Creating a genomic resource for *Grosmannia clavigera*

An investigation into the physiological and chemical interactions between *Grosmannia clavigera,* a mountain pine beetle fungal associate and lodgepole pine metabolites involved in tree defence

by

William Scott DiGuistini

B.Sc., Simon Fraser University, 2003

A thesis submitted in partial fulfillment of
the requirements for the degree of

Doctor of Philosophy

in

The Faculty of Graduate Studies

(Forestry)

The University of British Columbia
(Vancouver)

July 2010

Abstract

Rapid advances in DNA sequencing have created the possibility for probing deeply into the genomes and transcriptomes of organisms; however, bioinformatic tools for handling these massive quantities of data are changing rapidly. In this thesis I describe the bioinformatic methods and molecular resources developed for studying the ascomycete fungus Grosmannia clavigera (Gc), a lodgepole pine (Pinus contorta) pathogen specifically associated with the Mountain Pine Beetle (MPB), and results from initial analyses using these resources. We have developed genomic resources for Gc including: a) the whole genome sequence, b) expressed sequence tags (ESTs), and c) RNA-seq data. We have annotated the genome using the transcriptome resources (e.g. ESTs and RNA-seq) and computational (gene prediction software) data. In an initial analysis of the Gc genome we focused on aspects important for colonizing the host tree and for tolerance towards conifer defence chemicals. We showed that Gc is heterothallic, and report evidence for repeat-induced point mutation. Gene expression profiling provided insight into mechanisms for Gc's tolerance towards conifer defence chemicals, specifically oleoresin terpenoids. RNA-seq revealed a substantial antimicrobial stress response of Gc induced by terpenoids, and our data suggests that Gc may reduce the toxicity of these defence chemicals by utilizing them as a carbon source. Terpenoid treatment strongly activated a ~100 kb region of the Gc genome that contains a set of genes that may be important for detoxification of these host defence chemicals. Using a recently developed Gc gene knock-out system, we provide evidence that a PDR-type ABC transporter is important for the successful response of Gc against host terpenoids.

Table of contents

Abstract ii

Table of contents iv

List of tables..........ix

List of figures x

List of abbreviations..........xi

Acknowledgements..........xiv

Dedication..........xv

Co-authorship statement..........xvi

1 Creating genomic resources for a non-model organism 1

1.1 Introduction 1

1.2 Genome sequencing 3

1.3 Genome annotation 11

1.4 The Mountain pine beetle system..........15

1.4.1 Mountain pine beetle life cycle 15

1.4.2 Microorganisms associated with the mountain pine beetle..........16

1.4.3 Introduction to conifer defenses 19

1.4.4 Pathogenic strategies for overcoming host defenses 23

1.5 Overview and purpose of this thesis..........26

1.6 References 27

2 Generation and annotation of lodgepole pine and oleoresin induced expressed sequences from the blue-stain fungus Ophiostoma clavigerum, a mountain pine beetle associated pathogen..........39

2.1 Introduction..........39

2.2 Materials and methods 41

2.2.1 Fungal strain and culture conditions..41

2.2.2 RNA isolation, cDNA library normalization..42

2.2.3 DNA sequencing and analysis...42

2.2.4 Annotation of ESTs 42

2.2.5 Quantitative real-time PCR 43

2.3 Results...44

2.3.1 Generation and assembly of EST collections 44

2.3.2 EST analysis 45

2.4 Discussion...48

2.5 References 57

3 De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data 60

3.1 Introduction...60

3.2 Materials and methods 63

3.2.1 Library construction and sequencing 63

3.2.2 Filtering Sanger and 454 reads...64

3.2.3 Assembling Illumina data...65

3.2.4 Forge hybrid assembler and genome assembly analysis 67

3.2.5 Generating the GCgb1 genome sequence 69

3.3 Results...69

3.3.1 Generating sequence data..69

3.3.2 Initial assembly analysis..70

3.3.3 Optimizing Sanger/454 assemblies using 454 read filtering 72

3.3.4 Improving assemblies with Illumina PE reads by trimming and filtering73

3.3.5 Assessing the final assembly 76

3.3.6 Analysis of Illumina and 454 read data..78

3.4 Discussion..79

3.5 References 91

4 Genome and transcriptome analysis of the mountain pine beetle-fungal symbiont Grosmannia clavigera, a lodgepole pine pathogen,..93

4.1 Introduction 93

4.2 Materials and methods 95

4.3 Results..97

4.3.1 Genome sequence and protein coding annotations in G. clavigera.....97

4.3.2 Identification of protein coding sequence variations 98

4.3.3 Detection of G. clavigera gene orthologues 99

4.3.4 Identification and annotation of genes and proteins for inhabiting host pine...100

4.3.5 Identification and annotation of genes for detoxifying host defence metabolites 102

4.3.6 Response of G. clavigera to lodgepole pine phloem extract (LPPE) treatment 103

4.3.7 Response of G. clavigera to terpene treatment.................................105

4.4 Discussion...107

4.5 References..125

5 Conclusions...129

5.1 Genomics in a non-model system 130

5.1.1 Generating a genome sequence for G. clavigera..............................131

5.1.2 Generating EST and RNA-seq resources for annotating the G. clavigera genome sequence...132

5.1.3 Annotating the G. clavigera genome sequence 134

5.2.1 Inhabiting the wood substrate 135

5.2.2 Detoxifying host-specific defense metabolites 136

5.3 Perspective on future work..137

5.3.2 Further resource development for G. clavigera 137

5.3.3 G. clavigera host-specific metabolite detoxification............................138

5.4 References..139

Appendix A. Supplementary information supporting chapter 2 140

A.1 Fungal strain and culture conditions used in chapter 2 140

A.2 RNA isolation, cDNA library normalization used in chapter 2.....................140

A.3 Sequence annotation used in chapter 2 142

A.4 Quantitative real-time PCR...142

A.5 References 143

Appendix B. Supplementary information supporting chapter 3 144

B.1 454 SE read filtering...144

B.2 Trimming and filtering Illumina PE reads 146

B.3 Assessing Forge assemblies that integrate Illumina PE read data directly.150

Appendix C. Supplementary information supporting chapter 4 151

C.1 Genome finishing 151

C.1.1 Genome finishing stages 151

C.2 Predicting protein coding gene models for G. clavigera.............................157

C.3 Phylogenetic tree construction 158

C.4 Positive selection analysis of Methyltransferases 159

C.5 Peptide sequencing for the identification of extracellular proteins 159

C.5.1 Generating biological materials 159

C.5.2 Protein extraction...160

C.5.3 1D SDS-PAGE and LC-MS/MS 160

C.5.4 Data analysis 161

C.6 Protein coding SNP discovery using RNA-seq 163

C.6.1 Generating biological materials 163

C.6.2 Generating RNA-seq data...164

C.6.3 Predicting SNPs from the G. clavigera genome 165

C.6.4 Validating RNA-seq SNP predictions...166

C.7 RNA-seq expression experiments...168

C.7.1 Terpene and Phenolic treatments 168

C.8 Gene deletion and fungal growth conditions..174

C.8.1 Generating the deletion mutant...174

C.8.2 Culturing media...174

C.9 References..177

List of tables

Table 2.1 Primers used in chapter 2........53

Table 2.2 Summary of ESTs reported in chapter 2........53

Table 2.3 UPTs ranked by EST frequency........54

Table 2.4 UPTs with a similar sequence in the PHI-base database........55

Table 2.5 Putative CYP450 and ABC-type transporter UPTs identified in the *G. clavigera* EST collection........56

Table 2.6 UPTs putatively ascribed to the DHN-melanin biosynthesis pathway......57

Table 3.1 Velvet assemblies of the *G. clavigera* genome generated from Illumina GA$_{ii}$ read data........86

Table 3.2 Forge assemblies of *G. clavigera* generated using Illumina, 454 and Sanger read data........87

Table 4.1 Genome characteristics for *G. clavigera*........114

Table 4.2 Origin of strains used for RNA-seq SNP calling of *G. clavigera* protein-coding gene regions........115

Table 4.3 Enzymes identified by peptide sequencing *G. clavigera* extracellular protein fraction........116

Table 4.4 Lodgepole pine phloem extract (LPPE) induced gene expression clusters........117

Table 4.5 A terpene induced gene expression cluster........118

List of figures

Figure 3.1 Overview of the process for assembling the *G. clavigera* draft genome sequence........88

Figure 3.2 Effects of sequencing platforms on consensus sequence indels............89

Figure 3.3 Validation of draft assemblies by alignment to intermediate reference sequence........90

Figure 3.4 Quality control analysis of genomic read data used for assembling *G. clavigera* genome sequence 91

Figure 4.1 Overview of MPB biology and taxonomic position of *G. clavigera* 119

Figure 4.2 Growth of *G. clavigera* following treatment with LPPE or terpenes 121

Figure 4.3 Circular venn diagram describing gene expression relationships between different treatment experiments 122

Figure 4.4 RNA-seq profiling and expression clusters 123

Figure 4.5 Functional analysis of terpene catabolism 124

List of abbreviations

Numbers & Symbols	
kb	Kilobase
Mb	Megabase
Gb	Gigabase
µg	Microgram
GHz	Gigahertz
µL	Microlitre
mL	Millilitre
L	Litre
nM	Nanomolar
min	Minute
sec	Second
hr	Hour
cm	Centimeter
Species names	
A. nidulans	*Aspergillus nidulans*
D. jeffreyii	*Dendroctonus jeffreyii*
D. ponderosae	*Dendroctonus ponderosae*
E. coli	*Escherichia coli*
Gc	*Grosmannia clavigera*
G. clavigera	*Grosmannia clavigera*
M. grisea	*Magnaporthe grisea*
N. crassa	*Neurospora crassa*
O. clavigerum	*Ophiostoma clavigerum*
O. montium	*Ophiostoma montium*
A - Z	
A	Adenine
ABC	ATP- binding cassette
ATP	Adenosine-5'-triphosphate
B.C.	British Columbia
BAC	Bacterial artificial chromosome
BLAST	Basic local alignment search Tool
BLASTx	Nucleotide query-protein database BLAST
bp	Base pair
C	Cytosine
CAZy	Carbohydrate-active enzymes
CBM	Carbohydrate binding module
cDNA	Complementary DNA

CE	Carbohydrate esterase
CEGMA	Core eukaryotic genes mapping approach
CFEM	Eight cysteine-containing protein domain
CFS	Canadian Forest Service
CRT	Cyclic reversible termination
CYP450	Cytochrome P450
DHN	1,8-Dihydroxynaphthalene
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxyribonucleotide triphosphate
e-value	Expected value
EC	Expression cluster
EST	Expressed sequence tags
FGSC	Fungal genetic stock center
FNR	False negative rate
FOLy	Fungal oxidative lignin enzyme
FPC	Finger printed contigs
FPR	False positive rate
G	Guanine
GA	Genome analyzer
GHs	Glycoside hydrolases
GO	Gene ontology
HMG	High-mobility group
ID	Identity
KEGG	Kyoto encyclopedia of genes and genomes
LPPE	Lodgepole pine phloem extract
M	Mega
MAQ	Mapping and assembly with quality
MAT	Mating-type
MDR	Multidrug resistance
MEA	Malt extract agar
MFS	Major facilitator superfamily
MPB	Mountain pine beetle
mRNA	Messanger RNA
mtDNA	Mitochondrial DNA
Mtfse	Methyltransferase
N50	The scaffold (contig) length for which 50% of the assembled genome is in scaffolds (contigs) that are at least as long as N50
NAD(P)H	Nicotinamide adenine dinucleotide phosphate
NCBI	National Center for Biotechnology Information
NGS	Next generation sequencing
nr	Non-redundant
odw	Organic dry weight
OLC	Overlap-layout-consensus

PAGE	Polyacrylamide gel electrophoresis
PAML	Phylogenetic analysis by maximum likelihood
PCR	Polymerase chain reaction
PE	Paired-end
PFAM	Protein families
PHI	Pathogen-host interactions
qPCR	Quantitative-PCR
QRL	Quality region length
RAM	Random-access memory
RIP	Repeat-induced point
RNA	Ribonucleic acid
SBH	Sequencing by hybridization
SBL	Sequencing by ligation
SE	Single-end
SNA	Single-nucleotide addition
SNP	Single nucleotide polymorphism
STSs	Sequence-tagged sites
T	Thymine
UPTs	Unique putative transcripts
UTR	Untranslated region
YAC	Yeast artificial chromosome

Acknowledgements

They say it takes a community to raise a child and surely the same is true for a PhD candidate. I’ve been very fortunate for being raised amongst many excellent and dedicated scientists. First, my thanks for the people whom I’ve worked with directly while completing this research: Shelley Miller, Claire Oddy, Alex Plattner, Vincent Wang, Thomas Wang, Huang-Ju Chen, Jae Jin Kim, David Jack, Sangwon Lee, Remy Martin, Kristin Tangen, Deepa Pursuwaman, Uljana Hesse, Rob Kirkpatrick, Michael Seidel, Simon Chan, Rod Docking, Greg Taylor, Nancy Liao, Richard Varhol, Anthony Fejes, Sepideh Alamouti, Philippe Tanguay and Ye Wang. To everybody at the GSC, who make the GSC a fantastic place for doing research. To all the great folks who’ve developed the software used for analyzing data and especially to those who support it judiciously; with special thanks to Darren Platt. Thank you for the excellent mentorship I’ve received from: Gordon Robertson, Steven Ralph, Chris Keeling, Young Woon Lim and Martin Hirst. To Steven Jones, for providing key inputs at critical moments. To my committee members Renee Alfaro and Zamir Punja, who’ve been supportive and enthusiastic about my work. To my committee member and co-supervisor Joerg Bohlmann, for teaching me about being critical, prepared and strategic. To my supervisor, Colette Breuil who has supported, pushed and guided me with as much enthusiasm as any parent. To my friends for hanging in there. To my parents, who provided me with an environment that has allowed me to follow my dreams. To my wife and son, whom still love me despite the stress and long hours, my deepest respect, thanks and love.

To Sebastian and bean

The best is yet to come

Co-authorship statement

Chapter 2: SD designed the experiments and analysis. Experiments were performed by SD. cDNA library construction was carried out by SD with input from SR. EST sequencing was carried out under the direction of RAH. Website resources were developed under the direction of SJMJ. Data analysis was performed by SD with input from YWL. The manuscript was prepared by SD, JB and CB with assistance from SR.

Chapter 3: SD, NYL, RAH and SJMJ designed the analysis. Sanger sequencing was carried out under the direction of RAH. 454 sequencing was carried out under the direction of EM. Illumina sequencing was carried out under the direction of MH. Forge was developed by DP. Assemblies were performed by SD, NYL, MS, SKC and DP. Data analysis was performed by SD, SKC, TRD and NYL under the direction of IB. The manuscript was prepared by SD, CB, JB and SJMJ with assistance from GR.

Chapter 4: SD, SJMJ, JB and CB designed the analysis. Genome sequence finishing was performed by NYL, TRD, and GT under the direction of SD with input from IB and SJMJ. The genome annotation strategy was developed by SD. Genome annotations were performed by SD and SKC. RNA-seq experiments were performed by SD. Illumina sequencing was carried out under the direction of MH. RNA-seq experiments were analyzed by SD. Peptide sequencing experiments were designed by SD and PT. Peptide sequencing data analysis was performed by NF, PT and SD. Analysis of the MAT loci was performed by CT and NF. Figure 1 was created by SMA. Phylogenetic analysis was performed by SD. Additional EST resources were developed by SD and analysis was performed by UH. CAZy analysis was performed by BH. FOLy analysis was performed by AL. Transporter deletion mutants were prepared by YW. Growth experiments were performed by YW and SD. The manuscript was prepared by SD, SJMJ, JB and CB with assistance from GR.

Co-authors are abbreviated as: SD, William S. DiGuistini; SR, Stephen G. Ralph; YWL, Young W. Lim; SJMJ, Steven J.M. Jones; RAH, Robert A. Holt; JB, Jörg Bohlmann; CB, Colette Breuil; NYL, Nancy Y. Liao; EM, Elaine Mardis; MH, Martin Hirst; DP, Darren Platt; MS, Michael Seidel; SKC, Simon K. Chan; TRD, T. Rod Docking; IB, Inanc Birol; GR, Gordon Robertson; GT, Greg Taylor; PT, Philippe Tanguay; NF, Nicolas Feau; CT, Clement K. Tsui; SMA, Sepideh Massoumi Alamouti; UH, Uljana Hesse-Orce; BH, Bernard Henrissat; AL, Anthony Levasseur; YW, Ye Wang.

1 Creating genomic resources for a non-model organism

"Because the mechanisms of each trait of interest are manifested at lower levels of biological organization and the significance of a trait is only apparent at higher levels, understanding a given trait usually requires the simultaneous use of molecular, cellular, organismal, population and ecological approaches"

Martin Feder, in *New Directions in Ecological Physiology* Cambridge University Press 1987

1.1 Introduction

To date, the great scientific advances made in genomics apply to a relatively small number of biological systems and have been generated by a small number of relatively large organizations or consortiums.

New sequencing platforms, such as Roche's GS FLX and Illumina's GA, can process a bacterial or fungal genome sequence in a single day, and are generating sequences from an expanding list of organisms (Kyrpides, 2009). The decreasing costs and increasing throughput of DNA sequencing has prompted many researchers to start ambitious sequencing projects; however, the tools and methods for processing and interpreting this data are also changing rapidly.

The mountain pine beetle (MPB; *Dendroctonus ponderosae*) is one of the most important factors disturbing the management of lodgepole pine ecosystems. Due to the susceptibility of mature lodgepole pine stands to the recurring attacks by MPB, mature pine forests should not be left to stand for long periods, and managing timber flow and sustained yield is difficult. It is estimated that approximately 16 million hectares of lodgepole pine forest have been affected in B.C., where 900 million cubic meters of MPB-killed wood will have to be processed by the end of 2010 (www.for.gov.bc.ca/hfp/mountain_pine_beetle/).

While the MPB and its microorganism associates is both ecologically and economically significant, few molecular tools are available for studying this complex biological system. My research focuses on one of the MPB fungal associates, *Grosmannia clavigera,* previously reported as *Ophiostoma clavigerum* (Zipfel, de Beer, Jacobs, Wingfield, & Wingfield, 2006). The purpose of my work is to develop a set of genomics based resources and use them for performing a preliminary analysis of *G. clavigera's* tolerance for host-specific metabolites. This preliminary analysis focuses on the detoxification of the primary phloem chemical defenses of lodgepole pine: terpenes and MeOH:H_20 extractable metabolites. Because this fungus has a relatively compact genome for a eukaryote (~30 Mb) and belongs to the sordariomycetes, a class where early whole genome sequencing efforts were focused (Dean, et al., 2005; Galagan, et al., 2003), it is also a good system for developing methods that take advantage of the new sequencing platforms.

The biological understandings developed in the work described here begins to identify key genes and pathways that enable *Gc* host colonization. Variations in these should be characterized in populations over large geographic regions. In the long term, this research has potential to support development of new approaches for managing and controlling MPB outbreaks. Finally, the methods developed here highlight the possibilities for utilizing advanced sequencing technologies for studying non-model biological systems.

1.2 Genome sequencing

In the last 20 years much progress has been made towards improving the quality, speed and cost of genome sequencing; however, the sequencing and assembly of a eukaryotic genome remains a difficult task. In the past, the most commonly used approach for generating a genome sequence involved ‘shotgun’ sequencing and physically mapping individual clones. As computers, assembly algorithms and sequencing technologies have advanced, the strategies used for sequencing have changed; however, the essential reagents have remained the same.

To build a physical map, the genomic DNA must first be fragmented and cloned into a suitable vector. The primary vectors are: 1) phage (Olson, et al., 1986), 2) cosmid (Coulson, Sulston, Brenner, & Karn, 1986), 3) fosmid (Magrini, et al., 2004; Shizuya, et al., 1992), 4) BAC (Marra, et al., 1997; McPherson, et al., 2001; Shizuya, et al., 1992) and 5) YAC (Burke, Carle, & Olson, 1987; Schuler, et al., 1996). YACs contain cloned inserts larger than one megabase pair (Mb) in size and were used to construct the first-generation physical maps of the human and mouse genomes. However, for various reasons (Green, 2001; Venter, Smith, & Hood, 1996) YACs are no longer considered a good starting point for genome sequencing projects. By contrast, BACs, which carry 100-200 kb inserts, and fosmids, which carry 40 kb inserts, are extremely useful for constructing second-generation physical maps and/or for shotgun sequencing. Fosmids are ideal for constructing clone libraries and physical maps for genomes up to ~1 Gb. In the *G. clavigera* genome sequencing project presented in this thesis a fosmid cloning strategy was used.

The physical map is assembled from a large number of clones chosen at random from the clone library, by inferring overlaps between clones with sufficiently similar 'fingerprints'. The fingerprints are generated from unique DNA landmarks such as Sequence-Tagged Sites (STSs), restriction sites and/or other sequence-based elements. In a contemporary, high-throughput mapping pipeline informatic tools such as Image (www.sanger.ac.uk/Software/Image) and FPC (www.genome.arizona.edu/software/fpc/) are used for assembling the physical map. A selection of the overlapping clones — a minimum tiling path that spans the genome sequence — then becomes the foundation for sequencing. For each selected clone, the DNA is purified and subjected to random fragmentation and size fractionation; the DNA fragments in a defined size range (for example, 2-5 kb) are recovered and subcloned into a plasmid- or M13-based vector, or sequenced directly following a minimum number of preparatory steps as in 'next-generation' sequencing methods (NGS; i.e. contemporary sequencing methods not based on the technology developed by Fred Sanger)(Green, 2001).

Lander and Waterman (1988) provided a set of simple mathematical formulas that aid in planning a physical mapping project. They described how the expected number of mapped islands varies with the number of clones fingerprinted and the type of fingerprinting scheme used. It is important to note that these formulas also apply to shotgun sequencing, since the DNA sequence of the individual fragments can be considered the most detailed fingerprint.

Whole-genome shotgun sequencing involves the assembly of sequence reads generated in a random, genome-wide fashion. It was introduced by Sanger et al. (1977) and refined by Ansorge and colleagues (Edwards, et al., 1990). After shearing an organism's genomic DNA into defined size ranges, sequence reads are generated from both ends of the genome fragments (Fleischmann, et al., 1995). This approach produces highly redundant sequence coverage across the genome. A key aspect of the shotgun strategy is the generation of sequence reads from both fragment ends (paired-end sequencing, PE) with varying distances between the pairs. This is especially true for NGS, where abundant PE information is traded off against read length. The expected physical distances separating these PE reads are important and can often be used for determining the placement of ambiguous reads when repetitive genomic regions are encountered during assembly. The use of several size classes of DNA fragments is crucial, as each class has a different role to play in the assembly process, such as spanning repetitive sequence or providing long-range contig ordering (scaffolding). The main problem with the whole-genome shotgun strategy, especially when used in the absence of clone-derived mapping data, is the finishing of sequence gaps and misassemblies caused by repetitive sequences. Without contig ordering data and clones to work from it is difficult to determine the most effective finishing strategy.

The first eukaryotic whole-genome shotgun sequencing project involved *Drosophila* (Adams, et al., 2000). While virtually all euchromatic regions were assembled, achieving a high quality, ordered sequence required BAC-by-BAC finishing (Celniker, et al., 2002; Hoskins, et al., 2007). Eliminating the cloning step by sequencing the sheared DNA directly or with minimum preparation became practical with current NGS platforms (Li, et al., 2009) (see below). While these methods can be highly cost-effective for generating draft genome sequences, and not all organisms will require finishing to the 'comparative' or 'reference' levels (Blakesley, et al., 2004), a comprehensive finishing strategy that will yield such a sequence will likely still require clone-by-clone sequencing.

Whether a cloning or direct sequencing strategy is selected, a number of sequencing platforms, in addition to the Sanger/ABI platform, are now available. The variety of features they offer will likely ensure that multiple platforms will coexist as some have clear advantages for particular applications over others (Metzker, 2009). Sequencing methods are grouped broadly into categories of template preparation, sequencing and imaging, and data analysis. The unique combination of specific protocols distinguishes one technology from another and determines the type of data produced. The currently available platforms are: 1) Roche/454 (Margulies, et al., 2005), 2) Illumina/Solexa GA (Bentley, et al., 2008), 3) Life/APG SOLID (Valouev, et al., 2008), 4) Helicos BioSciences (Braslavsky, Hebert, Kartalov, & Quake, 2003) and the 5) Polonator (Shendure, et al., 2005). In addition, technology currently being developed by Pacific

Biosciences (Eid, et al., 2009; www.pacificbiosciences.com), and Ion torrent (www.iontorrent.com) will soon be available.

New sequencing chemistries are classified as cyclic reversible termination (CRT), single-nucleotide addition (SNA) and real-time sequencing (Metzker, 2009). In addition sequencing by ligation (SBL), an approach in which DNA polymerase is replaced by DNA ligase is also used. Two methods are used for preparing templates for NGS reactions: clonally amplified templates originating from single DNA molecules, and single DNA-molecule templates. Clonal amplification is necessary because most imaging systems cannot detect fluorescent events that reflect a single base addition to a single DNA fragment, so amplified templates are required. Although clonally amplified methods offer advantages over bacterial cloning in regards to sequence biases, additional steps in the protocol add complexity and require large amounts of starting material (3-20 μg of DNA). The preparation of single-molecule templates is more straightforward and requires less starting material (<1 μg). More importantly, these methods do not require PCR. PCR can create mutations in clonally amplified templates, and has AT- and GC-rich amplification biases that can result in under-representation of some genomic regions. Together, these factors can produce misleading and spurious genomic assemblies. A recent single nucleotide sequencing technology, developed by the company Ion Torrent (www.iontorrent.com) departs from fluorescence based detection methods by detecting pH changes resulting from the liberation of H+ ions that occur during nucleotide incorporation by the polymerase.

Although quality scores and accuracy estimates are provided by each manufacturer, there is no consensus that a high quality base from one platform is equivalent to one from another platform. Each platform has its unique sequencing problems. For instance, during Illumina CRT sequencing, substitutions are the most common error type, with a higher portion of errors occurring when the previous incorporated nucleotide is a ‘G’ base (Dohm, Lottaz, Borodina, & Himmelbauer, 2008; Harismendy, et al., 2009; Hillier, et al., 2008). These problems are usually not evident from the inspection of quality scores (personal observation). The 454's lack of cloning bias and ability to sequence through regions of the genome that exhibit strong secondary structure provide examples of the advantages of combining multiple technologies; however, the 454 platform has a well known propensity for misreporting bases in the regions surrounding homopolymer runs (Huse, Huber, Morrison, Sogin, & Welch, 2007). The strategy developed for sequencing the genome of *G. clavigera* described in this thesis included Sanger, 454 and Illumina sequencing.

Perhaps the most essential element for any whole-genome shotgun-sequencing strategy is a robust assembly program. Such a program should accommodate the inevitably large collection of sequence reads derived from many Gb of genomic DNA, and the varying error models that arise from combining read data from multiple sequencing platforms.

The ‘overlap-layout-consensus’ (OLC) method for assembling genomes is used by many currently available assemblers (Batzoglou, 2002; Bonfield, Smith, & Staden, 1995; Green, 1995; Huang & Madan, 1999; Sutton, White, Adams, & Kerlavage, 1995), and was the standard for all eukaryotic genome sequencing projects until the publication of the Giant Panda genome sequence (Li, et al., 2009). The calculation is described by Pevzner and Tang (2001) as a sequence string problem. Given a set of strings $S=\{s_i,...,s_n\}$, the solver tries to find the shortest string s such that each s_i appears as a substring of s. This difficult calculation was thought to limit the application of shotgun sequencing (Green, 2001). The introduction of heuristics and sophisticated algorithms (Myers, 1995) increased the size and complexity of problems that could be solved. This, and improvements in computing systems allowed OLC to remain the dominant method for assembling genomes, until NGS technologies became important. The implementation of homopolymer run collapses in the Celera assembler illustrates how algorithms become increasingly complex when multiple sequence technologies are used (Miller, et al., 2008). One proposed solution for reducing the impact of incorrectly called homopolymer runs on a genome sequence assembly utilizing 454 read data is to collapse these runs prior to identifying overlaps and then expand them following read layout. This example demonstrates how OLC is adapting to evolving sequencing technologies; however, all the strengths of NGS play to the weaknesses of the OLC method.

In general, the massive quantity of data, the short lengths of the reads and the higher error rates produced by NGS sequencing platforms amplify the cumbersome and complicated algorithms used by most OLC assemblers. To assemble the *G. clavigera* genome we used the assembler Forge (personal communication Darren Platt). Forge uses distributed memory hash tables and pruned overlap graphs, and is currently the only OLC assembly tool that is able to process realistic quantities of NGS read data. Nevertheless, to assemble the *G. clavigera* genome sequence Forge required 10-84 hours on a server cluster using 40 nodes that ranged from dual 2.0 GHz processors with 2 GB of RAM to quad-core 2.6 GHz processors with 16 GB of RAM. In contrast, a de Bruijn graph based assembler (see below) accomplished the same task in approximately three hours on a standard server with two, 2.2 GHz dual-core AMD Opteron 275, processors and 8 GB of RAM.

Using a Eulerian method for solving the genome assembly problem arose from concepts developed while working with Sequencing By Hybridization (SBH; Idury & Waterman, 1995). SBH assembly is similar to fragment assembly except that SBH, a progenitor of microarrays generates fixed-length ‘read’ data. Working with fixed length reads provides a computational benefit as it reduces the problem complexity (See Pevzner & Tang, 2001 for additional explanation). The Eulerian method places sequence reads or fragments of reads (k-mers) in a directed graph. In this graph, each node represents a read or fragment, and edges represent overlaps between reads. The resulting DNA sequence can then be inferred from walking along the edges of the graph (a path), as described by Pevzner, who first implemented these concepts in the software EULER (Pevzner, Tang, & Waterman, 2001). The primary challenge of this apparently simple solution arises from sequencing errors and repeats. In this situation the graph becomes too complex to be resolved simply, and many recent implementations of this method have attempted to resolve the graph in more sophisticated ways (Butler, et al., 2008; Simpson, Wong, Jackman, & Schein, 2009; Zerbino & Birney, 2008). As with the OLC method, large quantities of NGS data are challenging to manage, and so tracking read-to-k-mer relationships is not possible. In the latest de Brujin assemblers, k-mers generated from sequence reads become abstractions. This resolves the above problem, but prevents enforcing Eulerian paths that are consistent with either read paths or PE read relationships, so PE reads are used in a second stage of assembly, in which contigs are joined using read- or k-mer-to-contig alignments.

1.3 Genome annotation

After determining an organism's genomic sequence, the next task is locating the protein coding genes. This involves: 1) delimiting the gene regions, 2) defining the intron-exon boundaries, 3) locating the translation initiation and termination sites and 4) predicting the flanking UTR regions. Deeper analyses, often combined with additional experimental data, are then performed to more carefully identify the transcription start and stop sites and catalog alternative splicing, non-coding RNAs and small RNAs. Finally, quantifying how expression levels change under different conditions for these genome annotations adds an additional data dimension and leads to biological insight. A large literature exists describing these computational problems and many tools have been created for generating and analyzing these genome annotations (Brent, 2007; Cochrane & Galperin, 2010; Coghlan, et al., 2008). To develop and implement a genome annotation strategy it is necessary to have some idea of what the annotations should look like. This is important when training the gene predictors and assessing the results of computational analysis.

Fungi possess compact gene structures. Analyses of the currently sequenced fungal genomes indicate that protein coding gene collections range from ~4-25 K (Dietrich, 2004; Martin, et al., 2008). Coding sequence lengths average between 1.3 and 1.9 kb. Intron densities range from < 300 introns across the complete gene collection for the hemiascomycetes to 6 introns/gene in some basidiomycetes. The ascomycetes for which complete genome sequences have been generated have an average of ~1.5 introns/gene (personal observations). Average intron lengths are typically short, ranging from 68-150 bp, although examples exceeding 3000 bp have been observed (personal observation). Alternative splicing occurs in fungi and is estimated to impact 3-10 % of genes, this is a significantly lower level of splicing than in humans, where 40-80 % of genes have alternative isoforms (Chen & Manley, 2009). Although fungal gene models tend to have fewer introns and less splicing than their mammalian counterparts, there is a large diversity of gene structures. While simple gene structures facilitate accurate gene predictions, the large diversity in gene models means that organism-specific data is very important and effective training is crucial for gene predictors. Of course, the best predictions still cannot match a manually curated and validated collection of gene models (Coghlan, et al., 2008); however, detailed analysis such as this is labour intensive and requires well-trained technicians. For many projects this is simply not affordable.

One method for annotating a genome is to use a hybrid approach (personal communication Jason Stajich, UC Riverside). This approach integrates experimental (EST and RNA-seq; see below), comparative (proteins from the Swissprot database), and computational (gene prediction software) data. The method essentially involves four steps: 1) running gene prediction programs using best available parameters, 2) aligning protein and transcript data to the genome, 3) combining predictions into a set of composite gene models using the software GLEAN (Elsik, Mackey, Reese, Milshina, & al., 2007), 4) training the gene prediction software using the first set of consensus genes. Repeating steps 1 and 3 using refined prediction parameters yielding a final collection of predicted genes.

Expressed Sequence Tags (ESTs) are single-pass sequence reads generated from cloned cDNA fragments that reflect the relative abundances of expressed genes. Due to the cost of building cDNA libraries and sampling them deeply, it is not possible to generate enough read data for inferring gene expression levels reliably. As well, they typically only cover a portion of a transcript, and are prone to sequencing errors that cannot be compensated for by sequencing depth. As reagents for annotating the genome they also offer benefits such as 1) relatively long lengths, 2) UTR coverage, and 3) strand orientation (current RNA-seq protocols, see below, do not provide strand orientation). The ~600-800 bp of sequence generated in a single pass can provide reliable evidence for joining multiple exons (including non-adjacent exons).

RNA-seq is a relatively new approach for studying expressed genes in a genome wide fashion utilizing NGS platforms (Pepke, Wold, & Mortazavi, 2009; Wang, Gerstein, & Snyder, 2008). The method is relatively straightforward: a population of RNA (total or fractionated, such as poly(A)+) is converted to a library of cDNA fragments with adaptors attached to one or both ends. Each molecule, with or without amplification, is then sequenced on a high-throughput sequencing platform to obtain short sequences from one end (SE sequencing) or both ends (PE sequencing). The read lengths depend on the sequencing technology used. In principle, any high-throughput sequencing technology can be used for RNA-Seq. Currently, the Illumina GA , Applied Biosystems SOLiD (Cloonan & Grimmond, 2008), Helicos (Lipson, et al., 2009) and Roche 454 (Vera, et al., 2008) are being used. After sequencing, reads are either aligned to a reference genome or transcript collection (Wilhelm, Marguerat, Goodhead, & Bahler, 2010), or assembled *de novo* (Birol, et al., 2009) to produce a genome-scale transcription map consisting of both the transcriptional structure and/or level of expression for each gene (see SEQanswers: http://seqanswers.com/). Some reads may also span exon junctions or contain poly(A) ends and these can not be analyzed in the same way; however, methods have been developed for mapping these reads specifically (Wilhelm, et al., 2010). RNA-seq is replacing microarrays in the study of gene expression. It can: 1) be used for identifying and quantifying rare transcripts without prior gene knowledge, 2) provide information regarding sequence variants in the expressed transcripts and 3) be more sensitive to differentiating the expression of sequences sharing a high degree of similarity.

1.4 The Mountain pine beetle system

1.4.1 Mountain pine beetle life cycle

The MPB infests approximately ten different host tree species belonging to the genus *Pinus* including lodgepole pine (*Pinus contorta*) in pure or mixed pine forest ecosystems. During the endemic phase (less than 10 trees attacked per hectare) MPB colonizes weakened or dying trees in association with other beetles (‘near-obligate parasite’; Raffa, Phillips, & Salom, 1993), during the transition and in an epidemic (outbreak phase) the beetles attack healthy mature trees (Carroll, et al., 2009; Raffa, 2001). The MPB life cycle begins when beetles disperse during summer. The timing of beetle flight seems best correlated with temperature (Safranyik, 1978). The female insect bores through the bark, makes a vertical gallery in the phloem, mates and deposits in average between 40-60 eggs along the gallery wall. Eggs are laid singly and larvae emerge approximately one or two weeks later. The larvae feed in the phloem and mine galleries perpendicular to the main gallery. Generally, overwintering occurs in the second or third larval instar before cold temperatures cause winter dormancy. Larvae that survive the winter begin feeding again in April, completing their development into the fourth instar. Larvae pupate within the chambers excavated during their development. Pupae transform into adults during the early summer season from late June to early-September (Amman, 1978; Allan Carrol, CFS Canada, personal communication). At least three main factors are responsible for the MPB’s recent expansion: 1) food supply (with high concentrations of mature pines available), 2) climate changes (mild winters are responsible of low brood mortality and dryer summers increase stress on trees), and finally 3) forest management practices such as extensive fire suppression.

1.4.2 Microorganisms associated with the mountain pine beetle

A critical component of the MPB system is the 'complex' of pathogenic and saprophytic fungi vectored by the beetles (Paine, Raffa, & Harrington, 1997). Beetle surfaces and galleries host diverse microflora including, bacteria, yeasts, filamentous ascomycetes (including staining fungi) and basidiomycetes (Cardoza, Vasanthakumar, Suazo, & Raffa, 2009; Kim, Allen, Humble & Breuil, 2005; Lim, Kim, Lu, & Breuil, 2005). Some of these microorganisms have symbiotic associations with this beetle and are not found on any other beetles; They are present on the exoskeleton, gut and mycangia of MPBs. The most abundant fungal associates of MPB belong to the heterogenous group known as 'ophiostomatoid fungi' included in the genera *Grosmannia, Ophiostoma* and *Ceratocystiopsis*. The two primary associates are *Grosmannia clavigera* (Robinson-Jeffrey and Davidson) Zipfel de Beer & Wingfield and *Ophiostoma montium* (Rumbold) von Arx (Robinson, 1962; Rumbold, 1941; Six, 2003; Solheim, 1995; Zipfel, et al., 2006). These fungi not only grow in the beetle galleries, but also in the tree phloem and sapwood, which they discolour. More recently, two additional fungal associates of the MPB have been reported (Lee, Kim, & Breuil, 2005; Plattner, et al., 2009). The pathogenic *Leptographium longiclavatum* Lee, Kim & Breuil (Lee, Kim, & Breuil, 2005) which grows in and stains the sapwood of lodgepole pine subsequently killing it (Lee, Kim, & Breuil, 2006a), and a non pathogenic slow growing *Ceratocystiopsis* sp. that mainly colonizes beetle galleries.

Poorly understood factors govern the frequencies at which these fungal associates are found on the beetles, in the galleries and extended into the phloem and sapwood. *O. montium* is isolated more frequently than *G. clavigera* from beetle exoskeletons (Lee, Kim, & Breuil, 2006b; Robinson, 1962; Six, 2003a), and this is also sometimes true of *Ceratocystiopsis* sp. (Lee, Kim, & Breuil, 2006b). It is likely that the frequency of isolation varies with factors such as the beetle life cycle, environmental variations (e.g. temperature), and beetle population densities (especially in relation to host availability). As an example of environmental variation, Six and Bentz (2007) assessed variations in the fungal-beetle relationship by measuring the frequencies of association for *G. clavigera* and *O. montium* with MPB for different ambient temperatures. They found higher frequencies of *O. montium* at high temperatures. This could result in seasonal (temporal) or latitudinal (spatial) stratifications in the fungal population. Such phenomenon may decrease competition by reducing the amount of niche overlap. This observation is consistent with the physiological observation that *O. montium* has a higher temperature optima than *G. clavigera* (Six & Bentz, 2007) in vitro.

Little is known about how virulence and aggressiveness may vary within a species and in relation to other microorganisms vectored by the beetle. *G. clavigera* can kill trees in the absence of beetles (Yamaoka, Hiratsuka, & Maruyama, 1995). In contrast, inoculation of lodgepole pine with *O. montium* does not lead to tree death (Yamaoka, et al., 1995; personal observation). Rice, Thormann & Langor (2007) observed that six weeks after inoculation *G. clavigera* induced longer lesion lengths in Jack pine (*Pinus banksiana*) than on the primary MPB host, lodgepole pine. *G. clavigera* also induced longer lesion lengths than *O. montium* in lodgepole and hybrid lodgepole x Jack pines, whereas in Jack pine the induced lesion lengths were similar for both fungi (Rice, Thormann & Langor 2007). Plattner et al. (2008) inoculated a small number of *G. clavigera* strains into lodgepole pine and observed variations between the strains for several measurements of tree damage.

