Genomics studies of two cereal rust fungi with a focus on avirulence gene searches

Jiapeng Chen

School of Life and Environmental Sciences

The University of Sydney

A thesis submitted in fulfilment of the requirements for the degree of

Doctor of Philosophy

Augustus 2017

Statement of originality

This thesis has not been submitted for any degree or other purposes to this or any other University. I certify that the intellectual content of this thesis is the product of my own work and that all the technical assistance received in preparing this thesis and sources have been acknowledged.

Jiapeng Chen Jiapeng Chen

Digitally signed by Jiapeng Chen
Date: 2017.11.20
20:39:54 +08'00'

Acknowledgements

First of all, I would like to thank the Plant Breeding Institute (PBI) for providing the facilities and the valuable collection of rust isolates to conduct my research. Acknowledgement also goes to the Judith and David Coffey Life Lab, the Grains Research and Development Corporation (GRDC), Two Blades Foundation, and the Faculty of Agriculture and Environment/School of Life and Environmental Sciences, for research funding and scholarship.

I would like to express my appreciation for my supervisor, Prof. Robert Park, for giving me a great opportunity to study in the Rust Group, and showing great patience for the improvement of my writing skills. I really appreciate him for believing in me having the potential to perform a PhD study when I applied to upgrade my MPhil study to PhD track in 2016. Also, my study wouldn't have been possible without his efforts to find funding for my scholarship.

Sincere appreciation goes to my auxiliary supervisors Dr. Peng Zhang, Dr. Jingqin Wu, and Dr. Chongmei Dong for the critical discussions of my research results and feedbacks on my writings. Also, all the technical assistances from Dr. Peng Zhang, Dr. Sami Hoxha, Dr. Haydar Karaoglu, Dr. Chongmei Dong, and Dr. Ashlea Grewar for sequencing libraries constructions are much appreciated.

I would also like to thank my co-supervisor, Dr. Peter Dodds, who has provided valuable scientific advice to my studies and more importantly, taught me a highly productive workstyle. My work experience, including the seminars, presentations, and meetings, in the group has broadened my view of plant immunity and pathology. To the rest of colleagues at the

Crace site of CSIRO: all the talks and jokes we had in morning teas and lunch times is a sweet memory in the background of the PhD struggle.

To all the past and present members of the Rust Group: the sincerest thanks because of the fundamental resources and research platforms you have built, and all the knowledge of cereal rust pathogens you have left as a great legacy through all over the years. These were the starting point of my study. A special appreciation to the supporting staff in PBI who maintains the daily running of the institute and enable us to focus on our research. A special appreciation to Charlie, the coffee van lady, for providing coffee of great quality.

Thank you to my parents and sister, who have provided so much encouragement and pride for me. Specifically, my dad for teaching me a strong work ethic by setting a daily example throughout my childhood. To my mother and sister for the weekly phone calls to cheer me up.

Abstract

The rust fungi (phylum Basidiomycota, class Teliomycetes, order Pucciniales) are obligate biotrophic pathogens of a wide range of plant species and have been a substantial threat to crop security. Resistance breeding by incorporating resistance (R) genes into cultivars has been a cost-effective approach to protect cereal crops against rust diseases. Those resistance genes typically encode immune receptor proteins that can recognize avirulence (Avr) proteins from rust fungi. However, such resistance is frequently evaded by rust fungi via evolution of new virulence, typically through deletion or alteration of the Avr genes. Therefore, the identification and molecular cloning of R and Avr genes is an essential step for understanding the molecular arms-races between plants and rust fungi.

The research work in this thesis focused on searches for Avr gene in two rust fungi, the causal agent of wheat stem rust *Puccinia graminis* f. sp. *tritici* (*Pgt*), and the causal agent of barley leaf rust *Puccinia hordei* (*Ph*), with Next-generation sequencing (NGS) technologies. In **Chapter 1**, a general background about the importance of the host plants and the fungi was covered. Literature related to plant immunity and pathogen effector proteins was reviewed. In addition, a critical survey was performed to discuss bioinformatics tools for effector protein prediction, particularly those based on NGS techniques.

In **Chapter 2**, two *Pgt* isolates were studied in detail, one wildtype (ID 279) and one mutant derivative (ID 632) that differed in carrying virulence to wheat stem rust resistance gene *Sr50*. This phenotypic difference was assumed to be caused by abolishment or mutation of the corresponding Avr gene *AvrSr50* in *Pgt*632. Towards identification of this Avr gene,

genomes of the two isolates were sequenced. Analysis of sequence variation between the two isolates in genes encoding haustorially-expressed secreted proteins (HSPs) revealed amino acid-changing variations in 18 HSP genes. In these genes, allelic variants from one haplotype were missing in the mutant *Pgt*632. Genome wide comparisons identified a large chromosomal segment of about 2.5 Mbp with loss-of-heterozygosity (LOH) in the mutant derivative, which spanned all the 18 HSPs. Further analysis showed no reduction of genomic mass in the LOH contigs, indicating that it might have resulted from a rare somatic recombination event rather than a simple genomic deletion. This mutation event was potentially associated with the loss of *AvrSr50*. The LOH region harbored 25 annotated HSP genes, which were considered likely to include *AvrSr50*. Allele sequences were resolved for 20 of these loci that appeared to be single copy genes based on analysis of allele frequencies at variant sites. One of these constructed allele sequences was validated to encode the *AvrSr50* protein, showing for the first time the effectiveness of Avr gene search via comparative genomics in rust fungi.

The genomics analysis of *Pgt* was based on two reference genomes completed in 2011 and 2015. However, a reference genome was still lacking for *Ph*. In **Chapter 3**, a *de novo* genome assembly was performed for a *Ph* isolate *Ph*612, producing 15,913 scaffolds amounting to 127 Mbp. A scan of the assembly revealed that 55% of it comprised repetitive elements. Gene models were predicted using transcriptome sequencing of barley leaf tissues inoculated with *Ph*612 and homologous proteins from *Pgt*. A total of 16,354 genes were predicted, including 1,072 secreted protein-encoding genes that are potentially avirulence genes. Functional analyses of the predicted genes revealed important genomic contents related to various biological processes in this pathogen, including mating type loci, hallmarks associated with the obligate biotrophic lifestyle, and putative effector functions.

The completion of the *Ph* genome assembly allowed mapping of whole genome re-sequencing data, which is becoming a high-throughput and cost-effective method to identify Avr gene candidates. In **Chapter 4**, four additional *Ph* isolates derived from a same clonal lineage as *Ph*612 were studied. These isolates differed in virulence for three barley R genes *Rph3*, *Rph13* and *Rph19*. Towards the identification of the corresponding Avr genes (designated as *AvrRph3*, *AvrRph13*, and *AvrRph19*), the isolates were sequenced and mapped to the reference genome for genotype inference. DNA sequence variations in the secreted protein-coding genes were analyzed and related to differences in virulence, resulting in 99, 114 and 120 candidate genes for *AvrRph13*, *AvrRph3*, and *AvrRph19*, respectively. The identification of these candidates set a foundation for future functional studies targeting the isolation and characterization of the three Avr genes.

