies. Plate cells on SD-Trp medium (50,000 colonies/plate) (see Note 27), culture for 3–4 days at 30°C, and then collect transformants with sterile water using a plate scraper. Wash the cells twice with sterile water, resuspend them in one pellet volume of sterile glycerol solution (subdivided into 100 μL aliquots), and immediately freeze in liquid nitrogen for storage at –80°C.
4. In preparation for mating, inoculate four flasks of SD-Leu (30 mL ea.) with different amounts of freshly grown "bait strain" cells ($OD_{600} = 0.01$, 0.02, 0.04 and 0.08). Culture at 30°C in a rotary shaker (150 rpm) and monitor growth by checking absorbance at 600 nm. When cells are in the exponential growth phase (i.e. 1.5–2.5 OD_{600} after 16–20 h), take 30 OD cell-equivalents from the flask containing exponentially grown cells and discard the rest. At the same time, quickly thaw one aliquot of "prey strain" cells at 35°C, inoculate two flasks of SD-Trp (50 mL ea.) with different amounts of cells ($OD_{600} = 0.3$ and 0.1), and culture and check growth as above. Take 20 OD cell-equivalents from the flask containing exponentially grown cells and discard the rest of the cells.

5. Mix together 30 OD_{600} of "bait strain" and the 20 OD_{600} of "prey strain" cells, pellet by centrifugation, and perform mating as described by Soellick and Uhrig (18).
6. After mating, resuspend cells in 10 mL of sterile 1 M sorbitol, bring to a final volume of 400 mL with sterile water, and plate 4 mL of cell suspension (at least thirty 150 mm plates) on SD-Leu-Trp-His + 3AT. As a control, serially dilute 100 μL of cells from 100× to 10,000× in tenfold increments, plate 250 μL of each dilution on SD-Leu-Trp plates. Culture for 5–7 days at 30°C and count colonies to determine the number of zygotes (10^6–10^7 expected).
7. In preparation for testing the two-hybrid interaction, collect single colonies into liquid SD-Leu-Trp medium in 96-well plates and grow cells overnight at 30°C (see Note 28). Make a replicate on SD-Leu-Trp-His + 3AT and on SD-Leu-Trp (three plates for each starting plate) until homogeneous growth is obtained. At this stage, a 384-well microtiter plate can be assembled.
8. Using freshly grown colonies perform reporter gene assays to evaluate the strength of protein–protein interactions. Assay procedures are essentially the same as those described in Subheading 3.8 with a few modifications: (a) *HIS3* assays must be carried out at the previously determined 3AT concentration, and, if possible, at some higher concentration (≤200 mM); (b) the *URA3* reporter assay must be performed on SD-Leu-Trp-Ura plates. Colonies growing on SD-Leu-Trp-His + 3AT as well as on SD-Leu-Trp-Ura, and scoring positive to the β-Gal assay are likely to harbor a pair of interacting proteins.
9. Interactions must be confirmed in an independent replicate of the THS to make sure that no mutation has occurred in the "bait plasmid" or in the "bait strain" after transformation (e.g. a mutation in the "bait" that converts it into a self-activator). If bait/prey interactions are authentic, the phenotypes associated with reporter gene expression must be reproduced when the prey is reintroduced into MaV103 cells along with the original "bait plasmid," but not with the empty pDEST™32 plasmid. This can be done in two ways (see the ProQuest™ Two Hybrid System manual for details): (a) by retransforming the isolated "prey plasmid" (see Subheading 3.12 for yeast plasmid preparation) into MaV103 cells together with its cognate "bait plasmid" or pDEST™32 (very reliable, but time-consuming method); (b) by replica plating cells on SD-Trp + 5 μg/mL cycloheximide for 3–4 days in order to induce "bait plasmid" loss, followed by a further replicate on SD-Trp and SD-Leu-Trp to select colonies that have lost the plasmid (growth on the former, but not the latter medium), and then transformation with an

independent "bait plasmid" preparation and pDEST™32 (faster, but less reliable method). In both cases, reporter gene expression must be verified.

10. Proceed with sequence analysis of confirmed interactors (see below).

3.11. Sequence Analysis of Insert DNA Amplicons Produced by Yeast Colony PCR

1. Using a sterile pipette tip, resuspend one freshly grown yeast colony (2 mm diameter) in 10 µL of 20 mM NaOH (freshly diluted from a 2 M stock solution) and incubate for 15 min at 98°C in a thermal cycler. Transfer tube to ice, quickly add 10 µL of 0.25 M Tris–HCl (pH 8), mix and use immediately for PCR amplification.
2. Use 6 µL of each lysed cell preparation for a 35-cycle PCR reaction, carried out in a final volume of 50 µL with the oligonucleotide primers (10 µM each) and the annealing temperatures specified in Table 2.
3. Check amplification by loading 3 µL of the above reaction mixtures on a 1% agarose gel.
4. Purify PCR products and directly sequence individual amplicons using the oligonucleotide primers specified in Table 2.