The symbiotic relationship between the beetle and its fungal complex is critical for understanding the biology of the MPB. While beetles mine the host phloem under the bark, aggressive fungi such as *G. clavigera* propagate from beetle galleries and penetrate the underlying sapwood. The fungi benefit because beetles carry them through the tree bark, making available a fresh, moist, nutrient-rich wood environment that contains few or no competing microorganisms. The benefits to the beetle and beetle progeny are less clear. There is evidence and speculation that fungi 1) make nutrients available (particularly nitrogen and sterol) for beetles to complete their life cycle, 2) make the attacked environment more favorable for the beetles by detoxifying host defense metabolites and lowering the wood moisture content, 3) instigate and weaken the tree defenses (Lieutier, Yart, & Salle, 2009). This last point might be particularly important during the switch from endemic to epidemic conditions because fungal associates lower the critical threshold for beetle attacks required to overwhelm the tree defenses.

1.4.3 Introduction to conifer defenses

Conifers possess a complex chemical defense system that can be deployed against a wide range of pests and pathogens. The system has both constitutive and inducible elements composed from metabolites such as terpenes and phenolics. Preformed defense compounds are an immediate deterrent and serve to protect host tissues from primary colonizers. Major components of this preformed defense system are oleoresin, primarily synthesized in specialized anatomical structures of the phloem and xylem (Keeling and Bohlmann, 2006), and phenolics potentially synthesized and stored in polyphenolic parenchyma cells (PP; Franceschi, Krekling, Berryman, & Christiansen, 1998). Oleoresin is primarily composed of monoterpenes and diterpenes with small amounts of oxygenated derivatives (terpenoids), sesquiterpenes and other metabolites such as the phenylpropanoid, 4-allylanisole (Gambuel, Caves, Caffey-Moquin, & Paine, 1985; Joseph, Kelsey, Peck, & Niwa, 2001; Keeling & Bohlmann, 2006; Trapp & Croteau, 2001).

Elicitation by bark beetles, fungi or by wounding the stem of *Pinus* spp. can produce secondary responses that include increased oleoresinosis, tissue necrosis, and the development of a wound periderm surrounding the wound site (Hudgins, Christiansen, & Franceschi, 2003; Lieutier & Berryman, 1988). Modified cell walls and development of the wound periderm immediately adjacent to the beetle-fungal site of invasion can be reinforced through the deposition of phenolics, pectins and lignin (Bonello, Pearce, Watt, & Grime, 1991; Hammerschmidt, 1999). In some cases, traumatic resin ducts are formed and phenolic bodies in the secondary phloem appear larger and/or denser (Hudgins, Christiansen, & Franceschi, 2003). Induced defense reactions increase the abundances and may change the composition of the constitutive defenses described above.

Although tree phytochemistry plays a critical role in host selection by bark beetles (Raffa, 2001), it also governs the rate at which fungal spores deposited by the beetle will germinate and grow. In general, both terpenes and phenolics have been shown to have deleterious effects on microorganisms (Delorme & Lieutier, 1990; Himejima, Hobson, Otsuka, Wood, & Kubo, 1992; Hofstetter, Mahfouz, Klepzig, & Ayres, 2005; Savluchinske Feio, Gigante, Roseiro, & Marcelo-Curto, 1999) and bark beetles (Kopper, Illman, Kersten, Klepzig, & Raffa, 2005; Paine, et al., 1997). However, it is important to recognize that testing purified substances can result in misleading observations because of the physiological variations between individuals, chemical states that may vary depending on physicochemical factors leading to inappropriate test conditions, synergistic effects between metabolites that exist naturally as a complex mixture are lost in vitro, difficulty replicating biologically meaningful metabolite concentrations, and inadvertent solvent effects on insects and fungi.

Monoterpenes can be both fungistatic and fungicidal. Because of their hydrophobic nature it is typically expected that they interact with membranes and membrane bound enzymes (Sikkema, de Bont, & Poolman, 1994; Uribe, Ramirez, & Pena, 1985). Although this is likely an oversimplification considering the enormous diversity of terpenes and terpenoids in nature (Keeling & Bohlmann, 2006). Differences in activity arise from the differences in structures (skeletons), level and types of substitutions as well as differences between stereoisomers (Gershenzon & Dudareva, 2007). In general, studies of monoterpene antimicrobial activities have reported that oxygen containing functional groups, especially alcohols increase their toxicity; however, the addition of polar groups may also increase their rate of detoxification making them more amenable for enzymatic modifications or transport out of the cell (Dhar, Ayala, Andarge, Morisseau, & Snyder-Leiby, 2004; Knobloch, Weigand, Weis, & Brunke, 1986; Naigre, Kalck, Rogues, Roux, & Michel, 1996; Zukerman, 1951). In yeast, treatment with terpinolene, lead to gene expression changes in stress response genes, the ergosterol pathway, phospholipid biosynthesis, and cell wall organization (Parveen, et al., 2004).

Shrimpton and Whitney (1968) first investigated the effects of oleoresin on the growth of MPB associated blue stain fungi. They found that *G. clavigera* and *O. montium* have different responses to host metabolites. *O. montium* growth rate decreased with increasing oleoresin volatiles, whereas the volatile fraction had a fungistatic effect on *G. clavigera* (*Europhium*), delaying growth but not slowing it significantly. Paine and Hanlon (1994) assessed the toxicity of monoterpenes by adding them to the media or saturating them in the headspace for two strains of *G. clavigera* and one strain of *Leptographium terebrantis*. They measured a reduction in mycelial growth for some monoterpenes and surprisingly an increase in growth rate for other monoterpenes. By examining two host specific fungi and one host non-specific fungus they were testing for a host-specific metabolite relationship; however, they did not observe any host-pathogen specific relationships.

Phenolics have several modes of action that arise from differences in their oxidative state (e.g. protonated, ionized or oxidized). Factors affecting the physicochemical environment therefore influence their activity. Oxidative activation is the most common mode (Appel, 1993). Phenolic oxidation leads to the formation of quinones and free radicals that can inactivate fungal enzymes (Cowan, 1999). These actions can be direct through phenolic binding or indirect through free radical inactivation. Differences in activity arise from the differences in structural classes (e.g. hydrolyzable vs. condensed tannins), levels and kinds of substitutions as well as the differences between stereoisomers (Appel, 1993; Harborne, 1988; Harborne, 1991; Nicholson & Hammerschmidt, 1992).

In the MPB system there is currently no published work that I am aware of assessing the effects of lodgepole pine phenolics on the growth of any MPB-associated microorganisms. In this thesis, we developed a lodgepole pine phloem extract (LPPE) treatment. However, due to the uncertainty in the composition of this complex mixture and the uncertainty in the biologically active metabolites inducing *G. clavigera* gene expression we don’t currently consider it a proxy for phenolic treatment. In other tree-beetle-fungal systems, results reported have been inconsistent; compounds that appeared promising in the field did not provide in vitro evidence for their biological importance (Lieutier, et al., 2003). The divergent results from in vitro tests indicate that the antifungal toxicity of phenolics is specific to the compound-fungal interaction and that system-specific information is important for such work.

1.4.4 Pathogenic strategies for overcoming host defenses

In the struggle to colonize a host, pathogen strategies may involve degradation or conversion of toxic host molecules, transport of host defense molecules out of the cell, modification of cell structures to avoid/exclude/sequester toxic host molecules, and suppression or modification of host defense signaling (Katsir, Chung, Koo, & Howe, 2008). The pathogen likely uses a combination of these strategies.

Cellular detoxification often begins with oxidative, reductive or conjugating events that transform host chemicals into less toxic and more easily excreted metabolites (excretion can be both active and passive).However, the oxidation, N-acetylation and sulfate or glutathione conjugation of defense metabolites can also lead to metabolites with greater toxicity and mutagenic potential than the original host chemicals (Goldstone, et al., 2006).

To achieve transformations of host defense metabolites, it is unknown whether the *G. clavigera* strategy relies on a limited number of modifying genes with broad substrate specificities or on a large number of genes with narrow substrate specificities. Assessing the extent to which *G. clavigera* has evolved specialized enzymes for detoxifying lodgepole pine defense metabolites would increase our understanding of the mechanisms underlying this host-pathogen relationship.

Glycoside hydrolases (GHs) are a widespread group of enzymes that hydrolyse the glycosidic bond between two or more carbohydrates or between a carbohydrate and a noncarbohydrate moiety (www.cazy.org/). It has been demonstrated that GHs are important for detoxifying sugar-conjugated antimicrobial chemicals such as phenolics and saponins in fungi (Bouarab, Melton, Peart, Baulcombe, & Osbourn, 2002; Pareja-Jaime, Roncero, & Ruiz-Roldán, 2008; Zheng & Shetty, 2000). In fungi, GHs range in number from ~50-350 enzymes belonging to ~115 different sequence-based families (personal communication, Bernard Henrissat). Work on saponin detoxifying GHs has established that these enzymes typically belong to CAZy family 3. Work on phenolic detoxification in organisms such as the forest fungus *Lentinula edodes* has been primarily biochemical in nature, and has not yet established a functional identification to the nucleotide level (D'Annibale, Casa, Pieruccetti, Ricci, & Marabottini, 2004; Makkar, Tsuneda, Tokuyasu, & Mori, 2001).

Cytochrome P450s (CYP450) are important for the transformation and degradation of many environmental chemicals and phytotoxins (van den Brink, van Gorcom, van den Hondel, & Punt, 1998). In fungi, the number of CYP450 genes ranges from 3 to 350 (p450.riceblast.snu.ac.kr); the 150 CYP450s in *Phanerochaete chrysosporium* are divided into 12 families and 23 sub-families (Doddapaneni, Chakraborty, & Yadav, 2005). Fungi that are more closely related to *G. clavigera,* such as *Neurospora crassa* and *Magnaporthe grisea* are predicted to have 41 and 123 CYP450s, respectively (drnelson.utmem.edu/cytochromeP450.html). It is suggested that the higher number of CYP450s in *M. grisea* in contrast to the lower number in *N. crassa* is related to *M.*

grisea's pathogenic lifestyle (Deng, Carbone, & Dean, 2007). CYP450s are likely candidates for adding polar groups to highly insoluble aliphatic terpenes, and so increasing their cytosolic solubility and facilitating their degradation. This is likely analogous to the well-described degradation of alkanes by the yeast *Yarrowia lipolytica* (Fickers, et al., 2005).

Oxidation is often followed by reductive or conjugative modifications. These reactions are performed by a diverse range of enzymes such as glutathione-S-transferases, sulfotransferases, UDP-glucuronosyl transferases, N-acetyl transferases, aldo-keto reductases, epoxide hydrolases and NAD(P)H-quinone oxidoreductases (Goldstone, et al., 2006). These enzymes have not received as much attention as CYP450s, GHs or ABC transporters (described below) with regard to detoxification of host defense metabolites in fungi. However, work in humans and other animals provide relevant information for work in this area (Balogh, Roberts, Shireman, Greene, & Atkins, 2008).

Since the most ubiquitous form of xenobiotic resistance observed in microorganisms involves the over-expression of membrane transporters, especially those of the ATP-binding cassette (ABC) membrane transport family, tolerance to lodgepole pine terpenes (and phenolics) may involve their activity. The roles of ABC transporters in plant-pathogen interactions have been described for *Botrytis cinerea* (Stefanato, et al., 2009), *Mycosphaerella graminicola* (Zwiers, Stergiopoulos, Gielkens, Goodall, & De Waard, 2003)*, Gibberella pulicaris* (Fleissner, Sopalla, & Weltring, 2002) and *M. grisea* (Sun, Suresh, Deng, & Naqvi, 2006; Urban, Bhargava, & Hamer, 1999).

1.5 Overview and purpose of this thesis

In this thesis, we developed a genomic resource for the mountain pine beetle fungal associate, *Grosmannia clavigera.* We used this resource for conducting a preliminary investigation into the physiological and chemical interactions between the fungus and host metabolites involved in tree defenses. We generated genomic resources including: a) the whole genome sequence, b) ESTs describing different fungal strains and environmental conditions, and c) RNA-seq data describing fungal gene expression when the fungus was exposed to tree defense chemicals such as terpenes and phenolics. Then, we annotated the fungal genome using these genomic resources (e.g. ESTs and RNA-seq) and computational (gene prediction software) data. Finally, we provide a preliminary evaluation of the completeness of *G. clavigera's* genome as well as the first biological insights into how this fungus tolerates host defense metabolites, and so colonizes and ultimately kills the host.

This thesis is organized into a series of journal manuscripts that are presented as independent chapters. Each chapter includes an introduction, results, discussion and a reference section. The introduction and the conclusions follow the standard thesis requirements.

1.6 References

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., et al. (2000). The genome sequence of *Drosophila melanogaster*. *Science, 287*(5461), 2185-2195.

Amman, G. (1978). Biology, Ecology, and Causes of Outbreaks of the Mountain Pine Beetle in Lodgepole Pine Forests. In: A. AA Berryman, G. D., Stark, R. W., (Ed.), *Theory and Practice of Mountain Pine Beetle Management in Lodgepole Pine Forests* (pp. 39-53).

Appel, H. (1993). Phenolics in ecological interactions - the importance of oxidation. *J Chem Ecol, 19*(7), 1521-1552.

Balogh, L. M., Roberts, A. G., Shireman, L. M., Greene, R. J., & Atkins, W. M. (2008). The stereochemical course of 4-hydroxy-2-nonenal metabolism by glutathione S-transferases. *J Biol Chem, 283*(24), 16702-16710.

Batzoglou, S. (2002). ARACHNE: A Whole-Genome Shotgun Assembler. *Genome Research, 12*(1), 177-189.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature, 456*(7218), 53-59.

Birol, I., Jackman, S. D., Nielsen, C. B., Qian, J. Q., Varhol, R., Stazyk, G., et al. (2009). De novo transcriptome assembly with ABySS. *Bioinformatics, 25*(21), 2872-2877.

Blakesley, R. W., Hansen, N. F., Mullikin, J. C., Thomas, P. J., Mcdowell, J. C., Maskeri, B., et al. (2004). An intermediate grade of finished genomic sequence suitable for comparative analyses. *Genome Res, 14*(11), 2235-2244.

Bonello, P., Pearce, R., Watt, F., & Grime, G. (1991). An induced papilla response in primary roots of scots pine challenged in vitro with *Cylindrocarpon destructans*. *Physiological and Molecular Plant Pathology, 39*(3), 213-228.

Bonfield, J. K., Smith, K., & Staden, R. (1995). A new DNA sequence assembly program. *Nucleic Acids Res, 23*(24), 4992-4999.

Bouarab, K., Melton, R., Peart, J., Baulcombe, D., & Osbourn, A. (2002). A saponin-detoxifying enzyme mediates suppression of plant defenses. *Nature, 418*(6900), 889-892.

Braslavsky, I., Hebert, B., Kartalov, E., & Quake, S. R. (2003). Sequence information can be obtained from single DNA molecules. *Proc Natl Acad Sci U S A, 100*(7), 3960-3964.

Brent, M. R. (2007). How does eukaryotic gene prediction work? *Nat Biotechnol, 25*(8), 883-885.

Brignolas, F., Lieutier, F., Sauvard, D., Christiansen, E., & Berryman, A. (1998). Phenolic predictors for Norway spruce resistance to the bark beetle *Ips typographus* (Coleoptera : Scolytidae) and an associated fungus, *Ceratocystis polonica*. *Can J For. Res., 28*(5), 720-728.

Burke, D. T., Carle, G. F., & Olson, M. V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. *Science, 236*(4803), 806-812.

Butler, J., Maccallum, I., Kleber, M., Shlyakhter, I. A., Belmonte, M. K., Lander, E. S., et al. (2008). ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Research*, 12.

Cardoza, Y. J., Vasanthakumar, A., Suazo, A., & Raffa, K. F. (2009). Survey and phylogenetic analysis of culturable microbes in the oral secretions of three bark beetle species. *Entomologia Experimentalis et Applicata, 131*(2), 138-147.

Carroll, A. L., Aukema, B. H., Raffa, K. F., Linton, D. A., Smith, G. D., & Lindgren, B. S. (2009). Mountain pine beetle outbreak development: the endemic - Incipient epidemic transition. Mountain Pine Beetle Initiative, Canadian Forest Service, Working Paper Project 1.03, pp 1-27.

Celniker, S. E., Wheeler, D. A., Kronmiller, B., Carlson, J. W., Halpern, A., Patel, S., et al. (2002). Finishing a whole-genome shotgun: release 3 of the *Drosophila melanogaster* euchromatic genome sequence. *Genome Biol, 3*(12), RESEARCH0079.

Chen, M., & Manley, J. L. (2009). Mechanisms of alternative splicing regulation: insights from molecular and genomics approaches. *Nat Rev Mol Cell Biol, 10*(11), 741-754.

Cloonan, N., & Grimmond, S. M. (2008). Transcriptome content and dynamics at single-nucleotide resolution. *Genome Biol, 9*(9), 234.

Cochrane, G. R., & Galperin, M. Y. (2010). The 2010 Nucleic Acids Research Database Issue and online Database Collection: a community of data resources. *Nucleic Acids Research, 38* (Database), D1-D4.

Coghlan, A., Fiedler, T. J., Mckay, S. J., Flicek, P., Harris, T. W., Blasiar, D., et al. (2008). nGASP - the nematode genome annotation assessment project. *BMC Bioinformatics, 9*, 549.

Collins, C. M., Murray, P. G., Denman, S., Morrissey, J. P., Byrnes, L., Teeri, T. T., et al. (2007). Molecular cloning and expression analysis of two distinct beta-glucosidase genes, bg1 and aven1, with very different biological roles from the thermophilic, saprophytic fungus *Talaromyces emersonii*. *Mycol Res, 111*, 840-849.

Coulson, A., Sulston, J., Brenner, S., & Karn, J. (1986). Toward a physical map of the genome of the nematode *Caenorhabditis elegans*. *Proc Natl Acad Sci U S A, 83* (20), 7821-7825.

Cowan, M. M. (1999). Plant products as antimicrobial agents. *Clin Microbiol Rev, 12*(4), 564-582.

D'Annibale, A., Casa, R., Pieruccetti, F., Ricci, M., & Marabottini, R. (2004). *Lentinula edodes* removes phenols from olive-mill wastewater: impact on durum wheat (Triticum durum Desf.) germinability. *Chemosphere, 54*(7), 887-894.

Dean, R., Talbot, N. J., Ebbole, D., Farman, M. L., Mitchell, T. K., Orbach, M. J., et al. (2005). The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature, 434*(7036), 980-986.

Delorme, L., & Lieutier, F. (1990). Monoterpene Composition of the Preformed and Induced Resins of Scots Pine, and Their Effect on Bark Beetles and Associated Fungi. *European Journal of Forest Pathology, 20*(5), 304-316.

Deng, J., Carbone, I., & Dean, R. (2007). The evolutionary history of cytochrome P450 genes in four filamentous Ascomycetes. *BMC Evol Biol, 7*, 30.

Dhar, P., Ayala, U., Andarge, E., Morisseau, S., & Snyder-Leiby, T. (2004). Study of the structural changes on the antimicrobial activity of [3.1.1]-bicyclics. *Journal of Essential Oil Research, 16*(6), 612-616.

Dietrich, F. (2004). The *Ashbya gossypii* genome as a tool for mapping the ancient *Saccharomyces cerevisiae* genome. *Science, 304* (5670), 518-518.

Doddapaneni, H., Chakraborty, R., & Yadav, J. S. (2005). Genome-wide structural and evolutionary analysis of the P450 monooxygenase genes (P450ome) in the white rot fungus *Phanerochaete chrysosporium*: evidence for gene duplications and extensive gene clustering. *BMC Genomics, 6*, 92.

Dohm, J. C., Lottaz, C., Borodina, T., & Himmelbauer, H. (2008). Substantial biases in ultra-short read data sets from high-throughput DNA sequencing. *Nucleic Acids Research*, 11.

Edwards, A., Voss, H., Rice, P., Civitello, A., Stegemann, J., Schwager, C., et al. (1990). Automated Dna Sequencing Of The Human Hprt Locus. *Genomics, 6*(4), 593-608.

Eid, J., Fehr, A., Gray, J., Luong, K., Lyle, J., Otto, G., et al. (2009). Real-time DNA sequencing from single polymerase molecules. *Science, 323*(5910), 133-138.

Elsik, C., Mackey, A., Reese, J., Milshina, N., et al. (2007). Creating a honey bee consensus gene set. *Genome Biol*.

Fickers, P., Benetti, P. H., Wache, Y., Marty, A., Mauersberger, S., Smit, M. S., et al. (2005). Hydrophobic substrate utilisation by the yeast *Yarrowia lipolytica*, and its potential applications. *FEMS Yeast Res, 5*(6-7), 527-543.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., et al. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science, 269*(5223), 496-512.

Fleissner, A., Sopalla, C., & Weltring, K.-M. (2002). An ATP-binding cassette multidrug-resistance transporter is necessary for tolerance of *Gibberella pulicaris* to phytoalexins and virulence on potato tubers. *Mol Plant Microbe Interact, 15*(2), 102-108.

Franceschi, V., Krekling, T., Berryman, A., & Christiansen, E. (1998). Specialized phloem parenchyma cells in Norway spruce (Pinaceae) bark are an important site of defense reactions. *American Journal of Botany, 85*(5), 601-615.

Galagan, J. E., Calvo, S. E., Borkovich, K. A., Selker, E. U., Read, N. D., Jaffe, D., et al. (2003). The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature, 422*(6934), 859-868.

Gambuel, H. A., Caves, R. G., Caffey-Moquin, M. K., & Paine, T. D. (1985). Variation in the chemistry of Ioblolly pine in relation to infection by the blue-stain fungus. pp. 177-184, In: S.J. Branbam and R.C, Thatcher (eds.), Integrated Pest Management Research Symposium: The Proceedings. U.S. Department of Agriculture Forest Service, So. Forest Experiment sta- tion, General Technical Report SO-56.

Gershenzon, J., & Dudareva, N. (2007). The function of terpene natural products in the natural world. *Nat Chem Biol, 3*(7), 408-414.

Goldstone, J. V., Hamdoun, A., Cole, B. J., Howard-Ashby, M., Nebert, D. W., Scally, M., et al. (2006). The chemical defensome: environmental sensing and response genes in the *Strongylocentrotus purpuratus* genome. *Developmental Biology, 300* (1), 366-384.

Green, E. D. (2001). Strategies for the systematic sequencing of complex genomes. *Nat Rev Genet, 2*(8), 573-583.

Green, P. (1995). Phrap and Cross_Match. from http://www.phrap.org

Hammerschmidt, R. (1999). Phytoalexins: What Have We Learned After 60 Years? *Annu Rev Phytopathol, 37*, 285-306.

Harborne, J. B. (1988). Flavonoids in the environment: structure-activity relationships. *Prog Clin Biol Res, 280*, 17-27.

Harborne, J. B. (1991). *The chemical basis of plant defense* (R.T. Palo and C.T. Robbins ed.). Boca Raton, Florida: CRC Press.

Harismendy, O., Ng, P. C., Strausberg, R. L., Wang, X., Stockwell, T. B., Beeson, K. Y., et al. (2009). Evaluation of next generation sequencing platforms for population targeted sequencing studies. *Genome Biol, 10*(3), R32.

Hillier, L. W., Marth, G. T., Quinlan, A. R., Dooling, D., Fewell, G., Barnett, D., et al. (2008). Whole-genome sequencing and variant discovery in *C. elegans*. *Nat Methods, 5*(2), 183-188.

Himejima, M., Hobson, K., Otsuka, T., Wood, D., & Kubo, I. (1992). Antimicrobial Terpenes from Oleoresin of Ponderosa pine tree *Pinus ponderosa* - a Defense-Mechanism against Microbial Invasion. *J Chem Ecol, 18*(10), 1809-1818.

Hofstetter, R. W., Mahfouz, J. B., Klepzig, K. D., & Ayres, M. P. (2005). Effects of tree phytochemistry on the interactions among endophloedic fungi associated with the southern pine beetle. *J Chem Ecol, 31*(3), 539-560.

Hoskins, R. A., Carlson, J. W., Kennedy, C., Acevedo, D., Evans-Holm, M., Frise, E., et al. (2007). Sequence finishing and mapping of *Drosophila melanogaster* heterochromatin. *Science, 316*(5831), 1625-1628.

Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Res, 9*(9), 868-877.

Hudgins, J. W., Christiansen, E., & Franceschi, V. R. (2003). Methyl jasmonate induces changes mimicking anatomical defenses in diverse members of the Pinaceae. *Tree Physiol, 23*(6), 361-371.

Hudgins, J. W., Christiansen, E., & Franceschi, V. R. (2004). Induction of anatomically based defense responses in stems of diverse conifers by methyl jasmonate: a phylogenetic perspective. *Tree Physiol, 24*(3), 251-264.

Huse, S. M., Huber, J. A., Morrison, H. G., Sogin, M. L., & Welch, D. (2007). Accuracy and quality of massively parallel DNA pyrosequencing. *Genome Biol, 8*(7), R143.

Idury, R. M., & Waterman, M. S. (1995). A new algorithm for DNA sequence assembly. *J Comput Biol, 2*(2), 291-306.

Joseph, G., Kelsey, R. G., Peck, R. W., & Niwa, C. G. (2001). Response of some scolytids and their predators to ethanol and 4-allylanisole in pine forests of central Oregon. *J Chem Ecol, 27*(4), 697-715.

Katsir, L., Chung, H. S., Koo, A. J., & Howe, G. A. (2008). Jasmonate signaling: a conserved mechanism of hormone sensing. *Curr Opin Plant Biol, 11*(4), 428-435.

Keeling, C., & Bohlmann, J. (2006). Genes, enzymes and chemicals of terpenoid diversity in the constitutive and induced defense of conifers against insects and pathogens. *New Phytol, 170*(4), 657-675.

Kim, J.-J.,E. A. Allen, Humble, L. M., & Breuil, C. (2005). Ophiostomatoid and basidiomycetous fungi associated with green, red, and grey lodgepole pines after mountain pine beetle (*Dendroctonus ponderosae*) infestation. *Can J For. Res.* (35), 274-284.

Knobloch, K., Weigand, H., Weis, N., & Brunke, E. J. (1986). Action of Terpenoids on Energy Metabolism *Progress in Essential Oil Research* (pp. 429-442).

Kopper, B., Illman, B., Kersten, P., Klepzig, K., & Raffa, K. (2005). Effects of diterpene acids on components of a conifer bark beetle-fungal interaction: Tolerance by *Ips pini* and sensitivity by its associate *Ophiostoma ips*. *Environmental Entomology, 34*(2), 486-493.

Kyrpides, N. C. (2009). Fifteen years of microbial genomics: meeting the challenges and fulfilling the dream. *Nat Biotechnol, 27*(7), 627-632.

Lander, E. S., & Waterman, M. S. (1988). Genomic mapping by fingerprinting random clones: a mathematical analysis. *Genomics, 2*(3), 231-239.

Lee, S., Kim, J., & Breuil, C. (2005). *Leptographium longiclavatum* sp nov., a new species associated with the mountain pine beetle, *Dendroctonus ponderosae*. *Mycol Res, 109*, 1162-1170.

Lee, S., Kim, J., & Breuil, C. (2006a). Pathogenicity of *Leptographium longiclavatum* associated with *Dendroctonus ponderosae* to *Pinus contorta*. *Can J For. Res., 36* (11), 2864-2872.

Lee, S., Kim, J.-J., & Breuil, C. (2006b). Diversity of fungi associated with the mountain pine beetle, *Dendroctonus ponderosae* and infested lodgepole pines in British Columbia. *Fungal Diversity*(22), 91-105.

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., et al. (2009). The sequence and de novo assembly of the giant panda genome. *Nature*.

Lieutier, F. (2004). Host Resistance to Bark Beetles and Its Variations. In F. Lieutier, K. Day, A. Battisti, J. Grégoire & H. Evans (Eds.), *Bark and Wood Boring Insects in Living Trees in Europe, a Synthesis* (pp. 135-180). Netherlands: Springer.

Lieutier, F., & Berryman, A. (1988). Preliminary Histological Investigations of the Defense Reactions of 3 Pines to *Ceratocystis clavigera* and 2 Chemical Elicitors. *Can J For. Res., 18*(10), 1243-1247.

Lieutier, F., Brignolas, F., Sauvard, D., Yart, A., Galet, C., Brunet, M., et al. (2003). Intra- and inter-provenance variability in phloem phenols of *Picea abies* and relationship to a bark beetle-associated fungus. *Tree Physiol*, *23*(4), 247-256.

Lieutier, F., Yart, A., & Salle, A. (2009). Stimulation of tree defenses by Ophiostomatoid fungi can explain attack success of bark beetles on conifers. *Annals of Forest Science, 66*(8), 801.

Lim, Y. W., Kim, J., Lu, M., & Breuil, C. (2005). Determining fungal diversity on *Dendroctonus ponderosae* and *Ips pini* affecting lodgepole pine using cultural and molecular methods. *Fungal Diversity, 19*, 79-94.

Lipson, D., Raz, T., Kieu, A., Jones, D. R., Giladi, E., Thayer, E., et al. (2009). Quantification of the yeast transcriptome by single-molecule sequencing. *Nat Biotechnol, 27*(7), 652-U105.

Magrini, V., Warren, W., Wallis, J., Goldman, W., Xu, J., Mardis, E., et al. (2004). Fosmid-based physical mapping of the *Histoplasma capsulatum* genorne. *Genome Res, 14*(8), 1603-1609.

Makkar, R., Tsuneda, A., Tokuyasu, K., & Mori, Y. (2001). *Lentinula edodes* produces a multicomponent protein complex containing manganese (II)-dependent peroxidase, laccase and β-glucosidase. *FEMS microbiology letters, 200*(2), 175-179.

Margulies, M., Egholm, M., Altman, W. E., Attiya, S., Bader, J. S., Bemben, L. A., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature, 437*(7057), 376-380.

Marra, M. A., Kucaba, T. A., Dietrich, N. L., Green, E. D., Brownstein, B., Wilson, R. K., et al. (1997). High throughput fingerprint analysis of large-insert clones. *Genome Res, 7*(11), 1072-1084.

Martin, F., Aerts, A., Ahren, D., Brun, A., Danchin, E. G., Duchaussoy, F., et al. (2008). The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. *Nature, 452*(7183), 88-92.

McPherson, J. D., Marra, M., Hillier, L., Waterston, R. H., Chinwalla, A., Wallis, J., et al. (2001). A physical map of the human genome. *Nature, 409*(6822), 934-941.

Metzker, M. L. (2009). Sequencing technologies — the next generation. *Nature Reviews Genetics, 11*(1), 31-46.

Miller, J. R., Delcher, A. L., Koren, S., Venter, E., Walenz, B. P., Brownley, A., et al. (2008). Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics, 24*(24), 2818-2824.

Myers, E. W. (1995). Toward simplifying and accurately formulating fragment assembly. *J Comput Biol, 2*(2), 275-290.

Naigre, R., Kalck, P., Rogues, C., Roux, I., & Michel, G. (1996). Comparison of antimicrobial properties of monoterpenes and their carbonylated products. *Planta Medica, 62*(3), 275-277.

Nicholson, R., & Hammerschmidt, R. (1992). Phenolic-compounds and their role in disease resistance. *Annual review of phytopathology, 30*, 369-389.

Olson, M. V., Dutchik, J. E., Graham, M. Y., Brodeur, G. M., Helms, C., Frank, M., et al. (1986). Random-clone strategy for genomic restriction mapping in yeast. *Proc Natl Acad Sci U S A, 83*(20), 7826-7830.

Paine, T., & Hanlon, C. (1994). Influence of oleoresin constituents from *Pinus ponderosa* and *Pinus jeffreyi* on growth of mycangial fungi from *Dendroctonus ponderosae* and *Dendroctonus jeffreyi*. *J Chem Ecol, 20*(10), 2551-2563.

Paine, T., Raffa, K., & Harrington, T. (1997). Interactions among scolytid bark beetles, their associated fungi, and live host conifers. *Annu Rev Entomol, 42*, 179-206.

Pareja-Jaime, Y., Roncero, M. I. G., & Ruiz-Roldán, M. C. (2008). Tomatinase from *Fusarium oxysporum* f. sp. lycopersici is required for full virulence on tomato plants. *Mol Plant Microbe Interact, 21*(6), 728-736.

Parveen, M., Hasan, M., Takahashi, J., Murata, Y., Kitagawa, E., Kodama, O., et al. (2004). Response of *Saccharomyces cerevisiae* to a monoterpene: evaluation of antifungal potential by DNA microarray analysis. *J Antimicrob Chemother, 54*(1), 46-55.

Pepke, S., Wold, B., & Mortazavi, A. (2009). Computation for ChIP-seq and RNA-seq studies. *Nat Methods, 6*(11s), S22-S32.

Pevzner, P. A., & Tang, H. (2001). Fragment assembly with double-barreled data. *Bioinformatics, 17 Suppl 1*, S225-233.

Pevzner, P. A., Tang, H., & Waterman, M. S. (2001). An Eulerian path approach to DNA fragment assembly. *Proc Natl Acad Sci USA, 98*(17), 9748-9753.

Plattner, A., Kim, J., DiGuistini, S., & Breuil, C. (2008). Variation in pathogenicity of a mountain pine beetle–associated blue-stain fungus. *Can J Plant Pathology, 30* (3), 457-466.

Plattner, A., Kim, J.-J., Reid, J., Hausner, G., Lim, Y. W., Yamaoka, Y., et al. (2009). Resolving taxonomic and phylogenetic incongruence within species *Ceratocystiopsis minuta*. *Mycologia, 101*(6), 878-887.

Raffa, K. (2001). Mixed messages across multiple trophic levels: the ecology of bark beetle chemical communication systems. *Chemoecology, 11*(2), 49-65.

Raffa K.F., Phillips T.W., & Salom, S.M. (1993). Strategies and mechanisms of host colonization by bark beetles. In S. T. D. a. F. G.M. (Ed.), *Beetle-pathogen interactions in conifer forests* (pp. 103–128). San Diego: Academic Press.

Rice, A., Thormann, M., & Langor, D. (2007). Mountain pine beetle associated blue-stain fungi cause lesions on jack pine, lodgepole pine, and lodgepole x jack pine hybrids in Alberta. *Canadian Journal of Botany, 85*, 307-315.

Robinson, R. C. (1962). Blue stain fungi in lodgepole pine (*Pinus contorta* Dougl. var. latifolia Engelm.) infested by the mountain pine beetle (*Dendroctonus monticolae* Hopk.). *Can. J. Bot., 40*, 609-614.

Rumbold, C. T. (1941). A blue stain fungus, *Ceratostomella montium* n. sp., and some yeasts associated with two species of *Dendroctonus*. *Journal of Agricultural Research, 62*, 589-601.

Safranyik, L. (1978). Effects of Climate and Weather on Mountain Pine Beetle Populations. In A. AA Berryman, G. D., Stark, R. W. (Ed.), *Theory and Practice of Mountain Pine Beetle Management in Lodgepole Pine Forests* (pp. 77-84).

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., et al. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. *Nature, 265* (5596), 687-695.

Savluchinske Feio, S., Gigante, B., Roseiro, J. C., & Marcelo-Curto, M. J. (1999). Antimicrobial activity of diterpene resin acid derivatives. *J Microbiol Methods, 35* (3), 201-206.

Schuler, G. D., Boguski, M. S., Stewart, E. A., Stein, L. D., Gyapay, G., Rice, K., et al. (1996). A gene map of the human genome. *Science, 274*(5287), 540-546.

Shendure, J., Porreca, G. J., Reppas, N. B., Lin, X., McCutcheon, J. P., Rosenbaum, A. M., et al. (2005). Accurate multiplex polony sequencing of an evolved bacterial genome. *Science, 309*(5741), 1728-1732.

Shizuya, H., Birren, B., Kim, U., Mancino, V., Slepak, T., Tachiiri, Y., et al. (1992). Cloning and stable maintenance of 300-kilobase-pair fragments of human dna in *Escherichia coli* using an F-factor-based vector. *Proc Natl Acad Sci USA, 89*(18), 8794-8797.

Shrimpton, D., & Whitney, H. (1968). Inhibition of growth of blue stain fungi by wood extractives. *Can. J. Bot., 46*, 757-761.

Sikkema, J., de Bont, J., & Poolman, B. (1994). Interactions of cyclic hydrocarbons with biological membranes. *J Biol Chem, 269*(11), 8022-8028.

Simpson, J., Wong, K., Jackman, S., & Schein, J. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Res*.

Six, D. (2003a). A comparison of mycangial and phoretic fungi of individual mountain pine beetles. *Can J Forest Res, 33*(7), 1331-1334.

Six, D. (2003b). A comparison of mycangial and phoretic fungi of individual mountain pine beetles.

Six, D., & Bentz, B. J. (2007). Temperature Determines Symbiont Abundance in a Multipartite Bark Beetle-fungus Ectosymbiosis. *Microb Ecol, 54*(1), 112-118.

Solheim, H. (1995). Early stages of blue-stain fungus Invasion of lodgepole pine sapwood following mountain pine-beetle attack. *Can. J. Bot., 73*(1), 70-74.

Stefanato, F. L., Abou-Mansour, E., Buchala, A., Kretschmer, M., Mosbach, A., Hahn, M., et al. (2009). The ABC transporter BcatrB from *Botrytis cinerea* exports camalexin and is a virulence factor on *Arabidopsis thaliana*. *The Plant Journal, 58*(3), 499-510.

Sun, C. B., Suresh, A., Deng, Y. Z., & Naqvi, N. I. (2006). A multidrug resistance transporter in *Magnaporthe* is required for host penetration and for survival during oxidative stress. *Plant Cell, 18*(12), 3686-3705.

Sutton, G., White, O., Adams, M. D., & Kerlavage, A. R. (1995). TIGR Assembler: A New Tool for Assembling Large Shotgun Sequencing Projects. *Genome Science & Technology, 1*(1), 9-19.

Trapp, & Croteau. (2001). Defensive resin biosynthesis in conifers. *Annu Rev Plant Physiol Plant Mol Biol, 52*, 689-724.

Urban, M., Bhargava, T., & Hamer, J. E. (1999). An ATP-driven efflux pump is a novel pathogenicity factor in rice blast disease. *EMBO J, 18*(3), 512-521.

Uribe, S., Ramirez, J., & Pena, A. (1985). Effects of β-pinene on yeast membrane functions. *J Bacteriol, 161*(3), 1195-1200.

Valouev, A., Ichikawa, J., Tonthat, T., Stuart, J., Ranade, S., Peckham, H., et al. (2008). A high-resolution, nucleosome position map of *C. elegans* reveals a lack of universal sequence-dictated positioning. *Genome Res, 18*(7), 1051-1063.

van den Brink, H., van Gorcom, R., van den Hondel, C., & Punt, P. (1998). Cytochrome P450 enzyme systems in fungi. *Fungal Genet Biol, 23*(1), 1-17.

Venter, J. C., Smith, H. O., & Hood, L. (1996). A new strategy for genome sequencing. *Nature, 381*(6581), 364-366.

Vera, J. C., Wheat, C. W., Fescemyer, H. W., Frilander, M. J., Crawford, D. L., Hanski, I., et al. (2008). Rapid transcriptome characterization for a nonmodel organism using 454 pyrosequencing. *Mol Ecol, 17*(7), 1636-1647.

Wang, Z., Gerstein, M., & Snyder, M. (2008). RNA-Seq: a revolutionary tool for transcriptomics. *Nature Reviews Genetics*, 7.

Wilhelm, B. T., Marguerat, S., Goodhead, I., & Bahler, J. (2010). Defining transcribed regions using RNA-seq. *Nat Protoc, 5*(2), 255-266.

Wu, H., & Hu, Z. (1997). Comparative anatomy of resin ducts of the Pinaceae. *Trees-Struct Funct, 11*(3), 135-143.

Yamaoka, Y., Hiratsuka, Y., & Maruyama, P. (1995). The ability of *Ophiostoma clavigerum* to kill mature lodgepole pine trees. *European Journal of Forest Pathology, 25*(6-7), 401-404.

Zerbino, D. R., & Birney, E. (2008). Velvet: algorithms for de novo short read assembly using de Bruijn graphs. *Genome Research, 18*(5), 821-829.

Zipfel, R. D., de Beer, Z. W., Jacobs, K., Wingfield, B. D., & Wingfield, M. J. (2006). Multi-gene phylogenies define *Ceratocystiopsis* and *Grosmannia* distinct from *Ophiostoma*. *Studies in Mycology*(55), 75-97.