The research work detailed in this thesis revealed the existence of high genetic diversity in the candidate effector genes identified in both rust fungi. Such diversity increases the evolutionary potential of the pathogens to overcome host resistance. The bioinformatics methods and logical framework developed in this thesis successfully narrowed down the search of *AvrSr50* to 18 candidate genes in *Pgt*, and reported close to 100 candidates for three Avr genes in *Ph*. The results have contributed to the pool of knowledge about rust virulence evolution, which will assist in development of more durable resistance to these pathogens in both wheat and barley.

Abbreviations

AA	amino acid
Avr gene	avirulence gene
bp	base pair
ETI	effector-triggered immunity
ETS	effector-triggered susceptibility
GO	Gene Ontology
HR	hypersensitive response
HSP	haustoria-secreted protein
IPS	InterProScan
InDel	insertion and deletion
Kbp	kilo base pair
LOH	loss of heterozygosity
Mbp	million base pair
Mli	*Melampsora lini*
Mlp	*Melampsora larici-populina*
NCBI	The National Center for Biotechnology Information
NGS	Next-generation sequencing

Nr database	NCBI Non-redundant database
NSY	non-synonymous
PacBio	Pacific Biosciences
PAMP	pathogen-associated molecular patterns
PCR	polymerase chain reaction
Pgt	*Puccinia graminis* f. sp. *tritici*
Ph	*Puccinia hordei*
PRR	Pattern recognition receptors
Pst	*Puccinia striiformis* f. sp. *tritici*
Pt	*Puccinia triticina*
PTI	PAMP-triggered immunity
R gene	resistance gene
SMRT	single molecule real time
SNP	single nucleotide polymorphism
SNV	single nucleotide variation
SP	signal peptide
SYN	synonymous
TAL	Transcription activator-like
TGS	Third-generation sequencing
UTR	untranslated region

Contents

List of Figures

List of Tables

Chapter 1

General introduction and literature review

1.1 The Hosts

1.1.1 Wheat

Wheat is one of the most important cereal crops in the world. According to FAOSTAT (Food and Agriculture Organization of United Nations), about 760 million tons of wheat was produced in 2016 globally (FAOUN, 2017). Compared to other cereal grains, wheat has high nutritional value in carbohydrate, fiber and protein: 100 grams of hard red winter wheat contains about 327 kcal energy, 71g carbohydrate, 12g protein and 12g dietary fiber (USDA, 2017).

Cultivated wheat species include einkorn (*Triticum monococcum*), durum (*Triticum durum*), emmer (*Triticum dicoccoides*), spelt (*Triticum spelta*) and common wheat (*Triticum aestivum*). Based on DNA fingerprinting and archaeological evidence, wild einkorn wheat was shown to be domesticated in a region known as the Fertile Crescent (Heun et al., 1997). The earliest cultivation of wild emmer wheat to our current knowledge was in southern Leandvant dating back to 9,600 B.C. (Colledge and Conolly, 2007), while the earliest domestication of emmer wheat was believed to be in south east Turkey (Özkan et al., 2002).

As global population continues to increase, it is crucial that agricultural productivity be improved. A great amount of research effort has been directed at studying wheat to improve traits such as yield, disease resistance, drought resistance and tolerance to other biotic and abiotic stresses. The International Maize and Wheat Improvement Center (CIMMYT) located in Mexico is one of the key centers undertaking research on wheat improvement. CIMMYT has been tremendously successful in providing elite seeds to many developing countries. The Borlaug Global Rust Initiative (BGRI) is another key organization, established following the detection of a lineage of stem rust known colloquially as “Ug99” with the aim of protecting wheat from the three rust diseases, *viz.* stem rust, leaf rust and stripe rust.

To better study wheat on a molecular level, the International Wheat Genome Sequencing Consortium was established with the aim of sequencing and characterizing the genome of common wheat, *T. aestivum*. The common wheat genome is hexaploid with 42 chromosomes (AABBDD, 6x=42), and is about 17 Gbp in size (International Wheat Genome Sequencing Consortium, 2014). The hexaploid genome resulted from the hybridization of the tetraploid *T. durum* (AABB) and diploid *Aegilops tauschii* (DD) (Shewry, 2009). About 80% of the wheat genome is highly repetitive transposable elements, which have made the genome assembly extremely complicated. Shotgun sequencing and assembly of flow cell-sorted chromosome DNA identified gene orthologs in the three sub-genomes, but the assembly was highly fragmented (International Wheat Genome Sequencing Consortium, 2014). Recently, a near-complete genome assembly and high quality annotation was generated (Clavijo et al., 2017), providing a valuable reference for faster gene isolation, rapid genetic marker development, and precise breeding to meet the increasing need for more wheat to feed the world.

1.1.2 Barley

Barley (*Hordeum vulgare L.*) is another important cereal crop with large-scale cultivation. According to FAOSTAT, about 136 million tons of barley grain is produced from about 566,000 km^2 worldwide. The cultivation of barley can be traced back to two independent locations: one being the Near East Fertile Crescent, dating back to 8,000 B.C (Nevo, 1992), and another being Tibet (Dai et al., 2012). Genome comparisons of wild barley lines from these two locations suggested that they diverged about 2.76 million years ago.

Compared to wheat, barley is more stress tolerant and is adapted to a wider range of environments, and thus is an important food source in some poorer countries with harsh environmental conditions (Mayer et al., 2012). In more developed countries, barley is an important source of dietary fiber, which is helpful for lowering the risk of some serious diseases including type II diabetes, cardiovascular diseases and colorectal cancers (Mayer et al., 2012). A high quality reference genome (11.6 Gbp) for barley was recently published, which will facilitate gene cloning and comparative genomics for cereal genetics as a community reference (Mascher et al., 2017).

1.2 Rust pathogens

Rust pathogens cause diseases that are characterized by a rusted appearance on the surface of plants, the appearance of which is imparted by either asexual or sexual spores that are produced in pustules. While some studies have succeeded in growing several rust species in axenic culture (Boasson and Shaw, 1982), they are considered obligate biotrophic parasites because growth in culture is extremely poor and they can only reproduce on a living host.

Rust pathogens cause significant yield losses in both wheat and barley. In 1935, about 50% of wheat production was lost in North US due to wheat stem rust caused by fungus *Puccinia graminis* f. sp. *tritici* (*Pgt*) (Leonard, 2001). Dill-Macky et al. (1990) studied the impact of stem rust *Pgt* in barley, and recorded yield losses up to 45%. Cotterill et al. (1992) studied barley leaf rust caused by pathogen *Puccinia hordei* (*Ph*) in Queensland and recorded yield losses of up to 40%. Because of the threat posed by rust diseases, a great deal of effort has been put into breeding wheat and barley cultivars for rust resistance in Australia (Park, 2008).