3.12. Sequence Analysis from a Yeast Plasmid Miniprep

1. Grow cells in 5–10 mL of selective medium to an $OD_{600} \geq 1$.
2. Centrifuge 3 mL of the above cultures in 1.5 mL tubes, discard supernatant and resuspend pellets by vortexing in 1 mL of TE. Sediment cells by centrifugation and remove all residual liquid.
3. Resuspend TE-washed cells by vortexing in 250 µL of RNase A (10 µg/mL)-supplemented lysis solution I and incubate at 37°C for 30 min.
4. Add 250 µL of freshly made lysis solution II, mix gently by inverting the tubes four to eight times, and incubate for 3–5 min at room temperature (do not exceed 5 min).
5. Add 350 µL of ice-cold neutralization solution III and immediately mix by inverting the tubes four to eight times.
6. Centrifuge for 10 min in a microcentrifuge (maximum speed; 4°C), pour supernatants into new tubes, add 0.6 volumes of isopropanol, and vortex.
7. Centrifuge for 10 min at room temperature, pour off supernatants, and wash pellets with 1 mL of 70% ethanol.
8. Centrifuge for 5 min at room temperature, dry pellets, and resuspend in 4 µL of water.
9. Transform electrocompetent *E. coli* (DH10T1^{R}) cells with 1 µL of each plasmid preparation to amplify the plasmids.
10. Purify the plasmids and sequence the inserts using the oligonucleotide primers specified in Table 2.

3.13. Protein Extraction for Immunoprecipitation

1. Grind 0.1–1 g of freshly harvested mycorrhizal root tips in three times (w/v) ice cold extraction buffer.
2. To lyse the cells, leave the ground samples for 30 min at 4°C with shaking followed by sonication for five times, 5 s each on ice, or in ice water.
3. Centrifuge the macerate at maximum speed for 3 min at 4°C and use supernatant as described in Subheading 3.14.

3.14. Immunoprecipitation

1. Transfer 1.5 mg of Dynabeads to a 1.5 mL tube. Place the tube on the magnet and remove the supernatant (see Notes 38 and 39).
2. Add 5–10 μL of the antibody solution in 200 μL of PBS-T. Incubate for 10 min. Remove supernatant.
3. To cross-link the antibody to the bead and prevent co-elution later with your protein of interest, add 250 μL of 5 mM suberic acid bis(3-sulfo-*N*-hydroxysuccinimide ester) sodium salt (BS^3) in 20 mM sodium phosphate, 0.15 M NaCl (pH 7.4). Incubate in this solution for 30 min (see Note 40).
4. Stop the reaction by adding 12.5 μL of 1 M Tris–HCl (pH 7.5), and incubate for an additional 15 min.
5. Place the tube on the magnet, remove the supernatant, and wash the beads three times with 200 μL PBS-T buffer (pH 7.4).
6. After removing the last wash, add between 100 and 1,000 μL of the crude protein extract and use this to resuspend the Dynabeads–antibody complex.
7. Incubate the solution between 10 and 120 min (see Notes 41 and 42).
8. Wash the Dynabead–antibody–antigen complex gently with 200 μL PBS buffer (pH 7.4) three times. Resuspend the beads in 100 μL buffer and transfer solution to a new tube (see Note 43).
9. Elution of the proteins can be done in a number of different ways. Typically, to analyze the various components of the protein complex, a denaturing elution is used. This can be done by heating 20 μL of denaturing elution buffer with the bead complex at 70°C for 10 min (see Note 44).
10. Remove the supernatant and denature the proteins at 95°C for 5–10 min followed by incubation on ice for 2 min.
11. Separate samples by gel electrophoresis (routine conditions) and pass to staining.

3.15. Silver Staining Compatible with Mass Spectrometry Analysis

(Adapted from Ref. 19)

1. Fix a newly run gel in gel fixative solution for 20 min with light shaking.
2. Rinse the gel in 50% methanol for an additional 10 min.

3. Wash the gel in distilled water with four changes of water, 30 min each rinse (see Notes 45 and 46).
4. Following the washes, sensitize the gel in 0.02% sodium thiosulfate for 1 min (room temperature, in the dark).
5. Wash the gel in distilled water two times, leaving the gel in the bath 1 min each time.
6. Replace the water with cold 0.1% silver nitrate and incubate at 4°C for 20 min.
7. Wash the gel for 1 min in fresh distilled water.
8. Change the gel chamber and wash once more in distilled water (see Note 47).
9. Develop the bands in a solution of 0.04% formalin and 2% sodium carbonate (see Notes 48 and 49).
10. Terminate the staining in 5% acetic acid (see Note 50).
11. The gel can be stored until further use in 1% acetic acid or can be dried using a gel dryer.