Zukerman, I. (1951). Effect of oxidized d-limonene on micro-organisms. *Nature, 168* (4273), 517-518.

Zwiers, L.-H., Stergiopoulos, I., Gielkens, M. M. C., Goodall, S. D., & De Waard, M. A. (2003). ABC transporters of the wheat pathogen *Mycosphaerella graminicola* function as protectants against biotic and xenobiotic toxic compounds. *Mol Genet Genomics, 269*(4), 499-507.

2 Generation and annotation of lodgepole pine and oleoresin induced expressed sequences from the blue-stain fungus *Ophiostoma clavigerum*, a mountain pine beetle associated pathogen[1]

2.1 Introduction

Ophiostoma clavigerum (Robinson-Jeffrey & Davidson) Harrington, is a pathogenic fungus of the order Ophiostomatales. It is exclusively associated with the closely related bark beetles *Dendroctonus ponderosae* Hopkins (Mountain Pine Beetle, MPB) and *Dendroctonus jeffreyii* Hopkins (Kim, et al., 2005). When MPBs attack their primary conifer host, healthy lodgepole pine (*Pinus contorta* Douglas var. *latifolia* Engelmann), they introduce *O. clavigerum* and other associated fungi. *O. clavigerum* colonizes the sapwood rapidly (Robinson, 1962). During colonization a blue/black melanin pigment is produced, transpiration is blocked and tree death occurs; however, the strength of the wood is not affected. MPB attack and the development of staining fungi result in reduced timber value and an accumulation of unsalvageable trees. In the current MPB epidemic in British Columbia, Canada, tree and wood losses are estimated to be worth billions of dollars (http://www.for.gov.bc.ca/hfp/mountain_pine_beetle/).

The major component of lodgepole pine defense against *O. clavigerum* is oleoresin (Raffa & Berryman, 1982). Oleoresin is composed mainly of 10-carbon monoterpenes, 15-carbon sesquiterpenes, and 20-carbon diterpenes (Keeling & Bohlmann, 2006). It is

[1] A version of this chapter has been published. Scott DiGuistini, Stephen G. Ralph, Young W. Lim, Steven J.M. Jones, Robert A. Holt, Jörg Bohlmann and Colette Breuil. (2007) Generation and annotation of lodgepole pine and oleoresin-induced expressed sequences from the blue-stain fungus *Ophiostoma clavigerum*, a Mountain Pine Beetle-associated pathogen FEMS Microbiol Lett. 267(2):151-8

toxic towards some fungi and bark beetles (Delorme & Lieutier, 1990, Himejima, et al., 1992). Because *O. clavigerum* survives and grows in a saturated oleoresin environment, it likely possesses mechanisms for modifying or degrading the antimicrobial oleoresin components. While we anticipate that detoxifying enzymes such as P450s and broad specificity transport proteins such as ATP-Binding-Cassette (ABC) transporters will likely be important, additional components remain to be discovered. A prerequisite for fully characterizing *O. clavigerum*'s growth, development, and tolerance towards lodgepole pine oleoresin is its genomic sequence. As this is not currently available, we used cDNA libraries to generate large numbers of expressed sequence tags (ESTs), with which we identified genes expressed on media that simulated conditions *O. clavigerum* would encounter while growing in lodgepole pine. To facilitate identifying rarer transcripts, we used cDNA library normalization techniques to reduce the frequency of highly expressed genes (Bonaldo, et al., 1996). This paper describes an EST collection derived from four normalized cDNA libraries that were generated from filamentous fungal growth of *O. clavigerum* under three different conditions: 1) lodgepole pine sawdust, 2) malt extract agar (MEA), and 3) MEA supplemented with lodgepole pine monoterpene and diterpene metabolites. We have sequenced more than 6,500 ESTs representing 2,600 unique putative transcripts (UPTs).

2.2 Materials and methods

2.2.1 Fungal strain and culture conditions

The *O. clavigerum* (strain SLKw1407) used in this study was isolated from a *D. ponderosae*-infested lodgepole pine tree from a MPB epidemic region near Kamloops, BC. Mycelium for library construction were collected from solid media inoculated with a suspension containing 5x10^5 spores and were incubated for 7 days under ambient conditions. Mycelium were generated on three media: 1) ~10 g of fresh lodgepole pine sawdust that was electron beam sterilized (Iontron Sterilization, Coquitlam, BC) and mixed with 1.5 % (w/v) autoclaved agar, 2) 1 % MEA (Difco Laboratories, Detroit, MI; w/v), and 3) 1 % MEA supplemented with 75 µL of a diterpene blend (abietic, isopimaric and pimaric acids) spread onto the surface of 12.5 mL of media. The diterpene-supplemented cultures were then placed inside a 2 L gas-tight glass bell with a headspace saturated with 250 µL of a mixture of monoterpenes ((+/-)α-pinene, (-)β-pinene, 3-carene, β-phellandrene, (+/-)limonene, α-terpinolene and γ-terpinene). The selected metabolites were blended in a ratio similar to that described by Shrimpton (1973) based on extractive analysis of MPB attacked trees and is more fully described in appendix A.1.

2.2.2 RNA isolation, cDNA library normalization

For each of the three media conditions, total RNA was isolated from *O. clavigerum* mycelia pooled from 12-20 cultures. mRNA was isolated and 5 µg of poly(A^+) mRNA was reverse transcribed and directionally cloned into a 5'*Eco*RI and 3'*Xho*I pre-digested phagemid vector pBS II SK(+) (Stratagene, La Jolla, CA). The libraries were normalized using the method described by Bonaldo et al. (1996). The libraries were normalized to $C_ot = 5$, except for the sawdust library which was split into two fractions, one normalized to $C_ot = 10$ while the other was normalized to $C_ot = 5$. See appendix A.2 for an explanation of C_ot.

2.2.3 DNA sequencing and analysis

Randomly selected clones from all libraries were partially sequenced from the 3' end using the –21 M13 forward primer (Table 2.1). Chromatograms were processed and quality trimmed using PHRED v. 0.020425.c (Ewing & Green, 1998), and vector sequences were removed using CROSS-MATCH (http://www.phrap.org/). Finally, sequences with PHRED 20 of at least 100 bp were assembled using CAP3 (Huang & Madan, 1999) with default settings except for a minimum overlap of 40 bp and a minimum percent identity of 95.

2.2.4 Annotation of ESTs

ESTs were annotated using the results from BLASTx and BLASTn (Altschul, et al., 1990) analysis unless otherwise noted. The sequence databases used for *O.*

clavigerum EST annotation were: the non-redundant (*nr*) division of GenBank (downloaded on 2005-03-07 http://www.ncbi.nlm.nih.gov/BLAST/), Swiss-prot (downloaded on 2005-09-27 http://www.expasy.org/sprot/), the Gene Ontology (GO) v. 2006-01 (Ashburner, et al., 2000) and PHI-base v. 2.1 (Winnenburg, et al., 2006). Further analysis using GO annotations was performed with GoMiner (Zeeberg, et al., 2003) and results from this analysis were visualized using VennMaster (Kestler, et al., 2005). Further description can be found in appendix A.3.

2.2.5 Quantitative real-time PCR

Trizol (Invitrogen, Mississauga, ON) extractions were used to purify RNA for quantitative real-time PCR (qPCR) with the following modifications: 1) centrifugations were performed at 4°C and 2) 1-Bromo-3-chloro-propane (BCP) was substituted for chloroform. DNaseI (Fisher Scientific, Ottawa, ON) treatment of the Trizol extracted RNA ensured adequate removal of all genomic DNA contamination. cDNA was produced from 5 µg of total RNA using Superscript II (Invitrogen, Mississauga, ON) and oligo $(dT)_{12-18}$ following the manufacturer's protocol. qPCR was performed on a Stratagene M3000P (La Jolla, CA) and data analysis was performed within SAS (See appendix A.4 materials; Statistical Analysis Systems, Cary, NC). Primers for the analysis of the transcript of interest were CFEM1-F and CFEM1-R (Table 2.1); the amplicon size was 90 bp. Primers for the reference transcript were 1407btub-F and 1407btub-R (Table 2.1); the amplicon size was 135bp. Primer specificity (single product of expected length) was confirmed by analysis on a 2 % agarose gel and by melting curve analysis. PCR reactions were composed of forward and reverse primers, each at 300 nM or 600 nM (optimum primer concentration was determined using a dilution curve), 1x iQ supermix pre-mix (Bio-Rad, Mississauga, ON) and 50 ng of *O. clavigerum* cDNA in a total volume of 25 µl. Cycling parameters for qPCR were 95°C for 10 min followed by 40 cycles of 95°C for 10 sec, 62°C for 30 sec, 72°C for 30 sec and an observation step of 82°C for 18 sec followed by a melting point analysis. Each analysis was replicated three times with biologically and technically independent samples.

2.3 Results

2.3.1 Generation and assembly of EST collections

We generated a total of 6,528 3'-EST sequences from four normalized, uni-directional, cDNA libraries (MPB01-04) based on mycelial growth under three different culture conditions: 1) lodgepole pine sawdust, 2) MEA and 3) MEA supplemented with monoterpenes and diterpenes (Table 2.2). After quality filtering, sequences less than 100 bp were excluded, and 5,975 high quality sequences remained, with an average PHRED 20 for these reads of 786 bp. The ratio of high quality reads to total reads was 92% and this did not differ substantially between libraries. CAP3 analysis of the high-quality ESTs identified 2,620 UPTs: 4,497 sequences clustered into 1,142 contigs (the average number of contig members was 3.94) and 1,478 singletons (Table 2.2). Although most UPTs were represented by a single member, 965 contigs contained two to five members, 170 contigs contained six to twenty members and 7 contigs contained 21 or more members. The largest contig contained 72 ESTs, and was annotated as a gene encoding a 40S ribosomal protein (Table 2.3). Libraries MPB01-04 contained 365, 618, 378 and 400 unique transcripts respectively; only 52 UPTs were represented with ESTs from all four libraries. All sequences have been deposited into Genbank's dbEST database (Accession nos EE724403-EE730376).

2.3.2 EST analysis

Of the 2,620 UPTs, BLASTx of the *nr* database identified 67% (1,755 transcripts) that were similar to previously deposited protein sequences (score ≥ 100, Expected (*E*)-value ≤ 10^{-15}). Ninety-two percent of the transcripts with 'best matches' were similar to known or predicted protein sequences from *Magnaporthe grisea* (35 %), *Neurospora crassa* (34 %), or *Gibberella zeae* (23 %). Another 7 % (183) of these transcripts could be matched with proteins from other fungi, and only 1 % had 'best matches' that were not fungi.

Following this preliminary annotation, UPTs were ranked by frequency analysis. Frequency was calculated as the number of ESTs contributing to a UPT. The frequency analysis revealed that a small number of UPTs appeared to be over represented within individual libraries. For example, 20 of the 22 MPB0586 EST sequences originated from the MPB03 library (Table 3). The BLASTx 'best match' for MPB0586 was *A. fumigatus* EAL88523, which was annotated as a cysteine-containing domain present in fungal extracellular membrane proteins (CFEM). QPCR indicated that the expression of the CFEM-domain-containing gene, MPB0586, was induced by fungal growth in the presence of the selected oleoresin metabolites. Monoterpenes stimulated the expression of this gene by approximately 2.75 fold ($p<0.0001$, $\text{critical}_{\alpha}= 0.0167$)

compared with expression on MEA alone, while the solvent DMSO caused some increase in expression (1.64, p=0.0004, critical$_{\alpha}$= 0.0167) compared with MEA. In relation to the DMSO treatment, the diterpene abietic acid did not have an effect on MPB0586 expression (p=0.0121, critical$_{\alpha}$= 0.0167). We did not test the combined effects of diterpene acids with monoterpenes.

As no information is available for the genes involved in *O. clavigerum*'s pathogenicity*,* UPTs were compared against PHI-base, a curated sequence collection of fungal verified virulence and pathogenicity genes. To determine the EST collection's comprehensiveness we analyzed the cytochrome P450 and ATP-binding-cassette (ABC)-type transporter gene families, as these are likely involved in host defense chemical detoxification. Because we previously identified genes for most of the Ophiostomatoid enzymes in the 1,8-Dihydroxynaphthalene (DHN)-melanin biosynthetic pathway (Loppnau, et al., 2004), we used this pathway as a reference for assessing the EST collection's completeness. BLASTx was used to compare the *O. clavigerum* UPTs with PHI-base (score ≥ 100, *E* value ≤ 10^{-50}), and UPTs were ranked by frequency; Table 2.4 highlights the top ten sequences. Fourteen cytochrome P450s containing the heme binding site motif: Phe-X-X-Gly-X-Arg-X-Cys-X-Gly (Werck-Reichhart & Feyereisen, 2000) were identified and manually confirmed within the *O. clavigerum* EST collection (Table 2.5). We also examined ABC-transporter proteins, and found six sequences whose 3' ends contained elements of the ABC transporter signature motif, the linker peptide region (CDD: cd00267; http://www.ncbi.nlm.nih.gov/Structure/cdd/; Table 5). The *O. clavigerum* EST collection contained a number of genes from the DHN-melanin biosynthesis pathway. Sequences with similarity to genes from *Ophiostoma* and *Ceratocystis* species were found and included a polyketide synthase and scytalone dehydratase as well as tetra- and tri- hydroxynaphthalene reductases. This analysis also identified a sequence present in *Aspergillus fumigatus,* AYG1, not previously shown in *Ophiostoma* and *Ceratocystis* species (Table 2.6).

Comparing *O. clavigerum* UPTs against the Gene Ontology (GO) database extended the functional annotation. GO annotations were assigned using the single 'best match' BLASTx hit against the GO sequence database (score ≥ 100, *E* value ≤ 10^{-15}) and 905 (34%) of *O. clavigerum's* unique sequences could be associated with GO terms. Following the assignment of GO terms to *O. clavigerum's* UPTs, GoMiner was used to compare the distributions of ESTs amongst the biological process and molecular function categories and between the cDNA libraries (data not shown). ESTs with GO associations for biological processes like actin cytoskeleton binding and organization, as well as transcriptional and cell cycle regulation were over-represented in libraries MPB01 and MPB04, which originated from *O. clavigerum* growth on sawdust media. Libraries derived from fungal growth on MEA possessed roughly the same proportion of unique sequences as libraries generated from fungal growth on sawdust. However, libraries derived from fungal growth on MEA contained a more narrow GO distribution, with more unique sequences assigned to fewer biological process and molecular function categories. For the library derived from *O. clavigerum* growth on MEA supplemented with oleoresin terpenoids, ESTs with GO associations for the biological processes cytoplasm organization and biogenesis were over-represented, along with several shared terms related to vitamin metabolism; under molecular function, cytoskeleton binding proteins were over-represented, as well as a large group of proteins putatively identified as having oxidoreductase activity.

2.4 Discussion

O. clavigerum is a MPB associated pathogen that causes economic losses by discolouring wood and killing MPB infested lodgepole pine trees. Very limited gene sequence information is available for this fungus, despite its economic and ecological importance. In this work we generated and described a normalized EST collection for a strain of *O. clavigerum* isolated from a lodgepole pine tree in one of British Columbia's MPB epidemic regions. We sequenced more than 6,500 ESTs, which represented approximately 2,620 UPTs.

We used GoMiner with VennMaster to compare GO terms associated with unassembled ESTs with terms for assembled UPTs, and identified GO categories over-represented within biological process and molecular function hierarchies. For cDNA libraries derived from fungus grown on sawdust, over-represented biological processes included actin cytoskeleton re-organization as well as transcription and cell-cycle regulation. Fungal growth in conifer stems and on sawdust-derived media is nutrient-limited (Meerts, 2002). Consistent with this, cultures grown on lodgepole pine sawdust had over-represented GO categories with similarity to those for nutrient-limited yeast cells (Gasch & Werner-Washburne, 2002). Because specific nutrient-limitations are often developmental cues (Cullen & Sprague, 2000) and can induce virulence-associated gene expression (Snoeijers, et al., 2000), we anticipate that the ESTs derived from *O. clavigerum* grown on sawdust will contain sequences useful for characterizing fungal pathogenicity.

Oleoresin is a major component of conifer defense systems (Keeling & Bohlmann, 2006). However, *O. clavigerum* cultured with oleoresin shows only slightly decreased growth rates (Shrimpton & Whitney, 1968); further, some oleoresin components stimulate its growth (Paine & Hanlon, 1994). GO analysis of ESTs from fungi grown on oleoresin indicated that oxidoreductase enzymes were over-represented. Many plant pathogens utilize oxidoreductases for antimicrobial natural product detoxification and for host chemical perception (Idnurm and Howlett, 2001; Palmer et al., 2004). Since little is known about the molecular mechanisms utilized by *O. clavigerum* to colonize lodgepole pine, these observations represent good leads for dissecting the molecular genetics of oleoresin detoxification and pathogenicity.

P450-mediated phytoalexin detoxification is important for some plant-fungal interactions (Maloney & VanEtten, 1994). P450s have been shown in both trees and fungi to participate in terpenoid biosynthesis (Parker & Scott, 2005, Ro, et al., 2005). To identify genes contributing to *O. clavigerum* pathogenesis we screened the EST collection for members of the P450 protein family. Fourteen P450 UPTs were identified in *O. clavigerum* ESTs that contained the heme binding site motif. The majority of P450 clones were sequenced once; none were sequenced more than eight times. Alignment and phylogenetic analyses indicated no relationship between the four cDNA libraries and P450 CYP classification or P450 EST frequency (data not shown). Similarly, ABC transporters have been implicated in antibiotic resistance for a number of agriculturally important fungal pathogens (Urban, et al., 1999) and are also involved in plant antifungal terpenoid secretions (Jasinski, et al., 2001). We found six UPTs whose 3' ends were similar to known ABC genes. Because lodgepole pine is rich in antimicrobial oleoresin components, both P450s and ABC transporters represent interesting candidates for functional characterization in *O. clavigerum* interactions with lodgepole pine defense chemicals.

Because Ophiostomatoid fungi cause substantial economic losses to the forest industry by discoloring sapwood, DHN-melanin biosynthesis is their most extensively studied pathway (Loppnau, et al., 2004). As expected, the *O. clavigerum* EST collection contained DHN-melanin biosynthesis pathway genes previously characterized in Ophiostomatoids. In addition, we identified the AYG1 gene, a polyketide shortening hydrolase (Fujii, et al., 2004) not previously shown in Ophiostomatoid fungi. DHN-melanin has been shown to play an important role in many fungal pathosystems by strengthening cell walls and providing environmental protection against reactive oxygen species and UV. However, its role in *O. clavigerum* biology is uncertain. ESTs identified in this project will allow for future studies testing the roles of DHN-melanin in *O. clavigerum* interactions with lodgepole pine.

In summary, we generated the first genome-scale expressed sequence data for *O. clavigerum* and provided preliminary annotations for many of the UPTs. In addition, we identified candidate genes for further studies to test the role of oleoresin tolerance in the pathogenicity of *O. clavigerum* towards lodgepole pine. These results begin the characterization of the molecular interactions between this fungal pathogen and its host. The ESTs will serve as reagents for developing additional genomics tools for characterizing *O. clavigerum* gene expression, and will be broadly applicable to studies of Ophiostomatoid fungi vectored by bark beetles.

Tables

Table 2.1 Primers used in this study

Primer name	Application	Sequence 5’-3’	Primer Length
Anchored oligo d(T)	cDNA library construction	GAGAGAGAGAGAGAGAGAGAACTAGT**CTCGAG**T(T$_{16}$)VN[a]	51bp
-21M13F	cDNA library construction	TGTAAAACGACGGCCAGT	18bp
M13R	cDNA library construction	CAGGAAACAGCTATGAC	17bp
T7	Normalization	GTAATACGACTCACTATAGGGC	22bp
SKpBS	Normalization	CGCTCTAGAACTAGTGGATCC	21bp
CFEM1-F	qRT-PCR	AGCCACCGGGCTCAGCCAGACA	22bp
CFEM1-R	qRT-PCR	GGCAACTGCGGCACCGATCC	20bp
1407btub-F	qRT-PCR	TCTCGACAGCAATGGAGT	18bp
1407btub-R	qRT-PCR	CCCGAGGCCTCGTTGAAGTA	20bp

[a] XhoI (bold) restriction site added.

Table 2.2 *Ophiostoma clavigerum* EST summary.

Total 3' End Sequences	
MPB01 ($C_{o}t$ = 10, Sawdust)	1536
MPB02 ($C_{o}t$ = 5, MEA)	1920
MPB03 ($C_{o}t$ = 5, MEA+Terpenes)	1536
MPB04 ($C_{o}t$ = 5, Sawdust)	1536
Complete Assembly	
High Quality Reads	5975
Reads in Contigs[**]	4497
Contigs	1142
Singletons	1478
Unique Sequences	**2620**

[*] Average Phred 20 for all reads is 786; Phred 20 should be interpreted as a 1 in 100 probability that a base has been called incorrectly, in other words, sequences are considered 99% accurate.

[**] CAP3 assembly used default settings with minimum overlap set to 40 bp and the minimum percent identity set to 95.

Table 2.3 UPTs ranked by EST frequency. Frequency has not been normalized for differences in the number of reads per library.

Uniseq ID	EST Frequency	1	2	3	4	Best Match*; Accession No.	*E* value
MPB0888	72	3	34	33	2	40S ribosomal protein S3; XP_322575	6.00E-38
MPB0492	58	7	19	27	5	hypothetical protein; EAA54505	7.00E-20
MPB0391	37	3	16	16	2	cytochrome c reductase iron-sulfur subunit; CAA26308	1.00E-93
MPB0564	33	6	9	14	4	hypothetical protein; XP_331746.1	7.00E-60
MPB01049	30	4	15	9	2	No significant similarity	N/A
MPB0410	25	11	1	0	13	No significant similarity	N/A
MPB0586	22	0	2	20	0	CFEM domain protein; EAL88523	2.00E-09
MPB0924	19	1	8	7	3	hypothetical protein; EAA50425.1	3.00E-75
MPB0732	18	1	7	10	0	No significant similarity	N/A
MPB0205	18	7	5	3	3	No significant similarity	N/A
MPB0556	17	4	4	8	1	Vacuolar ATP synthase subunit E; XP_327732	3.00E-55
MPB0939	17	1	9	5	2	short chain dehydrogenase/reductase family; EAL84601	4.00E-36
MPB0278	17	6	3	0	8	malate dehydrogenase-like protein; AAX07691	1.00E-116
MPB0545	16	0	9	7	0	microtubule associated protein EB1; EAL92397	8.00E-70
MPB0652	16	3	5	7	1	dihydrolipoyllysine-residue acetyltransferase-like protein; AAX07694	1.00E-105
MPB0413	16	0	7	9	0	aspartate-semialdehyde dehydrogenase; EAL92885	6.00E-87
MPB0107	16	10	1	1	4	related to protein-tyrosine phosphatase; CAD70824	3.00E-30
MPB0365	16	6	2	0	8	hypothetical protein; XP_328441	2.00E-47
MPB0464	15	5	1	2	7	Cytochrome c oxidase subunit VIa family; EAL92949	3.00E-31

* ‘Best Match’ determined using results of a BLASTx analysis against the *nr* database.

Table 2.4 Comparison of *O. clavigerum* UPTs with the PHI-base database*.

Uniseq ID	EST Frequency	Accession; Best Match; Pathogen	*E* value
MPB0197	9	AAK98783; putative vacuolar ATPase MVP1; *Magnaporthe grisea*	2.00E-62
MPB0489	7	BAB85760; putative mitochondrial carrier protein; *Fusarium oxysporum*	4.00E-86
MPB01268	5	AAQ16572; putative mitochondrial cyclophilin 1; *Botryotinia fuckeliana*	2.00E-67
MPB0945	4	AAP68994; thiol-specific antioxidant protein 1; *Cryptococcus neoformans* var. *grubii*	1.00E-62
MPB0610	3	CAC17748; trehalose-6-phosphate phosphatase; *Candida albicans*	5.00E-72
MPB0132	3	CAA67930; putative mannosyl transferase; *Candida albicans*	2.00E-91
MPB01050	3	AAB86583; manganese-superoxide dismutase precursor; *Candida albicans*	1.00E-63
MPB0293	2	AAD47837; alanine racemase; *Cochliobolus carbonum*	2.00E-77
MPB0893	2	CAB56523; ornithine decarboxylase; *Phaeosphaeria nodorum*	2.00E-81
MPB0212	2	AAF09475; osmotic sensitivity MAP Kinase; *Magnaporthe grisea*	1.00E-100
MPB0720	1	CAD88591; superoxide dismutase; *Botryotinia fuckeliana*	2.00E-69

*Comparison was performed using BLASTx (score ≥ 100, *E* value ≤ 10^{-50}).

Table 2.5 UPTs belonging to the P450 and ABC gene families. ‘No significant hit’ indicates that the sequence contains a heme binding motif but could not be matched to a sequence from the public databases with significant similarity to suggest relationship.

Uniseq ID	EST Frequency	Swiss-prot ID; Description*	*E* value
P450s			
MPB0658	8	P79084; O-methylsterigmatocystin oxidoreductase	6.00E-21
MPB0465	4	Q8K4D6; Cytochrome P450 4X1	6.00E-24
MPB0943	3	P17549; Benzoate 4-monooxygenase	1.00E-141
MPB0897	2	P17177; Cytochrome P450 27, mitochondrial precursor	0.003
MPB1123	2	O13317; Isotrichodermin C-15 hydroxylase	4.00E-29
MPB2073	1	Q12645; Pisatin demethylase	3.00E-06
MPB1836	1	P17549; Benzoate 4-monooxygenase	7.00E-50
MPB1190	1	Q12612; Trichodiene oxygenase	3.00E-14
MPB1208	1	Q9VE01; Probable cytochrome P450 12a5, mitochondrial precursor	4.00E-06
MPB1396	1	P15540; Cytochrome P450 Steroid 21-hydroxylase	6.00E-07
MPB1828	1	No Significant Hit	N/A
MPB1650	1	Q92088; Cytochrome P450 2M1 Lauric acid omega-6-hydroxylase	6.00E-25
MPB1907	1	Q9LTM7; Cytochrome P450 71B16	8.00E-14
MPB2140	1	P54781; Cytochrome P450 61 C-22 sterol desaturase	7.00E-55
ABCs			
MPB00211	4	P32386; ATP-dependent bile acid permease	1.00E-15
MPB00109	2	P51533; ATP-dependent permease PDR10	5.00E-43
MPB01201	1	P43569; Probable ATP-dependent transporter YFL028C	2.00E-07
MPB02272	1	P53049; Oligomycin resistance ATP-dependent permease YOR1	9.00E-65
MPB02382	1	P36619; Leptomycin B resistance protein pmd1	8.00E-46
MPB01806	1	P43569; Probable ATP-dependent transporter YFL028C	4.00E-52

*Descriptions generated from Swiss-prot and belong to the ID of the ‘Best Match’.

Table 2.6 Analysis of UPTs identified to be involved in DHN-melanin biosynthesis*.

Uniseq ID	EST Frequency	Accession; Description; Fungus	*E* value	Reference
MPB0329	2	BAA18956; Polyketide Synthase; *Colletotrichum lagenarium*	3.00E-100	Takano et al., 1995
MPB0166	8	AAF03354; Serine protease-like hydrolase; *Aspergillus fumigatus*	2.00E-66	Tsai et al., 1999
MPB0789	10	AAK07185; Tetrahydroxynaphthalene Reductase; *Ophiostoma floccosum*	1.00E-123	Wang and Breuil, 2002
MPB0755	8	AAK11296; Scytalone Dehydratase; *Ophiostoma floccosum*	5.00E-86	Wang et al., 2001
MPB0851	9	AAK60499; Trihydroxynaphthalene Reductase; *Ophiostoma floccosum*	1.00E-113	Wang and Breuil, 2002

*UPTs were identified by comparing known *A. fumigatus* pathway members against the *O. clavigerum* EST collection using tBLASTn.

2.5 References

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., & Lipman, D.J. (1990). Basic local alignment search tool. *J Mol Biol, 215,* 403-410.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., Harris, M.A., Hill, D.P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J.C., Richardson, J.E., Ringwald, M., Rubin, G.M. & Sherlock, G. (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet, 25,* 25-29.

Bonaldo, M.F., Lennon, G., & Soares, M.B. (1996). Normalization and subtraction: two approaches to facilitate gene discovery. *Genome Res, 6,* 791-806.

Cullen, P.J., & Sprague, G.F., Jr. (2000). Glucose depletion causes haploid invasive growth in yeast. *Proc Natl Acad Sci U S A, 97,* 13619-13624.

Delorme, L., & Lieutier, F. (1990). Monoterpene composition of the preformed and induced resins of Scots pine, and their effect on bark beetles and associated fungi. *Eur J For Path, 20,* 304-316.

Ewing, B., & Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res*, *8,* 186-194.

Fujii, I., Yasuoka, Y., Tsai, H.F., Chang, Y.C., Kwon-Chung, K.J. & Ebizuka, Y. (2004). Hydrolytic polyketide shortening by ayg1p, a novel enzyme involved in fungal melanin biosynthesis. *J Biol Chem, 279,* 44613-44620.

Gasch, A.P., & Werner-Washburne, M. (2002). The genomics of yeast responses to environmental stress and starvation. *Funct Integr Genomics, 2,* 181-192.

Himejima, M., Hobson, K.R., Otsuka, T., Wood, D.L. & Kubo, I. (1992). Antimicrobial terpenes from oleoresin of Ponderosa pine tree *Pinus Ponderosa* - a defense mechanism against microbial invasion. *J Chem Ecol, 18,* 1809-1818.

Huang, X., & Madan, A. (1999). CAP3: A DNA sequence assembly program. *Genome Res, 9,* 868-877.

Idnurm, A., Howlett, B. J. (2001). Pathogenicity genes of phytopathogenic fungi. *Molecular Plant Pathology,* (2), 241-255.

Jasinski, M., Stukkens, Y., Degand, H., Purnelle, B., Marchand-Brynaert, J., & Boutry, M. (2001). A plant plasma membrane ATP binding cassette-type transporter is involved in antifungal terpenoid secretion. *Plant Cell, 13,* 1095-1107.

Keeling, C.I. & Bohlmann, J. (2006). Genes, enzymes and chemicals of terpenoid diversity in the constitutive and induced defense of conifers against insects and pathogens. *New Phytol, 170,* 657-675.

Kestler, H.A., Muller, A., Gress, T.M., & Buchholz, M. (2005). Generalized Venn diagrams: a new method of visualizing complex genetic set relations. *Bioinformatics, 21,* 1592-1595.

Kim, J.J., Allen, E.A., Humble, L.M. & Breuil, C. (2005). Ophiostomatoid and basidiomycetous fungi associated with green, red, and grey lodgepole pines after mountain pine beetle (*Dendroctonus ponderosae*) infestation. *Can J For Res, 35*, 274-284.

Loppnau, P., Tanguay, P. & Breuil, C. (2004). Isolation and disruption of the melanin pathway polyketide synthase gene of the softwood deep stain fungus *Ceratocystis resinifera*. *Fungal Genet Biol, 41,* 33-41.

Maloney, A.P., & VanEtten, H.D. (1994). A gene from the fungal plant pathogen *Nectria haematococca* that encodes the phytoalexin-detoxifying enzyme pisatin demethylase defines a new cytochrome P450 family. *Mol Gen Genet, 243,* 506-514.

Meerts, P. (2002). Mineral nutrient concentrations in sapwood and heartwood: a literature review. *Ann For Sci, 59,* 713-722.

Paine, T.D. & Hanlon, C.C. (1994). Influence of oleoresin constituents from *Pinus ponderosa* and *Pinus jeffreyi* on Growth of Mycangial Fungi from *Dendroctonus ponderosae* and *Dendroctonus jeffreyi*. *J Chem Ecol, 20,* 2551-2562.

Palmer, A. G., Gao, R., Maresh, J., Erbil, W. K., & Lynn, D. G. (2004). Chemical biology of multi-host/pathogen interactions: chemical perception and metabolic complementation. *Annu Rev Phytopathol*, (42), 439-64.

Parker, J.E., & Scott, D.B. (2005). Indole diterpene biosynthesis in ascomycetes fungi. *Handbook of Mycology,* Vol. 22 (An, Z. ed.) pp. 405-426. CRC Press, New York.

Raffa, K.F., & Berryman, A.A. (1982). Physiological differences between lodgepole pines resistant and susceptible to the Mountain Pine Beetle (Coleoptera, Scolytidae) and associated microorganisms. *Environ Ent, 11,* 486-492.

Ro, D.K., Arimura, G., Lau, S.Y., Piers, E., & Bohlmann, J. (2005). Loblolly pine abietadienol/abietadienal oxidase PtAO (CYP720B1) is a multifunctional, multisubstrate cytochrome P450 monooxygenase. *Proc Natl Acad Sci U S A, 102,* 8060-8065.

Robinson, R.C. (1962). Blue stain fungi in lodgepole pine (*Pinus contorta* Dougl. var. *latifolia* Engelm.) infested by the Mountain Pine Beetle (*Dendroctonus monticolae* Hopk.). *Can J Bot, 40,* 609-614.

Shrimpton, D.M., & Whitney, H.S. (1968). Inhibition of growth of blue stain fungi by wood extractives. *Can J Bot, 46,* 757-761.

Shrimpton, D.M. (1973). Extractives associated with wound response of lodgepole pine attacked by the mountain pine beetle and associated microorgansims. *Can J Bot, 51,* 527-534

Snoeijers, S.S., Pérez-Garcia, A., Joosten, M.H.A.J., & De Wit, P.J.G.M. (2000). The effect of nitrogen on disease development and gene expression in bacterial and fungal plant pathogens. *Eur J Plant Path, 106,* 493-506.

Urban, M., Bhargava, T. & Hamer, J.E. (1999). An ATP-driven efflux pump is a novel pathogenicity factor in rice blast disease. *EMBO J, 18,* 512-521.

Werck-Reichhart, D. & Feyereisen, R. (2000). Cytochromes P450: a success story. *Genome Biol, 1,* R3003.

Winnenburg, R., Baldwin, T.K., Urban, M., Rawlings, C., Kohler, J. & Hammond-Kosack, K.E. (2006). PHI-base: a new database for pathogen host interactions. *Nucleic Acids Res, 34,* D459-464.

Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S., Bussey, K.J., Riss, J., Barrett, J.C., & Weinstein, J.N. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol, 4,* R28.

3 *De novo* genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data[2]

3.1 Introduction

The efficiency of *de novo* genome sequence assembly processes depend heavily on the length, fold-coverage and per-base accuracy of the sequence data. Despite substantial improvements in the quality, speed and cost of Sanger sequencing, generating a high quality draft *de novo* genome sequence for a eukaryotic genome remains expensive. New sequencing-by-synthesis systems from Roche (454), Illumina (Genome Analyzer) and ABI (SOLiD) offer greatly reduced per-base sequencing costs. While they are attractive for generating *de novo* sequence assemblies for eukaryotes, these technologies add several complicating factors: they generate short (typically 450 bp for 454; 50 to 100 bp for Illumina and SOLiD) reads that cannot resolve low complexity sequence regions or distributed repetitive elements; they have system-specific error models; and they can have higher base-calling error rates. To this point, then, *de novo* assemblies that use either 454 data alone, or that combine 454 with Sanger data in a 'hybrid' approach have been reported only for prokaryote genomes, and no *de novo* assemblies that use Illumina reads, either alone or in combination with Sanger and 454 read data, have been reported for a eukaryotic genome.

[2] A version of this chapter has been published. Scott DiGuistini, Nancy Y. Liao, Darren Platt, Gordon Robertson, Michael Seidel, Simon K. Chan, T. Robert Docking, Inanc Birol, Robert A. Holt, Martin Hirst, Elaine Mardis, Marco A. Marra, Richard C. Hamelin, Jörg Bohlmann, Colette Breuil and Steven J.M. Jones.(2009) *De novo* genome sequence assembly of a filamentous fungus using Sanger, 454, and Illumina sequence data. Genome Biology, 10:R94 doi:10.1186/gb-2009-10-9-r94.

In principle, it should be possible to generate a *de novo* genome sequence for a eukaryotic genome by combining sequence information from different technologies. However, the new sequencing technologies are evolving rapidly, and no comprehensive bioinformatic system has been developed for optimizing such an approach. Such a system should flexibly integrate read data from different sequencing platforms, while addressing sequencing depth, read quality and error models. Read quality and error models raise two challenges. First, while it is desirable to identify a subset of high quality reads prior to genome assembly, and established read quality scoring methods exist for Sanger sequence data, there are no rigorous equivalents for 454 or Illumina reads (Huse et al, 2007). Second, error models differ between different sequencing technologies.

A number of genome assemblers are currently available for combining Sanger and 454 read collections, as well as specialized short read assembly programs like ALLPATHS, SSAKE, Velvet and ABySS (Butler et al, 2008; Warren et al, 2007; Zerbino & Birney, 2008; Simpson et al, 2008). However, short reads require greater sequencing depth to ensure specificity in read overlaps, as shorter overlaps cause ambiguities in the assembly stage. This increased sequence depth prevents both applying the traditional overlap-layout-consensus method directly and extending Sanger/454 hybrid assemblers to use ultra-short reads. Assemblers that are primarily intended for short reads can process deep coverage read data; however, because read length and software limitations restrict the unambiguous sequence regions that they can assemble and they currently lack the capacity for scaffolding contigs effectively, they are typically limited to ultra-short reads. When we assessed such assemblers, the above challenges - likely compounded by the high error rate in our earlier Illumina read collections - resulted in contigs that were either too short or too unreliable to support comparing homologous blocks of sequence between genomes.

The Forge genome assembler (Forge http://sourceforge.net/projects/forge/) was designed for assembling combinations of reads from Sanger and ‘next-generation’ sequencing technologies, and attempts to address the above challenges. Distributed memory hash tables and pruned overlap graphs allow its classical overlap-layout-consensus approach to handle large data sets with deep coverage. Simulation techniques embedded in the algorithm allow it to automatically adapt to varying read lengths and error characteristics to accommodate rapidly changing performance in next-generation sequencing platforms.

In the work described here, we developed a hybrid approach that uses Forge for generating *de novo* draft genome sequences, and applied the approach to a filamentous fungus, *Grosmannia clavigera (Gc)*. To generate the draft sequence, we combined: conventional, 40-kb fosmid paired-end (PE) Sanger reads from an ABI 3730xl sequencer, single-end (SE) 454 reads from Roche GS20 and GS-FLX sequencers; and PE reads from an Illumina Genome Analyzer (GA_{ii}) sequencer. The current sequence assembly is approximately 32.5 Mb in length and has an N50 scaffold size of approximately 782 kb. The assembly as well as the raw read data are available from National Center for Biotechnology Information (NCBI; see Materials and methods).

We describe how we prepared read data for assembly by filtering and trimming using an internally developed pipeline which we make available (Pipeline scripts ftp://ftp.bcgsc.ca/supplementary/Grosmannia_clavigera/tools/). We outline below our experience in assembling this eukaryotic genome using the Velvet and Forge assemblers. We also describe a bioinformatic approach for assessing the accuracy of such hybrid assemblies when no high quality reference sequence exists.

3.2 Materials and methods

3.2.1 Library construction and sequencing

Gc spores from strain kw1407 (Lee, Kim & Breuil, 2006) were spread onto cellophane overlaid on 1.5% agar containing 1% malt extract in 15 cm petri dishes. The fungal spores were incubated at 22°C in the dark for 8 days, and the mycelia were removed from the cellophane and pooled. DNA was extracted from mycelia following the method of Möller *et al.* (1992), but without first lyophilizing the mycelia. For constructing a 40-kb fosmid library, fungal DNA was randomly sheared, then blunt-end repaired and size-selected by electrophoresis on a 1% agarose gel. Recovered DNA was ligated to the pEpiFOS-5 vector (Epicentre Biotechnologies, Madison, WI), mixed with Lambda packaging extract and incubated with host *Escherichia coli* cells. Clones containing inserts were selected and paired-end-sequenced on an ABI 3730xl. For sequencing on the Roche GS20 or GS-FLX sequencers, DNA was prepared using the methods described by Margulies *et al.* (2005). For preparing the approximately 200-bp library on the Illumina GA$_{ii}$ sequencer, 5 µg of DNA was sonicated for 10 minutes, alternating 1 minute on and 1 minute off, using a Sonic Dismembrator 550 (Fisher Scientific, Ottawa, CAN). Sonicated DNA was then separated in an 8% PAGE. The library was constructed from the eluted 190- to 210-bp fraction of DNA using Illumina's genomic DNA kit, following their protocol (Illumina, San Diego, CA, USA). Four lanes in a single flow-cell were sequenced to 42 cycles using v.1 sequencing and cleavage reagents. Data was processed using Illumina's GA pipeline (v.0.3.0 beta3).

