DISCOVERY OF CANDIDATE EFFECTORS INVOLVED IN *Cronartium quercuum* f. sp. *fusiforme* INFECTION OF PINE AND OAK

By

KATHERINE E. SMITH

A THESIS PRESENTED TO THE GRADUATE SCHOOL
OF THE UNIVERSITY OF FLORIDA IN PARTIAL FULFILLMENT
OF THE REQUIREMENTS FOR THE DEGREE OF
MASTER OF SCIENCE

UNIVERSITY OF FLORIDA

2012

© 2012 Katherine E. Smith

For the late bloomers

ACKNOWLEDGMENTS

I would like to thank the Plant Molecular and Cellular Biology Program for having the flexibility to allow me to complete this degree on a part-time basis. I also would like to thank my committee members John Davis, Jeff Rollins and Jason Smith for their time and expert advice. I especially want to thank my colleagues Chris Dervinis and Alison Morse for their help, advice and encouragement during my tenure in the Forest Genomics Laboratory.

TABLE OF CONTENTS

page

LIST OF TABLES

LIST OF FIGURES

Abstract of Thesis Presented to the Graduate School
of the University of Florida in Partial Fulfillment of the
Requirements for the Degree of Master of Science

DISCOVERY OF CANDIDATE EFFECTORS INVOLVED IN *Cronartium quercuum* f. sp. *fusiforme* INFECTION OF PINE AND OAK

By

Katherine E. Smith

August 2012

Chair: John Davis
Major: Plant Molecular and Cellular Biology

Cronartium quercuum (Berk.) Miyabe ex Shirai f. sp. *fusiforme* (*Burds. & Snow*) – *Cqf* – is a biotrophic rust fungus that infects both pine and oak trees. It is the causative agent of fusiform rust which is the major disease of southern pine forests. For decades, the large economic impact of fusiform rust has motivated a large volume of research on managing the disease. Since the southern pine host is economically important to the sawtimber and pulpwood industries, the primary research focus has been on the pine host. To date, the major research advances have been the identification of pine families with genetic resistance and the discovery of specific interactions between the pine host and the pathogen.

Pathogen effectors are a diverse set of secreted proteins that enhance virulence, often by mechanistically thwarting host defenses. Fungal effectors that have been characterized in other pathosystems are typically small secreted proteins, localized to either the host cytoplasm or apoplast. Identification of the genes encoding secreted proteins provides insight into the pathogenicity of this fungus, as well as a source of candidate fungal effectors. Avirulence genes encode a subset of effectors that are

recognized by the host, which in turn allows the host to block disease development. Candidate effectors identified through the annotation of genome sequence can be a valuable complement to marker based avirulence gene identification strategies. A *Cqf* sequencing project has been undertaken at the United States Department of Energy (DOE), Joint Genome Institute (JGI), which provided a project midpoint draft, as well as a second and final,genomic sequence assembly enabling the annotation of a complete set of genes. This project contributed to the annotation and used it as a tool to identify secreted proteins and measure fungal transcriptome level gene expression in both hosts. Major contributions of this work include an *in silico* predicted secretome, a list of host-specific candidate effectors and a physical map of the *AVIRULENT TO FUSIFORM RUST 1* (*AVR1*) locus that contains avirulence gene candidates.

CHAPTER 1
INTRODUCTION AND LITERATURE REVIEW

Biotrophy and Fungal Effectors in Plants

Biotrophic fungi can persist inside their hosts for long periods of time causing large scale changes to host cell morphology, as is the case for *Cronartium quercuum* (Berk.) Miyabe ex Shirai f. sp*. fusiforme* (*Burds. & Snow*), referred to as *Cqf*. Unlike necrotrophs that kill their hosts and feed on dying or dead tissue, biotrophs have evolved mechanisms for evading host defenses and feeding off living cells. In order to obtain the nutrients necessary for proliferation within the host, rust fungi such as *Cqf* form host cell penetration structures called appressoria (Gray et al., 1982) that apply pressure to the host cell wall allowing specialized infection structures called haustoria to grow into the host cell without actually making contact with host cytoplasm (Figure 1-1). Discovery of expressed genes specific to haustoria have shed light on the acquisition of metabolites during the biotrophic interaction. Some of the planta-induced genes (PIGS) of the bean rust fungus *Uromyces fabae* include amino acid and hexose transporters, as well as ATPases (Hahn and Mendgen, 1997), which are suspected to provide the energy necessary for haustoria-specific amino acid permeases to transport metabolites into haustoria cells (Hahn and Mendgen, 2001). Other PIGS are involved in general defense, such as a metallothionein that may provide protection from oxidative stress and a cytochrome P-450 monooxygenase that may detoxify damaging plant compounds.

Fungi avoid host defenses and establish disease through the action of an array of secreted proteins referred to as effectors. These proteins interact with the host to alter host cell morphology and function (Gan et al., 2010). The *Cqf*-pine host interaction is an

obvious example since not only must the fungus suppress host defenses to grow within pine stems, but also host cells must change morphology to form galls. Some fungal effectors function outside of host cells, in the apoplastic space between plant cells or in the extrahaustorial matrix (Figure 1-1). For example, some effectors mechanistically block host defenses by inhibiting host chitinases and proteases. For example, *Cladosporium fulvum*, a biotrophic fungus that causes tomato leaf mold, secretes a chitinase inhibitor (Avr4) and a plant cysteine protease inhibitor (Avr2) (van Esse, 2008) during infection. Avr4 binds chitin in the fungal cell wall and apparently protects the wall from degradation by tomato chitinases (van den Burg et al., 2006). AvrP123, from the flax rust fungus *Melampsora lini*, is related to Kazal serine protease inhibitors (Catanzariti, 2006). In addition, there are many more fungal effectors that function inside host cells, sometimes entering the host nucleus and likely affecting gene expression. For example, the *Uromyces fabae* haustoria specific protein Rust Transferred Protein 1 was shown by immunofluorescence to be localized to the extrahaustorial matrix (Figure 1-1) and the interior of plant cells, including the nucleus (Kemen et al., 2005).

A subset of effectors conform to predictions made by the gene for gene hypothesis, meaning they are avirulence genes with alleles that specifically interact with host resistance gene alleles resulting in a resistance phenotype. This phenomenon was first reported by Flor in *Melampsora lini* (Flor, 1955), and subsequent genetic studies have revealed at least 30 *AVR* genes with corresponding resistance genes in flax (Ellis et al., 2007). Coevolution of rust fungi with their hosts has led to the proliferation of pathotype-specific avirulence genes and interacting host resistance genes that has been described as an “arms race.” Cloned avirulence genes of *Melampsora lini* are

contained in four loci, each with *AVR* genes similar to each other but with no known homologs in other species. Most known avirulence effectors, including those from *Melampsora lini*, enter host cells and interact directly with resistance gene products. Effector proteins encoded by genes on the *AvrL567* locus have variants that differ in amino acids on the protein surface and have been shown in yeast two hybrid studies to directly interact with resistance proteins (Stergiopoulos and de Wit, 2009).

Exactly how secreted effectors get into host cells remains unclear. In rust fungi, haustoria penetrate into the host cell such that the fungal cell wall does not make contact with host cytoplasm, but is separated from it by an extrahaustorial matrix surrounded by an extrahaustorial membrane (Figure 1-1). Secreted proteins containing an N-terminal secretion signal are delivered through the haustorial cell wall into the extrahaustorial matrix via the eukaryotic secretory system. They must then enter the plant cell by either a plant derived mechanism or a fungal derived mechanism. There is evidence in fungal-like oomycota for a plant derived system since the conserved effector domain, RXLR, is both necessary and sufficient for oomycte effectors to enter host cells and this domain is present in plant proteins involved in membrane trafficking (Panstruga and Dodds, 2009). However, biotrophic fungal effectors have no such conserved motif and most have little similarity to known proteins (Gan et al., 2010).

Cronartium quercuum* f. sp. *fusiforme

Heteroecious Lifecycle

Cronartium quercuum, defined as the rust fungus that causes galls on pine and colonizes both pine and oak as hosts, has been proposed to be divided into special forms (*formae speciales,* f.sp.) so that f.sp. *fusiforme* (loblolly, longleaf and slash pine), f.sp. *banksianae* (jack pine), f.sp. *echinatae* (shortleaf pine) and f.sp. *virginianae*

(Virginia pine) are each distinguished by the pine host (Burdsall and Snow, 1977). Phylogenetic analysis of pine stem rusts from the genera *Cronartium* and *Peridermium* based on nucleotide similarity of the internal transcribed region of nuclear ribosomal genes placed *Cqf* nearest to the other *Cronartium quercuum* special forms and *Endocronartium harknessii* (Vogler and Bruns, 1998). *E. harknessii* is a species that causes galls on pine, however, it is autoecious, meaning all stages of its lifecycle occur on pine.

Figure 1-2 illustrates the complex *Cqf* lifecycle that requires two hosts, oak and pine, includes five different spore stages and takes up to two years to complete (Phelps, 1978). Timing of the various stages is influenced by weather conditions, such as temperature, wind and rainfall. In the late spring, young pine trees are infected primarily by penetration of germinated basidiospores (Schmidt, 1998) into stems resulting in stem and branch galls within 6-9 months. In the fall, galls produce bright orange pycnial droplets that contain haploid spores. Presumably insects and perhaps other organisms move pycnia within a gall or from one gall to another causing spermatization (Kubisiak et al., 2005), leading to the dikaryotic (N + N) spore type, aeciospores, that are produced by galls in the spring. Aeciospores are bright yellow and coat the outside of the gall. These spores are dispersed by the wind and infect succulent leaves of species in the group red oak, primarily through stomata. Oaks are minimally affected by the fungus (Schmidt, 1998). Infected oak leaves have long cylindrical teliospores that germinate and produce basidiospores (Phelps, 1978). Basidiospores carried by the wind then infect young pines and the cycle begins again. Meiosis occurs in the telial columns (2N) while on the oak host giving rise to haploid basidiospores (N) that are

each the result of different recombination events of the original N + N aeciospore genotype that infected the oak. Oak leaf infection by aeciospores may also give rise to pustules that contain urediniospores, the dikaryotic repeating stage on oak. There is no repeating stage on pine, since only basidiospores can infect pine. High genetic diversity of the fungal population is thought to be maintained at least in part by yearly genetic recombination on oaks.

Fusiform Rust Disease

Fusiform rust disease caused by the fungus *Cronartium quercuum* f. sp. *fusiforme*, occurs solely on the North American continent and primarily in the southern United States, from northern Florida, west to the eastern edge of Texas and Oklahoma and north as far as Maryland. *Cqf* affects primarily the southern pine species, *Pinus elliottii* (slash pine) and *Pinus taeda* (loblolly pine), although *Pinus palustris* (longleaf pine) and *Pinus serotina* (pond pine) show moderate susceptibility (Dwinell, 1976). It has been estimated that over 35 million dollars are lost to fusiform rust disease annually (Anderson et al., 1986). Figure 1-3 and Figure 1-4 show the widespread incidence of fusiform rust, measured by the USDA Forest Service, across the entire natural range of slash and loblolly pine, respectively (Starkey, 1997). The warm and moist climate of the southeastern United States is ideal for *Cqf* spore production and germination. The major symptom is gall formation on the stems and branches of young, actively growing pines. These galls damage and deform trees often leading to death, but galls can also persist for many years. Galls that persist put trees in risk of breakage and wind damage (Phelps, 1978).

The management of fusiform rust disease has included two general strategies, amending silviculture practices and planting rust resistant pine families. Silviculture

strategies have included removing infected seedlings from nurseries, removing infected stems from young pines, as well as efforts to reduce the number of oaks at a field site. Perhaps the more effective and sustainable control will be planting resistant pines. At first, resistant trees came from field trials or were survivors from heavily rust infected areas. In 1973, the USDA Forest Service established the Rust Resistance Screening Center (Asheville, North Carolina) in response to the rust epidemic. There, an artificial and accelerated inoculation technique is used on six-week-old seedlings that allows quick testing of a large number of genotypes. Based on a performance index, greenhouse inoculation correlates well with field trials. The screening center has also aided research on the genetics of pine resistance (Schmidt, 2003).