1.2.1 Life cycles

Many rust pathogens have complex life cycles that include up to five different spore stages. The life cycle of the wheat stem rust pathogen *Pgt* was described in detail by Leonard and Szabo (2005). The pathogen has a heteroecious life cycle that involves two unrelated plant hosts, one primary host and one alternate host. Towards the end of the primary host's growth season, *Pgt* stops producing urediniospores and begins to produce teliospores (Figure 1.1). Teliospores are two-celled, with each cell containing two sets of homologous chromosomes (N + N). Teliospores geminate to produce a basidium, in which meiosis occurs to produce haploid (N) basidiospores. The basidiospores have two mating types, which are often designated as "+" and "-". Basidiospores may be dispersed by a variety of agents (wind, insect and animals) to the alternate host. In the case of *Pgt*, the alternate hosts are in the genus *Berberis* and *Mahonia* (e.g. common barberry *B. vulgaris*). Under suitable conditions, infection on the alternate host takes place and results in flask-shaped pycnia, which produce two different cell types: pycniospores (N) and “flexuous” hyphae (N). Pycniospores of one mating type can fuse with “flexuous” hyphae of the other mating type, and subsequent dikaryons (N + N) grow and differentiate into an aecium that releases aeciospores. The heteroecious life cycle is completed when aeciospores infect the primary host plant. *Pgt* undergoes the asexual part of its life cycle by successful infection of aeciospores, which results in a fungal colony within the host plant tissues and the development of urediniospores in pustules that rupture the epidermis. The urediniospores can be dispersed by wind, and initiate the infection cycle once again.

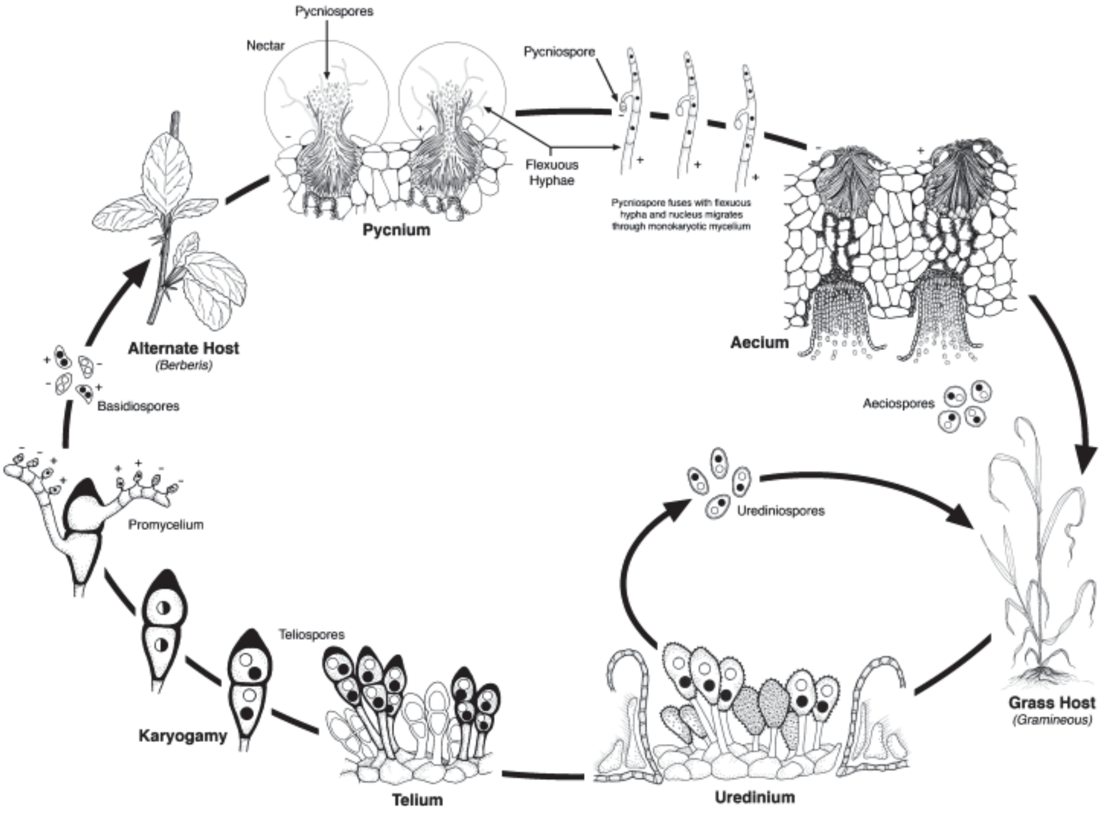


Figure 1.1: Life cycle of *Puccinia graminis* f. sp. *tritici* (Leonard and Szabo, 2005).

The barley leaf rust pathogen *P. hordei* also has a heteroecious life cycle, in which the primary hosts are members of genus *Hordeum*, and alternate hosts in *Liliaceae* family. Anikster et al. (1982) collected teliospores isolated from barleys including *Hordeum spontaneum*, *H. bulbosum*, and *H. murinum*, and inoculated the spores on four

Liliaceae species *Ornithogalum brachystachys, O. trichophyllum, Dipcadi erythraeum,* and *Leopoldia eburnean*. Pycnia and aecia could form on the *Liliaceae* species, aeciospores inoculated on the primary *Hordeum* hosts established successful infections and uredinium reproductions. In the dikaryotic stage, the infection of *P. hordei* shows parasitic specialization. Uredinial spores isolated from *H. vulgare* and *H. vulgare* spp. *spontaneum* can infect only the two host species, while urediniospores isolated from *H. bulbosum* and *H. murinum* infected only their original hosts. On the other hand, parasitic specialization was not observed in the monokaryotic stage in different alternate host species (Anikster, 1989). The author proposed that *P. hordei* had biogenetic host expansion from the dikaryotic host to monokaryotic hosts (Anikster, 1989). In Australia, the only alternate host of *P. hordei* found so far is the Star of Bethlehem (*Omithogalum umbellatum*) (Wallwork et al., 1992). *P. hordei* isolates collected from the alternate host in South Australia were found to comprise different pathotypes, suggesting that the alternate host may contribute to new barley virulence via sexual hybridization (Wallwork et al., 1992).

1.2.2 Infection by rust fungi

Under appropriate moisture and temperature conditions, urediniospores of rust fungi germinate on host plants and produce a germ tube, which elongates to reach open stomata of the plant. Rust infection is host specific; a rust pathogen inoculated onto a non-host plant may not germinate properly or may be unable to locate stomata because of a failure to recognize the topology of leaf surface (Webb and Fellers, 2006). Once the elongating germ tube locates a stoma, the end of the germ tube forms an appressorium, which produces a penetration peg that pushes through the stoma and enters into intercellular space. A substomatal vesicle is then formed, from which infection hyphae grow towards host cells. The tip of the infection hypha forms a cell structure termed the haustorial mother cell, which produces a penetration peg to penetrate the host cell wall and allow the development of haustoria between the cell wall and the plasma membrane (Mendgen et al., 1996; Wiethölter et al., 2003).