3.2.2 Filtering Sanger and 454 reads

For Sanger PE data, we removed reads that had less than 200 bp of continuous sequence with a minimum quality score of Phred 20; 14,522 reads with an average read length of approximately 600 bp remained. We discarded 454 reads that contained uncalled base positions (no-calls), then pooled reads into separate GS20 and GS-FLX sets. After assessing the two read length distributions, we discarded reads whose lengths were either less than 40 bp or longer than 200 bp, or less than 50 bp or longer than 350 bp from the GS20 and GS-FLX sets, respectively, as described by Huse *et al.* (2007). We then applied a low complexity filter to the 454 and Sanger reads using DUST with a 50% threshold (DUST ftp://ftp.ncbi.nlm.nih.gov/pub/tatusov/dust/). Contamination filtering was performed against a database of bacterial genome sequences. From the initial GS20 read collection approximately 3% of reads were identified with 98% or greater similarity to the genome sequence of *Anaerostipes caccae* and were removed. Lastly, 454 reads were mapped against the Univec database (NCBI http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html) using BLAST to trim and filter library adaptor sequence; 3% of reads were removed and approximately

7.5 Mb of sequence were trimmed from the read collection with no significant difference in the pre- and post- trimming read length (163 bp).

3.2.3 Assembling Illumina data

Version 7.31 of Velvet is able to generate scaffolded contigs, which results in larger N50 values; however, we were unable to observe scaffolding resulting from our hybrid Sanger/Illumina read assembly. Further, comparing Illumina-only assemblies generated from previous and current Velvet versions to our reference sequence indicated that the contig merging increased the number of assembly errors (data not shown). Given our assembly strategy, the limitations of the Velvet v. 7.31 release indicated that we should continue using Velvet v. 6.04 for our current work.

Because eukaryotic genomes pose an increasing number of ambiguous sequence regions compared with prokaryotes, and because we had generated relatively deep sequence coverage for the 200-bp Illumina library we used the highest available assembly *k*-mer parameter (hash length) of 31 for all Velvet assemblies reported here. We calculated expected coverage and the coverage cut-off parameters as described in the Velvet documentation.

We applied a simple paired-read analysis to identify chimeric pairs that we believed to be artifacts of library construction and sequencing. We have termed these ‘shadow’ reads. Briefly, we identified a shadow read pair when a read shares X identical starting bases with its mate, where we tested X equals 6, 8, 10, 12, 14, 16, 20 or 24. We discarded such read pairs with 6 bp or greater shared sequence.

We tested trimming and filtering on the Illumina reads used for assembly and developed a QRL metric using the calibrated Illumina Phred-like quality scores. We calculated a read’s QRL as follows. Moving from the 3' towards the 5' end of a read, we used the highest probability score value for each base position to determine a quality score for that base. The maximum possible value for this score is 40. For each read, the QRL was the length between the first and last bases that were above a quality score threshold.

We assessed the Velvet assemblies using four metrics: N50, the scaffold (contig) length for which 50% of the assembled genome is in scaffolds (contigs) that are at least as long as N50; the assembly size, calculated by adding the total length of retained contigs or scaffolds; alignment of the assembly contigs against a manually finished reference sequence; and alignments of expressed sequence tags (ESTs) to the assemblies, using a set of 7,169 unique ESTs (each EST was selected as the member with the longest PHRED 20 read length from multiple sequence alignments generated by clustering approximately 43 k EST reads) generated in ongoing and previous work (DiGuistini et al, 2007). We aligned ESTs using BLASTn with an E-value threshold of 1e-50, and differentiated complete alignments from resolvable and irresolvable partial alignments. Resolvable partial alignments were alignments occurring on a contig edge that could be merged with another partial alignment on a complimentary contig edge. Irresolvable partial alignments were alignments in which the partial EST alignments were isolated in the interior sequence region of a contig such that it was not possible to join the complementary alignments. For identifying small insertions and deletions from the same BLAST report, a custom PERL script was used to parse the alignment data; insertions were identified as gaps in the EST query alignment, deletions were identified from gaps in the target side of the alignment. Several of these contig assemblies were then tested in Forge and further assessed using the methodology described below.

Velvet assemblies took approximately 3 hours on a server with two 2.2 GHz dual-core AMD Opteron 275 processors and 8 GB of RAM. Velvet assemblies handled by Forge were assigned base quality scores as uniform PHRED 20 at each position.

3.2.4 Forge hybrid assembler and genome assembly analysis

The Forge output is a consensus sequence with quality scores and a complete multiple sequence alignment for all reads, with locations in a tabular format that can be converted into the Consed ‘ace’ file format (Gordon, Abajian & Green, 1998).

We assessed scaffold qualities using the 42-bp PE reads rejected by the filtering process described above in assembling the draft genome sequence. We aligned the reads to the draft assemblies using MAQ (Li, Ruan & Durbin, 2008) in paired-end mode. We processed the output and identified PE relationships using custom PERL scripts and the Vancouver Short Read Analysis Package (Fejes et al, 2008). We separated the aligned reads into three subsets: PE reads that were correctly spaced and oriented and aligned on the same scaffold; PE reads that were aligned on separate scaffolds; and unpaired read alignments. We used clusters of read-alignment pairs to identify pairs that could be used to merge scaffolds and to identify low quality assembly regions. The first type had a read cluster located at a scaffold edge and a mate-pair cluster located on a complementary scaffold edge. In the second type the complementary cluster was located in the interior scaffold sequence region such that the complementary clusters could not be joined. Because PE read mates can be incorrectly paired in the Illumina flowcell image analysis pipeline, and base-calling errors or low-complexity sequences can result in read placement errors by MAQ (Gordon, Abajian & Green,1998), we required cluster sizes of at least 10 before using a cluster to mark a potential scaffold merge or to identify a low-quality region.

As described above, the EST collection was aligned to the Forge assemblies for quality control and alignments were generated against the manually finished genome sequence using nucmer within the MUMmer package with the seed cluster parameter (-c) set to 750. Read coverage, repeat data and quality data were then combined and visualized using Circos (Krzywinski et al, 2009). RepeatMasker (www.repeatmasker.org) was used for preliminary filtering of repetitive elements against repbase (v. 14) with the species parameter set to ‘fungi/metazoa group’ prior to gene prediction. Gene prediction was done using Augustus (Stanke et al, 2006). The Forge hybrid assemblies were generated using the following settings: a genome size estimate of 35 Mb and a hash table size of 80 M for assemblies generated from Sanger/454 read data only or those that included preassembled Illumina PE read data and 260 M for the assembly with direct integration of the Illumina PE read data. The

Forge assemblies took 10 to 84 hours on a Linux server cluster using 40 nodes ranging from dual 2.0 GHz processors with 2 GB of RAM to quad-core 2.6 GHz processors with 16 GB of RAM.

3.2.5 Generating the GCgb1 genome sequence

We generated a reference genome sequence and used it for *de novo* assembly verification by using the methodology described above, we added 10,000 additional Sanger fosmid PE reads and approximately 7.6 M, 50 bp Illumina PE reads (see appendix B.3 in additional data file 1). After assembling these data with Forge and applying manual editing, primer walking and other standard finishing techniques; the largest and tenth largest contigs of the resulting genome sequence were 2.33 and 0.68 Mb long, respectively. The largest scaffold was approximately 2.9 Mb and the scaffold N50 was approximately 950 kb. Eighty five percent of the genome sequence was contained within the top 29 scaffolds.

3.3 Results

3.3.1 Generating sequence data

We assembled a genome sequence for *Gc* using the pipeline described below and in Figure 3.1. We first constructed a fosmid library, from which we generated 18,424 Sanger PE sequences (approximately 0.3-fold genome sequence coverage). We then used sheared genomic DNA to generate seven read sets on Roche GS20 and GS-FLX sequencers, producing 3,045,953 reads with 100.0 and 224.5 bp average lengths, respectively (250 Mb of sequence data; approximately 7.7-fold genome sequence coverage). Finally, we supplemented these data sets with PE, 42-bp reads (82,655,316) for a single library of approximately 200-bp sheared genomic DNA fragments on an Illumina GA$_{ii}$ (approximately 3.3 Gb of sequence data; approximately 100-fold genome sequence coverage).

3.3.2 Initial assembly analysis

Initially, Illumina PE read data required preassembly, as we were unable to complete a Forge (v.20090319) run using our entire read collection; we integrated these data by preassembling them with Velvet. We assembled the read data described above, alone or in combination, and devised a strategy for refining these assemblies. Using Velvet (v.6.04 and v.7.31) we assessed assemblies generated from Illumina PE read data and Illumina with Sanger PE read data (see Materials and methods: Assembling Illumina data); using Forge we assessed assemblies generated from 454 SE read data, 454 SE with Sanger PE read data and 454 SE and Sanger PE read data plus a Velvet-preassembled contig backbone. We used a collection of 7,169 unique EST sequences to do an initial assessment of these assemblies. From the EST-to-genome alignments, we determined the number of complete alignments as well as the number of times an alignment was split between contigs in a resolvable (‘partial’) or unresolvable (‘misassembly’) manner (described in Materials and methods), and also identified small insertions or deletions (termed indels). The Velvet assembly generated from Illumina PE data alone yielded an N50 contig length of approximately 24.5 kb, and covered approximately 26.7 Mb of the 32.5 Mb manually finished genome sequence (Table 3.1).

In contrast, a Forge assembly of the 454 read collection yielded an N50 contig length of approximately 7.8 kb and covered approximately 29.5 Mb of the complete genome sequence (Table 3.2). We checked the overlap between these assemblies, and found that 100% of the Velvet-Illumina assembly was contained within the Forge-454 assembly, while the 454 assembly contained an additional approximately 2.5 Mb of sequence that was not found in the Illumina assembly.

Comparing indels across assemblies indicated that the rate at which small (1 to 5 bp) insertions or deletions appeared in the assembled consensus sequence depended on the fraction of 454 data in the assembly (Figure 3.2). When we inspected the frequency of each base that was inserted or deleted across all assemblies that used 454 read data, the pattern was consistently A>T>C>G, while Velvet assemblies of Illumina reads produced a C>A>T>G indel pattern where A, C, G, and T represent indel frequencies for their corresponding bases. To assess whether these small insertions and deletions could disrupt the phasing of the assembled genome sequence (that is, the periodicity of nucleotide sequences within the assembly relative to *cis* factors), we examined the predicted protein collections from each of these assemblies. Average predicted protein sequences contained 401.1 versus 527.0 amino acids in assemblies that used only 454 or only Illumina data, respectively. Although this difference could be the result of an increased contig N50 length in the Illumina based assembly (Table 3.1 and 3.2), we observed that, in the NCBI non-redundant database (NCBI www.ncbi.nlm.nih.gov), the fraction of predicted protein sequences with at least one significantly similar sequence was 60% for the 454-only assembly but 70% for the Illumina-only assembly. This suggests that the shorter average protein lengths in assemblies with greater ratios of 454 reads were due to spurious peptide sequences and not contig end truncations. Assemblies that used 454 read data achieved greater amounts of total assembled DNA, including relatively more sequence annotated with repetitive elements, despite shorter contig N50 values; the 454 assembly and the Sanger-454-Illumina assembly were annotated with approximately equal numbers of repetitive elements, while the Velvet assembly had approximately half as many annotations. Because the 454 assemblies also had acceptably low EST-detectable misassembly rates, we concluded that a strategy that combined all three read types would be optimal. We assessed validating our assembly methodology using simulation, but found that the results did not accurately reflect the outcomes of working with real read data. This was likely due to the difficulty of accurately modelling read-specific sequence quality and errors (results not shown).

3.3.3 Optimizing Sanger/454 assemblies using 454 read filtering

Filtering 454 SE reads for no-calls, length and sequence complexity incrementally improved the overall quality of the *de novo* assembled *Gc* genome sequence relative to a manually finished sequence, which we will refer to as *GC*gb$_1$ (see Materials and methods for a description). For 454 SE reads, no-call filtering removed 95,833 (3%) reads, and length filtering further removed 141 (0.009%) GS20 reads and 3,583 (0.2%) GS-FLX reads. Applying these filtering strategies reduced both the contig and scaffold N50s, suggesting that when a hybrid assembly includes relatively low 454 SE sequence coverage, filtering reads by no-calls and length may be overly aggressive. However, for our strategy of assembling Sanger PE and 454 SE read data around high-coverage Illumina read data, the two filtering steps were worthwhile; applied together, they improved the integration of the different sequence types and reduced the number of chimeric contig ends by 20% (See appendix B.1).

Low complexity regions (that is, genome sequences with a simple repetitive composition) are expected features for a filamentous fungus. We found that reads containing such sequences were associated with misassemblies (data not shown). Using DUST (ftp://ftp.ncbi.nlm.nih.gov/pub/tatusov/dust/), we filtered 522 of the Sanger reads and 3,889 of the 454 reads containing such repetitive composition. Filtering 454 and Sanger reads for low complexity sequences marginally affected contig and scaffold N50; however, it reduced the number of scaffolds containing gaps from 685 to 666, and decreased the number of irresolvable split EST alignments by 7. Given this, we removed reads containing low complexity sequence from the draft assemblies. We intend to resolve such regions in the finishing stage of the sequencing project, using tools and resources that are better suited for such genomic elements.

3.3.4 Improving assemblies with Illumina PE reads by trimming and filtering

Given the promising initial assembly of the Illumina PE read data, we assessed trimming and filtering as a means to improve the Velvet assembly accuracy. Beginning with the 82.6 M, 42-bp PE reads, we discarded 1.1 M reads containing no-call bases and 1.9 M shadow reads (described in Materials and methods). To optimise the Velvet assembly, we used alignments with our preliminary 454 and Sanger sequence assembly to determine trimming and quality read length (QRL; described in Materials and methods) filtering parameters for removing low quality bases from reads (appendix B.2; Figure B4A).

As determined by EST alignments and alignments to *GC*gb$_1$, trimming and filtering improved the accuracy while only marginally reducing the total length of DNA assembled; however, more aggressive read trimming and filtering substantially reduced the contig N50s in Velvet assemblies (Table 3.1). Trimming Illumina reads from 42 bp to 38 bp (T38) and then to 36 bp (T36) reduced the assembly N50 to 10.7 kb and 2.9 kb, respectively. For the T36 assembly, trimming reduced the total amount of assembled sequence and the number of complete EST-to-assembly alignments, while also reducing the number of EST-detectable assembly errors from 29 to 11 (Table 3.1). Trimming Illumina reads also reduced the effective level of coverage, which likely explains why the N50 and complete EST-to-genome alignments were reduced. Given this, we assessed whether the improvements in EST-detectable assembly errors could also have resulted from arbitrary read trimming and subsequent shortening of the assembled contig lengths. We tested this by removing 6 bp from the 5' end of each read. In the resulting assembly the N50 and complete EST-to-genome alignment counts were approximately half of the corresponding values for the T36 assembly, and the EST-detectable error rate was five times higher, validating the efficiency of our trimming algorithm.

Filtering low quality data (QRL(Q10) = 28) resulted in an assembly that, relative to the T36 assembly, had a smaller N50 (1,299 bp) but only a marginally lower number of EST-detectable assembly errors. We then tested whether filtering by randomly removing the same number of reads that had been removed by QRL filtering changed the resulting assembly. We found that although random filtering did not substantially change N50, it tripled the number of EST-detectable errors and doubled the number of ESTs with no genome assembly alignment, validating the efficiency of our filtering algorithm.

Relative to *GC*gb_1, we found that this trimmed and filtered Illumina read collection yielded the most accurate Velvet contigs and that these contigs had approximately 15% fewer chimeric contig ends. Using the approximately 51 M Illumina PE reads resulting from trimming and filtering (approximately 56.5x genome sequence coverage) and the Sanger and 454 data reported above we attempted two assemblies using a revised version of Forge (v.20090526). We tested: incorporating the Illumina PE data following Velvet preassembly, (Sanger-454-IlluminaPA) and incorporating the Illumina PE data directly, (Sanger-454-IlluminaDA). EST-to-genome sequence alignments and Illumina PE read alignment cluster analysis showed that the Sanger-454-IlluminaDA genome sequence had a lower misassembly rate than the Sanger-454-IlluminaPA assembly (Table 3.2). However, alignment to *GC*gb_1 suggested that the Sanger-454-IlluminaPA was a more accurate assembly in regards to long range continuity (Figure 3.3). The Sanger-454-IlluminaDA assembly had greater contig N50 whereas the Sanger-454-IlluminaPA assembly had greater scaffold N50 (Table 3.1).

3.3.5 Assessing the final assembly

Assembly Sanger-454-IlluminaPA had 6,314 complete EST alignments and 40 EST-detected assembly errors. The number of scaffolds containing gaps greater than 1 kb, 163, was substantially lower than the 656 in the best assembly achieved without the Illumina PE read data. We assessed the quality of this Forge hybrid assembly using the consistency of the Sanger PE read pairings and 200-bp Illumina PE reads. Adding the Illumina PE read data increased the fraction of consistently-paired Sanger PE reads from 64 to 81% for Sanger-454-IlluminaPA versus the best assembly without Illumina PE read data; for Illumina PE alignment data, the numbers of unpaired reads decreased by 37% and those paired on different scaffolds decreased by 21%, while the number of paired reads on the same scaffold with an appropriate fragment length increased by approximately 1.5 M. The assembly contained 46 scaffolds longer than 100 kb, which represented 88.5% of the total genome sequence. These scaffolds had a G+C content of 53.2%. The 10 largest scaffolds contained 48 gaps with a total length of approximately 181 kb (appendix B, Figure B5). The longest scaffold was approximately 3.67 Mb and the tenth longest scaffold was approximately 782 kb.

The 454 read coverage and Sanger PE read placements for assembly Sanger-454-IlluminaPA indicate that the distribution of read data was generally uniform across the top ten scaffolds (appendix B, Figure B5). We noted 12 sequence regions with unexpectedly high read coverage. Preliminary analysis of these sequence regions indicates that, as expected, they were spanned by repetitive elements, primarily transposons. Large gene families with high levels of similarity were also problematic.

However, there is no evidence that such genomic elements necessarily ended up in misassemblies; rather, they sometimes caused early contig growth termination by making the collapsed sequence data unavailable to other appropriate genomic regions. Misassemblies primarily occurred when the repeat span was large and fosmid collapses brought incorrect contigs into adjacency during scaffolding. However, these are easily identified and corrected during sequence finishing.

Assessing the final draft assembly using the 200-bp Illumina PE read set highlighted genomic regions with collapsed repetitive elements, low coverage, misassemblies, and adjacent scaffolds. The PE alignment data were plotted by coverage and shown in Figure B5 in appendix B. Correctly paired read alignments had a mean outside distance of 193 bp and appeared to be evenly distributed across the scaffolds. However, approximately 1,500 anomalous PE read-alignment-clusters (that is, reads with overly stretched gap distances between pairs, unpaired reads or reads paired inappropriately on different scaffolds) highlight that automated rules can be applied to the current draft assembly, and we have implemented a semi-automated system in our finishing pipeline to leverage this data. In *GC*gb$_1$, we have currently resolved > 90% of the anomalous clusters identified in Sanger-454-IlluminaPA. As expected, many (approximately 85%) of the ambiguities that arose during our analysis of PE read clusters occurred at scaffold edges (< 3 kb), suggesting that scaffold growth termination was accurate in this assembly; further, scaffold growth was constrained by read ambiguity rather than by low coverage. Although greater sequencing depth could improve this by allowing better resolution of read overlap alignments, some types of genomic elements will likely continue to cause ambiguity in read overlaps, leading to premature truncation of scaffold growth.

By counting complete gene models for core eukaryotic proteins reported by CEGMA (Parra, Bradnum & Korf, 2007), we estimated that we have generated gene models for greater than 94% of the full genome's hypothetical gene model collection. For the preliminary Sanger-454-IlluminaPA gene predictions, the average gene density was approximately 1 gene/3.5 kb, the average gene length was approximately approximately 1.5 kb, the average transcript length was approximately 1.2 kb, and the average transcript G+C content was approximately 58%. Similar values have been reported for other ascomycetes from the order sordariomycetes (Galagan et al, 2003; Dean et al, 2005). A detailed description and annotation of the *Gc* genome will be published separately (manuscript in preparation).

3.3.6 Analysis of Illumina and 454 read data

We used the manually finished *GC*gb$_1$ assembly to assess the performance of the Illumina and 454 sequencing platforms (Figure 3.4). We quantified the efficiency of discovering new and useful sequence data, as well as the rate at which the new sequence data covered *GC*gb$_1$. We performed this analysis on all possible read substrings with length 28 bp (termed *k*-mers) generated from the raw reads rather than on the raw reads themselves. Although the rate at which novel *k*-mers were discovered was approximately the same for both technologies at lower numbers of *k*-mers, when we split the analysis of novel *k*-mers into those that appeared at least twice versus once, a greater error rate was observable in the Illumina *k*-mer collection (Figure 3.4A). Because the 454 read lengths were longer, the unique *k*-mers generated from this read collection overlapped each other more than *k*-mers generated from the Illumina reads. This was inherent in the *k*-mer sampling process and likely explains the slower gain in 454 genome coverage (Figure 3.4B). Our data was insufficient for systematically assessing library saturation; however, it was apparent that the large number of reads generated for either library captured the entire genome sequence we assembled (Figure 3.4B). Based on EST-to-genome alignments, approximately 0.6% of the protein coding sequence were missing or ambiguous in *GC*gb$_1$. This could suggest that a portion of the genome remains ambiguous to our assembly methodology or that read data is missing from our sequence set. Given the rapid development of wet lab methodologies, it will be interesting to see whether library saturation remains a challenge for *de novo* genome sequencing.

3.4 Discussion

We sought to rapidly generate a *de novo* genome assembly that supported high quality protein coding gene predictions, wet lab experiments, comparative genomics and sequence finishing for a eukaryotic organism. We used a hybrid approach for sequencing and assembly. We combined Sanger PE, 454 SE and Illumina PE sequence data, and developed an assembly strategy that was adaptable to evolving technologies, tools and methods. Using Forge we generated a draft genome sequence with a length of approximately 32.5 Mb, which had a contig N50 length of approximately 32 kb and a scaffold N50 length of approximately 782 kb. During this work, read lengths and read quality improved for 454 and Illumina platforms; as they changed, we evaluated different ways of processing Illumina sequence reads in order to integrate them into assemblies. We characterised the accuracy of the draft assemblies by aligning ESTs, Illumina PE reads and a manually finished sequence to them.

We chose Forge as the assembler for three reasons. First, it can flexibly integrate different sequencing technologies by automatically adapting alignment parameters for particular read error models. This facilitates using it with evolving sequencing technologies and variable, technology-specific read or contig preprocessing. Second, it is capable of integrating PE information directly into the contig-building and merging processes, making it ideally suited for processing abundant short paired reads. Finally, because it can be run on computer processors running in parallel, it can be applied to the relatively large data sets generated by next-generation platforms. From our initial observations, Forge assemblies were promising they integrated Illumina PE read data directly, yielded accurate assemblies with good long range continuity.

Although Forge was designed to accommodate the 454 scoring system, the vendor-supplied quality scores do not indicate the probability that a base is called correctly. While this shortcoming can be addressed by transforming the scores into a Phred-like scale similar to that used for Sanger reads (Brockman et al, 2008), we chose an empirical approach and rejected problematic data (Huse et al, 2007). We found that by aggressively applying no-call and length filtering we could improve the overall quality of the assembly, as measured by alignments to the *GC*gb$_1$ sequence, reduced gap sizes and fewer EST-detectable misassemblies. Low complexity filtering was especially useful for the 454 SE read data because, without read pairing information to anchor ambiguous overlaps, accurate read placement appeared difficult to resolve. Although we substantially improved the assemblies using these methods, 454 base calling inaccuracies in the vicinity of homopolymer runs continued to cause phasing problems that affected gene predictions in the assembled consensus sequence. We found that adding Sanger PE reads,Velvet contigs and then Illumina PE reads directly into the assembly progressively improved the consensus sequence by reducing the frequency of these indels. We also found that aligning a collection of Illumina-based assemblies back to the final assembly in a post-processing step accurately identified and resolved these homopolymers.

Given the promising initial assembly of Illumina PE reads, we further assessed how to improve the accuracy of Velvet-assembled contigs. Profiles of read quality and substitution error rate relative to the Sanger/454 preliminary assembly suggested that trimming the 42-bp Illumina reads would improve the assembly accuracy. While trimming reads at position 36 resulted in a lower N50, EST and reference sequence alignments showed that this assembly contained fewer errors; further, these contigs yielded a more accurate Forge assembly than either those with reads trimmed at position 38 or untrimmed. Importantly, adding the Illumina data to Forge assemblies substantially reduced the number of scaffolds and contigs, suggesting that these relatively inexpensive reads contributed additional data and encouraged contig growth and merging.

Forge uses a statistical model of overlap derived from internal simulations to determine the probability that two reads reliably overlap. This probability is systematically lowered or reduced to zero in repetitive regions, forcing Forge to rely on alternative information such as reads with mate pairs anchored in a scaffold, polymorphisms within a repeat family, or the combination of a low probability overlap and read-pair data. An important advance made with Forge during the course of our work was the ability to scale beyond 50 M reads, which enabled the direct integration of Illumina PE read data in a single Forge assembly stage. The increased accuracy of EST-to-genome alignments, Illumina PE read alignments and the significant increase in contig N50 of the resulting assembly likely resulted from the large amount of pairing information introduced by this data. This suggests that when abundant PE information is available, read sequence length is not as important a limitation as anticipated. Currently, one challenge of this assembly method appears to be in balancing out the PE information in the low coverage Sanger data versus the high coverage Illumina data. Although more Fosmid pairs were correctly assigned to the same scaffold in the Sanger-454-IlluminaPA assembly, a greater fraction of the fosmid read pairs had consistant pairing distances in the assembly generated from direct integration of the Illumina PE read data. We also detected fewer inconsistencies in the Sanger-454-IlluminaDA assembly using the Illumina PE alignment strategy. This could have resulted from working directly with the Illumina PE reads in the assembly stage versus working with read substrings (*k*-mers), which is typical in a short read assembler like Velvet. Working with read substrings is an abstraction that doesn’t enforce read integrity onto the contig consensus sequence. For the Illumina PE library reported here, read pairing distances were not distributed normally around the mean, and left hand tailing increased at greater pairing distances (appendix B, Figure B4B). Read pairs with zero gap distance were also noted and could cause occasional sequence deletions in Forge assemblies if not filtered out.

We also noted that although low quality reads did not improve the assembly of genome sequence and so should be filtered out, they remained valuable as PE alignments for assessing and finishing the draft genome sequence. We are assessing the use of additional Illumina PE sequence data to evaluate the quality of the draft genome assembly and to guide finishing. We identified high quality regions in the assembly by calculating the coverage of correctly paired Illumina PE reads, and used scaffold-spanning PE reads to identify possible ambiguities or misassemblies in the consensus sequence. For such assessments, Illumina PE data offer advantages over EST data: the large number of reads provides deeper coverage, and the sequence data include non-transcribed regions, which are typically more difficult to assemble. We were also able to use the PE data to map the boundaries of misassemblies and to link scaffold edges in the consensus sequence. Improved software tools for working with Illumina PE data will likely benefit both the assembly of draft genome sequences and the finishing of these drafts.

In conclusion, we assembled a draft genome sequence for a fungal pathogen using Illumina, 454 and Sanger sequence data. We found that the highest quality assemblies resulted from integrating the read and contig collections in a single round of assembly, using software that could coherently manage the varying read and contig lengths as well as the different error models. Aggressively filtering this high coverage data was an effective strategy for incrementally improving the resulting draft assemblies. We anticipate that the iterative approach that we describe will facilitate using rapidly improving sequencing technologies to generate draft eukaryotic genome sequences.

Tables and figures

Table 3.1 Velvet assemblies generated from Illumina GA$_{ii}$ read data. Assembly T42 was generated from the untrimmed, no-call and shadow filtered Illumina PE reads. Assemblies T38 and T36 were generated by trimming the last 4 and 6 bp respectively from the T42 read set. Assembly T36, QRL(Q10)=28 was generated with the T36 read set from which reads were removed if they failed the QRL(Q10)=28 quality region length filtering (see Materials and methods).

ID	T42	T38	T36	T36; QRL(Q10)=28
Total contigs	6,945	8,637	19,118	39,488
N50 contig	24,566 (N/A)	10,706	2,902	1,299
Total DNA (bp)	26,721,397	26,466,756	25,854,719	24,812,690
EST analysis*	6,585/29	6,204/24	4,657/11	2,923/9

[a] EST alignments are given as: Complete alignments/Misassemblies, see Materials and methods.

Table 3.2 Forge assemblies generated using Illumina, 454 and Sanger read data. The '454' assembly was generated using only 454 SE read data. The 'Sanger-454' assembly was generated by combining the Sanger PE and 454 SE read collections. The 'Sanger-454-IlluminaPA' assembly was generated by combining the Sanger PE and 454 SE read collections with preassembled (PA) contigs generated from Illumina PE reads with Velvet. The 'Sanger-454-IlluminaDA' assembly was generated by combining the Sanger PE and 454 SE read collections with Illumina PE reads (DA = direct assembly).

ID	**454**	**Sanger-454**	**Sanger-454-IlluminaPA**	**Sanger-454-IlluminaDA**
Total scaffolds[a]	7,860	4,805	2,293	1,374
N50 contig (scaffold)	5,773 (N/A)	7,440 (289,760)	32,214 (782,476)	187,326 (283,373)
Total DNA (bp)[b]	29,484,877	34,841,371	35,344,453	29,530,079
# of scaffolds with gaps[c]	0	656	163	75
Augustus predictions	10,555	10,230	8,912	8,476
EST analysis[d]	5,544/25	5,747/60	6,314/40	6,685/33

[a] Scaffolds included in this calculation contained two or more reads and were longer than 500 bp.

[b] Total DNA was calculated excluding gaps and was performed on scaffolds that contained two or more reads and were longer than 500 bp.

[c] Gaps included in this calculation were longer than 50 bp.

[d] EST alignments are given as: Complete alignments/Misassemblies, see Materials and methods.

Figure 3.1 Overview of the process for producing de novo assemblies.

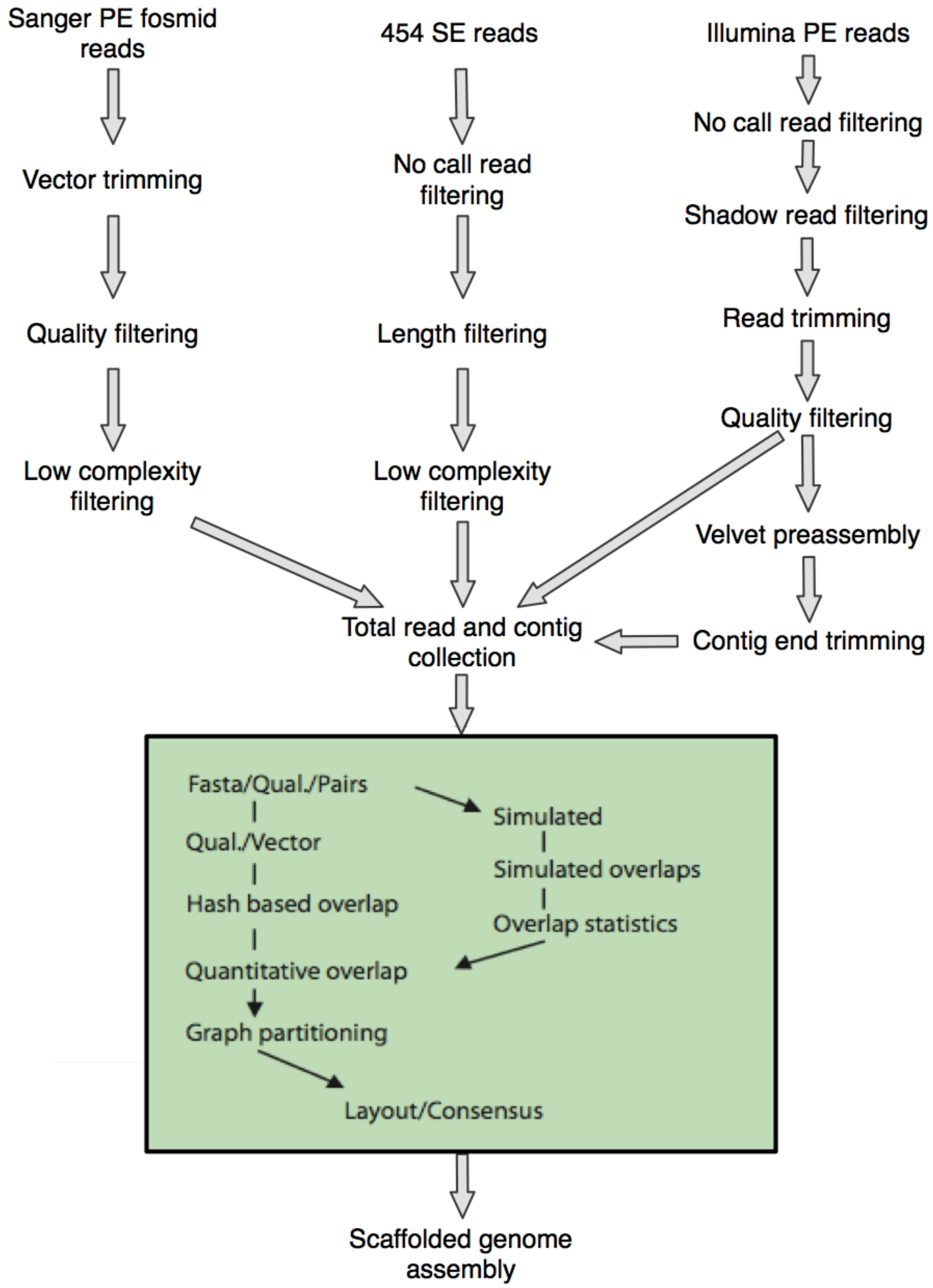


Figure 3.2 The proportion of 454 read data within the total read collection affected the number of small insertions and deletions (indels) based on analysis of 7,169 unique EST-to-genome alignments. The relative proportions of insertions (blue) and deletions (orange) in the assembly sequence are shown in the inset pie chart. Assemblies are described in Tables 3.1 and 3.2; those including 454 read data were assembled with Forge; the Illumina only assembly was generated with Velvet.

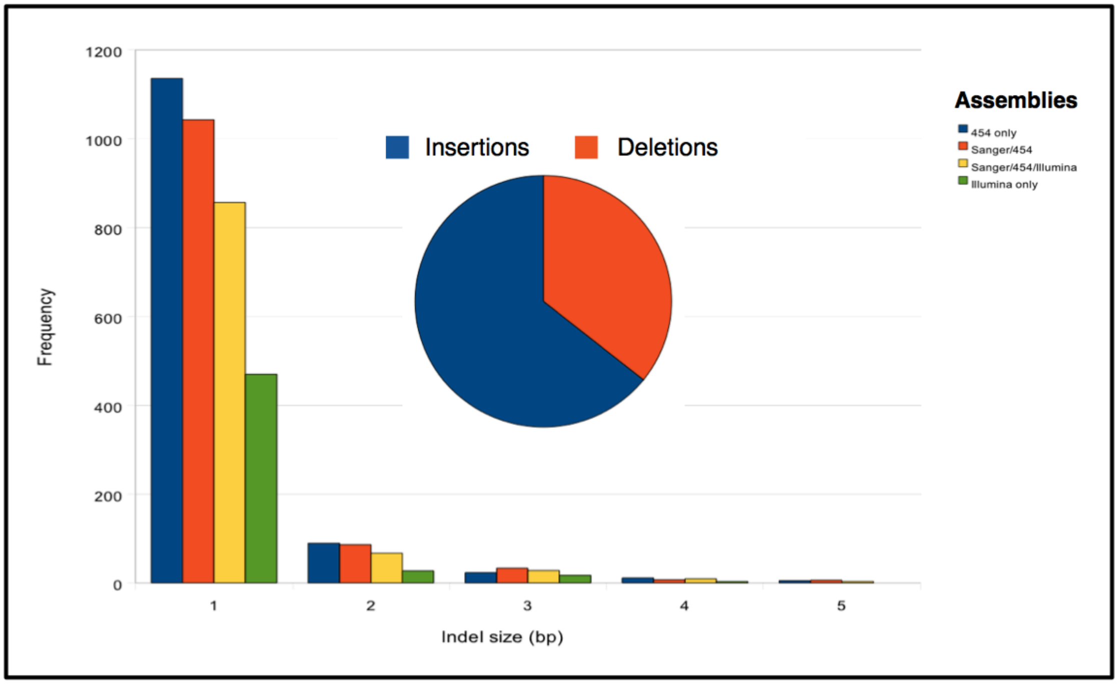


Figure 3.3 Alignments of scaffolds greater that 100 kb A. 'Sanger/454/IlluminaDA' (~24 Mb on 80 scaffolds) and B. 'Sanger/454/IlluminaPA' (~28.7 Mb on 46 scaffolds) on the Y-axis against the manually finished genome sequence (GCgb1) on the X-axis.