3.16. Transformation of L. bicolor

(As reported by Ref. 20)

1. Innoculate fresh colonies of *L. bicolor* on dialysis membranes (CelluSep T3, molecular weight cut-off of 12,000 ± 1,400) overlaying malt extract medium. Allow colonies to grow for 1 week at 20°C in the dark.
2. Transfer membranes to induction medium.
3. Pick and grow an *Agrobacterium tumefaciens* (Strain LBA1100 or AGL-1) colony containing the transformation vector until the culture has an $OD_{600} = 0.2$. Centrifuge the culture at 2250 × *g* at 15°C for 15 min, resuspend in an equal amount of liquid induction medium (without agar) and culture for an additional 6 h.
4. Apply 50 μL drops of bacterial culture directly on actively growing hyphae of *L. bicolor* (around the periphery of the colony) and incubate in the dark at 20°C for 4 days.
5. Transfer the colonies on dialysis membrane to selection medium (see Subheading 2.16). Grow in the dark at 20°C and transfer the cultures to fresh medium every 2 weeks (see Notes 51 and 52).
6. Once the wild-type control is dead or no longer growing, subculture actively growing *L. bicolor* transformants and transfer to fresh selection medium (see Note 53).

3.17. Populus Transformation

1. Once the gene of interest has been cloned into an appropriate plant expression vector (e.g. pORE; (21)), transfer the plasmid into Agrobacterium line GV101 (or another comparable strain). Grow a positive colony in 5 mL of YEP, with the

appropriate antibiotics for selection, for 48 h. On the same day that the colony is inoculated, harvest 50–60 internodes of in vitro 717-1B4 and cut each segment along the axis. Lay each segment, cut side down, on preculture medium (M1) (see Notes 54–57).

2. After the 48 h incubation from step 1, reinoculate 1–5 mL of the *Agrobacterium* culture in 100 mL of fresh YEP liquid medium supplemented with the appropriate antibiotics. Grow this culture until it reaches an OD_{660} of 0.2–0.3. Spin down the culture at 1,200 × *g* for 20 min at 4°C. Resuspend the bacterial pellet in 100 mL of modified liquid M1 medium (leave out agar and all components added after sterilization, supplement with 10 μM acetosyrigone) and shake for 1 h at 24°C.
3. Remove poplar internodes from preculture medium and place them in the *Agrobacterium* broth. Incubate explants in the broth at 24°C for 16 h in the dark with gentle agitation (125 rpm) (see Note 58).
4. Following incubation, decant the *Agrobacterium* supernatant and transfer the stem segments (cut side down) onto fresh M1 solid medium. Incubate in the dark at 24°C for 48 h.
5. After incubation, each explant should be surrounded by a bacterial colony. Remove the explants from the medium, trying not to take up too much of the bacteria, and transfer them to a sterile flask. Wash the explants with 100 mL of sterile water by shaking the flask vigorously for 1 min. Pass the solution and stem segments through sterile cheese cloth, recover the stems, and place them in a new sterile flask. Wash again with 100 mL of sterile water. Perform a total of seven to ten washes, with the last two washes carried out with water supplemented with cefotaxim (50 mg/100 mL).
6. After the last wash, transfer the segments onto M2 solid medium. Leave stem segments on this medium for 3 weeks at 24°C in the dark (see Notes 59 and 60).
7. After 3 weeks, transfer the stem segments to M3 medium and put the plates in the light, under long-day conditions (16 h light/8 h dark) (see Notes 61 and 62).
8. Once light green calluses reach a diameter of approximately 0.4 cm, transfer the calluses to regeneration medium.
9. When calluses start to produce shoots, transfer them to M3 medium supplemented with 0.01 μM TDZ.
10. As soon as the shoots are approximately 2 cm in length, cut them from the calluses and transfer them to ½ MS medium to root.

3.18. Analysis of Mycorrhization Potential

1. 2- to 3 months before the experiment is to be started, a fungal innoculum must be prepared. To do this, prepare 1 L of a 3:1 mixture of peat:vermiculite in a large glass jar with a vented cap. Autoclave this mixture dry.
2. Add 650 mL of Pachlewski medium and allow the medium to soak up into the peat for 1 h. Reautoclave this mixture.
3. Once the media has cooled, inoculate with roughly 50 2–4 mm^2 blocks of fungal covered agar. Try to inoculate many different levels within the jar and all around the circumference to get the most homogeneous growth of the fungus (see Note 63).
4. Leave the jar in the dark at 24°C for 2–3 months until the fungus has grown throughout the medium in the jar.
5. Once the culture has grown sufficiently, mix the fungal innoculum into well dampened Terra-Green at a dosing rate of 5–10% and transfer into 1 L pots (see Note 64).
6. Into each pot place a stem cutting of *Populus trichocarpa* that is one internode in length and lightly water the pot (see Notes 65–67).
7. Allow the plants to grow at 21°C in the light for 16 h during the day and at 18°C during the night for 2.5–3 months with light watering. After 1 month of coculture, fertilize the plants once per week with 20 mL of nutrient solution per 1 L of Terra-Green (see Notes 68 and 69).
8. After 3 months (in the case of *L. bicolor*), the plants should have an extensive root system with many active, healthy mycorrhizal root tips.

4. Notes

1. We have successfully used this extraction protocol with as few as five mycorrhizal root tips, but more tips are always preferable.
2. Generally, we fix *L. bicolor:P. trichocarpa* mycorrhiza for 16 h. The ability of paraformaldehyde to fix tissues is dependent on the pH of the solution. It is essential to check the final pH of the fixative before use; otherwise, there is a suboptimal preservation of tissue and proteins.
3. Due to the thick consistency of an agarose solution at this concentration, we delay the solidification of the solution by floating 1.5 mL tubes in a 60°C water bath filled with molten agarose solution. Remove tubes as needed from the water bath, add samples, and place immediately on ice to prevent movement of the sample.