1.2.2.1 Haustorium

The haustorium is the main interface between a rust pathogen and its host, and it is the organ for nutrient uptake, an important function for rust pathogen's biotrophic life style (Garnica et al., 2014). The development of a technique for haustorial purification has enabled the study of biochemical activities within haustoria. Hahn et al. (1997) constructed a cDNA library from purified haustoria of *Uromyces fabae* (*Uf*) and provided a catalogue of *in planta* expressed genes, several of which were shown to share sequence similarity with genes functional for nutrient acquisition and transportation. One of those genes, *PIG2*, was expressed exclusively in haustoria and encoded an amino acid transporter, supporting the hypothesis of the nutrient uptake function of haustoria. A later study by Voegele et al. (2001) showed that the gene *HXT1* in *Uf* encodes a protein secreted into the extra-haustorial space for catalyzing the hydrolysis of sucrose, suggesting that haustoria function in sugar uptake. The amino acid transporters AAT1, AAT2 and AAT3 were also found in haustoria, suggesting that haustoria are also involved in amino acid uptake (Mendgen and Hahn, 2002; Struck et al., 2002, 2004). In a more recent genomics study by Duplessis et al. (2011), transcriptomic sequencing of purified haustoria from two rust pathogens, *Melampsora larici-populina* (*Mlp*) and *Pgt*, showed that homologs of *Hxt1*, *ATT1, AAT2* and *AAT3*, and other nutrient transporter genes were expressed *in planta*, suggesting that haustoria in the two pathogens have functions in sugar and amino acid uptake.

1.2.3 Origins of variation in rust fungal populations

Variation in a rust pathogen populations can arise from the introduction of an exotic isolate, simple mutation, or somatic hybridization.

1.2.3.1 Migration

Rust spores can migrate between different regions with the help of a variety of agents (e.g. wind and animals). The wheat stripe rust pathogen *Puccinia striiformis* f. sp.

tritici (*Pst*) was first detected in Australia in 1979, and was believed to have originated from Europe (Wellings, 2007). In another case, a different form of *P. striiformis* that can infect wild barley grass was introduced to Australia in 1998 (Wellings, 2007). A third exotic incursion of *Pst* was detected in Western Australia in 2002, and subsequently spread to Eastern Australia resulting in a major shift in the composition of the *Pst* pathogen population (Wellings, 2007). In at least two of these cases, human mediated transport of urediniospores is believed to have been the means by which the rusts moved to Australia (Wellings, 2007).

1.2.3.2 Somatic hybridization

Somatic hybridization has been documented between isolates of the same or different rust species, and was reviewed by Park and Wellings (2012). One example of interspecific hybridization involved the two rust species *Cronartium ribicola* and *Cronartium comandrae*, which respectively cause white pine blister rust and comandra blister rust (Joly et al., 2006). The researchers analyzed 12 co-dominant polymerase chain reaction (PCR) loci in aecial samples of white pine blister rust, and found some had heterozygous alleles and novel alleles at all 12 loci, one of which was suggested to be from *C. ribicola*, and another from *C. comandrae*. Scanning electron microscopy showed that the morphology of the putative hybrid was intermediate between the two parents (Joly et al., 2006).

In a study of intraspecific somatic hybridization, Nelson et al. (1955) mixed different races of *Pgt* under lab conditions. Races 38 and 56 were mixed and inoculated onto the resistant host Khapli emmer, and two variant isolates virulent on this host were recovered. In later generations of the new variant isolates, virulence on Khapli emmer was gradually lost. Park et al. (1999) documented a spontaneous intraspecific hybridization event in nature within the wheat leaf rust pathogen *P. triticina*. The authors found that wheat leaf rust pathotype 64-(6),(7),(10),11 combined pathogenic and isozymic features only present in pathotypes 104-2,3,(6),(7),11 and 53-1,(6),(7),10,11. Studies of random amplification of polymorphic DNA (RAPD) in 64-(6),(7),(10),11 found three bands only in pathotype 53-1,(6),(7),10,11, and one band only in pathotype

104-2,3,(6),(7),11. These discoveries were consistent with pathotype 64-(6),(7),(10),11 arising from somatic hybridization between 104-2,3,(6),(7),11 and 53-1,(6),(7),10,11. Intraspecific hybridization between different formae speciales of rust pathogens has also been documented. Watson and Luig (1959) mixed urediniospores of isolates of *Pgt* and *P. graminis* f. sp. *secalis* (the rye stem rust fungus), and recovered isolates that combined pathogenic characteristics of the two parental isolates and were considered to be most likely somatic hybrids.

Based on the genetic evidences for somatic hybridization, it was proposed that it involved germ tube anastomosis after urediniospore germination (Burdon and Silk, 1997; Nelson et al., 1955). Wang and McCallum (2009) observed germ tube anastomosis in 27 *Pt* isolates, each of which represented a distinct virulence phenotype. In their study, germ tube fusion bodies, formed at the tips of germ tubes, were affected by the urediniospore mixture density and illumination length during germination. One of their microscopic experiments demonstrated that four nuclei were present in close proximity in the fused bodies resulted from the germ tubes of *Pt* isolates MBDS-3-115 and TBBJ-5-11. This result provided the first evidence for the correlation of germ tube anastomosis and nuclei reassortment. However, it was unknown whether chromosomal rearrangement or exchange between nuclei was involved and lead to somatic recombination. This question remains un-addressed and may be answered with fluorescence *in situ* hybridization in future studies.

1.2.3.3 Mutation

Simple mutation is implicated as the major cause of the evolution in cereal attacking rust pathogens in situations where sexual recombination is rare or absent. The extent to which mutation contributes to the development of variation in traits such as virulence depends upon the rate of mutation, the ploidy of the pathogen, the size of the pathogen population, and the selective advantage conferred by the mutant genotype (Burdon and Silk, 1997). Steele et al. (2001) analyzed 18 Australasian isolates of *Pst*, which had different virulence patterns that had been attributed to stepwise mutations. Random amplified polymorphic DNAs (RAPDs) were produced from the isolates, and amplified fragment length polymorphisms (AFLPs) were analyzed. No molecular variation was detected with either RAPDs or AFLPs, supporting the hypothesis that the isolates arose via simple mutation.

Hovmøller and Justesen (2007) estimated rates of evolution in three lineages of *Pst* in northwest Europe, which were believed to represent asexually reproducing populations. Association of 14 avirulence/virulence alleles and AFLPs from 17,000 AFLP fragments was studied. The pattern in gain and loss of AFLP fragments led the authors to estimate a mutation rate from one in 1.4 million to 4.1 million per locus per generation in each clonal lineage. The effective mutation rate from avirulence to virulence was estimated to be three magnitudes lower (Hovmøller and Justesen, 2007).

1.3 Plant immunity

Despite the evolution of sophisticated pathogenesis mechanisms in rust pathogens, plants have evolved resistance to pathogens, often in the form of a hypersensitive response (HR) characterized by localized cell death that arrests the growth of an invading pathogen Jones and Dangl (2006). HR is triggered by recognition of infecting pathogens, typically involving an avirulence gene from the pathogen and a corresponding resistance gene from the host plant.