Studies have shown that the *Cqf*-pine host interaction operates in a gene-for-gene manner (Flor, 1955) and the first resistance gene, Fr1, was mapped in loblolly pine (Wilcox et al., 1996). Recently, the corresponding avirulence gene, *AVR1*, was mapped in the fungus (Kubisiak et al., 2011). This work produced 421 mapped markers throughout the *Cqf* genome, 14 of which are linked to *AVR1* and define a genetic interval for the gene. In addition to Fr1, 8 more resistance genes have been mapped in pine (Amerson et al. in preparation). Taken together, the research surrounding major gene resistance in pine has led to the proposal of a new approach to fusiform rust disease management. Nelson et al. (2010) proposed pine resistance screening combined with monitoring of *AVR* gene allele frequencies in the field by geographic region to guide selection of resistant pine families. The authors presented two methods by which this could be achieved. One is through the identification and use of single genotype isolates of the fungus that are informative when tested on pines that are

heterozygous and homozygous recessive for the corresponding resistance genes. These materials could be used to monitor genes present in *Cqf* samples or to screen pine families for breeding programs. The second method would be to expand the number of markers on the genetic maps of *Cqf* and loblolly pine such that resistance genes and avirulence genes, respectively, could be definitively identified in test subjects (Nelson et al., 2010). This is much easier to achieve in the fungus, compared to the pine host. The huge genome size of loblolly pine, 21,658Mb, compared to ~90Mb in *Cqf*, makes it more difficult to obtain the marker coverage necessary to definitively identify resistance genes. In addition, the *Cqf* reference genome sequence will make identification of avirulence genes and subsequent development of fully diagnostic markers more straight forward.

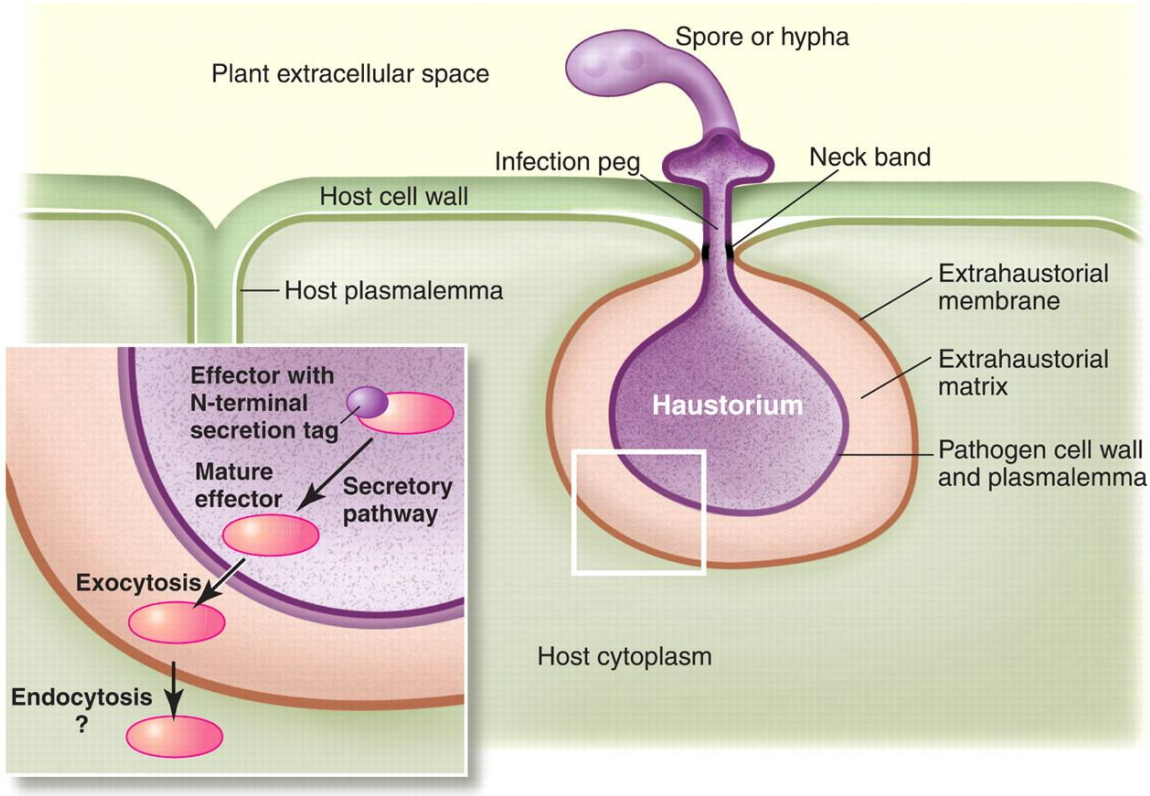

Figure 1-1. Schematic of the haustorium infection structure taken from Panstruga and Dodds, 2009.

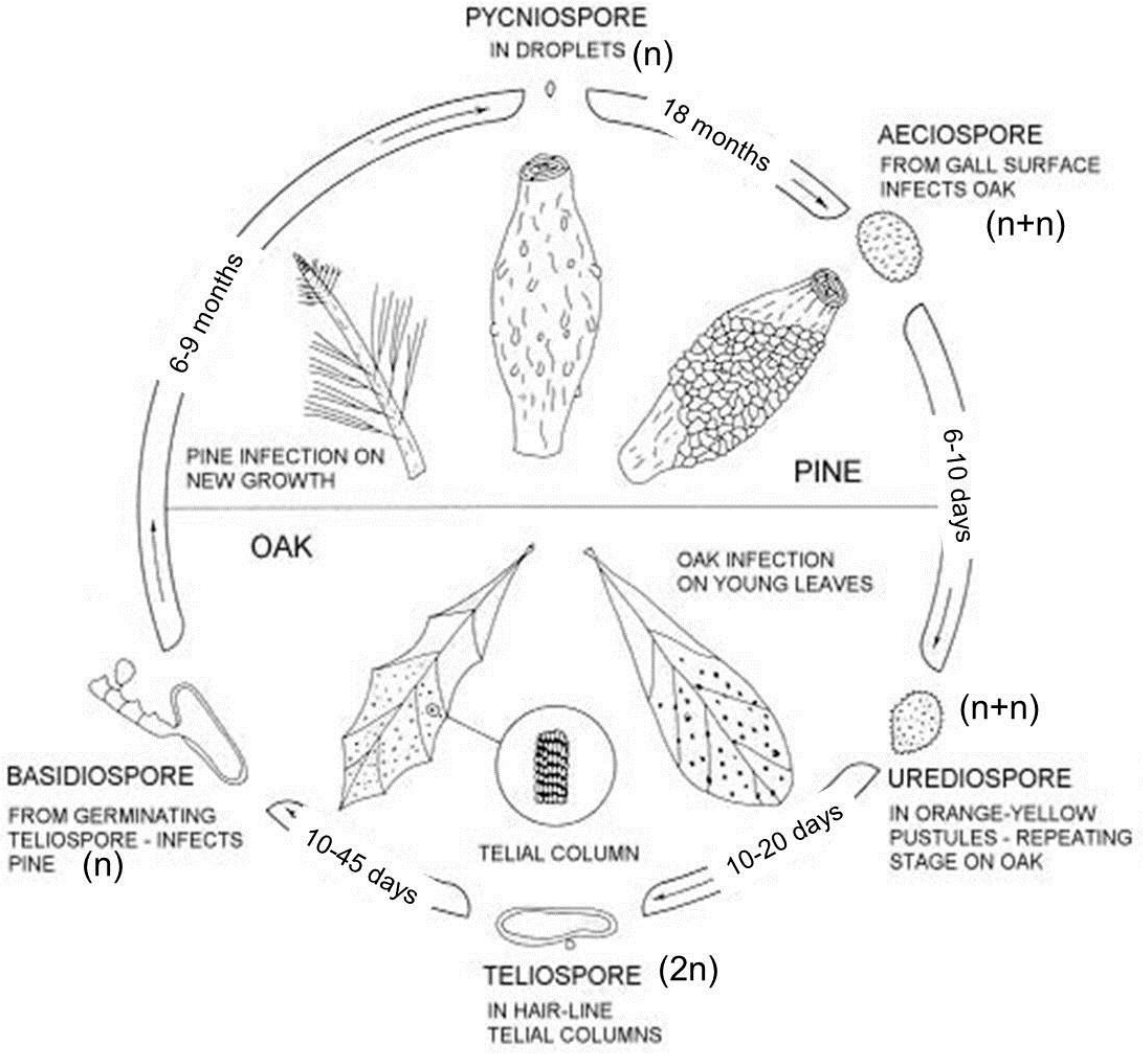

Figure 1-2. The lifecycle of *Cronartium quercuum* modified from Phelps and Czabator 1978.

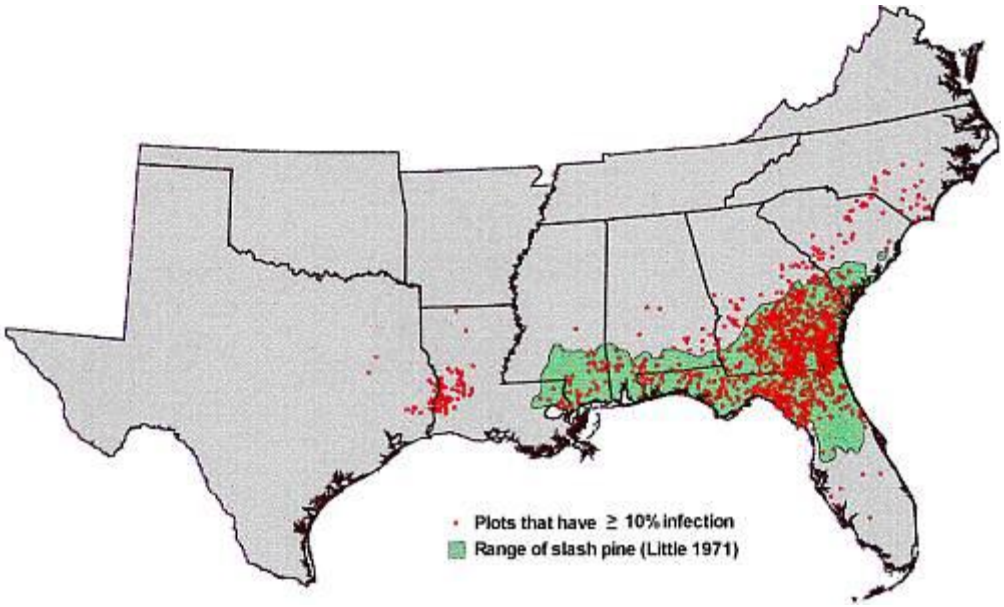

Figure 1-3. Location of slash pine containing plots with ≥10% infection rate from Starkey et al. 1997.

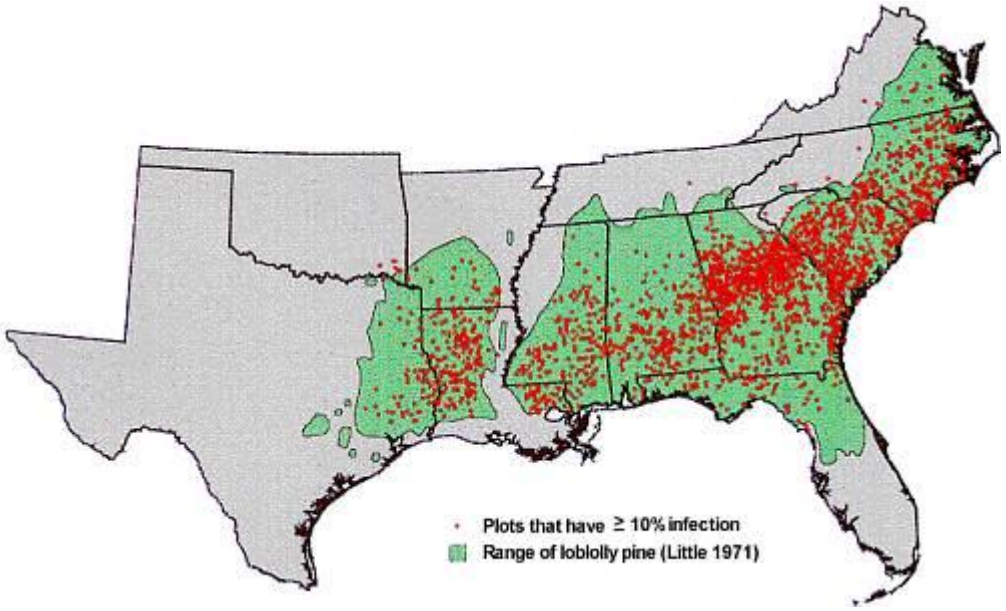

Figure 1-4. Location of loblolly pine containing plots with ≥10% infection rate from Starkey et al. 1997.

