Exploring Genome Structure and Gene Regulation Related to Virulence in Fungal Phytopathogens
Using Next Generation Sequencing Techniques

Dissertation

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the
Graduate School of The Ohio State University

By
Jinnan Hu, M.S.
Graduate Program in Plant Pathology

The Ohio State University
2013

Dissertation Committee:
Thomas K. Mitchell, Adviser
Michael J. Boehm
Guo-liang Wang
Kun Huang
Abstract

Over the last decade, a technological revolution has provided enormous advances in the knowledge of complex biological processes largely enabled by the development of next-generation sequencing (NGS) techniques. Applications of NGS include studies of entire genomes, characterization of the entire transcriptome (RNA-Seq), and detection of protein-DNA binding sites (ChIP-Seq). The cost of sequencing a fungal genome, for example, has decreased from more than one hundred thousand dollars to currently only three thousand dollars. With the development of the applications and the affordable cost, NGS is changing the way biologists designing and carrying out research. This dissertation describes developed analysis pipelines of NGS data in the field of fungal phytopathogens using four different projects as examples. In a genome comparison project, sequenced short reads are \textit{de novo} assembled to form a genome draft, then gene models are predicted either \textit{ab initio} or assisted by RNA-Seq, followed by the comparison between genomes at different resolutions such as the nucleotide level and the genome structural level. Two chapters in this dissertation serve as examples of our genome comparison pipeline put to work to address biological questions. In Chapter 2, the \textit{Alternaria arborescens} sequences of the unique conditionally dispensable chromosome (CDC) were separated from essential chromosomes (EC) using a novel bioinformatics approach. A pair-wise
comparison between the CDC and ECs showed that CDC sequences had significant variation and that it may have been originally acquired through a horizontal gene transfer event. In Chapter 4, seven field isolates of *Magnaporthe oryzae* were sequenced and their genome content compared. Over 10,000 SNP and Indel locations were identified as well as genes under strong positive selection, which are considered potential virulence related genes. While in a RNA-Seq analysis pipeline, sequence reads are first assembled *de novo* or mapped to a reference genome, and the expression level for individual genes in each sequencing library is calculated to identify differentially expressed genes. Chapter 3 describes a RNA-Seq analysis project, in which the transcriptome profile in the dollar spot pathosystem has been sequenced using a combination of Illumina and Roche 454 NGS technologies. A large number of genes were found up-regulated during the interaction between *Sclerotinia homoeocarpa* and *Agrostis stolonifera* with some having annotations suggesting their roles in virulence related processes. With regard to protein-DNA binding, reads from ChIP-Seq experiment are mapped to the reference genome and the “peak regions” of mapped reads are identified as candidate binding regions, within which binding motifs are predicted. Chapter 5 describes the identifications of the binding sites and motifs of the *M. oryzae* transcription factor MoCRZ1 using a combination of ChIP-chip and microarray data, and then the prediction accuracy is improved by a novel approach utilizing the spatial distribution pattern of the candidate motifs. Finally, the last chapter summarizes a large
scale mutagenesis project to identify two avirulence genes in *M. oryzae* by generating random mutants, followed by a pathogenicity screen on rice cultivars containing the corresponding resistance genes. Although imperfections and challenges remain, this dissertation shows four successful NGS applications in fungal phytopathogens. With the continued development of sequencing techniques and bioinformatics tools, NGS based projects with more sophisticated experimental designs will undoubtedly produce larger and more accurate data to biologists in the near future.
Dedication

Dedicated to the students at The Ohio State University
Acknowledgments

I would like to gratefully and sincerely thank Dr. Thomas Mitchell for his guidance, understanding, patience, and most importantly, his friendship during my PhD studies. His mentorship was paramount in providing a well-rounded experience consistent with my long-term career goals. He encouraged me to not only grow as a plant pathologist but also as an instructor and an independent thinker. For everything you’ve done for me, Dr. Mitchell, I thank you.

I would also like to thank all the three professors served in my dissertation committee. Dr. Michael Boehm, Dr. Guo-Liang Wang, and Dr. Kun Huang, for their assistance and guidance in getting my graduate career started on the right foot and providing me with the foundation for becoming a qualified scientist.

I would like to thank the faculties of Department of Plant Pathology, and students of PPGSA for their valuable discussions and friendship. In particular, I would like to thank Dr. Stephen Opiyo for his data analysis help, Maria Bellizzi and Dr. Songbiao Chen for the guidance and support in my bench experiments, and Dr. Angela Orshinsky, with whom I worked closely and finished a publication together.

Finally, and most importantly, I would like to thank my wife Chenxi Chen. Her support, encouragement, quiet patience and unwavering love were undeniably the bedrock upon which the past 4 years of my life have been built. I would never reach this far without her company and support. I thank my parents, for their faith in me and allowing me to be as ambitious as I wanted.
Vita

2008........................................................................................................B.S. Biochemistry, Nanjing University, China

2008 to present...................................................................................Graduate Research Associate, Department of

Plant Pathology, The Ohio State University

Publications


Fields of Study

Major Field: Plant Pathology
## Table of Contents

Abstract ........................................................................................................................................... ii  
Dedication ...................................................................................................................................... v  
Acknowledgments ....................................................................................................................... vi  
Vita ................................................................................................................................................ vii  
List of Tables .................................................................................................................................... ix 
List of Figures ............................................................................................................................ xii  
Chapter 1: Review of the knowledge obtained from whole genome studies of fungal phytopathogens by applying next-generation sequencing technologies ........................................ 1  
Chapter 2: Genomic characterization of the conditionally dispensable chromosome in *Alternaria arborescens* provides evidence for horizontal gene transfer .............................................. 39  
Chapter 3: RNA-Seq Analysis of the *Sclerotinia homoeocarpa* – Creeping Bentgrass Pathosystem ........................................................................................................................................ 87  
Chapter 4: Whole genome sequencing of seven *Magnaporthe oryzae* field isolates reveals genome content variation and predicted pathogenicity determinants ................................................. 127  
Chapter 5: A Distribution Pattern Assisted Method of Transcription Factor Binding Site Discovery for Both Yeast and Filamentous Fungi ........................................................................................................ 158  
Chapter 6: Identification of novel Avr genes by Screening of Gain-of-virulence Mutants in rice blast fungus *Magnaporthe oryzae* ........................................................................................................ 187  
Conclusions .................................................................................................................................. 211  
References ...................................................................................................................................... 215
List of Tables

Table 1.1. Genomes sequenced of plant pathogen fungi and oomycetes .............................................. 32
Table 1.2. Comparison of the three mainstream 2nd generation sequencing technologies and PacBio which is regarded as the 3rd generation sequencing technique. ................................. 33
Table 1.3. Comparison of the technique specification and application of RNA-Seq, tilling microarray, and low-throughput techniques .......................................................................................... 34
Table 2.1. Velvet de-novo assembly statistics ............................................................................................. 67
Table 2.2. Alignment of A. arborescens marker gene contigs and A. brassicicola contigs .................. 68
Table 2.3. Repeat region identification ..................................................................................................... 69
Table 2.4. Codon usage correlation analysis .............................................................................................. 70
Table 2.5. Differences in codon usage between CDC and EC genes ....................................................... 71
Table 2.6. Ka/Ks ratio of CDC protein conversed domains ........................................................................ 72
Table 2.7. General statistical facts of A. arborencens and other sequenced fungi genome ................. 73
Table 2.8. Conserved domains in CDC putative PKS genes .................................................................... 74
Table 2.9. Primers used for Southern hybridization .................................................................................. 75
Table 3.1. Characteristics of the 454 RNA-seq reads .............................................................................. 110
Table 3.2. Mapping characteristics of the Sclerotinia homoeocarpa (SH) and Agrostis stolonifera
(AS) SBS reads to the SH and AS transcript assemblies ................................................................. 111

Table 3.3. Summary of select Sclerotinia homoeocarpa transcript types that are significantly increased at 96 hpi after inoculation onto creeping bentgrass ...................................................... 112

Table 3.4. Summary of select Agrostis stolonifera transcript types that are significantly increased at 96 hpi after inoculation with Sclerotinia homoeocarpa ......................................................... 113

Table 3.5. Top enriched Sclerotinia homoeocarpa domains at 96 h post inoculation on creeping bentgrass ........................................................................................................................................... 114

Table 3.6. Top enriched Agrostis stolonifera domains at 96 h post inoculation with Sclerotinia homoeocarpa ........................................................................................................................................... 115

Table 3.7. Simple sequence repeat (SSR) markers from the Sclerotinia homoeocarpa and Agrostis stolonifera transcriptome ............................................................................................................. 116

Table 4.1. General statistic numbers of the genome assembly and gene prediction ..................... 148

Table 4.2. Top eight genes containing the copy number variation .............................................. 149

Table 4.3. Number of SNPs & IDPs identified, their location, and the number of those causing large impact on gene function ........................................................................................................ 150

Table 5.1. List of yeast TFs involved in training and testing ............................................................ 174

Table 5.2. TFs in training group ........................................................................................................ 175

Table 5.3. Fkh1 consensus motifs testing result .............................................................................. 176
Table 5.4. Test results of all testing group ................................................................. 177

Table 5.5. Statistical enrichment test for the three candidate motifs .............................................. 178

Table 5.6. The occurrence of the three candidate motifs in ORFs and upstream regions ............... 179

Table 5.7. MoCrz1 consensus motifs testing result ........................................................................ 180

Table 6.1. Pathogenicity test conducted in Dr. Mitchell’s lab .......................................................... 199

Table 6.2. Pathogenicity test conducted in Dr. Pan’s lab .................................................................. 200

Table 6.3. Summary of TAIL-PCR results of different clones generated from Iso_2 and Iso_9 ........ 201

Table 6.4. Features of 6 candidate Avr genes .................................................................................. 202

Table 6.5. Calculation of possibility to identify one specific gene or two genes using random
insertion approach in different number of mutants generated ...................................................... 203

Table 6.6. Primers used in TAIL-PCR........................................................................................... 204
List of Figures

Figure 1.1. Number of genomes available for the fungi and oomycete species ......................... 35
Figure 1.2. Sequencing cost released by The National Human Genome Research Institute ....... 36
Figure 1.3. Contribution of different stages to the total cost of a genome sequencing project ... 37
Figure 1.4. A general pipeline for genome comparison analysis used for this dissertation ......... 38
Figure 2.1. Annotation of the initial two CDC contigs ................................................................. 76
Figure 2.2. Southern hybridization to validate CDC contigs prediction ..................................... 77
Figure 2.3. Global contig alignment between \textit{A. arborescens} contigs (blue blocks) and \textit{A. brassicicola} contigs (red blocks) .......................................................... 78
Figure 2.4. The Codon Adaptation Index (CAI) distribution of \textit{A. arborescens} CDC genes compared to EC genes .......................................................... 79
Figure 2.5. BLAST taxonomy report of all CDC genes against NCBI “nr/nt” database .............. 80
Figure 2.6. GO term of CDC genes ............................................................................................. 81
Figure 2.7. Structure of predicted secondary metabolite biosynthesis (SMB) clusters .......... 82
Figure 2.8. Average Ka/Ks ratio, Ka, and Ks for all aligned EC genes and CDC genes .......... 83
Figure 2.9. Phylogenetic relationship of proteins in \textit{A. arborescens} CDC shows discordance compared to EC proteins .......................................................... 84
Figure 2.10. Scatter plots of BLAST Score Ratio (BSR) of EC proteins (A) and CDC proteins (B) ... 85

Figure 2.11. Diversity of marker genes on CDC and ECs among different Alternaria isolates .... 86

Figure 3.1. Read length of 454 sequencing data. The range of sequence read length (bp) of 454 data is shown as the shaded area ................................................................. 119

Figure 3.2. Signal tracks of SBS reads for two genes (SH3549 and SH8608) from either the PDA, PDB, or interaction library ........................................................................................................... 120

Figure 3.3. Venn diagram of RNA-Seq reads unique and common to the SH library, AS library, and Interaction library .................................................................................................................. 121

Figure 3.4. Top Species Blast hits: (a) SH transcript library (b) AS transcript library ................. 122

Figure 3.5. GO terms associated with up-regulated SH transcripts ......................................... 123

Figure 3.6. GO terms associated with up-regulated AS transcripts ........................................ 124

Figure 3.7. Comparison of semi-quantitative RT-PCR data to RPKM values ......................... 125

Figure 3.8. Validation of RPKM data for selected SH and AS transcripts using real time PCR .... 126

Figure 4.1. Dot plots of the alignments among 7 sequenced isolates and reference 70-15 sequence ...................................................................................................................................... 151

Figure 4.2. Phylogenetic tree of different isolates constructed using genome-wise SNP data .................. 152

Figure 4.3. Gene pool analysis among the four isolate(s) groups .............................................. 153

Figure 4.4. Copy number variation of the 7 sequenced isolates along genome ...................... 154
Figure 4.5. Number of Genes with higher/lower copy number compared to 70-15 .................. 155

Figure 4.6. SNPs & IDPs density (number per 100KB) along genomes of 7 sequenced isolates.156

Figure 4.7. Ka/Ks ratio distribution along genome in the 7 sequenced isolates ....................... 157

Figure 5.1. The MoCRZ1 binding region identification experiments design ............................. 181

Figure 5.2. Analysis Pipeline for generating model of MDP approach and test it using yeast transcription factor data ................................................................................................................. 182

Figure 5.3. Distribution frequency curve of training group .......................................................... 183

Figure 5.4. Frequency curves of Fkh1 candidate motifs .............................................................. 184

Figure 5.5. Predicted MoCRZ1 binding motifs identified ......................................................... 185

Figure 5.6 Frequency curves of MoCrz1 candidate motifs ....................................................... 186

Figure 6.1. Screening Pipeline ................................................................................................. 205

Figure 6.2. Lesions in the first re-inoculation of the 5 isolates from pooled inoculation .......... 206

Figure 6.3. Lesions from pathogenicity test conducted in Dr. Mitchell’s lab ......................... 207

Figure 6.4. All four candidate mutants showed strong GFP signal ....................................... 208

Figure 6.5. Position of the T-DNA insertion in Iso_2 and Iso_4 mutants ................................. 209

Figure 6.6. Structure of the T-DNA inserted ............................................................................. 210
Chapter 1: Review of the knowledge obtained from whole genome studies of fungal phytopathogens by applying next-generation sequencing technologies
With the rapid development of the next-generation sequencing (NGS) technologies, genome sequencing of many organisms in every branch of the life tree has been completed. With regard to fungi and oomycetes, which were often regarded as containing the simplest eukaryotic genomes, the genome of *Saccharomyces cerevisiae* was the first sequenced in 1996 [1] followed shortly by *Neurospora crassa* and *Magnaporthe grisea* in 2000 and 2002 respectively [2, 3]. Following on the success of these early efforts, more than 100 fungi and oomycete species have had their genomes sequenced, with most completed using NGS technologies. This wealth of data enabled many genome comparison studies, including species that cause diseases of crop plants. Specifically, much work has focused on the identification of virulence determinants and their evolution. NGS technologies have also been applied in studies focusing on the transcriptome (RNA-Seq) or DNA-protein interactions (ChIP-Seq). This chapter reviews some of the important projects of these efforts, summarizes the biological discoveries, and notes the technical advances made along the way. Also discussed is the accuracy, reproducibility and cost of NGS studies.

**Discoveries made using next-generation sequencing technologies**

**Genomes sequenced**

As of January 2013, there are 2,298 genomes uploaded in the National Center of Biotechnology Information (NCBI) database for Eukaryotic organism, with 720 of them belong to fungi, of which 406
are evolved to the “Complete” category. Of the fungal genomes, 306 belong to the phyla Ascomycota, 70 belong to Basidiomycota, and 32 placed in the “Other” category. Some species have multiple genomes updated for different isolates (Figure 1.1). These are mostly model organisms such as *Saccharomyces cerevisiae*, which has genomes available for as many as 52 isolates.

**Large variation of genome size**

It has been widely known from long before the application of NGS, that fungal genomes are variable with regard to length, but the degree of variability is only now being confirmed following NGS projects. (Table 1.1) Fungal phytopathogens and oomycetes have extreme variability in genome size. For example, the genome size difference between smut fungi *Ustilago maydis* (~19-21Mb) and *Phytophthora infestans* (~220-280Mb) is as large as nearly 15-fold [4-7]. Generally, filamentous fungal phytopathogens have larger genome than their yeast type relatives [4, 8, 9]. For example, the ascomycete fungi powdery mildews (*Golovinomyces orontii*) have the largest genome in that phyla – around 160 Mb [8]. For basidiomycetes, the rust fungus (*Melampsora larci-populina*) has the largest genome at over 89 Mb [9]. Some oomycete pathogens have genomes that are around or larger than 100 Mb, including downy mildew (*Hyaloperonospora arabidopsis*) with ~100 Mb genome [10], and the *P. infestans* clade (~240 Mb) [4, 7]. Compared with pathogens, the genomes of non-pathogenic fungi sequenced so far are more likely to stay at the size of 40 Mb, such as *Aspergillus oryzae* at 37 Mb [11], *Neurospora crassa* at 41 Mb [3] and *Schizophyllum commune* at 39 Mb [12]. The closest
non-pathogen relatives of oomycetes – diatoms – have the genome size ranging from 56 Mb for *Aureococcus anophagefferens* [13] to 27 Mb for *Phaeodactylum tricornutum* [14]. However, the trend of fungal phytopathogens towards larger genomes is not absolute, and some filamentous pathogens actually have relatively small genomes, possibly because of gene loss as in *Albugo laibachii* [15], intron loss as in *U. maydis* [5], or reduced transposon content in *Sclerotinia sclerotiorum* [16].

An interesting point discovered through NGS projects was the number of gene models do not correlate with the genome size. An example is the *Blumeria graminis* which has the genome about 6-fold larger than that of *U. maydis* but contains fewer gene models [5, 8]. Reasons for this lack of correlation is the fact that most expansion in phytopathogen genomes is not the coding regions but repetitive DNA such as low complexity repeats or transposons [2, 4, 8-10, 17]. In some extreme cases, the amount of repetitive DNA can reach as high as 74% in *P. infestans* and 65% in *Leptosphaeria maculans*.

“Lineage-specific” chromosomes & sequences

An interesting finding associated with the genome size variation is “lineage-specific” chromosomes, genome regions, and gene families. Firstly, comparative genome studies revealed that large genomic regions may be variable among isolates of a given species. Category of these variable regions is
unique chromosomes are referred to as supernumerary or conditionally dispensable, because they are not typically required for saprophytic growth [18-20]. These chromosomes have been identified in many fungi including *Magnaporthe oryzae* [2, 21, 22], *Fusarium oxysporum* [23], *Nectria haematococca* [24, 25], *Mycosphaerella graminicola* [26], *Cochliobolus heterostrophus* [27], *Leptosphaeria maculans* [28], and *Alternaria alternata* [29, 30]. In addition, “lineage-specific” regions can also be found as part of a chromosome, or even distributed in multiple chromosomes as single or multiple copies [25, 31-34]. Some gene families were identified only in the pathogenic species but not in their non-pathogen relatives, which are then regarded as pathogenicity related genes [32]. Such expanded genes may encode putative transporters and lytic enzymes necessary for pathogens to infect hosts. A recent study on the the poplar leaf rust *M. larci-populina* and wheat stem rust *P. graminis f. sp. tritici* found that around 26% of the predicted genes are lineage-specific, including peptidases, lipases, glycosyl hydrolases, ion superoxide dismutases, oligopeptide membrane transporters, kinases and transcription factors [9]. Additionally secreted protein coding genes were found upregulated during infection [9]. By comparison, biotrophic fungal species often lack or have a reduced number of these gene families [5].

In some cases, the lineage-specific genes are found sequestered on the “lineage specific” chromosomes, making the question of their evolution more straight forward. For example, in a study of *Magnaporthe oryzae*, a 1.68 Mb isolate specific genome region was identified, with 3 AVR genes were found in the region [21]. Another example is the pea-pathogenic form of *N. haematococca*, which contains genes involved in host nutrient utilization and resistance to plant antimicrobials, that
all locate on the “lineage-specific” chromosome [24, 35, 36].

Pathogenicity gene cluster

In fungal phytopathogens, genes encoding or being involved in pathogenicity sometimes co-locate in clusters [37, 38]. A typical cluster may include a set of co-regulated genes, or genes that code for biosynthetic enzymes, transcription factors or transporters [39]. An example is the gene cluster in *Fusarium graminearum*, which contains 10-12 genes locates next to each other and 2 genes locate outside the cluster [40]. In *Alternaria alternata*, metabolite gene clusters were found on the “lineage-specific” chromosomes, such as AF-toxin, and AK-toxin [41, 42]. A slightly different case in *U. maydis* genes encoding secreted proteins were not located in a large cluster but in small sub-clusters containing 3-26 co-regulated genes dispersed throughout the genome. Deletion of nine genes in the clusters resulted changing of virulence towards its host maize [5, 6].

Not all the pathogenicity related genes are clustered, but most of them show some preference in genome location, including “lineage-specific” chromosome or “conditionally dispensable chromosome” (CDC), and telomere-proximal regions. One shared feature of these regions rich with virulence related genes is the break or absence of synteny with their related species when the genome sequences are aligned [43]. As mentioned above, host-specific toxin genes in *Alternaria* pathogens are located at the CDC [44], while three AVR genes were found in the 1.68 Mb additional
genomic regions of the rice blast fungus *Magnaporthe oryzae* [21]. Subtelomeric regions are also locations preferred by virulence related genes, since usually they tend to evolve at a higher rate than sequences closer to the centromere [43, 45, 46]. An example also from *M. oryzae* is the avirulence gene Pita (Avr-Pita), which was found originally to reside in an unstable location only 48bp from the telomeric repeats of Chromosome 3 [45]. It was then reported recently that multiple translocation events were identified for Avr-Pita in multiple *M. oryzae* species, with this gene found most of the time located near or within retrotansposon-rich subtelomeric regions [47].

**Genome structures**

Focusing on the regions outside coding sequences, researchers found some interesting points of genome structure: gene-sparse regions, isochore-like regions, and transposable elements. Several filamentous fungal phytopathogens were found to contain a discontinuous distribution pattern of gene density along genome: gene-dense regions and gene-sparse regions. This pattern is particularly obvious in the genome of the oomycete *Phytophthora infestans*, in which two types of regions could be distinguished by the length of flanking noncoding regions. Roughly 2,000 gene-sparse regions, which typically contain less than 10 genes, were identified [4, 7, 48]. It should be noted that gene-dense and gene-sparse regions do not differ in GC-content. Differences in GC-content, however, generated “isochore-like regions” in the genome of *L. maculans* – pathogen of Brassica spp. In this case a sharply changing of GC-content in the isochore-like region was identified compared to the rest
of genome, with these regions lacking of coding genes, and have only 148 out of 12,469 (1.2%) genes located in them [17]. It has been proposed that both gene-sparse regions and isochore-like regions were created by invasion of transposable elements [17, 49].

Transposable elements (TE) were often proposed as a driver of phytopathogen genome evolution because of the fact that they mediate chromosomal rearrangements by ectopic recombination. So studies applying NGS, and focusing on TE location were performed, with the goal to elucidate mechanism of TE events and the genome changes they cause. The first finding is the extreme variation of the proportion of TEs in different genomes. In F. graminearum, only 0.1% of total genome and only two classes of TEs were identified [43]. U. maydis is another example of a genome that lacks a large number TEs with only 1.1% presented [5]. Other fungi show modest amounts of TEs, such as 9.7% in M. oryzae [2], and 28% in F. oxysporum [23]. On the other side, powdery mildew fungi show about 64% TEs which caused huge genome expansion and possible gene losses as it has only ~5900 genes identified [8]. TEs were found enriched in supernumerary chromosomes and telomere-proximal regions [43, 50], as being reported in F. oxysporum f.sp. lycopersici with 74% of the TEs located on the supernumerary chromosomes [23].
Mechanisms leading to genome modification

Single nucleotide mutations

Genome structures can be modified by a variety of mechanisms throughout evolutions. Although the mechanism could not be directly studied by NGS genome comparison studies, data from analyzing genome sequences can be used as evidence to support or reject assumptions of what the driving force is. DNA point mutations are one of the most direct mechanisms that lead to genome sequence changing and the alternation of protein function. Usually more than 10,000 Single-nucleotide polymorphisms (SNPs) can be identified in a fungal phytopathogen genome when multiple isolates are compared to a reference. SNPs are not distributed evenly along genome, but in fact several “hot spots” can be identified near telomeres and discrete AT-rich chromosome regions, as showed in the *F. graminearum* genome [43, 51]. Additionally the ratio of synonymous to non-synonymous SNPs (dN/dS) is valuable and can be calculated to serve as a signature of local selection force. The dN/dS ratio can be calculated for individual genes, a genome regions, or the whole genome, with the higher the ratio representing the stronger positive selection force [7, 15, 26, 52]. Usually high values are observed in regions correlated with high TE density or in the CDCs. Individual genes, or domains of particular genes, showing high dN/dS values can often be the ones associated with pathogenicity or virulence effectors. As reported in the RXLR containing effector genes in oomycetes, C-terminal domains showed much higher dN/dS values than the other domains [53]. Another mechanism driving genome sequence change is “repeat-induced point” (RIP) mutations. RIP occurs in a process
that mutates a duplicated DNA sequences from CpA to TpA and from TpG to TpA in meiosis [54, 55]. This mutation process has been found in most ascomycete fungi but not oomycetes to date, and only one basidiomycetes *Pucciniomycotina subphylum* [56]. The function of RIP mutations were generally regarded as deactivation of the duplicated genes by introducing premature stop codons or non-synonymous substitutions. However, this process sometimes affects the flanking sequences outside the duplicated region and thus introduces mutation to the “wild type” sequences [57, 58]. In *Fusarium spp.*, the higher mutation rate contributed by RIP is speculated to accelerate the evolution of the effector genes [57]. RIP has also being found in *L. maculans* and may contribute to the generation of isochore-like regions in the genome [17]. Conversely, powdery mildew fungi do not have RIP mutation machinery, and its expanded genome size may be a result of this omission [8].

**Horizontal gene transfer**

Horizontal gene transfer (HGT) is another important mechanism leading to large genome reorganization. HGT is the movement, without recombination, of stable genetic material between two individuals [59]. HGT may not only occur between different individuals of the same species, but also between species or even between bacteria and fungi or between fungi and oomycetes [60, 61]. In fungi, the movement of plasmids, mycoviruses, transposable elements, gene clusters, and whole chromosomes have been demonstrated from one individual to another [62]. The most well studied example of HGT in fungi is the movement of the *ToxA* gene from the wheat blotch pathogen
Stagonospora nodorum to Pyrenophora tritici-repentis, the causal agent of tan spot of wheat [63, 64]. This horizontal transfer event was identified by nucleotide sequence similarity and structural comparisons between genes from both species. The direction of transfer was inferred by the fact that the ToxA gene consisted of a single haplotype in P. tritici-repentis but 11 haplotypes in S. nodorum isolates.

Technical development of the next-generation sequencing technologies

The mainstream NGS techniques

Three sequencing methods developed by Roche, Illumina, and SOLid are generally regarded as the second generation sequencing techniques. Other techniques also exist in the market, such as Pacific Biosciences, but have not been widely adopted. The core of these three approaches is the amplification of input DNA as single molecule so the number copies can reach a scale large enough to be detected. There are millions of DNA molecules to be amplified simultaneously but each molecule has to be amplified within a closed system. (Table 1.2)

The Roche’s 454 sequencing platform amplifies DNA inside water droplets in an oil solution. Each droplet containing a single DNA template connected to a single primer-coated bead, which then forms a colony by so called “emulsion PCR”. The sequencing platform contains a large number of picolitre-volume wells with the size that only one single bead could fit in with the requisite
sequencing enzymes. Pyrosequencing uses luciferase to generate a light signal to detect individual nucleotides added to the nascent DNA strand, and the combined light signal records are then used to generate sequence read-outs. 454 sequencing usually generates long sequencing reads length, usually around 400-500 bp, easy for de novo assembly and detection of RNA alternate splicing. But it is difficult to read polymers (continues A/C/G/T) because the fluorescent intensity and length of polymers is not in linear relationship. (Information acquired from official website http://www.454.com)

Illumina’s DNA sequencing technology is called “sequence-by-synthesis”. DNA molecules are first connected to primers on a flow cell surface. Then single molecules are amplified to form clusters, each of which contains thousands of the same nucleotide molecules. Then the four types of Dideoxy nucleoside triphosphates (ddNTPs) are added and non-incorporated nucleotides are washed away. In the next step, DNA is extended by one nucleotide at a time and images of the fluorescently labeled nucleotides are captured by a camera. Then a dye with the 3’ terminal chemically removed is added so that the nucleotide sequence is ready for another cycle. The advantage of Illumina’s SBS sequencer includes its fully automation procedures and relatively low sequencing cost compared to other methods. Sequencing coverage can reach very deep level, so it is good for RNA-Seq and mixed sample sequencing. However, it’s based on a reversible reaction, so sequencing quality goes low when number cycle is high; in other words it only generates short reads which makes it difficult for de novo assembly. (Information acquired from official website http://www.illumina.com)
SOLiD sequencing technology uses “sequence-by-ligation”. The process is similar to that used by 454 sequencing, where the DNA is first amplified by emulsion PCR before nucleotide detection. Therefore, each of the beads containing copies of the same DNA molecule are then deposited on a glass slide. All oligonucleotides of a fixed length are pooled and labeled based on the first two nucleotides. Oligonucleotides are then annealed and ligated; the preferential ligation by DNA ligase for matching sequences results in a signal detection of the nucleotide at that position. In SOLiD, every position in genome is detected twice, so it possesses higher sequencing quality and accuracy, which makes it a good option for SNP detection. But like Illumina’s sequencer, reads are too short for de novo assembly. (Information acquired from official website http://www.appliedbiosystems.com)

Transcriptome detection techniques

While sequencing detects and reports genome information, it does not contain expression data, which at times can be more impacting to the phenotype of the orgasm. Many types of transcriptome detection techniques exist, including SAGE and MPSS, which are tag-based approaches, and microarray and RNA-Seq which are so called “high throughput” approaches. (Table 1.3)

To perform SAGE (Serial Analysis of Gene Expression) a researcher must first isolate short tags from all mRNA transcripts in a cell, then link multiple tags in tandem and clone the newly formed chimeric DNA strand into E. coli. Following sequencing, tags are mapped to genes and expression level is
determined by enumerating tags for each gene. One most important advantage of SAGE is it does not require a priori knowledge of genome structure as only part of genome sequences are required. Other advantages include its high efficient and ability to detect unpredicted ORFs, anti-sense transcripts and alternate splicing forms. However, not all genes have the required NlaIII site, and different genes may contain the same SAGE tag, which will mislead the downstream analysis [65].