1.3.1 The gene-for-gene hypothesis

The gene-for-gene relationship was first formulated by Flor to explain inheritance of avirulence factors in flax rust fungus, *Melampsora lini* (Flor, 1971). In crosses of different races of *M. lini*, pathogenicity for specific resistance genes segregated into a 3:1 (avirulence/virulence) ratio in the F_2 progeny (Flor, 1946), indicating that the avirulence of the pathogen was determined by a single dominant gene known as avirulence gene (Avr gene). In the study of host resistance to *M. lini*, Flor (1947) crossed resistant flax and susceptible flax, and a 3:1 (resistant/susceptible) segregation ratio was found in the F_2 progeny when inoculated with the same race. These results indicated that resistance in the flax line was controlled by a dominant resistance gene (R gene). With further genetic studies of flax and flax rust interaction, Flor proposed a gene-for-gene hypothesis: "for every gene in the plant that confers resistance, there is a corresponding gene in the pathogen that confers avirulence" (Flor, 1971). For example, the flax variety Dakota has resistance genotype *MM*, and the Dakota-virulent rust Race 22 has pathogenicity genotype $a_m a_m$. In contrast, flax rust Race 1 has pathogenicity genotype $a_m A_m$ and thus is avirulent on Dakota. Crossing of Race 22 and 1 resulted in a F_1 culture A with pathogenicity genotype $a_m A_m$, which was shown to be avirulent on Dakota (Flor, 1946). Similar gene-for-gene interactions have been observed in many other plant-pathogen interactions including *Triticum-Pgt, Avenae-Ustilago avenae, Hordeum-Erysiphe graminis*, etc (Flor, 1971). A biochemical explanation of the theory is a receptor-ligand model in which the plant activates immunity based on R-gene-mediated recognition of Avr gene products secreted from the pathogen (Van Der Biezen and Jones, 1998). In this model, Avr genes encode effector proteins that can enter a plant, overcoming host resistance and initiating disease.

1.3.2 Molecular components in the host-pathogen interactions

As more molecular data have accumulated, the understanding of molecular interactions between microbial pathogens and plants has become clearer. Jones and Dangl (2006) described two branches of the plant immune system, one of which uses trans-membrane pattern recognition receptors (PRR) that recognize conserved pathogen-associated molecular patterns (PAMP) and then activate what is known as PAMP-triggered immunity (PTI). Another branch of the plant immune system employs proteins with conserved nucleotide binding (NB) and leucine rich repeat (LLR) domains to recognize effector proteins secreted by pathogens inside host plants, a process known as effector-triggered immunity (ETI). The recognitions of effectors by NB-LRR receptors can be in either direct or indirect mechanisms (Dodds and Rathjen, 2010). In the direct recognitions, the effectors bind to the receptors in physical associations, most of which can be demonstrated by yeast two-hybrid assays. In the indirect recognitions, the effectors modify accessory proteins and these modifications are recognized by the NB-LRR proteins. The accessory proteins may be virulence targets of the effectors or target mimics as decoys to the effectors, in a “guard model” or a “decoy model”, respectively. The interaction of NB-LRR proteins and effectors underlies the gene-for-gene interaction, and thus ETI was adopted to replace the traditional term ‘gene-for-gene resistance’ (Gassmann and Bhattacharjee, 2012).

Jones and Dangl (2006) further proposed a four phase “zigzag” model to show that the “zigzag” resistance outcome of plants was determined by the interaction of host and pathogen molecules (Figure 1.2). In the first phase, host PPRs recognize pathogen PAMPs and trigger PTI. But in phase two, pathogen effectors could interfere with the PTI, resulting in host susceptibility or effector-triggered susceptibility (ETS). In the third phase, host NB-LRR resistance proteins recognize some of the effectors in the process of ETI, which leads to localized cell death or HR to stop pathogen colonization. In the final phase, under selection pressure to avoid ETI, pathogens either discard or diversify the recognized effector genes to evade host recognition, and successful pathogens will be able to dampen the host immunity again.

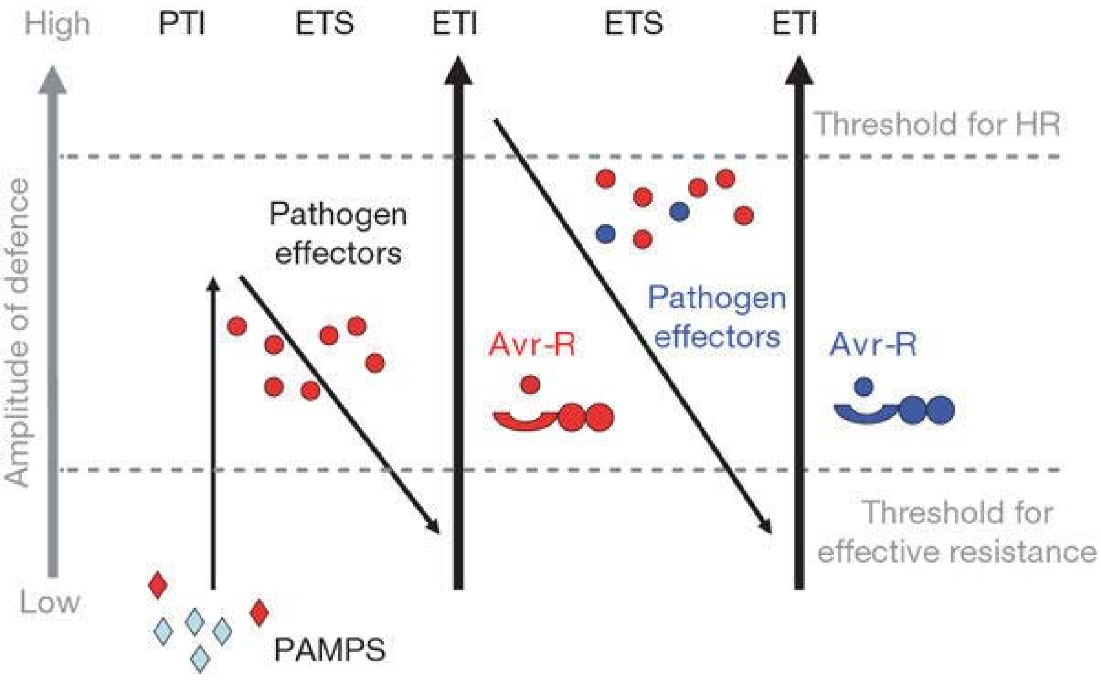


Figure 1.2: The zigzag model of molecular interactions in plant-pathogens which have additive effect on host immunity outcome (Jones and Dangl, 2006).