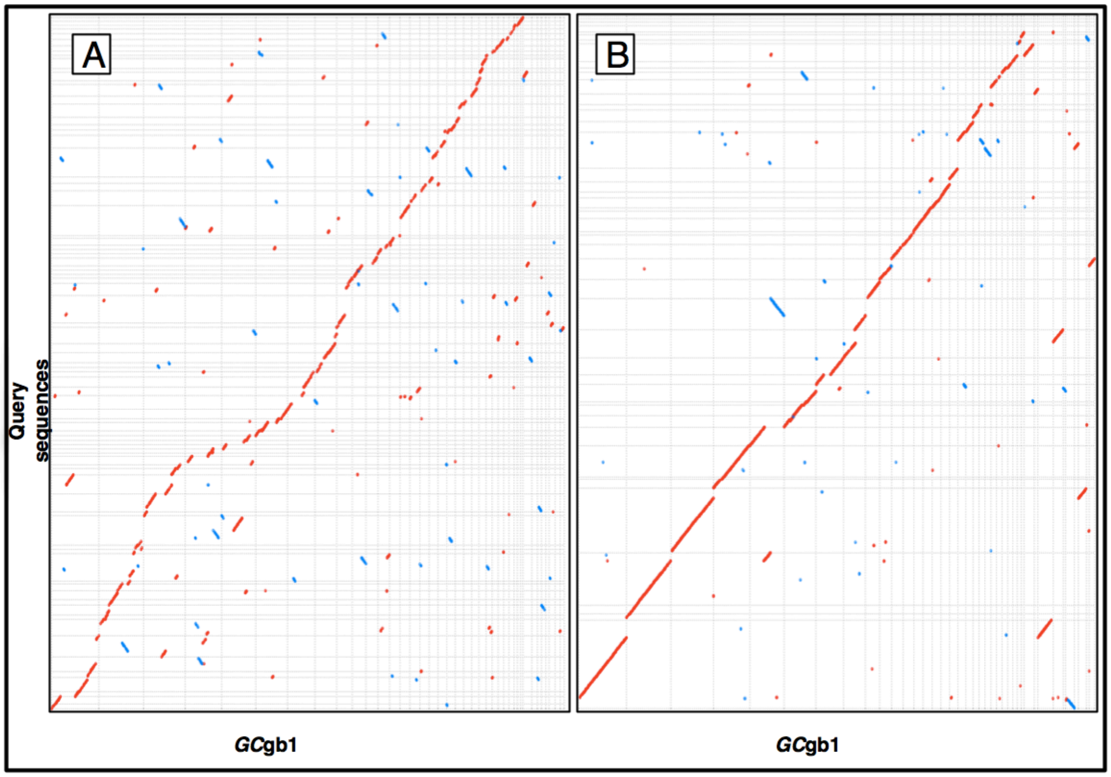
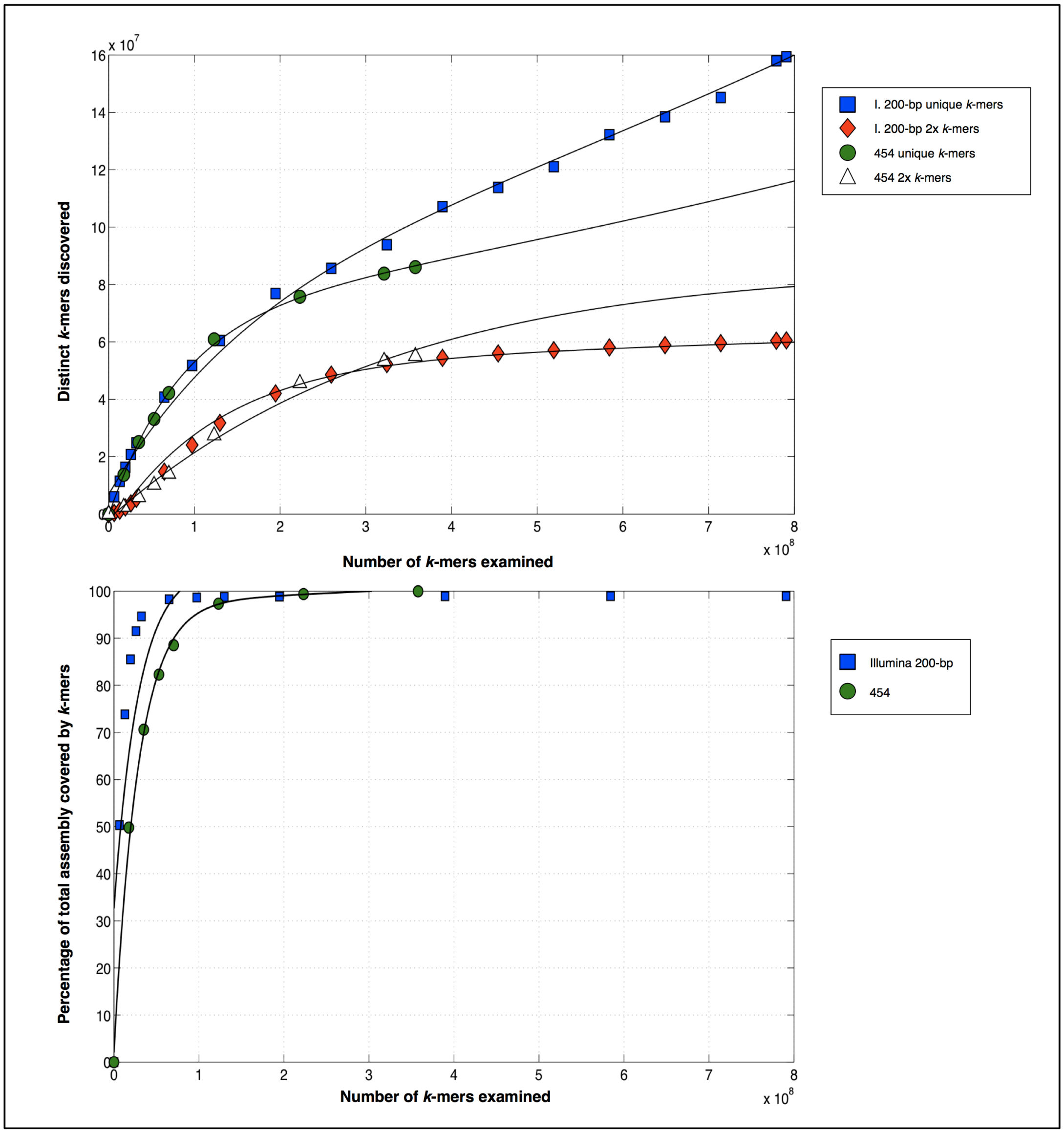


Figure 3.4 A. Raw reads were processed into overlapping 28 bp *k*-mers, and any *k*-mer that varied from all other *k*-mers by at least one bp was accepted as new sequence information. The analysis was done separately for unique *k*-mers and those that occurred at least twice (2x *k*-mers). B. MAQ was then used to map these *k*-mers to the reference genome sequence and the rate at which new coverage was generated was plotted against the number of *k*-mers examined.



3.5 References

Brockman, W., Alvarez, P., Young, S., Garber, M., Giannoukos, G., Lee, W.L., Russ, C., Lander, E.S., Nusbaum, C., Jaffe, D.B. (2008): Quality scores and SNP detection in sequencing-by-synthesis systems. *Genome Res*,*18,* 763-770.

Butler, J., Maccallum, I., Kleber, M., Shlyakhter, I., Belmonte, M., et al. (2008). ALLPATHS: De novo assembly of whole-genome shotgun microreads. *Genome Research, 18*, 810-820.

Dean, R.A., Talbot, N.J., Ebbole, D., Farman, M.L., Mitchell, T.K., et al. (2005). The genome sequence of the rice blast fungus Magnaporthe grisea. *Nature, 434,* 980-986.

DiGuistini, S., Ralph, S.G., Lim, Y.W., Holt, R., Jones, S., et al. (2007). Generation and annotation of lodgepole pine and oleoresin-induced expressed sequences from the blue-stain fungus Ophiostoma clavigerum, a Mountain Pine Beetle-associated pathogen. *FEMS Microbiology Letters, 267*, 151-158.

DUST [ftp://ftp.ncbi.nlm.nih.gov/pub/tatusov/dust/]

Fejes, A., Robertson, G., Bilenky, M., Varhol, R., Bainbridge, M., et al. (2008). FindPeaks 3.1: a tool for identifying areas of enrichment from massively parallel short-read sequencing technology. *Bioinformatics, 24*, 1729-1730.

Forge [http://sourceforge.net/projects/forge/]

Galagan, J.E., Calvo, S.E., Borkovich, K.A., Selker, E.U., Read, N.D., et al. (2003). The genome sequence of the filamentous fungus Neurospora crassa. *Nature, 422,* 859-868.

Gordon, D., Abajian, C., & Green, P. (1998). Consed: a graphical tool for sequence finishing. *Genome Research, 8,* 195-202.

Huse, S., Huber, J., Morrison, H., Sogin, M., & Welsh, M.D. (2007). Accuracy and quality of massively- parallel DNA pyrosequencing. *Genome Biology, 8,* R143.

Krzywinski, M., Schein, J., Birol, I., Connors, J., Gascoyne, R., et al. (2009). Circos: An information aesthetic for comparative genomics. *Genome Research, 19,* 1639-1645.

Lee, S., Kim, J., & Breuil, C. (2006). Pathogenicity of Leptographium longiclavatum associated with Dendroctonus ponderosae to Pinus contorta. *Canadian Journal of Forest Research, 36,* 2864-2872.

Li, H., Ruan, J., & Durbin, R. (2008). Mapping short DNA sequencing reads and calling variants using mapping quality scores. *Genome Research, 118,* 1851-1858.

Margulies, M., Egholm, M., Altman, W.E., Attiya, S., Bader, J.S., et al. (2005). Genome sequencing in microfabricated high-density picolitre reactors. *Nature, 437,* 376-378.

Möller, E.M., Bahnweg, G., Sandermann, H., Geiger, H.H. (1992). A simple and efficient protocol for isolation of high molecular weight DNA from filamentous fungi, fruit bodies, and infected plant tissues. *Nucleic Acids Research, 20,* 6115-6116.

NCBI [http://www.ncbi.nlm.nih.gov]

Parra, G., Bradnam, K., & Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics, 23,* 1061-1067.

Pipeline scripts [ftp://ftp.bcgsc.ca/supplementary/Grosmannia_clavigera/tools/]

RepeatMasker [http://www.repeatmasker.org/]

Stanke, M., Schöffmann, O., Morgenstern, B., & Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *BMC Bioinformatics, 7,* 62.

The Tria Project [http://www.thetriaproject.ca/index.php]

Simpson, J., Wong, K., Jackman, S., Schein, J., Jones, S.J.M., & Birol, I. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Research, 19,* 1117-1123.

Warren, R., Sutton, G., Jones, S., & Holt, R. (2007). Assembling millions of short DNA sequences using SSAKE. *Bioinformatics, 23,* 500-501.

Zerbino, D. & Birney, E. (2008). Velvet: Algorithms for *de novo* short read assembly using De Bruijn graphs. *Genome Research, 18,* 821-829.

4 Genome and transcriptome analysis of the mountain pine beetle-fungal symbiont *Grosmannia clavigera*, a lodgepole pine pathogen[3,4]

4.1 Introduction

Bark beetles and their fungal associates have inhabited conifer hosts since the Mesozoic era (Seybold, Bohlmann, & Raffa, 2000), and are the most economically and ecologically significant insect pests of forests in the northern hemisphere. The current mountain pine beetle (MPB, *Dendroctonus ponderosae*) outbreak in Western North America is the largest since the early 1900s, and has killed an estimated 630 million cubic meters (~16.3 M hectares) of lodgepole pine (*Pinus contorta* subsp. *latifolia* Engelm.) forest in British Columbia (www.for.gov.bc.ca/hfp/mountain_pine_beetle/). Climate change is thought to be a contributing factor to the current MPB epidemic, and the devastation of large areas of pine forests is predicted to have major consequences that include disturbing the global balance of atmospheric carbon emission and sequestration (Kurz et al., 2008).

Among the MPB-associated microflora (Lee, Kim, & Breuil, 2006a), the ascomycete *Grosmannia clavigera* (*Gc*) is a critical component of this large-scale epidemic (Figure 4.1). This pathogenic fungus can kill lodgepole pine when inoculated without the beetle (Lee, Kim, & Breuil, 2006b). The association between bark beetles and vectored fungi is

[3] A version of this chapter has been submitted as DiGuistini, et al. (2010) Genome and transcriptome analysis of the mountain pine beetle-fungal symbiont *Grosmannia clavigera*, a lodgepole pine pathogen. PNAS

[4] Data reported in this chapter will be available at NCBI and accessible through the *Grosmannia clavigera* genome page (http://www.ncbi.nlm.nih.gov/, Project ID 39837) upon submission of the manuscript to a peer reviewed journal.

symbiotic. The fungi benefit from beetles which carry them through the tree bark to a nutrient-rich wood environment of a new host. The benefits to the beetle and its progeny are less clear, but may include making nutrients available, detoxifying host defence metabolites, and weakening tree defences (Ayres, Wilkens, Ruel, & Lombardero, 2000; Bleiker, & Six, 2007; Lieutier, Yart, & Salle, 2009). While fungi and bark beetles must overcome physical and chemical defences to become established in conifer hosts, their relative contributions to this process are not well defined.

Oleoresin terpenoids and phenolics are key chemical defence components in conifers (Franceschi, et al., 2005; Keeling, & Bohlmann, 2006). In lodgepole pine, phenolics are stored in specialised polyphenolic parenchyma (PP) cells of the bark and phloem. Monoterpenoids and diterpene resin acids of the toxic oleoresins are formed and accumulate in resin ducts of the phloem and sapwood.

To understand the interactions underlying colonization of lodgepole pine by the MPB and its fungal associates, and to identify possible biochemical mechanisms by which *Gc* overcomes conifer defences, we reported a draft eukaryotic genome sequence assembled primarily from next generation sequencing (NGS) data (DiGuistini et al., 2009). Here, we report the finished 29.8 Mb *Gc* genome sequence, its protein coding gene annotations, and results from applying these resources to begin to clarify the interaction of *Gc* with host tree defence chemicals. We describe a set of 8,305 annotated protein coding sequences; preliminary annotation of protein coding sequence polymorphisms; proteins secreted in response to growth on wood; changes in the fungal transcriptome induced by exposure to lodgepole pine phloem extract (LPPE) or oleoresin terpenoids; and genes and pathways involved in the modification, transport and metabolism of these conifer defence components.

4.2 Materials and methods

4.2.1 Strains

Gc strain kw1407 (NCBI Taxonomy ID: 655863) is deposited into the University of Alberta Mycological Herbarium (UAMH) 11150 along with the additional isolates used in this study (11151-11156).

4.2.2 Genome finishing, ESTs, genome annotations

Finishing was performed on the draft assembly described previously (DiGuistini et al., 2009; appendix C.1) with additional data: one lane of Illumina Genome Analyzer (GA_{ii}) from a 3-kb long insert library (Bentley et al., 2008) and 1,299 finishing reactions performed for filling gaps. Telomeric repeats were identified using the sequence TTAGGG. Expressed Sequence Tags (ESTs) were reported earlier (DiGuistini et al., 2007; Hesse-Orce et al., submitted). Gene models are a composite of *ab initio* and homology based predictions generated using GLEAN (Elsik et al., 2007; appendix C.2). Putative gene function assignments were generated from searches of the NCBI NR and Swissprot databases using BLAST and combined with PFAM domain assignments. GO annotations were assigned using Blast2GO (Conesa et al., 2005). Predicted protein localizations were determined using SignalP, TMHMM and WolfPsort.

4.2.3 Peptide sequencing

To obtain extracellular proteins for *Gc,* the fungus was grown on sawdust-agar plates overlayed with cellophane. After three days of growth, mycelia and cellophane were transferred to acetate buffer, centrifuged and filtered (appendix C.5). The protein solution was concentrated and separated by 1D SDS-PAGE (appendix C.5). In-gel protein digests were performed for 16 bands cut from the 1D gel (appendix C.5). Peptide analysis was performed by tandem mass spectrometry (appendix C.5). Bioinformatic analysis was performed with custom scripts (appendix C.5, available upon request).

4.2.4 RNA-seq and variant detection

RNA-seq data was generated with an Illumina (GA_{ii}) from poly(A^{+}) mRNA (appendix C. 6). Sequence clusters were generated on an Illumina cluster station. Lanes were sequenced to 36 cycles. Post run analysis was performed with the Illumina GA pipeline (v.1.0). Paired-end (PE) reads were aligned to the reference genome sequence using CLC Genomics workbench (http://www.clcbio.com/; CLCbio, DK); SNP prediction was also performed within this software package with additional post-prediction filtering (appendix C.6).

4.2.5 Terpene and LPPE treatment expression analysis

Culture conditions for mycelia generated for expression analysis are described in the appendix C.7. Terpene and LPPE treatment preparations are described in the appendix C.7. Treatments for transcriptome analysis were carried out using a TLC sprayer applying the treatment directly to culture surfaces with filtered nitrogen gas as the carrier.

4.3 Results

4.3.1 Genome sequence and protein coding annotations in *G. clavigera*

Manually finishing the genome of *Gc* (kw1407; NCBI, Genome PID: 39837) yielded 18 supercontigs (SC) with a total length of 29.8 Mb (Table 4.1; appendix C.1). Telomeric sequences suggested that the SCs belonged to 7 chromosomes. We achieved 64x sequence coverage across 90 % of the finished genome sequence (Figures C1 and C2). We validated the assembly by aligning to it 99.4 % of 7,169 unique Expressed Sequence Tag (EST) sequences (method described in DiGuistini et al., 2009). We assembled the mitochondrial genome into a single ~90 kb circularized sequence; alignment to related fungal mitochondrial sequences validated the accuracy of this assembly (Figure C3).

Before predicting gene models, we masked the assembled genome sequence for repetitive elements identified using similarity to repeat databases (repbase v.20090120) and *de novo* repeat detection using RepeatScout (Price et al., 2005). In total, 10.4 % of the finished genome was found to be composed of repeats or low complexity sequences. Evidence for repeat-induced point mutation (RIP) was identified using RIPCAL (Hane & Oliver, 2008), and was found almost exclusively within transposable elements. After excluding mitochondrial DNA, we predicted 8,305 protein coding gene models from the assembled genome sequence accounting for 46 % of the total genome length. We found introns in 77.2 % of the gene predictions with an average of 1.86 introns per gene. Locations of introns within gene models were distributed uniformly with a small bias towards both the 5' and 3' ends. The predicted gene models were supported and validated with EST, RNA-seq, and peptide sequences. We annotated the translated set of sequences using available public sequence databases and assigned functional descriptions for approximately 85 % of the total predicted protein collection.

4.3.2 Identification of protein coding sequence variations

To assess single nucleotide polymorphisms (SNPs) in the protein coding regions of the *Gc* genome and to provide additional gene model support we annotated the genome using RNA-seq read data from a collection of seven additional *Gc* strains (Table 4.2; 42 different culture-treatment combinations: appendix C.6). For this purpose we generated cDNA from polyA$^+$ purified total RNA and sequenced it using a paired-end read approach on the Illumina Genome Analyzer platform. We predicted 17,236 SNPs from the tag-to-genome alignments (FPR = 0.0045, FNR = 0.16) of which 12,160 occurred within ~14.5 Mb of protein coding gene model sequence covering 92 % of the total transcriptome length (~15.77 Mb). Only a small number of variants were located in predicted intron regions (741; 6 %). These 741 SNPs could have resulted from incompletely spliced transcripts, alternatively spliced transcripts or inappropriately predicted introns. We found a SNP density of 1 variant per 1,189 bp across the predicted genes and an average minor allele frequency (MAF) of 25.1 %. Transitions were favored over transversions by a ratio of 3:1 and amino acid sequence variations for 5,689 of the predicted SNPs.

4.3.3 Detection of *G. clavigera* gene orthologues

We next used orthoMCL (Li, Stoeckert, & Roos, 2003) for identifying *Gc* gene orthologues. We clustered ~186.4 K predicted protein sequences from 17 fungal taxa and identified 6,780 ortho-groups (groups of putative orthologous genes) that contained at least one member from *Gc*. Of these, 1,940 contained a representative from all taxa and 692 possessed a strict single copy orthologous relationship (i.e. clusters contained exactly one member per species). Phylogenetic analysis was performed with these 692 genes to confirm the phylogenetic position of *Gc* within the class Sordariomycetes (Figure 4.1C; appendix C.3). We identified the mating-type (MAT) gene, suggesting that the sequenced strain belongs to the MAT-1-2 idiomorph. The high-mobility group (HMG) domain of the *Gc* MAT protein was similar to those in MAT loci of other filamentous ascomycetes. We detected the MAT-1-1 idiomorph alpha-domain in other *Gc* isolates, but not in the sequenced strain.

Using CAFE (De Bie et al., 2006) we identified *Gc* gene family expansions for methyltransferases (Mtfse), Major Facilitator Superfamily (MFS) transporters, and serine-peptidases, whereas gene family contractions occurred for Na^+/Ca^{2+} transporting ATPases, glycoside hydrolases (GHs), zinc-type alcohol dehydrogenases and cytochrome P450s (CYP450s). The largest *Gc* specific gene family expansion was for O-methyltransferases (Mtfse), for which we identified 199 methyltransferase-like sequences (PFAM: PF08241-2). Using a phylogenetic analysis including a subset from the other fungal taxa we observed a clade containing seven *Gc* Mtfse sequences that showed significant support for branch-specific differences in synonymous vs. non-synonymous substitution rates using a likelihood ratio test ($p < 0.001$) indicating that these methyltransferases may be under positive selection (appendix C.4).

4.3.4 Identification and annotation of genes and proteins for inhabiting host pine

To identify genes and mechanisms used by *Gc* to grow in the host sapwood, we isolated proteins secreted by *Gc* during mycelial growth on pine sawdust-supplemented agar medium. Peptide sequencing supported 214 of the *Gc* gene models described above, for which we identified enriched Gene Ontology (GO) terms. Ninety percent of these annotated genes (162 genes) belonged to metabolic processes, with the greatest enrichment occurring within ‘carbohydrate metabolism’ (GO:0005975) and ‘proteolysis’ (GO:0006508). The deduced protein sequences were enriched in secretion signal peptides, which were predicted for 135 (63 %) of the 214 genes but for only 939 (11 %) of the entire gene model collection.

We identified 231 carbohydrate-active enzymes in the *Gc* genome, using the CAZy classification system (www.cazy.org). This number is smaller than previously reported for *Neurospora crassa* (277) or *Magnaporthe grisea* (378). The *Gc* genome contained 139 GHs, 17 of which we detected by peptide sequencing (described above; Table 4.3). These GHs included enzymes likely involved in maintaining cell wall plasticity during growth and morphogenesis and in acquiring carbohydrates. However, GHs involved in degrading host ligno-cellulose structures were notably absent from both the proteome and genome collections (e.g. GH6 cellulase). *Gc* has only two carbohydrate binding modules (CBMs) assigned to family CBM1, one was attached to a GH12 plant cell wall digesting enzyme and the other was attached to a chitinase. Carbohydrate esterases (CEs) were also sparse, in particular families CE5 and CE1. Whereas *M. grisea* and *N. crassa* respectively have 10 and 7 CE1s and 15 and 3 CE5's, *Gc* has only one of each. In addition, we noted that *Gc's* CE1 family contained only a cinnamoyl esterase, and no type A feruloyl esterase. Peptide sequencing and signal peptide analysis indicated that the CE5 and CE8 enzymes were secreted during growth on the sawdust-agar plates (Table 4.3).

We used the MEROPS database (merops.sanger.ac.uk) and identify 287 putative peptidases in *Gc*. Twenty of these peptidases, belonging to the A1, S8, S28 and S53 families, were also identified in the peptide sequencing data. The top five ranked by peptide-spectra abundance are reported in Table 4.3. We identified a lineage-specific gene expansions within the peptidase families S53 (10 genes). S53 genes were among the most abundant peptidases secreted during growth on the sawdust-agar plates.

4.3.5 Identification and annotation of genes for detoxifying host defence metabolites

In the phloem and sapwood of a pine host, *Gc* grows in an environment with high concentrations of terpenoid and phenolic defence metabolites. Growth of *Gc* on malt extract agar (MEA) was reduced in the presence of lodgepole pine phloem extract (LPPE). In the presence of terpenoids *Gc* grew with an initial lag phase (24 hr) followed by growth at nearly the same rate as untreated controls (Figure 4.2A and 4.2B). In contrast, *N. crassa* growth was substantially reduced when challenged with the LPPE treatment and completely inhibited by the terpene treatment (data not shown), highlighting *Gc's* tolerance for these conifer defence compounds and supporting the hypothesis that overcoming terpenoid defences could be an important pathogenicity factor.

To identify possible mechanisms used by *Gc* to overcome host chemical defences we used Illumina expression profiling (RNA-seq) on mycelia samples collected at 12 and 36 hr after LPPE or terpene treatments. In total, 4,690 gene models showed a > 2-fold increase in transcript abundance in at least one of the four treatment-time comparisons (Figure 4.3). We plotted expression levels for genes induced by the LPPE and terpene treatments in 50 kb windows and noted regions with high transcriptional activity (co-expression clusters (ECs); see below and Figure C5).

4.3.6 Response of *G. clavigera* to lodgepole pine phloem extract (LPPE) treatment

GOMiner (Zeeberg et al., 2003) analysis of the 12 hr LPPE gene expression data identified several enriched biological processes including carbohydrate metabolism ($p < 0.001$), alcohol metabolism ($p < 0.001$), glycolysis ($p < 0.001$), external encapsulating structure organization ($p < 0.001$), cellular protein metabolic processes ($p < 0.001$), and cellular aromatic compound metabolic processes ($p < 0.001$).

We used the CAZy classifications for inspecting GH gene expression 12 hr following LPPE treatment. This analysis indicated upregulation of *Gc* GHs targeting the plant cell wall (families: GH51, GH78, GH61, GH53, GH43) and an α, α-trehalase (GH37). In addition, genes encoding proteins from families GH3, GH5 and GH39 were also induced following the LPPE treatment. Although the substrate specificity of these GHs is unknown the GH3 and GH39 proteins are likely intracellular as they do not possess extracellular signal peptide sequences; whereas, the GH5 has a secretion signal and no GPI anchor.

Overall, gene expression data indicated that following LPPE treatment a strong oxidative stress was generated. We observed substantial induction of both Mn/Fe$^-$ and Cu/Zn$^-$ superoxide dismutases, peroxidases and a thioredoxin and thioredoxin reductase. Upregulation of the eight subunits of the T-complex polypeptide, involved in actin and tubulin folding, as well as the induction of the actin and tubulin genes themselves suggests the importance of cytoskeleton reorganization. Strong oxidative stresses would also result in damage to existing proteins and this is corroborated by the upregulation of 19 genes encoding proteasome and proteasome regulatory subunits.

Overall, gene expression data at 36 hr indicated that host chemical detoxification may be a critical process for growth on the LPPE extract. Many of the genes initially upregulated by the 12 hr LPPE treatment were no longer induced. GO analysis provided no evidence for significantly enriched biological processes; we observed that this was not due to a lack of upregulated genes at this time point but rather to the lack of GO annotations in the genes that were induced. Many of the highly expressed genes belonged to gene families with known roles in detoxification. In particular, we observed the induction of numerous transcription factors, oxidoreductases, and CYP450s (Table C1). Genes known to be involved in aromatic degradation such as the *N. crassa* 3-carboxy-cis,cis-muconate cyclase (G_7783) and genes involved in the catabolism of phenylacetate and its hydroxy derivatives in *Aspergillus nidulans* were induced at low levels in *Gc* (Kajander et al., 2002; Fernández-Cañón & Peñalva, 1995). As in *A. nidulans,* we noted that the genes of the phenylacetate catabolic pathway were clustered although the cluster is expanded in *Gc* and includes the phenylacetate 2-hydroxylase (GCSC_179: 1.126-1.135 Mb).

We investigated the genome regions surrounding the putative detoxification genes within the ECs (Figure C5). Two ECs were previously identified by digital profiling ESTs following LPPE treatment (Hesse-Orce et al., submitted), and our current expression data validates these results. The additional gene annotation and expression data reported here allow us to extend one of these clusters consisting of six loci by four loci to 10 gene models (GCSC_140; 1.13-1.15 Mb; Cluster I, Table 4.4, Figure C6). The region with the highest average expression levels over a 50-kb window in the LPPE treated data (GCSC_173; 1.84-1.90 Mb; Cluster II, Table 4.4) contained 12 genes, all of which responded strongly to the LPPE treatment.

4.3.7 Response of *G. clavigera* to terpene treatment

In the 12 hr mycelial cultures treated with terpenes, RNA degradation was consistently observed (Figure 4.2A inset) while LPPE, oxidative, osmotic, temperature and nitrogen starvation treatments, did not induce this degradation process. We observed two overlapping GO clusters within the biological process hierarchy for the 12 hr post-terpene treated sample. In apparent agreement with a putative role for RNA biogenesis and turnover, the first cluster included genes annotated to 'mRNA processing' ($p < 0.001$) and 'ribosome biogenesis' ($p < 0.001$), while the second included genes annotated to 'amino acid biosynthesis' ($p < 0.001$). We observed the induction of genes encoding DNA repair, recombination, stability and replication proteins such as helix destabilizing proteins, topoisomerases, ss-DNA binding protein, DNA repair nucleases, mismatch repair proteins, DNA ligases, a DNA glycosylase and DNA polymerases. In addition, we observed the induction of genes encoding histones H2A, H2B and H4, while alternate variants for H2A and H4 and histones H3 and H1 were strongly repressed. Strong induction of an H4 arginine methyltransferase, H3-K79 methyltransferase, ubiquitin conjugase and SIR2-like deacetylase implicates chromatin remodeling in the process of changing gene expression.

As *Gc* emerged from the lag phase of growth (36 hr), there was a marked change in its gene expression pattern compared to 12 hr following treatment. We observed among the differentially expressed genes two overlapping clusters of GO-terms within the biological process hierarchy. The clusters encompass ‘lipid metabolic processes’ ($p < 0.001$) and ‘alcohol metabolism’ ($p < 0.001$). The molecular function hierarchy contained several small clusters falling primarily within ‘catalytic activity’ ($p < 0.001$) and ‘microtubule based processes’ ($p < 0.001$). Within these classifications noteworthy members were ‘oxidoreductase activity’ ($p < 0.001$), ‘aldehyde dehydrogenase activity’ ($p < 0.001$), and ‘electron carrier activity’ ($p < 0.001$). Encompassed within the microtubule-based processes are ‘cytoskeleton organization and biogenesis’ ($p < 0.001$), ‘cytoskeleton based intracellular transport’ ($p < 0.001$), ‘cellular localization’ ($p < 0.001$) and ‘transport’ ($p < 0.001$). KEGG annotations supported the GO analysis, indicating induction of the fatty acid and glyoxylate pathways. We examined the β-oxidation capacity of *Gc* and found that the *FOX2* multifunctional β-oxidation enzyme (G_6203) was induced; however, the mitochondrial short-chain enoyl CoA hydratase (G_647) was more strongly induced. In addition, we observed strong induction at both 12 and 36 hr for carnitine acyl transferase and for carnitine acetyl transferase, indicating that β-oxidation in the mitochondria may be favored over the peroxisome. We were not able to readily identify a peroxisomal acyl-CoA oxidase and we observed no increase in expression for peroxisomal catalases indicating that this fungus likely uses a non-forming H_2O_2 pathway for peroxisomal β-oxidation (Thieringer & Kunau, 1991).

We identified a 100-kb EC on supercontig GCSC_108 (0.9 - 1.1 Mb). Within an ~85 kb core section of this genome region (Table 4.5) 35 gene models were predicted, 18 were induced in response to the terpene treatment, 4 were repressed and 12 were unchanged (Figure 4.4). The most strongly induced genes in this region were a flavoprotein monooxygenase, an FMO-like monooxygenase containing a lipocalin signature and a short-chain dehdrogenase/reductase enzyme. In addition to these oxidoreductases were enzymes such as an epoxide hydrolase, alcohol dehydrogenase and aldehyde dehydrogenase, which may be important for activating terpenoids or their intermediates for β-oxidation.

4.4 Discussion

In the work reported here we have created genomic and molecular resources for a bark beetle-symbiotic fungus and tree pathogen and applied these resources for beginning to characterize this biological system. We completed the sequencing and assembly of the *Gc* genome using both traditional finishing strategies and NGS data. This approach resulted in a high quality, comparative grade (Blakesley et al., 2004) genome sequence. The genome size, repeat content and gene collection are similar to other saprophytic and pathogenic fungi in the class Sordariomycetes. We established that *Gc* is heterothallic and generated data for future studies on the relationship between sexual reproduction and epidemic dynamics. We also identified evidence for the fungus-specific genome defence mechanism RIP, primarily in *Gc* transposon sequences. In *N. crassa* RIP occurs before or during meiosis, causing C•G to T•A mutations within duplicated sequences (Galagan & Selker, 2004). This result indicated that *Gc* has sexual potential despite the *Gc* sexual cycle being rarely reported in field data and not yet achieved under laboratory conditions.

Like other sapstaining fungi that colonize conifers, *Gc* is unable to degrade the structural components of wood, lignin and cellulose (Zabel, & Morell, 1992). Instead, such fungi utilize non-structural components, including but not limited to proteins, amino acids, starch, soluble sugars, triglycerides and fatty acids (Abraham et al., 1998; Gao & Breuil, 1995). In support of this we found that *Gc* secretes a number of GHs associated with the utilization of soluble sugars and starch and *Gc* possesses relatively more plant cell wall degrading enzymes that could be involved in pectin degradation of the cell wall or the tracheid bordered pit membranes, which would allow the fungus to colonize the sapwood (Lieutier & Berryman, 1988). In addition, we identified a lipase that is likely involved in using pine triglycerides, a major carbon source for this fungus. Triglycerides account for ~ 2-2.5 % oven-dry weight (odw) of lodgepole pine stems (Gao, Chen & Breuil, 1995). As well, we showed that *Gc* secretes a number of peptidases. Some of these are necessary for the fungus to retrieve nitrogen from pine stem proteins, which occur at levels varying between 0.3 % odw for phloem to 0.05 % odw for sapwood (Abraham & Breuil, 1996; Bleiker & Six, 2007). The peptidase gene family expansion in *Gc*, highlights the importance of organic nitrogen acquisition in the pathogen life cycle. A similar expansion has been reported for the human pathogenic fungus *Coccidiodes* (Sharpton et al., 2009). In contrast the *Gc* genome is deficient in a number of GH, CE and CBM families. For example, GH6 commonly used by saprophytes for degrading cellulose, is absent from the *Gc* genome. Similarly, *Gc* appears to have no type A feruloyl esterase and only a single type B feruloyl esterase which does not possess a signal peptide and is therefore likely not secreted. Without a secreted feruloyl esterase *Gc* cannot hydrolyze the diferulate cross-links in plant cell walls.

Detoxification of host defence metabolites such as terpenoids and phenolics begins with oxidative, reductive or conjugating transformations that modify the chemicals into less toxic and more easily excreted metabolites (Duffy, Schouten & Raaijmakers, 2003). To achieve these transformations, it is unknown whether *Gc* relies on a limited number of modifying genes with broad substrate specificities, or a large number of genes with narrow substrate specificities. The expansion of the O-methyltransferase gene family within the *Gc* genome described here may represent evolutionary specialization to address detoxification. O-methyltransferases have roles in plant phenolic biosynthesis (Preisig, Matthews & VanEtten, 1989; Liu & Dixon, 2001) and in fungal and animal phenolic detoxification (Jeffers, McRoberts & Harper, 1997; Männistö & Kaakkola, 1999; Feltrer et al., 2010). GHs are also factors for fungi to detoxify sugar-conjugated antimicrobial chemicals such as phenolics and saponins (Bouarab et al., 2002; Pareja-Jaime, Roncero & Ruiz-Roldán, 2008; Zheng & Shetty, 2000). Using CAZy classifications we assessed GH gene expression in the LPPE treatment and found several GHs that were induced. From these, we have identified a small number of candidate GHs that may be involved in hydrolyzing aryl-glycosides. In contrast, we noted that *Gc* had a relatively low number of CYP450s (54) and no identifiable CYP450 gene family expansions. SNP analysis of the pooled RNA-seq data combined with results of the gene expression studies yielded a number of CYP450s inducible by host defence chemicals that warrant further investigation. These CYP450s are now being explored within a larger collection of *Gc* strains in ongoing work.

Following a MPB attack or *Gc* inoculation the concentration of host defence chemicals increases (Raffa & Berryman, 1983). However, the relative contribution of terpenes and phenolics in host defence is unclear. The LPPE treatment has a higher chemical complexity than the terpene blend used in the gene expression experiments (Figure C4) and captures more than just the defensive chemicals encountered by the fungal propagules when initially deposited into the tree phloem; however, the total number of changed genes induced was substantially greater following terpene treatment. Comparing *Gc's* LPPE and terpene responses, 41 genes were induced by both treatments at 12 hr. This collection was enriched for general and chemical stress responders, including heat shock and chaperonin proteins, MFS transporters, and enzymes involved in oxidative biotransformation. This collection also included a putative DNA glycosylase and cytidine deaminase, suggesting that changes in DNA methylation or RNA/DNA editing may be important in early chemical stress responses. In RNA extracted from cells treated for 36 hr, 58 genes were induced by both treatments. The induced genes included ABC and MFS transporters, glycoside hydrolases and xenobiotic activating enzymes such as CYP450s, hydrolases and general oxidases.

The LPPE extract is a complex mixture of methanol-water soluble compounds (Figure C4), which, from our initial extract analysis, contains free sugars, phenolics and glycosylated phenolics (data not shown). When *Gc* was treated with LPPE the mycelia became pink, suggesting that oxidized phenolic derivatives such as quinones and free radicals were generated. Consistent with this, genes involved in sugar utilization and response to oxidative stresses were strongly induced. The activation of genes and gene clusters that may be involved in the detoxification or degradation of host antimicrobial compounds suggests that *Gc* needs to detoxify its environment. The induction of transcription factors may reflect the regulatory coordination required to process the diverse collection of antimicrobial compounds in this phloem extract. Catabolic genes that degrade aromatics can occur as gene clusters in fungi (Fernández-Cañón & Peñalva, 1995), and, in *Gc*, LPPE treatment induced gene clusters. We anticipated finding genes involved in β-ketoadipate metabolism in these clusters because this pathway is commonly used for the aerobic detoxification of phenolics in microorganisms (Kajander et al., 2002). However, neither the *Gc* orthologue to *N. crassa* 3-carboxy-cis,cis-muconate cyclase or the *A. nidulans* phenylacetate catabolic gene cluster, nor genes identified in the TCA cycle were strongly induced in response to the LPPE treatment. This suggests that *Gc* may use unique catabolic pathways to detoxify host-specific pine defence chemicals.

When *Gc* was treated with terpenes we observed a lag phase in growth, substantial amounts of mRNA degradation, and large fluxes in cellular protein content as indicated by GO term enrichment in amino acid biosynthesis. At 12 hr, in contrast to the LPPE treatment, we also noted that histone and histone-related proteins were induced. These changes are consistent with terpene treatment inducing transcriptome reprogramming that is mediated by chromatin remodeling. At 36 hr following the terpene treatment, genes involved in fatty acid metabolism were induced. This may indicate that terpenoid detoxification occurs, in part, through the β-oxidation pathway; consistent with this, we confirmed that *Gc* is able to utilize terpenoids as a sole carbon source (Figure 4.5). Genes involved in both peroxisomal and mitochondrial β-oxidation pathways were induced. It is likely that both pathways contribute towards terpenoid degradation and that shuttling occurs between the organelles.

When averaging co-expression over 50 kb windows at 36 hr, we observed a gene cluster that spanned approximately 100 kb and responded strongly to the terpene treatment. Annotation of the induced genes in this region suggested that this cluster might contribute to early enzymatic steps in terpene catabolism. Co-expression clusters that span large genomic regions have been reported for higher eukaryotes (Hurst, Pál, & Lercher, 2004), but not for fungi. Such clusters may originate either through selection for increased coordination of gene expression, or for the advantage of the clustered state itself (Hurst, Pál, & Lercher, 2004; Walton, 2000). For *Gc,* toxic metabolite intermediates would likely accumulate as a result of a suboptimal detoxification pathway, and precise gene regulation and reliable gene transmission are likely important in the origin and maintenance of this genome region. In support of this, in *Burkholderia xenovorans* nearly half of the 93 genes induced by the diterpene dehydroabietic acid (DHA) occur in an ~80 kb region, and the genes in this region participate directly in DHA catabolism (Smith, Park, Tiedje & Mohn, 2007). Finally, we have begun exploring the contribution of the most strongly induced pleiotropic drug resistance (PDR) transporter, GLEAN_8030, to *Gc's* terpenoid tolerance. Deleting this gene using our recently developed split-marker *Agrobacterium*-mediated transformation system (Wang et al., 2010) prevented mycelial growth of the fungus on the terpene-supplemented media (Figure C7). Terpenoids are antimicrobials in their natural plant and tree defence context. By applying the genomic and molecular resources developed in this work we have begun to clarify the specialized mechanisms that *Gc* has adapted that allow it to tolerate terpenoids and grow in its pine host --- an evolutionary adaptation that is an important factor in the interaction between host tree, the fungal pathogen, and its beetle vector.

Tables and figures

Table 4.1 Genome characteristics for *G. clavigera.*

Genome characteristics for *G. clavigera*	
Genome size	29.8 Mb
Chromosome number	7
SuperContig number	18
Contig N50	1.2 Mb
GC% genome	53.4%
GC% gene	59.0%
GC% transcript	60.5%
GC% intron	50.3%
GC% intergenic	48.7%
Protein coding gene number	8303
Gene density	1/3,517 bp
Mean gene length	1,903 bp
Mean transcript length	1,641 bp
Mean intergenic distance	1,466 bp
Percent multi-exonic genes	77.2%
Mean exon length	574.5 bp
Mean number of introns/gene	1.86
Median intron length	70 bp

Table 4.2 Origin of strains described in this study.

Species	Isolate	Location of origin	Host	Isolated from	Year
G. clavigera	kw1407	Kamloops	*Pinus contorta*	Gallery	2003
G. clavigera	ATCC18086	BC/Cache Creek	*Pinus ponderosae*	Sapwood	1965
G. clavigera	B5	Banff	*Pinus contorta*	MPB body	2003
G. clavigera	B10	Banff	*Pinus contorta*	MPB body	2003
G. clavigera	H55	Houston	*Pinus contorta*	MPB body	2003
G. clavigera	0200-01-14	Kamloops	*Pinus contorta*	Sexual spore	2004
G. clavigera	DPCHMC3	Cypress Hill	*Pinus contorta*	MPB mycangia	2007
G. clavigera	DPLKGT1B	Kelowna	*Pinus contorta*	MPB body	2007

Table 4.3 Genes identified by proteome sequencing following mycelial growth of *Gc* on pine sawdust-supplemented agar media. All members of this gene list also possess signal peptides as identified by signalP.