4. A number of different antibody dilutions (usually between 1:50 and 1:2,000) should be tested.
5. As a control, some samples should also be incubated in IgGs purified from preimmune serum.
6. Primary antibodies against the target protein and the IgGs from the preimmune serum *must* be at the same concentration. A more concentrated preimmune serum may result in nonspecific binding.
7. As an additional control, samples should also be incubated overnight in PBS buffer containing 0.5% BSA without the primary antibody and then probed with the secondary antibody on the following day.
8. Perform this step in the dark to avoid excitation of the fluorochrome.
9. Alternatively, if appropriate, wash the samples once more in water and mount in 100 μM propidium iodide. This stains both plant cell walls as well as nuclei. Use *extreme* caution when using propidium iodide as it is a potential carcinogen.
10. This is a fairly stringent fixative and is meant to maintain subcellular structure. If the protein being probed is undetectable with this method, and if subcellular localization is not necessary, it is also possible to use 4% paraformaldehyde in sodium phosphate buffer (pH 7.4) with the addition of 0.5% glutaraldehyde or 3% paraformaldehyde plus 1% glutaraldehyde. Representative images of both a light fixation and a strong fixation are shown in Fig. 3.
11. Black staining of the tissue during this step also increases contrast during electron microscopy observation, making some cellular components easier to visualize.
12. Do not fix for longer than 1 h, as this will over-fix the tissue.
13. If a "softer" fixation method is being used to preserve antigenicity of the protein, this step can be left out. To test how fixation might affect the antigenicity of your protein, mix different concentrations of your protein (in recombinant form) with the chosen fixative and spot it on a nitrocellulose membrane. Allow the spot to dry and perform a standard Western dot-blot. Should fixation affect antigenicity, the detection limits of the "fixed" protein are lowered compared to a nonfixed control.
14. Depending on time constraints, it is possible to leave samples in 70% ethanol overnight at 4°C.
15. Make sure to add the appropriate amount of benzoyl peroxide as recommended by the manufacturer. This reagent catalyzes cross-linking of the resin in subsequent steps. After adding benzoyl peroxide, stir the solution overnight before use.

16. To ensure proper mixing of LR white, it is recommended to stir the LR white solution for 30 min, followed by 15 min still to eliminate bubbles, before use.
17. Since oxygen prevents polymerization of LR white, fill the tubes to the top and cap them tightly. Some of the top resin might remain unpolymerized, but after 24 h, the main part of the block should be properly hardened.
18. Before adding the sample to the capsule, put a thin layer of LR white along the bottom of the capsule. This helps to make sure that there is always a thin layer of LR white around all sides of the sample. Without this layer, some samples may adhere to the tube and fall out of the block during sectioning.
19. In the case of sensitive antigens, it is possible to polymerize the block at low temperature (from 4 to –20°C) under a UV lamp. With this approach, omit osmium tetroxide postfixation.
20. Due to the frequent presence of reverse transcription/amplification inhibitors in RNA preparations from mycorrhizal root tips, smaller quantities may improve the efficiency of the RT reaction.
21. Because of the different orientations of the fused ends of vector- and cDNA library-encoded polypeptides utilized for the different screens (N-term–C-term for SST; C-term–N-term for TAT and THS; see Table 1), two different starting libraries must be constructed: for SST a 5′-enriched cDNA library is prepared with an oligo(dN) primer in order to avoid stop-codons while for TAT, NTT, and THS a full-length library is prepared using an oligo(dT) primer.
22. cDNA libraries can be constructed conveniently in Gateway™ vectors (Invitrogen) that allow: (a) cloning in the desired orientation and without the use of restriction enzymes that may cut the cDNA of interest; and (b) easy transfer of an entire library from one vector to another, thus drastically reducing the amount of starting mRNA (and labor-time) that would be required to construct different libraries independently.
23. Once the optimal number of cycles (N) has been determined, the reaction is run again for N–1 cycles and the products from this reaction are used in the following steps.
24. In the case of pSUC2TM13ORI, the Gateway cassette must be cloned into EcoRI/XhoI sites in order to remove the entire polylinker and obtain the Gateway-compatible pSUC-GW vector.
25. In order to optimize transformation conditions, pilot-scale transformation experiments are initially performed with 1 μg of library DNA and different polyethylene glycol concentrations (33.3% ±2% and ±4%).