1.4 Effector biology

Microbial plant pathogens (including bacteria, oomycetes, and fungi) secrete effectors that target a variety of cellular components in plants such as plasma membranes, chloroplasts, and nuclear components (Bozkurt et al., 2012; Deslandes and Rivas, 2012; Dodds and Rathjen, 2010). For instance, transcription activator-like (TAL) effectors target host genes by binding to promoters of these genes, activating their transcription for pathogenesis. The TAL effector AvrBs3, from the bacterium *Xanthomonas*, was found to bind the promoter of gene *upa20* in pepper, and reprogrammed host cell development (Kay et al., 2007). Römer et al. (2007) showed that AvrBs3 also acted as a transcription factor for another gene *Bs3*, the induction of which resulted in a R gene-mediated HR. In another case, effector HopN1, a cysteine protease protein from the bacterial pathogen *Pseudomonas syringae*, could target tomato protein PsbQ in chloroplasts, with the proposed function of suppressing host reactive oxygen species and callose deposition (Rodríguez-Herva et al., 2012). This effector was also shown to suppress host *Arabidopsis* and non-host *Nicotiana tabacum* cell death, possibly favoring pathogen survival and reproduction (López-Solanilla et al., 2004). Another effector HopI1 from the same pathogen *P. syringae* bound to transcripts of its target gene Hsp70s via a J domain and stimulated Hsp70 ATP hydrolysis activity (Jelenska et al., 2010).

In the oomycete pathogen *Phytophthora sojae*, which causes root rot of soybean, GIP1 and GIP2 are glucanase inhibitor proteins secreted into extracellular spaces where they inhibit host endoglucanase activity presumably to prevent degradation of glucan in the pathogen cell wall (Rose et al., 2002). Another extracellular effector is protease inhibitor EPI1 from *Phytophthora infestans* that was shown to inhibit serine protease P69B in its host tomato, probably to execute a counter defense function (Tian et al., 2004). In another study, two cystatin-like proteins EPIC1 and EPIC2B, active in the host apoplastic space, were found secreted by *P. infestans* to target host tomato cysteine protease protein C14, of which silencing will increase host susceptibility to *P. infestans* (Tian et al., 2007).

In the fungal pathogen *Ustilago maydis* that causes maize smut disease, Pep1 inhibits a host peroxidase POX12, which is a key component in maize reactive oxygen species generating system (Hemetsberger et al., 2012). Pit2, another secreted effector from the same pathogen, inhibits host apoplastic cysteine proteases and interfere with their salicylic acid-associated plant defenses (Mueller et al., 2013). *U. maydis* also secretes chorismate mutase Cmu1 to counteract maize salicylic acid-induced immune activities (Djamei et al., 2011). This is achieved in the cytoplasm of plant cells where Cmu1 reduces the abundance of chorismate, a precursor for the synthesis of salicylic acid.

1.4.1 Rust fungi effectors

Compared with other microbial pathogens, many biochemical activities of rust fungal effectors are still undetermined. Effector cloning in rust fungi is challenging, mainly because they are obligate biotrophs that are difficult to culture *in vitro* (Petre et al., 2014). Up until now, only eight effector proteins from three species of rust fungi have been identified: RPT1 in the bean rust fungus *U. fabae*, PGTAUSPE-10-1 in the wheat stem rust fungus *Pgt*, and AvrM, AvrL567, AvrP123, AvrP4, AvrL2 and AvrM14 in the flax rust fungus *M. lini* (Anderson et al., 2016; Catanzariti et al., 2006; Dodds et al., 2004; Kemen et al., 2005; Upadhyaya et al., 2014). All of the reported rust effectors except RPT1 were identified by their ability to trigger HR responses, but their pathogenic functions remain unknown.

Studies of *U. fabae* by Kemen et al. (2005) provided first direct evidence of effector protein transfer from rust haustoria to the host. The protein RTP1p of *U. fabae* was detected by immunofluorescence and electron microscopy in the extra-haustorial matrix (ETM) at the initial stage of infection, and was shown to be subsequently transferred into the host cell cytoplasm.

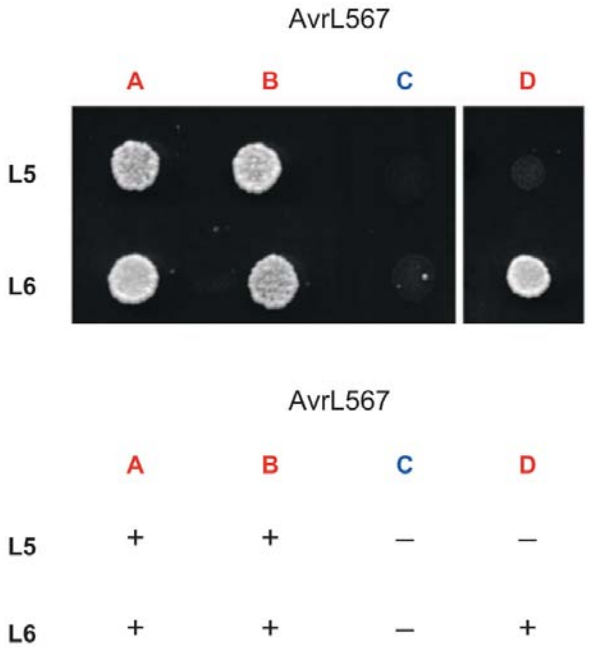
The *AvrL567* genes were expressed in haustoria and their proteins were recognized inside plant cells, suggesting the Avr gene product was translocated from haustoria into host cells (Dodds et al., 2004). Later, a study of a haustorium-specific cDNA library showed that effectors are enriched in haustoria (Catanzariti et al., 2006). Among 429 unigenes recovered from the cDNA library, 21 encoded a signal peptide and included four Avr genes: *AvrL567*, *AvrM, AvrP123,* and *AvrP4*. The mRNA of *AvrL567* was found in infected leaf tissue and in the haustorial cDNA library, but was not detected in un-germinated spores or spores that had been germinated on water (Ellis et al., 2007). These studies indicated that Avr gene products should be enriched in haustoria and that a haustorial-specific cDNA library should be a good resource to identify new Avr genes. This approach was adopted in several later studies that sequenced the transcriptome from purified haustoria for an initial data set to search for effector genes (Cantu et al., 2013; Garnica et al., 2013; Upadhyaya et al., 2015).

Dodds et al. (2004) found elevated polymorphism in Avr gene variants *AvrL567-A, B* and *C*, presumably the result of high selection pressure to avoid R gene recognition. A total of 30 nucleotide differences were found in the 450 bp coding sequences of the three variants *AvrL567-A, B* and *C* (Dodds et al., 2004). In a sharp contrast, only 25 nucleotide differences were found in the 6907 bp regions surrounding the gene variants. The hypothesis was further confirmed by structure analysis of the AvrL567-A and AvrL567-B proteins, which identified that the protein polymorphisms of the two variants mapped to surface residues where interactions with R protein occurred (Wang et al., 2007). For example, a change of amino acid I to T at the position 50 of AvrL567-A abolished the recognition by L5 and L6 proteins.