G_ID	Spectra abundance	DB description	Description
Carbohydrate active enzymes[a]			
GLEAN_8298	413	Glycoside hydrolase Family 55	β-1,3-glucanase
GLEAN_4590	191	Glycoside hydrolase Family 16 w. GPI anchor	Multiple activities
GLEAN_6913	161	Glycoside hydrolase Family 5 w. GPI anchor	Xylanase
GLEAN_6037	135	Glycoside hydrolase Family 28	Polygalacturonase
GLEAN_7572	115	Glycoside hydrolase Family 3	β-glucosidase
GLEAN_5244	80	Glycoside hydrolase Family 18 w. 2 CBM50 and 1 CBM18	Chitinase
GLEAN_4940	76	Glycoside hydrolase Family 35	β-galactosidase
GLEAN_6363	56	Glycoside hydrolase Family 17 w. GPI anchor	Fungal specific endo-1,3-β-glucosidase
GLEAN_3441	46	Glycoside hydrolase Family 3	β-glucosidase
GLEAN_8284	37	Glycoside hydrolase Family 18 w. CBM18	Chitinase
GLEAN_3913	28	Glycoside hydrolase Family 76 w. GPI anchor	α-1,6-mannanase
GLEAN_6144	23	Glycoside hydrolase Family 28	Polygalacturonase
GLEAN_2653	17	Glycoside hydrolase Family 11	Xylanase
GLEAN_3235	15	Glycoside hydrolase Family 12	Endoglucanase or xyloglucan hydrolase
GLEAN_6459	15	Glycoside hydrolase Family 78	α-L-rhamnosidase
GLEAN_4814	14	Glycoside hydrolase Family 51	α-L-arabinofuranosidase
GLEAN_8240	7	Carbohydrate esterase Family 5	Cutinase
GLEAN_1923	4	Carbohydrate esterase Family 8	Pectin esterase
GLEAN_679	4	Polysaccharide Lyase Family 1	Pectin lyase
Lipase			
GLEAN_4804	7	Triacylglycerol lipase	Lipase (class 3)
Peptidases[b]			
GLEAN_2700	341	Peptidase S41-like	Unassigned function
GLEAN_2281	87	Peptidase A1	Aspartic endopeptidase
GLEAN_229	77	Peptidase S28	Prolyl endopeptidase
GLEAN_4866	60	Peptidase S8	Serine protease
GLEAN_4258	45	Peptidase S8	Serine protease
Extracellular oxidative enzymes[c]			
GLEAN_6725	49	LDA3	Glyoxal oxidase
GLEAN_3074	21	LO2	Lignin oxidase
GLEAN_6781	4	LO1	Lignin oxidase

[a] See CAZy (www.cazy.org) for information regarding the descriptions provided.

[b] See MEROPS (merops.sanger.ac.uk) for information regarding the descriptions provided.

[c] See FOLy (foly.esil.univ-mrs.fr) for information regarding the descriptions provided.

Table 4.4 Lodgepole pine phloem extract (LPPE) induced gene expression clusters. Cluster I is located on supercontig GCSC_140 and Cluster II is located on supercontig GCSC_173.

Gene model	Approx. contig pos.	36 Hr LPPE test statistic	Putative function	InterPro annotation
Cluster I				
GLEAN_2439	1130422	0.88	Salicylaldehyde dehydrogenase	IPR015590
GLEAN_7592	1132087	6.41	Histidinol dehydrogenase	IPR001692
GLEAN_7591	1133873	1.2	Short-chain dehydrogenase/reductase	IPR002198
GLEAN_2440	1135126	3.7	FAD-linked oxidase	IPR006094
GLEAN_7590	1136717	1.75	Short-chain dehydrogenase/reductase	IPR002198
GLEAN_2441	1138138	9.51	FAD binding monooxygenase	IPR006076; IPR013096
GLEAN_2442	1141798	0.67	Zinc ion binding transcription factor	IPR001138; IPR007087
GLEAN_7589	1144843	5.1	Hypothetical protein	NA
GLEAN_2443	1145684	5.97	Aldehyde dehydrogenase	IPR016161
GLEAN_2444	1147504	5.32	Aromatic ring-opening dioxygenase	IPR004183
Cluster II				
GLEAN_1289	1881185	30.62	Short chain dehydrogenase/reductase	IPR002198
GLEAN_5641	1872722	17.46	Duf1446 domain protein	IPR010839
GLEAN_5638	1887927	16.62	Dimethylaniline monooxygenase	IPR020946
GLEAN_5640	1876832	16.01	Aldehyde dehydrogenase	IPR015590
GLEAN_1288	1875555	8.29	CoA-transferase family III	IPR003673
GLEAN_5642	1868904	8.23	Alcohol dehydrogenase	IPR013149; IPR013154

GLEAN_5639	1882536	5.56	Heavy metal translocating p-type ATPase	IPR005834;
				IPR006121;
				IPR008250

GLEAN_1287	1870426	5.15	Aldehyde dehydrogenase	IPR015590
GLEAN_1290	1893037	4.11	Methyltransferase	IPR013216
GLEAN_1286	1867499	2.79	Alpha-beta hydrolase	IPR013094
GLEAN_1291	1895364	1.63	MFS transporter	IPR011701
GLEAN_5643	1865026	1.40	MFS transporter	IPR011701

Table 4.5 A terpene induced gene expression cluster on supercontig GCSC_108.

Gene model	Approx. contig pos.	36 Hr Terpene test statistic	Putative function	InterPro annotation
GLEAN_6910	921864	3.70	Aconitase	IPR000573; IPR001030
GLEAN_2126	925332	5.03	Hexachloro-cyclohexane acid phosphatase	IPR000560
GLEAN_6909	927093	-6.84	Hypothetical protein w. transcriptional support	N/A
GLEAN_2128	930328	0.00	Collagen superfamily protein	IPR008160
GLEAN_6908	932463	0.00	F-box domain containing protein	IPR001810
GLEAN_6907	934148	-1.60	MFS transporter	IPR011701
GLEAN_2129	937255	-0.43	Arylsulfatase	IPR000917
GLEAN_6906	939501	1.65	Hypothetical protein w. transcriptional support	N/A
GLEAN_6905	942915	14.05	Zinc ion binding transcription factor	IPR001138; IPR007219
GLEAN_6904	945888	40.41	Short-chain dehydrogenase/reductase	IPR002198
GLEAN_2130	947572	27.71	Lipase/esterase	IPR013094
GLEAN_6903	949183	24.33	Duf1446 domain protein	IPR010839
GLEAN_2131	952701	0.45	Hypothetical protein	N/A
GLEAN_2132	954736	0.00	Methyltransferase	IPR013217
GLEAN_2133	957350	-3.43	Hypothetical protein w. transcriptional support	N/A
GLEAN_2134	991260	0.00	Hypothetical protein	N/A
GLEAN_2135	1004184	2.79	Zinc ion binding transcription factor	IPR001138; IPR007219
GLEAN_6902	1006716	29.42	Acyl-CoA ligase/synthetase	IPR000873
GLEAN_2136	1010511	13.29	MFS transporter	IPR011701
GLEAN_6901	1012354	35.52	Zinc-type alcohol dehydrogenase	IPR013149; IPR013154
GLEAN_2137	1013826	34.82	Aldehyde dehydrogenase	IPR015590
GLEAN_6900	1015751	-1.73	Hypothetical protein w. transcriptional support	N/A
GLEAN_2138	1019440	-2.74	Efflux transporter	N/A
GLEAN_6899	1023799	-3.36	Transcription factor	N/A
GLEAN_2139	1026243	1.06	Peptidoglycan-binding lysin protein	IPR002482; IPR008816; IPR011058
GLEAN_2140	1027715	-0.55	NADP oxidoreductase, coenzyme F420-dependent	IPR004455
GLEAN_6898	1028706	24.01	Epoxide hydrolase	IPR000073
GLEAN_2141	1030359	60.10	Flavoprotein monooxygenase	IPR002937
GLEAN_2142	1033713	5.38	Tripeptidyl peptidase A	IPR015366
GLEAN_6897	1036183	9.89	Zinc ion binding transcription factor	IPR001138; IPR007219
GLEAN_2143	1039739	55.73	FMO-like monooxygenase w. lipocalin signature	IPR013027
GLEAN_6896	1041776	49.10	Short-chain dehydrogenase/reductase	IPR002198
GLEAN_2144	1043474	46.86	FMO-like monooxygenase	IPR013027

Figure 4.1 (A) MPBs disperse during early summer, both sexes of MPB carry blue stain fungi. Beetles bore through bark, make their galleries in the phloem and deposit eggs along the gallery walls. During this process they introduce *G. clavigera* (*Gc*) and other associated microorganisms. The filamentous fungi, yeast and bacteria begin colonizing the tree, the staining fungi penetrate the xylem. The larvae feed on phloem creating galleries at right angles to the main galleries, completing their development after the fourth instar. Larvae pupate within the excavated chambers, and pupae transform into adults during early summer. While feeding the larvae and beetles accumulate associated microorganisms in their guts, on their exoskeletons and in specialized maxillary structures known as mycangia. This ensures that the appropriate microorganisms are transmitted to the next host. *Gc* colonizes the sapwood rapidly and produces a blue/black melanin pigment. Fungal growth blocks water and nutrient flow in the sapwood and phloem, and tree death occurs. (B) Some observed phenotypes of *Gc*. Light micrographs of asexual stage characterized with mononematous (i) and synnematous (ii) conidiophores reproducing conidia. (iii) light micrograph of sexual structure characterized by a spherical ascocarp oozing ascospores. (iv) stereomicrograph of conidiophores that grow inside the MPB gallery and ascocarps (inset) on the inner bark of lodgepole pine. (C) A phylogenetic tree showing the positioning of *Gc* within the pezizomycotina.

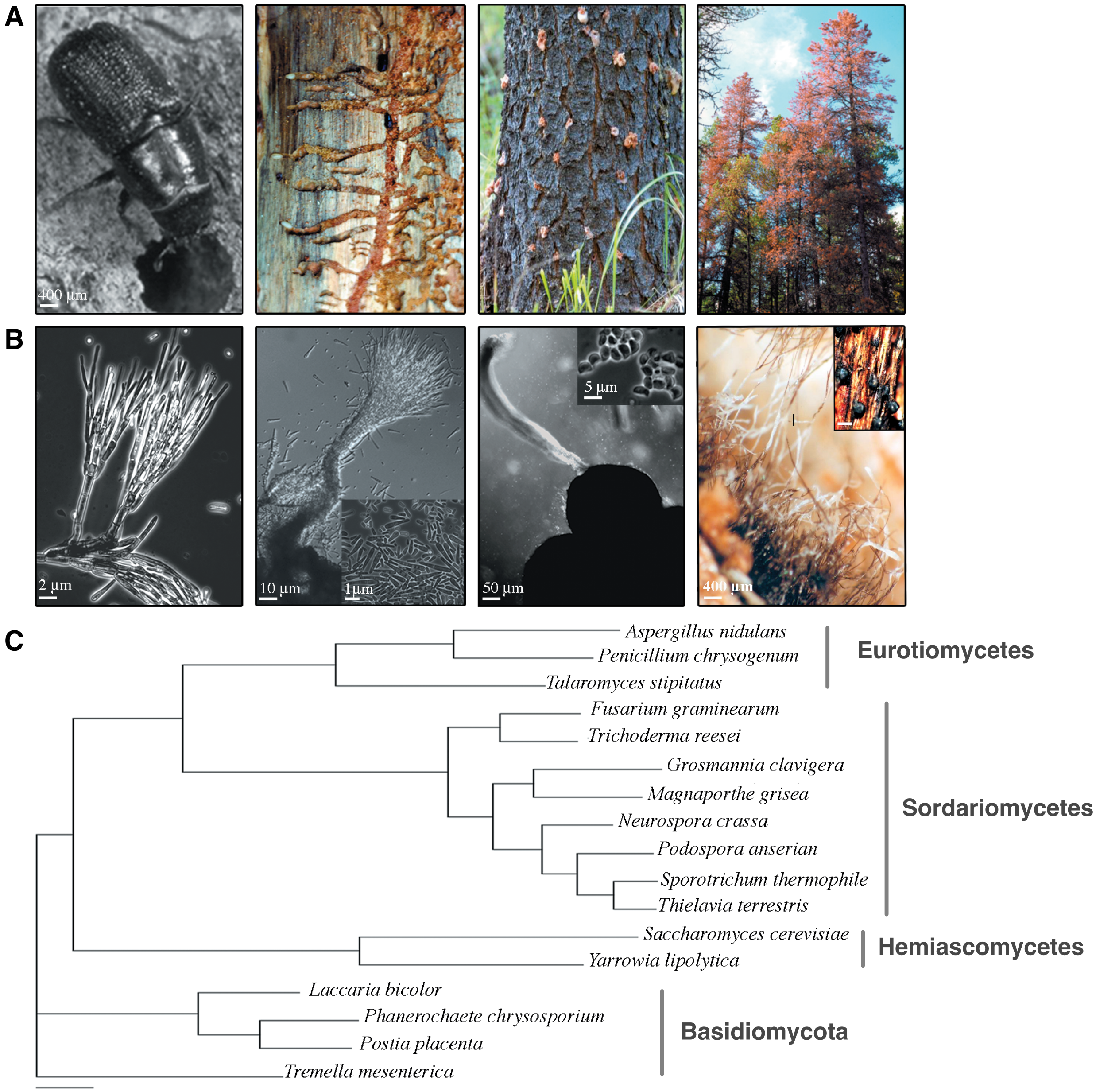

A
400 µm
B
2 µm
10 µm
1µm
5 µm
50 µm
400 µm
C
Aspergillus nidulans
Penicillium chrysogenum
Talaromyces stipitatus
Eurotiomycetes
Fusarium graminearum
Trichoderma reesei
Grosmannia clavigera
Magnaporthe grisea
Neurospora crassa
Podospora anserian
Sporotrichum thermophile
Thielavia terrestris
Sordariomycetes
Saccharomyces cerevisiae
Yarrowia lipolytica
Hemiascomycetes
Laccaria bicolor
Phanerochaete chrysosporium
Postia placenta
Tremella mesenterica
Basidiomycota

Figure 4.2 (A) Growth of *Gc* with (Red) and without (Blue) terpene treatment. Top left inset is an example 'virtual gel' generated from total RNA extracted from *Gc* mycelia 12 hr after terpene treatment. (B) Growth of *Gc* with (Red) and without (Blue) lodgepole pine phloem (LPPE) treatment. Top left inset is an example 'virtual gel' generated from total RNA extracted from *Gc* mycelia 12 hr after LPPE treatment.

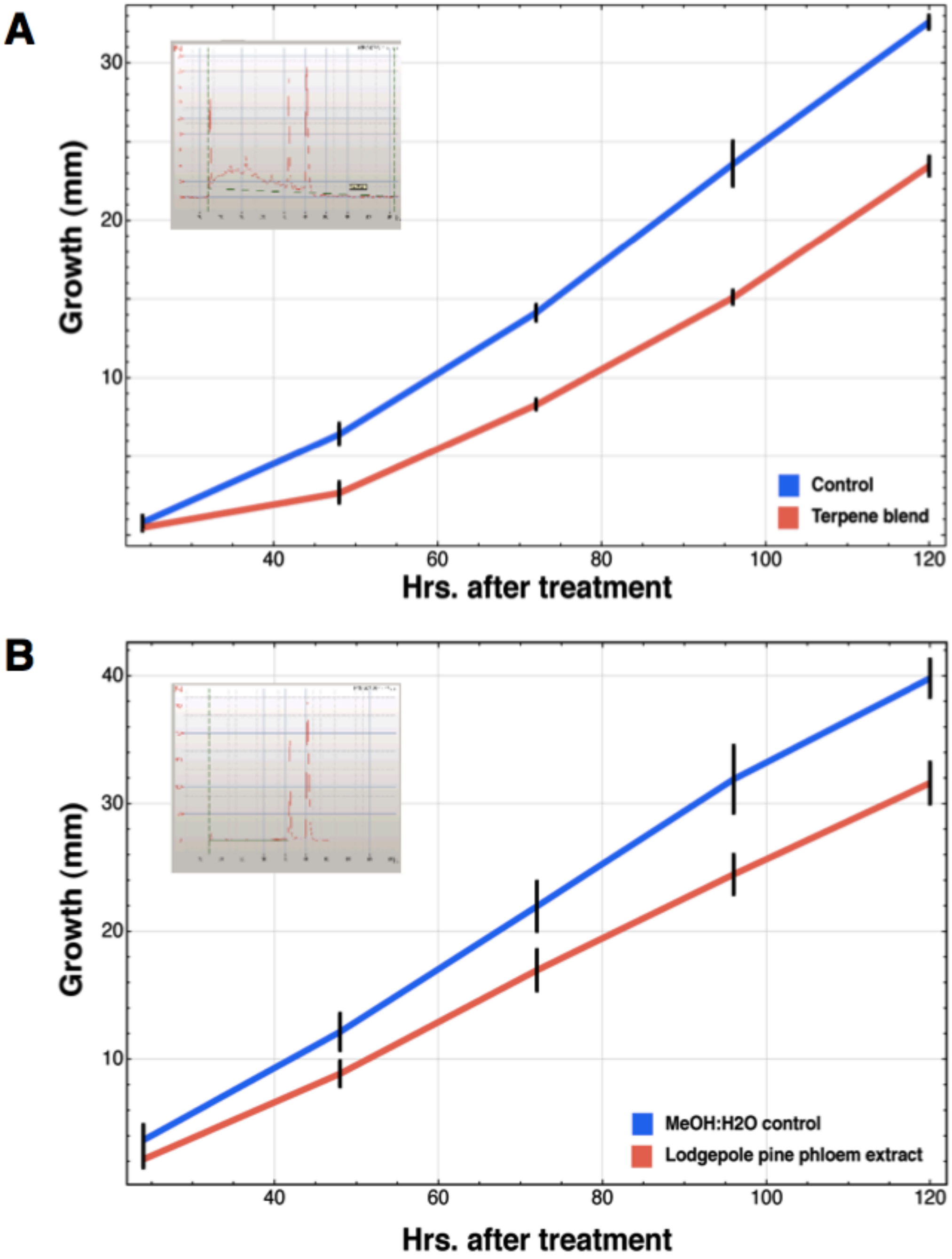


Figure 4.3 Circular venn diagram describing relationships within and between terpene- and LPPE-treated expression data. The outer ring describes the number of up- (Blue) and down- (Orange) regulated genes within a single control vs. treatment analysis. The ribbons connecting fragments of the outer ring describe the relationships shared between different analyses such that the width of the ribbons at the ring connecting point describes the number of shared relationships and the colour of the ribbons describes whether the shared relationships are induced (Blue) or repressed (Orange).

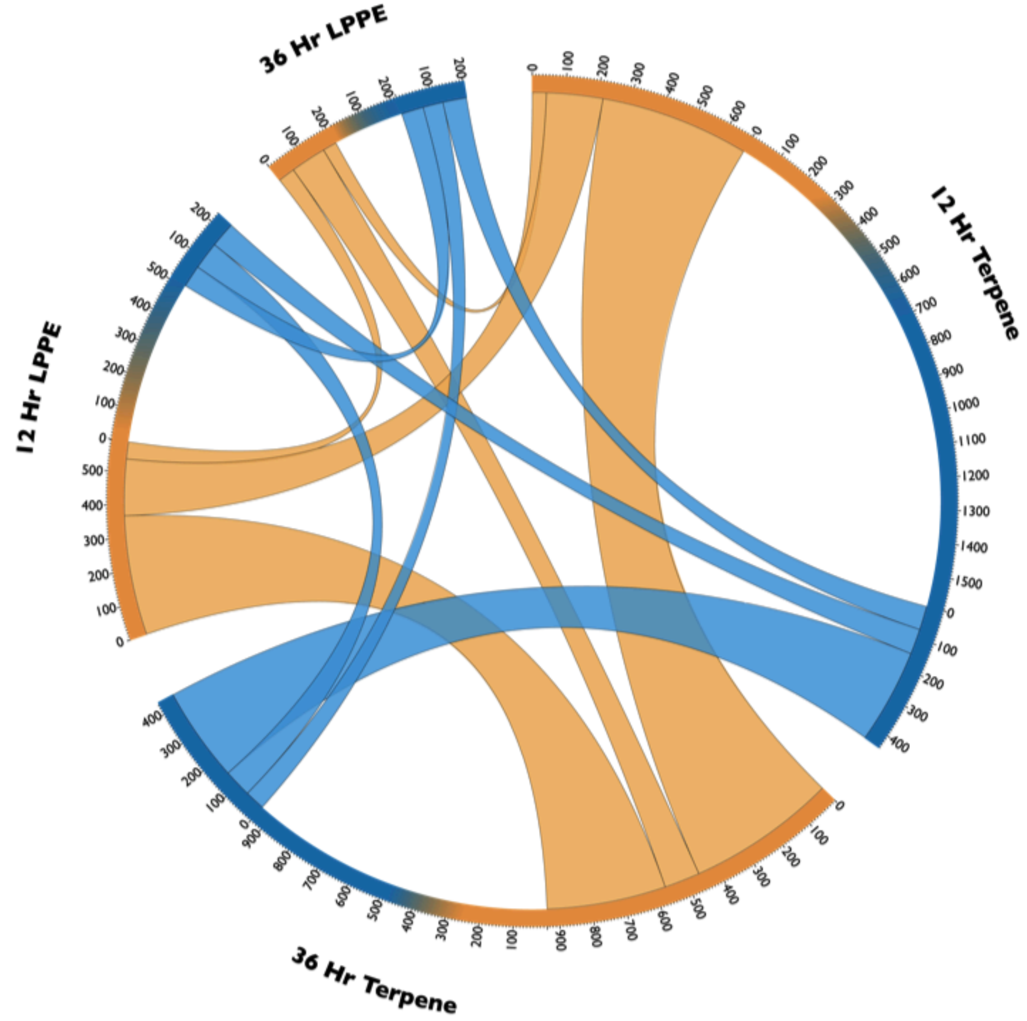


Figure 4.4 RNA-seq profiling reveals a cluster of co-expressed genes on supercontig GCSC_108. For complete details and the results of genome wide mapping data see Figure C5. From top to bottom: transposons detected during genome wide repeat masking using de novo and reference based methods. Expression analysis results derived from comparison of control vs. treatment for the 36 hr terpene-treated data averaged in 50 kb windows across GCSC_108. **Enlargement.** Log-transformed alignment coverage for the peak region of co-expression showing the agreement between predicted gene models and RNA-seq data.

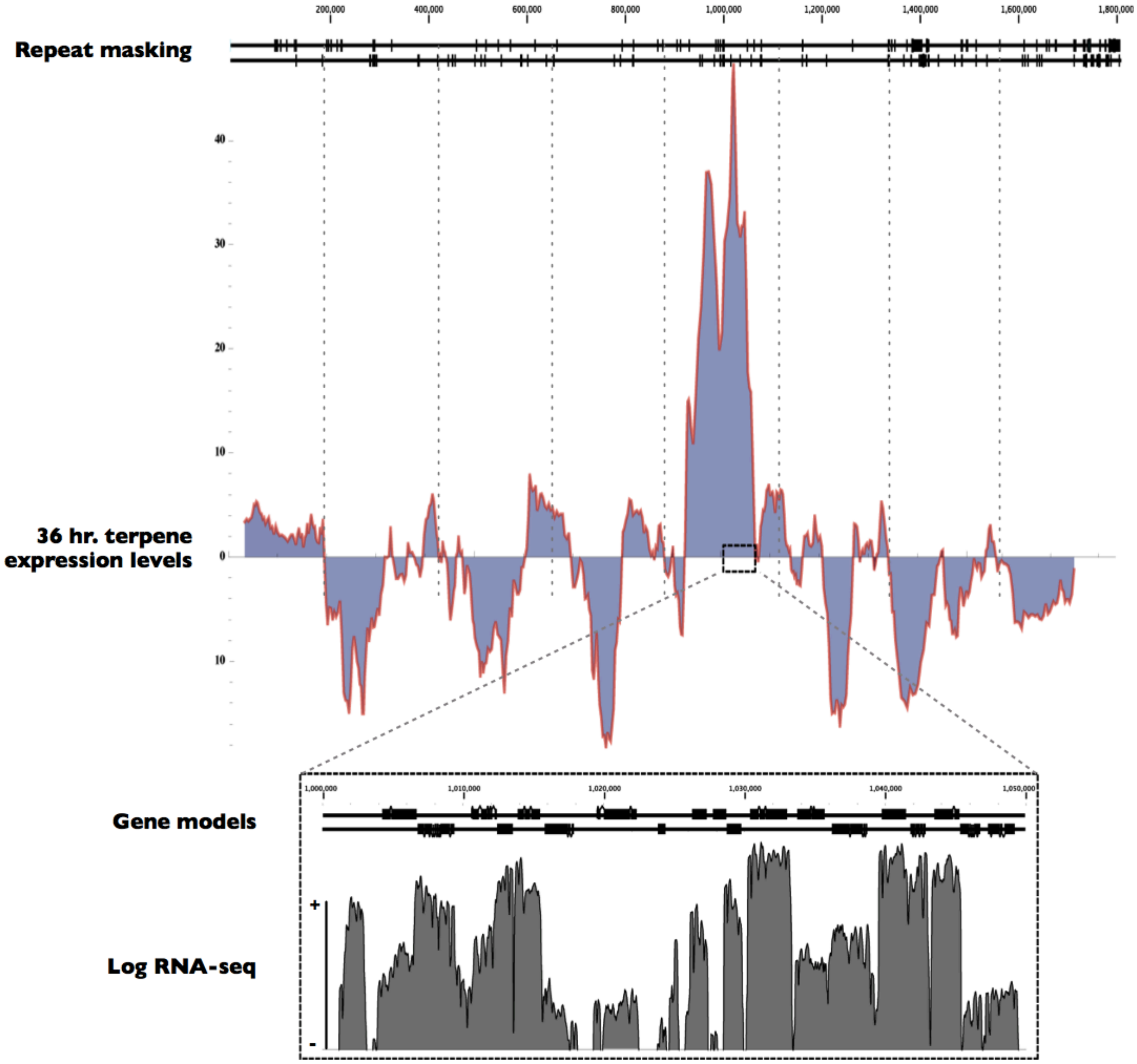
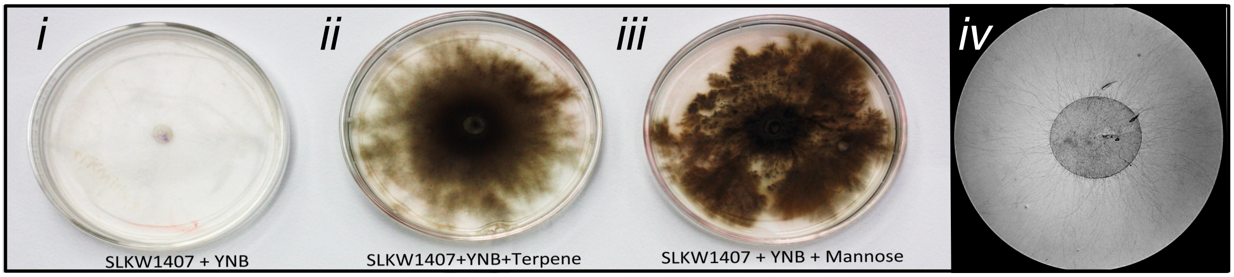


Figure 4.5 Phenotypes observed for the growth of *Gc* on minimal media with or without a carbon source. The terpene blend and maltose were tested as carbon sources. From left to right (i) minimal media without a carbon source, (ii) minimal media amended with the terpene blend, and (iii) minimal media amended with mannose. Additional plates: *iv*) close-up of the *Gc* growth on minimal media without a carbon source, highlighting the sparse searching growth phenotype.



4.5 References

Abraham, L., Hoffman, B., Gao, Y., & Breuil, C. (1998). Action of *Ophiostoma piceae* proteinase and lipase on wood nutrients. *Can J Microbiol 44*, 698-701.

Abraham, L., & Breuil, C. (1996). Isolation and characterization of a subtilisin-like serine proteinase secreted by the sap-staining fungus *Ophiostoma piceae. Enzyme Microb Technol 18*(2), 133-140.

Ayres, M., Wilkens, R., Ruel, J., & Lombardero, M. (2000). Nitrogen Budgets of phloem-feeding bark beetles with and without symbiotic fungi. *Ecology 8*(18), 2198-2210.

Bentley, D. R., Balasubramanian, S., Swerdlow, H. P., Smith, G. P., Milton, J., Brown, C. G., et al. (2008). Accurate whole human genome sequencing using reversible terminator chemistry. *Nature 456*(7218), 53-59.

Blakesley, R. W., Hansen, N. F., Mullikin, J. C., Thomas, P. J., Mcdowell, J. C., Maskeri, B., et al. (2004). An intermediate grade of finished genomic sequence suitable for comparative analyses. *Genome Res 14*(11), 2235-2244

Bleiker, K., & Six, D. (2007). Dietary benefits of fungal associates to an eruptive herbivore: potential implications of multiple associates on host population dynamics. *Environ Entomol 36*(6), 1384-96.

Bouarab, K., Melton, R., Peart, J., Baulcombe, D., & Osbourn, A. (2002). A saponin-detoxifying enzyme mediates suppression of plant defences. *Nature 418*(6900), 889-892.

Conesa, A., Götz, S., García-Gómez, J. M., Terol, J.,Talón, M., & Robles, M. (2005). Blast2GO: a universal tool for annotation visualization and analysis in functional genomics research. *Bioinformatics 21*(18), 3674-3676.

De Bie, T., Cristianini, N., Demuth, J. P., & Hahn, M. W. (2006). CAFE: a computational tool for the study of gene family evolution. *Bioinformatics 22*, 1269-1271.

DiGuistini, S., Ralph, S. G., Lim, Y. W., Holt, R., Jones, S., Bohlmann, J., et al. (2007). Generation and annotation of lodgepole pine and oleoresin-induced expressed sequences from the blue-stain fungus *Ophiostoma clavigerum* a Mountain Pine Beetle-associated pathogen. *FEMS Microbiol Lett 267*(2), 151-158.

DiGuistini, S., Liao, N. Y., Platt, D., Robertson, G., Seidel, M., Chan, S., et al. (2009). De novo genome sequence assembly of a filamentous fungus using Sanger 454 and Illumina sequence data. *Genome Biol 10*(9), R94.

Duffy, B., Schouten, A., & Raaijmakers, J. M. (2003). Pathogen self-defence: mechanisms to counteract microbial antagonism. *Annu Rev Phytopathol 41*, 501-538.

Elsik, C., Mackey, A., Reese, J., Milshina, N. V., Roos, D. S., & Weinstock, G. M. (2007). Creating a honey bee consensus gene set. *Genome Biol 8*(1), R13

Feltrer, R., Álvarez-Rodríguez, M. L., Barreiro, C., Godio, R. P., Coque, J-J. R. (2010). Characterization of a novel 246-trichlorophenol-inducible gene encoding chlorophenol O-methyltransferase from *Trichoderma longibrachiatum* responsible for the formation of chloroanisoles and detoxification of chlorophenols. *Fungal Genet and Biol 47*, 458–467.

Fernández-Cañón, J. M., & Peñalva, M. A., (1995). Fungal metabolic model for human type I hereditary tyrosinaemia. *Proc Natl Acad Sci USA 92*(20), 9132-6

Franceschi, V. R., Krokene, P., Christiansen, E., & Krekling, T. (2005). Anatomical and chemical defences of conifer bark against bark beetles and other pests. *New Phytol 167* (2), 353-375.

Galagan, J. E., & Selker, E. U., (2004). RIP: the evolutionary cost of genome defence. *Trends Genet 20*(9), 417-423.

Gao, Y., & Breuil, C. (1995). Extracellular lipase production by a sapwood-staining fungus *Ophiostoma piceae. World J Microbiol Biotechnol 11*(6), 638-642.

Gao, Y., Chen, T., & Breuil, C. (1995). Identification and quantification of nonvolatile lipophilic substances in fresh sapwood and heartwood of lodgepole pine (*Pinus-contorta* Dougl). *Holzforschung 49*(1), 20-28.

Hane, J. K., & Oliver, R. P., (2008). RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. *BMC Bioinformatics 9*, 478.

Hesse-Orce, U., DiGuistini, S., Keeling, C. I., Wang, Y., Docking, T. R., Liao, N. Y., Robertson, G., Holt, R. A., Jones, S. J. M., Bohlmann, J., & Breuil, C. (2010). Gene discovery for the bark beetle-vectored fungal tree pathogen *Grosmannia clavigera. BMC Genomics* (submitted).

Hurst, L. D., Pál, C., & Lercher, M. J. (2004). The evolutionary dynamics of eukaryotic gene order. *Nat Rev Genet 5*(4), 299-310.

Jeffers, M. R., McRoberts, W. C., & Harper, D. B. (1997). Identification of a phenolic 3-O-methyltransferase in the lignin-degrading fungus *Phanerochaete chrysosporium. Microbiology 143*, 1975-1981.

Kajander, T., Merckel, M. C., Thompson, A., Deacon, A. M., Mazur, P., Kozarich, J. W., & Goldman, A. (2002). The structure of *Neurospora crassa* 3-carboxy-cis,cis-muconate lactonizing enzyme a β-propeller cycloisomerase. *Structure 10*, 483-492

Keeling, C. I., & Bohlmann, J. (2006). Genes enzymes and chemicals of terpenoid diversity in the constitutive and induced defence of conifers against insects and pathogens. *New Phytol 170*(4), 657-675.

Kurz, W. A., Dymond, C. C., Stinson, G., Rampley, G. J., Neilson, E. T., Carroll, A. L., et al. (2008). Mountain pine beetle and forest carbon feedback to climate change. *Nature 452*(7190), 987-990.

Lee, S., Kim, J-J., & Breuil, C. (2006a). Diversity of fungi associated with the mountain pine beetle *Dendroctonus ponderosae* and infested lodgepole pines in British Columbia. *Fungal Diversity 22*, 91-105.

Lee, S., Kim, J-J., & Breuil, C. (2006b). Pathogenicity of *Leptographium longiclavatum* associated with *Dendroctonus ponderosae* to *Pinus contorta. Can J For Res 36*(11), 2864-2872.

Li, L., Stoeckert, Jr. C., & Roos, D. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res 13*, 2178-2189.

Lieutier, F., & Berryman, A. (1988). Preliminary histological investigations of the defence reactions of 3 pines to *Ceratocystis-clavigera* and two chemical elicitors. *Can J For Res 18*(10), 1243-1247.

Lieutier, F., Yart, A., & Salle, A. (2009). Stimulation of tree defences by Ophiostomatoid fungi can explain attack success of bark beetles on conifers. *Annals of Forest Science 66*(8), 801-823.

Liu, C. J., Dixon, R. A., (2001). Elicitor-induced association of isoflavone O-methyltransferase with endomembranes prevents the formation and 7-O-methylation of daidzein during isoflavonoid phytoalexin biosynthesis. *Plant Cell 13*(12), 2643-58.

Männistö, P. T., and Kaakkola, S. (1999). Catechol-O-methyltransferase (COMT): biochemistry molecular biology pharmacology and clinical efficacy of the new selective COMT inhibitors. *Pharmacol Rev 51*, 593-628.

Pareja-Jaime, Y., Roncero, M. I. G., & Ruiz-Roldán, M. C. (2008). Tomatinase from *Fusarium oxysporum* f sp lycopersici is required for full virulence on tomato plants *Mol Plant Microbe Interact 21*(6), 728-736.

Preisig, C. L., Matthews, D. E., & VanEtten, H. D. (1989). Purification and characterization of S-adenosyl-L-methionine: 6-a-hydroxymaackiain 3-O-methyltransferase from *Pisum sativum. Plant Physiol 91*, 559-566.

Price, A. L., Jones, N. C., & Pevzner, P. A., (2005). De novo identification of repeat families in large genomes. Proceedings of the 13th Annual International conference on Intelligent Systems for Molecular Biology (ISMB-05), Detroit Michigan.

Raffa, K., & Berryman, A. (1983). Physiological-aspects of lodgepole pine wound responses to a fungal symbiont of the mountain pine-beetle *Dendroctonus ponderosae* (Coleoptera Scolytidae). *Canadian Entomologist 115*(7), 723-734.

Seybold, S., Bohlmann, J., & Raffa, K. (2000). Biosynthesis of coniferophagous bark beetle pheromones and conifer isoprenoids: Evolutionary perspective and synthesis. *Canadian Entomologist 132*(6), 697-753.

Sharpton, T., Stajich, J., Rounsley, S., Gardner, M., Wortman, J., Jordar, V., et al. (2009). Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives. *Genome Res 19*(10), 1722-31.

Smith, D. J., Park, J., Tiedje, J. M., & Mohn, W. W. (2007). A large gene cluster in *Burkholderia xenovorans* encoding abietane diterpenoid catabolism. *J Bacteriol 189* (17), 6195-6204.

Thieringer, R., & Kunau, W. H. (1991). The β-oxidation system in catalase-free microbodies of the filamentous fungus *Neurospora crassa. J Biol Chem 266*, 13110-13117.

Walton, J. D. (2000). Horizontal gene transfer and the evolution of secondary metabolite gene clusters in fungi: An hypothesis. *Fungal Genet Biol 30*, 167-171.

Wang, Y., DiGuistini, S., Wang, T-C. T., Bohlmann, J., & Breuil, C. (2010). *Agrobacterium*-meditated gene disruption using split-marker in *Grosmannia clavigera* a mountain pine beetle associated pathogen. *Curr Genet 56*(3), 297-307.

Zabel, R. A., & Morell, J. J. (1992). Wood stains and discolorations. In: Zabel, R. A., & Morell, J. J. (Eds) *Wood microbiology: decay and its prevention,* pp 326-343.

Zeeberg, B. R., Feng, W., Wang, G., Wang, M. D., Fojo, A. T., Sunshine, M., et al. (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol 4*(4), R28.

Zheng, Z., & Shetty, K. (2000). Solid-state bioconversion of phenolics from cranberry pomace and role of *Lentinus edodes* β-Glucosidase. *J Agric Food Chem 48*(3), 895-900.

5 Conclusions

A genome sequence captures most of an individual's hereditary information including its transcript and protein encoding sequences and substantial 'non-coding' information that play important regulatory roles in the control of the expression of these transcripts and proteins. Importantly, such a sequence does not fully inform the range of possibilities that exists for a given species or the ways in which these components will manifest in response to the environment. However, a genome sequence is a powerful collection of data for understanding the mechanisms underlying a trait of interest and ultimately the traits underlying the biology of a given organism.

For researchers working within one of the relatively small number of species for which a reference genome sequence is available, new research paradigms have been established. Genome annotation predicts where the functional elements of the genome are located and expression analysis provides a starting point for forward genetic and functional experiments. This creates a tremendous opportunity for rapidly driving wet lab experiments using strong, data driven hypothesis.

Genome-based resources and the tools for analyzing them are changing rapidly. In this thesis we developed a set of genomics-based resources for *G. clavigera* using publicly available bioinformatic tools; however, because these resources were built using the newest sequencing platforms, in addition to the publicly available software custom methods and tools were also necessary. The methods developed highlight the possibilities for utilizing advanced sequencing technologies for studying non-model biological systems. The resources that were created will become a foundation for future bioinformatic and laboratory based studies that explore the biology of *G. clavigera*.

The best method for validating the *G. clavigera* genomic resources was to perform preliminary analysis of *G. clavigera* biology. We chose to focus this analysis on *G. clavigera's* tolerance for host-specific metabolites. This analysis focused on the detoxification of the primary chemical defenses in lodgepole pine phloem extracts and oleoresin and attempted to link the new genomic resources to laboratory-based experiments.

This preliminary work has begun to identify key genes and pathways that will need to be assessed and validated. The effects of these genome variations on phenotypes should be characterized in populations over large geographic regions. In the long term, this approach has the potential to support development of new ways of managing and controlling MPB outbreaks.

5.1 Genomics in a non-model system

In this thesis we reported on genomic resources created for *G. clavigera* including the whole genome sequence, ESTs, and RNA-seq data. We annotated the fungal genome using these genomic resources (e.g. ESTs and RNA-seq) and computational (gene prediction software) data.

5.1.1 Generating a genome sequence for *G. clavigera*

Due to the novelty of assembling a genome sequence from mixed read data as reported in chapter three, it was necessary to develop new bioinformatic methods such as data encoding, quality filtering and iterative assembly methods. They were required for handling the enormous quantities of read data, mixing read data originating from different sequencing platforms, and facilitating quality control (QC) analysis that leveraged the benefits provided by the enormous sequencing depth that the NGS platforms produced.

We demonstrated the success of our assembly methodology by several means described in chapter three. In addition, to the standard QC methods previously applied for assessing genome assembly quality, we developed a novel method that utilized NGS data. This method is useful for assembly QC and finishing; however, the approach is also adaptable for semi-automated genome assembly and was used as an assembly methodology for the panda genome sequence (Li et al., 2009), which was the first eukaryotic genome reported to be assembled purely from ultra-short NGS read data.