The main Principle behind MPSS (Massively Parallel Signature Sequencing) is to isolate short tags from all mRNA transcripts in a cell, then ligate with an adapter and sequencing primers. Finally after PCR is performed within beads, they are loaded onto a flow cell for high throughput sequencing. By applying high throughput sequencing methods, the cost is reduced for the amount of data obtained. However, this technique has the similar limitations as SAGE [66].

In a tiling microarray application in which probes were designed along genome with a certain density, researchers firstly generate a chip containing oligonucleotide probes in each spot, then extract mRNA and label with fluorescence. Dyes cDNAs are then loaded to the premade chips. At this time, the probe binds specifically to target cDNA sequence. After washing away the unbound cDNA sequences, the amount of cDNA bound to each cell is detected by the fluorescence intensity [67]. There are many types of chips on the market, including spotted vs in situ synthesized arrays (such as provided by Agilent and Affymetrix), and two-channel vs. one-channel detections. The general advantages of microarray technologies are that they are relatively mature techniques, which means it is easier be designed to detect specific gene groups and easy for bioinformatics analysis. However,
a successful application of microarray need to know gene sequences to design probes. Also, microarrays designed for a whole transcriptome detection are expensive compared to RNA-seq.

The RNA-Seq technique is a newly developed technique that enables researchers to get insight of whole transcriptome in a single sequencing row [68-71]. To perform RNA-Seq, a cDNA library is made from extracted mRNA, then the cDNA is fragmented and ligated with adapters and primers to be sequenced by NGS sequencing, typically 454 or illumina, to generate millions of short reads representing the transcriptome. Following sequencing, there are generally two ways to process the short reads, either align them to the reference genome or perform a de-novo assembly, if no proper reference genome is available. From the mapping data or assembly contigs, researchers derive gene models and the expression level for individual genes in genome scales [72].

Although RNA-Seq is a recently developed technology and is now under active development, it provides some key advantages over other transcriptome detection methods. First, and maybe the most important advantage, is it can be applied to organisms without a mature genome sequence available. For example, a 454-based RNA-Seq approach has been used to sequence the transcriptome of the Glanville fritillary butterfly, and transcriptome contig libraries were de novo assembled to aid in gene model annotation [73]. Another advantage is the accuracy provided by RNA-Seq: the resolution can reach single base pair sensitivity, which makes it especially useful to detect exon boundaries. The third advantage is that it has a wide detection range compared to microarrays. Because mapping of 75bp reads could be unique if less than two mismatches are
allowed, so RNA-Seq has very low background noise. Additionally there’s no upper limitation to the expression level, since the number of mapped reads can be highly correlated to the expression level. As a result, RNA-Seq has a very large dynamic range of expression level detection, estimated to be greater than 9,000-fold in a recent yeast transcriptome study [70].

Despite the advantages of RNA-Seq, there are still challenges both in the wet lab and dry lab studies. Unlike small RNAs, which can be sequenced directly, mRNA usually requires a fragmentation step before sequencing and those manipulations complicate further analysis. For example, bias created during RNA fragmentation over the transcript body [70] and bias during cDNA fragmentation towards the identification of sequences from the 3’ ends [68]. Also, in the amplification stage, PCR artifacts may be generated and it’s very difficult to distinguish these from the highly expressed transcripts reads. There are also bioinformatics challenges, like other NGS technologies, that include retrieving and processing the large amount of data efficiently. Further, accurate tools specifically designed to map reads, remove bias, and calculate expression level need to be developed.

Cost of Next-generation sequencing study

The cost of sequencing the first genome of human reached roughly $3 billion, and human labor costs were as high as hundreds of researchers in several international institutes all spending 13 years to complete the project (National Human Genome Research Institute
The advance of NGS technology has now made the price drop dramatically during the last decade. James Watsons’ genome was completed for less than $1 million in the year 2007 [74]. By 2009, cost to conduct a human genome sequencing dropped to $100,000 (http://www.genomesunzipped.org/2011/06/guest-post-by-clive-brown-the-disruptive-power-of-cheap-dna-sequencing.php). Today, after nearly 10 year development, the cost to generate a genome assembly of fungi is as low as surprisingly $3,000. The goal to sequence a genome for only $1,000 [75, 76] seems very close.

One method to analyze the cost of NGS projects is to use the National Human Genome Research Institute (NHGRI) database, which tracks the cost of sequencing projects it funds (Figure 1.2). The data shows that the cost of sequencing one base pair dropped quickly after 2008, much faster than predicted by Moore’s law, which states that the number of transistors of an integrated circuit doubles roughly every 2 years. However, the NHGRI only calculates the direct cost of sequencing while ignoring other project components, which include continuously developing and maintaining sequencing and bioinformatics tools to analyse data, as well as project control to ensure the quality of analysis.

The cost of a NGS project can be divided roughly into 4 parts: 1) project design, sample collection, and preparation; 2) sample sequencing; 3) data storage and reduction; 4) downstream analysis. After sequencing costs decreased, the cost from other 3 parts (1,3,4) of a project, especially the sample preparation and downstream analysis, became the largest contributors to the total cost (Figure 1.3).
NGS projects have currently passed the stage of simply generating a draft genome assembly or even simple genome comparisons of two related species, but now encompass large scale genome comparisons including hundreds of individuals or transcriptome analyses involving multiple individuals and more than ten conditions with replications. In such projects, samples are carefully collected and prepared to ensure accuracy, which often requires collaboration of several labs and escalating costs. The downstream analysis is the most important step as this is where scientific inquiring happens. However, downstream analysis not only requires significant computational resources, but also expensive manpower to install and configure the tools to build pipelines and explain the output results. Confounding the matter, proper tools are often not available and thus new ones need to be developed, adding significant time and cost.

**Challenge in the bioinformatics analysis of NGS projects**

One important evaluation of scientific research is the reproducibility, which means others should be able to repeat a procedure and obtain the same results based on the methods and materials in the publication. However, for NGS projects, it’s often difficult or even impossible for others to repeat published works. In a survey included 19 recent NGS projects, only 6 provided the name and version of the tools they used in the analysis stages, while others either failed to provide the version of tools or failed to even provide name of tools they used [77]. In another survey of 50 publications using BWA (a popular aligner), only 7 of them provided detailed version and parameter setting used in the
project [77]. Obviously, others could not repeat the analysis without knowing the name of tools, but version and parameters used in the project are necessary. Since some of the widely used tools are under continual development, newly released versions are not only for fixing bugs, but also modification of algorithms occurs. Parameters are also playing fatal roles in analysis, and sometimes even a modest change will dramatically change the results. For example, the mismatches allowed in mapping reads will change the number of predicted SNPs. As such, publications need to include detailed parameters, and the actual default settings.

Another problem in NGS projects is that the customized analysis pipeline makes the results difficult to compare with each other. Since each tool is designed to function well for specific organisms, specific sequencing techniques, or answering specific questions, the result may be not reliable. One possible method to enhance the reproducibility and comparability is the adoption of the analysis pipeline applied and recommended by the 1000 Genome project [78]. If every research team carries the exact same pipeline from 100 Genome project and standardizes every step, then the results will be easily repeated and comparable. However, it’s impracticable for everyone to adopt the same pipeline, either because it is difficult to install and configure the particular tools, or in some stages additional information is required, which is actually not available for most of the non-model organisms.

A compromised solution is to choose the most reasonable tools, but provide detailed documentation for every step in the analysis pipeline. Luckily some tools are being developed to integrate the most
used functions and generate documentation with parameters automatically. These “frameworks”,
include: Bioextract [79], Galaxy [80], GenePattern [81], GeneProf [82], Mobyle [83] and others,
which actually bring together diverse tools under a unified interface.

A general pipeline for genome comparison analysis used for this dissertation

As mentioned above, there are no “one-fit-all” analysis pipelines for all NGS research, since the
sequencing technique, organism, and research questions vary dramatically between projects. Here is
a generalized pipeline generated and used in the following chapters of this dissertation (Figure 1.4).
Each stage in the pipeline is described, with the hurdles at that stage highlighted.

Stage 1: Sequencing reads manipulation

1) Quality checking and filtering: It’s important to check the average quality score of FastQ files to
insure they are within the normal range (>35). If a deep enough coverage of sequencing reads is
achieved, 20%-50% low quality score reads and especially those at each end of short reads are
trimmed, this process is especially important if these reads are going to map to a reference
genome assembly.

2) Adapter removal: Some reads contain sequencing adapter nucleotides. It is important to remove
these sequences if mapping is the next step.

Hurdle at this step: It’s difficult to perform the filtering work while still keep paired-end sequences in
the right order. An additional script in Velvet [84] could pick out paired reads from filtered files in their original paired order and input them into downstream assembly.

Stage 2: *De novo* assembly

1) Find the most appropriate assembler that performs best and is most compatible with the sequencing files. For the projects in this dissertation, all available assemblers were tested including Velvet, Edena [85], and SOAPdenovo [86].

2) Optimize parameters: The range of major parameters for assembly were decided based on the manual description and reads length. Then testing runs were performed to optimize the parameters.

Hurdle at this step: Formatting problems arise when merging results from Velvet and Edema using Minimus2 [87].

Stage 3: Gene prediction & annotation

1) Find the most appropriate gene predictor: the gene predictor should have a high prediction accuracy in comparison to other available tools and have a pre-trained matrix. For the projects in this dissertation, Augustus [88] for *Magnaporthe* species and FGENESH [89] for *Alternaria* species were used.

2) Convert gene prediction result to Fasta file format.

3) Annotate genes using Blast2GO [90], a batch blasting software.
4) Assign GO terms to the gene sequences using Blast2GO.

Hurdle at this step: FGENESH is online only and has a limited upload volume. Blast2GO becomes unstable when too many genes were analyzed at the same time. It is sometimes necessary to divide data into small pieces and combined them after processing.

Stage 4: Genome characteristic comparison

1) Calculate general statistics such as GC-content, GC3-content [49], average gene length, average exon numbers by using CLC sequence viewer or in-house written script.

2) Codon usage bias detection: calculate indicator numbers such as RSCU [91] and CAI [92].

Stage 5: Alignment & Structure variation identification

1) Find proper tool for alignment: not many alignment tools are capable enough to conduct contig verses contig alignment, Mummer suite [93] was chosen for the research in this dissertation.

2) Identification of Structure variation: There are two types of variation: 1) present/absent variation which is easily detected by alignment; 2) relocation variation: this can be identified by first aligning entire contigs and then dividing contigs into smaller pieces and aligning each to compare results with the original whole contig. Any difference may suggest a relocation event.

Hurdle at this step: It’s subjective to a set cut-off to determine alignment; and segmented contigs make it difficult to detect relocation events at a large scale.
Stage 6: SNP identification

1) Mapping short reads to reference: this is now pretty mature process; BWA [94], bowtie [95], and SOAPaligner [96] all perform well and rapidly complete mapping since they use basically the same algorithms.

2) Calling SNP: It’s better to use a SNP caller that takes quality scores of each base as extra information such as SOAPsnp.

Hurdle at this step: It’s pretty easy and fast to generate SNP files, however, there are usually more than 10,000 SNPs at the genome scale for fungi, so the hurdle is how to analyze them? For the work in this dissertation, each SNP was labeled as synonymous/non-synonymous.

Stage 7: Phylogenetic analysis

1) Choose sequences: It’s important to choose sequences from different species that have homology and can be aligned. It is ideal to select several sequences from each species to avoid bias. For this dissertation, 6 genes were selected in the *Alternaria* chapter and more than 10,000 SNP markers were selected in the *Magnaporthe* comparative chapter.

2) Choose methods: There are multiple algorithms to generate a phylogenetic tree, choose one that best fit your purpose. For this dissertation, maximum likelihood approach was selected for all phylogenetic analysis.

Stage 8: Selection force calculation
1) Extract sequences showing homology from different isolates identified by BLAST.

2) Perform a codon alignment: which translate ORFs to protein first and do alignment, then go back to align nucleotide sequences based on protein alignment. Prank [97] is the alignment tool used in this dissertation.

3) For projects in this dissertation, input alignment files into Codeml tools in the PAML suite [98] and choose most appropriate model (M7 or M8) to calculate dN/dS.
References


40. Proctor RH MS, Alexander NJ, Desjardins AE.: Evidence that a secondary metabolic


63. Timothy L Friesen EHS, Zhaohui Liu, Steven Meinhardt, Hua Ling, Justin D Faris1, Jack B Rasmussen, Peter S Solomon, Bruce A McDonald & Richard P Oliver: Emergence of a new disease as a result of interspecific virulence gene transfer. *Nature Genetics* 2006, **38**:953-956.


<table>
<thead>
<tr>
<th>Organism/Name</th>
<th>SubGroup</th>
<th>Natural Host</th>
<th>Genome Size (Mb)</th>
<th>GC%</th>
<th>Predicted Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Phytophthora ramorum</td>
<td>Oomycete</td>
<td>Oak</td>
<td>54.427</td>
<td>53.9</td>
<td>-</td>
</tr>
<tr>
<td>Phytophthora sojae</td>
<td>Oomycete</td>
<td>Soybean</td>
<td>79.327</td>
<td>54.6</td>
<td>28142</td>
</tr>
<tr>
<td>Phytophthora infestans</td>
<td>Oomycete</td>
<td>Potato,Tomato</td>
<td>190.142</td>
<td>51</td>
<td>19344</td>
</tr>
<tr>
<td>Ustilago maydis S21</td>
<td>Basidiomycetes</td>
<td>Maize</td>
<td>19.8578</td>
<td>53.7</td>
<td>6671</td>
</tr>
<tr>
<td>Sporisorium reilianum SRZ2</td>
<td>Basidiomycetes</td>
<td>Maize</td>
<td>18.4769</td>
<td>59.5</td>
<td>6806</td>
</tr>
<tr>
<td>Puccinia striiformis f. sp. tritici</td>
<td>Basidiomycetes</td>
<td>Wheat</td>
<td>64.799</td>
<td>44.5</td>
<td>-</td>
</tr>
<tr>
<td>PST-130</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Melampsora larici-populina 98AG31</td>
<td>Basidiomycetes</td>
<td>Poplar</td>
<td>97.683</td>
<td>41</td>
<td>16380</td>
</tr>
<tr>
<td>Pyrenophora teres f. teres 0-1</td>
<td>Ascomycetes</td>
<td>Barley</td>
<td>33.581</td>
<td>50.9</td>
<td>11799</td>
</tr>
<tr>
<td>Leptosphaeria maculans JN3</td>
<td>Ascomycetes</td>
<td>Crucifers</td>
<td>45.1246</td>
<td>-</td>
<td>12469</td>
</tr>
<tr>
<td>Sclerotinia sclerotiorum 1980 UF-70</td>
<td>Ascomycetes</td>
<td>Multiple</td>
<td>38.204</td>
<td>41.8</td>
<td>14637</td>
</tr>
<tr>
<td>Erysiphe pisi</td>
<td>Ascomycetes</td>
<td>Pea</td>
<td>69.276</td>
<td>39.2</td>
<td>-</td>
</tr>
<tr>
<td>Blumeria graminis f. sp. hordei DH14</td>
<td>Ascomycetes</td>
<td>Grasses</td>
<td>98.284</td>
<td>44</td>
<td>-</td>
</tr>
<tr>
<td>Magnaporthe oryzae 70-15</td>
<td>Ascomycetes</td>
<td>Rice</td>
<td>40.9791</td>
<td>-</td>
<td>13032</td>
</tr>
<tr>
<td>Verticillium albo-atrum VaMs.102</td>
<td>Ascomycetes</td>
<td>Multiple</td>
<td>30.331</td>
<td>56</td>
<td>10488</td>
</tr>
<tr>
<td>Verticillium dahliae VdLs.17</td>
<td>Ascomycetes</td>
<td>Multiple</td>
<td>32.975</td>
<td>55.8</td>
<td>10811</td>
</tr>
<tr>
<td>Fusarium oxysporum f. sp. lycopersici</td>
<td>Ascomycetes</td>
<td>Multiple</td>
<td>61.4707</td>
<td>48.2</td>
<td>-</td>
</tr>
<tr>
<td>4287</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Nectria haematococca mpVI 77-13-4</td>
<td>Ascomycetes</td>
<td>Multiple</td>
<td>51.23</td>
<td>50.8</td>
<td>15708</td>
</tr>
<tr>
<td>Alternaria arborescens EGS 39-128</td>
<td>Ascomycetes</td>
<td>Tomato</td>
<td>33.89</td>
<td>50.9</td>
<td>-</td>
</tr>
</tbody>
</table>

Table 1.1. Genomes sequenced of plant pathogen fungi and oomycetes.
### Table 1.2. Comparison of the three mainstream 2nd generation sequencing technologies and PacBio which is regarded as the 3rd generation sequencing technique.

<table>
<thead>
<tr>
<th>NGS technique</th>
<th>Company</th>
<th>Sequencing strategy</th>
<th>Amplification</th>
<th>Read length</th>
<th>Reads per run</th>
<th>Claim to fame</th>
</tr>
</thead>
<tbody>
<tr>
<td>454</td>
<td>Roche</td>
<td>Synthesis</td>
<td>emPCR</td>
<td>400 bp</td>
<td>1 million</td>
<td>First Next-Gen Sequencer long reads</td>
</tr>
<tr>
<td>Hi-Seq</td>
<td>Illumina</td>
<td>Synthesis</td>
<td>BridgePCR</td>
<td>100 bp</td>
<td>Up to 640 million</td>
<td>First short-read sequencer current leader in advantages</td>
</tr>
<tr>
<td>SOLiD</td>
<td>Life Technologies</td>
<td>Ligation</td>
<td>emPCR</td>
<td>50 bp</td>
<td>Up to 1.4 billion</td>
<td>Second short-read sequencer low error rates</td>
</tr>
<tr>
<td>PacBio</td>
<td>Pacific Biosciences</td>
<td>Synthesis</td>
<td>None</td>
<td>2 Kb</td>
<td>45,000</td>
<td>First real-time single-molecule sequencing</td>
</tr>
<tr>
<td>Technology</td>
<td>RNA-Seq</td>
<td>Tilling Microarray</td>
<td>cDNA or EST sequencing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>----------------------------------</td>
<td>---------</td>
<td>--------------------</td>
<td>------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Principle</td>
<td>NGS</td>
<td>Hybridization</td>
<td>Sanger sequencing</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Resolution</td>
<td>Single base</td>
<td>~ 10-100 bp</td>
<td>Single base</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Throughput</td>
<td>High</td>
<td>High</td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Genomic sequences</td>
<td>Not necessary</td>
<td>Necessary</td>
<td>Not necessary</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Background noise</td>
<td>Low</td>
<td>High</td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Detection range</td>
<td>High</td>
<td>Low</td>
<td>Low</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Isoforms detection</td>
<td>yes</td>
<td>limited</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Allelic detection</td>
<td>yes</td>
<td>limited</td>
<td>Yes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Amount of input RNA</td>
<td>Low</td>
<td>High</td>
<td>High</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 1.3. Comparison of the technique specification and application of RNA-Seq, tilling microarray, and low-throughput techniques.
Figure 1.1. Number of genomes available for the fungi and oomycete species. These species all have multiple isolates being sequenced.
Figure 1.2. Sequencing cost released by The National Human Genome Research Institute (NHGRI). (http://www.genome.gov/sequencingcosts/) Blue line with dots in it represents the cost per raw MB of DNA sequencing performed in the institute. The while line represents the expectation made by Moore’s Law.
Figure 1.3. Contribution of different stages to the total cost of a genome sequencing project over time. Note the sharp shrink of the component of sequencing stage. The three time points have interval of approximately ten years. (Sboner A, et al.: The real cost of sequencing: higher than you think! Genome Biology 2011, 12:125)
Figure 1.4. A general pipeline for genome comparison analysis used for this dissertation. Blue square frames represent analysis steps, Oval frames represent outside data source, purple frames represent tools used for analysis in this dissertation.
Chapter 2: Genomic characterization of the conditionally dispensable chromosome in *Alternaria arborescens* provides evidence for horizontal gene transfer
Introduction

The rapid development of next-generation sequencing technologies over the past decade has led to a flood of both de-novo sequencing and re-sequencing projects in almost every branch of the tree of life. Within the fungal kingdom, comparative genome studies have led to the unexpected finding that large genomic regions may be variable among isolates of a given species. One category of these variable regions is unique chromosomes referred to as supernumerary or conditionally dispensable because they are not typically required for saprophytic growth [1-3]. These chromosomes have been identified in many fungi including Magnaporthe oryzae [4-6], Fusarium oxysporum [7], Nectria haematococca [8, 9], Mycosphaerella graminicola [10], Cochliobolus heterostrophus [11], Leptosphaeria maculans [12], and Alternaria alternata [13, 14].

Plant pathogenic fungi in the genus Alternaria infect a remarkable range of host plants and are major causes of agricultural yield losses [15]. Conditionally dispensable chromosomes (CDCs) are carried by several of the small-spored, plant-pathogenic Alternaria species [13, 14, 16]. These chromosomes are generally less than 2.0 MB in size, and may be transmitted horizontally between isolates in a population, potentially conferring new pathogenic attributes to the receiving isolate [17-20]. Loss of the CDC can also occur during repeated sub-culturing, resulting in the transition from a pathogenic to saprophytic form of the fungus [13]. Several genes coding host specific toxins (HSTs) have been located to gene clusters on CDCs, including those producing AF-toxin from the strawberry pathotype [21], AK-toxin from the Japanese pear pathotype [22], and ACT-toxin from the tangerine pathotype.
These toxins share a common 9,10-epoxy-8-hydroxy-9-methyl-decatrienoic acid structural moiety, with the genes encoding each toxin sharing a high degree of homology [21-25]. In addition, the AMT gene from the apple pathotype, a gene involved in host-specific AM-toxin cyclic peptide biosynthesis, is located on a small chromosome of 1.1 to 1.7 Mb [13, 26], with at least four copies involved in AM-toxin biosynthesis [27]. The only other gene sequences identified to date on CDCs are extended families of tranposon-like sequences (TLSs) [14].

Horizontal gene transfer (HGT) is the movement, without recombination, of stable genetic material between two individuals [28]. HGT may not only occur between different individuals of the same species, but also between species or even between bacteria and fungi or between fungi and oomycetes [29, 30]. In fungi, the movement of plasmids, mycoviruses, transposable elements, gene clusters, and whole chromosomes have been demonstrated from one individual to another [31]. The first theory to explain gain and loss of HSTs was proposed in 1983 [32]. It has then been hypothesized that the genome content of CDCs in Alternaria species were acquired through HGT events [14]. The most well studied example of HGT in fungi is the movement of the ToxA gene from the wheat blotch pathogen Stagonospora nodorum to Pyrenophora tritici-repentis, the causal agent of tan spot of wheat [33, 34]. This horizontal transfer event was identified by nucleotide sequence similarity and structural comparisons between genes from both species. The direction of transfer was inferred by the fact that the ToxA gene consisted of a single haplotype in P. tritici-repentis but 11 haplotypes in S. nodorum isolates.
*Alternaria arborescens* (synonym *A. alternata* f. sp. *lycopersici*), the fungus that produces host-specific AAL toxin, is the causal agent of stem canker of tomato [35, 36]. It has been observed in pulsed field gel electrophoresis (PFGE) studies that *A. arborescens* carries one CDC of 1.0-Mb [16, 37]. To date, only two genes have been reported to be carried on this CDC including *ALT1*, which is a PKS gene involved in AAL toxin biosynthesis [38, 39], and *AaMSAS*, also a PKS gene [40, 41]. A CDC deletion mutant of *A. arborescens* generated through restriction enzyme mediated integration (REMI) showed a toxin- and pathogenicity-minus phenotype [41]. In addition, in protoplast fusion experiments, a CDC from *A. arborescens* was observed to transfer into the strawberry pathotype, and subsequently introduced new tomato pathogenicity to the fusant [41].

In this study, we used a next generation sequencing approach to produce a (90X) draft sequence of the *A. arborescens* genome and used a novel bioinformatic approach to separate CDC contigs from the essential chromosome (EC) contigs. The gene content of the CDC was analyzed to answer the following questions: (1) What is the difference between the CDC and EC genome content at the nucleotide level? (2) Are CDC genes under positive selection and could they represent additional virulence factors in addition to the known toxin encoding genes? (3) Is the evolutionary history of the CDC the same as that of the ECs, and is there any evidence of a HGT event? In answering these questions, we confirmed a different genome content pattern of the *A. arborescens* CDC and found evidence for HGT.
Results

Sequencing & assembly

* Alternaria arborescens * strain EGS 39-128 (CBS 102605) [42] was sequenced by a whole genome shotgun approach using the Illumina Genome Analyzer II, which resulted in ~50 million paired-end short reads of 75 bp representing 90X average coverage of the predicted genome content. *De-novo* assembly was performed using Velvet [43] (version 0.7), and confirmed by Edena [44] and Minimus2 [45]. The assembly resulted in 1,332 contigs with a N50 of 624KB and total size of 34.0MB (Table 2.1; Assembly has been deposited at DDBJ/EMBL/GenBank under the accession AIIC00000000. The version described in this paper is the first version, AIIC01000000.) One hundred thirty-seven large contigs with lengths greater than 10 KB and representing 98% of the genome assembly content were chosen for further analysis.

Marker-assisted identification of contigs carrying toxin biosynthetic genes

The first challenge in analyzing the CDC was to isolate its assembly contigs away from EC contigs. For this genome, the process was made more challenging as there is no defined reference genome, few genetic markers, and no optical mapping. It is known from previous studies of Alternaria CDC that most *Alternaria* species, including the isolate used in this study, have a single CDC [16, 37]. To begin assembly of this chromosome, two previously identified CDC genes that belong to the toxin
biosynthetic cluster, \textit{ALT1} and \textit{AaMSAS} [38, 40], were used as markers to search all contigs. Through this strategy, two putative CDC contigs of 15 KB and 48 KB in length were identified as containing \textit{ALT1} and \textit{AaMSAS}, respectively. These two contigs were annotated to identify PKS genes, other toxin biosynthetic genes, as well as genes with orthologs in other fungi and bacteria (Figure 2.1). Multiple putative HST genes were identified on both contigs, consistent with predictions based on previous reports [14].

**Identification of the remaining CDC contigs and validation by Southern hybridization**

To identify additional CDC contigs, the \textit{Alternaria brassicicola} (\textit{Ab}) genome sequence was used as a reference (downloaded from The Genome Institute at Washington University). \textit{A. brassicicola} is a related species to \textit{A. arborescens} (\textit{Aa}) but does not carry CDCs. All contigs from \textit{Aa} were aligned to \textit{Ab} contigs using MUMmer [46] as the alignment tool with an identity cut-off of 90%. Eight previously identified marker genes from \textit{Aa}, 6 from the EC and 2 from CDC, were used to set criteria to distinguish contigs belonging to CDCs versus ECs [41] (Table 2.2). After comparing the alignments of contigs containing the 8 markers genes, we set a CDC contig cutoff as those contigs with less than 20% coverage of sequence aligned to \textit{Ab} with higher than 90% identity (including both coding and non-coding regions). Through this method, 29 predicted CDC contigs were identified with total length of 1.0 MB, same as the expected size from clamped homogenous electric fields (CHEF) gel analysis [16, 37] (Figure 2.2). The remaining 108 contigs were considered essential chromosomes.
(ECs) with total genome size of 32.3 MB (Figure 2.3). Validation that the selected contigs belong to the CDC was performed by Southern hybridization using genes predicted to reside on the CDC and EC contigs respectively as probes (Figure 2.2C-F). Five probes were hybridized including 4 from the CDC and 1 from the EC. Each probe gene is predicted to be present in a single copy. The first probe was \textit{ALT1}, a toxin gene known to reside on the CDC [41]. This was followed by hybridizing with three CDC genes predicted from the \textit{de-novo} assembly, including a transporter CDC\_92, a polysaccharide export protein CDC\_102, and an o-methyltransferase CDC\_147. The fifth probe was EC\_97\_90\_g721, a gene annotated as PKS and predicted to reside on the ECs.

**Gene prediction, length, GC3-content, and repeat identification**

Nine thousand, one hundred sixty-seven genes were predicted by FGENESH [47] using pre-trained \textit{Alternaria} parameters, of which 209 genes were assigned to CDC contigs and 8958 to EC contigs. The average length of each predicted gene was 1.8 KB, and the gene density was 3.7 KB per gene. Compared to gene predictions for \textit{A. brassicicola} (average gene length=1.3KB, gene density=3.0KB per gene), \textit{A. arborescens} genes were longer and present in lower density. To evaluate the origin of the CDC, the predicted genes residing on the CDC and EC contigs were compared at the nucleotide level, including gene length, GC3-content [48], repeat load, and codon usage bias. This analysis showed that CDC genes are about 200 bp shorter on average than EC genes ($P=2.36\times10^{-9}$) and have significantly lower GC3-content ($P=0.028$). Repeat regions composed 5.3\% of CDC contigs while only
0.6% of EC contigs (Table 2.3). It should be noted that some repeat regions could be lost in short read sequences *de-novo* assemblies, however even with possible suppressed numbers, this result indicates approximately 10X repeat enrichment in the CDC compared to the EC.

**Codon usage analysis**

Codon usage comparisons of CDC and EC genes were used to determine whether a bias in codon usage exists between the two groups. Both the Codon Adaptation Index [49] (CAI, \( P=0 \), Figure 2.4) and Relative Synonymous Codon Usage (RSCU) [50] correlation \( (P=1.14E-14, \text{Table 2.4}) \) from the two groups were significantly different, suggesting a different origin. The largest codon usage bias was observed for the amino acids Tyrosine, Lysine, and Asparagine, with a preference for A over G and T over C in CDC genes (Table 2.5). These three amino acids were not biased for CDCs in *Fusarium* [7], indicating there’s no universal codon usage bias pattern between CDCs from *Alternaria* and *Fusarium* and ECs.