CHAPTER 2
RNA SEQUENCING SUPPORTS PREDICTION OF *CQF* GENE MODELS AND A SECRETOME

Background

Researchers from the USDA Forest Service, Southern Institute of Forest Genetics and the University of Florida School of Forest Resources and Conservation, were awarded a sequencing grant by the Joint Genome Institute (JGI) Community Sequencing Program, to sequence the *Cronartium quercuum* f. sp*. fusiforme* genome (http://www.jgi.doe.gov/sequencing/why/cronartium.html). Pine trees are of interest to JGI because they are fast growing and produce large amounts of biomass; therefore, hold promise as a source of biofuel. Fusiform rust greatly diminishes yields of southern pine forests and a complete genome sequence of *Cqf* could provide major insights leading to improved control of the disease. Elucidating genes involved in the pathogenicity of *Cqf* is a first step in the development of new disease control strategies (Nelson et al., 2010).

In July 2011, the project midpoint draft assembly of a genomic reference sequence was completed, which consists of 2084 scaffolds totaling 89.1MB, very close to the ~90Mb genome size predicted using flow cytometry (Anderson et al., 2010). Gene models, including translation start and stop sites, introns, exons and untranslated regions, were annotated on those scaffolds with the support of RNA sequencing data. The JGI *Cqf* sequencing project was supplied with 6 RNA samples from various fungal stages for sequencing. The goal was to provide gene expression data to support *ab initio* gene model prediction with the most complete coverage possible. More stages of the fungus sampled would enable better coverage. Expressed transcript sequences were compared to gene models to delineate introns, exons and translation start and

stop sites. It is likely that an essentially complete set of genes was annotated using the project midpoint assembly. This is because the total scaffold length was similar to the expected genome size and stretches of unknown nucleotides (N's) present in the assembly were considered to be repetitive non-genic sequence.

Given an ever increasing number of sequenced genomes, it is now a common strategy to look for fungal effectors, as well as other proteins important to the disease process, in the secretome of a pathogen. Secretomes are predicted *in silico*, by searching for N-terminal signal peptides in protein sequences derived from gene models. Seventy-nine of the 426 secreted proteins of the biotrophic smut fungus, *Ustilago maydis*, were located in clusters throughout the genome (Kamper et al., 2006). In this paper, information gained from cluster expression patterns during infection and virulence of cluster deletion mutants supported the utility of predicting secreted proteins. The rice blast fungus, *Magnaporthe grisea*, was predicted to have 739 secreted proteins, 8 of which may encode cutinases (Dean et al., 2005). One of those 8 putative cutinases, CUT2, was later shown to be important in host penetration and required for full virulence (Skamnioti, 2007). A greater number of secreted proteins were predicted in the larger rust genomes of *Melampsora larici-populina* and *Puccinia graminis*, 1549 and 1852 respectively. Hierarchical clustering of these proteins identified 8 protein families as likely effectors and good candidates for further functional testing (Saunders et al., 2012). Here, a set of *Cqf* secreted proteins is identified from the project midpoint assembly gene models and described using BLAST (Basic Local Alignment Search Tool) comparison and functional enrichment analysis.

Materials and Methods

Fungal and Plant Material

Samples for RNA sequencing were collected from both the pine and the oak hosts of *Cqf*. Galls were collected from 5 year old slash pine trees in October as pycniospores were forming and again in April as aeciospores were forming. Galls were freeze dried for 1 week. To obtain tissue for RNA extraction, yellow-colored aeciospore (spring) or orange-colored pycniospore (fall) hymenial layers were chipped away from the outside of freeze dried galls using a scalpel. Aeciospores were collected by knocking them off the surface of spring collected pine galls. They were then stored at -20°C. Oak associated tissues included infected oak leaves with attached telial columns, telial columns removed from oak leaves and basidiospores collected onto pH 2.0 water wetted filters (to prevent germination). Oak leaves and telial columns were stored at -20°C and basidiospores were stored in pH 2.0 water, at 4°C for up to 4 days. Pine associated tissues were collected in the field, at the University of Florida in Gainesville, Florida. Oak associated tissues were collected from greenhouse inoculated, open pollinated wild northern red oak (*Quercus rubra*) seedlings at the USDA Forest Service, Resistance Screening Center in Asheville, North Carolina.

RNA Extraction

RNA was extracted from the following tissues: aeciospores, basidiospores, telial columns, infected oak leaves and the spring and fall hymenial layers of pine galls. A previously described cetyltrimethylammonium bromide (CTAB) buffer method (Chang et al., 1993) was used, with different grinding procedures for each tissue. Infected oak leaves were frozen and ground in liquid nitrogen. The following tissues were ground in CTAB buffer pre-warmed to 65°C using a Geno/Grinder 2000 homogenizer (BT&C

Incorporated): 1) Aeciospores were ground in 4ml round bottom vials containing ~20mg of spores and a 1.0cm stainless steel ball; 2) Basidiospores were ground in 1.5ml Eppendorf tubes containing ~20mg of spores, 150mg zircon beads and 12.5mg diatomaceous earth; 3) Telial columns were ground in 1.5ml Eppendorf tubes containing ~50mg of telial columns, 150mg zircon beads, 12.5mg diatomaceous earth and a single 2mm tungsten bead; 4) Chips of both pycniospore and aeciospore hymenial layers were preprocessed in a coffee grinder and then ground in 4ml round bottom vials containing CTAB buffer and a 1.0cm steel ball. In all cases extracted RNA was treated for 30 minutes at 37°C with RQ1 RNase-free DNase (Promega, M6101) and then purified using a Qiagen RNeasy Mini Spin Column.

RNA Sequencing

RNA for library production and sequencing was analyzed on an Agilent 2100 Bioanalyzer and only RNA with a minimum RNA integrity score (RIN) of 6.3 was provided to JGI. Libraries were constructed using RNA from the following five samples: 1) Pycnial hymenial layer; 2) Aecial hymenial layer; 3) Aeciospores; 4) Basidiospores; and 5) Infected oak leaves with telial columns. All libraries were constructed to enable Illumina 75-base pair, paired end high throughput sequencing, which generated a total of ~406 million reads.

Gene Model Prediction

Gene models were predicted as part of the *Cqf* genome sequencing project by the Yandell laboratory (www.yandell-lab.org/) at the University of Utah. The genome annotation pipeline Maker2 was used (Holt and Yandell, 2011). Maker2 was designed to predict gene models from genome sequence data without the benefit of preexisting known gene models to use as training data for gene finding programs. The pipeline

uses SNAP, Augustus and GeneMark to produce gene models, then incorporates RNA sequence data and BLAST data to refine gene models.

Functional Annotation

The publicly available genomic research tool, BLAST2GO, was used to functionally annotate genes in the *Cqf* genome (Conesa et al., 2005). BLAST2GO performed a BLAST similarity search on each protein sequence with a significance level cutoff of E^{-6}. BLAST2GO compiles candidate GO (Gene Ontology, http://www.geneontology.org) terms associated with gene identifiers (GI) of the hits, along with the accompanying evidence codes (EC). Annotation assignments are made by applying an annotation rule that takes into account sequence similarity and node relatedness among the GO term candidates, as well as experimental evidence. Experimental evidence is weighted more highly than electronic evidence in the assignment of GO terms. Enrichment for specific GO terms in the *Cqf* secretome was tested using a module within BLAST2GO that integrates the protein structure comparison tool GOSSIP (Global Structural Superposition of Proteins), to compute Fisher's Exact Test with a default false discovery rate (FDR) of 0.05.

Bioinformatic Secretome

Proteins were designated as secreted using previously published methods (Joly et al., 2010). Signal peptides were identified using both TargetP 1.1 (www.cbs.dtu.dk/services/TargetP/) and SignalP 3.0 (www.cbs.dtu.dk/services/SignalP-3.0) online software. Proteins with transmembrane domains were removed using the online software TMHMM 2.0 (www.cbs.dtu.dk/services/TMHMM/). To account for the fact that TMHMM 2.0 may have difficulty distinguishing transmembrane domains from signal peptides (Krogh et al., 2001), proteins with a signal peptide predicted by TargetP

and SignalP and a single transmembrane domain occurring within 40 amino acids of the N terminus were designated as secreted (Joly et al., 2010). The Joly et al. method was validated using a second method developed by Min (2010). This method combined SignalP 4.0 (www.cbs.dtu.dk/services/SignalP), WoLF PSORT (http:/wolfpsort.org) and Phobius (http:/phobius.sbc.su.se), online software, to locate signal peptides. TMHMM 2.0 was used to remove transmembrane proteins and ScanProsite (http:/prosite.expasy.org/scanprosite) was used to remove proteins targeted to the endoplasmic reticulum (Min, 2010).

Results

Cqf Gene Models

The *Cqf* sequencing project generated 8782 total *in silico* gene models. Nearly all of these (8161) were present in the RNA sequencing libraries and therefore had evidence of expression as well as support for the determination of a coding region**.** The 621 gene models that were not present in the RNA sequence libraries were designated as predicted by *ab initio* evidence only. One hundred and fifty nine of the gene models in this category encoded proteins with homology to proteins in the NCBI (National Center for Biotechnology Information) database (Figure 2-1). Approximately 75% had no hits in the database and five with no hits could be assigned GO terms.

Proteins encoded in all predicted gene models were compared to the NCBI protein database by BLASTP. Top-hits broken down by species reflect that the two fungi that are most closely related to *Cqf* and have sequenced genomes are *Melampsora larici-populina* and *Puccinia graminis* (Figure 2-2). Both of these fungi are rusts and *Melampsora larici-populina* alternates infection between *Populus*, an angiosperm tree species like *Quercus*, and *Larix*, a gymnosperm tree species like *Pinus*. Roughly half of

all *Cqf* proteins were similar to proteins of known function, while the other half either had no similarity at all (~25%) or similarity to proteins present in other organisms but with no known function (~25%) (Figure 2-3). BLAST protein sequence comparison and GO term assignment using the BLAST2GO platform resulted in the functional annotation of 3841genes or 47% of genes in the *Cqf* genome.

Cqf Secretome

The 804 member *Cqf* secretome was determined by a bioinformatic prediction method used by Joly et al. to identify secreted proteins in several closely related *Melampsora* leaf rust species (Joly et al., 2010). A second validation method was also applied. This method was developed in an effort to predict secretion while minimizing false positives and false negatives, by using a manually curated set of known fungal proteins (Min, 2010). A Venn diagram comparison shows the methods to have similar results and a large degree of overlap, with the validation method predicting more secreted proteins, 1053 in total (Figure 2-4). The validation method is less restrictive and therefore likely to contain more false positives.

The *Cqf* secreted proteins were characterized by sequence comparison and by function enrichment testing based on GO terms. BLASTP results for the secretome compared to the entire proteome show an increase in the proportion of proteins that share similarity to unknown proteins or share no similarity to any proteins. Fifty percent of small secreted peptides (SSPs, less than 300 amino acids) have no hits in the NCBI protein database (Figure 2-5). Secreted proteins with known functions were heavily enriched for GO terms associated with carbohydrate metabolism, lipases and proteases (Table 2-1).

Discussion

The RNA sequencing data provided good support for defining gene models. Overall 93% of gene models were expressed in the six fungal stages sampled. *Melampsora larici-populina* and *Puccinia graminis*, two closely related and sequenced rust fungi accounted for about 71% of the top hits by species (Figure 2-3). The total number of *Cqf* gene models, 8782, is curiously less than the number reported for *Melampsora larici-populina* and *Puccinia graminis*, 16,399 and 17,773 respectively (Duplessis et al., 2011b). However, the proportion of *Cqf* proteins predicted to be secreted, 9.2%, is similar to *Melampsora larici-populina* and *Puccinia graminis*, which have 9.2% and 10.1% predicted secreted proteins respectively (Saunders et al., 2012). It remains to be seen what accounts for the large difference in number of genes in these similarly-sized rust genomes. The current gene count for *Cqf* is well supported, since Maker2 uses three different *ab initio* gene finding programs and the predicted genes have excellent evidence for expression.