26. The SST assay is based on the expression of a modified invertase, lacking the signal peptide and thus unable to support growth on sucrose, in an invertase-less (*suc2*Δ) yeast mutant strain. If the protein of interest, fused to the N-terminus of the signal-less invertase, contains a functional secretion signal, invertase secretion and the ability to grow on a sucrose-containing (or otherwise restricted medium) is restored. This screen allows the experimental validation of putative secreted proteins identified by standard bioinformatic analyses (e.g. PSORT, SignalP, and TargetP) as well as the discovery of novel secreted proteins (reviewed in Ref. 22).
27. To ensure adequate spacing of the colonies, plate cells on 10–40 Petri dishes (150 mm diameter).
28. Growth in liquid SD can be avoided by spotting single colonies on a selective SD-agar plate in a 96×-format. This can be done by gently pressing the sterile 96-pin device on the agar plate to mark the position of evenly spaced spots.
29. The TAT assay is based on nuclear expression of the DNA binding domain (DBD) of yeast Gal4 with its NLS in a modified yeast strain harboring three reporter genes under the control of Gal4-regulated promoters. Due to the lack of an activation domain (AD), the DBD is not capable of autonomous transcription, but may regain this capacity when fused to a cDNA library-provided surrogate of the AD.
30. The TAT screen can be used to validate gene annotation, but also for the identification of novel transcription factors. The latter is especially valuable for at least two reasons: (a) while DBDs usually share fairly conserved folds, ADs are much more eclectic and are thus not so easily recognizable by bioinformatic analysis; (b) since an extensive gene expression reprograming is likely to occur upon fungus–plant interaction and mycorrhiza formation, transcriptional activators may act as master genes controlling this developmental transition.
31. ~90% of the initially selected colonies usually confirm their growth phenotype.
32. At this step, a 384-well microtiter plate can be assembled, which can then be conveniently replicated using a 384-multipinner device.
33. Despite this fairly stringent criterion, about 70% of the clones that pass initial selection are usually retained and subjected to sequence analysis.
34. In our experience, ~15–20% of TAT-positive clones code for proteins that are not capable to enter the nucleus autonomously (23). These false positives, which originate from the presence of a NLS in the standard Gal4-DBD construct used for TAT and likely reflect the lack of strict structural

requirements for activation, can be filtered out with a secondary screen named "NTT."

35. The NTT is based on the expression of the protein of interest in frame with an NLS-less artificial transcription factor. If the resulting fusion protein enters the nucleus, transcription of reporter genes (*HIS3* and *LacZ*) is activated, thus indicating that the TAT-positive clone is indeed a true transcriptional activator.
36. One of the hybrids, the "bait," contains the Gal4-DBD (as for the TAT) fused to a protein of interest; the other hybrid, the "prey," comes from a pDEST™22 library in which each foreign cDNA, from the tissue (or life-cycle stage) of interest, is fused to the Gal4-AD. If the "bait" and the "prey" interact, Gal4-dependent transcription factor activity is restored, reporter gene expression is activated, and can be detected on specific media.
37. The THS is based on the expression of two different hybrid proteins in the nucleus of a specific yeast strain. The 3AT concentration used for this screen must be optimized so that at the lowest 3AT concentration a particular "bait strain" fails to grow.
38. Add 1–5 μg of the primary antibody diluted in 200 μL of PBS-T buffer (pH 7.4). Allow the primary antibody to bind to the bead for 10 min at room temperature.
39. Due to the weight of the beads, they quickly separate out of the mixture. Therefore, gently mix the solution periodically throughout the incubation time, both for this step and for all the following incubation steps.
40. This cross-linking step is not necessary, but may help avoid masking of the proteins of interest by co-elution with the primary antibody.
41. Incubation time depends on the affinity of the antibody for its target protein. If longer times of incubation are utilized, it is best to perform this step at 4°C to help prevent protein degradation.
42. Longer incubation times result in a very quick increase in nonspecific binding; therefore, try to minimize contact time with the antibodies.
43. This transfer is to prevent elution of proteins which remained bound to tube walls despite washes.
44. If time does not permit elution and immediate analysis by gel electrophoresis, add the elution buffer to the Dynabead-antibody–antigen complex and freeze at –20°C. For subsequent elution, defrost the tube, heat at 70°C, and continue with step 11.

45. More washes or overnight washing will decrease the background, if this is an issue.
46. During the change from methanol to water the gel may float – make sure that at the end of the washes the gel sinks to the bottom of the incubation vessel.
47. If the chamber is not changed, silver that adheres to it will react with the developer in the next step.
48. The length of time it takes before the bands become evident depends on the expected recovery of protein.
49. It is important to watch the gel during the developing process and to change the buffer as soon as it turns slightly yellowish.
50. The staining reaction does not stop immediately, so as a general rule block the reaction slightly before the desired band darkness is achieved.
51. Ensure that a wild-type control is treated similarly to be certain that selection is working appropriately.
52. Typically subculture two to three times to ensure proper selection.
53. To ensure maintenance of transgene expression, always maintain stock cultures on selection medium.
54. It is advisable to prepare stock solutions of these compounds grouped according to macro- and microelements, vitamins, etc.
55. It is essential to use the proper type of iron-containing EDTA salt.
56. The poplar genotype affects the success of the transformation process, as well as how transformation is carried out. The protocol described here is especially suited for the poplar INRA clone 717-1B4, as in our hands it shows routine transformation success and is long-lived in in vitro culture.
57. It must be noted that poplar transformation and regeneration of a whole plant is a long process that takes many months. Therefore, it is important to carefully decide which gene(s) are to be transformed.
58. It is also possible to use light vacuum conditions to help infiltrate *Agrobacterium* into plant tissues. If vacuum-assisted infiltration is used, shorten incubation time to 4 h.
59. Cefotaxime is a bacteriostatic compound; thus, it will only inhibit *Agrobacterium* growth without actually killing it. For this reason, it is necessary to keep a very careful eye on any stem segments contaminated with *Agrobacterium* as they may contaminate all the other tissues in the same plate.
60. This dark period can be extended to 30 days.