**Annotation of EC genes**

The assembly results show the size of essential chromosomes region collectively to be 33.0 MB with 8958 predicted genes. RepeatMasker identified only 0.12% of the EC region as simple repeats (about 50bp in length) and 0.08% as low complexity, indicating that short repeats may be lost during *de*
*novo* assembly of Illumina sequencing reads. For secreted protein identification, 1099 (12.2%) of the EC proteins were predicted to contain signal peptides, and were functionally annotated using BLAST to the NCBI database with more than 98% of the genes returning at least one hit with an E-value < 1.0E-3. From the BLAST results, we identified 212 transcription factors, 98 oxidase proteins, 202 kinase proteins, 279 transporters, 81 Cytochrome P450s, and 45 different proteases.

**Annotation of CDC genes**

Several host-specific toxin genes and transposon-like sequences have been reported to be carried by CDCs in *Alternaria* [14]. We used two methods to annotate the functions of resident CDC genes: (1) they were blasted against the NCBI non-redundant database as well as Pfam [51] and NCBI CDD [52] to search for functional domains; (2) they were scanned to identify transcription factors, PKS genes, NRPS genes, P450s, transporters, and pathogenicity related genes.

From 160 NCBI BLASTN hits of putative CDC genes (Figure 2.5), the top five species matches were *Pyrenophora teres*, *Pyrenophora tritici-repentis*, *Phaeosphaeria nodorum*, *Leptosphaeria maculans*, and *Nectria haematococca*, all of which are fungal phytopathogens. Interestingly, *A. alternata* was ranked 7th in this list, demonstrating that CDC genes were more similar to genes present in other fungal species rather than other *Alternaria* spp. Moreover, besides *N. haematococca*, all these other fungi are closely related taxonomically belonging to the class, Dothideomycetes.
Gene ontology terms were assigned to CDC genes based on BLAST matches with sequences whose function was previously characterized [53]. Ninety CDC genes were assigned to a biological process, 51 for molecular function, and 15 for cellular component (Figure 2.6). Among the biological process assignments, 54% of genes were assigned to “metabolic process”, and 10% to “biosynthetic process”. Enrichment of metabolic and biosynthetic process in CDC genes as compared to EC genes supported the observation that *Alternaria* CDC genes are enriched for polyketide synthases (PKS) and toxin synthases. Molecular function terms showed a significant percentage (39%) to “nucleotide/nucleic acid binding”, which shows an enrichment of transcription factors and gene regulation elements.

To provide a more detailed characterization of putative CDC genes, each was translated to identify protein families. Among the 209 predicted CDC proteins, 31 were identified as carrying PKS domains. Two proteins were found to carry highly modular domains: KS-AT-KR-ACP on CDC_141 and KS-AT-DH-ER-KR-ACP on CDC_165. The remaining 29 PKS proteins each carried 1 or 2 ACPs (Acyl carrier protein) domains. Seven proteins were found to carry NRPS domains: 3 Enterobactin domains, 2 Bacitrancin domains, 1 Pyochelin domain, and 1 CDA1 domain. Two proteins were identified as hybrid PKS-NRPS. Seven proteins were identified as P450 monooxygenase proteins. For transcription factors, 24 proteins were characterized to contain TF domains, in which Zn2Cys6 was the prominent group. Multiple ADP/ATP transporters, ABC transporters, ion transporters and major facilitator superfamily (MFS) transporters were also found in CDC protein group. Additionally, it was found that multiple proteins carrying FAD binding domains and oxidoreductases. Finally, 37 proteins were identified as putative pathogenicity related genes through scanning CDC genes in the pathogen-host
Secondary metabolite biosynthetic gene clusters

In fungi, it has been reported that genes responsible for secondary metabolite biosynthesis (SMB) may be clustered [4, 55]. Typically these include PKS or NRPS genes, as well as genes responsible for structural modifications of initial metabolites, for transport, and for transcription regulation [56]. In this study, we screened each CDC gene and those surrounding them, looking for evidence of clustering of PKS, NRPS, transcription factors, transporters, P450 proteins, FAD binding proteins, transferases, and oxidoreductases. We identified 10 putative SMB clusters (Figure 2.7). A typical SMB cluster is formed by 3-6 genes, with 1 or 2 PKS or NRPS genes, and other metabolite syntheses related genes.

Evolutionary selection of CDC genes and domains

To estimate selection on CDC and EC genes, Ka/Ks ratios were calculated, with the assumption that genes with Ka > Ks were likely under positive selection, genes with Ka = Ks were likely evolving neutrally, and genes with Ka < Ks were likely under purifying (negative) selection. Twenty-eight CDC and 6,036 EC genes were successfully aligned to A. brassicicola genes and Ka and Ks values were
calculated for each. It was observed CDC genes had about a double $\text{Ka} (0.08/0.043)$ and larger $\text{Ka/Ks}$ ratios ($0.133/0.084$) than EC genes (Figure 2.8), possibly indicating greater positive selection on CDC genes. The two CDC genes with highest $\text{Ka/Ks}$ ratio was CDC_102 (PKSs) and CDC_146 (phosphotransferase). However, no CDC genes showed $\text{Ka/Ks}>1$, suggesting that in these two 
*Alternaria* species strong positive selection may only occur in specific regions of a protein. The selection ratio only at conserved domains of CDC genes was then estimated. Domains of aligned CDC genes were identified using the NCBI CDD database. Each individual domain was extracted then the $\text{Ka/Ks}$ ratio was calculated and compared to that from same full length protein (Table 2.6). We found two domains from CDC_151 with a higher $\text{Ka/Ks}$ ratio compared to whole length protein: a 12x increase for the haloacid dehalogenase-like hydrolases (HAD_like) domain, which uses a nucleophilic aspartate in their phosphoryl transfer reaction, and 2x increase for heavy-metal-associated (HMA) domain, which transports or detoxifies heavy metals. Another interesting example was CDC_144, whose domain patatin-like phospholipase (Pat17, belonging to the alpha-beta hydrolase family) showed a 1.5x increase compared to whole length protein.

**Origin of CDC**

In the taxonomy report of CDC BLAST results, top hits came primarily from closely related dothideomycete fungi, indicating CDC may have fungal origin, or a transfer event occurred between CDC content and one or more fungal genomes. To test whether CDCs have the same phylogenetic
placement with ECs, a phylogenetic analysis was conducted, including EC and CDC genes from *A. arborescens*, genes from *A. brassicicola*, and from three other ascomycete species: *P. tritici-repentis*, *L. maculans*, and *A. oryzae*. Proteins coded by 6 genes showing homology in all 6 groups were used to build a distance tree (Figure 2.9) using the neighbor-joining method [57]. Results show the CDC clade was within but basal to the two *Alternaria* clades.

A BLAST score ratio (BSR) [58] analysis was performed to test whether individual proteins on the *A. arborescens* CDC had more similarity to *A. brassicicola* or other fungi, and the result was compared with the same analysis to EC proteins. Complete genome protein sequences of three fungal species: *P. tritici-repentis*, *L. maculans*, and *A. oryzae* were extracted and built into a library called “3-fungi”, representing proteins from closely related fungal species. Then proteins from the CDC and ECs were compared to the *A. brassicicola* protein library and “3-fungi” protein library respectively (Figure 2.10). It was clear, that there was less divergence between EC proteins and *A. brassicicola* proteins compared to that with other fungi (35.5% vs 15.1%), consistent with the species phylogeny. In contrast, CDC proteins are more similar to proteins from other fungi (16.7% vs 10.0%), suggesting they have different evolutionary history other than EC proteins.
Discussion

Comparison of CDCs in other sequenced fungal genomes

Sequenced genomes of phytopathogenic fungi have been increasing quickly due to the availability of the next-generation sequencing technologies and lowered costs. Genome size of phytopathogenic fungi ranges from ~20MB (Sporisorium reilianum) [59] to 150MB (Erysiphe pisi) [60], but the average are around 40MB. Biotrophs usually have larger genome size than necrotrophs or hemibiotrophs [61]. Within the Dothideomycete class, there are 13 genome projects with public data available including A. brassicicola from the JGI fungal genomics program (Table 2.7). Most species have genome sizes ranging from 30MB to 45MB except Mycosphaerella fijiensis which has a 74 MB genome (produced at the Joint Genome Institute of the United States Department of Energy in 2007). A. arborescens has a 34MB genome, smaller than most other species but larger than A. brassicicola, and also has the smallest gene numbers of model and relatively short gene lengths.

Compared to other recently published assemblies of CDCs in filamentous fungi, A. arborescens has a relatively small number of CDCs (one) and CDC size (1.0Mb). M. graminicola, is reported to have the highest number of dispensable chromosomes with upwards of 8 ranging in size from 0.39 to 0.77MB [62]. Three CDCs in N. haematococca [9, 63], and 4 complete CDCs and partial sequences of another 2 in F. oxysporum [7] were identified. In other Alternaria species, identified CDCs are relatively larger such as 1.05Mb in the strawberry pathotype [14], 1.1 to 1.7 Mb (depending on strains) in the apple pathotype [13], and 4.1 Mb in the Japanese pear pathotype [64, 65]. In A. arborescens, only 1
dispensable chromosome is present, representing only 3% of the genome content, which is significantly smaller than other cases and may suggest a more recent acquisition or different origin.

**PKS and NRPS clusters**

Phytopathogenic fungi produce a diverse array of secondary metabolites, including host-selective toxins conferring pathogenicity [66]. It was reported in two basidiomycete maize pathogens, the candidate effector genes are located in small clusters that are dispersed throughout the genome [67]. However, in other fungi, especially ascomycetes, genes coding for toxins can co-locate in clusters consisting of more than 10 contiguous genes. A well-known example is the trichothecene biosynthetic gene cluster in *F. graminearum*, which contains 10-12 genes including a terpene synthase gene, P450 monooxygenase genes, acyl transferase genes, regulatory genes, and transporter genes [55]. While in *A. fumigatus* 26 SMB clusters were identified, each contains 5-48 genes [68]. In our study, 29 PKS, 5 NRPS, and 2 hybrid PKS-NRPS genes were found on the CDC, with larger density compared to other fungi. However, among 10 predicted SMB clusters, most were relatively small and only carried 3-8 genes, which may not represent the true cluster size due to short contigs length that may divide one large cluster into two or more. One example of an identified cluster was located on contig Node_309, which consists of 5 genes, including 2 PKSs, 1 NRPS putatively coding for enterobactin, a phosphate transferase gene, and a MFS transporter. It lacks regulators, P450s, and transporters compared to other typical clusters. However, this cluster locates
at the edge of the contig. Only 5 genes away from this cluster, another small cluster was identified containing a PKS, NRPS, P450s and an ABC transporter, suggesting these two could be part of a larger cluster. In this study, PKS genes were identified by screening the PKS sequence database, especially the domain database, which include: ketoacyl synthase (KS), acyl transferase (AT), ketoreductase (KR), dehydratase (DH), enoyl reductase (ER), and acyl carrier protein (ACP, also known as PP domain). The KS, AT, and ACP domains are essential for PKS function [69]. Two PKS genes were identified to have multiple domains above: KS-AT-KR-ACP in CDC_141, and KS-AT-DH-ER-KR-ACP in CDC_165. The remaining 29 PKS genes each carries 1 or 2 ACP domains. Despite these conserved domains, other domains carried by these genes were divergent, indicating variance and multifunction of each PKS genes (Table 2.8). However, at least 3 domain families were found to be enriched in the identified PKS genes: ABC_membrane (4 identified), NADB_Rossmann (7 identified), and P-loop NTPase (6 identified), suggesting these proteins are transmembrane and may catalyze enzymatic reactions. In the NRPS and hybrid PKS-NRPS gene group, enterobactin, bacitracin, pyoverdine, syringomycin, and CDA1 domains were identified, 4 of which were reported from bacteria [70-73]. We eliminated the possibility of these genes originating from bacterial sequencing contamination by BLAST comparing all assembly contig sequences against the NCBI All Bacterial database with 2017 genome sequences, and found that the species with most hits was *Streptomyces coelicolor* with > 80% identity. However, only 0.7% of the entire *S. coelicolor* genome was covered. Indicating that either these genes have an origin from bacteria or their product proteins interact with each other and require a highly conserved structure that was retained during evolution.
Comparison of CDC genes from different isolates

After comparing CDC genes with EC genes and other species, the divergence of CDCs in different A. alternata isolates was examined. Sequences of 8 genes, 2 from CDC and 6 from EC [39] were downloaded from NCBI GeneBank. For each gene, homologous were extracted for each from isolates collected worldwide, including both pathogens and non-pathogens (for EC gene only). We also used homologs of marker genes in A. arborescens and a phylogenetic analysis was conducted to calculate the mean distance among different isolates (Figure 2.11). While the 6 EC genes showed some diversity, the 2 CDC genes were almost identical among all isolates, suggesting they may be transmitted among different isolates and had a different evolutionary history than EC genes.

Horizontal gene transfer

According to the horizontal gene transfer hypothesis, A. arborescens may have acquired its CDC from another Alternaria species, from a fungus other than Alternaria, or possibly from a bacterium or virus [74]. There are at least two possible explanations for its origin: (1) CDCs were present in an Alternaria ancestor, but were independently lost during vertical transmission in other non-pathogenic Alternaria species. (2) CDCs arose from the essential chromosome as a copy first but then went under divergence so no obvious orthology could be detected. To test which of the three models fits this case best, we built a complete EC protein library and blasted all CDC proteins against
it to detect any possible orthology. Out of 209 CDC proteins, we found 12 (5.7%) showing orthology to EC proteins. Although the low orthology percentage alone could not exclude the “duplication and divergence” model, taken together with differences on GC3-content and codon usage bias, the possibility that this model fits is minimal.

To distinguish between HGT and vertical transmission hypothesis, we identified differences between *A. arborescens* CDC and EC genes in length, GC3-content, and codon usage bias. There was limited orthology detected between two groups; CDC genes showed discordant phylogenetic relation with EC, and had higher similarity to other fungi than *A. brassicicola*. From previous phylogenetic analysis of 13 *A. alternata* isolates collected worldwide, CDC genes from different isolates were almost identical despite diverse EC background [41]. Taken together these results, we concluded that the HGT model may serve as the best fit model in this case. Additionally, these data support the 1983 theory proposed by Nishimura that Alternaria species acquired HSTs by HGT [32].

In this study, we identified evidence for the possibility of HGT event occurred in *A. arborescens*. For *Alternaria*, this strategy has its advantages. First, as a pathogen with a wide host range, as observed in nature, transportable pathogenicity chromosome may increase pathogen's adaptation to environment. Second, loss of a CDC when there's no host may reduce the cost of carrying extra genome content. Third, as asexual fungi, horizontal transfer may compensate the lack of genetic recombination.

We were also trying to identify the donor in *A. arborescens's* HGT events, as HGTs have been
reported previously from bacteria to fungi, and from fungi to oomycetes [29, 30]. Phytopathogenic fungi were also reported to carry exogenous genes from plants [75]. Here, we tried to find evidence of CDC genes originating from either bacteria or its host tomato by blasting CDC genes against bacteria (available genomes on NCBI) and tomato genome sequences. Two CDC genes were identified showing homology to the tomato genome sequences but no homologs to *A. arborescens* EC sequences: CDC_7 to gi47104509 (E_value = 2e-26), and CDC_36 to gi47104495 (E_value=1.1e-20). These two genes are annotated as enolase and NRPS. We were not able to identify any bacterial origin CDC genes. However, when we blasted complete CDC contig sequences to the NCBI bacteria database, we found that *Streptomyces* as a genus had the most number of hits. The genome sequence of *Streptomyces coelicolor A3* (8.6MB) [76] was downloaded from genebank then used as reference for CDC contigs blast. Setting 80% identity as cutoff, we found 7.0% of CDC contigs had homology with the *Streptomyces coelicolor* genome. Although these homologous regions were usually fragmented into small hits, bias was observed as in Node_136, the contig containing ALT1 toxin gene, 14.0% aligned with a *Streptomyces* genome region. This suggests a possible HGT event between *Alternaria* and *Streptomyces*. Fragmentation structure of *Streptomyces* homologs indicates either the horizontal transfer in this case did not include large continuous sequence or it is an ancient event.
Conclusions

In this study, we identified \textit{A. arborescens} CDC sequences through a whole genome sequencing and \textit{de-novo} assembly process. By comparing nucleotide usage between CDC and EC contigs, we found evidence supporting HGT in \textit{A. arborescens}. We also identified some predicted CDC genes under positive selection that may serve as virulence factors. However, questions still remain, such as the similarity and difference among CDCs from different \textit{A. arborescens} isolates. To better understand CDC characteristics and mechanisms of HGT, other isolates need to be sequenced.

Materials and methods

Sequencing, assembly & alignment

\textit{A.arborescens} DNA was extracted following a protocol described [77] and the sequencing library was prepared using the Illumina Paired-End DNA Sample Prep Kit. Sequencing was performed using Illumina Genome Analyzer II. Short reads were assembled \textit{de-novo} using Velvet, and quality improved by a pipeline including two alternate assemblers: Edena [44], and Minimus2 [45]. Parameters including k-mer length for Velvet and hash length for Edena were optimized by sequential step changes. The alignment between \textit{A.arborescens} and \textit{A.brassicicola} was conducted using the Nucmer program in the MUMMER suit, with parameter c=15, l=10. Alignments with identities lower than 90% or lengths shorter than 100bp were removed.
Southern hybridization

On the CHEF gel membrane presented in Figure 2.2, lane 1 contains size markers, lane 2 contains A. *arborescens* chromosomes that had degraded, and lane 3 contains intact A. *arborescens* chromosomes. Southern hybridization was conducted using the GE health CDP-Star kit with 5 gene probes, including 1 CDC marker gene *Alt1*, 3 predicted CDC genes, and 1 predicted EC gene. Primers (Table 2.9) were designed using Primer3 [78] (v0.4.0). Blots were stripped between hybridizations to ensure no probes from previous hybridization remained. Film was exposed for 48 hours.

Gene prediction, codon usage analysis & repetitive DNA identification

Gene prediction was conducted using FGENESH [47], an *ab initio* gene predictor provided in the Softberry website. A pre-trained *Alternaria* matrix was used to optimize predictions. Both CDSs and protein sequences were generated and converted into fasta format files. ACUA [79] was used for calculating CAI and RSUC for each gene, and CAI distribution curves from the CDC group and EC group were compared to each other. Student's f-test was used to analyse data. RepeatScout [80] was used for *de-novo* identification of repeat sequences in both CDC and EC sequences. The repeat libraries were then aligned back to CDC and EC contigs using Nucmer to calculate the repeat percentage for each group.
Gene annotation

Blast2go [81] was used to annotate genes by “BLASTX” to the NCBI non-redundant protein database and then GO term assignment from the gene ontology database. Annotation of conserved domains was identified by scanning proteins through Pfam and NCBI CDD. PKS and NRPS genes were identified through scanning an online database SBSPKS [82]. The Fungal transcription factor database (FTFD) [83] was used to identify transcription factors. Transporters, P450s, and oxidoreductases were identified based on BLAST and domain inspection. Potential secreted proteins were predicted using Signal 3.0 [84]. Pathogenicity and virulence factors were identified through scanning CDC genes in the pathogen-host interactions database (PHI-base) [54].

Estimating Ka/Ks Ratios

A. arborescens proteins were blasted against A.brassicicola proteins to generate a match list between the two groups with a bits score cut-off at 300. The gene sequences coding for aligned proteins were extracted by an in-house PERL script. Prank [85] was used to conduct codon alignment, in which two protein sequences were aligned first and then DNA sequences were aligned based on the corresponding protein alignments. The codon alignment result was then entered into “Codem” in PAML [86] (v4.0) for Ka and Ks calculation with model M0. In calculating, the Nei and Gojobori [87] method and Yang and Nielsen [88] method were used.
References


25. Nakatsuka S, Feng BN, Goto T, Tsuge T, Nishimura S: Biosynthesis of Host-Selective Toxins Produced by Alternaria-Alternata Pathogens .2. Biosynthetic Origin of (8r,9s)-9,10-Epoxo-8-Hydroxy-9-Methyl-Deca-(2e,4z,6e)-Trienoic Acid, a Precursor of


33. Timothy L Friesen EHS, Zhaohui Liu, Steven Meinhardt, Hua Ling, Justin D Faris1, Jack B Rasmussen, Peter S Solomon, Bruce A McDonald & Richard P Oliver: Emergence of a new disease as a result of interspecific virulence gene transfer. *Nature Genetics* 2006, 38:953-956.


45. Sommer DD, Delcher AL, Salzberg SL, Pop M: Minimus: a fast, lightweight genome assembler. BMC bioinformatics 2007, 8:


<table>
<thead>
<tr>
<th>K</th>
<th>Ins_length (bp)</th>
<th>Exp_cov</th>
<th>Cut-off</th>
<th>Min_lgth (bp)</th>
<th>N50 (bp)</th>
<th>Max_lgth (bp)</th>
<th>Total size (Mb)</th>
<th>Reads used</th>
</tr>
</thead>
<tbody>
<tr>
<td>31</td>
<td>350</td>
<td>42</td>
<td>5</td>
<td>100</td>
<td>84390</td>
<td>485954</td>
<td>33.9</td>
<td>46910423</td>
</tr>
<tr>
<td>41</td>
<td>350</td>
<td>33</td>
<td>5</td>
<td>100</td>
<td>390559</td>
<td>1219988</td>
<td>33.9</td>
<td>46820556</td>
</tr>
<tr>
<td>51</td>
<td>350</td>
<td>24</td>
<td>5</td>
<td>100</td>
<td>624394</td>
<td>1544694</td>
<td>34.0</td>
<td>46047345</td>
</tr>
<tr>
<td>57</td>
<td>350</td>
<td>16</td>
<td>5</td>
<td>100</td>
<td>441130</td>
<td>2135139</td>
<td>34.0</td>
<td>45564058</td>
</tr>
</tbody>
</table>

Table 2.1. Velvet *de-novo* assembly statistics. Results from assembly attempts using varying parameters. The row in shadow shows the optimized assembly.
<table>
<thead>
<tr>
<th>Contig ID</th>
<th>Length (bp)</th>
<th>Alignment coverage</th>
<th>Marker gene</th>
<th>New ID</th>
<th>GC%</th>
</tr>
</thead>
<tbody>
<tr>
<td>NODE_9</td>
<td>48862</td>
<td>13.6%</td>
<td>AaMSAS</td>
<td>CDC_contig_23</td>
<td>53.015%</td>
</tr>
<tr>
<td>NODE_136</td>
<td>14729</td>
<td>15.7%</td>
<td>ALT1</td>
<td>CDC_contig_27</td>
<td>24.942%</td>
</tr>
<tr>
<td>NODE_758</td>
<td>174048</td>
<td>26.1%</td>
<td>AKS17</td>
<td>EC_contig_009</td>
<td>51.127%</td>
</tr>
<tr>
<td>NODE_8</td>
<td>199480</td>
<td>44.1%</td>
<td>AKS21</td>
<td>EC_contig_038</td>
<td>51.143%</td>
</tr>
<tr>
<td>NODE_58</td>
<td>833831</td>
<td>61.9%</td>
<td>MAT1-2-1</td>
<td>EC_contig_063</td>
<td>51.209%</td>
</tr>
<tr>
<td>NODE_82</td>
<td>1201916</td>
<td>69.9%</td>
<td>ALM</td>
<td>EC_contig_085</td>
<td>51.495%</td>
</tr>
<tr>
<td>NODE_274</td>
<td>1472031</td>
<td>72.1%</td>
<td>VKS2</td>
<td>EC_contig_091</td>
<td>51.489%</td>
</tr>
<tr>
<td>NODE_151</td>
<td>812677</td>
<td>73.5%</td>
<td>AaTUB</td>
<td>EC_contig_094</td>
<td>51.705%</td>
</tr>
</tbody>
</table>

Table 2.2. Alignment of *A. arborescens* marker gene contigs and *A. brassicicola* contigs. The first two contigs contain two CDC marker genes and the following six contigs contain six EC marker genes.
<table>
<thead>
<tr>
<th>Chromosome</th>
<th>Total length (bp)</th>
<th>Repeat length (bp)</th>
<th>Repeat coverage</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDC</td>
<td>906,671</td>
<td>47,990</td>
<td>5.29%</td>
</tr>
<tr>
<td>ECs</td>
<td>32,675,713</td>
<td>186,054</td>
<td>0.57%</td>
</tr>
</tbody>
</table>

Table 2.3. Repeat region identification.
Table 2.4. Codon usage correlation analysis. Numbers in red show lower correlation scores between CDC and EC contigs compared to correlation within EC groups or within CDC groups, indicating existence of codon usage bias of CDC genes.

<table>
<thead>
<tr>
<th></th>
<th>CDC_1</th>
<th>CDC_2</th>
<th>EC_1</th>
<th>EC_2</th>
<th>EC_3</th>
<th>EC_4</th>
<th>EC_5</th>
<th>EC_6</th>
<th>EC_7</th>
<th>EC_8</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDC_1</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC_2</td>
<td>0.995</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EC_1</td>
<td>0.988</td>
<td>0.991</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EC_2</td>
<td>0.992</td>
<td>0.995</td>
<td>0.999</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EC_3</td>
<td>0.992</td>
<td>0.994</td>
<td>0.999</td>
<td>0.999</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EC_4</td>
<td>0.989</td>
<td>0.991</td>
<td>1.000</td>
<td>0.999</td>
<td>0.999</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EC_5</td>
<td>0.990</td>
<td>0.992</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
<td>1.000</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>EC_6</td>
<td>0.991</td>
<td>0.995</td>
<td>0.999</td>
<td>1.000</td>
<td>0.999</td>
<td>0.999</td>
<td>0.999</td>
<td>1.000</td>
<td></td>
<td></td>
</tr>
<tr>
<td>EC_7</td>
<td>0.992</td>
<td>0.994</td>
<td>0.999</td>
<td>0.999</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
<td></td>
</tr>
<tr>
<td>EC_8</td>
<td>0.992</td>
<td>0.994</td>
<td>0.999</td>
<td>0.999</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
<td>0.999</td>
<td>1.000</td>
<td>1.000</td>
</tr>
<tr>
<td>Codon</td>
<td>Amino Acid</td>
<td>Frequency in CDC</td>
<td>Frequency in EC</td>
<td>Significance</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>------------</td>
<td>------------------</td>
<td>-----------------</td>
<td>--------------</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAA</td>
<td>K</td>
<td>35.0%</td>
<td>29.6</td>
<td>P=0.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAG</td>
<td>K</td>
<td>65.0%</td>
<td>70.4</td>
<td>P=0.73</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAT</td>
<td>N</td>
<td>36.8%</td>
<td>32.3</td>
<td>P=0.62</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>AAC</td>
<td>N</td>
<td>63.3%</td>
<td>67.7</td>
<td>P=0.72</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAT</td>
<td>Y</td>
<td>37.7%</td>
<td>32.2</td>
<td>P=0.54</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAC</td>
<td>Y</td>
<td>62.4%</td>
<td>67.8</td>
<td>P=0.66</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 2.5. Differences in codon usage between CDC and EC genes. Top three amino acids showing the most bias are listed.
<table>
<thead>
<tr>
<th>Domain_ID</th>
<th>Protein_Ka/Ks</th>
<th>Domain/pr</th>
<th>Domain_ID</th>
<th>Protein_Ka/Ks</th>
<th>Domain/pr</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDC_151_HAD_like</td>
<td>67.10</td>
<td>0.74</td>
<td>CDC_151_HMA</td>
<td>51.70</td>
<td>0.13</td>
</tr>
<tr>
<td>CDC_84_GAL4</td>
<td>22.00</td>
<td>0.21</td>
<td>CDC_151_E1-E2_ATPase</td>
<td>103.80</td>
<td>0.10</td>
</tr>
<tr>
<td>CDC_144_Pat17_PNPLA8_PNP</td>
<td>113.50</td>
<td>0.41</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC_178_AbfB</td>
<td>327.90</td>
<td>0.12</td>
<td>CDC_68_p450</td>
<td>442.30</td>
<td>0.14</td>
</tr>
<tr>
<td>CDC_161_RTA1</td>
<td>184.00</td>
<td>0.12</td>
<td>CDC_118_MOZ_SAS</td>
<td>441.40</td>
<td>0.03</td>
</tr>
<tr>
<td>CDC_46_TA_like</td>
<td>395.20</td>
<td>0.10</td>
<td>CDC_92_Aa_trans</td>
<td>900.30</td>
<td>0.07</td>
</tr>
<tr>
<td>CDC_28_SDR_c</td>
<td>533.30</td>
<td>0.03</td>
<td>CDC_66_P21-Arc</td>
<td>438.40</td>
<td>0.03</td>
</tr>
<tr>
<td>CDC_183_Syja_N</td>
<td>916.00</td>
<td>0.02</td>
<td>CDC_169_Rubis_subs-bind</td>
<td>106.90</td>
<td>0.09</td>
</tr>
<tr>
<td>CDC_161_Peptidases_S8_S53</td>
<td>256.10</td>
<td>0.09</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC_102_CAP59_mtransfer</td>
<td>163.00</td>
<td>0.34</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC_188_Peptidases_S53</td>
<td>957.50</td>
<td>0.03</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC_178_ArabFuran-catal</td>
<td>629.00</td>
<td>0.06</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC_67_Rep_fac_C</td>
<td>196.40</td>
<td>0.02</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC_184_Rab5_related</td>
<td>240.70</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC_146_APH</td>
<td>72.30</td>
<td>0.20</td>
<td>CDC_107_PLN02844</td>
<td>469.90</td>
<td>0.06</td>
</tr>
<tr>
<td>CDC_161_PA_PoS1_like</td>
<td>245.50</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC_161_Peptidases_S8_5</td>
<td>405.40</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC_161_DUF1034</td>
<td>237.50</td>
<td>0.05</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC_51_ACC_central</td>
<td>210.10</td>
<td>0.01</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>CDC_67_AAA</td>
<td>336.40</td>
<td>0.01</td>
<td>CDC_189_Na_Ca_ex</td>
<td>277.80</td>
<td>0.00</td>
</tr>
</tbody>
</table>