Gene models without RNA sequence support are possibly genuine *Cqf* genes, but expressed in stages of the fungus that were not sampled, for example time points earlier in infection. Once validated, these genes could be involved in early host recognition or act as host range determinates. Among the 75 proteins in this category that are predicted to be secreted and therefore have greater potential as effectors, 69 have no homology to known proteins or are homologous to proteins of unknown function (Figure 2-2). Only six proteins have known function and all six could be involved in infection by one or more of the following modes; cell wall degradation, nutrient acquisition or signaling. These proteins include are 3 glycoside hydrolases from

3 different hydrolase families, 1 subtilisin protease, one triacylglycerol lipase and one serine/threonine kinase.

The predicted secretome of *Cqf* provides a list of potential virulence determinants and potential effectors that can be used to gain insight into the pathogenicity of the fungus. One *Cqf* secreted protein shares 29% identity with the TRI14 protein of the hemibiotroph *Colletotrichum higginsianum*. TRI14 is a virulence determinant in *Fusarium graminearum* involved in the synthesis of the mycotoxin trichothecene (Dyer et al., 2005). Four were similar to rust transferred protein from *Uromyces viciae-fabae*, (Kemen et al., 2005) with amino acid identities between 41% and 43%. Two proteins were similar to two of the haustorially expressed secreted proteins of flax rust (*Melampsora lini*) (Catanzariti, 2006), hesp379 with 60% identity and hesp 735 with 58% identity. In addition, the *Cqf* secretome is enriched for candidate avirulence effectors. These proteins coevolve with hosts and known fungal avirulence effectors commonly have no sequence similarity to proteins in other fungi and are short secreted, cysteine-rich peptides (Stergiopoulos and de Wit, 2009). The *Cqf* secretome includes 514 SSPs (small secreted peptides less than 300 amino acids), 220 of which have no similarity to other proteins (Figure 2-6). Taken as a group the SSPs are more cysteine-rich, with an average percent of cysteine residues for the proteome as whole of 1.43 compared to 1.74 for those proteins in the secretome and 2.24 for the SSPs. The 118 SSPs with greater than 3% cysteine residues, 106 of which have no assignable function, are good candidates for avirulence effectors. Five out of the six secreted proteins that contain a cysteine-rich CFEM domain (CFEM, common in fungal extracellular and membrane) are under 325 amino acids. CFEM domains are suspected

to be involved in pathogenesis as cell-surface receptors, signal transducers or adhesion molecules (Kulkarni et al., 2003). Using an e-value cut of $1E^{-50}$, 16% of *Cqf* genes encoding secreted proteins have paralogs, an indication of diversifying selection. The most paralogous genes are 1 set of five paralogs and 2 sets of four, all of which encode unknown proteins. This secretome data combined with avirulence gene mapping could be useful in the identification of the 9 (at least) expected avirulence genes in *Cqf*-pine interaction.

Proteins potentially important during infection exist among the 210 secreted proteins that could be functionally annotated by the BLAST2GO platform. Using a false discovery rate (FDR) of .001 these proteins were heavily enriched in process and activity GO terms involved in cell wall degradation and carbohydrate metabolism. These enriched GO terms support the idea that *Cqf* can restructure its own cell wall as it grows within a host, indicated by the enrichment of terms for both chitin metabolic and catabolic processes. In addition, protein degradation functions such as peptidase and serine peptidase activity are enriched. These functions are common in fungal pathogens because they obtain nutrient amino acids from their hosts.

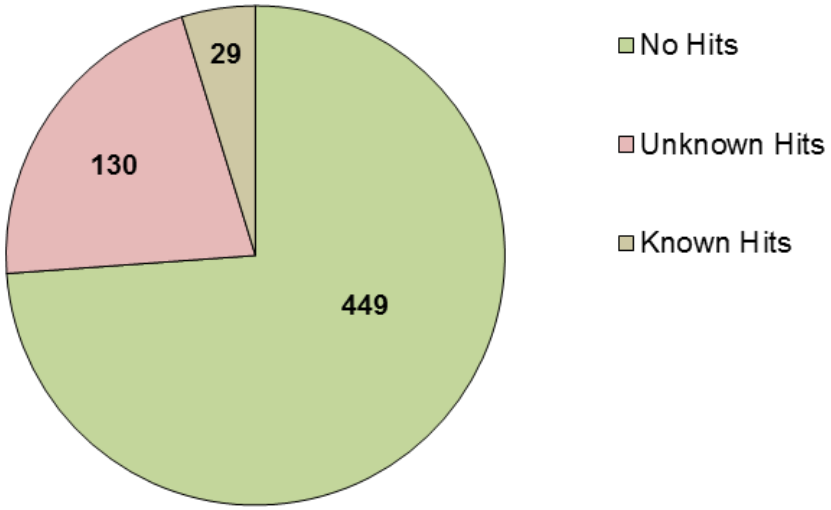

Figure 2-1. BLASTP results of *ab intio* only gene models compared to the NCBI protein database (expect-value cutoff E^{-6}).

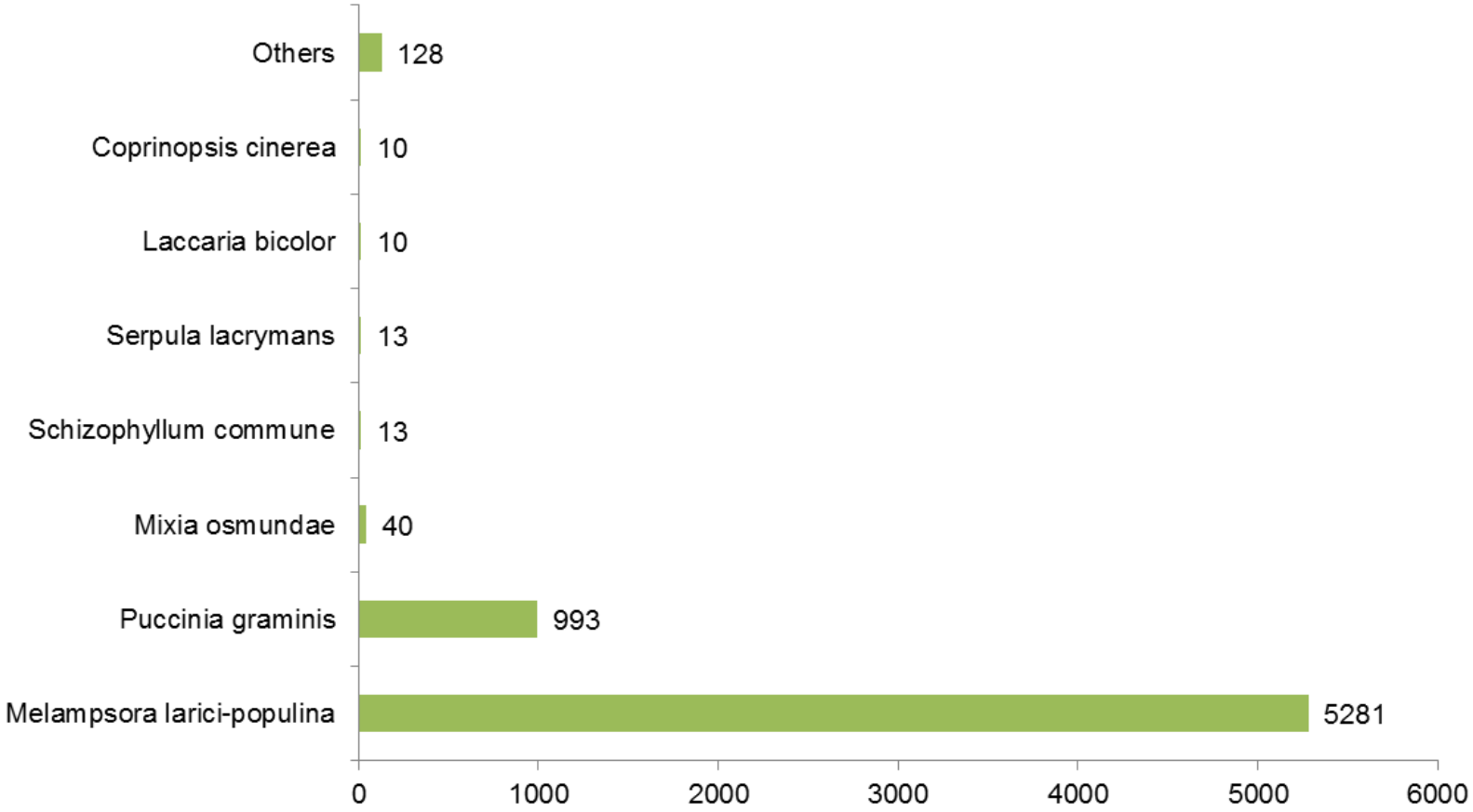

Figure 2-2. The number of top BLASTP hits by species. All species shown have a sequenced genome. The others category includes all species with less than 10 top-hits (expect-value cutoff E^{-6})

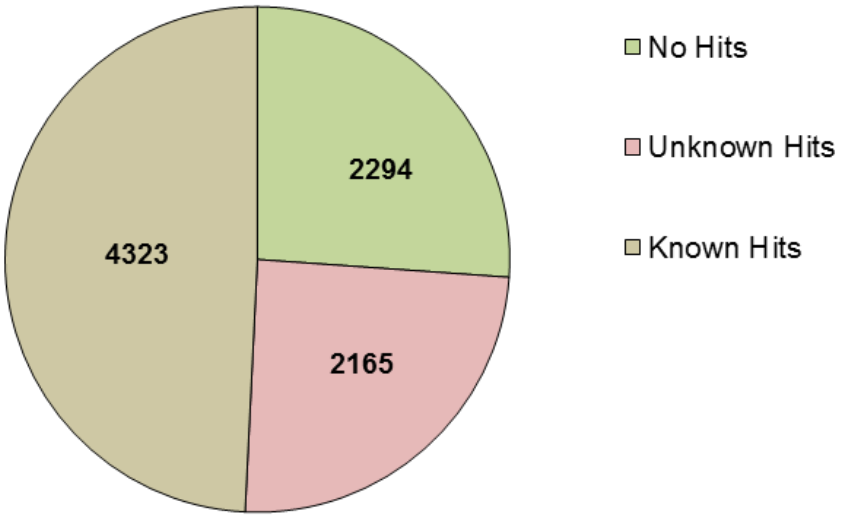

Figure 2-3. BLASTP comparison of *Cqf* gene models to the NCBI protein database (expect-value cutoff E^{-6}).

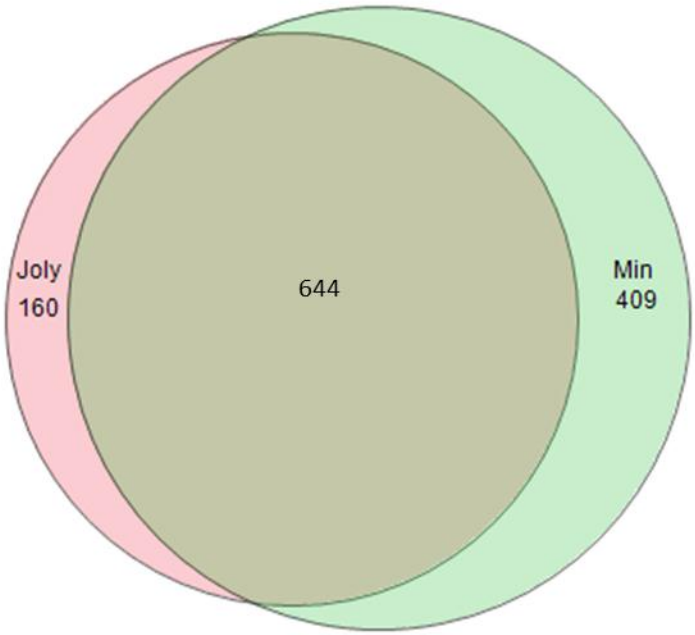

Figure 2-4. Comparison of the number of *Cqf* secreted proteins predicted by two different *in silico* methods from Joly et al., 2010 and Min, 2010.