61. As light will degrade antibiotics, the medium must be changed every 10–14 days to prevent *Agrobacterium* regrowth.
62. Calluses which start to turn brown must be immediately removed from the plates and discarded as they are not expressing the transgene of choice.
63. The size of innoculum blocks depends on the fungus being propagated. For *L. bicolor,* small-sized blocks are best since the fungus still grows well, while for other fungi larger blocks have to be used as smaller innocula grow very slowly.
64. In our experience, a fungal innoculum higher than 10% inhibits the formation of mycorrhizal root tips.
65. We typically use *P. trichocarpa* as it forms mycorrhizal root tips in between 40 and 50% of infected root tips.
66. Stems used for this application are typically harvested in winter when trees are dormant, stored at –6°C, and slowly warmed up to room temperature before use.
67. As an alternative to dormant cuttings, we have also prerooted *P. trichocarpa* stem segments in hydroponic solution for 1 week and then planted these plants into the Terra-Green: *L. bicolor* mixture. This is particularly useful when performing time courses in which the roots must come into contact with the fungus immediately.
68. If *L. bicolor* is the fungus under study, do *not* overwater the pots. Excessive watering will kill the fungus and a functional symbiosis will not be established.
69. Do not use too much nitrogen as the plant will not set up a symbiotic relationship with the fungus.

References

1. Molina, R., Massicotte, H.B., and Trappe, J.M. (1992) Specificity phenomena in mycorrhizal symbiosis: community-ecological consequences and practical implications. *In* Mycorrhizal functioning. E.M. Allen (editor). Chapman & Hall, New York, N.Y. pp. 358–423.
2. Buée, M., Reich, M., Murat, C., Morin, E., Nilsson, R.H., Uroz, S., and Martin, F. (2009) 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytol.* **184,** 449–456.
3. Martin, F., and Nehls, U. (2009) Harnessing ectomycorrhizal genomics for ecological insights. *Curr. Opin. Plant Biol.* **12,** 508–515.
4. Duchesne, L.C., Peterson, R.L., and Ellis, B.E. (1987) The accumulation of plant-produced antimicrobial compounds in response to ectomycorrhizal fungi: a review. *Phytoprotection.* **68,** 17–27.
5. Duchesne, L.C., Peterson, R.L., and Ellis, B.E. (1989) The time course of disease suppression and antibiosis by the ectomycorrhizal fungus *Paxillus involutus. New Phyt.* **111:**693–698.
6. Colpaert, J.V., and Van assche, J.A. (1992) Zinc toxicity in ectomycorrhizal *Pinus sylvestris. Plant Soil* **143,**201–211.
7. Galli, U., Schuepp, H., and Brunold, C. (1994) Heavy metal binding by mycorrhizal fungi. *Physiol. Plant.* **92,** 364–368.
8. Martin, F., Aerts, A., Ahren, D., Brun, A., Danchin, E.G.J., Duchaussoy, F., Gibon, J., Kohler, A., Lindquist, E., Pereda, V. *et al.* (2008) The genome of *Laccaria bicolor* provides

insights into mycorrhizal symbiosis. *Nature* **452**, 88–92.