Table 2.6. Ka/Ks ratio of CDC protein conversed domains. The last column shows the enrichment of Ka/Ks domains as compared to the whole protein.
<table>
<thead>
<tr>
<th>Species</th>
<th>Genome Size</th>
<th>Gene Model Number</th>
<th>Average Transcript Length</th>
</tr>
</thead>
<tbody>
<tr>
<td>Alternaria arborescens</td>
<td>34</td>
<td>9167</td>
<td>1300</td>
</tr>
<tr>
<td>Cochliobolus heterostrophus</td>
<td>34.9</td>
<td>9633</td>
<td>1639</td>
</tr>
<tr>
<td>Dothistroma septosporum</td>
<td>30.2</td>
<td>12580</td>
<td>1778</td>
</tr>
<tr>
<td>Hysterium pulicar</td>
<td>38.4</td>
<td>12352</td>
<td>1524</td>
</tr>
<tr>
<td>Leptosphaeria maculans</td>
<td>44.9</td>
<td>12469</td>
<td>1258</td>
</tr>
<tr>
<td>Mycosphaerella fijiensis</td>
<td>74.1</td>
<td>13107</td>
<td>1833</td>
</tr>
<tr>
<td>Mycosphaerella graminicola</td>
<td>39.7</td>
<td>10952</td>
<td>?</td>
</tr>
<tr>
<td>Pyrenophora teres f. teres</td>
<td>33.6</td>
<td>11799</td>
<td>1390</td>
</tr>
<tr>
<td>Pyrenophora tritici-repentis</td>
<td>37.8</td>
<td>12171</td>
<td>?</td>
</tr>
<tr>
<td>Rhytidhysteron rufulum</td>
<td>40.2</td>
<td>12117</td>
<td>1505</td>
</tr>
<tr>
<td>Septoria musiva</td>
<td>29.4</td>
<td>10233</td>
<td>1841</td>
</tr>
<tr>
<td>Setosphaeria turcica</td>
<td>43</td>
<td>11702</td>
<td>1490</td>
</tr>
<tr>
<td>Stagonospora nodorum</td>
<td>37.2</td>
<td>16597</td>
<td>?</td>
</tr>
<tr>
<td>Alternaria brassicicola</td>
<td>31</td>
<td>10688</td>
<td>1333</td>
</tr>
</tbody>
</table>

Table 2.7. General statistical facts of *A. arborescens* and other sequenced fungi genome.
<table>
<thead>
<tr>
<th>Query</th>
<th>From</th>
<th>To</th>
<th>Bitscore</th>
<th>Accession</th>
<th>Short name</th>
</tr>
</thead>
<tbody>
<tr>
<td>CDC_4</td>
<td>2</td>
<td>295</td>
<td>212.038</td>
<td>cl02872</td>
<td>DHQ_Fe-ADH superfamily</td>
</tr>
<tr>
<td>CDC_6</td>
<td>41</td>
<td>299</td>
<td>123.824</td>
<td>cl10444</td>
<td>Ras_like_GTPase superfamily</td>
</tr>
<tr>
<td>CDC_23</td>
<td>23</td>
<td>257</td>
<td>189.418</td>
<td>cl09931</td>
<td>NADB_Rossmann superfamily</td>
</tr>
<tr>
<td>CDC_23</td>
<td>3</td>
<td>259</td>
<td>208.475</td>
<td>PRK05653</td>
<td>fabG</td>
</tr>
<tr>
<td>CDC_28</td>
<td>50</td>
<td>299</td>
<td>211.374</td>
<td>cl09931</td>
<td>NADB_Rossmann superfamily</td>
</tr>
<tr>
<td>CDC_28</td>
<td>43</td>
<td>303</td>
<td>220.031</td>
<td>PRK05653</td>
<td>fabG</td>
</tr>
<tr>
<td>CDC_28</td>
<td>695</td>
<td>303</td>
<td>220.031</td>
<td>PRK05653</td>
<td>fabG</td>
</tr>
<tr>
<td>CDC_36</td>
<td>154</td>
<td>381</td>
<td>314.006</td>
<td>cl09099</td>
<td>P-loop NTPase superfamily</td>
</tr>
<tr>
<td>CDC_36</td>
<td>154</td>
<td>381</td>
<td>314.006</td>
<td>cl09099</td>
<td>P-loop NTPase superfamily</td>
</tr>
<tr>
<td>CDC_36</td>
<td>695</td>
<td>381</td>
<td>313.621</td>
<td>cl09099</td>
<td>P-loop NTPase superfamily</td>
</tr>
<tr>
<td>CDC_36</td>
<td>695</td>
<td>381</td>
<td>313.621</td>
<td>cl09099</td>
<td>P-loop NTPase superfamily</td>
</tr>
<tr>
<td>CDC_37</td>
<td>67</td>
<td>496</td>
<td>169.38</td>
<td>cl02078</td>
<td>p450 superfamily</td>
</tr>
<tr>
<td>CDC_41</td>
<td>457</td>
<td>662</td>
<td>196.457</td>
<td>cl09099</td>
<td>P-loop NTPase superfamily</td>
</tr>
<tr>
<td>CDC_41</td>
<td>150</td>
<td>688</td>
<td>161.391</td>
<td>COG1126</td>
<td>GlnQ</td>
</tr>
<tr>
<td>CDC_41</td>
<td>694</td>
<td>1102</td>
<td>421.667</td>
<td>cl09099</td>
<td>P-loop NTPase superfamily</td>
</tr>
<tr>
<td>CDC_41</td>
<td>694</td>
<td>1102</td>
<td>421.667</td>
<td>cl09099</td>
<td>P-loop NTPase superfamily</td>
</tr>
<tr>
<td>CDC_41</td>
<td>26</td>
<td>439</td>
<td>227.546</td>
<td>cl12078</td>
<td>p450 superfamily</td>
</tr>
<tr>
<td>CDC_41</td>
<td>466</td>
<td>1102</td>
<td>260.325</td>
<td>COG0369</td>
<td>CysJ</td>
</tr>
<tr>
<td>CDC_49</td>
<td>180</td>
<td>552</td>
<td>285.031</td>
<td>cl00285</td>
<td>Aconitase superfamily</td>
</tr>
<tr>
<td>CDC_50</td>
<td>1022</td>
<td>1205</td>
<td>116.368</td>
<td>cl00160</td>
<td>LbetaH superfamily</td>
</tr>
<tr>
<td>CDC_141</td>
<td>33</td>
<td>454</td>
<td>533.223</td>
<td>cl00383</td>
<td>PKS</td>
</tr>
<tr>
<td>CDC_141</td>
<td>33</td>
<td>454</td>
<td>533.223</td>
<td>cl09938</td>
<td>cond_enzymes superfamily</td>
</tr>
<tr>
<td>CDC_141</td>
<td>1237</td>
<td>1608</td>
<td>202.13</td>
<td>cl08955</td>
<td>KR_2_FAS_SDR_x</td>
</tr>
<tr>
<td>CDC_141</td>
<td>1237</td>
<td>1608</td>
<td>202.13</td>
<td>cl09931</td>
<td>NADB_Rossmann superfamily</td>
</tr>
<tr>
<td>CDC_141</td>
<td>564</td>
<td>879</td>
<td>270.822</td>
<td>cl08282</td>
<td>Acyl_transf_1 superfamily</td>
</tr>
<tr>
<td>CDC_141</td>
<td>34</td>
<td>457</td>
<td>623.228</td>
<td>smart00825</td>
<td>PKS_KS</td>
</tr>
<tr>
<td>CDC_144</td>
<td>271</td>
<td>596</td>
<td>127.448</td>
<td>cl07199</td>
<td>Pat17_PNPLA8_PNPLA9_like</td>
</tr>
<tr>
<td>CDC_144</td>
<td>271</td>
<td>596</td>
<td>127.448</td>
<td>cl11396</td>
<td>Patatin_and_cPLA2 superfamily</td>
</tr>
<tr>
<td>CDC_147</td>
<td>172</td>
<td>365</td>
<td>121.98</td>
<td>cl14604</td>
<td>Methyltransf_2 superfamily</td>
</tr>
<tr>
<td>CDC_165</td>
<td>30</td>
<td>447</td>
<td>519.356</td>
<td>cl00833</td>
<td>PKS</td>
</tr>
<tr>
<td>CDC_165</td>
<td>30</td>
<td>447</td>
<td>519.356</td>
<td>cl09938</td>
<td>cond_enzymes superfamily</td>
</tr>
<tr>
<td>CDC_165</td>
<td>1909</td>
<td>2144</td>
<td>276.372</td>
<td>cl05195</td>
<td>enoyl_red</td>
</tr>
<tr>
<td>CDC_165</td>
<td>1909</td>
<td>2144</td>
<td>276.372</td>
<td>cl14614</td>
<td>MDR superfamily</td>
</tr>
<tr>
<td>CDC_165</td>
<td>2158</td>
<td>2381</td>
<td>171.314</td>
<td>cl08955</td>
<td>KR_2_FAS_SDR_x</td>
</tr>
<tr>
<td>CDC_165</td>
<td>2158</td>
<td>2381</td>
<td>171.314</td>
<td>cl09931</td>
<td>NADB_Rossmann superfamily</td>
</tr>
<tr>
<td>CDC_165</td>
<td>640</td>
<td>940</td>
<td>283.123</td>
<td>cl08282</td>
<td>Acyl_transf_1 superfamily</td>
</tr>
<tr>
<td>CDC_165</td>
<td>1005</td>
<td>1179</td>
<td>110.38</td>
<td>cl11739</td>
<td>PKS_DH superfamily</td>
</tr>
<tr>
<td>CDC_165</td>
<td>32</td>
<td>449</td>
<td>595.108</td>
<td>smart00825</td>
<td>PKS_KS</td>
</tr>
<tr>
<td>CDC_165</td>
<td>1913</td>
<td>2144</td>
<td>293.921</td>
<td>smart00829</td>
<td>PKS_ER</td>
</tr>
<tr>
<td>CDC_183</td>
<td>57</td>
<td>410</td>
<td>355.778</td>
<td>cl11995</td>
<td>Syja_N superfamily</td>
</tr>
</tbody>
</table>

Table 2.8. Conserved domains in CDC putative PKS genes.
<table>
<thead>
<tr>
<th>Probe</th>
<th>Primer_F</th>
<th>Primer_R</th>
</tr>
</thead>
<tbody>
<tr>
<td>ALT1</td>
<td>TGCAGTCGAGCTGTCACTTT</td>
<td>GCGATCAGAGATGACGAACA</td>
</tr>
<tr>
<td>CDC_92</td>
<td>CGTCGGTTATCCTGGTCACT</td>
<td>GAATCGCAGATGCAATGATG</td>
</tr>
<tr>
<td>CDC_102</td>
<td>CCTCCGCAGCTCTACGATA</td>
<td>CCTATGCCGTCCCAACAAC</td>
</tr>
<tr>
<td>CDC_147</td>
<td>ATGATTCGGCAAATCTCTGG</td>
<td>CTTAGGTAGGCAGGCAAAG</td>
</tr>
<tr>
<td>EC_97_90_g721</td>
<td>GCCTATCTGCACCGCTCTAC</td>
<td>GATGGCGACTGCTAGACCTC</td>
</tr>
</tbody>
</table>

Table 2.9. Primers used for Southern hybridization.
Figure 2.1. Annotation of the initial two CDC contigs. The two lines represent two contigs containing CDC marker genes. Each box represents a BLAST annotation with the direction indicated by arrows.
Figure 2.2. Southern hybridization to validate CDC contigs prediction. (A) Chromosomes of *A. arborescens* separated by CHEF. Hybridization of (B) ALT1 gene, three predicted CDC genes including (C) CDC_92, (D) CDC_102, (E) CDC_147, and (F) a predicted EC gene EC_97_90_g721 are shown. It should be noted that the membrane used for hybridization was prepared in 2004, so only a limited number of hybridization experiments were possible.
Figure 2.3. Global contig alignment between *A. arborescens* contigs (blue blocks) and *A. brassicicola* contigs (red blocks). The first 11 lines contain EC contigs alignment and the last line (and the magnified inset) contains CDC contigs alignment. Alignment identity cut-off was set as 70%. Contig numbers are given above blocks. The corresponding alignment identity is shown as spots under contigs.
Figure 2.4. The Codon Adaptation Index (CAI) distribution of *A. arborescens* CDC genes compared to EC genes. The CAI derived from the RSCU estimations was computed using “Automated Codon Usage Analysis Software” (ACUA).
Figure 2.5. BLAST taxonomy report of all CDC genes against NCBI “Nucleotide collection (nr/nt)” database. Species of BLAST hits (E-value < 1.0E-3) were ranked by their occurrence.
Figure 2.6. GO term of CDC genes. GO terms were assigned to all CDC genes and level three terms in the categories Biological Process (A), Molecular Function (B), and Cellular Component (C) are showed in the corresponding pie chart.
Figure 2.7. Structure of predicted secondary metabolite biosynthesis (SMB) clusters. Each block represents a predicted CDC gene with the annotation listed above the block. Contig (NODE) numbers and color legend are given. Type I PKSs are multifunctional enzymes that are organized into modules, each harbors a set of distinct, non-iteratively acting activities responsible for the catalysis, while type II PKSs are multienzyme complexes carrying a single set of iteratively acting activities.
Figure 2.8. Average Ka/Ks ratio, Ka, and Ks for all aligned EC genes and CDC genes. 6323 EC proteins and 34 CDC proteins were involved and codem program in the PAML suite used.
Figure 2.9. Phylogenetic relationship of proteins in *A. arborescens* CDC shows discordance compared to EC proteins. The maximum-likelihood tree was constructed by “MEGA 5” using six protein sequences showing homology in all six groups.
Figure 2.10. Scatter plots of BLAST Score Ratio (BSR) of EC proteins (A) and CDC proteins (B). The numbers in yellow regions indicate the percentage of genes that either lack homologous sequences (lower left corner) or contain homologous sequences (upper right corner) in both *A. brassicicola* (*Ab*) library and 3-fungi library (*P. tritici-repentis*, *L. maculans*, and *A. oryzae*). The numbers in pink regions (upper left) indicate the percentage of genes with homologous sequences in 3-fungi library but not in *Ab* library, while numbers in green regions (lower right) indicate the percentage of genes with homologous sequences in the *Ab* library but not in the 3-fungi library.
Figure 2.11. Diversity of marker genes on CDC and ECs among different *Alternaria* isolates. The first 6 genes in black color are EC marker genes and the two in red are CDC markers.
Chapter 3: RNA-Seq Analysis of the *Sclerotinia homoeocarpa* – Creeping Bentgrass Pathosystem
Introduction

*Sclerotinia homoeocarpa* is an ascomycete fungus, which causes dollar spot disease on turfgrass all around the world [1]. *S. homoeocarpa*'s hosts include all species of turfgrass and some of dicot plants [1]. One of its hosts, creeping bentgrass (also known as *Agrostis stolonifera* L.), is a cool-season turfgrass commonly planted on golf course greens in both the United States and Canada [2]. Many commonly planted creeping bentgrass cultivars contain no resistance to dollar spot disease and the frequent, low mowing practices applied to them in a golf course setting promote disease outbreaks [1, 3].

Straw-colored, hourglass-shaped lesions with characteristic reddish-brown borders can be observed in dollar spot infected plants. Diseased areas may grow to about 2.5 cm wide [1], and the diseased areas will merge to form larger patches of diseased turf under environmental conditions for the pathogen [4-6]. However, predictive models designed to reduce fungicide applications have developed slowly because of a lack of a full understanding of *S. homoeocarpa* lifestyle, epidemiology, and disease etiology [1, 4].

The overwinter structure of *S. homoeocarpa* is a stroma and the pathogen becomes active in the spring. The fungus infects the newly emerging leaf tissue through the stomate, wounds, and directly with an appressorium [7-9]. From reports, only the infertile apothecia have been seen for North American isolates [10], but population studies indicate the possibility of genetic recombination in this fungus [11]. Early studies recorded root browning and cell death through the production of
diffusible toxins [12]. Recently, several tetranorditerpenoid compounds have been identified that can cause root-browning via extremely phytotoxic properties; however, a potential correlation between the production of these compounds and disease symptoms was still missing [13].

To control dollar spot, cultural management includes maintaining adequate nitrogen balance, sufficient air flow to remove dew, and planting moderately resistant cultivars or species of turf [14, 15]. Fungicides are often applied biweekly to weekly on highly maintained areas such as golf courses because cultural practices are not sufficient for management of the disease. As a result, high amounts of fungicide use have led to pathogen resistance to several chemical classes commonly used on turf [1]. Except for traditional chemical fungicides, other disease controlling products have been used including plant defense activators that function by stimulating two different plant defense pathways: systemic acquired resistance and induced systemic resistance. However, it is unclear which pathway is more efficient for preventing dollar spot epidemics. To develop more sustainable and practical management strategies, it will be essential to elucidate the molecular interactions between *S. homoeocarpa* and creeping bentgrass, which may greatly accelerate the development of plant defense activators and the cultivars containing increased resistance to *S. homoeocarpa*.

One of the available methods that enable researchers to investigate molecular change of pathogen and host in a transcriptome scale level is the 2nd generation sequencing (NGS) technology, which could also reduce the cost to a relatively low level in return for a vast amounts of data with quantitative properties [16-19]. In this chapter, two NGS technologies were used to generate
sequence data for RNA-seq analysis: Illumina’s sequencing-by-synthesis (SBS) and Roche’s 454-pyrosequencing. The 454 reads were used for the de novo assembly of *S. homoeocarpha* and creeping bentgrass transcriptome libraries. SBS reads were mapped to the 454 assemblies to calculate transcript levels from *S. homoeocarpha* and creeping bentgrass during dollar spot disease development. The objective of this study was to identify transcripts that may be important for fungal virulence and creeping bentgrass defense. The results of the analysis will be used to form testable hypotheses for future studies on dollar spot etiology and turfgrass defense mechanisms.

**Results**

**Transcriptomic analysis**

Six cDNA libraries for 454 sequencing and 4 cDNA libraries for SBS sequencing were prepared by Dr. Venu Reddyvari-Channarayappa. The *S. homoeocarpha* (SH) 454 sequencing data contained 600,760 reads in total, and the *A. stolonifera* (AS) 454 sequence data contained 205,40 reads in total (Table 3.1). The 454 read lengths ranged from 50 bp to > 500 bp, with a majority of the reads between 40 and 500 bp in length (Figure 3.1). The transcriptome coverage for each of the 454 assemblies was 3.3x coverage for AS and 17.2x coverage for the SH assembly. These sequencing coverage numbers were calculated by dividing the total number of sequence reads by the size of the respective, assembled transcriptome libraries.
The SBS reads, which was used for calculating significant differences in transcript levels between libraries, resulted in 4.3 – 7. Million reads (Table 3.2); however, these reads were only 16 bp long, because of the first generation of Illumina sequencing protocols we used. The SBS reads provided ample coverage of the entire transcript assemblies: 14x coverage for SH and 12x for AS. The 454 library construction resulted in 9,319 SH transcripts with length ranging from 400 bp to 3,500 bp, with a distribution peak at 500 bp. Construction of the 454 transcript library for AS resulted in 20,293 transcript sequences with length ranging from 200 bp to 8,500 bp with a distribution peak at 480 bp. The full length of transcripts was not estimated since very few annotated coding sequences from either SH or AS were available at Genebank. The mapping percentage of SBS transcripts was quite low. Only 67.4% of SH PDA-only reads, 63.8% of SH PDB-only reads, and 33.3% of Interaction reads were mapped to the SH transcripts assembly, respectively. Only 16.8% of AS only reads, and 5.7% of the Interaction reads mapped to the AS transcripts library (Table 3.2). This is likely due to the short length of the SBS reads and stringent mapping parameters used in this experiment, which discarded reads mapped to multiple locations and did not allow any base mismatches.

The quality of the SBS mapping was confirmed by visualizing coverage of previously annotated transcripts, elongation factor-1 and calmodulin-type gene (Figure 3.2). The signal tracks demonstrated that both transcript types provided a coverage of mapped reads from SH from PDA, from PDB, and from the interaction library.

A Venn diagram was constructed to identify the distribution of the unique and common reads in
SH-only, AS-only, and interaction libraries for both SBS and 454 data (Figure 3.3). There are discrepancies in the reads distribution between the SBS and 454 libraries. The most obvious example is the increase in the number of reads that are unique to the SH-only library from 6.2% in the SBS to 58.9% in the 454 reads. This is because there were more total SH reads in the 454 dataset, where four of the six libraries (69.8% of the reads) were constructed using RNA from *S. homoeocarpa* grown in pure culture. In the SBS dataset, only two out of four (31.7% of the reads) of the libraries were constructed from *S. homoeocarpa* grown in pure culture. The number of reads shared by the SH-only and AS-only libraries was consistently low in both SBS and 454 datasets. These shared reads are likely a result of similar sequences in both organisms, including those of transposable elements found in both organisms.

**Comparative analysis**

The transcript assemblies were blasted against the NCBI protein database for functional annotation using Blast2GO [20] ‘blastx’ function. There were 57% (5,257 of the 9,319) SH transcripts returned at least one annotation hit. Likewise, 57% (11,551 of the 20,293) AS transcripts returned at least one hit. From the ‘blastn’ taxonomy report, top hits of the *S. homoeocarpa* sequences were primarily from *Sclerotinia sclerotiorum* (44.1%) and *Botryotinia fuckeliana* (34.9%) (Figure 3.4a). This is not surprising considering their taxonomic similarity to *S. homoeocarpa*, and the abundance of *S. sclerotiorum* and *B. fuckeliana* sequence data available on GeneBank. The 26 fungal species
represented in the taxonomy report belong to various classes of the subphylum Pezizomycotina, with seven of the 26 species being pathogens of monocot plants. There were a small number of sequences related to monocots including rice (*Oryza sativa*, 1.6%), bamboo (*Phyllostachys edulis*, 0.3%), and barley (*Hordeum vulgare*, 0.4%). We attributed this either to contamination of the fungal sequences by *A. stolonifera* sequences in the interaction library, to the presence of very highly conserved sequences, or to the presence of sequences occurring in both organisms, such as those of transposable elements. The top blast hit species for the *A. stolonifera* sequences included predominantly monocot grass species (Figure 3.4b). Fungal sequences that were found included those of *Pyrenophora teres* Smed.-Pet. And *Phaeosphaeria nodorum* (E. Mull) Hedjar, pathogens of barley and wheat, respectively.

**Differentially expressed gene Identification**

Reads Per Kilobase of exon model per Million mapped reads (RPKM) value was calculated of all SH and AS transcripts in different libraries and the Log2 ratio was calculated to reflect the fold changes of interaction library compared to control libraries (SH PDA and AS-only).

To identify transcripts that were up-regulated in the interaction condition, transcripts with statistically significant changes of RPKM values during infection were examined to determine if there were patterns in the types of genes with increased or decreased transcription level. Table 3.3
summarizes some of the types of SH transcripts of interest during infection including a long list of glycosyl hydroalase transcripts and various proteases. Table 3.4 summarizes transcripts of interest in the AS library including a long list of retrotransposons, defense-related proteins including 13 different germin-like proteins, and various enzymes involved in plant hormone synthesis and regulation. Interestingly, all jasmonate-induced protein transcripts were significantly down regulated at 96 hpi.

**Gene ontology terms analysis**

The list of significantly up-regulated SH and AS transcripts were sorted into gene ontology (GO) term categories for molecular function, biological processes, and cellular component using the Blast2GO [20] program (Figures 3.5 and 3.6). The molecular function terms associated with up-regulated SH transcripts include a large percentage of plant cell-wall degrading enzymes categories (Figure 3.5a) including polygalacturonase activity (4.0%), cellulose activity (5.1%), endopeptidase activity (4.0%), α-arabinofuranosidase activity (2.9%), and serine-type peptidase activity (5.1%). Other up-regulated molecular function of GO terms included those for xenobiotic-transporting activity (5.1%) and oxidoreductase activity (7.4%). Biological process of up-regulated SH transcript GO terms included auxin biosynthesis (11.4%), transmembrane transport (18.9%), and oxidation reduction (25.1%) (Figure 3.5b). A majority of the up-regulated SH cellular component GO terms were for those integral to the membrane (37.2%) and those for extracellular regions (34.0%) (Figure 3.5c), perhaps
reflecting the secretion of cell wall degrading enzymes and other secondary metabolites important for pathogenicity and virulence of \( S. \) homoeocarpa.

The molecular function of GO terms associated with the up-regulated AS transcripts included those associated with hydrolase activity (19.2%), DNA binding activity (10.9%), and nucleotide binding (24.3%) (Figure 3.6a). The Biological process GO terms associated with the up-regulated AS transcript library included those related to translation (10.8%), transport (14.6%), and DNA metabolic processes (11.1%) (Figure 3.6b). Finally, an overwhelming number of cellular process GO terms from the AS up-regulated library were associated with the mitochondrion (30.8%) andoplasts (31.8%) (Figure 3.6c).

Conserved domain analysis

Protein domains were identified in the \( S. \) homoeocarpa and \( A. \) stolonifera transcripts libraries using HMMER software (v3.0) [21]. Also among SH transcripts, 1588 were predicted to contain a signal peptide by SignalP (v4.0) [22], and thus could be candidates for secreted proteins. Three proteases with secretion signals included two subtilisin/sedolisin proteases with log-fold increases of 64 and 6.2 as well as a cuticle degrading serine protease with a log-fold increase of 4.0. Conserved domains significantly enriched in the interaction library (P<0.01) are listed in Table 3.5 (SH conserved domains) and Table 3.6 (AS conserved domains). Domains enriched in the SH interaction library include a
variety of glycosyl hydrolases, proteases domains, and transporters. Domains of interest in the enriched AS interaction library include cytochrome P450, various ABC transporters, and cupin domains, which include germin-class enzymes.

**Identification of microsatellite markers**

Simple sequence repeats (SSRs) were identified using Microsatellite Identification Tool (MISA). In the AS transcriptome library, 380 SSR markers were identified. The distribution in size of these SSRs is shown in Table 3.7. The largest groups of SSR markers in grass were the single and tri-nucleotide repeats that made up approximately 47.7% and 27.2% of the identified markers, respectively. There were 715 SSR sequences identified in the SH transcripts library (Table 3.7). Similar to the AS SSRs, the largest groups were the single and tri-nucleotide repeats that made up 37.2% and 30.9% of the identified markers, respectively.

**Validation of RPKM value by RT-PCR**

To verify the RPKM calculation, RT-PCR and semi-quantitive PCR of selected genes were performed by Dr. Angela Orshinsky. Firstly, Five SH transcripts up-regulated in infection stage were selected for semi-quantitive PCR, which all showed positive correlation between RPKM fold change and relative intensity of gel bands (Figure 3.7). To further confirm RPKM calculations, real-time RT-PCR assays
were performed for genes with significantly different transcription levels during infection. The relative transcription data verified that the RPKM calculation and statistical analysis of transcriptome data was accurate (Figure 3.8). The real time RT-PCR data confirms the up-regulation of two fungal xylanases (SH_5411 and SH_5726), a fungal laccase (SH_7961), and a creeping bentgrass germin protein (AS_608). Down-regulated transcripts SH_6925 (scyatalone dehydrogenase) and SH_8369 (β-mannosidase) were also verified.

Discussions

Both SBS and 454 technologies have been reported to show high reproducibility and accuracy for transcriptome studies [18, 23], and reached higher-quality data generation than microarray and cloning based EST method [18, 19]. Compared to 454, SBS technology is considered to be superior because of the greater sequencing coverage for a similar cost. In this study, we obtained a mapping percentage of SBS to the *S. homoeocarpa* assembly ranging from 33.3-67.4%, and *A. stolonifera* mapping percentages ranging from 5.7-16.8%. The low mapping percentage could be explained by the highly stringent parameters we applied, which eliminated all reads with even a single mismatch and multi-locus mapping reads, contributing factors may also be transcriptome misassembly. The use of our analysis pipeline and the statistics applied was verified with real-time PCR. It should be noted that mapping uncertainty exists and accounts for a large portion of the reads in the mapping, especially for SBS reads, since short sequence lengths make it difficult to distinguish paralogous
genes, alternatively spliced isoforms, and low complexity sequences, all of which result in multi-locus mapping reads that are discarded in process [24]. Low mapping percentage may also be due to the incompletely assembled transcriptome library using 454 reads.

It is now possible to identify potential pathogenicity factors highly transcribed by *S. homoeocarpa* during infection of *A. stolonifera* by RNA-seq analysis. The most surprising discovery was the number of enriched *S. homoeocarpa* transcripts within the interaction library that encoded glycosyl hydrolase enzymes. In total, 52 of the up-regulated SH transcripts encoded glycosyl hydrolase genes and 22.3% of molecular function GO terms were associated with glycosyl hydrolase activity. A keyword search using search terms “glycosyl hydrolase”, “cellulose” and “pectinase” in the SH transcript library revealed over 100 transcripts annotated as glycosyl hydrolase, representing 85 different families of the total 125 glycosyl hydrolase families [25, 26]. The enriched number and such a varied group of glycosyl hydrolase families is consistent with the ability of this pathogen to cause disease on a wide range of plant hosts. From the data in this project, *S. homoeocarpa* glycosyl hydrolase genes showing up-regulation were mostly xylanases and arabinose. This was expected since grass cell walls contain roughly 40% xylans as well as glucuronoxynoxylans (GAX) which make up a majority of monocot hemicellulose [27]. It was reported that pectinases, xylanases, and celluases were often virulence and pathogenicity factors for fungal phytopathogens. For example, *Botrytis cinera* Pers. and *Septoria nodorum* (Berk.), showed reduced virulence in a mutant of a cell wall degrading enzyme [28-30]. However, the large number of up-regulated and functionally redundant cell wall degrading enzymes in the interaction library will make it very difficult to verify the role of any single enzyme.
A large number of proteinases, mostly serine proteases, were also found associated with pathogen inoculation. Secretion signals were detected on three of the proteases of interest, including two subtilisin/sedolisin proteases and one identified as a cuticle degrading serine protease. There were a 4.0-6.4 log-fold up-regulation for these secreted proteases in the interaction library. In the molecular function GO terms, 5.1% were associated specifically with serine-proease activity. Serine proteases are known as potential pathogenicity determinants for a variety of fungal phytopathogens including nematode-parasitizing fungi, entomopathogens [31].