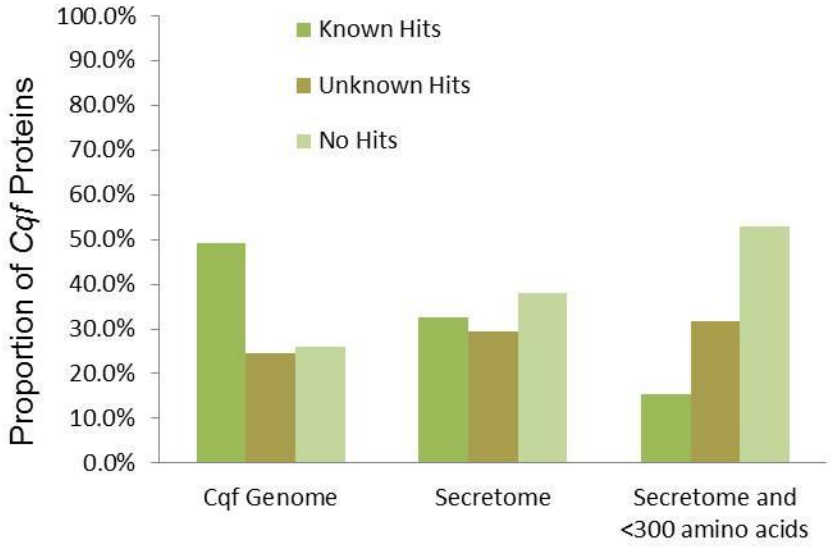

Figure 2-5. The *Cqf* secretome is enriched for proteins with no similarity to known proteins (BLASTP with the NCBI protein database, expect-value cutoff E^{-6}).

Table 2-1. List of GO terms enriched in the *Cqf* secretome (FDR .001)

Category	Term	FDR	# in test group	# in reference group
Process	Carbohydrate Metabolic Process	8.61E-33	76	281
	Metabolic Process	9.38E-13	171	2152
	Polysaccharide Catabolic Process	1.77E-10	14	22
	Primary Metabolic Process	3.80E-09	139	1669
	Polysaccharide Metabolic Process	4.77E-05	16	46
	Carbohydrate Catabolic Process	1.72E-04	15	60
	Disaccharide Metabolic Process	2.78E-04	12	40
	Glycoside Metabolic Process	3.17E-04	12	41
	Oligosaccharide Metabolic Process	3.66E-04	12	42
	Aminoglycan Catabolic Process	3.72E-04	6	9
	Chitin Catabolic Process	3.72E-04	6	9
Function	Hydrolase Activity, Acting on Glycosyl Bonds	3.24E-21	32	64
	Catalytic Activity	4.95E-18	168	1892
	Hydrolase Activity	1.48E-17	97	728
	Hydrolase Activity, Hydrolyzing O-Glycosyl Compounds	6.09E-16	24	47
	Cellulase Activity	2.82E-04	7	12
	Peptidase Activity	3.66E-04	23	140
	Serine-Type Peptidase Activity	3.72E-04	11	36
	Serine Hydrolase Activity	3.72E-04	11	36
	Chitinase Activity	3.72E-04	6	9
	Lipase Activity	4.57E-04	9	24
	Phospholipase Activity	5.82E-04	8	19

CHAPTER 3
MEASURING *CQF* GENE EXPRESSION IN PINE AND OAK INFECTION

Background

Studies of *Cqf* gene expression to date have focused on only a handful of genes (Warren and Covert, 2004; Baker et al., 2006). Using the gene models obtained from the *Cqf* sequencing project, whole genome microarrays were used to measure gene expression during the vegetative stage of infection in both the oak and pine hosts. The fungus infects the pine stem, leading to gall formation. By contrast, the fungus infects oak through leaf stomata and does not stimulate gall formation. The vegetative growth stage of the non-sporulating fungus was chosen as the most comparable between these two different infections. The fungus is actively obtaining nutrients from the host to support its own growth and in order to do so, actively circumventing host defenses. Since effector loci coevolve with host resistance loci it would not be surprising if separate sets of effectors evolved for each host in the genome of a single fungus. Recent microarray studies in the closely related *Melampsora larici-populina* have focused on the telial host only, comparing infection time points, different zones of infected leaves or infected versus uninfected tissues. These data show that small secreted “effector-like” protein encoding genes are expressed more highly in infected tissues compared to spores (Duplessis et al., 2011b). In addition, candidate effectors are expressed to a higher level in haustoria-containing host mesophyll cells compared to areas where sporulation is taking place (Hacquard et al., 2010). *Melampsora larici-populina* also expresses distinct sets of these genes along the time course of infection (Duplessis et al., 2011a; Hacquard et al., 2012). The mircoarray experiment described

in this thesis offers a unique direct comparison of the same fungus actively growing in two different hosts.

In order to obtain *Cqf* gene expression information while the fungus is interacting with each host, the sampled tissues contained both *Cqf* and host transcripts, either pine stem or oak leaf. The use of microarrays offers advantages over high throughput RNA sequencing in this experimental design. Since essentially all genes in the genome have probes on the microarrays all genes were sampled. By contrast, RNA sequencing would require enough sequencing depth to accommodate the highly expressed genes, both host and pathogen, and detect less common transcripts. Lack of sequencing depth could result in skewing the observed global expression pattern. The microarrays "filtered" host transcripts because host transcripts do not hybridize to *Cqf* probes. We used the microarray data to test whether or not the *Cqf* transcriptome is substantially different, or substantially similar, as the fungus infects its two hosts.

Materials and Methods

Fungal and Plant Material

Inoculations of both oak and pine seedlings were done at the USDA Forest Service, Resistance Screening Center in Asheville, North Carolina. A single uredinial spore isolate called SC20-21 was used to inoculate 15 open-pollinated northern red oak (*Quercus rubra*) seedlings. Three infected leaves were pooled and harvested from each of 8 plants five days after inoculation before lesions were visible and before telia formed. The 3 leaves were combined into a single sample and immediately frozen on dry ice. Inoculated oak plants that were not harvested were used: 1) To monitor the subsequent level of infection; and 2) To collect basidiospores for pine inoculations. Heavy infection was observed and ample basidiospores were collected. Fifty open-

pollinated susceptible slash pine (*Pinus elliottii*) seedlings were inoculated. Galled stem sections, of uniform size, were collected from 8 pine seedlings, 18 weeks after inoculation and well before sporulation. Individual stem samples were immediately frozen on dry ice.

RNA Extraction

Pine stem samples were freeze dried for 4 days before extraction. They were broken into small pieces using a coffee grinder. The small pieces were further ground to a fine powder using three 5/32-inch stainless steel balls in microcentrifuge tubes processed in a Geno/Grinder 2000 homogenizer. Oak leaf samples were ground in liquid nitrogen to a fine powder. RNA was extracted from approximately 200mg of ground oak or pine sample using a previously described cetyltrimethylammonium bromide (CTAB) buffer method (Chang et al., 1993).

Microarray Experimental Design

The microarray experiment was a two dye control design (Churchill, 2002). Two Agilent 4 X 44K microarray slides populated with custom probes were used. Probes were designed using Agilent's web-based eArray software. Of the 8782 *Cqf* gene models, eArray designed from one to five 60-mer oligonucleotide probes for 8692 genes. All probes were used and each microarray had 43803 gene features with 26525 probes present once and 8639 probes present twice. Labeled target cRNA (complementary RNA) was generated using Agilent's Low Input Quick Amp Labeling Kit, such that oak and pine samples were labeled with either cy3 or cy5 an equal number of times across the experiment. Each microarray was hybridized with labeled cRNA target derived from a single oak sample and labeled cRNA target derived from a single pine sample. There were a total of eight oak sample replications and eight pine

sample replications. Target hybridization and scanning was performed by the University of Florida's Interdisciplinary Center for Biotechnology Research using standard procedures and an Agilent Microarray Scanner.

Statistical Analysis

Agilent's Feature Extraction Software (version 10.7) was used to extract features and to calculate background subtracted feature signal intensities based on local background. Features were flagged and removed from differential expression analysis if any of the following software determined conditions were met: 1) The feature could not be found; 2) The feature was saturated; 3) The feature was non-uniform; 4) The feature was an outlier; or 5) The feature was not positive and significant over background. A gene was categorized as "not detected" if all oak hybridized and pine hybridized gene features were removed. A gene was categorized as "expressed in oak only" if all pine hybridized gene features were removed. A gene was categorized as "expressed in pine only" if all oak hybridized gene features were removed. All remaining genes were categorized as "expressed in both hosts."

Background subtracted signal intensities were $\log_2$ transformed and normalized by setting microarray means to zero with a standard deviation of 1. Least squared means of gene expression levels in oak and pine were calculated with a mixed model analysis of variance (ANOVA) using PROC MIXED in SAS (SAS Institute, Cary, NC, USA) where the effect of host was fixed and the effects of probe, dye and array were random. Q-values on the estimates between oak and pine were calculated using the statistical software R (The R Foundation for Statistical Computing) with a false discovery rate of 0.01 (Storey and Tibshirani, 2003).

Results

Figure 3-1 depicts the tissue expression patterns of all 8692 genes on the microarrays based on results from the feature extraction software. The majority of genes are expressed in both tissues. The mixed tissue samples, oak leaf/*Cqf* and pine stem/*Cqf*, had the effect of limiting the amount of *Cqf* target hybridizing to the microarrays. Signal intensities across all microarrays were observed to be low based on the fact that 40% of gene features were removed as not positive and significant above background. Despite the large number of features below background only 111 genes were completely removed by flags and designated undetected. This was due to the large number of sample replicates (8) for each host and the fact that most genes had multiple probes with 6865 genes having 3 or more probes after flagged features were removed. In addition, there was evidence for expression of 572 out of 621 *ab initio* genes obtained with microarrays, 95 of which were expressed 2-fold or more in one host over the other.

Figure 3-2 shows the distribution of log2-transformed, background subtracted signal intensities of *Cqf* expression in pine and oak hosts. The mean transcript abundance was higher in pine compared to oak, and the variability of transcript abundance was higher in pine compared to oak (i.e., there was a wider bell-shaped curve in pine compared to oak). A total of 5077 genes showed evidence of significantly higher expression in one host compared to the other, with 3068 transcripts higher in pine and 2009 transcripts higher in oak. Applying a stringent criterion of statistical significance (FDR<0.01) in addition to 4-fold higher transcript abundance in one host compared to the other, 73 genes were expressed to a higher level in pine and 100 genes were expressed higher in oak. Genes encoding secreted proteins, as well as

lineage specific SSPs, were enriched among these highly differentially expressed genes (Figure 3-3, Figure 3-4).

Discussion

The microarray experiment was designed to test whether transcript abundance reflected involvement of distinct, nonoverlapping sets of *Cqf* effectors during development of leaf rust (on oak) and stem galls (on pine). However, almost all *Cqf* genes showed evidence of expression in both *Quercus* and *Pinus* hosts, suggesting that the same set of genes is involved in both infections (Figure 3-1). It therefore seems feasible that the heteroecious life cycle of *Cqf* is enabled by similar genes acting to establish disease states in both hosts. Perhaps *Cqf* effectors target host processes that are conserved between angiosperms and gymnosperms. Alternatively, the host alternation of *Cqf* may be conditioned by post-transcriptional processes that are not reflected by differential transcript abundance.

Interestingly, *Cqf* colonizing oak leaves appears to show a lower mean and less variation in transcript abundance compared to *Cqf* colonizing pine (Figure 3-2). The reason for the observed difference is not obvious, and could be due to *Cqf* abundance and RNA extractability in the samples. Alternatively there could be an underlying biological explanation, perhaps reflecting distinct host-specific constraints on *Cqf* gene regulation.