In a later study, genomic clones containing the *AvrL567* loci were sequenced for seven flax rust strains, namely CH5, H, C, I, BsI, Fi and 339 (Dodds et al., 2006). A total of 12 variants of *AvrL567* were identified and named as *AvrL567-A, -B, -C, -D, –, -L*. In order to test the avirulence function of each variant, they were cloned into *Agrobacterium* and delivered into flax leaves via infiltration. Leaf HR was observed on flax lines with the R genes *L5, L6* or *L7*, but not on flax lines lacking these R genes. In terms of the Avr genes, seven variants have avirulence function (inducing HR) in flax plants, while the other five do not induce HR. The avirulence shows specificity in a gene-for-gene manner. For example, variant *AvrL567*-D triggers obvious HR in a flax line containing the resistance gene *L6*, but induces a weak response only in flax lines with the resistance genes *L5* or *L7*, suggesting *AvrL567*-D is specific to *L6*. *AvrL567*-A can trigger HR in flax with *L5, L6* or *L7*. Therefore, the avirulent variants can be differentiated based on their interaction with the R genes.

The avirulence specificity of *AvrL567* to R genes *L5, L6* and *L7* was shown to be determined by direct interactions of the Avr-R proteins (Dodds et al., 2006). The

AvrL567 gene copies and R genes were co-expressed in a Yeast two-hybrid assay system to test their interaction specificity. If an Avr protein and an R protein interact directly, a reporter gene expresses and enables yeast growth, indicating direct interaction. The results showed AvrL567-A and AvrL567-B have direct interaction with both L5 and L6 proteins, AvrL567-C has no direct interaction with both L5 and L6 proteins, and AvrL567-D has direct interaction with the L6 protein but not with the L5 protein (Figure 1.3 upper panel). The direct interaction specificity is consistent with the recognition specificity of the *Avr567* variants by flax resistance genes (Figure 1.3 lower panel).



AvrL567

	A	B	C	D
L5	+	+	–	–
L6	+	+	–	+

Figure 1.3: Interaction specificity of R-Avr proteins in the flax-flax rust pathosystem (figure adapted from Ellis et al., 2007). The upper panel shows direct interactions of AvrL567-A, B, C, D with R protein L5, L6 in a Yeast-two hybrid system. The lower panel shows presence (+) or absence (-) of HR after transient expression of *AvrL567*-A, B, C, D in flax containing the R gene *L5* or *L6*.

1.5 Pathogen genetics and genomics

1.5.1 Diversity study

As rust pathogens have a rapid evolutionary ability, high genetic diversity has been observed in the environment (Hovmøller and Justesen, 2007; Hubbard et al., 2015; Karaoglu and Park, 2014). Traditional approaches to study genetic structure in rust populations have been based on PCR-based molecular tools and pathotype profiling. PCR-based investigations are limited and biased towards targeted regions, and thus provide low resolution on the genomic space. Keiper et al. (2003) studied the potential of three PCR-based tools, AFLP, selectively amplified microsatellites, and sequence-specific amplification polymorphisms, to distinguish genetic relationships of rust isolates of *Pgt*, *P. triticina, Pst, P. graminis* f. sp. *avenae, P. striiformis* f. sp. *pseudohordei*. Three of the tools could discriminate the five different taxa, but they differed in the amount of variation detected among isolates within a specific taxon. In a study of population structure of *P. recondita* f. sp. *tritici* (syn. *P. triticina*) in Western Europe with RAPD markers, only 18 RAPD phenotypes were identified among 61 isolates, while 35 pathotypes were identified for the same isolates based on their infection types on 24 differential lines (Park et al., 2000). These two studies provided insight into the genetic structure of rust pathogen populations, but were limited by the technologies used.

Application of whole genome or transcriptome sequencing has provided higher resolution for diversity studies, with the ability to capture millions of nucleotides rather than hundreds of markers. Hubbard et al. (2015) sequenced the transcriptomes of 39 *Pst*-infected leaf samples collected from the field. The Illumina sequencing reads were aligned to a reference genome built from isolate PST-130. Based on the read alignments, phylogenetic relationships were analyzed for the 39 isolates, which were clustered into four clades. The result was consistent with the four phenotypic groups determined based on virulence profiles of the isolates. Rust population studies using whole genome sequencing include the studies by Cantu et al. (2013); Upadhyaya et al. (2015); Wu et al. (2017); Zheng et al. (2013). These high-resolution genotypic data shed light on genetic substructure within field populations, which in turn provides significant surveillance information for agriculture.

1.5.2 Evolutionary history

Obligate biotrophic pathogens, including rust fungi, have evolved a parasitic life style that is characterized by nutrient uptake from the host and sustained suppression of the host immune response. The evolution of this life style is reflected at the genome level, and has been studied to gain insights into its molecular basis. Spanu et al. (2010) characterised the genome of *Blumeria graminis* f. sp. *hordei* (*Bgh*), an obligate biotrophic pathogen responsible for powdery mildew disease of barley, and found that genome size expansion resulted from a massive proliferation of transposable elements in contrast with reduction of genes for inorganic nitrate and sulfur assimilation. Genome expansion was also attributed to diversifying evolution of putative effector genes that were specific to *Bgh* and shared no homologs in two other closely related powdery mildew fungi *Erysiphe pisi* and *Golovinomyces orontii*. In a more recent study, obligate biotrophic features (including impaired nitrate and sulfur assimilation pathways) were revealed again in the genomes of the poplar rust pathogen *Mlp* and the wheat stem rust fungus *Pgt* (Duplessis et al., 2011). In comparing the two genomes with other necrotrophic species in the Basidiomycota phylum, a total gain of 774 gene families and loss of 424 gene families were found. About 1,000 gained genes encoded small secreted proteins, which were considered to be potential effectors related to pathogenesis. Several classic genes typically found in saprotrophic basidiomycetes responsible for nitrate and sulfate assimilation were missing in the genomes, contrasting the two biotrophic pathogens with saprotrophic fungi.

In another comparative genomics study, genome differences of the powdery mildew pathogens *Blumeria graminis* f. sp. *tritici* (*Bgt*, wheat attacking) and *Bgh* were used as a molecular clock, giving an estimate of the time of divergence of the two pathogenic forms (namely *Bgt* and *Bgh*) of 6.3 (±1.1) million years ago (Wicker et al., 2013). In comparing gene sequences of the two formae speciales, the non-synonymous-to-synonymous substitution ratios (dN/dS) of genes provided a measure of selection pressure imposed on the genes. The average dN/dS of 437 predicted effector genes of 0.8 was much larger than the dN/dS of 5,258 non-effector genes (majority less than 0.5, average 0.24), indicating that the predicted effector genes may be under diversifying selection for adaption to different hosts. To further study *Bgt* evolution, the authors sequenced isolate JIW2 collected in England, isolate 70 collected in Israel, and two isolates (isolate 96224 and 94202) collected in Switzerland. In comparing their genomes, some regions had much higher SNP frequencies (>10 fold) than other regions, suggesting that the genomes of the isolates are recombinational mosaics of different haplogroups.

1.5.3 Effector identification

Under positive selection, avirulence effectors evolve quickly to avoid recognition by R genes, giving rise to pathogens with new virulence. Therefore, effector identification and monitoring effector reservoirs in pathogen populations are important topics in agricultural research.