It was observed that multi-drug resistance (MDR) ATP binding cassette (ABC) transporter proteins were also up-regulated in the S. homoeocarpa interaction condition. Members of the MDR-ABC transporter family play important roles in resistance to plant defense compounds and fungicide, as well as efflux of endogeneous toxins [32, 33]. In B. cinerea, both ABC and MFS transporters have been implicated in resistance of the pathogen to a wide range of fungicides including phenylpyrroles, azoles, anilopyrimidines, dicarboximides, and strobilurins [34-36]. Also, the enrichment of different transport proteins with redundant functions in the S. homoeocarpa interaction library may explain the development of isolates that show resistance to multiple fungicide class.

On the plant side, some of the most prominent up-regulated defense-related transcripts include the germin and germin-like proteins. Germins belong to the cupin superfamily and one important defense enzyme – oxalate oxidase [37] – were put in this category. Germin-type oxalates are enzymes which are very unique to Gramineaeious plants [37]. Oxalate oxidase has been known to be
more active in moderately resistant creeping bentgrass cultivars such as L-93, compared to Crenshaw which is a highly susceptible cultivar [38]. In addition to the plant defense-related transcripts, other significantly up-regulated creeping bentgrass transcripts were annotated as transposons and retrotransposons, which is not surprising since it has been reported that transposons make up about 80% of the DNA in cereal crops [39]. Retrotransposons were reported to be transcriptionally active in plants undergoing stress such as during wounding and in the presence of salicylic acid and jasmonate [40]. Although the clear connection between transposons and plant defense was not discovered, their increased transcription level during infection should be further investigated.

The RNA-seq analysis presented in this project showed the value of next generation sequencing technology for helping elucidate mechanism of a plant-host interaction system. The rapid development of the sequencing technology since the completion of this project has already increase reads lengths and provided deeper sequencing coverage with a relatively lower cost. Also, the development and upgrade of analysis tools has made the assembly process faster and more accurate. This resulting data analysis will be used to support future research efforts to characterize *S. homoeocarpa* virulence factors and creeping bentgrass defense-related transcripts.
Materials and Methods

Preparation RNA for SBS sequencing libraries

RNA libraries for SBS and 454 sequencing were prepared by Dr. Venu Reddyvari-Channarayappa in this project.

*Sclerotinia homoeocarpa* isolate MB01 was isolated from creeping bentgrass at The Ohio State University Turfgrass Research and Education Facility. The fungus was cultured on PDA media and 5mm plugs of actively growing mycelium were transferred to PDB medium and incubated at 26°C with shaking at 160 rpm for 96 h. Cultures grown on PDA overplayed with cellophane membranes were also used for SBS library construction. At 96 h, mycelia were filtered from PDB or scraped from the cellophane membrane on PDA, and total RNA was extracted using TriReagent (Sigma) according to manufacturer’s instructions. For interaction and creeping bentgrass libraries, creeping bentgrass cv. Crenshaw was grown in a growth chamber at 26°C day and 22°C night temperatures with a 12 h photoperiod and at 70% relative humidity. Three-week old seedlings were challenged with millet seeds colonized by *S. homoeocarpa* MB01. These plants were incubated in clear bags to create humidity for 48 h after which the plants were uncovered during the day and bagged again for the night. After 96 h, leaves from inoculated and non-inoculated plants were harvested, homogenized in liquid nitrogen, and RNA was extracted using TriReagent according to the manufacturer’s instruction.

The SBS library templates were prepared using the Illumina Duplex-Specific thermostable nuclease
normalization kit and were analyzed using the Illumina Genome Analyzer I (GAI) at the Molecular and Cellular Imaging Center of The Ohio State University in Wooster, Ohio. The resulting libraries included a 96 h PDB culture control, a 96 h PDA culture control, a 96 h creeping bentgrass control, and a 96 h \textit{S. homoeocarpa} – creeping bentgrass interaction library.

\textbf{Preparation RNA for 454 sequencing libraries}

\textit{S. homoeocarpa} isolate MB01 was cultured on PDA and 5 mm plugs of actively growing mycelium were transferred to PDB medium and incubated at 26°C with shaking at 160 rpm for 48, 96, or 144h. At each time point, mycelia were filtered and total RNA was extracted using TriReagent according to manufacturer’s instructions. For interaction and creeping bentgrass libraries, creeping bentgrass cv. Crenshaw was grown in a growth chamber as described previously. Three-week old seedlings were inoculated in one of two ways: with a 5mm plug of actively growing mycelium from a PDA culture or misting of a mycelia homogenate, in water, from a 96h PDB culture of \textit{S. homoeocarpa} MB01. These plants were incubated as described for SBS library preparation. After 96 h, leaves from inoculated and non-inoculated plants were harvested, homogenized in liquid nitrogen, and RNA was extracted using TriReagent according to the manufacturer’s instructions. Total RNA was sent to the Core Genomics Facility at Purdue University, Indiana for preparation and sequencing using Roche GS-FLX (454) sequencer with Titanium chemistry. Seven RNA-seq libraries were prepared that included \textit{S. homoeocarpa} PDB cultures at 48, 96, and 144 h after inoculation, \textit{S. homoeocarpa} grown on PDA at
96 h, creeping bentgrass non-inoculated controls, and for creeping bentgrass incubated for 96 h with

*S. homoeocarpa* applied as a PDA culture plug or mycelia homogenate in water.

**Sequencing and transcripts library assembly**

The 454 raw sequence reads were assembled using GS Data Analysis Software (v2.5, Roche, Indianapolis, IN) after removal of adaptor sequences. *S. homoeocarpa* (SH) reads from the PDA, PDB, and 96 hip interaction libraries, and the *A. stolonifera* (AS) reads from the 96h control and 96 h interaction libraries were assembled using GS *de-novo* assembler (v2.5, Roche) with parameter setting as 90% identity and a minimum 40 bp overlap. To eliminate SH reads in AS assembly library, reads from the SH interaction libraries that matched AS GeneBank EST reads were removed from the SH transcript assembly. Then reads from the AS interaction library matching SH assembly were also removed. Both SH and AS transcriptome assembly as used in this study was submitted to GeneBank TSA database (Accession: PRJNA84359).

Full length SBS reads were directly mapped to the SH and AS transcriptome libraries using the Maq (v0.7) [41] as mapping tool. Mapping parameters allowed for only unique mapped reads with no mismatches. The number of mapped SBS reads for each transcript was counted by “soap.coverage” tool in SOAPAligner package[42].
Transcript expression analysis

RPKM value were calculated for each transcript by dividing number of mapped reads by length of the transcripts and number of total sequenced reads in this library. Fisher’s Exact Test, a test used for analyzing unreplicated transcript data [43] was used to determine statistical significance of transcript RPKM value change. The test was done using a pairwise comparison of interaction RPKM value versus RPKM from fungus grown on PDA, or for interaction RPKM values non-inoculated control grass. The SH assembly contained 9,319 unique annotated transcripts and the AS assembly contained 20,293 unique annotated transcripts. Therefore, the statistical significance of the Fisher’s Test was evaluated against a Bonferroni corrected P-value of $1.86 \times 10^{-5}$ and $2.43 \times 10^{-6}$ for fungal and grass transcripts, respectively. Furthermore, only transcripts with at least 2-log fold change in transcript abundance were selected as up-regulated. These conservative criteria were applied to avoid false positives.

The log-fold RPKM values of selected transcripts were validated using real-time RT-PCR reaction on an iQ5 real time thermocycler. RNA was extracted from fungus grown on PDA, non-inoculated grass, and infected grass as descript above. RNA was extracted using Trizol (Invitrogen) and RNA extracts were treated with RQ1 DNAse (Promega) according to the manufacturer’s directions, except that RNAse Inhibitor (Promega) was added to each 100ul of RNA extract. Complementary DNA was created from 1ug of total RNA using iScript (BioRad). PCR was conducted using the SYBR supermix (Biorad) as per manufacturer’s directions. The cycle conditions were 95°C for 3 minutes followed by
40 cycles of 95°C for 30 s, 58.5°C for 30 s, and 72°C for 30 s. The melting temperature profiles and gel electrophoresis was used to evaluate the specificity of the reactions and the absence of primer dimers. Real-time RT-PCR was conducted with three biological replicates for each of grass, interaction, and fungal cDNA samples. Three technical replicates of each cDNA sample were used in each experiment.

Reaction efficiency and relative expression data were analyzed using the relative expression software tool (REST) program (Qiagen). Log base two values of relative expression ratios calculated in the REST program and the corresponding log base two RPKM expression ratios were compared graphically.

Transcripts library annotation

Taxonomic and functional annotation of the SH and AS transcripts were conducted by using Blast2GO [20] software to run “blastx” and “blastn” algorithms against non-redundant nucleo-tide/protein database from NCBI. Within the SH library, any transcripts that resulted in top blastn hit species of *Hordeum, Oryza, Sorghum, Agrostis, or Vitis* were removed. Combined graphs for GO terms associated with statistically up-regulated transcripts are presented in the results section.
**Conserved domain analysis**

Transcript sequences were translated to proteins in all 6 possible frames, and conserved protein domains were identified using HMMER (v3.0) [21] to screen the Pfam-A database. All the predicted domains with E-value <0.01 were reported. The enrichment ratio and significant value was calculated based on a Fiser’s Exact Test. Descriptions for enriched domains were found on the Pfam website.

**Identification of microsatellite markers**

SSR markers were identified using Micro Satellite identification tool (MISA) in both SH and AS transcripts libraries. Input files were the complete sequences from SH and AS libraries. Parameters allowed for the inclusion of mononucleotide repeats that were a minimum of 10ntd long, dinucleotides with minimum of 6 repeats, trinucleotide minimum of 5 repeats, tetranucleotide minimum of 5 repeats, pentanucleotides with a minimum of 5 repeats, and hexanucleotides with a minimum of 5 repeats. The maximum interruption distance between 2 SSRs was 10 ntd.
References


43. Pfaffl MW, Horgan GW, Dempfle L: Relative expression software tool (REST (c)) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic acids research 2002, 30(9).
## Table 3.1. Characteristics of the 454 RNA-Seq reads.

<table>
<thead>
<tr>
<th>Transcript Assembly</th>
<th>Sequencing Libraries Included</th>
<th>Number of Reads</th>
<th>Number of Isotigs</th>
<th>Mean of Isotig Length (bp)</th>
<th>Number of Singletons</th>
</tr>
</thead>
<tbody>
<tr>
<td>Total SH</td>
<td>SH PDA 96h</td>
<td>600,760</td>
<td>10,101</td>
<td>1,172</td>
<td>51,502</td>
</tr>
<tr>
<td></td>
<td>SH PDB 48h / 96h / 144h</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Interaction 96h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total AS</td>
<td>AS 96h</td>
<td>205,403</td>
<td>5,017</td>
<td>898</td>
<td>58,446</td>
</tr>
<tr>
<td>Interaction 96h</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Table 3.2. Mapping characteristics of the *Sclerotinia homoeocarpa* (SH) and *Agrostis stolonifera* (AS) SBS reads to the SH and AS transcript assemblies.

<table>
<thead>
<tr>
<th>SBS Library</th>
<th>SH PDA</th>
<th>SH PDB</th>
<th>AS 96h</th>
<th>Interaction</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of Reads</td>
<td>6,101,988</td>
<td>4,350,510</td>
<td>6,885,250</td>
<td>7,182,868</td>
</tr>
<tr>
<td>Mapping Percentage to SH Assembly</td>
<td>67.4%</td>
<td>63.8%</td>
<td>0.9%</td>
<td>33.3%</td>
</tr>
<tr>
<td>Mapping Percentage to AS Assembly</td>
<td>2.6%</td>
<td>2.4%</td>
<td>16.8%</td>
<td>5.7%</td>
</tr>
<tr>
<td>Transcript Group</td>
<td>Description</td>
<td>Transcripts</td>
<td>Log fold change</td>
<td></td>
</tr>
<tr>
<td>---------------------------</td>
<td>-----------------------------------</td>
<td>-------------</td>
<td>-----------------</td>
<td></td>
</tr>
<tr>
<td>Glycosyl hydrolases</td>
<td>ferruloyl esterase</td>
<td>2</td>
<td>4.8 - 11.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>dihydroceramidase</td>
<td>1</td>
<td>3.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>amylase</td>
<td>1</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>arabinofuranosidase</td>
<td>7</td>
<td>2.4 - 9.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rhamnosidase</td>
<td>1</td>
<td>2.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>endoglucanase</td>
<td>4</td>
<td>2.9 - 5.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>glucosidase</td>
<td>4</td>
<td>3.1 - 4.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>xyllosidase</td>
<td>5</td>
<td>2.6 - 5.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cellobiohydrolase</td>
<td>4</td>
<td>4.3 - 6.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cellobiose dehydrogenase</td>
<td>1</td>
<td>4.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cellulase</td>
<td>1</td>
<td>3.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>cutinase</td>
<td>1</td>
<td>3.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mannosidase</td>
<td>2</td>
<td>3.1 - 4.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>xylanase</td>
<td>3</td>
<td>4.6 - 8.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>glucanase</td>
<td>8</td>
<td>2.2 - 7.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>polygalacturonase</td>
<td>3</td>
<td>3.0 - 8.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>rhamnogalacturonase</td>
<td>2</td>
<td>2.9 - 8.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td>lysozyme</td>
<td>1</td>
<td>5.7</td>
<td></td>
</tr>
<tr>
<td></td>
<td>pectinesterase</td>
<td>1</td>
<td>2.0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>52</td>
<td>2.0 - 11.5</td>
<td></td>
</tr>
<tr>
<td>Proteases</td>
<td>serine protease</td>
<td>4</td>
<td>2.9 - 9.4</td>
<td></td>
</tr>
<tr>
<td></td>
<td>aspergillopepsin-2</td>
<td>1</td>
<td>2.3</td>
<td></td>
</tr>
<tr>
<td></td>
<td>peptidases</td>
<td>8</td>
<td>2.4 - 6.5</td>
<td></td>
</tr>
<tr>
<td></td>
<td>neutral protease</td>
<td>1</td>
<td>5.8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>14</td>
<td>2.3 - 9.4</td>
<td></td>
</tr>
<tr>
<td>Transporters</td>
<td>mdr-like abc transporter</td>
<td>7</td>
<td>4.1 - 6.6</td>
<td></td>
</tr>
<tr>
<td></td>
<td>mfs multidrug</td>
<td>2</td>
<td>2.7 - 2.9</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Total</td>
<td>9</td>
<td>2.7 - 6.6</td>
<td></td>
</tr>
</tbody>
</table>

Table 3.3. Summary of select *Sclerotinia homoeocarpa* transcript types that are significantly increased at 96 hpi after inoculation onto creeping bentgrass.
<table>
<thead>
<tr>
<th>Transcript Group</th>
<th>Product</th>
<th>Transcripts</th>
<th>Log fold change</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Transposon</strong></td>
<td>ana transposon</td>
<td>1</td>
<td>5.8</td>
</tr>
<tr>
<td></td>
<td>copia-type retroelement</td>
<td>2</td>
<td>2.6-4.3</td>
</tr>
<tr>
<td></td>
<td>gag-pol polyprotein</td>
<td>10</td>
<td>2.4-4.9</td>
</tr>
<tr>
<td></td>
<td>mutator-like transposase</td>
<td>1</td>
<td>4.6</td>
</tr>
<tr>
<td></td>
<td>retroelement pol poly</td>
<td>2</td>
<td>2.2-2.7</td>
</tr>
<tr>
<td></td>
<td>retroelement line</td>
<td>1</td>
<td>3.0</td>
</tr>
<tr>
<td></td>
<td>retrotransposon ty1-copia</td>
<td>6</td>
<td>3.4-5.8</td>
</tr>
<tr>
<td></td>
<td>retrotransposon ty3-gypsy</td>
<td>13</td>
<td>2.5-8.6</td>
</tr>
<tr>
<td></td>
<td>transposon en spm</td>
<td>15</td>
<td>2.0-5.2</td>
</tr>
<tr>
<td></td>
<td>unidentified transposon</td>
<td>2</td>
<td>2.3-3.7</td>
</tr>
<tr>
<td></td>
<td>retrotransposon unclassified</td>
<td>18</td>
<td>2.4-7.6</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td>72</td>
<td>2.0-8.6</td>
</tr>
<tr>
<td><strong>Defense</strong></td>
<td>antranilate synthase</td>
<td>1</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>calcineurin</td>
<td>1</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>cinnamoyl reductase</td>
<td>1</td>
<td>3.5</td>
</tr>
<tr>
<td></td>
<td>cytochrome p450</td>
<td>5</td>
<td>2.0-5.5</td>
</tr>
<tr>
<td></td>
<td>disease resistance nbs-lrr</td>
<td>3</td>
<td>2.3-4.0</td>
</tr>
<tr>
<td></td>
<td>e3 ubiquitin-protein ligase</td>
<td>2</td>
<td>3.7-4.0</td>
</tr>
<tr>
<td></td>
<td>fusarium resistance, i2c-5-like</td>
<td>1</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>germin a</td>
<td>13</td>
<td>3.4-6.2</td>
</tr>
<tr>
<td></td>
<td>mdr abc transporter</td>
<td>6</td>
<td>5.9-7.7</td>
</tr>
<tr>
<td></td>
<td>pathogenesis associated pep2</td>
<td>1</td>
<td>9.1</td>
</tr>
<tr>
<td></td>
<td>rust resistance kinase lr10</td>
<td>1</td>
<td>4.5</td>
</tr>
<tr>
<td></td>
<td>terpene synthase</td>
<td>1</td>
<td>2.9</td>
</tr>
<tr>
<td></td>
<td>ubiquitin</td>
<td>2</td>
<td>2.1-4.8</td>
</tr>
<tr>
<td></td>
<td>ubiquitin-conjugating enzyme</td>
<td>1</td>
<td>2.1</td>
</tr>
<tr>
<td></td>
<td>ubiquitin-specific protease</td>
<td>1</td>
<td>3.8</td>
</tr>
<tr>
<td></td>
<td>zingiberene synthase</td>
<td>1</td>
<td>2.8</td>
</tr>
<tr>
<td></td>
<td><strong>Total</strong></td>
<td>47</td>
<td>2.0-9.1</td>
</tr>
</tbody>
</table>

Table 3.4. Summary of select *Agrostis stolonifera* transcript types that are significantly increased at 96 hpi after inoculation with *Sclerotinia homoeocarpa*.
<table>
<thead>
<tr>
<th>Domain name</th>
<th>Enrichment</th>
<th>P-value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Glyco_hydro_45</td>
<td>17.0</td>
<td>0.0030</td>
<td>Endoglucanase</td>
</tr>
<tr>
<td>Glyco_hydro_62</td>
<td>17.0</td>
<td>0.0030</td>
<td>Alpha- L-arabinofuranosidase</td>
</tr>
<tr>
<td>Glyco_hydro_61</td>
<td>15.3</td>
<td>&lt;0.0001</td>
<td>Endoglucanase</td>
</tr>
<tr>
<td>Peptidase_S28</td>
<td>13.6</td>
<td>0.0010</td>
<td>Serine carboxypeptidase S28</td>
</tr>
<tr>
<td>CBM_1</td>
<td>13.4</td>
<td>&lt;0.0001</td>
<td>Fungal cellulose binding domain</td>
</tr>
<tr>
<td>E1_DerP2_DerF2</td>
<td>12.8</td>
<td>0.0003</td>
<td>Lipid recognition, recognition of pathogen products</td>
</tr>
<tr>
<td>Pectinesterase</td>
<td>12.8</td>
<td>0.0003</td>
<td>Pectinesterase</td>
</tr>
<tr>
<td>Eno-Rase_NADH_b</td>
<td>12.2</td>
<td>0.0003</td>
<td>NAD(P)H binding domain, trans-2-enoyl-CoA reductase</td>
</tr>
<tr>
<td>Glyco_hydro_43</td>
<td>11.4</td>
<td>&lt;0.0001</td>
<td>Arabinanase</td>
</tr>
<tr>
<td>COX3</td>
<td>11.4</td>
<td>0.0016</td>
<td>Cytochrome c oxidase subunit III</td>
</tr>
<tr>
<td>Flavodoxin_2</td>
<td>11.4</td>
<td>0.0016</td>
<td>Flavodoxin-like fold</td>
</tr>
<tr>
<td>Glyco_hydro_11</td>
<td>11.4</td>
<td>0.0016</td>
<td>Xylanases</td>
</tr>
<tr>
<td>Glyco_hydro_12</td>
<td>11.4</td>
<td>0.0016</td>
<td>Endoglucanase and xyloglucan hydrolase</td>
</tr>
<tr>
<td>Nucleoaplasmin</td>
<td>11.4</td>
<td>0.0016</td>
<td>Chromatin decondensation proteins</td>
</tr>
<tr>
<td>PLA2_B</td>
<td>11.4</td>
<td>0.0016</td>
<td>Lyso phospholipase catalytic domain</td>
</tr>
<tr>
<td>Glyco_hydro_28</td>
<td>11.0</td>
<td>&lt;0.0001</td>
<td>Polygalacturonase, Rhamnogalacturonase A</td>
</tr>
<tr>
<td>ABC_ATPase</td>
<td>10.8</td>
<td>&lt;0.0001</td>
<td>Predicted ATPase of the ABC class</td>
</tr>
<tr>
<td>Syja_N</td>
<td>10.2</td>
<td>0.0006</td>
<td>SacI homology domain</td>
</tr>
<tr>
<td>Cellulase</td>
<td>9.7</td>
<td>0.0023</td>
<td>Cellulase</td>
</tr>
<tr>
<td>CFEM</td>
<td>9.7</td>
<td>0.0023</td>
<td>Cysteine rich, putative role in fungal pathogenesis</td>
</tr>
<tr>
<td>Flavodoxin_5</td>
<td>8.5</td>
<td>0.0030</td>
<td>Flavodoxin</td>
</tr>
<tr>
<td>FMN_red</td>
<td>8.5</td>
<td>0.0030</td>
<td>NADPH-dependent FMN reductase</td>
</tr>
<tr>
<td>Glyco_hydro_7</td>
<td>8.5</td>
<td>0.0030</td>
<td>Endoglucanase; cellulbiohydrolase</td>
</tr>
<tr>
<td>Glyco_hydro_92</td>
<td>8.5</td>
<td>0.0030</td>
<td>Alpha-1,2-mannosidases</td>
</tr>
<tr>
<td>Mannosyl_trans3</td>
<td>8.5</td>
<td>0.0030</td>
<td>Mannosyltransferase putative</td>
</tr>
<tr>
<td>Yos1</td>
<td>8.5</td>
<td>0.0030</td>
<td>Transport between ER and the Golgi complex</td>
</tr>
<tr>
<td>7tm_1</td>
<td>8.5</td>
<td>0.0089</td>
<td>7 transmembrane receptor</td>
</tr>
<tr>
<td>Chorismate_synt</td>
<td>8.5</td>
<td>0.0089</td>
<td>Chorismate synthesis</td>
</tr>
<tr>
<td>COX6B</td>
<td>8.5</td>
<td>0.0089</td>
<td>Cytochrome oxidase c subunit Vlb</td>
</tr>
</tbody>
</table>

Table 3.5. Top enriched *Sclerotinia homoeocarpa* domains at 96 h post inoculation on creeping bentgrass.
Table 3.5 continued

<table>
<thead>
<tr>
<th>Name</th>
<th>E-value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cu-oxidase_2</td>
<td>8.5</td>
<td>Multicopper oxidase</td>
</tr>
<tr>
<td>Cupin_5</td>
<td>8.5</td>
<td>Cupin superfamily</td>
</tr>
<tr>
<td>DHDPS</td>
<td>8.5</td>
<td>Dihydrodipicolinate synthetase family</td>
</tr>
<tr>
<td>Fig1</td>
<td>8.5</td>
<td>Ca2+ regulator and membrane fusion protein Fig1</td>
</tr>
<tr>
<td>Metallophos_2</td>
<td>8.5</td>
<td>Calcineurin-like phosphoesterase</td>
</tr>
<tr>
<td>Pex2_Pex12</td>
<td>8.5</td>
<td>N terminal of a number of known and predicted peroxins</td>
</tr>
<tr>
<td>Ribosomal_L35Ae</td>
<td>8.5</td>
<td>Ribosomal protein L35Ae</td>
</tr>
<tr>
<td>SAPS</td>
<td>8.5</td>
<td>SIT4 phosphatase-associated protein</td>
</tr>
<tr>
<td>UbiA</td>
<td>8.5</td>
<td>UbiA prenyltransferase family</td>
</tr>
<tr>
<td>ABC_membrane</td>
<td>8.2</td>
<td>ABC transporter transmembrane region</td>
</tr>
<tr>
<td>Fungal_lectin</td>
<td>6.8</td>
<td>Fungal fucose-specific lectin</td>
</tr>
<tr>
<td>OPT</td>
<td>6.2</td>
<td>OPT oligopeptide transporter protein</td>
</tr>
<tr>
<td>Peptidase_S8</td>
<td>6.1</td>
<td>Subtilase family</td>
</tr>
<tr>
<td>ABC_tran</td>
<td>5.8</td>
<td>ABC transporter</td>
</tr>
<tr>
<td>Domain name</td>
<td>Enrichment</td>
<td>P-value</td>
</tr>
<tr>
<td>---------------------</td>
<td>------------</td>
<td>----------</td>
</tr>
<tr>
<td>eIF-5_eIF-2B</td>
<td>20.2</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ATS3</td>
<td>20.2</td>
<td>0.002</td>
</tr>
<tr>
<td>BDS_I_II</td>
<td>20.2</td>
<td>0.002</td>
</tr>
<tr>
<td>CSD</td>
<td>20.2</td>
<td>0.002</td>
</tr>
<tr>
<td>Folate_rec</td>
<td>20.2</td>
<td>0.002</td>
</tr>
<tr>
<td>MaoC_dehydratas</td>
<td>20.2</td>
<td>0.002</td>
</tr>
<tr>
<td>SH3_1</td>
<td>20.2</td>
<td>0.002</td>
</tr>
<tr>
<td>TAFII28</td>
<td>20.2</td>
<td>0.002</td>
</tr>
<tr>
<td>ABC2_membrane_3</td>
<td>15.1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Fer4</td>
<td>13.5</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Fer4_2</td>
<td>13.5</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Fer4_6</td>
<td>13.5</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Ribosomal_S2</td>
<td>13.5</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>CALCOCO1</td>
<td>13.5</td>
<td>0.0077</td>
</tr>
<tr>
<td>iPGM_N</td>
<td>13.5</td>
<td>0.0077</td>
</tr>
<tr>
<td>tRNA-synt_1c</td>
<td>12.6</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>TMF_DNA_bd</td>
<td>12.1</td>
<td>0.0004</td>
</tr>
<tr>
<td>UDPGT</td>
<td>12.1</td>
<td>0.0004</td>
</tr>
<tr>
<td>Cupin_1</td>
<td>11.9</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>Cupin_2</td>
<td>11.1</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ABC2_membrane</td>
<td>10.9</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>GrpE</td>
<td>10.1</td>
<td>0.0011</td>
</tr>
<tr>
<td>EFG_C</td>
<td>9.0</td>
<td>0.0002</td>
</tr>
<tr>
<td>GTP_EFTU_D3</td>
<td>8.8</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>ABC_ATPase</td>
<td>8.7</td>
<td>&lt;0.0001</td>
</tr>
<tr>
<td>IF-2B</td>
<td>8.7</td>
<td>0.0025</td>
</tr>
</tbody>
</table>

Table 3.6. Top enriched *Agrostis stolonifera* domains at 96 h post inoculation with *Sclerotinia homoeocarpa*.
### Table 3.6 continued

<table>
<thead>
<tr>
<th>Protein Name</th>
<th>Log2 Fold Change</th>
<th>p-value</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>RNA_pol_Rpb1_5</td>
<td>8.7</td>
<td>0.0025</td>
<td>RNA polymerase Rpb1, domain 5</td>
</tr>
<tr>
<td>Cytochrom_B_C</td>
<td>7.6</td>
<td>0.0049</td>
<td>Cytochrome b/b6, C-terminal</td>
</tr>
<tr>
<td>SMC_N</td>
<td>7.3</td>
<td>&lt;0.0001</td>
<td>Structural maintenance of chromosomes protein</td>
</tr>
<tr>
<td>ABC_membrane</td>
<td>6.7</td>
<td>&lt;0.0001</td>
<td>ABC transporter transmembrane region</td>
</tr>
<tr>
<td>ATP-synt_C</td>
<td>6.7</td>
<td>0.0086</td>
<td>ATP synthase subunit C</td>
</tr>
<tr>
<td>ATP-synt_ab_C</td>
<td>6.7</td>
<td>0.0012</td>
<td>ATP synthase alpha/beta chain</td>
</tr>
<tr>
<td>DUF1602</td>
<td>6.1</td>
<td>&lt;0.0001</td>
<td>Protein of unknown function</td>
</tr>
<tr>
<td>GTP_EFTU</td>
<td>5.8</td>
<td>&lt;0.0001</td>
<td>Elongation factor Tu GTP binding domain</td>
</tr>
<tr>
<td>ABC_tran</td>
<td>5.5</td>
<td>&lt;0.0001</td>
<td>ABC transporters</td>
</tr>
<tr>
<td>Terpene_synth_C</td>
<td>5.4</td>
<td>0.0049</td>
<td>Terpene synthase family, metal binding domain</td>
</tr>
<tr>
<td>p450</td>
<td>5.0</td>
<td>&lt;0.0001</td>
<td>Oxidative degradation of compounds</td>
</tr>
<tr>
<td>GTP_EFTU_D2</td>
<td>5.0</td>
<td>0.0005</td>
<td>Elongation factor Tu domain 2</td>
</tr>
<tr>
<td>Type of SSR</td>
<td>Number</td>
<td>Percentage</td>
<td>Length (bp)</td>
</tr>
<tr>
<td>--------------------------</td>
<td>--------</td>
<td>------------</td>
<td>-------------</td>
</tr>
<tr>
<td><strong>Sclerotinia homoeocarpa</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mononucleotide repeat</td>
<td>266</td>
<td>37.2%</td>
<td>10 – 23</td>
</tr>
<tr>
<td>Dinucleotide repeat</td>
<td>100</td>
<td>14.0%</td>
<td>12 – 42</td>
</tr>
<tr>
<td>Trinucleotide repeat</td>
<td>221</td>
<td>30.9%</td>
<td>15 – 84</td>
</tr>
<tr>
<td>Tetrtnucleotide repeat</td>
<td>56</td>
<td>7.8%</td>
<td>20 – 72</td>
</tr>
<tr>
<td>Pentanucleotide repeat</td>
<td>25</td>
<td>3.5%</td>
<td>25 – 45</td>
</tr>
<tr>
<td>Hexanucleotide repeat</td>
<td>24</td>
<td>3.4%</td>
<td>30 – 54</td>
</tr>
<tr>
<td>Complex</td>
<td>23</td>
<td>3.2%</td>
<td>21 – 112</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>715</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Agrostis stolonifera</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Mononucleotide repeat</td>
<td>186</td>
<td>47.7%</td>
<td>10 – 14</td>
</tr>
<tr>
<td>Dinucleotide repeat</td>
<td>66</td>
<td>16.9%</td>
<td>12 – 52</td>
</tr>
<tr>
<td>Trinucleotide repeat</td>
<td>106</td>
<td>27.2%</td>
<td>15 – 36</td>
</tr>
<tr>
<td>Tetrtnucleotide repeat</td>
<td>12</td>
<td>3.1%</td>
<td>24 – 28</td>
</tr>
<tr>
<td>Pentanucleotide repeat</td>
<td>0</td>
<td>0.0%</td>
<td>n/a</td>
</tr>
<tr>
<td>Hexanucleotide repeat</td>
<td>2</td>
<td>0.5%</td>
<td>48 – 54</td>
</tr>
<tr>
<td>Complex</td>
<td>18</td>
<td>4.6%</td>
<td>21 – 157</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>390</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3.7. Simple Sequence Repeat (SSR) markers from the *Sclerotinia homoeocarpa* and *Agrostis stolonifera* transcriptome.
Figure 3.1. Read length of 454 sequencing data. The range of sequence read length (bp) of 454 data is shown as the shaded area. The length of the SBS reads is 16 bp indicated with the arrow.
Figure 3.2. Signal tracks of SBS reads for two genes (SH3549 and SH8608) from either the PDA, PDB, or interaction library.
Figure 3.3. Venn diagram of RNA-Seq reads unique and common to the SH library, AS library, and Interaction library. A: Reads unique to the Inoculation library, B: Reads unique to the SH only library, C: Reads unique to the AS only library, D: Reads common to the Inoculation and SH library, E: Reads common to the Inoculation and AS library, F: Reads common to the SH and AS libraries, G: Reads common to the Inoculation, SH, and AS libraries.
Figure 3.4. Top Species Blast hits: (a) SH transcript library (b) AS transcript library.
Figure 3.5. GO terms associated with up-regulated SH transcripts. (a) Molecular Function, (b) Biological Process, (c) Cellular Component.
Figure 3.6. GO terms associated with up-regulated AS transcripts. (a) Molecular Function, (b) Biological Process, (c) Cellular Component.
Figure 3.7. Comparison of semi-quantitative RT-PCR data to RPKM values. Creeping bentgrass was infected with *Sclerotinia homoeocarpa* from a PDA culture or sprayed on as a mycelia homogenate in PDB.
Figure 3.8. Validation of RPKM data for selected SH and AS transcripts using relative expression real time PCR.
Chapter 4: Whole genome sequencing of seven *Magnaporthe oryzae* field isolates reveals genome content variation and predicted pathogenicity determinants
Introduction

Rice has been served as major food source worldwide, especially in Asia and Africa, however, a large portion of yields are lost through disease annually. Rice blast, caused by fungal pathogen \textit{Magnaporthe oryzae}, is one of the most severe rice diseases and is found almost everywhere rice is grown. Severe disease outbreaks can destroy up to 90\% of rice yields for an entire field, region, or country causing a dramatic impact on human welfare and regional economies [1-3]. Condium of \textit{Magnaporthe oryzae} are spread by rain or water flush, and accomplish the infection by penetrating into rice leaves using a special structure called an appressorium. Mycelia then explore inside host tissue and eventually cause cell death [4, 5]. The most effective method to control of rice blast comes from the introduction of resistance (R) gene into elite rice cultivars, which will recognize the corresponding avirulence (Avr) genes and trigger “Hyper sensitive response” [6]. However, this resistance is often broken within a few years [7, 8], mostly due to the mutation or function loss of the Avr genes under strong selection force.