The most highly differentially expressed *Cqf* genes between oak and pine colonized tissues are likely to be important to host specific infection. This assertion is supported by the fact that these categories are enriched in genes encoding secreted proteins (Figure 3-3). Of the 100 proteins encoded by genes expressed 4-fold or more in oak over pine, 37 are predicted to be secreted and 22 are lineage specific SSPs

(small secreted proteins less than 300 amino acids). Of the 73 proteins encoded by genes expressed 4-fold or more in pine over oak, 28 are predicted to be secreted and 10 are lineage specific SSPs (Figure 3-3, Figure 3-4). These are candidate effectors and include known putative effectors. The 4 rust transferred protein paralogs show an interesting expression pattern; 2 are overexpressed in oak with 3.7 and 6.6 fold increases in expression and 2 are overexpressed in pine with 2.6 and 4.5 fold increases in expression. In bean rust this effector is known to enter the nucleus of host cells, presumably to alter host gene expression in some way (Kemen et al., 2005). In *Cqf* different paralogs of the same effector might be preferentially altering host gene expression in two different hosts.

Gene differentially expressed between oak and pine, yet not predicted to be secreted may still be important to the infection process. Among the 63 genes encoding proteins with 4-fold or more increased expression in oak and not predicted to be secreted there are 44 unknown genes and 19 genes with putative functions. These include a multi-copper oxidase laccase-like protein involved in lignin degradation, 2 proteins involved in transport and 3 proteins encoding signaling proteins. The genes with unknown functions include 1 with an RNA binding domain and 1 with the transcription factor activity GO term. Among the 45 genes encoding proteins with 4-fold or more increased expression in pine and not predicted to be secreted there are 24 unknown genes and 21 genes with putative functions. These include six transporters and 1 identified as an mRNA binding post-transcriptional regulator. In addition, the gene encoding the known mycotoxin synthesis protein TRI14 is expressed 4-fold more in pine

than oak. Relative gene expression determined with microarrays has identified genes that putatively function preferentially during infection of one host or the other.

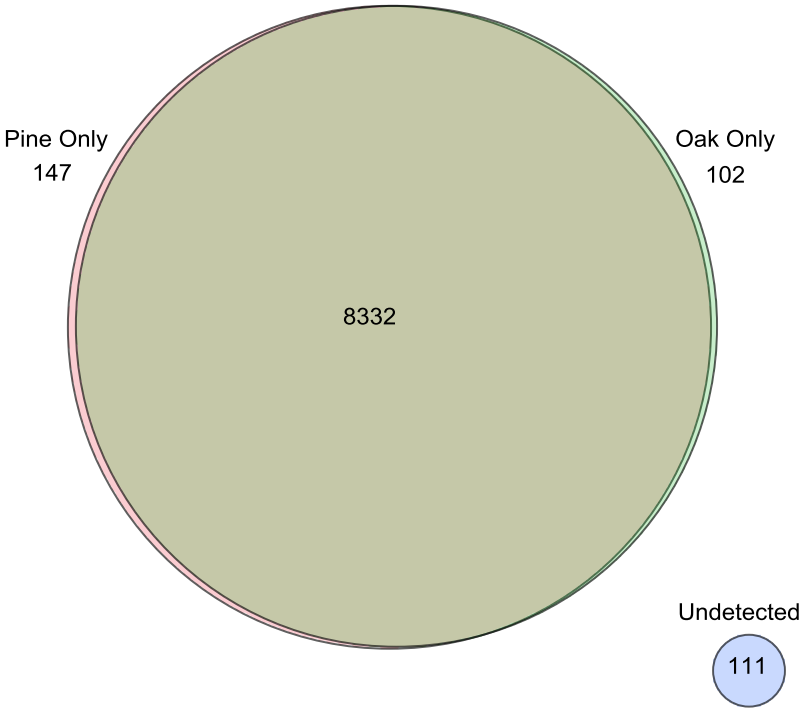

Figure 3-1. The majority of *Cqf* genes are expressed in both hosts.

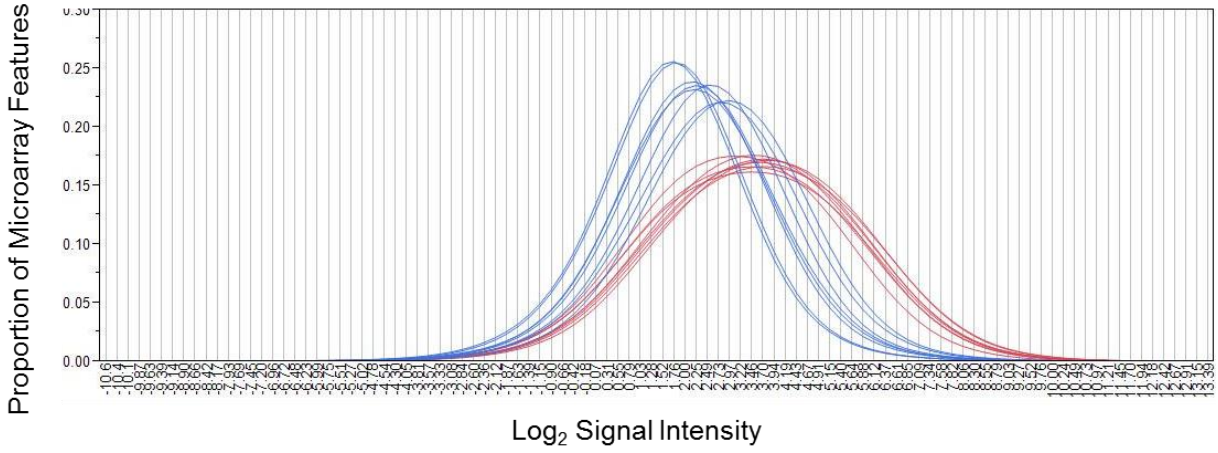

Figure 3-2. Distributions of log2-transformed pine (red) and oak (blue) background subtracted signal intensities. Each line is a single replicate sample.

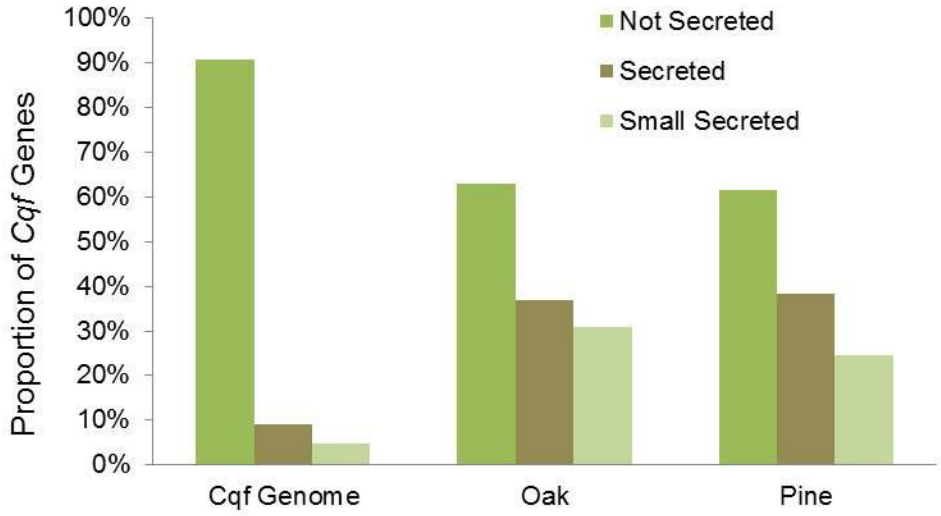

Figure 3-3. Genes encoding secreted proteins are enriched among genes with greater than 4-fold increased expression in one host over the other.

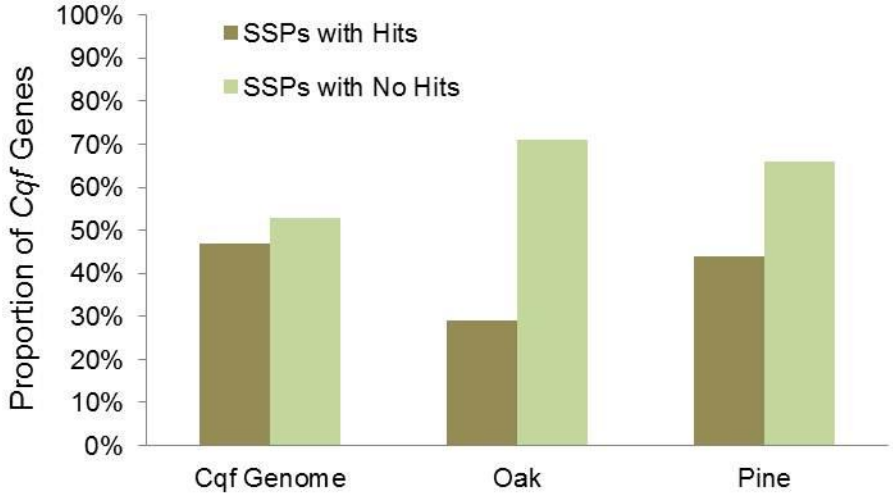

Figure 3-4. Genes encoding lineage-specific SSPs are enriched among genes with greater than 4-fold increased expression in one host over the other (BLASTP, expect-value cutoff E^{-6}).

CHAPTER 4
MAPPING *AVR1* CANDIDATE GENES

Given that the most promising strategy for fusiform rust disease control is the monitoring of *AVR* alleles in natural populations, one of the expected outcomes of the *Cqf* sequencing project was faster identification of *AVR* genes. The release of the JGI final assembly led to the construction of a physical map containing the *Avr1* locus (Figure 4-1) and the identification of *Avr1* candidate genes (Table 4-1).

The generation of additional sequence data included in JGI's final assembly has reduced the number of scaffolds from 2084 to 1198 by joining previously separate scaffolds. Marker sequences of polymorphic bands for 6 amplified fragment length polymorphism (AFLP) markers and two randomly amplified polymorphic DNA (RAPD) markers shown to be linked to *Avr1* (Kubisiak et al., 2011) were previously obtained, along with additional sequence surrounding the RAPD marker sequences, generated using the GenomeWalker Universal kit (Clontech #63894). These sequences were compared to the scaffold assembly and the results are summarized in Table 4-2. Two AFLP-derived and one large RAPD-derived marker sequence, as well as a simple sequence repeat (SSR) were localized to scaffold 20 in an order predicted by the *Cqf* genetic map (Kubisiak et al., 2011; Figure 4-1). The RAPD marker BB07 matched to scaffold 20 with 2229 of 2235 bp showing 100% identity. The AFLP marker E13M6 matched to scaffold 20 with an exact match in 74 of 74bp, while E6M7 matched to scaffold 20 with 99 of 461 bp and 96% identity. The primer sequences of SSR DN_058 occurred in scaffold 20 at a distance and orientation that would predict a polymerase chain reaction (PCR) band of 331 bp that includes 20 repeats of the dinucleotide pair TG, as expected based on marker development experiments in *Cqf* (Burdine et al.,

2007). These data provide good evidence that the *Avr1* gene is likely to reside between markers BB07 and E13M6 on the physical map. This interval contains 25 genes, 3 of which are predicted to encode secreted proteins (Table 4-1). The most logical candidates are genes encoding secreted proteins since avirulence proteins must leave fungal cells in order to interact with the host proteins and lead to the no disease phenotype, however, prediction methods are not foolproof and all genes should be considered as candidates.