5.1.2 Generating EST and RNA-seq resources for annotating the *G. clavigera* genome sequence

To annotate the *G. clavigera* genome sequence we used both experimental and computational data. The experimental data were a combination of EST, RNA-seq and protein coding sequences. The protein coding sequences were obtained from the Swissprot public database. EST and RNA-seq data were specific to *G. clavigera* and generated under biologically relevant conditions ensuring that transcript sequences reflected our interests and planned analyses. EST normalization procedures increased the representation of rarer transcripts. The sequence resources generated in chapter two of this thesis are novel because they were the first molecular resources generated for *G. clavigera* in a genome wide fashion; they represented the first opportunities for assessing the effects of terpenes on fungal growth and the effects of growth on wood on gene expression. Following publication of the 'EST manuscript' an additional ~50 K ESTs were generated extending the original collection and these are touched upon in chapter four. The additional resources provided first insight into population level variations as they were generated, in part, from seven additional *G. clavigera* strains. The second EST collection, due to its increased size and breadth, also provided opportunities for alternative transcript analysis, digital profiling, and further experimental treatment development. A key aspect in developing EST sequencing resources was also the development of experimental treatments and an experimental process: 1) we developed experimental treatments and performed preliminary physiological experiments, 2) used the experimental treatments for generating fungal cultures and building cDNA libraries, 3) sequenced and performed digital profiling, 4) validated digital profiling with qRT-PCR, 5) built larger experiments for genome wide expression profiling based on new information generated above, 6) QC'd expression profiling experiments using qRT-PCR validation experiment, 7) committed fungal RNA for RNA-seq expression profiling, 8) analyzed genome wide expression profiling data, 9) developed new physiological experiments for exploring genome wide expression data and 10) developed functional experiments such as gene knockouts for validating new physiological experiments.

Development and analysis of RNA-seq data is an active area of research in both the development of wet lab and data analysis methodologies. The RNA-seq data in this manuscript were generated and analyzed through a period of extremely active progress in this field, and as such the data were analyzed on several occasions. Due to the evolving nature of the genome sequence and annotations and the need for homogenous read data and statistical treatments; each analysis was performed on the entire data collection. Early data analysis efforts were entirely customized and all analyses were performed with custom scripts developed on an as needed basis. One change over this period was the development of new and improved tools that simplify the processes of mapping and analyzing the RNA-seq read data. The initial novelty of this work was in developing methods for integrating RNA-seq data into gene model predictions. Integration of the RNA-seq data into gene model predictions was tested in three different ways: 1) using a read-mapping strategy that included the development of splice junction libraries, 2) de novo assembly methodologies with direct integration into composite gene model predictions and 3) integrating into ab initio gene prediction by providing gene location hints. Work in this field by many others has since superseded these efforts (one example: Zheng et al., 2010). In addition to methodology development, this work generated biological insights, protein coding SNP sequences, and a rich gene expression profiling resource describing the responses of *G. clavigera* to LPPE and terpene treatments in the early stages of host colonization.

5.1.3 Annotating the *G. clavigera* genome sequence

A significant accomplishment of chapter four is the description of *G. clavigera* genome annotations, including the methodology that underlies their generation and functional descriptions. This work has facilitated all subsequent analyses. The methods applied to annotating the *G. clavigera* genome were similar to those described for other fungal genome sequencing projects. The functional descriptions ascribed to the protein coding gene regions predicted from the *G. clavigera* genome are a good starting place for inferring the function of any given gene. As the numbers and types of databases used in the initial functional descriptions were limited by time and resources, future work will certainly yield improved gene function annotations. Of course, as the data available in the public domain increases (i.e. more computational and experimental resources), revisiting these annotations will also yield substantial improvements. Nevertheless, the current annotations have proven to be a valuable resource, and inferences made from expression profiling data based on theses annotations have resulted in successful laboratory experiments.

5.2 Examining *G. clavigera* growth on sapwood and lodgepole pine defense metabolites

Most of the work undertaken in this thesis involved the development of methods and resources aimed at creating a rich genomic resource for supporting laboratory based experiments. The primary motivation that underpinned my interest was to explore the biological interactions of *G. clavigera* with defense chemicals in its primary host, lodgepole pine. To this end, preliminary analysis was performed with the new genomic resources, and findings from these analyses led to the work and data reported in chapter four.

5.2.1 Inhabiting the wood substrate

To facilitate our understanding of *G. clavigera's* growth in trees we undertook computational and laboratory based experiments for examining the enzymes involved in carbon and nitrogen acquisition by this fungus. In *G. clavigera*, we identified glycoside hydrolases, peptidases, lipases and esterases. We found a lipase that is likely involved in using pine triglycerides, a major carbon source for this fungus. We identified several peptidase families whose importance were suggested by orthoMCL cluster analysis. We used peptide sequencing to complement signal peptide analysis and to validate the significance of candidates identified from the genome sequence. We looked for genes that we anticipated to be involved in either cellulose or lignin degradation. This analysis suggested that genes missing from the genome sequence may explain why *G. clavigera* is not able to use the cellulose and lignin in lodgepole pine cell walls or affect the structural properties of the colonized wood.

5.2.2 Detoxifying host-specific defense metabolites

I began this research with the hypothesis that understanding *G. clavigera's* growth on host specific metabolites would draw my attention to interesting aspects of *G. clavigera* biology and then to the interesting elements of its genome sequence, since, I believed that *G. clavigera's* interaction with host-specific metabolites such as phloem extracts (phenolics) and terpenes is a critical feature underpinning its host-vector specificity. The preliminary analysis indicates that this hypothesis likely holds true or, at least, warrants further investigation. Findings, such as the large domain of co-expressed genes in the *G. clavigera* genome are indicating that genomic organization could in part be influenced by *G. clavigera's* relationship with host-defense metabolites. The PDR-type ABC transporter identified in the terpene treatment gene expression data indicate that a multidrug resistance phenotype is likely important in the early survival of this fungus following deposition by the MPB in lodgepole pine. The co-ordinated expression of genes in the beta-oxidation pathway provide evidence that the metabolism of terpenes by this fungus may be an interesting phenotype which facilitates its symbiotic relationship with the mountain pine beetle.

G. clavigera's responses to phloem extracts (phenolics) and terpenes are largely different. This observation is not surprising, given the differences in the basic chemical natures of these different types of natural product families. The LPPE blend is complex, and the compounds responsible for inducing the observed reduction in growth are not clear at this point in the research.

5.3 Perspective on future work

Although costs for sequencing are decreasing rapidly and recent signs indicate that a sequencer on every lab bench may become a reality in the not-distant future, the primary challenge for researchers in non-model organisms is to develop tools that can manage the substantial quantities of data emerging from these sequencers. The tools currently available are too difficult to use for a bench scientist and too naive for being useful to manage the enormous quantities of multidimensional data that can be generated. To explore the richness of an ever expanding collection of data new algorithms will be required that create the opportunity for automatic experimental analysis. A system whereby each new data collection becomes incorporated into previous collections and new reports and analysis are generated on the fly that allow scientists to make a day's experimental decisions in real time based on the results of the previous days sequencing runs is necessary.

5.3.2 Further resource development for *G. clavigera*

To improve upon the current *G. clavigera* genome resources, there are several directions that could be pursued in the near future: 1) improving upon and enriching the current reference genome sequence annotations, this would likely involve substantial amounts of manual labour, 2) expanding our knowledge of the *G. clavigera* genome sequence across more individuals from the *G. clavigera* population, and 3) adding genome sequences for additional species with strategic phylogenetic positions that will facilitate further bioinformatic annotation of the reference genome sequence such as for inferring important genome regulatory elements and key evolutionary events that have led to increasing bark beetle specialization.

5.3.3 *G. clavigera* host-specific metabolite detoxification

Our preliminary analysis of *G. clavigera* growth following treatment with LPPE or terpenes indicates a number of interesting possibilities worth pursuing in future research: 1) aromatic degradation pathways may be ordered into discrete units, 2) growth on terpenes appears to involve chromatin remodeling for facilitating the expression of large regions within the genome which might normally be located in heterochromatic regions, nucleosome occupancy experiments might clarify the role of chromatin remodeling in the response of *G. clavigera* to terpenes, 3) the toxic effects of terpenes on *G. clavigera* are dependent on environmental factors such as nutrient availability and assessing these factors in relationship to increasing terpene toxicity and within the lodgepole pine host population could provide an alternative approach for selecting lodgepole pine trees with increased resistance.

5.4 References

Li, R., Fan, W., Tian, G., Zhu, H., He, L., Cai, J., et al. (2009). The sequence and de novo assembly of the giant panda genome. *Nature*. 10.1038/nature08696

Zhang, G., Guo, G., Hu, X., Zhang, Y., Li, Q., Li, R., et al. (2010). Deep RNA sequencing at single base-pair resolution reveals high complexity of the rice transcriptome. *Genome Res*, 1-10.

Appendix A. Supplementary information supporting chapter 2

A.1 Fungal strain and culture conditions used in chapter 2

The diterpene blend consisted of abietic (Sigma, Oakville, ON), isopimaric and pimaric acids (Orchid-Helix Biotech, Vancouver, BC). Each diterpene was separately dissolved in DMSO at 1% (w/v), then the 3 solutions were mixed in a 2:1:1 ratio. The monoterpenes, (+/-)α-pinene, (-)β-pinene, 3-carene, β-phellandrene, (+/-)limonene, α-terpinolene and γ-terpinene, were blended, aliquoted onto watch-glasses and placed into the bottoms of 2 L glass bells. The relative composition for the selected metabolites was similar to that described by Shrimpton (Shrimpton, 1973) with the exception of high limonene levels. The composition of the blended monoterpenes was verified by sampling the headspace directly using Solid Phase Micro Extraction (SPME) and gas chromatography (data not shown). The supplied β-phellandrene was mixed equally with (+/-)limonene (kindly donated by Millenium Chemicals, Jacksonville, FL), and therefore the limonene was about 30 times higher than reported by Shrimpton (1973). All other monoterpenes were purchased from Sigma (Oakville, ON).

A.2 RNA isolation, cDNA library normalization used in chapter 2

RNA was isolated using the RNeasy Plant RNA isolation kit (Qiagen, Mississauga, ON). This was followed with a DNaseI (Fisher Scientific, Ottawa, ON) treatment and LiCl precipitation to yield a genomic-DNA-free RNA pellet (≥1 mg of RNA) whose purity was confirmed by spectrophotometer and agarose gel analysis. The RNA pellet was re-suspended and poly(A^+) RNA was obtained by two rounds of purification using the Poly

(A) Purist mRNA purification kit (Ambion, Austin, TX). The Stratagene cDNA library was constructed following the manufacturer's protocol except for four modifications: 1) first-strand synthesis was performed using Superscript II reverse transcriptase (Invitrogen, Mississauga, ON), 2) an anchored oligo d(T) primer (Table 1) was substituted, 3) gel size selection in a 1% Nusieve GTG low melting point agarose gel (BioWhittaker Molecular Applications, Rockland, ME) was used to maintain libraries with inserts of 650 bp or greater and, 4) β–agarase (New England Biolabs, Pickering, ON) was used to recover cDNA following gel fractionation.

C_ot normalization relies upon the assumption that cDNA re-annealing follows second-order kinetics. Because rarer transcripts will anneal less rapidly, a) a cDNA library can be hybridization with a pool of amplified partial cDNA sequences originating from the primary library, and b) single stranded cDNAs can be recovered following incubation periods that yield progressively more normalized cDNA pools. C_ot can be approximated from the equation:

$$(\text{O.D}/2)*0.45*\text{time} = C_ot.$$

Where 1µg DNA/µl = 20 O.D. Incubations were performed at 30°C. To confirm successful normalization, both the single stranded and double stranded fractions were collected and titered. A comparison of titer between these two sub-libraries reveals approximately the degree to which cDNAs have re-annealed.

A.3 Sequence annotation used in chapter 2

Sequences were filtered for a) excessively long PolyA tails (if >85% of the sequence 62 bp past the anticipated 18 bp polyA tail is composed of A's then the sequence was disregarded) and b) contaminating *E. coli*, *Agrobacterium*, gymnosperm (using a custom assembly of gymnosperm sequences downloaded from GenBank) and *Saccharomyces cerevisiae* genes. For the analysis of P450 and ABC proteins additional *Gc* ESTs were identified using RPS-BLAST (Altschul, et al., 1997), and motif-based searches using Fuzztran from EMBOSS v. 2.8.0 (Rice, et al., 2000). BLAST comparisons were done in reciprocal format; i.e. using known P450s or ABC sequences as queries against a BLAST-formatted database of *Gc* ESTs and by using *Gc* ESTs as queries against a BLAST-formatted database of known P450s or ABC proteins. TBLASTn was used to compare *Aspergillus fumigatus* DHN-melanin pathway genes (Langfelder, et al., 2003) against the *Gc* unique sequences.

A.4 Quantitative real-time PCR

Following is a brief description of the data analysis for qPCR. Cycle threshold (ct) values were manually assigned to both the gene-of-interest and an internal control gene, and a relative measure of gene expression, Δct, was calculated using the formula $\Delta ct = 2^{-(ct_{gene\text{-}of\text{-}interest} - ct_{reference\text{-}gene})}$. Biological and technical replicates were analyzed by a nested analysis of variance (ANOVA) and between-treatment differences were subjected to a Bonferroni's multiple comparison. The critical α-value was determined for three comparisons.

A.5 References

Altschul, SF, Madden, TL, Schaffer, AA, Zhang, J, Zhang, Z, Miller, W & Lipman, DJ (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389-3402.

Langfelder, K, Streibel, M, Jahn, B, Haase, G & Brakhage, AA (2003) Biosynthesis of fungal melanins and their importance for human pathogenic fungi. Fungal Genet Biol 38: 143-158.

Rice, P, Longden, I & Bleasby, A (2000) EMBOSS: the European Molecular Biology Open Software Suite. Trends Genet 16: 276-277.

Shrimpton, DM (1973) Extractives associated with wound response of lodgepole pine attacked by the mountain pine beetle and associated microorgansims. Can. J. Bot. 51: 527-534.

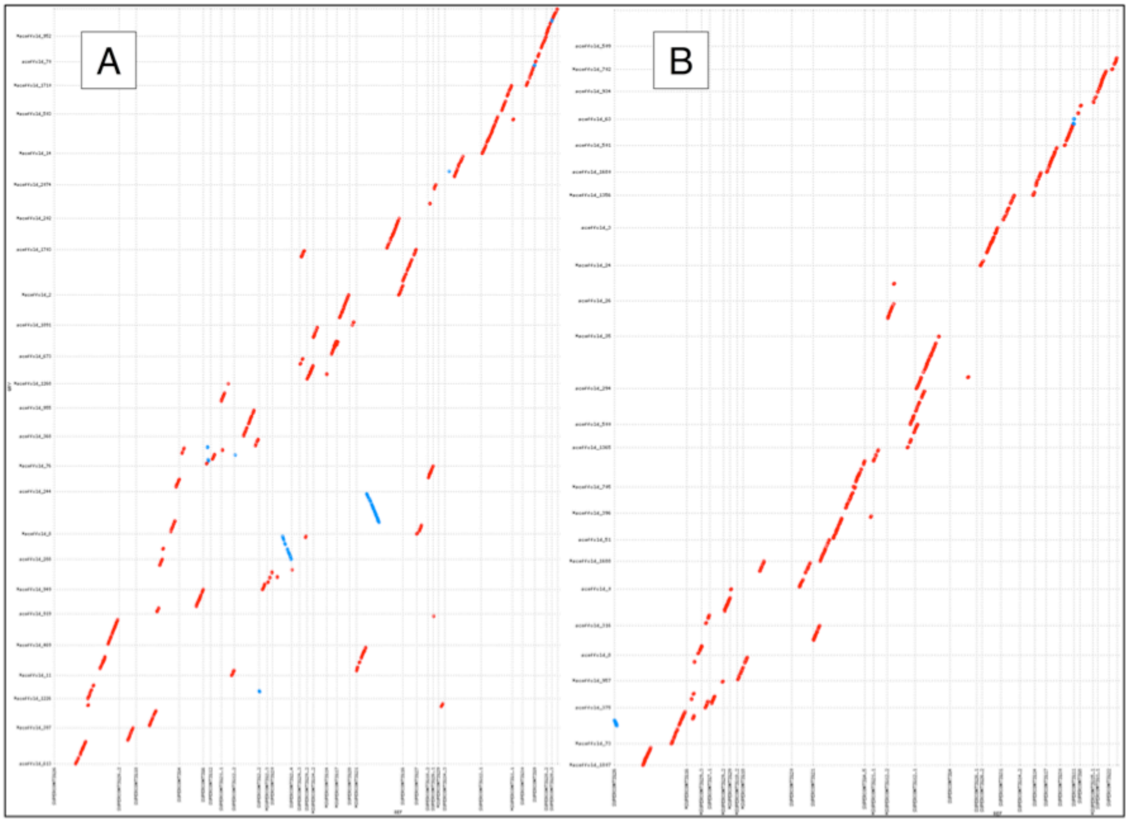
Appendix B. Supplementary information supporting chapter 3

B.1 454 SE read filtering

Following the methods developed by Huse et. al. [1] we used an empirical system for removing low quality 454 SE read data. In the Sanger/454 series of assemblies filtering reads by no-calls or length increased the number of consistently paired Sanger reads on the same scaffold by 54 and 27, and decreased the number of EST-detectable misassemblies by 9 and 3 respectively. Additionally, length filtering reads reduced the 454 associated indel rate by ~10%. Although applying both filtering methods reduced the number of EST-detectable misassemblies by 4, it also decreased the number of consistently paired Sanger reads on the same scaffold by 72, compared with the unfiltered 454 read assembly. As filtering reduced the 454 read coverage, the number of contigs assembled with Sanger reads decreased. This occurred because removing 454 read data prevented overlap alignments with Sanger reads and subsequent integration of the two data types. However, the higher quality data improved the assembled contigs, because more accurately placed Sanger PE reads improved assembly scaffolding (Figure S2). This became more critical as the quantity and variety of the read data increased in later assemblies.

Figure B1. Filtering by no-calls, length and complexity improved Sanger/454 hybrid assemblies.

A. Alignment of the top 25 scaffolds from the assembly generated using filtered and trimmed Sanger reads and unprocessed 454 reads (Table 1) against the manually finished reference genome sequence. The top 25 scaffolds account for X bp of sequence, and the range from largest to smallest was X to X respectively. The contig N50 for this draft assembly was X kb. **B.** As for (A), but for an assembly generated with the 454 read collection after filtering for no-calls, length and complexity. The top 25 scaffolds accounted for X bp of sequence, and the range from largest to smallest is X to X respectively. The contig N50 for this draft assembly was X.



B.2 Trimming and filtering Illumina PE reads

To remove low quality Illumina PE read data we profiled average quality scores (Q) and substitution rates along reads (Figures B3, B4). , we used MAQ [15] with Illumina's GA_{ii} calibrated quality scores to align the reads to an intermediate assembly that we had generated with Forge using only the filtered Sanger and 454 read data. From these read mappings, we calculated the average rate of substitution at each read position. For these calculations, we used reads that passed Illumina's default filtering (chastity $\geq$ 0.6) and had up to three mismatches in total (seed length = 28 bp; seed alignment $\leq$ 2 mismatches). Substitution rates were evaluated against the fraction of reads within three quality score ranges at the same read position. For the ~200-bp library (Figure B4A) quality scores decreased at a moderate rate along a read, then decreased sharply between 31/32 and then again at read positions 35/36. These sharp transitions, which are likely artefacts of the Illumina base-calling pipeline, suggested positions where trimming might improve the Velvet assembly by removing low quality bases at the 3' ends of reads. Both the substitution rate and the rate of lowest quality base calls increased gradually towards read ends. Low quality reads (quality score (Q) $\leq$ 10) had a higher rate of base substitutions relative to the reference assembly (Figure S5A). This rate increased by ~7% between read positions 28 and 29, likely due to an artefact of the MAQ 28-bp alignment seed length, and then increased steadily towards the read end. Empirically we found that filtering reads containing low quality base calls beyond position 28 (*i.e.* QRL(Q10)=28) improved the quality of the Velvet assembly.

For the Illumina ~650-bp library used to generate the reference assembly we started with a collection of ~24.2 M 50 bp PE reads. Shadow filtering removed ~2.9 M reads and purity filtering removed ~7.3 M reads. In the initial read mapping results we observed a large number of reads that mapped with a zero gap distance and required filtering. We removed ~2.3 M reads that had a zero gap distance and then applied a QRL(Q10)=42 filter based on the results plotted in Figures B3, 4 and 5B, removing ~4.1 M reads.

Figure B2. Assessing the quality of Illumina PE read data by average quality score (Q) at each read position

An average read quality score was calculated from quality scores (Q) extracted from the Illumina base-calling pipeline's 'export' file and plotted for each read position.

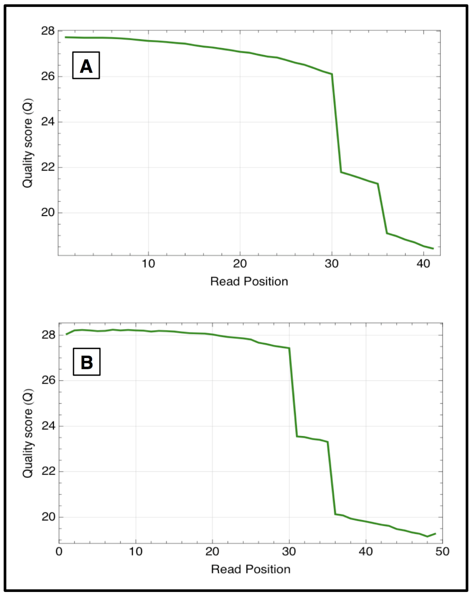


Figure B3. Assessing the quality of Illumina PE read data with profiles of substitution errors and read quality scores.

Read quality scores were extracted from the Illumina base-calling pipeline’s ‘export’ file. Base substitution sequencing errors were estimated as follows. MAQ was used to align reads to the preliminary Sanger/454 read assembly. From these read alignments, mismatches were tabulated and compared with the Illumina quality scores as a function of read position. The inset shows the library fragment length distribution determined from distances between mapped read pairs (measured from 5’ to 3’ of the mapped reads).

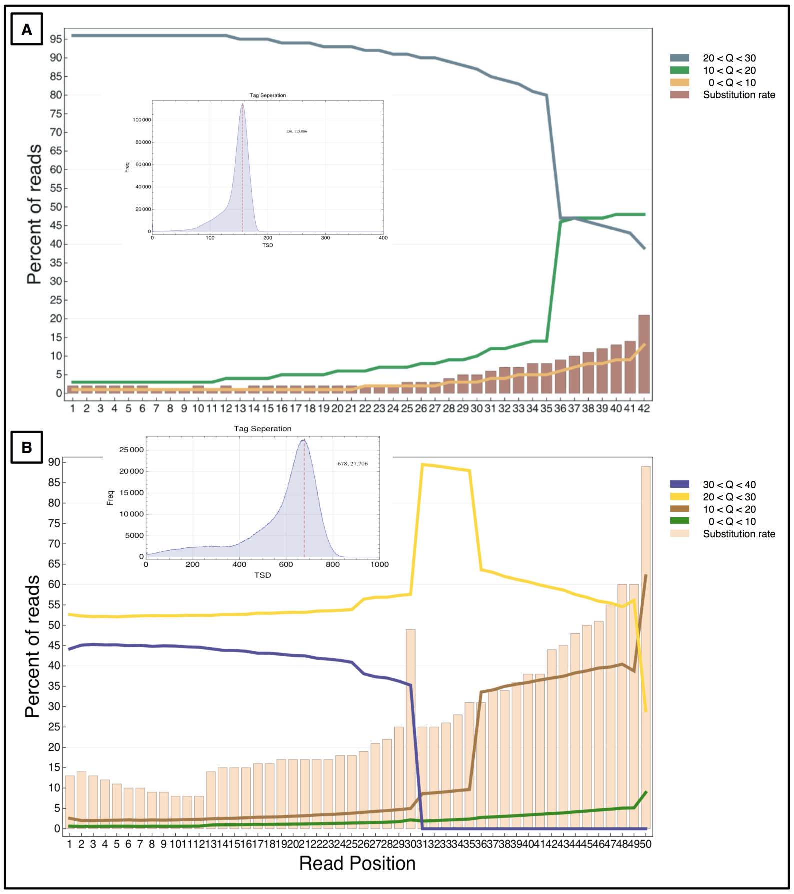
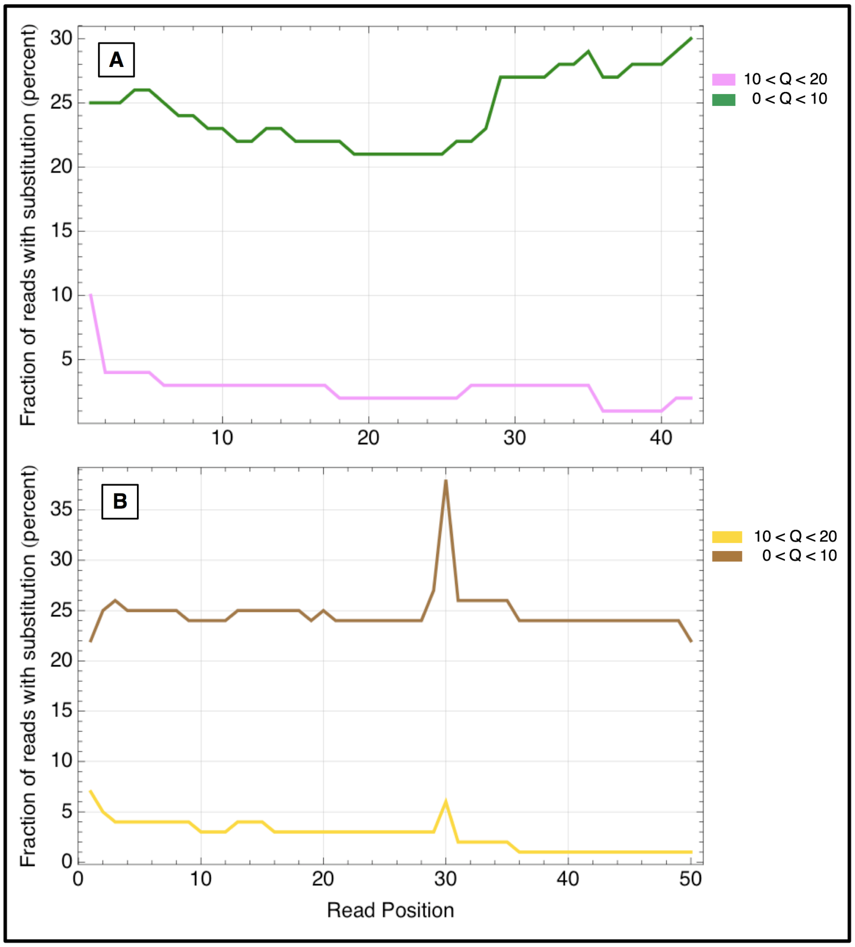


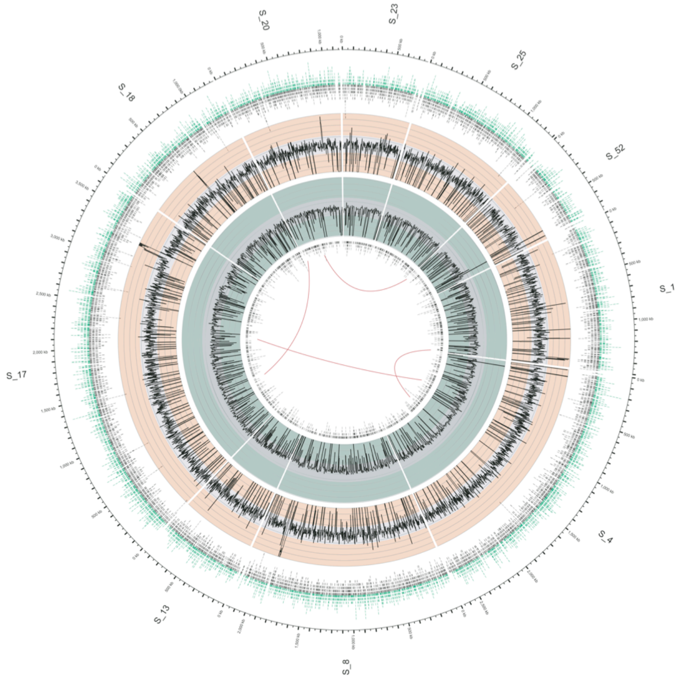
Figure B4. Fraction of reads with substitutions by read position

Using the MAQ data reported in Figure B4, we plotted the fraction of reads within each quality bin reported to have a sequence variation compared with the reference sequence.



B.3 Assessing Forge assemblies that integrate Illumina PE read data directly.

Figure B5. Overview of the largest 10 scaffolds from the best Forge assembly. From the outside towards the inside of the Circos plot (mkweb.bcgsc.ca/circos): 1) Bold numbering indicates the scaffold IDs. 2) Solid black lines represent scaffolds, and are marked by a kb length scale. 3) Green tiles indicate density of Sanger PE reads placed by Forge. 4) Orange tiles indicate density of the preassembled Velvet contigs placed by Forge. 5) Black plot/orange background indicates 454 read coverage averaged over a 2.5 kb window. 6) Black plot/green background indicates the coverage of correctly paired Illumina GA$_{ii}$ data averaged over a 2.5-kb window. 7) Black tiles indicate the density of sequences masked by RepeatMasker. 8) Arcs indicate pooled Illumina PE alignment clusters spanning different assembly scaffolds; these clusters indicate scaffold misassemblies.



Appendix C. Supplementary information supporting chapter 4

C.1 Genome finishing

We used a hybrid shotgun sequencing approach that combined data from the Sanger, 454 FLX and Illumina GA_{ii} sequencing platforms. We achieved at least 64x genome sequence coverage across 90 % of the finished genome sequence and ~9x scaffolding coverage resulting from Sanger generated read pairs from a 40 kb fosmid library. Sanger read data supported 44 % of this completed consensus sequence (Figure 4.2). The longest contig was 3.6 Mb and the contig N50 was 1.2 Mb (N90: 187,604 bp). We confirmed that the assembly was complete by aligning to it 99.4 % of 7,169 unique Expressed Sequence Tag (EST) sequences. Only 304 (4.2 %) of the EST-to-genome were partial alignments and located at contig edges. Using alignments of Illumina paired-end (PE) reads generated from 200, 700, and 3000-bp libraries, we identified and manually resolved clusters of read-pair discrepancies that originated within the draft assembly.

C.1.1 Genome finishing stages

1. Editing of Forge assembly and first round of fosmid walks followed by additional manual editing.
2. Additional Forge output brought into finishing assembly, manual merges based on 700-bp Illumina library, continued manual editing.
3. Second plate of fosmid walks plus additional unique contigs brought into the finishing

sequence, further manual editing. Unique contigs generated from a series of ABySS (Simpson et al, 2009) short read assemblies.

4. Incorporation of mini-assemblies generated from reads and mates mapped to contig ends.
5. Extensive manual editing.
6. Third plate of fosmid walks followed by continued manual editing.
7. Incorporation of 3-kb insert PE Illumina library.
8. Variant, insertion and deletion correction performed using Consed's autocorrection and by aligning series of Illumina short read assemblies (genomic and RNA-seq) and Sanger EST data to the finished assembly and identifying consensus variants.

Figure C1. A. Tracking contig N50, scaffold N50 and total contig number (inset) following each finishing stage. **B.** After completing the contig breaking, joining and scaffolding phase of finishing, insertions and deletions (indels) were identified and a semi-automated method for removing them was established. Indels were removed iteratively because each round of indel removal improved the detection of further indels by improving sequence-to-genome alignments. **C.** The effect of indel removal on genome sequence phasing could be measured through changing lengths of exons predicted by the ab initio gene prediction tool Augustus (Stanke & Waack, 2003). We observed that the removal of indels from the *Gc* genome led to increased lengths of predicted exon sequences.

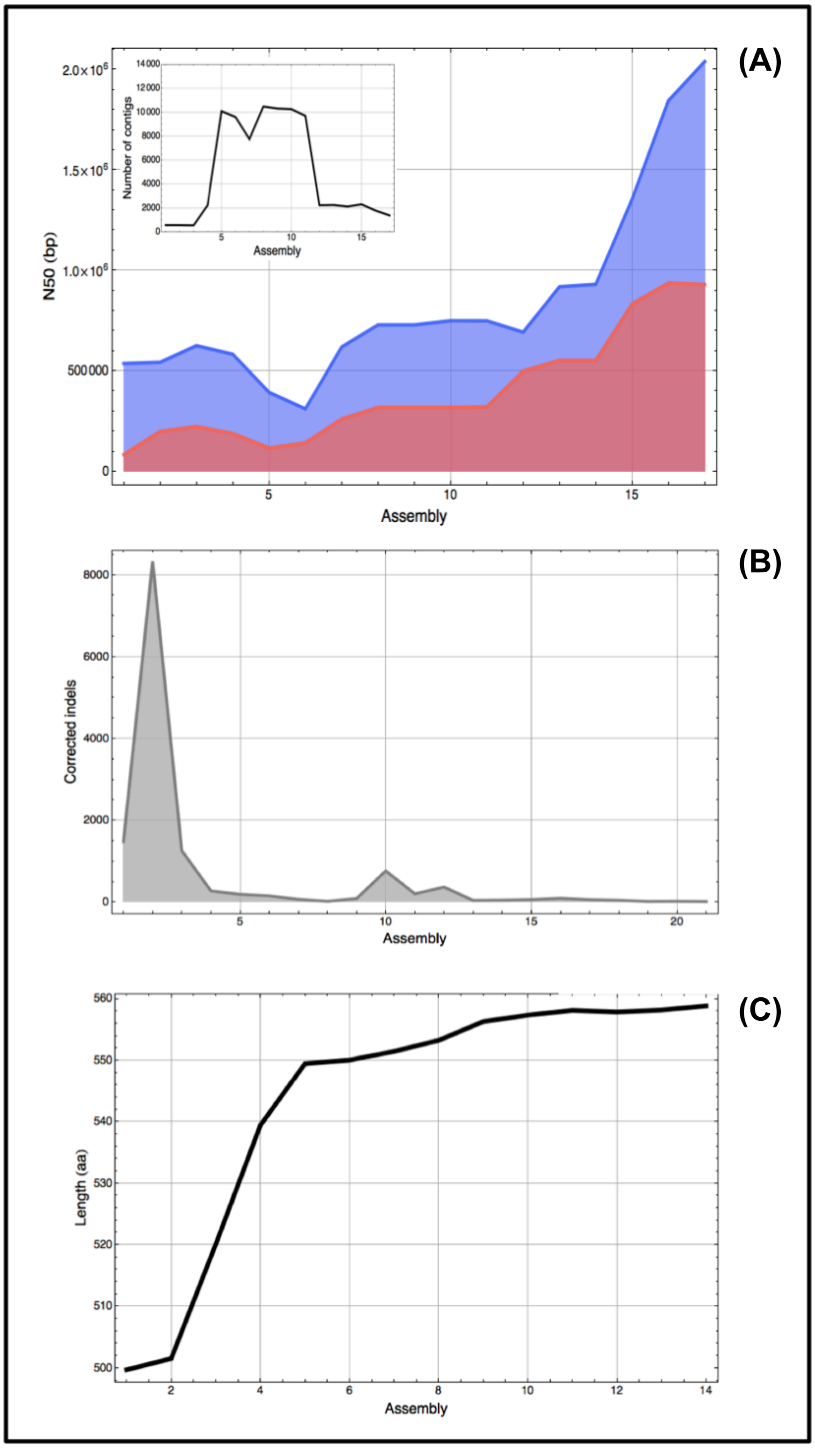

(A)
N50 (bp)
2.0×10⁶
1.5×10⁶
1.0×10⁶
500000
0
5
10
15
Assembly
Number of contigs
14000
12000
10000
8000
6000
4000
2000
0
5
10
15
Assembly
(B)
Corrected indels
8000
6000
4000
2000
0
5
10
15
20
Assembly
(C)
Length (aa)
560
550
540
530
520
510
500
2
4
6
8
10
12
14
Assembly

Figure C2 Genome read coverage was calculated for each sequencing platform at each position in the consensus sequence. Less than half (44%) of the completed genome is supported by Sanger read data. 454 read coverage is uniformly distributed. Illumina read data tails significantly to the left suggesting bias in the library preparation or sequencing.

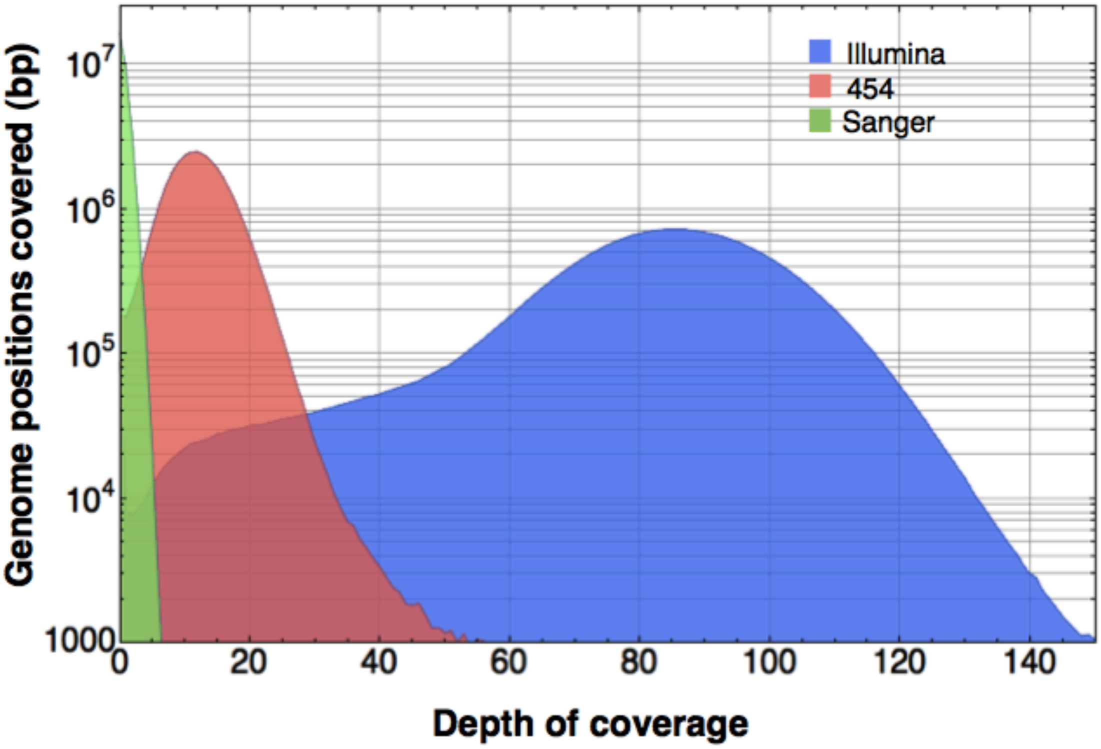
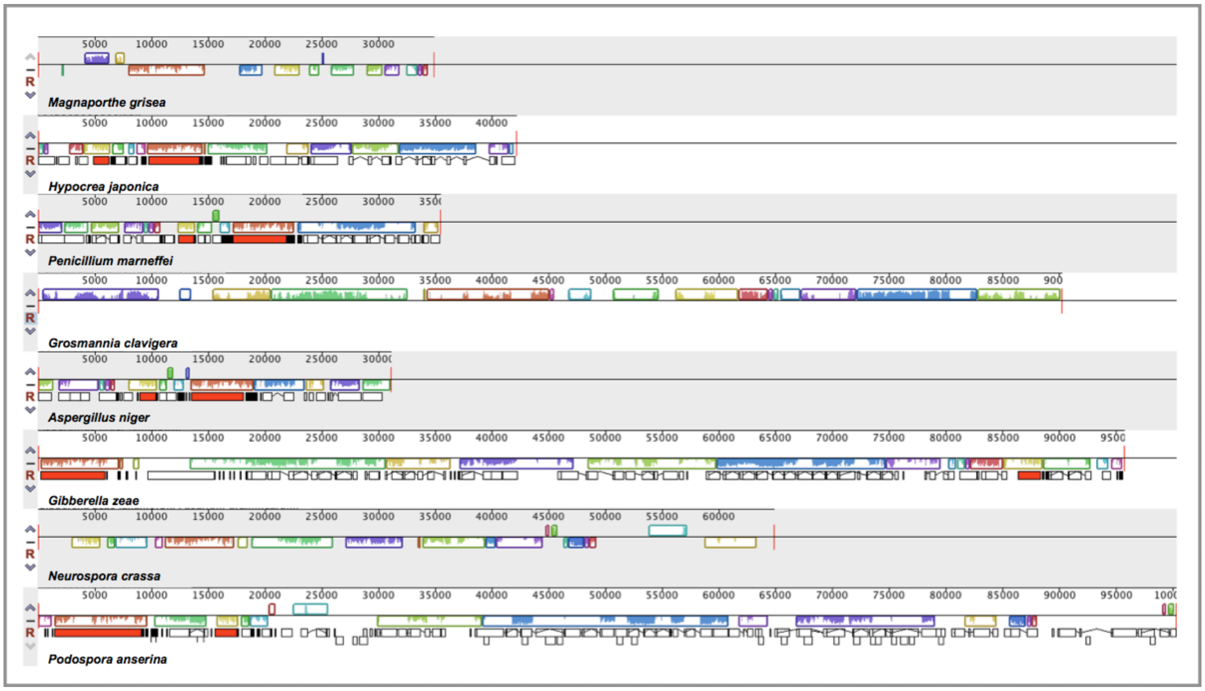


Figure C3. The mitochondria of *Gc* was assembled into a single ~90-kb sequence. Alignment of the *Gc* mitochondrial DNA sequence using Mauve (Darling et al., 2004) to other ascomycete fungi indicated that the assembly was accurate. Coloured blocks reflect co-linear regions of the alignment shared amongst the mitochondrial sequences. Coloured blocks appearing above the line indicate sense strand alignments vs coloured blocks below the line indicate complimentary strand alignments.