\textit{M. oryzae} and rice serve as model organisms for studying host-pathogen interaction since both genomes are publically available [9]. This pathosystem generally controlled by the gene-for-gene model [10-13], which was proposed by Flor in 1971 [14]. While many R genes in rice have been identified and cloned, the identification of Avr genes in the fungus has been limited. There are three Avr genes involved in host species specificity: PWL1 [15], PWL2 [16], and Avr-CO39 [17], and at least six Avr genes being involved in rice cultivar specificity, including Avr-Pita [18], ACE1 [19], AvrPiz-t
While R proteins are reported to contain some conserved functional domains - such as NBS-LRR domains, no obvious sequence or structural patterns can be identified for Avr genes, which make their identification difficult. However, “genome-wide association analysis” (GWAS) can help in identifying novel Avr genes based on the association between Avr genes and cultivar specific virulence.

In one study focusing on the Avr-Pita family that includes AvrPita1, AvrPita2, and AvrPita3, it was found that AvrPita1 and AvrPita2 were associated with transposon elements. These located randomly in different chromosomes in various field isolates, which may be a result of high selection imposed by the corresponding R gene [23]. This dramatic location variation, as well as high diversifying selection, makes it almost impossible to identify avirulence genes without genome sequencing and association analysis.

The reference strain of *M. oryzae* is 70-15, which was also used as the first rice blast isolate sequenced in 2002 [9]. This strain was generated from a cross between a rice pathotype and a weeping love grass (*Eragrostis curvula*) pathotype which was then backcrossed with rice isolate Guy11 from French Guyana [24, 25]. The resulting 70-15 isolate has been used as reference strain in many laboratories and used for many fungal genome comparison studies. It was reported that in 70-15, some weeping love grass pathogen sequences were retained.

With the fast development of next-generation sequencing technology, genome re-sequencing and comparative studies have been reported on multiple fungal phytopathogens including: *Alternaria*
[26], *Fusarium* [27], *Aspergillus* [28], and *Utilagus* [29] et al. Some of common findings from these studies include high level of variation among the genomes of different isolates and unique genome regions that contain isolate specific virulence factor genes.

In *M. oryzae*, several genome comparison projects have been initiated recently. The first identification of novel Avr genes via genome comparison was reported in 2009 [22]. A research group for that report first tried to screen the secreted proteins in 70-15 in different field isolates and associate polymorphisms with pathogenicity. However, no orthologous genes were found. Then they sequenced field isolate “Ina168” which is known to contain 9 Avr genes, and identified a 1.68Mb region unique to 70-15. Through an association study for genes located on this specific region, three novo Avr genes were identified. The project demonstrated the value of 70-15 reference genome, which can be used as an Avr gene minimum reference for comparison to field isolates.

Recently, another genome comparison between two field isolates of *M. oryzae* has been performed with 70-15 included as a reference [30]. The comparison discovered many isolate-specific genomic regions. For those shared genes, many predicted to be involved in host-pathogen interactions were expanded and found to be under diversifying selection. This study also identified a large number of transposon-like elements, with less than 30% of them conserved among all three isolates, suggesting recent active transposition events in this pathogen.

While the genome content of *M. oryzae* has been characterized through multiple sequencing
projects, other important genome-wide transcript profiling techniques have been investigated. In 2012, the transcriptome libraries of appressorium development have been sequenced using a NGS technique called “SuperSAGE”. By analyzing the expression level change during appressorium formation, Pmk1 MAP kinase was revealed as a key global regulator in appressorium formation and related process [31]. Other noticeable examples included a ChIP-Seq study to examine the regulation network of a calcium signal transcription factor MoCrz1 [32]. These successfully high throughput studies on Magnaporthe, as well as other organisms, demonstrated the power of these approaches, and built base of applying NGS studies to studying pathogen-host interactions.

This study investigated “isolate-specific” genome content. Major genome content differences in isolates of M. oryzae have been reported. The first report was in Ina168 by comparison to the reference 70-15 genome [22]. Differences were also identified easily through CHEF gels in an AVR-Pita study [23]. Both studies confirmed that Magnaporthe oryzae contains supernumerary chromosomes that vary in size and AVR genes may locate on them. Supernumerary chromosomes are sometimes called “conditionally dispensable chromosome”. They were firstly discovered and achieved notoriety in fungi in the genus Alternaria [33, 34]. Subsequently, through genome sequencing studies, similar chromosomes were reported in other fungi included Fusarium oxysporum [27], Nectria haematococca [35, 36], Mycosphaerella graminicola [37], Cochliobolus heterostrophus [38], Leptosphaeria maculans [39].

In this study, we performed whole genome sequencing on seven Magnaporthe oryzae field isolates
CHL1, CHL2, CHL42, CHL43, CHL346, CHL359, and CHL381, and then compare the de novo assembly
and aligned to the reference sequence of 70-15. Through this analysis we attempted to answer the
following questions: 1) Do isolate-specific genes or genome regions exist between the seven field
isolates? What are the functions of unique genomic regions and can we determine where were they
originated? 2) How many variations can be found between each isolate and 70-15? Do
SNPs/Indels/CNVs (copy number variation) serve as a major mutation sources and how many genes
do they affect? 3) How many secreted proteins are under positive selection force and showed
polymorphisms that suggest a function as potential virulence effectors?

Results and Discussions

Genome sequencing and assembly

Genomes of seven M. oryzae isolates strains CHL1, CHL2, CHL42, CHL43, CHL346, CHL359, and
CHL381 were sequenced by a Illumina Genome Analyzer II, a next-generation sequencing approach.
The resulting 75bp paired-end short reads were assembled by Velvet [40] in a hybrid manner. De
novo assembly was followed by aligning to the 70-15 finished sequence Version 8 as reference to
build a scaffold. It was observed that the genome size of the 7 field isolates showed significant
variation ranging from 37.2MB of CHL43 to 39.3MB of CHL346 (Table 4.1). The N50 of CHL42 and
CHL43 showed a significantly lower level compared to other 5 isolates, which may be caused by a
higher percentage of repeat sequences that made \textit{de novo} assembly more difficult. All field isolates have average 10.5K predicted ORFs. Gene density and average CDS length also showed a similar number as the 70-15 genome. The number of predicted gene models was about 2000 less than 70-15 on average\cite{9}, and also less than the number in another two publically available sequenced field isolates P131 and Y34 \cite{30}, which can be explained by the different gene prediction tools used. FGENESH \cite{41} was used for the prediction of P131 and Y34, and it was reported that the predictor tended to produce a significant number of false positive predictions. Another possible reason for our reduced predicted genes number was the lower quality of the genome assembly where many paralogous genes may have assembled into one location in the assembly.

\textbf{Genome content comparison}

Genome assembles comparison among all the 7 sequenced isolates and the reference 70-15 showed that each of sequenced genome could be aligned with each other with identity of at least 82.4%. There’s much more noise in the dot plot of comparison between 70-15 and field isolates, compared to the alignment between field isolates themselves, suggesting that more diversifying sequences exist between 70-15 and field isolates. While each strain contains unique genome content, strains collected from same geographical location also shared unique content within the group, but not in other geographic locations (Figure 4.1).
**Phylogenetic tree construction**

The phylogenetic relationships of the 7 sequenced field strains were determined using the genetic distances calculated from the whole genome SNPs, which contain about 10K SNP locations. The resulting neighbor-joining tree showed three divergent groups generally segregating according to their geographical origins, with field isolates CHL1 and CHL2 grouping more closely to the reference 70-15. (Figure 4.2)

It was surprising to find the close relation between the reference 70-15 and the CHL1/CHL2 group as 70-15 was generated from rice pathotype in France and CHL1/CHL2 were collected from Barley in Thailand. The possible explanation could be that 70-15 contains some genome content from love grass pathogen, while CHL1 and CHL2 are barley pathogens, thus they could share some sequences.

**Unique genes / Presence-absence variation**

Among the ~10K predicted ORFs in each of the 7 sequenced strains and 70-15, unique genes were counted. Since the 7 filed isolates belong to 3 phylogenetic and geographical groups, to simplify the Venn diagram, CHL1, CHL42, CHL346, and 70-15 were chosen to represent all 4 groups. As can be seen in Figure 4.3, most genes (84.7%) fell into “core genes”, and 5% were shared among all field isolate but not 70-15. An interesting finding was that 2,914 genes were found unique in 70-15 but not in any of the field isolates. As mentioned above, the large number of unique 70-15 genes is most
likely the result of the extra predicted genes compared to the field isolates. Examination of these 2,914 genes found that most of them were short fragmented genes with the average length of only 618 bp, compared to the average length of all the 70-15 genes at 1,317bp. However, despite the fact the 70-15 unique genes may be generated by false positive gene predictions, it’s still worth considering their functions. Another important fact is the field isolates specific genes, where 1,525 such genes were not found in 70-15. In a previous study aiming to identify Avr genes, 330 candidate genes were found only in the field isolate “Ina168” with 3 of them being confirmed as Avr genes [22].

**Copy number variation analysis**

Copy number variation (CNV) is now regarded as having significant effects on gene function, which in this case was calculated using sequenced reads coverage along the genome. The coverage was counted based on 70-15 as the reference. CNV was calculated for each sequenced strain individually and results are showed in Figure 4.4. It can be observed that 1) copy number variation does exist along the genome especially at the ends of chromosomes; 2) All field isolates shared a similar CNV distribution; 3) there are some unique CNV positions that only exist in some of the genomes but not the others. We also calculated the read coverage of all the predicted ORFs using 70-15 as the reference genome (Figure 4.5). When five-fold was used as a cut-off, typically a field isolate contained about 300 genes showing an expansion of gene copy numbers and 200 genes showing
reduced gene copy number. Table 4.2 shows the top 8 70-15 genes, which were under obvious CNVs. MGG_15335 in CHL346 had a copy number of 3, while this gene in other isolates showed no copy number difference.

**SNP and Indel polymorphisms**

SNP profiles were generated by mapping short reads to the reference 70-15 genome sequence, with one mismatch allowed, using SOAPsnp [42] to call SNPs. IDPs were identified by short reads mapped allowing gaps. As showed in Table 4.3, all isolates had about 10,000 to 15,000 SNPs and 3,000-10,000 Indels with CHL346, CHL359, and CHL381 showing the most SNPs and Indels as compared to 70-15. Among the SNPs, we identified those that have a potentially disabling effect on gene function as showed in Table 4.3. In total 355 SNPs were expected to introduce premature stop codons, 752 were expected to alter initiation methionine residues, and 132 were expected to disrupt splicing donor or acceptor sites. Large effect IDPs were identified as those cause frame shifts. There were 47,083 IDPs identified for all strains and 910 of them were large effect IDPs. Genes within large effect SNPs and IDPs in their coding regions were regarded as losing the original functions, and thus could be a good candidate for virulence determinates.

It was reported in previous studies that SNPs and IDPs were not distributed evenly along the genome, but in fact they were found enriched in localized regions and thus make sequences in those regions
highly variable. Here we calculated the density of SNPs and IDPs in the genomes to identify enrichment regions and compare them among isolates. Figure 4.6 shows the SNP density along genome. The result is that 1) SNP hot spots do exist and majority of them locate at the telomere regions; 2) Different isolates have unique hot spots but also share common locations where the SNPs and IDPs were enriched.

**Selection force**

Ka/Ks ratios reflect the selection force of either a genomic region or a specific gene with the assumption that Ka/Ks ratio=1 means it is under neutral selection and a higher ratio means it’s under positive selection. Similar as SNP enriched regions, Ka/Ks ratio hot spots were also reported previously. As showed in Figure 4.7 as sharp peaks, most selection force hot spots were common in all sequenced genomes, however some unique peaks were observed in one or several genomes, indicating events of isolate-specific selection enriched regions.

Compared to SNP enriched regions, the Ka/Ks ratio peaks were less in number but much sharper, indicating some regions were under strong diversifying selection. It should be noted that more than half of such peaks occur at the telomere regions, which is consistent with previous reports that the edge of chromosomes and subtelomere regions are unstable.
Genes under positive selection

Evolution occurs when fungal pathogens interact with host and adapt to environmental change. We are interested in detecting genes involved in this adaptation. We found several genes containing a Ka/Ks ratio > 1 and thus under positive selection associated with compelling functional annotations.

MGG_15380 encodes a salicylate hydroxylase. Endogenesis of salicylic acid has been reported to be required for systemic acquired resistance in plants [43]. Salicylate hydroxylase from pathogen hydroxylates salicylic acid, convert to catechol, which suppresses systemic acquired resistance in plants and favors infection of pathogens. The salicylate hydroxylase locus shows positive selection during co-evolution with rice plants harboring salicylic acid, which indicates a correlation between virulence and activity of salicylate hydroxylase in *Magnaporthe*.

MGG_17750 contains a SET structural domain. SET proteins methylate lysine residues on histone tails [44]. A previous study showed that SET gene families were involved in epigenetic modification and relevant to growth and development of multi-cellular organisms. SET family members SET1-6, SET9, SET-JmJc and SET-TPR genes have been reported in fungi. The analysis of fungal SET gene family phyletic evolution shows that SET-domain proteins were highly conserved in plants, fungi and mammals. Interestingly, SET-domain proteins differentiate species-specific lineages. JmJc-SET, SET-TPR and SET5/6 are specific to filamentous fungi and their homologous genes in rice blast isolates (MGG_09180, MGG_08578, MGG_09389, MGG_06798, MGG_04672) are classified in the same cluster during systemic phyletic evolution analysis, while there is no homologous genes in
*Neurospora crassa* and other non-pathogenic fungi [45]. This fact suggests that SET gene families may be involved in pathogenicity in rice blast isolates. Yeast SET3 orthologous gene SET3 in *Magnaporthe* serves as important component of histone HDACs, which interacts with Tig1.

**Conclusion**

In this study, we sequenced the genome of 7 *Magnaporthe oryzae* field isolates using next generation sequencing technology and performed a reference assisted *de novo* genome assembly. We found the genome statistics of the sequenced isolates were similar to the reference 70-15, but less gene models were predicted, probably due to the different set of prediction tools that we used. One interesting finding was the close relationship between 70-15 and two barley pathotype *Magnaporthe oryzae*. Variations at the gene level in the form of SNPs, Indels, CNVs were identified, and it was surprising to find most of these polymorphisms were enriched in select of the genome regions, which we call “hot spots”, and that different isolates contain unique “hot spots”. The selection ratios in genes were also calculated. Thus, those genes unique to 70-15 located in the “hot spots” and containing high Ka/Ks ratio we predict to be good candidates for novel virulence related and Avr genes.
Materials and Methods

Strains

All sequenced strains were provided by Dr. Qinghua Pan from China South Agricultural University. CHL1 and CHL2 were collected from Barley in Thailand; CHL42 and CHL43 were collected from rice in Yunnan Province, China; CHL346, CHL359, and CHL381 were collected from rice in Jiangsu Province, China.

Sequencing

Genome DNA was extracted from the plate cultivation following an in-house DNA extraction protocol. Whole genome sequencing library was generated with Illumina kit with an insert length around 300bp. Sequencing was performed using Illumina Genome Analyzer II at MCIC, the Ohio State University. 78pb paired-end reads were generated with coverage of about 100X for each isolate.

Assembly

Since we have both a polished reference genome 70-15 and a deep sequenced short reads library, a hybrid assembly method: “velvet with Columbus module”, [40] which firstly map reads to reference genome to guide the assembly and then perform a de novo assembly. Assembly parameters
including k-mer were optimized by test running. In the assembly we tried to avoid misguidance by the reference so we compared sequences of this hybrid assembly to a pure de novo assembly and found most sequences were identical between the two methods except hybrid assembly generated longer contigs. Gene prediction was performed using Augustus [46] with pre-trained “Magnaporthe grisea” matrix.

**Genome content comparison**

Genome of different isolates were compared with each other by performing a genome-wide sequence alignment using “Nucmer” tool in “MUMmer” package [47]. The alignment identity cut-off was set as 80%, and then the alignment results were visualized as plot dots pictures using “mummerplot” tools in the package.

**Unique gene identification**

To identify the common and unique genes, whole genome gene set of CHL1, CHL42, CHL346, and 70-15 were firstly input into orthologous proteins were acquired by “Inparanoid” [48] to perform a pair-wised alignment. Then the result was processed by “Quickparanoid” (http://pl.postech.ac.kr/QuickParanoid/) to generate gene ortholog cluster with default parameter. A Venn diagram was then drawn based on this ortholog cluster.
**Copy number variation (CNV) analysis**

Short sequencing reads from individual isolates were mapped to reference 70-15 genome using SOAPaligner2 [49] with only 1 mismatch allowed. Then the reads coverage depth of genome was scanned in a 1KB window to search for CNV. We also checked the average coverage depth of individual predicted genes using Cufflinks [50]. Then standard deviation and coefficient of variation were calculated for each gene in order to pick genes showing strong CNV.

**SNPs & IDPs analysis**

To identify the SNPs, short reads were mapped to reference 70-15 genome by SOAPaligner2 [49]. After sorting alignment file, SOAPsnp was used to identify SNP locations with default parameters. Then the SNP identification file was converted into VCF format and input into snpEff [51] together with gene prediction to predict the SNP effect. SNP density was calculated along genome in a 10KB window and was visualized to show the existence of SNP hot spots. IDPs were identified in a similar way, but pipeline including SAMtools for alignment and BCFtools for Indel identification were used. IDP effects were also predicted by snpEff.
Phylogenetic tree construction

All the nucleotide bases (about 10KB) in SNP locations were extracted and form an artificial sequence for each of the 7 isolates and 70-15 genome. Then these 10KB sequences which contain all the SNP information were input into MEGA5 [52] to generate a phylogenetic tree using maximum-likelihood algorithm.

Selection force calculation

From the SNP analysis, every SNP in coding region was counted for either synonymous mutation of non-synonymous mutation. And the ratio between number of non-synonymous and synonymous mutation was calculated in every 10KB region window. Then a genome wide Ka/Ks ratio picture was drawn based on this calculation.
References


45. Veerappan CS, Avarama Z, Moriyama EN: Evolution of SET-domain protein families in the


<table>
<thead>
<tr>
<th>Isolates</th>
<th>Genome size (Mb)</th>
<th>Contig number</th>
<th>N50 (Mb)</th>
<th>GC%</th>
<th>predicted ORFs</th>
<th>Gene density (Kb/per gene)</th>
<th>Average CDS length (Kb)</th>
</tr>
</thead>
<tbody>
<tr>
<td>CHL1</td>
<td>38.3</td>
<td>4227</td>
<td>116.7</td>
<td>51.56</td>
<td>10666</td>
<td>3.59</td>
<td>1524</td>
</tr>
<tr>
<td>CHL2</td>
<td>36.8</td>
<td>2990</td>
<td>221.2</td>
<td>51.88</td>
<td>10286</td>
<td>3.58</td>
<td>1543</td>
</tr>
<tr>
<td>CHL42</td>
<td>38.0</td>
<td>4282</td>
<td>50.9</td>
<td>51.57</td>
<td>10674</td>
<td>3.56</td>
<td>1531</td>
</tr>
<tr>
<td>CHL43</td>
<td>37.2</td>
<td>4254</td>
<td>52.3</td>
<td>51.61</td>
<td>10457</td>
<td>3.56</td>
<td>1513</td>
</tr>
<tr>
<td>CHL346</td>
<td>39.3</td>
<td>7523</td>
<td>151.9</td>
<td>51.00</td>
<td>10352</td>
<td>3.80</td>
<td>1523</td>
</tr>
<tr>
<td>CHL359</td>
<td>38.5</td>
<td>5870</td>
<td>175.2</td>
<td>51.70</td>
<td>10447</td>
<td>3.69</td>
<td>1536</td>
</tr>
<tr>
<td>CHL381</td>
<td>38.4</td>
<td>5619</td>
<td>175.2</td>
<td>51.73</td>
<td>10473</td>
<td>3.67</td>
<td>1536</td>
</tr>
</tbody>
</table>

Table 4.1. General statistic numbers of the genome assembly and gene prediction of sequenced isolates.
Table 4.2. Top eight genes containing the copy number variation. Copy number of genes in each isolates were calculated using 70-15 as reference, standard Deviation (SD) and Coefficient of Variation (C.V.) were calculated for each genes and sorted by C.V.
<table>
<thead>
<tr>
<th>Strains</th>
<th>SNPs</th>
<th>Indels</th>
<th>Region Distribution of SNPs</th>
<th>Number of High Impact Alternation</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>Exon</td>
<td>Intron</td>
</tr>
<tr>
<td>CHL1</td>
<td>9397</td>
<td>4704</td>
<td>22.9%</td>
<td>5.8%</td>
</tr>
<tr>
<td>CHL2</td>
<td>13188</td>
<td>7020</td>
<td>18.4%</td>
<td>5.2%</td>
</tr>
<tr>
<td>CHL42</td>
<td>12643</td>
<td>5610</td>
<td>23.3%</td>
<td>5.3%</td>
</tr>
<tr>
<td>CHL43</td>
<td>10590</td>
<td>3259</td>
<td>21.4%</td>
<td>5.2%</td>
</tr>
<tr>
<td>CHL346</td>
<td>14787</td>
<td>8456</td>
<td>20.0%</td>
<td>5.1%</td>
</tr>
<tr>
<td>CHL359</td>
<td>13462</td>
<td>8757</td>
<td>21.3%</td>
<td>5.3%</td>
</tr>
<tr>
<td>CHL381</td>
<td>13752</td>
<td>9277</td>
<td>21.0%</td>
<td>5.1%</td>
</tr>
</tbody>
</table>

Table 4.3. Number of SNPs & IDPs identified, their location, and the number of those causing large impact on gene function. (change of start codon, stop codon, splicing place frame shift)
Figure 4.1. Dot plots of the alignments among 7 sequenced isolates and reference 70-15 sequence. Percentage of genome region that could be successfully aligned were noted under each dot plots. Much more noises dots (non-synteny regions) could be observed in the alignments between 70-15 and field isolates.
Figure 4.2. Phylogenetic tree of different isolates constructed using genome-wise SNP data. This tree was calculated and drawn by software “MEGA” with maximum-likelihood algorithm. It should be noted that Reference 70-15 has relatively close relationship with CHL1/CHL2 group, even closer compared to other filed isolates.
Figure 4.3. Gene pool analysis among the four isolate(s) groups. Number of gene families was marked in every individual region. It should be noted that majority of gene families were shared by all isolates, and 70-15 has about 3000 genes unique to all the field isolates.
Figure 4.4. Copy number variation of the 7 sequenced isolates along genome. The base line represents no CNV in the region, and Peaks show the increased copy number enriched in those regions.
Figure 4.5. Number of Genes with higher/lower copy number compared to 70-15. Red bar represents the number of genes showing a 5 folds higher copy number compared to the gene in 70-15, blue bar represents the genes with reduced copy number.
Figure 4.6. SNPs & IDPs density (number per 100KB) along genomes of 7 sequenced isolates. Peaks represent enrichments of the SNPs and IDPs in those regions.
Figure 4.7. Ka/Ks ratio distribution along genome in the 7 sequenced isolates. Peaks represent regions under strong diversifying selections.
Chapter 5: A Distribution Pattern Assisted Method of Transcription Factor Binding Site Discovery

for Both Yeast and Filamentous Fungi
Introduction

Transcription factors (TFs) are proteins containing one or more DNA-binding domains, which bind to specific DNA sequences to activate or repress the recruitment of RNA polymerase, thereby up or down regulating transcription of downstream genes [1, 2]. In fungi, TFs bind to either enhancer or promoter regions, usually 1-1500bp upstream to the ORFs they regulate [3-5].

Understanding networks of transcriptional regulation is one of the most challenging yet important tasks in genome analysis. Transcription factors function by binding to recognition sites in gene regulatory regions, which are generally degenerate motifs of 5−15 base pairs [6]. Extensive research has focused on identifying transcription factor binding sites (TFBSs) by biological validation. Nevertheless, experiments identifying TFBS are usually time-consuming and laborious, leaving the binding sites of most transcription factors unclear [7]. Therefore, prediction of potential TFBSs using bioinformatics approaches has become an essential tool to explore gene structure and regulation.

Different approaches have been tried and applied to discover novel TFBSs over years, which could be put into two categories. Prediction tools in the first category search for the over-represented motifs in a given set of sequences - usually promoter regions of co-regulated genes or ChIP-chip/ChIP-Seq identified binding regions. This strategy has been used widely since it does not require additional information other than the sequences. In the searching process, TFBSs could be treated as either a Position Weight Matrix as in MEME [8] and GLAM [9], or consensuses as in SMILE [10] or Weeder [11], when heuristic approach was applied and exhaustive enumeration was avoided. Top candidates
after scoring and sorting are then represented in the form of a position frequency matrix (PFM) [12], or a position weight matrix (PWM). However, application of this approach was limited by the computational speed and scale of input data, thus could hardly be applied on the thousands of sequences generated by ChIP-Seq. As a result, new tools such as DREME [13] and WordSeeker [14] were designed specifically for large input dataset, which utilize new mapping algorithms and multi-nodes computation platform to significantly increase speed. Another way to search TFBSs focuses on biological information such as evolutionary conservation. This strategy is based on the assumption that the conserved non-coding regions across related species are likely to be under negative selection force and thus contain functional motifs. Due to the fast increase of available genome sequences, this method has been developed quickly in recent years, and generate lower rate of false-positive compared to other methods. [15, 16]. To date, various software have been developed to analyze possible binding motifs, however, the multifaceted biochemical interactions between proteins and DNA may easily lead to false-positive results and make theoretical predictions of TFBSs error prone [17].

Yeast transcription factors were used in this study as training and testing data sets. *Saccharomyces cerevisiae* is the most widely used and most well studied yeast species, whose genome serves as one of the most thoroughly analyzed to date [18]. Recently, multiple yeast databases have been built and published, including two transcription factor databases used in this study [19, 20].