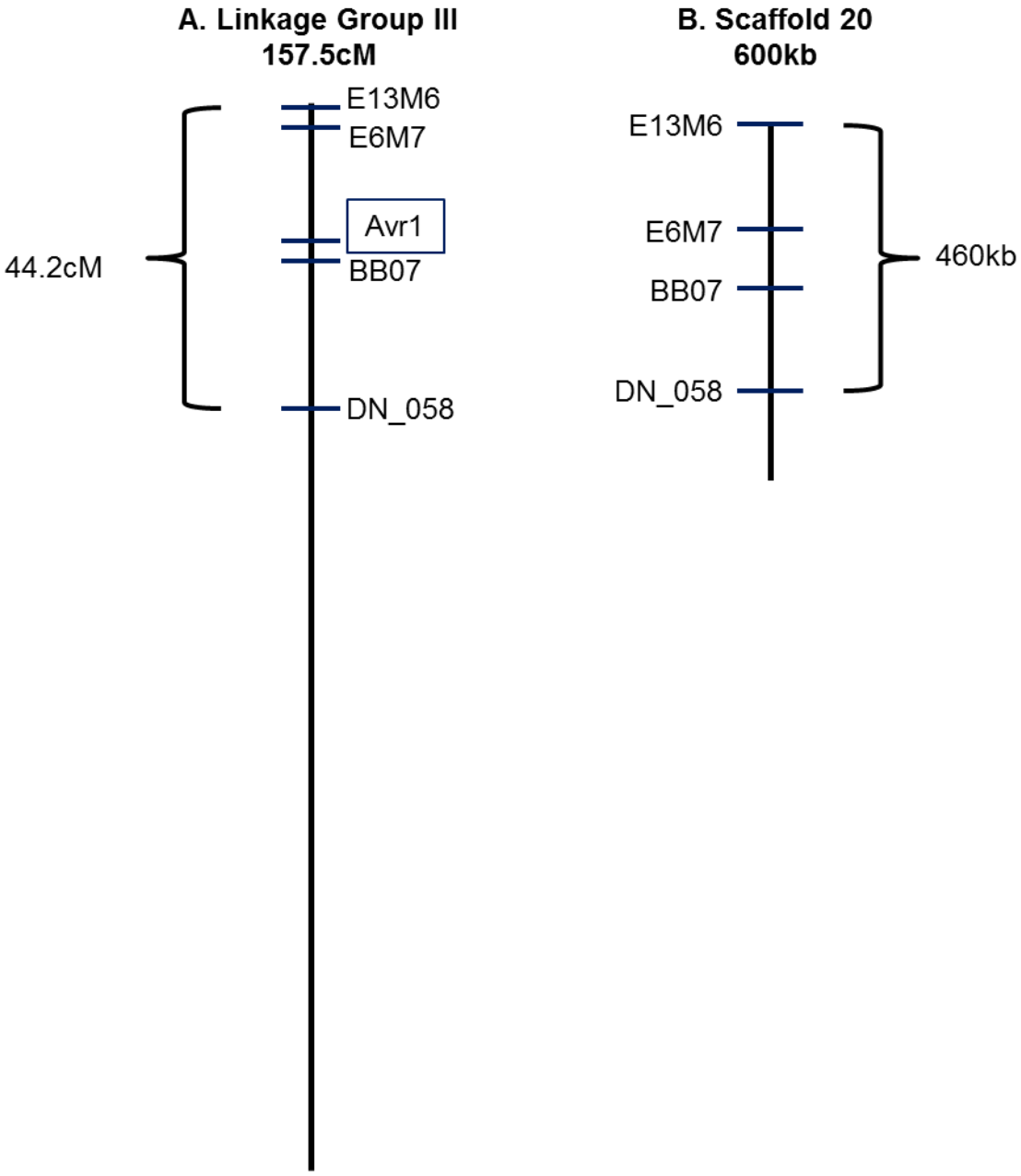

Figure 4-1. Genetic (A) and physical (B) maps of the *Avr1* locus indicate a physical/genetic distance ratio of 10.4kb/cM (460kb/44.2cM).

Table 4-1. *AVR1* candidate genes by location on scaffold 20

Scaffold 20 Coordinates	BLASTP Description	BLASTP E-value	Protein Length	Secretion Prediction	Microarray Expression (FDR 0.01)
4676-4749	E13M6 Marker				
11036-5318	hypothetical protein [Melampsora larici-populina]	$1.6e^{-173}$	1397		expressed in pine and oak
13459-13033	no hits		26		not tested
18285-22499	nuclear elongation and deformation protein 1	0	1115		expressed in pine and oak
41077-43578	hypothetical protein [Melampsora larici-populina]	0	312		expressed in pine and oak
74051-70543	developmental regulator	$6.1e^{-31}$	371		expressed in pine and oak
79309-77744	ribosomal RNA-processing protein 8	$2.0e^{-144}$	377		expressed in pine and oak
85282-86634	hypothetical protein [Puccinia graminis f. sp. tritici]	$5.7e^{-103}$	356	yes	expressed in pine and oak
106082-109381	hypothetical protein [Melampsora larici-populina]	$1.4e^{-101}$	763		expressed in pine and oak
118826-116211	aspartyl aminopeptidase	0	521		expressed in pine and oak
143532-133336	cation-transporting ATPase	0	2418	yes	expressed in pine and oak
145410-144897	no hits	-	103		4-fold up in oak
146710-147654	unknown protein	$6.4e^{-20}$	173		not tested
164104-164978	peptidyl-prolyl cis-trans isomerase NIMA-interacting 4	$2.9e^{-53}$	140		expressed in pine and oak
165764-168672	no hits	-	508	yes	2-fold up in pine

Table 4-1. Continued

Scaffold 20 Coordinates	BLASTP Description	BLASTP E-value	Protein Length	Secretion Prediction	Microarray Expression (FDR 0.01)
175775-177281	unknown protein	$6.6e^{-22}$	336		2-fold up in oak
178998-179097	E6M7 Marker				
194077-195314	no hits	-	103		4-fold up in oak
203178-200574	hypothetical protein [Melampsora larici-populina]	$1.75e^{-75}$	582		2-fold up in oak
205083-204227	ring-box protein 1	$9.9e^{-37}$	93		expressed in pine and oak
214869-213003	g protein alpha subunit	$1.6e^{-141}$	376		expressed in pine and oak
225168-218906	myosin 5	0	1689		expressed in pine and oak
227391-228236	guanine nucleotide-binding protein alpha-2 subunit	$1.1e^{-21}$	121		expressed in pine and oak
241466-242464	uracil phosphoribosyltransferase	$6.0e^{-115}$	199		expressed in pine and oak
243491-244153	hypothetical protein [Melampsora larici-populina]	$7.8e^{-9}$	220		expressed in pine and oak
252449-251812	hypothetical protein [Melampsora larici-populina]	$2.7e^{-59}$	186		expressed in pine and oak
253915-255134	scf complex subunit skp1	$1.0e^{-91}$	158		expressed in pine and oak
251681-254246	BB07 Marker				

Table 4-2. *AVR1* marker sequence identity to final assembly scaffolds

Marker	Type	Sequence Length	Top Hit Scaffold	Top Hit E-value	Next Best Hit Scaffold/E-value
E6M7*	AFLP	461	906	0	204/8E^{-52}
E7M6	AFLP	470	6	0	229/E^{-163}
E9M7	AFLP	437	39	0	202/0.10
E8M22	AFLP	473	47	0	53/0
E12M8	AFLP	201	398	E^{-110}	34/0.72
E13M6	AFLP	74	20	2E^{-35}	34/0.23
BB07	RAPD	2235	20	0	22/E-^{117}
AZ17	RAPD	903	32	0	309/0
DN_058 forward primer**	SSR	20	20	1.4	26/5.7
DN_058 reverse primer**	SSR	20	20	0.092	28/1.4

*E6M7 has homology to scaffold 20 with and E-value of 2E^{-40}
**Scaffold 20 contains a 331bp interval between the forward and reverse SSR primers that includes the dinucleotide repeat TG, repeated 20 times.

CHAPTER 5
CONCLUSIONS

This project has provided gene expression evidence that *Cqf* uses essentially one set of genes to cause two very different infections; pine gall formation and oak leaf colonization. The identification of encoded secreted proteins in the *Cqf* genome will serve as a reference for future experiments evaluating effectors and their roles in pathogenicity. Secretome information and gene expression results can be combined to provide valuable insight into candidate effectors. For example, two different families of paralogous genes that encode secreted proteins (one set of 4, and one set of 5) show at least 2-fold increased expression during pine infection compared to oak infection. The fact that these genes are unique to *Cqf* and show expression bias make them good candidates for effectors important to fusiform rust disease.

In addition, this work has provided a sequence interval that greatly enhances the possibility of identifying the fusiform rust avirulence gene, *Avr1* (Kubisiak et al., 2011), that interacts genetically with the resistance gene *Fr1* in loblolly pine (Wilcox et al., 1996). Two strategies are being taken to pinpoint this important effector. First, experiments are underway to narrow the candidate gene interval by identifying additional genetic markers linked to *Avr1* using high throughput sequencing data. Also in progress is a re-sequencing effort to identify alleles of candidate genes associated with the disease resistance and susceptibility by screening unrelated pycniospores from pines genotyped for the resistance locus. In a broader context, there are plans to integrate the previously published genetic map with the scaffold sequence in order to obtain a clearer picture of the *Cqf* genome and aid in the generation of physical maps for other *AVR* genes (Kubisiak et al., 2011).

APPENDIX
FUNCTIONALLY ENRICHED GO ANNOTATIONS IN THE *CQF* SECRETOME

Table A-1. Complete list of GO terms enriched in the *Cqf* secretome (FDR .05)

Category	Term	FDR
Process	Carbohydrate Metabolic Process	8.61E-33
	Metabolic Process	9.38E-13
	Polysaccharide Catabolic Process	1.77E-10
	Primary Metabolic Process	3.80E-09
	Polysaccharide Metabolic Process	4.77E-05
	Carbohydrate Catabolic Process	1.72E-04
	Disaccharide Metabolic Process	2.78E-04
	Glycoside Metabolic Process	3.17E-04
	Oligosaccharide Metabolic Process	3.66E-04
	Aminoglycan Catabolic Process	3.72E-04
	Chitin Catabolic Process	3.72E-04
	Cellular Carbohydrate Metabolic Process	0.00121
	Phospholipid Catabolic Process	0.001306
	Cell Wall Macromolecule Catabolic Process	0.001435
	Cellular Polysaccharide Metabolic Process	0.002117
	Proteolysis	0.003736
	Chitin Metabolic Process	0.003945
	Cell Wall Macromolecule Metabolic Process	0.004693
	Lipid Catabolic Process	0.005398
	Cellular Lipid Catabolic Process	0.006946
	Starch Metabolic Process	0.006946
	Sucrose Metabolic Process	0.006946
	Cellular Glucan Metabolic Process	0.015064
	Glucan Metabolic Process	0.015064
	Macromolecule Catabolic Process	0.016761
Function	Hydrolase Activity, Acting on Glycosyl Bonds	3.24E-21
	Catalytic Activity	4.95E-18
	Hydrolase Activity	1.48E-17
	Hydrolase Activity, Hydrolyzing O-Glycosyl Compounds	6.09E-16
	Cellulase Activity	2.82E-04
	Peptidase Activity	3.66E-04
	Serine-Type Peptidase Activity	3.72E-04
	Serine Hydrolase Activity	3.72E-04
	Chitinase Activity	3.72E-04
	Lipase Activity	4.57E-04
	Phospholipase Activity	5.82E-04
	Ion Binding	0.002612
	Cation Binding	0.002612
	Peptidase Activity, Acting on L-Amino Acid Peptides	0.003028
	Endopeptidase Activity	0.046954
Compartment	Extracellular Region	0.001435

LIST OF REFERENCES

Anderson, C.L., Kubisiak, T.L., Nelson, C.D., Smith, J.A., and Davis, J.M. 2010. Genome size variation in the pine fusiform rust pathogen *Cronartium quercuum* f.sp. *fusiforme* as determined by flow cytometry. Mycologia 102:1295-1302.

Anderson, R.L., McClure, J.P., Cost, N., and Uhler, R.J. 1986. Estimating fusiform rust losses in five southeast states. Southern Journal of Applied Forestry 10:237-240.

Baker, L.G., Spaine, P., and Covert, S.F. 2006. Effect of surface wettability on germination and gene expression in *Cronartium quercuum* f. sp. *fusiforme* basidiospores. Physiol Mol Plant P 68:168-175.

Burdine, C.S., Kubisiak, T.L., Johnson, G.N., and Nelson, C.D. 2007. Fifty-two polymorphic microsatellite loci in the rust fungus, *Cronartium quercuum* f.sp. *fusiforme*. Molecular Ecology Notes 7:1005-1008.

Burdsall, H.H., Jr., and Snow, G.A. 1977. Taxonomy of *Cronartium quercuum* and *C. fusiforme*. Mycologia 69:503-508.

Catanzariti, A.-M., Dodds, Peter N., Lawrence, Gregory J., Ayliffe, Michael A. and Ellis, Jeffrey G. 2006. Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. The Plant Cell 18:243-256.

Chang, S., Puryear, J., and Cairney, J. 1993. A simple and efficient method for isolating RNA from pine trees. Plant Molecular Biology Reporter 11:113-116.