9. Tuskan, G.A., DiFaxio, S., Jansson, S., Bohlmann, J., Grigoriev., I., Hellsten, U., Putnam, N., Ralph, S., Rombauts, S., Salamov, A. *et al.* (2006) The genome of the black cottonwood, Populus trichocarpa (Torr. & Gray). *Science* **313,** 1596–1604.
10. Busov, V.B., Meilan, R., Pearce, D.W., Ma, C., Rood, S.B., and Strauss, S.H. (2003) Activation tagging of a dominant gibberellins catabolism gene (GA2-oxidase) from poplar that regulates tree stature. *Plant Physiol.* **132,** 1283–1291.
11. Harrison, E.J., Bush, M., Plett, J.M., McPhee, D.P., Vitez, R., O'Malley, B., Sharma, V., Bosnich, W., Séguin, A., MacKay, J., and Regan, S. (2007) Diverse developmental mutants revealed in an activation-tagged population of poplar. *Can. J. Bot.* **85**, 1071–1081.
12. Jacobs, K.A., Collins-Racie, L.A., Colbert, M., Duckett, M., Golden-Fleet, M., Kelleher, K., Kriz, R., LaVallie, E.R., Merberg, D., Spaulding, V. *et al,* (1997) A genetic selection for isolating cDNAs encoding secreted proteins. *Gene* **198**, 289–296.
13. Balestrini, R., Hahn, M.G., Faccio, A., Mendgen, K., and Bonfante, P. (1996) Differential localization of carbohydrate epitopes in plant cell walls in the presence and absence of arbuscular mycorrhizal fungi. *Plant Physiol.* **111,** 203–213.
14. Gietz, R.D., and Schiestl, R.H. (2002) Transformation of yeast by lithium acetate/single-stranded carrier DNA/polyethylene glycol method. *Methods Enzym.* **350,** 87–96.
15. Titz, B., Thomas, S., Rajagopala, S.V., Chiba, T., Ito, T., and Uetz, P. (2006) Transcriptional activators in yeast. *Nucleic Acids Res.* **34,** 955–967.
16. Walhout, A.J., and Vidal, M. (2001) High-throughput yeast two-hybrid assays for large-scale protein interaction mapping. *Methods* **24,** 297–306.
17. Marshall, K.S., Zhang, Z., Curran, J., Derbyshire, S., Mymrik, J.S**.** (2007) An improved genetic system for detection and analysis of protein nuclear import signals. *BMC Mol. Biol.* **8,** 6.
18. Soellick, T.R., and Uhrig, J.F. (2001) Development of an optimized interaction-mating protocol for large-scale yeast two-hybrid analyses. *Genome Biol.* **2,** 52.
19. Mortz, E., Krogh, T.N., Vorum, H., and Görg, A. (2001) Improved silver staining protocols for high sensitivity protein identification using matrix-assisted laser desorption/ionization-time of flight analysis. *Proteomics* **1,** 1359–1363.
20. Kemppainen, M., Circosta, A., Tagu, D., Martin, F., and Pardo, A.J. (2005) *Agrobacterium*-mediated transformation of the ectomycorrhizal symbiont *Laccaria bicolor* S238N. *Mycorrhiza* **16,** 19–22.
21. Coutu, C., Brandle, J., Brown, D., Brown, K., Miki, B., Sinnomds, J., and Hegedus, D.D. (2007) pORE: a modular binary vector series suited for both monocot and dicot plant transformation. *Transgenic Res.* **16,** 771–781.
22. Nombela, C., Gil, C., and Chaffin, W.L. (2006) Non-conventional protein secretion in yeast. *Trends Microbiol.* **14,** 15–21.
23. Montanini, B., Levati, E., Bolchi, A., Kohler, A., Morin, E., Tisserant, E., Martin, F., and Ottonello, S. (2011) Genome-wide search and functional identification of transcription factors in the mycorrhizal fungus Tuber melanosporum. *New Phytol.* **189,** 736–750.

INDEX

A

Affinity purification 157–165, 200, 231
Affymetrix GeneChip 61, 63, 64, 66, 80, 84–85, 92–94, 97
Agrobacterium tumefaciens 213–223, 272
Annotation of repetitive sequences 33–47
ATMT (*Agrobacterium tumefaciens* mediated transformation) 213–223

B

Beta-galactosidase 14
BLAST2GO 107, 114–116, 119
BLAST search 107, 111–115, 119, 147, 151

C

Calf Intestinal Alkaline Phosphatase (CIAP) 14, 22, 170, 174, 227, 230–231
CAP3 ..... 106, 109, 149–151
Casein kinase II (CK2) 227, 230–231
cDNA library 252, 255–257, 263, 276–277
Cellular proteins 134, 136
ChIP-chip 121–130
ChIPed DNA 124, 127–130
ChIP-seq 59
Chromatin immunoprecipitation (ChIP) 123–124, 126–127, 130
Co-cultivation 216, 218–219, 223
Comparative genomics 58
Creeping bentgrass 167–169, 171, 172, 175, 177
Cryphonectria parasitica 225–235
C-value paradox 34

D

Digital northern analysis 175–176
DNA fragmentation 7, 123–124, 126–127, 130
DNA sequencing
  Illumina/Solexa sequencing 6, 53–54, 118, 167–177, 260
  next generation sequencer (NGS) 2, 5–8, 52–53, 57–59, 168
  pyrosequencing 5–6, 52–53, 106, 143
  Sanger sequencing 2–5, 8, 51–53, 56, 106, 142, 143, 148, 153
  454 sequencer 5–6, 52–53, 106, 116–118
  single molecule real time sequencing (SMRT) 7
  single molecule sequencing 7–8
  SOLiD 6–7, 54–58, 106, 168
Dollar spot 167–168, 176, 177

E

Ectomycorrhizae (EM) 250–252
Ectomycorrhizal symbiosis 249–280
Electroblotting 13, 16–17
Epichloë festucae 20
Epitope tag 160, 199–211
EST analysis 103–119
Expressed sequence tag (EST) 103–119
Expression analysis 59, 79–99, 251
Expression profiling 59, 65, 122

F

Filamentous Fungal Gene Expression Database (FFGED) 62, 64
FLAG tag 158, 160, 203, 210
Functional annotation 107, 118
Fungal biodiversity 141–153
Fusarium GeneChip 80, 84, 95
Fusarium graminearum 79–99, 200
Fusarium verticillioides 237–247
Fusion construct
  GFP 123, 129, 199–211
  3 × FLAG 158, 160, 164, 203, 204, 210
  13 × Myc 200, 203, 204, 210

G

Gas chromatography (GC) 238
GC-MS analysis 240–242
Gene
  deletion 157, 180, 191–197, 199, 238
  disruption 199, 238
  knockout 180