To expand this method to filamentous fungi, the MoCRZ1 transcription factor from *Magnaporthe*
*oryzae*, a fungal pathogen that cause severe rice blast disease worldwide, was included [21]. MoCRZ1 is a C2H2 zinc-finger type transcription factor that is activated by calcineurin dephosphorylation and functions as a mediator of calcineurin signaling [22]. In a project conducted by Dr. Soonok Kim – a senior researcher in Dr. Thomas Mitchell’s lab, binding sequences of MoCRZ1 have been identified by applying both ChIP-chip to identify the binding region of this transcription factor and microarray approaches to identify the regulated genes and their promoters [23] (Figure 5.1), which was then used to test different binding motif prediction approaches.

In this study, we developed a strategy to improve the predictions made by MEME, using the transcription factor binding motif distribution pattern (MDP) information. This MDP approach was then tested with data from both yeast and *Magnaporthe* transcription factors binding sites and predicted motifs. Although similar spatial distribution of yeast TFBSs has been reported [24], we focused on TFBS prediction improvement by using a novel MDP approach.

**Results and Discussions**

**The general distribution frequency curve**

As showed in the analysis pipeline (Figure 5.2), we checked all the 113 TFs with their 301 documented binding motifs from the yeast transcription factor databases “YEASTRACT” [19]. Then in the filtering step 32 TFs which have more than 50 documented regulated genes were chosen. Some
TFs had multiple binding motifs but similar to each other, so we referred to the positional-weight matrix provided by the database “JASPAR” [20] and merged those similar motifs to generate a consensus motif sequence (details in methods). In total, 32 TFs with 63 target motifs were included in the analysis dataset (Table 5.1). 25 TFs with 37 validated binding motifs were randomly chosen to build the training dataset (Table 5.2), and the other 7 TFs formed the testing dataset. For each motif, their occurrence locations in the validated regulated gene models were scanned in 1000bp upstream of transcription start site (TSS). Next, a general distribution curve was drawn from the average distribution of all motifs (Figure 5.3). It was observed from the curve that the lowest frequency was 2.3% at a region -50bp to 0bp, while the highest frequency was 9.2% at a region -200bp to -150bp. A peak was observed from the region -275bp to -100bp, with center at about -200bp. This distribution pattern was similar to those from previous studies in yeast [24, 25] and human [26].

**Estimating the reliability of motif prediction**

It was assumed from the pattern of the distribution curve, that majority of transcription factors worked as short distance cis-elements binding specifically at -300 to -100bp - a region referred to as the “PR” (peak region), but not at the other two regions: -150 to 0bp regarded as the “NBR” (non-binding region) where most transcription initiation complexes bind; and -1000 to -250bp regarded as the “DR” (distal region). To quantify pattern fitness, a DP (distribution pattern) value was introduced to estimate fitness of any TF binding motif to the general frequency curve. The DP value
was generated from shape of the general distribution curve and an assumption: a “true” binding motif should occur more often in the PR, but not NBR or DR, while a random over-represented motif sequences may not have any specific distribution preference in the 1KB upstream region. So if you compare the average occurrence/frequency of PR and that of NBR and DR, the former should be higher in a “true” binding motif and a larger difference represents a more reliable prediction. We thus proposed a formula to calculate the DP value which was expected to be close to zero in a random motif. A motif with DP value zero or negative then has lower possibility to be the biologically “true” TF binding motif.

The DP value was calculated as the following formula:

\[
\text{DP value} = (\text{Average PR Occurrence} \times 2) - (\text{Average NBR Occurrence}) - (\text{Average DR Occurrence})
\]

**Testing on yeast**

Seven TFs were chosen to be involved to test if the utilization of DP value could assist in improving TFBS predictions. Four public motif finding tools were included in the performance comparison: MEME (published in 1994), MDscan (published in 2002), WordSeeker (published in 2010), and DREME (published in 2011). These four tools all search for statistically over-represented motifs in a given sequences set. MEME uses the expectation maximization to fit a two-component finite mixture model to the input sequences, and multiple motifs are found by probabilistically erasing the
occurrences of the top motif and then repeating the process [8]. MDscan combines the advantages of two motif search strategies: position-specific weight matrix updating and word enumeration to enhance the success rate [27]. DREME [13] and WordSeeker [14] were developed in recent years and specifically designed to process large size of ChIP-chip/ChIP-Seq datasets on scalable analysis platforms.

For each TF, the 1000bp upstream sequences of their documented regulated genes were firstly selected as input into MEME for consensus motif search with the target motif length parameter set from 5bp to 9bp. The top ten consensus motifs in the results were then processed to calculate their OR and DP value.

The OR (over-represent) value, or the observed:expected frequency ratios (O/E) described and utilized in a previous study [28] reflects the statistical over-represent of these consensus motifs. OR value was calculated as following formula, where O refers to the overall occurrence of a motif across the 1KB upstream sequences set, ‘ln’ is the natural logarithm, and Eo represents the expected occurrence of that motif:

$$\text{OR value} = O \times \ln \left( \frac{O}{E_o} \right)$$

Taking transcription factor “Fkh1” for example, the top ten consensus motifs from MEME were originally ranked by P-value which represents the possibility of obtaining this motif solely by chance. The first two motifs AAA[AG]A[AG]AAA and TT[TC][TC]TT[CT][CT], were likely to be simple sequence
repeats and thus being removed from the ranked results. The 2nd motif [GT]GTAAACAA and the 3rd motif [TCG]TTGTTTAC were reverse complimentary to each other and matched documented Fkh1 motifs [AG][CT]AAACA[AT][AT] [29] and [AG]TAAA[CT]AA [30]. The remaining 6 consensus motifs were not Fkh1 documented binding motifs (Figure 5.4, Table 5.3).

The MDP approach utilize both the over-representation information – measured by OR value, and the distribution pattern information – measured by DP value. Since both values represent the reliability of the motif prediction – higher value represents higher reliability - a Re-rank value was introduced to combine both values so the “true” motifs with over-represented occurrence and distribution pattern fitting the general curve will obtain higher value and thus be picked up from all the candidates. To even the contribution of the two values, we checked the OR and DP value in all the training dataset, and decide to amplify DP value by 1000 times so the average and standard deviation of both value is close to each other.

The Re-rank value was calculated as following formula:

\[
\text{Re-rank value} = \text{OR value} + \text{DP value} \times 1000
\]

After re-ranking based on the Re-rank value, the motif originally in the 1st place ranked by MEME showed a negative DP value and thus dropped to 5th in the MDP rank, since its distribution pattern showed little similarity to the general TF frequency curve, while the two documented Fkh1 target motifs were raised from 2nd/3rd to 1st /2nd. Since these two target motifs are reverse
complimentary to each other, we recorded the rank change as 2nd in MEME and 1st in MDP. 

Same as described in “Fkh1”, the upstream sequences of the 7 TFs were input into MDscan, WordSeeker and DREME, seeking for motifs with expected length around 8bp and other parameters were left as default. Out of 7 tested TFs, 2 motifs (Xbp1 and Gzf3) failed to find validated target binding motifs in the top 10 consensus motifs predicted by MEME or other predictors. Summary of the remaining 5 TFs were shown in table 5.4. Among the 5 target motifs, MDP predicted 4 motifs as the 1st and one motif as the 2nd in rank. While MEME and DREME only predicted two target motifs as the 1st in rank; MDscan and WordSeeker each failed to predict two target motifs in the top ten results. Some repeat-like sequences were noticed in MDscan and WordSeeker results, indicating their detections were somehow disturbed. Overall, MDP generated a better rank for the target motifs compared to other four tools.

**Prediction of Magnaporthe MoCRZ1 binding motifs using ChIP-chip data**

MoCRZ1, a transcription factor involved in Ca2+/Calcineurin signaling in *Magnaporthe oryzae*, was also used to estimate this MDP approach. The strategy for the test was to firstly predict binding motifs from ChIP-chip identified binding regions using traditional methods and then apply our new MDP approach without ChIP-chip data to see if a similar prediction could be reached. In the other word, we identify the “true” motifs first and then tested if the MDP approach could identify “true”
MoCRZ1 binding motifs without the need of ChIP-chip data.

To identify the binding regions and regulated promoters of MoCRZ1, both ChIP-chip and microarray experiments were conducted previously. Then the exact binding sequences of MoCRZ1 (50–1247 bp) revealed from ChIP-chip studies were retrieved from the promoters of genes in common between ChIP-chip and microarray expression studies and subjected to motif signature analyses (Figure 5.5A). Initially, 106 sequences from each of the 83 genes differentially regulated in the WT/Dmocrz1 comparison of the microarray data were analyzed with MEME [8] and MDScan [27]. There were more sequences than genes as 21 genes had 2 peaks in their promoter region and 2 genes had 3 peaks. Candidate motifs from both algorithms were manually interrogated and enumerated to identify the top 3 candidates, which were subsequently screened against randomly retrieved 106 intergenic sequences with an average length of 509 bp (Figure 5.5A). The most enriched motif of CAC[AT]GCC was identified in 33 sequences in front of 24 genes, a 16X enrichment in MoCRZ1 bound sequences. The most common motif of TTGNTTG was found in 68 promoter sequences in front of 42 genes with 4X enrichment. Motif TAC[AC]GTA occurred in 22 promoter sequences of 18 genes with 4X enrichment (Table 5.5). Fifty-six genes had at least one motif, while all three of these motifs occurred in front of 5 genes. These three candidate motifs were also screened against the Magnaporthe genome to check the occurrence positions. Table 5.6 showed that motif TAC[AC]GTA occurred 1900 times in ORF but 4727 times in promoter region, but CAC[AT]GCC showed a higher occurrence in ORF regions. These motifs were searched against yeast motif database using TOMTOM [31]. The top match for CAC[AT]GCC was MET28 (p-value = 0.0013), while the second match was
CRZ1 with significant p-value (0.0022), showing Crz1p of *S. cerevisiae* has this motif in its promoters although it was not previously identified as a calcineurin dependent response element (CDRE) (Figure 5.5B). Pbx1b (P-value = 0.00039) and Zec (P-value = 0.00045) were best two matches for TTGNTTG, while no significant match was returned for TAC[AC]GTA. Binding of MoCRZ1 to the promoter region was confirmed by Electrophoretic Mobility Shift Assay (EMSA) conducted by Dr. Soonok Kim. A 209 bp PCR fragment having 1 TTGNTTG and 2 CAC[AT]GCC motifs from the MoCRZ1 promoter region was bound to purified MoCRZ1 protein (Figure 5.5C, left panel). A 325 bp fragment of CBP1 (MGG_03218) having 1 TTGNTTG and 1 CAC[AT]GCC motifs was also shown to bind to purified MoCRZ1 (Figure 5.5C, right panel).

**Testing MDP approach to predict MoCRZ1 binding motifs**

Then we tested if the MDP approach could reach the same binding motif predictions with only the microarray but not the use of ChIP-chip data. From the microarray data[23], 190 genes were picked as predicted MoCRZ1 regulated genes as they all showed a 2 fold or greater expression change between the control and all three conditions including Ca^{2+} deficiency, MoCrz1 inhibitor added, and the MoCRZ1 deletion background. Results of the top ten consensus motifs predicted by MEME from 1000bp upstream of MoCRZ1 regulated genes were shown in figure 5.6 and Table 5.7 with their distribution pattern. Three target motifs (TTGNTTG, CAC[AT]GCC, TAC[AC]GTA) were originally ranked as 3rd, 7th, and 8th, with simple repeats at the 1st and 2nd in rank. After re-ranking, these three
target motifs went up to 4th, 6th, and 7th. It was observed that another two motifs originally ranked at 5th and 10th went to 3rd and 5th, showing a significant rank improvement. These two motifs were next searched in the TOMTOM [31] database and identified as “Rim101” binding motifs, which was reported as a TF involved in a pathway acting in parallel to Crz1 via calcineurin [32].

Conclusions

In this study, we developed the MDP approach to improve TFBS prediction. Genome-wide TFBS analysis is generally challenging with both experimental validation and computational analysis required to refine TFBS predictions. The use of TFBS distribution profiles improves the accuracy of predictions by estimating both the over-representative level of the candidate motifs and their general distribution pattern as well. The major originality here is that we are more focusing on improving TFBSs prediction by using distribution pattern.

Methods

To select transcription factors, all the 113 TFs from the yeast transcription factor database “YEASTRACT” were checked and the number of their documented regulated genes was counted. Those TFs having less than 50 documented regulated genes were filtered out. We noticed that some TFs recorded in “YEASTRACT” had multiple documented binding motifs, however, regular expression
sequences of some binding motifs from the same TF showed high similarity, and thus could be clustered into a single motif. In those cases, the PWM (position weight matrix) was checked from the “JASPAR” database. If the clustering was supported by the PWM, then the different motifs were merged into a new one. For example, TF “Aft1” had 4 recorded binding motifs in “YESTRACT” database: YRCACCCR, TGCACCC, GGCACCC, and TGCACCCA, while only one matrix is recorded in “JASPAR”. So in the clustering process, consensus binding motif sequences of Aft1 was generated as “TRCACCY”. After clustering, each of the target motifs was scanned in the 1kb upstream TSS region of all genes in the yeast genome to check for number of occurrence. Any motif with less than 100 or more than 2000 occurrence were removed. The remaining 32 TFs were randomly divided into two groups: 25 TFs in training group and 7 in testing group.

List of documented regulated genes of the 32 TFs was downloaded from “YEASTRACT” database. Their 1KB upstream sequences were extracted from the yeast genome sequence and were used to scan for the binding motif sequences. The 1KB upstream region was divided into twenty 50bp windows and motif occurrence in each window was counted. Then the average occurrence of each window was calculated and a general frequency curve was generated.

For each TF in the testing group, the 1KB upstream sequences of their regulated genes were input into MEME running on a local cluster, to search for consensus motifs with their expected length around 8bp, as well as into MDscan, WordSeeker, and DREME. Then the top ten consensus motifs reported from MEME were used as queries to define the MDP.
The two commonly used motif discovery programs, MEME and MDScan, were used to identify the MoCRZ1 binding motif. Input data consisted of the exact binding sequences retrieved from the promoters of 83 genes with differential expression in the WT/Dmocrz1 comparison. Candidate motifs from both algorithms were manually interrogated and enumerated to identify the 3 top candidates, and queried against the yeast motif database using TOMTOM [31]. Enrichment was calculated over the 106 background sequence set which was randomly retrieved from intergenic region of the whole genome. Consensus motif sequences were calculated using WebLogo server at http://weblogo.berkeley.edu.
References


<table>
<thead>
<tr>
<th>TF Name</th>
<th>TF Structure</th>
<th>Number of Regulated Genes</th>
<th>TF Name</th>
<th>TF Structure</th>
<th>Number of Regulated Genes</th>
</tr>
</thead>
<tbody>
<tr>
<td>Rgt1</td>
<td>Fungal Zn cluster</td>
<td>63</td>
<td>Xbp1</td>
<td>Rel</td>
<td>501</td>
</tr>
<tr>
<td>Azf1</td>
<td>Beta-beta-alpha zinc finger</td>
<td>127</td>
<td>Reb1</td>
<td>Myb</td>
<td>512</td>
</tr>
<tr>
<td>Rtg1</td>
<td>Helix-loop-helix (bHLH)</td>
<td>129</td>
<td>Pdr8</td>
<td>Fungal Zn cluster</td>
<td>547</td>
</tr>
<tr>
<td>Mot3</td>
<td>Beta-beta-alpha zinc finger</td>
<td>135</td>
<td>Hsf1</td>
<td>E2F</td>
<td>571</td>
</tr>
<tr>
<td>Gzf3</td>
<td>GATA</td>
<td>149</td>
<td>Tec1</td>
<td>Homeo</td>
<td>571</td>
</tr>
<tr>
<td>Rlm1</td>
<td>Beta-beta-alpha zinc finger</td>
<td>205</td>
<td>Swi4</td>
<td>Rel</td>
<td>614</td>
</tr>
<tr>
<td>Gis1</td>
<td>Beta-beta-alpha zinc finger</td>
<td>223</td>
<td>Gln3</td>
<td>GATA</td>
<td>668</td>
</tr>
<tr>
<td>Fkh1</td>
<td>Forkhead-associated (FHA)</td>
<td>241</td>
<td>Abf1</td>
<td>Helix-loop-helix (bHLH)</td>
<td>669</td>
</tr>
<tr>
<td>Fkh2</td>
<td>Forkhead-associated (FHA)</td>
<td>313</td>
<td>Msn</td>
<td>Beta-beta-alpha zinc</td>
<td>740</td>
</tr>
<tr>
<td>Cbf1</td>
<td>Helix-loop-helix (bHLH)</td>
<td>337</td>
<td>Aft1</td>
<td>No confident structure</td>
<td>1114</td>
</tr>
<tr>
<td>Pho4</td>
<td>Helix-loop-helix (bHLH)</td>
<td>379</td>
<td>Msn</td>
<td>Beta-beta-alpha zinc</td>
<td>1187</td>
</tr>
<tr>
<td>Nrg1</td>
<td>Beta-beta-alpha zinc finger</td>
<td>399</td>
<td>Gcn4</td>
<td>Leucine zipper</td>
<td>1260</td>
</tr>
<tr>
<td>Mcm</td>
<td>MADS</td>
<td>403</td>
<td>Met4</td>
<td>Leucine zipper</td>
<td>1260</td>
</tr>
<tr>
<td>Adr1</td>
<td>Beta-beta-alpha zinc finger</td>
<td>443</td>
<td>Yap1</td>
<td>Leucine zipper</td>
<td>1824</td>
</tr>
<tr>
<td>Leu3</td>
<td>Fungal Zn cluster</td>
<td>495</td>
<td>Ste1</td>
<td>Homeo</td>
<td>2142</td>
</tr>
<tr>
<td>Mbp</td>
<td>Rel</td>
<td>498</td>
<td>Yap3</td>
<td>Leucine zipper</td>
<td>59</td>
</tr>
</tbody>
</table>

Table 5.1. List of yeast TFs involved in training and testing.
<table>
<thead>
<tr>
<th>TF</th>
<th>Documented Motif</th>
<th>Motif Length</th>
<th>Number of Binding Sites</th>
<th>Peak Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abf1</td>
<td>RTCRYBN(4)ACG</td>
<td>13</td>
<td>212</td>
<td>175</td>
</tr>
<tr>
<td></td>
<td>RTCRN(6)ACGNR</td>
<td>15</td>
<td>156</td>
<td>175</td>
</tr>
<tr>
<td></td>
<td>TNRCGN(6)TGAT</td>
<td>16</td>
<td>61</td>
<td>175</td>
</tr>
<tr>
<td></td>
<td>TCN(7)ACG</td>
<td>12</td>
<td>507</td>
<td>175</td>
</tr>
<tr>
<td>Adr1</td>
<td>TTGGRG</td>
<td>6</td>
<td>146</td>
<td>275</td>
</tr>
<tr>
<td>Aft1</td>
<td>TRCACCY</td>
<td>7</td>
<td>138</td>
<td>125</td>
</tr>
<tr>
<td>Azf1</td>
<td>AAMRGHA</td>
<td>7</td>
<td>221</td>
<td>75</td>
</tr>
<tr>
<td>Fkh2</td>
<td>RYAAACAWW</td>
<td>9</td>
<td>85</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>RTAAAYAA</td>
<td>8</td>
<td>118</td>
<td>100</td>
</tr>
<tr>
<td>Gcn4</td>
<td>TGASTCAY</td>
<td>9</td>
<td>85</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>RRTGACTC</td>
<td>8</td>
<td>79</td>
<td>175</td>
</tr>
<tr>
<td></td>
<td>CACGTG</td>
<td>8</td>
<td>168</td>
<td>275</td>
</tr>
<tr>
<td>Gis1</td>
<td>AGGGGG</td>
<td>5</td>
<td>120</td>
<td>300</td>
</tr>
<tr>
<td>Gln3</td>
<td>GATWDG</td>
<td>6</td>
<td>618</td>
<td>Unidentifiable</td>
</tr>
<tr>
<td></td>
<td>TTTNNGAA</td>
<td>8</td>
<td>240</td>
<td>200</td>
</tr>
<tr>
<td></td>
<td>GAANNTTC</td>
<td>8</td>
<td>188</td>
<td>225</td>
</tr>
<tr>
<td>Mcm1</td>
<td>DCCYWWWVNRRG</td>
<td>11</td>
<td>201</td>
<td>225</td>
</tr>
<tr>
<td></td>
<td>CCYWWWVNRRG</td>
<td>10</td>
<td>234</td>
<td>200</td>
</tr>
<tr>
<td>Met4</td>
<td>TCACGTG</td>
<td>7</td>
<td>79</td>
<td>375</td>
</tr>
<tr>
<td>Mot3</td>
<td>AGGYA</td>
<td>5</td>
<td>203</td>
<td>325</td>
</tr>
<tr>
<td>Msn2</td>
<td>CCCCT</td>
<td>5</td>
<td>536</td>
<td>175</td>
</tr>
<tr>
<td>Msn4</td>
<td>CCCCT</td>
<td>5</td>
<td>336</td>
<td>275</td>
</tr>
<tr>
<td>Nrg1</td>
<td>CCCCT</td>
<td>5</td>
<td>167</td>
<td>325</td>
</tr>
<tr>
<td></td>
<td>CCCCT</td>
<td>5</td>
<td>165</td>
<td>325</td>
</tr>
<tr>
<td>Pdr8</td>
<td>TCCGHHGA</td>
<td>8</td>
<td>51</td>
<td>325</td>
</tr>
<tr>
<td>Pho4</td>
<td>CACGTK</td>
<td>6</td>
<td>182</td>
<td>300</td>
</tr>
<tr>
<td>Rgt1</td>
<td>WWNNTCCCK</td>
<td>8</td>
<td>105</td>
<td>350</td>
</tr>
<tr>
<td>Rlm1</td>
<td>TAWWWWTAGM</td>
<td>10</td>
<td>52</td>
<td>325</td>
</tr>
<tr>
<td>Rtg1</td>
<td>GTCAC</td>
<td>5</td>
<td>233</td>
<td>Unidentifiable</td>
</tr>
<tr>
<td>Ste12</td>
<td>TGAAACA</td>
<td>7</td>
<td>210</td>
<td>225</td>
</tr>
<tr>
<td>Swi4</td>
<td>CRCGAAW</td>
<td>7</td>
<td>150</td>
<td>300</td>
</tr>
<tr>
<td>Tec1</td>
<td>RMATTCYY</td>
<td>8</td>
<td>234</td>
<td>275</td>
</tr>
<tr>
<td></td>
<td>CATTCTT</td>
<td>7</td>
<td>89</td>
<td>Unidentifiable</td>
</tr>
<tr>
<td></td>
<td>CATTCT</td>
<td>6</td>
<td>217</td>
<td>Unidentifiable</td>
</tr>
<tr>
<td></td>
<td>CATTCC</td>
<td>6</td>
<td>137</td>
<td>Unidentifiable</td>
</tr>
<tr>
<td>Yap1</td>
<td>TTACGTAA</td>
<td>8</td>
<td>50</td>
<td>175</td>
</tr>
<tr>
<td>Yap3</td>
<td>TGACTCA</td>
<td>7</td>
<td>88</td>
<td>175</td>
</tr>
</tbody>
</table>

Table 5.2. TFs in training group.
<table>
<thead>
<tr>
<th>Candidate Motif</th>
<th>Annotation</th>
<th>Rank by MEME</th>
<th>OR Value</th>
<th>DP Value (x1000)</th>
<th>Re-rank Value</th>
<th>Rank by MDP</th>
<th>Re-rank Process</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAA[AG]A[AG]AAA</td>
<td>Simple repeat</td>
<td>removed</td>
<td>671.5</td>
<td>47.2</td>
<td>718.7</td>
<td>removed</td>
<td></td>
</tr>
<tr>
<td>TT[TC][TC]TT[TC]</td>
<td>Simple repeat</td>
<td>removed</td>
<td>255.8</td>
<td>60.3</td>
<td>316.1</td>
<td>removed</td>
<td></td>
</tr>
<tr>
<td>[TC]TG[TC]TG[TC]TG</td>
<td></td>
<td>1</td>
<td>115</td>
<td>-37.0</td>
<td>78.0</td>
<td>5</td>
<td>2</td>
</tr>
<tr>
<td>[GT]GTAACAA</td>
<td>Fhk1</td>
<td>2</td>
<td>79.1</td>
<td>114.9</td>
<td>194.0</td>
<td>2</td>
<td></td>
</tr>
<tr>
<td>[TCG]TTGTAC</td>
<td>Fhk1</td>
<td>3</td>
<td>71.8</td>
<td>147.6</td>
<td>219.4</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>CAGC[AG]GC</td>
<td></td>
<td>4</td>
<td>54.6</td>
<td>-28.5</td>
<td>26.1</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td>CAAGAAA</td>
<td></td>
<td>5</td>
<td>61.1</td>
<td>-26.6</td>
<td>34.5</td>
<td>7</td>
<td></td>
</tr>
<tr>
<td>T[AG]TATATAT</td>
<td></td>
<td>6</td>
<td>43.2</td>
<td>56.5</td>
<td>99.7</td>
<td>4</td>
<td></td>
</tr>
<tr>
<td>GAAAAAG</td>
<td></td>
<td>7</td>
<td>53.8</td>
<td>11.2</td>
<td>65.0</td>
<td>6</td>
<td></td>
</tr>
<tr>
<td>T[TG][TG]CC[CT]T</td>
<td></td>
<td>8</td>
<td>72.6</td>
<td>98.4</td>
<td>171.0</td>
<td>3</td>
<td></td>
</tr>
</tbody>
</table>

Table 5.3. Fkh1 consensus motifs testing result. The two motifs in shadow represent two documented Fkh1 binding motifs.
<table>
<thead>
<tr>
<th>TF Name</th>
<th>Documented binding Motif</th>
<th>Rank by MDP</th>
<th>Rank by MEME</th>
<th>Rank by MDscan</th>
<th>Rank by WordSeeker</th>
<th>Rank by DREME</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fkh1</td>
<td>[TCG]TTGTTTAC</td>
<td>1</td>
<td>2</td>
<td>N/A</td>
<td>3</td>
<td>2</td>
</tr>
<tr>
<td>Cbf1</td>
<td>[TG]CACGTG[AC][TC]</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>Leu3</td>
<td>G[CG]C[AG][CAT]GGCC</td>
<td>2</td>
<td>6</td>
<td>5</td>
<td>N/A</td>
<td>8</td>
</tr>
<tr>
<td>Mbp1</td>
<td>[AT]GC[TGA]GC[TAG][CA]</td>
<td>1</td>
<td>3</td>
<td>N/A</td>
<td>N/A</td>
<td>8</td>
</tr>
<tr>
<td>Reb1</td>
<td>[GA]TTACCCG[GC]</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

Table 5.4. Test results of all testing group.
Table 5.5. Statistical enrichment test for the three candidate motifs. The statistical expectation was calculated based on the length and frequency of each nucleotide of the motif; the Occurrence in random region was calculated based on a set of randomly extracted intergenic region with the same length of the binding region sequences.