Churchill, G.A. 2002. Fundamentals of experimental design for cDNA microarrays. Nat Genet.32:490-495.

Conesa, A., Götz, S., García-Gómez, J.M., Terol, J., Talón, M., and Robles, M. 2005. BLAST2GO: a universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics 21:3674-3676.

Dean, R.A., Talbot, N.J., Ebbole, D.J., Farman, M.L., Mitchell, T.K., Orbach, M.J., Thon, M., Kulkarni, R., Xu, J.-R., Pan, H., Read, N.D., Lee, Y.-H., Carbone, I., Brown, D., Oh, Y.Y., Donofrio, N., Jeong, J.S., Soanes, D.M., Djonovic, S., Kolomiets, E., Rehmeyer, C., Li, W., Harding, M., Kim, S., Lebrun, M.-H., Bohnert, H., Coughlan, S., Butler, J., Calvo, S., Ma, L.-J., Nicol, R., Purcell, S., Nusbaum, C., Galagan, J.E., and Birren, B.W. 2005. The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature 434:980-986.

Duplessis, S., Hacquard, S., Delaruelle, C., Tisserant, E., Frey, P., Martin, F., and Kohler, A. 2011a. *Melampsora larici-populina* transcript profiling during germination and timecourse infection of poplar leaves reveals dynamic expression patterns associated with virulence and biotrophy. Mol Plant Microbe In 24:808-818.

Duplessis, S., Cuomo, C.A., Lin, Y.-C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., Joly, D.L., Hacquard, S., Amselem, J.l., Cantarel, B.L., Chiu, R., Coutinho, P.M., Feau, N., Field, M., Frey, P., Gelhaye, E., Goldberg, J., Grabherr, M.G., Kodira, C.D., Kohler, A., Kües, U., Lindquist, E.A., Lucas, S.M., Mago, R., Mauceli, E., Morin, E., Murat, C., Pangilinan, J.L., Park, R., Pearson, M., Quesneville, H., Rouhier, N., Sakthikumar, S., Salamov, A.A., Schmutz, J., Selles, B., Shapiro, H., Tanguay, P., Tuskan, G.A., Henrissat, B., Van de Peer, Y., Rouze, P., Ellis, J.G., Dodds, P.N., Schein, J.E., Zhong, S., Hamelin, R.C., Grigoriev, I.V., Szabo, L.J., and Martin, F. 2011b. Obligate biotrophy features unraveled by the genomic analysis of rust fungi. Proceedings of the National Academy of Sciences.

Dwinell, L.D. 1976. Biology of fusiform rust. in management of fusiform rust in southern pines, Symposium Proceedings of the Southern Forest Disease and Insect Research Council, R.J.S. Dinus, Robert A., ed (University of Florda, Gainesville, Flordia), pp. 18-24.

Dyer, R.B., Plattner, R.D., Kendra, D.F., and Brown, D.W. 2005. *Fusarium graminearum* TRI14 Is Required for High Virulence and DON Production on Wheat but Not for DON Synthesis in Vitro. J Agr Food Chem 53:9281-9287.

Ellis, J.G., Dodds, P.N., and Lawrence, G.J. 2007. Flax rust resistance gene specificity is based on direct resistance-avirulence protein interactions. Annual Review of Phytopathology 45:289-306.

Flor, H.H. 1955. Host-parasite interaction in flax rust - its genetics and other implications. Phytopathology 45:680-685.

Gan, P.H.P., Rafiqi, M., Hardham, A.R., and Dodds, P.N. 2010. Effectors of biotrophic fungal plant pathogens. Functional Plant Biology 37:913-918.

Gray, D.J., Amerson, H.V., and Dyke, C.G.V. 1982. An ultrastructural comparison of monokaryotic and dikaryotic haustoria formed by the fusiform rust fungus Cronartium quercuum f.sp. fusiforme. Canadian Journal of Botany 60:2914-2922.

Hacquard, S., Delaruelle, C., Legué, V., Tisserant, E., Kohler, A., Frey, P., Martin, F., and Duplessis, S. 2010. Laser capture microdissection of uredinia formed by *Melampsora larici-populina* revealed a transcriptional switch between biotrophy and sporulation. Mol Plant Microbe In 23:1275-1286.

Hacquard, S.p., Joly, D.L., Lin, Y.-C., Tisserant, E., Feau, N., Delaruelle, C., Legue, V., Kohler, A., Tanguay, P., Petre, B., Frey, P., Van de Peer, Y., Rouzé, P., Martin, F.M., Hamelin, R.C., and Duplessis, S.b. 2012. A comprehensive analysis of genes encoding small secreted proteins identifies candidate effectors in *Melampsora larici-populina* (poplar leaf rust). Mol Plant Microbe In.

Hahn, M., and Mendgen, K. 1997. Characterization of in planta-induced rust genes isolated from a haustorium-specific cDNA library. Mol Plant Microbe In 10:427-437.

Hahn, M., and Mendgen, K. 2001. Signal and nutrient exchange at biotrophic plant-fungus interfaces. Curr Opin Plant Biol 4:322-327.

Holt, C., and Yandell, M. 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. Bmc Bioinformatics 12:491.

Joly, D., Feau, N., Tanguay, P., and Hamelin, R. 2010. Comparative analysis of secreted protein evolution using expressed sequence tags from four poplar leaf rusts (*Melampsora* spp.). BMC Genomics 11:422.

Kamper, J., Kahmann, R., Bölker, M., Ma, L.-J., Brefort, T., Saville, B.J., Banuett, F., Kronstad, J.W., Gold, S.E., Muller, O., Perlin, M.H., Wosten, H.A.B., de Vries, R., Ruiz-Herrera, J., Reynaga-Pena, C.G., Snetselaar, K., McCann, M., Perez-Martin, J., Feldbrugge, M., Basse, C.W., Steinberg, G., Ibeas, J.I., Holloman, W., Guzman, P., Farman, M., Stajich, J.E., Sentandreu, R., Gonzalez-Prieto, J.M., Kennell, J.C., Molina, L., Schirawski, J., Mendoza-Mendoza, A., Greilinger, D., Munch, K., Rossel, N., Scherer, M., Vranes, M., Ladendorf, O., Vincon, V., Fuchs, U., Sandrock, B., Meng, S., Ho, E.C.H., Cahill, M.J., Boyce, K.J., Klose, J., Klosterman, S.J., Deelstra, H.J., Ortiz-Castellanos, L., Li, W., Sanchez-Alonso, P., Schreier, P.H., Hauser-Hahn, I., Vaupel, M., Koopmann, E., Friedrich, G., Voss, H., Schluter, T., Margolis, J., Platt, D., Swimmer, C., Gnirke, A., Chen, F., Vysotskaia, V., Mannhaupt, G., Guldener, U., Münsterkötter, M., Haase, D., Oesterheld, M., Mewes, H.-W., Mauceli, E.W., DeCaprio, D., Wade, C.M., Butler, J., Young, S., Jaffe, D.B., Calvo, S., Nusbaum, C., Galagan, J., and Birren, B.W. 2006. Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. Nature 444:97-101.

Kemen, E., Kemen, A.C., Rafiqi, M., Hempel, U., Mendgen, K., Hahn, M., and Voegele, R.T. 2005. Identification of a Protein from Rust Fungi Transferred from Haustoria into Infected Plant Cells. Mol Plant Microbe In 18:1130-1139.

Krogh, A., Larsson, B.r., von Heijne, G., and Sonnhammer, E.L.L. 2001. Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. J Mol Biol 305:567-580.

Kubisiak, T.L., Amerson, H.V., and Nelson, C.D. 2005. Genetic interaction of the fusiform rust fungus with resistance gene Fr1 in loblolly pine. Phytopathology 95:376-380.

Kubisiak, T.L., Anderson, C.L., Amerson, H.V., Smith, J.A., Davis, J.M., and Nelson, C.D. 2011. A genomic map enriched for markers linked to Avr1 in *Cronartium quercuum* f.sp. *fusiforme*. Fungal Genet Biol 48:266-274.

Kulkarni, R.D., Kelkar, H.S., and Dean, R.A. 2003. An eight-cysteine-containing CFEM domain unique to a group of fungal membrane proteins. Trends Biochem Sci 28:118-121.

Min, X.J. 2010. Evaluation of computational methods for secreted protein prediction in different eukaryotes. Journal of Proteomics and Bioinformatics 3:143-147.

Nelson, C.D., Kubisiak, T.L., and Amerson, H.V. 2010. Unravelling and managing fusiform rust disease: a model approach for coevolved forest tree pathosystems. Forest Pathol 40:64-72.

Panstruga, R., and Dodds, P.N. 2009. Terrific protein traffic: the mystery of effector protein delivery by filamentous plant pathogens. Science 324:748-750.

Phelps, W.R., Czabator, F.L. 1978. Fusiform rust of southern pines. in forest insect and disease Leaflet, Forest Service, United States Department of Agriculture, ed.

Saunders, D.G.O., Win, J., Cano, L.M., Szabo, L.J., Kamoun, S., and Raffaele, S. 2012. Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi. PLoS ONE 7:e29847.

Schmidt, R.A. 1998. Fusiform rust disease of southern pines: biology, ecology and management. In University of Florida, Institute of Food and Agriculture, Technical Bulletin, pp. 1-14.

Schmidt, R.A. 2003. Fusiform rust of southern pines: a major success for forest disease management. Phytopathology 93:1048-1051.

Skamnioti, P., Gurr, Sarah J. 2007. *Magnaporthe grisea* Cutinase2 mediates appressorium differentiation and host penetration and is required for full virulence. The Plant Cell 19:2674-2689.

Starkey, D.A., Anderson, Robert L. Young, Carol H., Cost, Noel D., Vissage, John S., May, Dennis M., Yockey, Edwin K. 1997. Monitoring incidence of fusiform rust in the south and change over time. In Protection Report (Atlanta, Georgia: United States Department of Agriculture, Forest Service).

Stergiopoulos, I., and de Wit, P.J.G.M. 2009. Fungal effector proteins. Annual Review of Phytopathology 47:233-263.

Storey, J.D., and Tibshirani, R. 2003. Statistical significance for genomewide studies. Proceedings of the National Academy of Sciences 100:9440-9445.

van den Burg, H.A., Harrison, S.J., Joosten, M.H.A.J., Vervoort, J., and de Wit, P.J.G.M. 2006. *Cladosporium fulvum* Avr4 protects fungal cell walls against hydrolysis by plant chitinases accumulating during infection. Mol Plant Microbe In 19:1420-1430.

van Esse, H.P., van't Klooster, John W., Bolton, Melvin D, Yadeta, Koste A., van Baarlen, Peter, Boeren, Sjef, Vervoort, Jacques, de Wit, Pierre J.G.M. and Thomma, Bart P.H.J. . 2008. The *Cladosporium fulvum* virulence protein Avr2 inhibits host proteases required for basal defense. The Plant Cell 20:1948-1963.

Vogler, D.R., and Bruns, T.D. 1998. Phylogenetic relationships among the pine stem rust fungi (*Cronartium* and *Peridermium* spp.). Mycologia 90:244-257.

Warren, J.M., and Covert, S.F. 2004. Differential expression of pine and *Cronartium quercuum* f. sp. *fusiforme* genes in fusiform rust galls. Appl Environ Microb 70:441-451.

Wilcox, P.L., Amerson, H.V., Kuhlman, E.G., Liu, B.H., OMalley, D.M., and Sederoff, R.R. 1996. Detection of a major gene for resistance to fusiform rust disease in loblolly pine by genomic mapping. P Natl Acad Sci USA 93:3859-3864.

BIOGRAPHICAL SKETCH

Katherine Smith has held a diverse array of laboratory technician positions since her 1988 graduation from Virginia Tech, with a Bachelor of Science in biology. She has been working in the field of plant pathology since 1996 and began working on fusiform rust disease when she was hired by the USDA Forest Service in 2000. She contributed to the work on genetic markers that surround the first *Cqf Avr* gene, *Avr1*. That research in combination with genomic sequencing could eventually lead to the cloning of *Avr1*. While working for the Forest Service, she began taking graduate courses in the fall of 2006 and entered the PMCB program in spring 2010, where she earned a Master of Science degree studying *Cqf* fungal effectors.