C.2 Predicting protein coding gene models for *G. clavigera*

Prior to generating predictions, mitochondrial and repetitive DNA sequence were removed or masked respectively. Briefly, RepeatMasker was used for screening against RepBase (Kohany, Gentles, Hankus, & Jurka, 2006) and for filtering *de novo* repetitive elements identified using RepeatScout (Price, Jones, & Pevzner, 2005). *Ab initio* predictions were performed using Augustus, SNAP and Twinscan (Korf, 2004; Korf, Flicek, Duan, & Brent, 2001). The reference used for running Twinscan was the *Magnaporthe grisea* genome sequence. During iteration one, software were run with the best available parameters. Proteins from the SwissProt database were aligned to the genome using Exonerate (Slater & Birney, 2005; percent 50; minintron 15; maxintron 3000; bestn 1) with the model protein2genome. BLAT was used for aligning *Gc* EST sequences to the reference genome (Kent, 2002; tileSize 11; fine; maxIntron 3000; extendThroughN; minIdentity 92; trimHardA; trimT) and then filtered using pslCDnaFilter (localNearBest 0.005; ignoreNs; bestOverlap). All predictions and alignments were then combined using GLEAN (Elsik, et al., 2007) to generate a single consensus set of gene models. Using this GLEAN collection as a training set a second iteration was performed for Twinscan and Augustus after refining the model parameters. 5’ and 3’ UTR predictions were generated by combining overlapping Augustus *de novo* predictions with expressed sequence evidence and reporting the median value where they aligned to GLEAN genes.

Approximately 50 K ESTs supported ~5.0 K gene models. Using CAP3 to combine a filtered (100-bp minimum contig size) ABySS assembly of ~29 M paired-end (PE) RNA-seq tags from the reference strain with the ~50 K ESTs resulted in 7,147 contigs of which 6,717 aligned to the genome sequence and provided support for ~5.5 K gene models. Evidence generated by aligning protein sequences from the SwissProt database to the genome using exonerate resulted in gene model support for 943 GLEANs. In total ~6.0 K GLEAN-generated gene models were supported by evidence from one of EST, EST/ABySS, or SwissProt.

C.3 Phylogenetic tree construction

Orthologous relationships were determined for a 17-way clustering from the complete genomes of the 17 fungal taxa: *Aspergillus nidulans, Fusarium graminearum, Hypocrea jecorina (T. reesei), Penicillium chrysogenum, Grosmannia clavigera, Magnaporthe grisea, Neurospora crassa, Podospora anserina,* Myceliophthora *thermophila, Talaromyces stipitatus, Thielavia terrestris, Laccaria bicolor, Phanerochaete chrysosporium, Postia placenta, Tremella mesenterica, Yarrowia lipolytica, Saccharomyces cerevisiae*. All predicted protein sequences for the genomes of these fungi were searched against each other using BLASTP and clustered into orthologous groups using OrthoMCL (Li et al., 2003) with default parameters. Single-copy orthologs were identified as the clusters with exactly one member per species. Phylogenetic relationships were determined from these single-copy orthologs and were aligned with MAFFT (Katoh and Toh, 2008). Alignments were pruned with Gblocks (Castresana, 2000), and data were combined into a single alignment file containing 191,447 characters (phylogenetic informative sites: 121,656). Phylogenetic trees were constructed using three methods: parsimony (Phylip v. 3.6.7; Felsenstein, 1989), Bayesian (MrBayes; Huelsenbeck & Ronquist, 2001), and maximum likelihood (RAxML;

Stamatakis et al. 2006). Divergence time estimation was performed using nonparametric rate smoothing (NPRS) (Britton et al., 2007) using the maximum likelihood tree as a starting point.

C.4 Positive selection analysis of Methyltransferases

Candidate methyltransferases were identified from the translations of the predicted gene model collections from the genomes of *A. nidulans, F. graminearum, G. clavigera, M. grisea, N. crassa* and *S. cerevisiae* using hmmsearch (hmmer.janelia.org; PFAM: PF08241-2). Protein sequences were aligned using MAFFT and the resulting phylogenies were used for choosing sequences for further analysis. Transcript sequences were substituted into the protein alignments (pal2nal; Suyama, Torrents, & Bork, 2006) and positive selection was assessed by comparing the fixed (where ω, the ratio of the dNonsynonymous and dSynonoymous substitution rates, is constant across all phylogenetic branches) and free-ratio (where ω was estimated for each phylogenetic branch from the data) evolutionary models using PAML (v. 4; Yang, 1997) in a likelihood ratio test of the model's maximum likelihood scores.

C.5 Peptide sequencing for the identification of extracellular proteins

C.5.1 Generating biological materials

Gc strain kw1407 was used for protein extraction. Lodgepole pine branches (5-7 cm in diameter) were debarked, cut, and pulverized in a Thomas Wiley Standard Model 4 mill with a 2 mm screen. The resulting sawdust was autoclaved, scooped into 100 X 15 mm Petri dishes, and submerged with water-agar 1.5 %. A conidial suspension was inoculated on the solidified sawdust medium overlaid with a sterile cellophane sheet (10^6 conidia / plate), and incubated at 21°C for three days.

C.5.2 Protein extraction

Extracellular proteins were obtained from *Gc* grown on 10 sawdust-agar plates. Cellophane sheets colonized with *Gc* were transferred to 50 ml of 50 mM acetate buffer (pH 5.0) containing 100 mM NaCl, and a protease inhibitor cocktail (Complete Mini; Roche), and incubated on a tube rotator for one hour at 21°C. The suspension was centrifuged at 4000 RPM for 5 min to remove the large particles and the supernatant was filtered through a 0.2 µm filter (Millipore). The protein solution was concentrated by centrifugal ultrafiltration with an Amicon Ultra 10,000-molecular-weight-cutoff filter (Millipore). The protein concentration was determined by the Bio-Rad *DC* protein assay.

C.5.3 1D SDS-PAGE and LC-MS/MS

The protein sample (15 µg) was mixed with β-mercaptoethanol-supplemented SDS-PAGE loading buffer (ratio 1:1), heated for 5 min at 65°C, and loaded in a single lane on a 0.75 mm-thick 10% SDS-acrylamide gel. After separation, the gel was stained with Coomassie blue and the lane was cut into 16 slices.

In-gel protein digests were performed on a MassPrep™ liquid handling station (Micromass Ltd) according to the manufacturer's specifications and using sequencing grade modified trypsin (Promega). Extracted peptides were completely dried using a SpeedVac and resuspended in 10 µl of 0.1% formic acid in water. Final extracts were analyzed by tandem mass spectrometry (LC MS/MS) using an LTQ (ThermoElectron) quadrupole ion trap mass spectrometer equipped with a nanospray source and a Surveyor autosampler and HPLC system (Thermo Electron Corporation).

C.5.4 Data analysis

A Python (v2.5) script was designed to split large genomic contigs into 100-kb fragments. The genome sequence and ESTs (DiGuistini et al. 2007) were translated in all six frames using Transeq (Rice et al. 2002). Gene models were added to the predicted protein database. This database was reversed, and a concatenated database containing the forward and reversed database was created. MS/MS spectra were interpreted using the Mascot® (Perkins et al. 1999) and SEQUEST® (Eng et al. 1994) programs and searched against the *Gc* forward and reversed concatenated database. The resulting SEQUEST and Mascot files were loaded into Scaffold® 2.0 (Proteome Software). For all fractions, we grouped all files from Mascot together as one biological sample, and all files from SEQUEST as another biological sample. The Scaffold software was used to validate protein identifications derived from MS/MS sequencing results. This software verifies peptide identifications assigned by SEQUEST and Mascot using the X!Tandem database searching program (Craig and Beavis 2003) and then probabilistically validates these peptide identifications using PeptideProphet (Keller et al. 2002) and derives corresponding protein probabilities using ProteinProphet (Nesvizhskii et al. 2003). Ninety-five percent peptide and 50% protein identification thresholds were used as the cut-offs; protein identification required two peptides per protein for identification.

The Scaffold peptide report was exported to MS Excel. Python scripts were written to automatically search and annotate the identified proteins using the Blastp and InterProScan programs against the UniProt database and InterPro (v. 17.0; Mulder et al. 2007) databases respectively. Gene Ontologies (GO) definitions were mapped to the identified proteins according to both the InterPro and UniProt accession numbers obtained. GO terms were then classified into functional groups according to molecular function, cellular component and biological process. Each peptide was mapped back to the genome and formatted to allow visualization in the Artemis genome browser (http://www.sanger.ac.uk/Software/Artemis).

Signal peptides were predicted using neural network (NN) and hidden Markov model (HMM) algorithms implemented on the SignalP 3.0 server (Nielsen et al. 1997). Protein with a statistically significant positive signal peptide as analyzed by the NN and the HMM (P>95%) algorithms received a score of 2. A score of 1 indicates that the signal peptide was predicted by only one of the algorithms, while a protein with a score of 0 was not predicted to be secreted by NN and HMM.

C.6 Protein coding SNP discovery using RNA-seq

C.6.1 Generating biological materials

Mycelia and spores used for extracting RNA and construct libraries were collected from solid media inoculated with a suspension containing 5x10^5 spores and were incubated for ~5-7 days (depending on library) under ambient temperature in the dark. Mycelia for each of the eight isolates were grown on seven media-treatment combinations independently; spores were generated using a single set of conditions from kw1407 exclusively. Treatments were as follows: 1) wood (W), 10g/plate lodgepole pine sawdust, 1.5% granulated agar; 2) starch (S), 0.17% YNB, 1.5% granulated agar, 1% starch, 0.1% potassium hydrogen phthalate (PHP) and 0.3% asparagine; 3) organic nitrogen (ON), 0.17% YNB, 1.5% granulated agar, 1% maltose, 0.1% PHP and 0.3% asparagine; 4) inorganic nitrogen (IN), 0.17% YNB, 1.5% granulated agar, 1% maltose, 0.1% PHP and 0.3% $NaNO_3$; 5) olive oil (OO); 0.17% YNB, 1.5% granulated agar, 1% (v/v) olive oil, 0.1% PHP, 0.5% tergitol and 0.3% asparagine (media was autoclaved without olive oil and the detergent tergitol, olive oil and tergitol were added using a sterile blender mixing at low speed before aliquoting to each plate); 6) lodgepole pine methanol extract (LPPE), 0.17% YNB, 1.5% granulated agar, 1% maltose, 0.1% PHP, 0.3% asparagine and 200 µl of the crude lodgepole pine methanol extract (50:50, MeOH:H2O). The crude lodgepole pine methanol extract was prepared as follows: a lodgepole pine bolt from a freshly cut tree was brought to the lab and froze to -20˚C. While still frozen, the bolt was cut into disks and the phloem separated. The phloem was then ground in a mill with liquid nitrogen. The ground phloem was then extracted in 80:20 methanol:water (2.5 mL/g) and sonicated at 4 °C for 2 hr. After centrifugation, the supernatant was removed and concentrated by 1/3 under a gentle flow of nitrogen gas. This concentrated crude extract was stored at –20°C for no longer than 8 months prior being use. For the spore (Sp) treatment, mycelia were grown on 1% MEA (Difco, England) and after seven days of incubation spores were washed from the culture surface using 3 ml of sterile water.

C.6.2 Generating RNA-seq data

RNA-seq data was generated with an Illumina (GA_{ii}) using the following methodology. Poly(A^+) mRNA was purified from 10 µg of DNaseI treated total RNA using the MACS^{TM} mRNA Isolation Kit (Miltenyi Biotec, Germany). Double-stranded (DS) cDNA was then synthesized from this poly(A^+) mRNA using the Superscript^{TM} DS cDNA synthesis kit (Invitrogen, USA) with 5 µM of random hexamer primers (Invitrogen). The cDNA was sheared for 10 minutes using a Sonic Dismembrator 550 (cup horn, Fisher Scientific, Canada), and then size separated in an 8% PAGE gel. The 220-to-250 bp fraction was excised and eluted from the gel slice overnight at 4 °C in 300 µl of elution buffer (5:1, LoTE buffer (3 mM Tris-HCl, pH 7.5, 0.2 mM EDTA)-7.5 M ammonium acetate), and was purified using a Spin-X filter tube (Fisher Scientific) followed by ethanol precipitation. The library was prepared using Illumina's paired-end library construction protocol (Illumina Inc., USA). Briefly, the cDNA was end-repaired and phosphorylated using treatment with the T4 and Klenow DNA Polymerases, and the T4 polynucleotide kinase in a single reaction. The 3’ adenine overhang was then ligated to the Illumina PE adapters, which contain 5’ thymine overhangs. The adapter-ligated cDNAs were purified on a Qiaquick spin column (Qiagen), then PCR-amplified with Phusion DNA Polymerase for 10 cycles using Illumina’s PE primer set (Illumina). PCR-products were purified on a Qiaquick MinElute column (Qiagen) and the quality and quantity were determined using both an Agilent DNA 1000 series II assay and a Nanodrop 1000 spectrophotometer (Nanodrop, USA) then diluted to 10nM. Clusters were generated from this purified material on the Illumina cluster station and then run on the GA$_{ii}$. Post run analysis was performed with version 1.3 of the Illumina software pipeline.

C.6.3 Predicting SNPs from the *G. clavigera* genome

PE reads were aligned to the reference genome sequence using CLC Genomics workbench (CLCbio, DK) using the following parameters: mismatch score cut-off = 8, mismatch cost = 2, alignment = ungapped, min distance = 25, max distance = 350. Using these read-to-genome alignments, SNPs were predicted. Predictions were performed within the Genomics workbench with the following parameters: average quality = 20, central quality = 20, maximum gaps & mismatches = 2, minimum coverage = 8, maximum coverage = 5000, minimum minor frequency = 10 %, minimum minor count = 3, windows length = 15. Minor allele frequency and minor allele read counts were combined with an “or” rule within the software package and we found post-prediction filtering for SNPs with allele frequencies higher than 10 % “and” minor read counts higher than 3 to be beneficial.

C.6.4 Validating RNA-seq SNP predictions

To validate the predicted RNA-seq variants we sequenced 22 gene model regions from the genomes of each strain. From 71 predicted variants within the amplified sequence regions, 70 (98.6 %) were validated. During this validation procedure an additional 13 variants were observed. Analysis of the tag-to-genome alignments indicated that these variants were present in the RNA-seq collection but overlooked during prediction due to low minor allele frequencies.

For each amplicon, a forward and a reverse primer was designed (18 to 24 bp long), with melting temperature ranges from 48 to 58 °C, GC content from 45 to 65 % and 2 G/C (i.e. 2G, 2C or GC) in the 3' end triplet. Before amplifying genes, the concentration of each primer was adjusted to 10 pmolµL^{-1}. *Gc* spores were spread onto cellophane overlaid on 1.5% agar containing 1% malt extract in 15 cm petri dishes. The fungal spores were incubated at 22 °C in the dark for eight days, and mycelia were removed from the cellophane and pooled. DNA was extracted from mycelia following the method of Möller et al. (1992) but without first lyophilizing the mycelia. Briefly fungi were frozen and then ground with liquid nitrogen. TES buffer was added to the fungal powder in a tube. Proteinase K, NaCl, CTAB, phenol, chloroform were added successively and the mix was spun many times. At the end of the protocol, isopropanol was added to reveal DNA, the solution was washed with ethanol, dried in the air and put in nano water.

Primers were tested on reference DNA to determine the optimum melting temperature (Tm), a Tm of 55.5 °C was chosen. For each set of primers, a master mix of 23 µL was prepared: 1 µL of DNA, 1 µL of dNTP, 4 µL of BSA, 2.5 µL of Buffer, 0.65 µL of DMSO, 0.65 µL of formamide, 13 µL of nanopure water. 0.2 µL of Taq polymerase was added at the end in the master mix. Then 1 µL of each primer was added. The parameters for the amplification were one cycle for 5 min at 95 °C (separation of the DNA strands); 33 cycles at 95.0 °C for 30 sec, at 55.5 °C for 1 min (primer binding step), and at 72 °C for 1 min and 20 sec (elongation step); and one cycle at 72 °C for 10 min. Then PCR products were stored at 4 °C before use. For the electrophoresis gel, a mix of 100 mL of buffer and 0.7 g of agarose was prepared. After heating, 1.5 µL of EtBr were added. 1 µL of dye was added to 5 µL of PCR products. The mix is load in the gel and the electrophoresis was run for 20 min with a 110 mV tension. Results were revealed under UV light.

C.7 RNA-seq expression experiments

C.7.1 Terpene and Phenolic treatments

Mycelia generated for RNA-seq experiments were started from a suspension of 5x10^5 spores on media prepared from 0.17% YNB, 1.5% granulated agar, 1% maltose, 0.1% PHP, 0.3% asparagine overlayed with cellphane. Cultures were incubated 3 days under ambient conditions. Both lodgepole pine methanol extract (LPPE) and terpene treatments were applied with a TLC sprayer using 0.22 µm filtered nitrogen gas as a carrier. LPPE treatment was applied as 300 µl of the 50:50, MeOH:H2O LPPE extract sprayed directly onto the culture surface. Crude extract was prepared as follows: a lodgepole pine bolt from a freshly cut tree was brought to the lab and froze to -20˚C. While still frozen, the bolt was cut into disks and the phloem separated on dry ice blocks. The phloem was then ground in a mill with liquid nitrogen. The ground phloem was then extracted in 80:20 methanol:water (2.5 mL/g) and sonicated at 4 °C for 2 hr. After centrifugation, the supernatant was removed and concentrated by 1/3 under a gentle flow of nitrogen gas. This concentrated crude extract was stored at –20°C for no longer than 8 months prior to use. An LC-MS profile for the crude extract is presented in Figure C4. The terpene treatment was carried out as above with the exception that 200 µL was applied. The terpene blend was prepared as follows: Monoterpenes, (+/-)α-pinene, (-)β-pinene, 3-carene, β-phellandrene, (+/-) limonene, α-terpinolene and γ-terpinene, were blended and aliquoted into a chemically resistant falcon tube. Diterpenes are then dissolved directly, abietic (Sigma, Oakville, ON), dehydroabietic, levopimaric, isopimaric and pimaric acids (Orchid-Helix Biotech, Vancouver, BC) in a 2:1:1:1:1 ratio. The relative composition for the selected metabolites was similar to that described by Shrimpton (1973) except limonene levels are disproportionately high (30x). The blended composition was verified by sampling the headspace directly using Solid Phase Micro Extraction (SPME) and gas chromatography (Figure C4). β-phellandrene was donated by Millenium Chemicals, (Jacksonville, FL) mixed equally with (+/-) limonene. Additional monoterpenes were purchased from Sigma (Oakville, ON).

Figure C4. Treatment profiles. **A.** lodgepole pine phloem extract (LPPE) **B.** volatile (i.e. monoterpene; see methods) portion of the terpene blend.

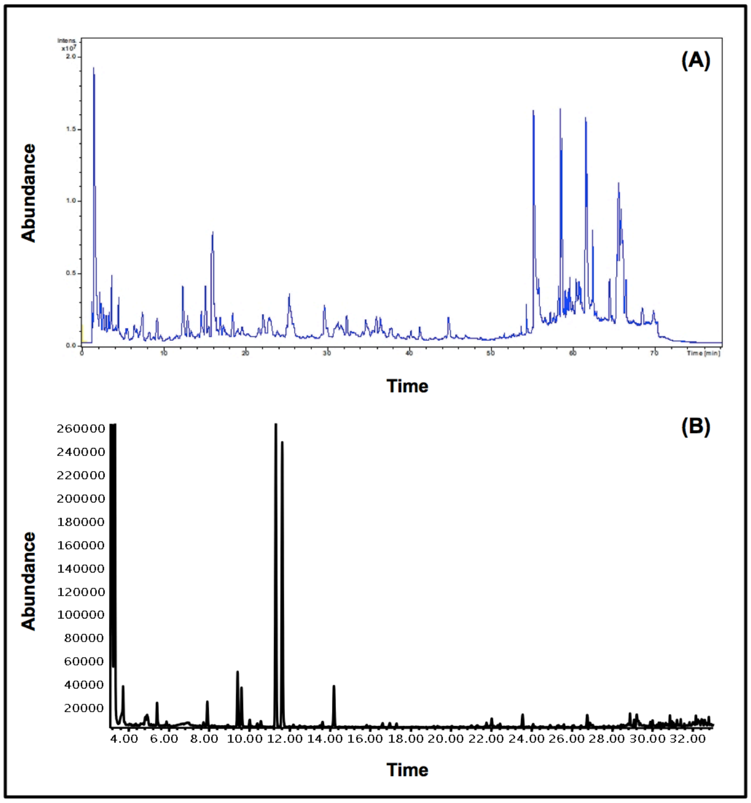


Figure C5. (next page) A genome wide exploration of the RNA-seq expression data. Starting from the center: A plot of the minor allele frequency (MAF; 75-kb window). This track is calculated by taking the average MAF for each gene adding them up within a given 75-kb window and then dividing by the length of the window. On the second track (purple/orange; 50-kb window), dN and dS have been plotted. dS is in orange and is plotted with orientation towards the outside of the circle. dN is in purple and is plotted oriented towards the inside of the circle. On the third track the expression data for the lodgepole pine phloem sample has been plotted (green/yellow). In yellow oriented towards the inside of the circle is the 12 hr sample. Oriented towards the outside of the circle in green is the 36 hr sample. In the fourth track is the expression data for the terpene samples (red/blue). Oriented towards the inside in blue is the 12 hr sample. Oriented towards the outside in red is the 36 hr sample. Around the perimeter linked to the ideogram (black) by red connectors is a subset of the induced gene annotations.

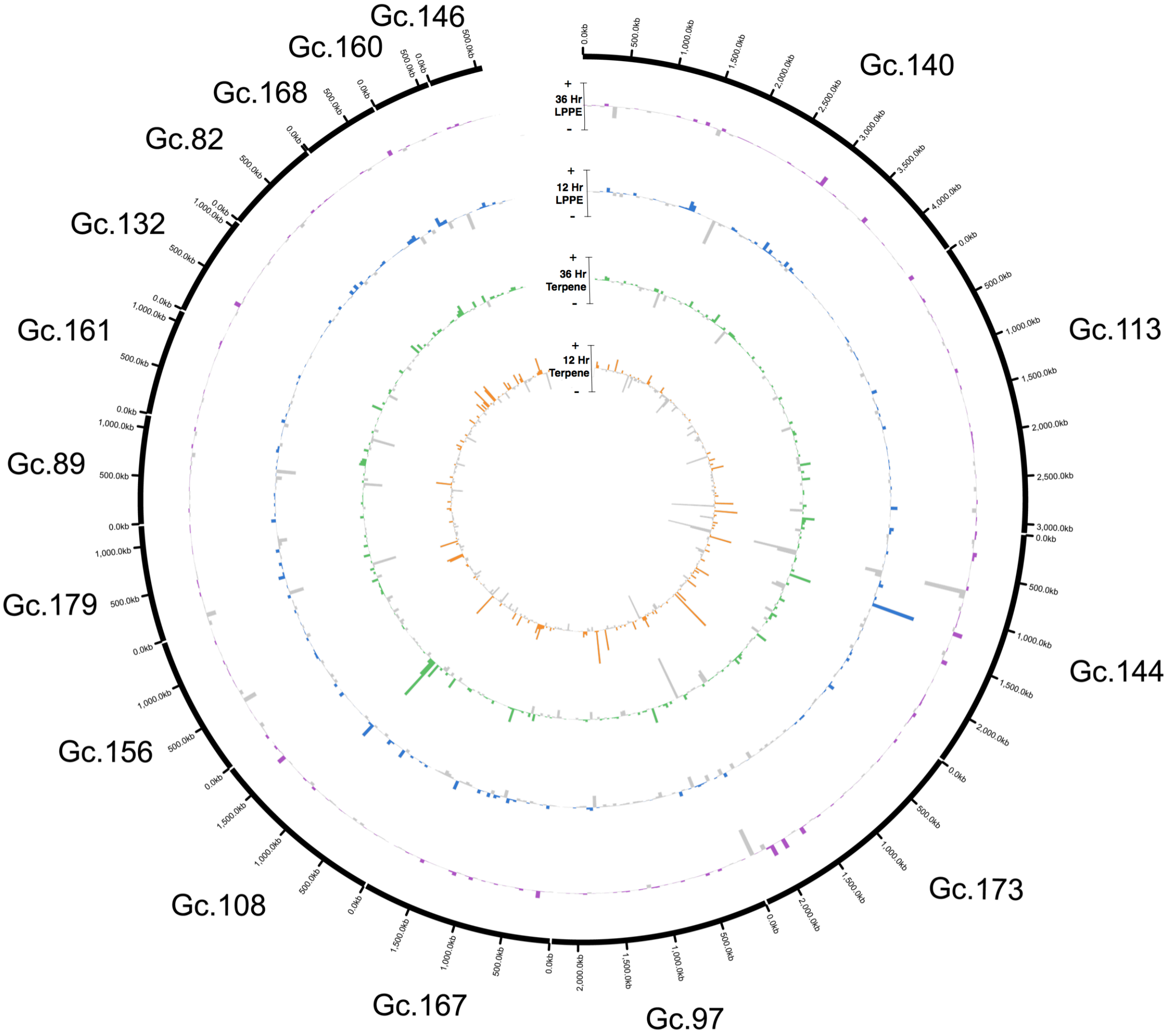
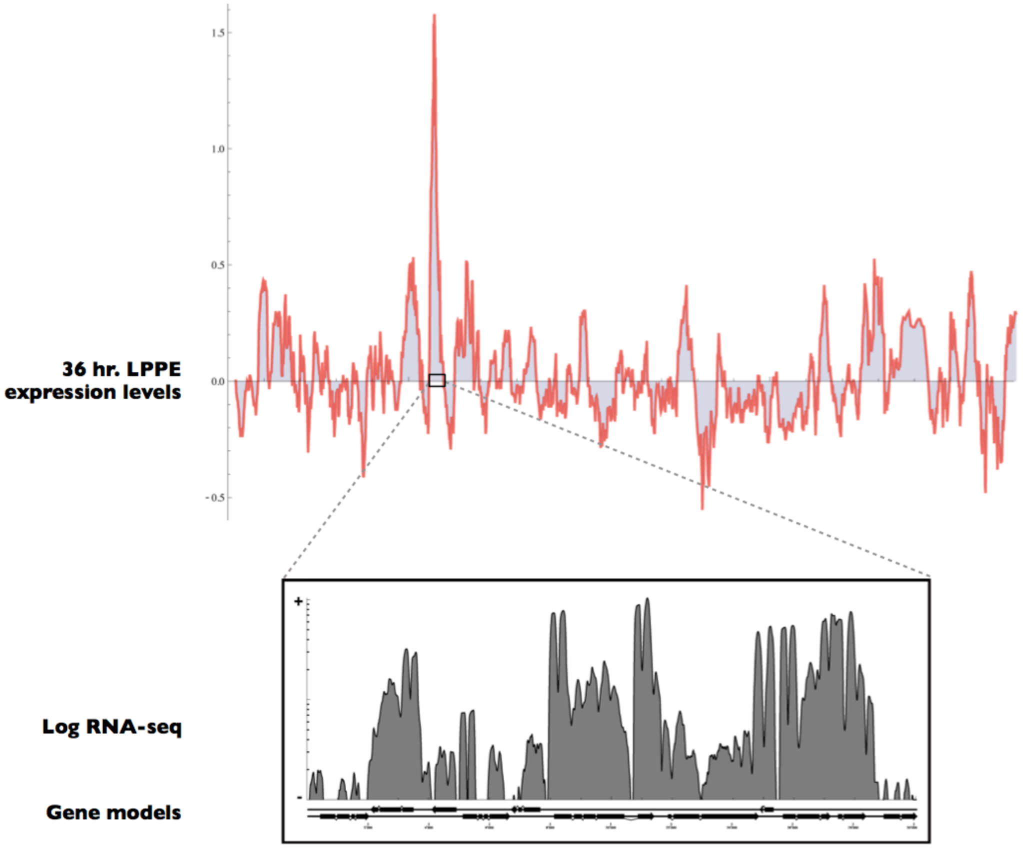

Gc.146
Gc.160
Gc.168
Gc.82
Gc.132
Gc.161
Gc.89
Gc.179
Gc.156
Gc.108
Gc.167
Gc.97
Gc.173
Gc.144
Gc.113
Gc.140
+
36 Hr
LPPE
-
+
12 Hr
LPPE
-
+
36 Hr
Terpene
-
+
12 Hr
Terpene
-

Figure C6. RNA-seq profiling reveals a cluster of co-expressed genes on supercontig GCSC_140. From top to bottom: Expression analysis derived from comparison of control vs. treatment for the 36 hr. LPPE-treated experiment averaged over 50 kb windows across GCSC_140. **Enlargement.** Log-transformed tag-to-genome alignment coverage for the induced co-expression region.



C.8 Gene deletion and fungal growth conditions

C.8.1 Generating the deletion mutant

ABC transporter mutants were generated using an *Agrobacterium*-mediated gene disruption procedure (Wang, DiGuistini, Wang, Bohlmann, & Breuil, 2010). The entire gene open reading frame (ORF) was replaced with the selective marker gene, hygromycin. Complete replacements were confirmed by PCR and southern blot. *Gc* growth assays and routine culturing were on 1% Malt Extract Agar (MEA) media.

C.8.2 Culturing media

Terpene utilization experiments were conducted on Yeast Nitrogen Base (YNB) minimal medium amended with mannose or terpenes (methods for preparing the terpene blend are reported above) as the single carbon source.

Figure C7. Phenotypes observed for the growth *Gc* ABC knockout (7L3) on minimal media with or without a carbon source. The terpene blend and maltose were tested as carbon sources. From left to right **A.** minimal media without a carbon source, **B.** minimal media amended with the terpene blend, and **C.** minimal media amended with mannose. **D.** close-up of the 7L3 minimal media plate amended with the terpene blend showing the complete absence of growth.

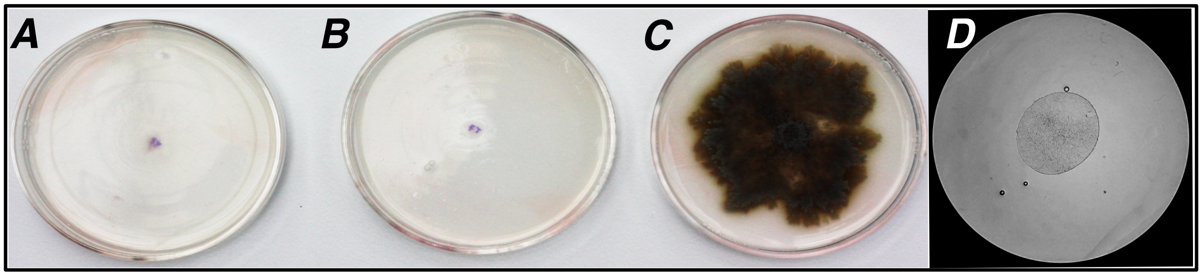


Table C1. A selection of *Gc* genes that encode transcription factors and oxidoreductases induced 36 hr following treatment with LPPE.

Seq. Name	Seq. Description	InterProScan	36 hr difference	P-value
Transcription factors				
GLEAN_3369	mads box transcription factor mcm1	IPR002100	1.46E-04	1.14E-06
GLEAN_7165	bzip transcription factor	IPR011616	8.12E-05	8.57E-06
GLEAN_18	c2h2 transcription factor		9.87E-05	3.57E-05
GLEAN_2085	bzip transcription factor	IPR011616	8.89E-05	4.91E-05
GLEAN_6345	zinc finger transcription factor ace1	IPR007087	1.19E-04	1.47E-04
GLEAN_3582	c2h2 transcription factor	IPR007087	8.01E-05	1.77E-04
GLEAN_2846	bzip transcription factor	IPR011616	4.21E-05	8.13E-04
GLEAN_5975	hlh transcription factor	IPR001092	5.17E-05	8.89E-04
GLEAN_2647	bZIP transcription factor	IPR011616	4.45E-05	3.22E-03
GLEAN_1777	bzip transcription factor (ap-1)	IPR011616; IPR013910	5.63E-05	3.30E-03
GLEAN_124	bzip transcription factor		3.96E-05	4.36E-03
GLEAN_4228	transcription factor	IPR007219	2.49E-05	6.57E-03
GLEAN_3927	transcription factor	IPR000818	3.24E-05	8.91E-03
GLEAN_7264	c2h2 transcription factor	IPR007087	5.41E-05	0.01
GLEAN_6400	bzip transcription factor	IPR011616	4.50E-05	0.01
GLEAN_7988	c2h2 transcription factor	IPR007087	3.83E-05	0.01
Oxidoreductases				
GLEAN_5640	aldehyde dehydrogenase	IPR015590	2.87E-04	0.00E+00
GLEAN_2441	FAD binding domain protein	IPR006076; IPR013096	1.16E-04	1.33E-15
GLEAN_5485	cytochrome p450 monooxygenase	IPR001128	4.25E-04	1.33E-15
GLEAN_684	lignostilbene dioxygenase	IPR004294	2.15E-04	2.22E-15
GLEAN_8113	pisatin demethylase	IPR001128	8.54E-05	5.77E-15

GLEAN_838	nadph cytochrome p450 reductase	IPR001433;
		IPR003097;
		IPR008254
			5.56E-04	9.44E-15

GLEAN_1289	short chain dehydrogenase reductase	IPR002198	1.03E-03	2.26E-14
GLEAN_7953	cytochrome p450 monooxygenase	IPR001128	4.09E-05	1.12E-09
GLEAN_2444	aromatic ring-opening dioxygenase	IPR004183	7.74E-05	1.04E-07
GLEAN_1287	aldehyde dehydrogenase	IPR015590	2.88E-05	2.62E-07
GLEAN_3974	dioxygenase	IPR004360	4.30E-05	5.78E-07
GLEAN_7183	cytochrome p450 monooxygenase	IPR001128	2.46E-05	4.40E-06
GLEAN_7015	alcohol dehydrogenase	IPR013154	5.53E-05	7.06E-06
GLEAN_8277	aryl-alcohol dehydrogenase	IPR001395	4.29E-05	3.48E-05
GLEAN_5484	steroid monooxygenase	IPR013027	3.10E-05	3.65E-05
GLEAN_2484	alpha-ketoglutarate dependent xanthine dioxygenase	IPR003819	3.48E-05	7.56E-05
GLEAN_2304	benzoate 4-monooxygenase cytochrome p450	IPR001128	3.11E-05	1.42E-04
GLEAN_4309	alpha-ketoglutarate-dependent taurine dioxygenase	IPR003819	3.06E-05	1.96E-03
GLEAN_7960	2og-fe oxygenase family protein	IPR005123	1.45E-05	3.75E-03
GLEAN_4251	cytochrome p450 phenylacetate 2-monooxygenase	IPR001128	1.31E-05	4.00E-03

C.9 References

Britton, T., Anderson, C.L., Jacquet, D., Lundqvist, S., & Bremer, K. (2007). Estimating divergence times in large phylogenetic trees. *Systematic Biology 56*(5), 741-752.

Castresana, J. (2007). Topological variation in single-gene phylogenetic trees. *Genome Biol, 8*(6), 216.

Craig, R., & Beavis, R.C. (2003). A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun Mass Spectrom*, 17(20): 2310-2316.

Darling, A. C. E., Mau, B., Blattner, F. R., & Perna, N. T. (2004). Mauve: multiple alignment of conserved genomic sequence with rearrangements. *Genome Res, 14*(7), 1394-1403.

DiGuistini, S., Liao, N., Platt, D., Robertson, G., Seidel, M., Chan, S., et al. (2009). De novo genome sequence assembly of a filamentous fungus using Sanger, 454 and Illumina sequence data. *Genome Biol, 10*(9), R94.

Elsik, C., Mackey, A., Reese, J., Milshina, N.V., Roos D. S., & Weinstock G.M. (2007). Creating a honey bee consensus gene set. *Genome Biol*, *8*(1):R13

Eng, J.K., McCormack, A.L., & Yates, J.R. (1994). An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom*, *5*(11): 976–989.

Felsenstein J. (1989). PHYLIP-phylogeny inference package. Version 3.5. *Cladistics 39*, 783-791

Huelsenbeck, J.P., Ronquist, F. (2001). MrBayes: Bayesian inference of phylogeny. *Bioinformatics 17*, 754-755.

Katoh, K. & Toh, H. (2008). Recent developments in the MAFFT multiple sequence alignment program. *Briefings in Bioinformatics, 9*(4), 286-298.

Keller, A., Nesvizhskii, A.I., Kolker, E., & Aebersold, R. (2002). Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search. *Anal Chem*, *74*(20): 5383-5392.

Kent, W. J. (2002). BLAT--the BLAST-like alignment tool. *Genome Res, 12*(4), 656-664.

Kohany, O., Gentles, A. J., Hankus, L., & Jurka, J. (2006). Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. *BMC Bioinformatics, 7*, 474.

Korf, I. (2004). Gene finding in novel genomes. *BMC Bioinformatics, 5*, 59.

Korf, I., Flicek, P., Duan, D., & Brent, M. R. (2001). Integrating genomic homology into gene structure prediction. *Bioinformatics, 17 Suppl 1*, S140-148.

Li, L., Stoeckert Jr, C., & Roos, D. (2003). OrthoMCL: Identification of ortholog groups for eukaryotic genomes. *Genome Res, 13*, 2178-2189

Möller, E. M., Bahnweg, G., Sandermann, H., & Geiger, H. H. (1992). A simple and efficient protocol for isolation of high molecular weight DNA from filamentous fungi, fruit bodies, and infected plant tissues. *Nucleic Acids Research, 20*(22), 6115-6116.

Mulder, NJ., Apweiler, R., Attwood, T.K., Bairoch, A., Bateman, A., Binns, D., et al., (2007). New developments in the InterPro database. *Nucleic Acids Res. 35* (Database Issue): D224-228.

Nesvizhskii, A.I., Keller, A., Kolker, E., & Aebersold, R. (2003). A statistical model for identifying proteins by tandem mass spectrometry. *Anal Chem*, *75*(17): 4646-4658.

Perkins, D.N., Pappin, D.J., Creasy, D.M., & Cottrell, J.S. 1999. Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis*, *20*(18): 3551-3567.

Price, A. L., Jones, N. C., & Pevzner, P. A. (2005). De novo identification of repeat families in large genomes. *Bioinformatics, 21 Suppl 1*, i351-358.

Rice, P., Longden, I., & Bleasby, A. (2002). EMBOSS: The european molecular biology open software suite. *Trends Genet*, *16*(6): 276-277.

Shrimpton, D. (1973). Extractives associated with wound response of lodgepole pine attacked by the mountain pine beetle and associated microorgansims. *Can. J. Bot., 51*, 527-534.

Simpson, J., Wong, K., Jackman, S., & Schein, J. (2009). ABySS: A parallel assembler for short read sequence data. *Genome Res*. 10.1101/gr.089532.108

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics, 6*, 31.

Stamatakis, A. (2006). RAxML-VI-HPC: Maximumlikelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics 22*, 2688-2690.

Stanke, M., & Waack, S. (2003). Gene prediction with a hidden Markov model and a new intron submodel. *Bioinformatics, 19 Suppl 2*, 215-225.

Suyama, M., Torrents, D., & Bork, P. (2006). PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res, 34* (Web Server issue), W609-612.