<table>
<thead>
<tr>
<th>Candidate motifs</th>
<th>Occurrence in binding region</th>
<th>Statistical expectation</th>
<th>Enrichment to expectation</th>
<th>Occurrence in random region</th>
<th>Enrichment in binding region</th>
</tr>
</thead>
<tbody>
<tr>
<td>TTGNTTG</td>
<td>68</td>
<td>15</td>
<td>4.5</td>
<td>17</td>
<td>4</td>
</tr>
<tr>
<td>CAC[AT]GCC</td>
<td>33</td>
<td>5</td>
<td>6.6</td>
<td>2</td>
<td>16.5</td>
</tr>
<tr>
<td>TAC[AC]GTA</td>
<td>22</td>
<td>7</td>
<td>3.1</td>
<td>6</td>
<td>3.7</td>
</tr>
<tr>
<td>Candidate Motifs</td>
<td>Total occurrence</td>
<td>Occurrence in ORFs</td>
<td>Occurrence in upstream regions</td>
<td></td>
<td></td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------</td>
<td>-------------------</td>
<td>-----------------------------</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TTGNTTG</td>
<td>38124</td>
<td>12246</td>
<td>14854</td>
<td></td>
<td></td>
</tr>
<tr>
<td>CAC[AT]GCC</td>
<td>12895</td>
<td>7275</td>
<td>3253</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TAC[AC]GTA</td>
<td>9648</td>
<td>1900</td>
<td>4727</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 5.6. The occurrence of the three candidate motifs in ORFs and upstream regions.
<table>
<thead>
<tr>
<th>Candidate Motif</th>
<th>Annotation</th>
<th>Rank by MEME</th>
<th>OR Value</th>
<th>DP Value (x1000)</th>
<th>Re-rank Value</th>
<th>Rank by MDP</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAAAAAAAAAAA</td>
<td>Simple repeat</td>
<td>removed</td>
<td>356.2</td>
<td>-1.3</td>
<td>354.9</td>
<td>removed</td>
</tr>
<tr>
<td>TTTT[TC]TTTT</td>
<td>Simple repeat</td>
<td>removed</td>
<td>369.5</td>
<td>1.1</td>
<td>370.6</td>
<td>removed</td>
</tr>
<tr>
<td>TA[GC][GC]TACCT</td>
<td>MoCRZ1</td>
<td>1</td>
<td>165.7</td>
<td>40.4</td>
<td>206.2</td>
<td>2</td>
</tr>
<tr>
<td>AGGTAGGTA</td>
<td></td>
<td>2</td>
<td>84.8</td>
<td>0</td>
<td>84.8</td>
<td>8</td>
</tr>
<tr>
<td>[GT]CTTGGC</td>
<td>Rim101</td>
<td>3</td>
<td>136.3</td>
<td>88.5</td>
<td>224.8</td>
<td>1</td>
</tr>
<tr>
<td>CTAG[AT]CTAG</td>
<td></td>
<td>4</td>
<td>68.1</td>
<td>48.6</td>
<td>116.8</td>
<td>6</td>
</tr>
<tr>
<td>CACAGCC</td>
<td>MoCRZ1</td>
<td>5</td>
<td>85.5</td>
<td>46.6</td>
<td>132.2</td>
<td>5</td>
</tr>
<tr>
<td>[TG][GT][TG][TG]G</td>
<td>MoCRZ1</td>
<td>6</td>
<td>77.2</td>
<td>56.6</td>
<td>133.9</td>
<td>4</td>
</tr>
<tr>
<td>TTT[GT][GCT]TTGC</td>
<td></td>
<td>7</td>
<td>101.3</td>
<td>0.7</td>
<td>102.1</td>
<td>7</td>
</tr>
<tr>
<td>TGCCAAG</td>
<td>Rim101</td>
<td>8</td>
<td>65.7</td>
<td>78.7</td>
<td>144.4</td>
<td>3</td>
</tr>
</tbody>
</table>

Table 5.7. MoCrz1 consensus motifs testing result. The five motifs in shadow are three MoCRZ1 binding motifs and two Rim101 binding motifs.
Figure 5.1. The MoCRZ1 binding region identification experiments design. (A) ChIP-chip experimental design to identify MoCRZ1 targets activated by calcium treatment. The Ca²⁺/FK506 treated sample served as the negative control treatment. (B) Expression microarray design. Wild type strain and MoCRZ1 deletion mutant (Dmocrz1) were treated with Ca²⁺ and/or FK506 as depicted. After global normalization of signal intensities to the average expression level of all probes among the replicated data sets, pairwise comparison between treatments was conducted. (C) Venn diagram showing number of genes identified from ChIP-chip and up-regulated / down-regulated in transcriptome profiling. Number of genes with more than 2 fold differential expression with P-value < 0.05 were noted in Ca²+ treated wild type samples in each comparison.
Figure 5.2. Analysis Pipeline for generating model of MDP approach and test it using yeast transcription factor data.
Figure 5.3. Distribution frequency curve of training group. Each block coded by colors represents the frequency of one training motif in each scanning window. The curve was generated as average of all blocks.
Figure 5.4. Frequency curves of Fkh1 candidate motifs.
Figure 5.5. Predicted MoCRZ1 binding motifs identified. (A) Analysis pipeline to identify predicted MoCRZ1 binding motifs. (B) WebLogo of Top 2 motifs and the best hits in the yeast motif database. Best hit from yeast motif database was displayed for comparison. (C) Electrophoretic mobility shift assay. Probe DNA was amplified by PCR from the promoter regions of MoCRZ1 (left panel, lanes 1 to 3) and CBP1 (right panel, lanes 4 to 6). Lanes 1 and 4, Biotin-labeled probe DNA; lanes 2 and 5, probe DNA with purified MoCRZ1 protein; lanes 3 and 6, competition reaction with 200 fold molar excess of unlabeled DNA. FP: free probe; SP: shifted probe.
Figure 5.6 Frequency curves of MoCrz1 candidate motifs.
Chapter 6: Identification of novel Avr genes by Screening of Gain-of-virulence Mutants in rice blast

fungus Magnaporthe oryzae
Introduction

*Magnaporthe oryzae* is a filamentous fungus that causes rice blast disease, which is also an important model organism for investigating fungal-plant interactions. Unlike other fungal pathogens of plants, the genomic sequences of both of the pathogen and the rice host are available [1, 2], providing a unique opportunity to study a host–parasite interaction from both sides using functional genomic approaches. The gene-for-gene model has been widely accepted and utilized in agricultural disease control since it was proposed decades ago (Flor 1971). To control rice blast, introducing a variety of confirmed resistance genes (R genes) into elite rice cultivars is considered as an effective strategy. For this purpose, it becomes essential to identify a large number of R genes in rice and corresponding Avr genes in the fungus, as well as understand their interactions. While more than 73 rice blast R genes to date have been identified and physically mapped [3], only a limited number of Avr genes have been identified [4-11]. This is partially because there are no obvious sequence patterns that can be easily recognized.

This study focuses on the corresponding Avr genes of the two R genes Pi-2 and Pi-9. The Pi-2 gene was introgressed from a highly resistant indica cultivar 5173 into the susceptible cultivar CO39 and named C101A51 [12]. Through wild hybridization and repeated backcrossing, the resistance gene Pi-9 was transferred from *O. minuta* into the elite breeding line IR31917 [13]. To examine the resistance spectra of Pi-2 and Pi-9 to isolates from other rice-growing areas, Dr. Guo-liang Wang’s group in the Ohio State University has tested them on 43 *M. oryzae* isolates collected from 13
different countries. The inoculation experiments confirmed that these two genes conferred high resistance to many blast isolates. Using the identified RAPD markers and BAC ends, a high-resolution map of the Pi-9 locus was constructed [14]. Interestingly, Pi-2 was found tightly linked to all of the Pi-9 markers, suggesting that Pi-2 and Pi-9 are co-localized on rice chromosome 6, or may actually be allelic [15].

To generate a library of *Magnaporthe* random mutants efficiently, *Agrobacterium tumefaciens*-mediated transformation (ATMT) was used in this study. ATMT is used to transfer genes to a wide variety of plants [16] as well as filamentous fungi [17]. Recently, several fungi were transformed using an ATMT approach to generate banks of random mutants [18, 19]. For insertion mutagenesis, this technique offers a huge potential [20, 21]. One of the principal advantages of ATMT over conventional transformation techniques is the versatility in choosing transformation system. *A. tumefaciens* is able to transform using protoplasts, hyphae, spores, and blocks of mushroom mycelia tissue [17, 22, 23]. Furthermore, ATMT generates a high percentage of transformants with only one single insert of T-DNA in the fungal genome, which will facilitate the subsequent isolation of tagged genes [24].

In this study, we generated random insertion transformants of *M. oryzae* through ATMT approach. By conducting pathogenicity assay on rice cultivars containing Pi-2 and Pi-9 R genes, we re-isolated 4 mutants that were able to break Pi-2 or Pi-9 resistance (Pipeline in Figure 6.1). Then we used a TAIL-PCR (Thermal Asymmetric Interlaced PCR) approach to amplify the insertion flanking
sequences to clone the affected locus. Several candidate genes were identified. However, the complementary experiments to confirm their avirulence gene functions remain.

**Results**

**ATMT transformants generation and pathogenicity screening**

A pipeline was built to produce ATMT mutants with the speed of 200 mutants a week. In total, roughly 5,000 ATMT mutants were generated during six months of continual transformations. The transformation cassette used to insert into *M. oryzae* isolate KJ201 genome contains a hygromycin resistant gene and a GFP expression gene with proper promoters. Each transformant was labeled and stored, before being put through pathogenicity tests on susceptible Nipponbare (NPB), and resistant Pi-2 and Pi-9 rice plants.

During pathogenicity screens, each possible lesion was transferred onto selection media to increase sensitivity. Among the 5,000 tested mutants, only 5 mutated strains were successfully re-isolated from lesions (Figure 6.2). They were then re-inoculated onto rice cultivars again, where 4 of them were re-isolated from lesions and thus became candidate mutants. They were renamed as Iso_1, Iso_2, Iso_3, and Iso_4 (Figure 6.3).

All the 4 candidate gain-of-virulence mutants were confirmed to have GFP signal gained from the ATMT inserted vectors under fluorescence microscope, with strong signal observed in conidia (Figure
Pathogenicity tests of four candidate mutants

The 4 candidate mutants were inoculated to rice cultivars to test their pathogenicity. The pathogenicity test has been performed both in Dr. Thomas Mitchell’s lab at The Ohio State University (Table 6.1, Figure 6.3) as well in Dr. Qinghua Pan’s lab in South Agricultural University in China to independently verifying results (Table 6.2). The pathogenicity test performed at OSU included *M. oryzae* isolate KJ201 as the negative control and NPB as the susceptible rice cultivar control, as well as rice lines containing the Pi-2 and Pi-9 resistant genes respectively. Lesions from each pots were counted. Positive controls for infection for all assays NPB showed average lesion counts more than 46. Negative control with wild KJ201 failed to make a single lesion on Pi-2 and Pi-9 containing plants, with an average lesion count less than one. Iso_1 had 7 average lesions on Pi-2, while this number for Iso_2 and Iso_4 were 15 and 5 respectively, while Iso_3 only generated tiny black spots and no typical lesions (Table 6.1, Figure 6.3). This pathogenicity test supported Iso_1, Iso_2 and Iso_4 strong Avr gene candidate mutants. For pathogenicity tests conducted in Dr. Qinghua-Pan’s lab, all controls performed as expected with Iso_4 showing virulence on Pi-2, Iso_2 showing virulence on Pi-9, and Iso_3 showing modest virulence on Pi-2. To summarize pathogenicity test from both labs, Iso_2 and Iso_4 were supported to gain the virulence to Pi-2 or Pi-9, and were advanced for further analysis and identification of the mutated gene.
TAIL-PCR to identify insertion regions

Direct TAIL-PCR of Iso_2 and Iso_4 showed multiple bands and Sanger sequencing results showed mixed nucleotide calls, so a cloning procedure was conducted prior to TAIL-PCR. After cloning and isolating, TAIL-PCR and sequencing was performed successfully in Dr. Qinghua Pan’s lab (Table 6.3, Figure 6.5).

Characterization of candidate genes

Sequences of TAIL-PCR products were aligned to the *M. oryzae* 70-15 reference genome using The Broad Institute’s web service. As a result, 6 candidate genes were identified. Features of these 6 candidate genes were characterized as showed in table 6.4. The six candidate genes were located at chromosomes I, III, and IV. Although none of them were predicted to contain a signal peptide, 4 of them were predicted to localize extracellular. The insertion position of T-DNA varies, as 2 were inserted at 5’ upstream regions, 3 were inserted at 3’ downstream regions, and 1 was inserted in the coding region. Although inserting in upstream or coding region may have larger effect on gene function, the 3 downstream insertions may also change gene expression. We cannot confirm or eliminate any of these candidate genes based on the characterization information so far.
**Future work**

This is an ongoing project and there’s still work to do before the identification of Avr-Pi2 and Avr-Pi9 genes. Future direction includes:

1. Functional annotation of all the candidate genes, estimate locus structure in all sequenced *M. oryzae* genomes, and predict the candidate Avr gene conferring virulence.

2. Transfer the candidate Avr-Pi2 genes into rice cultivar Chinos or other isolates which do not contain AvrPi2 gene to confirm its function.

**Methods and Materials**

**Strains used**

The fungal strain used to generate mutants is *M. oryzae* strain KJ201. The *Agrobacterium tumefaciens* strains used in ATMT approach is TMC55, which contained Kanamycin resistance gene as selection marker. The Rice strains used in this project were Nipponbare as a positive control and two strains containing Pi-2 and Pi-9, respectively provided by Dr. Guo-liang Wang’s group.
Number of mutants necessary to generate

Given the fact that the average gene length of *M. oryzae* is 1.7kb and the whole genome size is around 38Mb [1], possibility of transformants covering any one particular gene is calculated in Table 6.5. So if we want to reach an acceptable possibility to cover both Avr-Pi2 and Avr-Pi9, 30,000 transformants is the number that we consider. There’s limited information from previous study focusing on this screen of gain-of-virulence, which made it difficult to expect how many among 30,000 transformants will make visible changes in pathogenicity. However, in one reported study, 639 transformants of *M. oryzae* were obtained by REMI and 4 of them showed gain of ability to infect a rice strain which was resistant to wild-type [25]. If this proportion works in our case, then 31 transformants containing virulence ability should be expected among 5,000 transformants.

*Agrobacterium tumefaciens*-mediated (ATMT) transformation protocol

1. Grow *Agrobacterium* strain containing an appropriate binary vector in 5 ml Minimal medium (MM) with an appropriate antibiotics (Kanamycin for our binary vectors) for 1-2 days at 28 °C.

2. Inoculate the culture to Induction Medium (IM), containing Kan (75 ppm) and acetosyringone (AS), make O.D.600=0.15.

3. Culture the *Agrobacterium* cells for 6 hours at 28 °C by shaking at 200 rpm.
(Prepare Co-cultivation media with membrane on it, we can make 3 square plates with 100ml, 5 membranes per plate; Prepare TB3 agar with 200ppm hygromycin in square plates)

4. Prepare the *M. oryzae* spore suspension with dH₂O at concentration of 10⁶ spores/ml.

5. Mix 100 µl of the spore suspension with 100 µl of *Agrobacterium* culture.

6. Spread the 200 µl mixture onto the 3 nitrocellulose membrane (Whatman Cat. #7141 104; 47mm in diameter; 0.45 um in pore size) placed on the co-cultivation medium.

7. Incubate the plates for 26 hrs (24 to 48 hrs) under room temperature.

8. Transfer the membrane to the selection medium for 4 days.

**Pathogenicity assay**

Pathogenicity of the mutants generated was tested on pot-grown rice plants at the 6-8 leaf stage. In order to scan large number of mutants efficiently, 50 ATMT mutants were transformed to one V8 medium for growth and sporulation. Conidial suspensions (5 × 10⁴ conidia/ml) from 1 week cultures on V8 agar plates were collected and spray-inoculated onto rice plants grown in commercial potting soil in vinyl pots (25cm³). Each mixture of conidial suspension were inoculated on a group of rice including 2 pots of Pi-2, 2 pots of Pi-9 and 1 pots of *Nipponbare*. Inoculated plants were kept in a dew chamber for 24 hrs and allowed to grow in a growth chamber at the Ohio Department of
Agriculture. Disease symptoms were examined 1 week after inoculation [8].

Lesions on each plant were counted and scanned for documentation. Then the edge of lesion was cut into 0.5cm x 0.5cm pieces to process through a surface sterilization and then plated on a water agar plate. Mycelia growing out from the lesions were then transferred to selection medium containing hygromycin. Mycelia growing on the selection medium were then transferred to a V8 medium agar for sporulation and then inoculated again to Pi-2, Pi-9 and Nipponbare plants to confirm its gain-of-virulence.

**Recovery of flanking sequences**

The insertion sequences and their associated flanking sequences were recovered from ATMT lines using TAIL-PCR (Thermal Asymmetric Interlaced PCR) following the protocol described [26]. Genomic DNA of the confirmed mutants were extracted and amplified by TAIL-PCR using primers in Table 6.6 [27]. The PCR products were then purified and sequenced using traditional Sanger sequencing (Figure 6.6). Sequencing results were then mapped to *M. oryzae* reference 70-15 genome sequences downloaded from the Broad Institute using their online blast service to localize and identify the insertion regions [28].
References


### Table 6.1
Pathogenicity test conducted in Dr. Mitchell’s lab. Result indicates Iso_1, Iso_2, and Iso_4 may gain the virulence to break Pi-2 resistance.

<table>
<thead>
<tr>
<th>Host</th>
<th>KJ201</th>
<th>Iso_1</th>
<th>Iso_2</th>
<th>Iso_3</th>
<th>Iso_4</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Ave</td>
<td>Ave</td>
<td>Ave</td>
<td>Ave</td>
<td>Ave</td>
</tr>
<tr>
<td>NPB</td>
<td>63</td>
<td>71</td>
<td>101</td>
<td>51</td>
<td>115</td>
</tr>
<tr>
<td></td>
<td>52</td>
<td>46</td>
<td>43</td>
<td>50</td>
<td>75</td>
</tr>
<tr>
<td></td>
<td>28</td>
<td>45</td>
<td>53</td>
<td>51</td>
<td>93</td>
</tr>
<tr>
<td>Pi-2</td>
<td>0</td>
<td>0</td>
<td>23</td>
<td>8</td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>6</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td></td>
<td>0</td>
<td>17</td>
<td>22</td>
<td>8+spot</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.1. Pathogenicity test conducted in Dr. Mitchell's lab. Result indicates Iso_1, Iso_2, and Iso_4 may gain the virulence to break Pi-2 resistance.
<table>
<thead>
<tr>
<th>Host</th>
<th>KJ201</th>
<th>Iso_1</th>
<th>Iso_2</th>
<th>Iso_3</th>
<th>Iso_4</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPB</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>Pi-2</td>
<td>ND</td>
<td>R</td>
<td>R</td>
<td>ND</td>
<td>S</td>
</tr>
<tr>
<td>Pi-9</td>
<td>R</td>
<td>R</td>
<td>ND</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>Piz-t</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>Q1063</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
<td>S</td>
</tr>
</tbody>
</table>

S: Susceptible; R: Resistant; ND: Not determined;

Table 6.2. Pathogenicity test conducted in Dr. Pan’s lab. Result indicates Iso_2, Iso_3, and Iso_4 may gain the virulence to break Pi-2 or Pi-9 resistance.
Table 6.3. Summary of TAIL-PCR results of different clones generated from Iso_2 and Iso_4.

<table>
<thead>
<tr>
<th>T-DNA</th>
<th>Clone</th>
<th>Flanking sequences</th>
<th>T-DNA sequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>LB</td>
<td>Mu1</td>
<td>Supercontig 12:390652-391443</td>
<td>207-bp</td>
</tr>
<tr>
<td></td>
<td>Mu2</td>
<td>No specificity</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td>Mu3</td>
<td>Supercontig 12:390652-391443</td>
<td>207-bp</td>
</tr>
<tr>
<td>RB</td>
<td>Mu13</td>
<td>Surpercontig 22: 490226-491286</td>
<td>211-bp</td>
</tr>
<tr>
<td></td>
<td>Mu18</td>
<td>Surpercontig 21: 1734002-1734834</td>
<td>No</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Iso_2</td>
<td>Mu9</td>
<td>Supercontig 27:1359766-1360697</td>
<td>160-bp</td>
</tr>
<tr>
<td></td>
<td>Mu10</td>
<td>Supercontig 27:1360719-1361435</td>
<td>208-bp</td>
</tr>
<tr>
<td>RB</td>
<td>Mu29</td>
<td>Surpercontig 22: 490349-491286</td>
<td>213-bp</td>
</tr>
<tr>
<td></td>
<td>Mu30</td>
<td>Surpercontig 16: 245651-246865</td>
<td>No</td>
</tr>
</tbody>
</table>

Table 6.3. Summary of TAIL-PCR results of different clones generated from Iso_2 and Iso_4.
<table>
<thead>
<tr>
<th>Target gene</th>
<th>AvrPi9</th>
<th>AvrPi9</th>
<th>AvrPi2</th>
<th>AvrPi2</th>
<th>AvrPi2/9</th>
<th>AvrPi2/9</th>
</tr>
</thead>
<tbody>
<tr>
<td>Locus</td>
<td>MGG_04211.6</td>
<td>MGG_12175.6</td>
<td>MGG_08831.6</td>
<td>Unnamed locus</td>
<td>MGG_04842.6</td>
<td>MGG_04843.6</td>
</tr>
<tr>
<td>Gene product</td>
<td>Predicted protein</td>
<td>Tyrocidine synthetase 1</td>
<td>hypothetical</td>
<td>Unknown</td>
<td>Homoaconitate</td>
<td>C6 zinc finger domain-containing protein</td>
</tr>
<tr>
<td>Chromosome</td>
<td>IV</td>
<td>IV</td>
<td>I</td>
<td>I</td>
<td>III</td>
<td>III</td>
</tr>
<tr>
<td>Length Amino acid</td>
<td>109</td>
<td>4837</td>
<td>269</td>
<td>19</td>
<td>776</td>
<td>599</td>
</tr>
<tr>
<td>Signal Peptides</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>Subcellular location (score)</td>
<td>Extracellular (9.7)</td>
<td>Extracellular (4.4) Cytoplasmic (1.09)</td>
<td>Extracellular (4.43) Plasma membrane (2.20)</td>
<td>No</td>
<td>Extracellular (4.43) Plasma membrane (2.20)</td>
<td>Mitochondrial (3.16) Cytoplasmic (2.86)</td>
</tr>
<tr>
<td>transmembrane helices</td>
<td>No</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>GPI</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
<td>No</td>
</tr>
<tr>
<td>T-DNA inserted position</td>
<td>5’ upstream 2183-bp</td>
<td>5’ upstream 695-bp</td>
<td>Coding region 835-bp</td>
<td>3’ downstream 105-bp</td>
<td>3’ downstream 2512-bp</td>
<td>3’ downstream 790-bp</td>
</tr>
</tbody>
</table>

Table 6.4. Features of 6 candidate Avr genes.
Table 6.5. Calculation of possibility to identify one Avr gene or two genes using random insertion approach when different number of mutants generated.

<table>
<thead>
<tr>
<th>Number of mutants</th>
<th>Chance to identify one gene (Pi2 or Pi9)</th>
<th>Chance to identify both genes (Pi2 and Pi9)</th>
<th>Chance to identify at least one gene</th>
</tr>
</thead>
<tbody>
<tr>
<td>5,000</td>
<td>20.0%</td>
<td>4.0%</td>
<td>36.0%</td>
</tr>
<tr>
<td>10,000</td>
<td>36.1%</td>
<td>13.0%</td>
<td>59.2%</td>
</tr>
<tr>
<td>20,000</td>
<td>59.1%</td>
<td>34.9%</td>
<td>83.3%</td>
</tr>
<tr>
<td>30,000</td>
<td>73.9%</td>
<td>54.6%</td>
<td>93.2%</td>
</tr>
<tr>
<td>40,000</td>
<td>83.3%</td>
<td>69.4%</td>
<td>97.2%</td>
</tr>
<tr>
<td>50,000</td>
<td>89.3%</td>
<td>79.7%</td>
<td>98.9%</td>
</tr>
<tr>
<td>Code</td>
<td>Type</td>
<td>Sequence</td>
<td></td>
</tr>
<tr>
<td>-------</td>
<td>-----------</td>
<td>-------------------------------</td>
<td></td>
</tr>
<tr>
<td>TMP633</td>
<td>LB1_T_DNA</td>
<td>GTCCGAGGGCAAAGAAATAGAGTA</td>
<td></td>
</tr>
<tr>
<td>TMP634</td>
<td>LB2_T_DNA</td>
<td>CATGTGTGGAGCATATAAGAAACCCT</td>
<td></td>
</tr>
<tr>
<td>TMP635</td>
<td>LB3_T_DNA</td>
<td>GAATTAATTCGGCGTTAATTCAGT</td>
<td></td>
</tr>
<tr>
<td>TMP636</td>
<td>RB1_T_DNA</td>
<td>TTACAACGTCGTGACTGGAAAAAC</td>
<td></td>
</tr>
<tr>
<td>TMP637</td>
<td>RB2_T_DNA</td>
<td>CTGGCGTAATAGCGAAGAAAAGG</td>
<td></td>
</tr>
<tr>
<td>TMP638</td>
<td>RB3_T_DNA</td>
<td>CCCCTCCACACAGTTGCACCA</td>
<td></td>
</tr>
<tr>
<td>TMP639</td>
<td>AD1_T_DNA</td>
<td>NGTCGASWGANAWGGA</td>
<td></td>
</tr>
<tr>
<td>TMP640</td>
<td>AD2_T_DNA</td>
<td>TGWGNAGSANCASAGG</td>
<td></td>
</tr>
<tr>
<td>TMP641</td>
<td>AD3_T_DNA</td>
<td>AGWGNAGWANCAWAGG</td>
<td></td>
</tr>
<tr>
<td>TMP642</td>
<td>AD4_T_DNA</td>
<td>WAGTGNAGWANCANGA</td>
<td></td>
</tr>
<tr>
<td>TMP643</td>
<td>AD6_T_DNA</td>
<td>WGTGNAGWANCANAGA</td>
<td></td>
</tr>
<tr>
<td>TMP644</td>
<td>AD-1_T_DNA</td>
<td>WAGTGNAGWANCANAGA</td>
<td></td>
</tr>
<tr>
<td>TMP904</td>
<td>LAD1-1</td>
<td>ACGATGGACTCCAGAGCGGCVNVNNNGGA</td>
<td></td>
</tr>
<tr>
<td>TMP905</td>
<td>LAD1-3</td>
<td>ACGATGGACTCCAGAGCGGCVNVNNNCCA</td>
<td></td>
</tr>
<tr>
<td>TMP906</td>
<td>AC1</td>
<td>ACGATGGACTCCAGAG</td>
<td></td>
</tr>
</tbody>
</table>

Table 6.6. Primers used in TAIL-PCR.
Figure 6.1. Screening Pipeline. ATMT mutants were generated and pooled together, then the mixtures were inoculated onto rice leaves. Candidate mutants were reisolated from lesions and selected by hygromycin selection medium.
Figure 6.2. Lesions in the first re-inoculation of the 5 isolates from pooled inoculation. Note Iso_5 showed only minimum lesion, similar to the wild type KJ201, so Iso_5 were eliminated as candidates from this step.
Figure 6.3. Lesions from pathogenicity test conducted in Dr. Mitchell’s lab. Both positive and negative control works and Iso_1, Iso_2, and Iso_4 all made lesions on Pi-2 containing rice strain.
Figure 6.4. All four candidate mutants showed strong GFP signal. (A) Iso_1; (B) Iso_2; (C) Iso_3; (D) Iso_4;
Figure 6.5. Position of the T-DNA insertion in Iso_2 and Iso_4 mutants. The nucleotide in the brackets was deleted by T-DNA insergration. The arrows (▼) indicate the T-DNA (2.2 kb) insertion positions in Iso_2 and Iso_4, respectively. The thick arrows represent the coding regions and the thin line joining these coding regions indicates the position of the introns, or the intergenic regions.
Figure 6.6. Structure of the T-DNA inserted. In TAIL-PCR, multiple primers from LB or RB were matched with Arbitrary primers to capture the outer sequences of T-DNA.
Conclusions

This dissertation focuses on the application of next generation sequencing (NGS) approaches on the studies of fungal phytopathogens. The first chapter summarizes the development of NGS techniques and their successful application to studies to understand the molecular underpinning of fungal virulence to plants. NGS plays an important role in investigating biological questions at the genome scale, such as comparisons of between genomes, identification of differentially expressed genes, and characterization of binding regions of transcription.

Chapter 2 and Chapter 4 described two genome sequence comparison projects. In the first project, we identified *Alternaria arborescens* conditionally dispensable chromosomes (CDC) sequences through whole genome sequencing and a *de novo* assembly process. By comparing nucleotide variations between CDC and essential chromosome (EC) contigs, we found evidence supporting horizontal gene transfer of the CDC to *A. arborescens*. We also identified predicted CDC genes under positive selection that may serve as virulence factors. In Chapter 4 the *M. oryzae* comparative study, genomes of 7 *M. oryzae* field isolates were sequenced and a reference assisted *de novo* assembly was performed. We found the general genome content of the sequenced field isolates to be similar to the reference 70-15, but with fewer gene models predicted. One interesting finding was the close
relationship between 70-15 and two barley pathogens. Variations at the gene level in the form of SNPs, Indels, copy number variations were identified, and it was interesting to find that most of these polymorphisms were enriched in local genome neighborhoods, which we call “hot spots”. The selection ratios (Ka/Ks) for individual genes were also calculated. Thus, those genes unique to 70-15 located in the “hot spots” and containing Ka/Ks ratio >1 we predict to be candidates for novel virulence related and avirulence genes.

A RNA-Seq project is described in Chapter 3, where two NGS platforms were used to generate transcriptome sequencing data: Illumina’s sequencing-by-synthesis (SBS) and Roche’s 454-pyrosequencing. The 454 reads were used for the de novo assembly of S. homoeocarpa and creeping bentgrass transcriptome libraries, while SBS reads were mapped to the 454 assemblies to calculate transcript levels in order to identify differentially expressed genes associated with the infection. Fungal genes associated with infections include glycosyl hydrolases, proteases, and ABC transporters. Particularly, a large number of glycosyl hydrolase transcripts that target a wide range of plant cell wall compounds were found. Transcripts in creeping bentgrass associated with infections include germins, ubiquitin transcripts involved in proteasome degradation, and cinnamoyl reductase, which is involved in lignin production.

In Chapter 5, we proposed a novel motif distribution pattern (MDP) approach to improve transcription factor binding site (TFBS) predictions using data generated from ChIP-Seq techniques. In the MDP approach, the distance to the transcription start site of every documented binding motif
was collected and utilized to build a spatial model. The MDP approach improves the accuracy of predictions by estimating both the over-representative level of the candidate motifs and their general distribution pattern. The improved performance of MDP was observed in both 5 transcription factors in yeast and MoCRZ1 in *M. oryzae*.

All these high through-put projects using NGS techniques showed enormous variations in genome content between and within species and the dramatic expression changes in the transcriptome expression profiles during infection related development. For example, the number of SNP and Indel locations among different *M. oryzae* field isolates can reach as high as over ten thousand, and over 20% of genes in the RNA-Seq projects showed alternate splicing that generate multiple isoforms with different expression profiles. Studying systems with this extreme complexity can only be approached through NGS techniques to provide data at the whole genome scale. Concurrently, advanced bioinformatics tools are needed to reduce the complexity so that questions can be asked of the data. When the proper analysis pipeline is built, another issue researchers need to cope with is uncertainty, which is represented by the extensive number of parameters and cut-off options. By setting these values, researchers unavoidably introduce a bias, which is necessary but can be subjective and possibly misleading. To arrive at a robust conclusion, biological and technical replications as well as statistical support are required but rarely performed. Another way to validate the *in silico* analysis is to carry out experimentally validation with smaller data sets, such as Southern hybridizations for confirming the CDC gene identification in Chapter 2, RT-PCR for confirming the RPKM values as in Chapter 3, and gel shift assays for confirming the binding motif predictions as in
Chapter 5. Overall, projects in this dissertation demonstrated the practicability of applying NGS techniques to the analysis of fungal genomes and transcriptomes to explore the underlying mechanisms of virulence.
References


synthase peptide synthetase from Magnaporthe grisea signals pathogen attack to resistant rice. The Plant cell 2004, 16(9):2499-2513.


80. Han Y, Liu X, Benny U, Kistler HC, VanEtten HD: Genes determining pathogenicity to pea are clustered on a supernumerary chromosome in the fungal plant pathogen Nectria


102. Kars I vKJ (ed.): Extracellular enzymes and metabolites involved in pathogenesis of Botrytis.


180. Pfaffl MW, Horgan GW, Dempfle L: Relative expression software tool (REST (c)) for group-wise comparison and statistical analysis of relative expression results in real-time PCR. Nucleic acids research 2002, 30(9).


SALAMIAH YF-N, Hajime AKAMATSU, Hiroshi OTANI, Keisuke KOHMOTO and Motoichiro KODAMA: Genetic Analysis of Pathogenicity and Host-specific Toxin Production of Alternaria


228. Stefanato FL, Abou-Mansour E, Buchala A, Kretschmer M, Mosbach A, Hahn M, Bochet CG, Metraux JP, Schoonbeek HJ: The ABC transporter BcatrB from Botrytis cinerea exports...