Jin-Rong Xu and Burton H. Bluhm (eds.), *Fungal Genomics: Methods and Protocols*, Methods in Molecular Biology, vol. 722,
DOI 10.1007/978-1-61779-040-9, © Springer Science+Business Media, LLC 2011

Genome assembly 1–8, 11, 20, 58
GFP 122, 123, 129, 130, 199–211
Gibberella zeae 79

H

Hartig net 250
Head blight 79
High-throughput gene knockout 180, 186–188
Homokaryon 182, 185–187, 192
Homologous recombination
in *Magnaporthe* 200
in *Neurospora* 180, 183
HPLC/MS analysis 135, 137–139
Hypovirulence 226
Hypovirus 225–235

I

Immunofluo-localization 254, 260
Immunogold localization 254–255, 261–262
Insertional mutagenesis 213–223
Interactome 176
Internal transcribed spacer (ITS) 142, 143, 146–151
InterProScan 107, 110, 113–114
In vitro transcription 227–228

K

Knockout cassettes 180, 184, 186

L

Laccaria bicolor 250
LSU (large subunit) rRNA 143, 147–148, 152

M

Magnaporthe oryzae 64, 122, 157, 200, 213–223
Mass spectrometry (MS) 134, 157–165, 200, 238, 258, 271–272
Metabalome 246
Metabolic fingerprinting 237–247
Metabolomics 238
Metagenomics 146, 147, 149, 151
Microarray analysis
BAGEL 63, 64, 70, 72–77
experimental design 62
Genespring 62
hybridization 62, 65, 68, 72, 74
hypothesis tests 61, 62, 70
Limma 62–72, 76–77
normalization and scaling 61, 62, 69
validation 61, 63
MoCRZ1 binding sites 122, 130
mus–51 or *mus–52* 180, 183, 184, 186
Mutagenesis 213–223
Mycotoxin 79

N

Neurospora crassa 37, 179–188, 191–197
Nuclei isolation 123, 126

O

Operational taxonomic unit (OTU) 147, 149–152

P

Peptide identification 138
Perithecia 95, 96, 192, 196
Phenotypic analysis 191–197
Polymerase chain reaction (PCR) 4, 13, 17, 52, 55, 88, 124, 127–132, 141, 142, 144, 146, 160, 170–171, 174, 177, 180, 183, 184, 201, 204, 206, 208–210, 217, 220, 221, 232, 254, 255, 257, 260, 263, 264, 266, 270
Protein complex 59, 157–165, 271
Protein phosphorylation
de-phosphorylation 174, 227, 230–231
re-phosphorylation 227, 230–231
Protein-protein interaction 157, 158, 165, 253, 269
Proteome studies 133–139
Proteomics 134, 135, 138
Protoperithecia 95, 192, 196
Protoplast
preparation 82, 96
transformation 82, 87–89
Pseudogene 34, 35, 46, 142

Q

qRT-PCR 38, 97

R

RECON 39, 43
Repetitive sequence
annotation 33–47
identification 33–47
Resequencing 55, 57, 58
Rice blast fungus. *See Magnaporthe oryzae*
RNase
digestion 231, 233
mapping 231–233
RNase A 127, 228, 231–233, 270
RNase I 228, 231–233
RNase T1 .....228, 231–233
RNase V1 .....228, 231, 232

S

Sclerotinia homoeocarpa 167–169, 171–173, 175–177
Secondary metabolism 237, 238
Second generation sequencer 55
Secreted proteins 134, 136, 251, 277

Segmental duplications (SD) 34, 45–46
Selection marker
bleomycin 200, 203, 204, 209–211
geneticin 200, 203, 204, 210
hygromycin 82, 89, 160, 180, 182, 186, 193, 194, 200, 203, 204, 210, 211, 214, 216, 219–220
Sequence assembly 4–5, 118, 143, 144, 149
Sequencing-by-synthesis (SBS) 167–177
Shuttle vector 200, 203, 204, 207, 209
Silver staining 258, 271–272
SSU (small subunit) rRNA 141, 142, 147

T

TAIL-PCR 214, 217, 220–222
Tandem repeats 4, 34, 41, 45
T-DNA 214, 220–223
Telomere
cloning 11–30
containing clones 26, 27
hybridization 13, 15, 18–22, 25, 26, 29
oligonucleotides 13, 29
probe 17–22, 25, 26
sequences 12
Telomeric fragment 20, 23, 28
TERMINUS 12
Transcriptome
analysis 177
sequencing 53, 58–59
Transcriptomics 76
Transposable elements (TEs)
DNA transposon 36–37
retrotransposon 35–37
Trap assay
nuclear target 252
signal sequence 251
transcriptional activator 251, 267, 277, 278
Trichoderma 82, 105
Tryptic digestion 134, 136–137

W

Whole genome amplification 124, 128–129

Y

Yeast
colony PCR 208, 257, 270
competent cell 201, 207, 210
gap repair 164, 200, 210
plasmid miniprep 257, 270
recombinational cloning 180
transformation 158, 160, 201–202, 207–208, 210, 263–265
two hybrid 252, 256, 263, 268
in vivo recombination 180