The *AvrL567* genes from flax rust were isolated via a map-based cloning approach (Dodds et al., 2004). In detail, rust strain H (*L5, L6* and *L7* avirulent) was crossed with rust strain C (*L5, L6* and *L7* virulent) to produce a hybrid strain CH5, and selfing of CH5 produced 74 F_2 individuals. In a cDNA library of flax leaf tissues infected by an avirulent F_2 individual, a cDNA probe IU2F2 detected a restriction fragment length polymorphism (RFLP) that co-segregated with the *AvrL567* locus in the progenitor strain H. A 25.6 Kbp genomic contig containing the marker IU2F2 was obtained from rust strain CH5, and its RFLPs corresponded to the H derived IU2F2 RFLP locus, suggesting that the contig contained the avirulence allele of *AvrL567* from rust strain H. The virulence allele of *AvrL567* derived from rust strain C was also amplified and sequenced to an 11.5 Kbp contig. Sequence analysis of the H contig and the C contig revealed a DNA segment with two homologous copies in the H contig, and one homologous copy in the C contig. The segment contained three genes, one encoding a protein with a signal peptide, one encoding a protein related to yeast protein Sec14 with membrane fusion function, and the third gene had no homologue in public databases.

Large protein sequence variation was found for the first signal peptide-encoding gene, while little sequence variations were found for the homologs of the other two genes. Transient expression of the first gene in flax lines containing *L5, L6* or *L7* triggered HR, confirming it is *AvrL567*.

The traditional map-based cloning method has been used to clone most rust effectors up until now. However, it is labor intensive and time consuming. With the advances of Next-generation sequencing (NGS), an increasing number of genome sequences of plant pathogens, including rust fungi, are becoming available, providing opportunities to identify effectors *in silico*.

1.5.3.1 Protein features

Signal peptide Effector proteins must be secreted from pathogens into the host in order to target host immunity components. Secreted proteins in oomycete and fungal pathogens are exported via a secretory pathway based on a short N-terminal signal peptide (Sonah et al., 2016). Therefore, the presence of a signal peptide can be used as a feature to identify a protein as a candidate effector. However, no conserved sequence patterns were shown in signal peptides of identified effectors, and it is thus impractical to predict signal peptides of unknown proteins based on sequence similarity searches. Algorithms for predicting signal peptides and cleavage sites have been developed and have proven to be highly sensitivity and accurate (Menne et al., 2000).

Transmembrane domain Proteins with signal peptides are not necessarily secreted outside of cells, as some of them may be anchored in membranes. To avoid reporting a transmembrane protein as a secreted protein, it is necessary to predict transmembrane domains in candidate secreted proteins. However, distinguishing the two motifs is difficult, as both of them have hydrophobic segments (Sonah et al., 2016). TMHMM is a software program developed based on a hidden Markov model to predict transmembrane domains, which has been used in several effector prediction pipelines (Cantu et al., 2013; Krogh et al., 2001; Saunders et al., 2012). Tools that combine algorithms for transmembrane domain and signal peptide prediction are also available (e.g. ProtComp, Phobius, SignalP and SPOCTOPUS; Clark et al., 2003; Käll et al., 2004; Petersen et al., 2011; Viklund et al., 2008).

Conserved domains The cloning of four Avr genes (namely *ATR1NdWsB*, *ATR13*, *Avr1b-1* and *Avr3a*) in oomycete pathogens led to the discovery of two conserved motifs, RXLR and dEER, at the N terminals (Kamoun, 2006). The motifs were later confirmed necessary for the translocation of effectors into host plants (Whisson et al., 2007). The empirical motifs have provided a useful guide to discovering over 700 candidate RXLR effectors in the genomes of *P. infestans* and *P. ramorum* (Jiang et al., 2008). [YFW]xC is a conserved motif identified in many small secreted proteins from haustoria-forming fungi *Pgt*, *Pt* and *Blumeria graminis* f.sp. *hordei* (*Bgh*) (Godfrey et al., 2010). Within Bgh, the [YFW]xC containing proteins show poor sequence similarity but a conserved exon-intron structure, indicating a single remote origin in the powdery mildew fungi (Godfrey et al., 2010). As more candidate effectors are predicted from increasing numbers of fungal pathogen genomes, more conserved domains may be identified and contribute to finding new effectors.

Other criteria One approach to predict fungal effectors is protein sequence analysis based on various features already found in known effectors. Currently, only eight rust effectors have been characterized, and thus it is difficult to identify general criteria for computational prediction of rust fungal effectors. Despite a lack of general features in rust effectors, several computational pipelines for effector mining have been developed and reported. Saunders et al. (2012) designed a scoring method to calculate the probability of a protein being an effector. The method filters proteins that contain a secretion signal and no transmembrane domain, with length shorter than 150 amino acids, containing at least 3% cysteine, and without an annotated domain in PFAM database. Sperschneider et al. (2014) tried an alternative strategy in which diversifying selection imposed on effector genes is emphasized. This strategy successfully predicted a positive control effector gene *PGTAUSPE-10-1* as an effector, consistent with the study by Upadhyaya et al. (2014) that showed the gene product can cause HR in wheat containing gene *Sr22*.

More recently, a machine learning based *in silico* effector prediction software EffectorP was developed (Sperschneider et al., 2015). The software was trained to learn protein features including AA frequencies, AA length, molecular weight, and protein net charge. A total of 58 fungal effectors validated experimentally from 16 fungal species were used as a positive set to train the software. A set of randomly selected secreted proteins was used as negative training data. The trained EffectorP achieved sensitivity of 84.5% and specificity of 82.8% in predicting new effectors. However, the lack of a large training set of known rust fungal effectors may introduce bias in the machine learning, and result in prediction biased towards protein features only present in the known effectors.

1.5.3.2 Comparative genomics

Besides considering solely genomic features, researchers have also shown successful effector identification by comparative genomics. According to the gene-for-gene model (Flor, 1971), pathogenic strains virulent to a host containing a specific R gene should lack the cognate Avr gene. Comparing the genomes of virulent strains and avirulent strains has the potential to find absence/presence, or sequence differences in the Avr genes. Schmidt et al. (2016) compared genome sequencing data of 10 strains of *Fusarium oxysporum* f. sp. *melonis*, a melon root infecting pathogen. Some of the sequenced strains were of race 1 because of their ability to infect hosts containing R genes *Fom-1*, the other strains were of race 2 because they are virulent to hosts with *Fom-2*. In mapping sequencing data to the reference genome, 11 genomic sub-regions were absent in all race 2 isolates, but present in all race 1 isolates. One open reading frame encoding a signal peptide was found in the regions and thus considered to be a candidate for *AvrFom2*. Transformation of the candidate *AvrFom2* to race 2 isolates and inoculation of the transformed strains to melon containing R gene *Fom-2* showed no loss of fresh weight (loss of virulence) after 10 days' growth, confirming the candidate as *AvrFom2*. Absence/presence polymorphism of Avr gene in virulent/avirulent strains was shown useful in another study for the identification of *Ave1*, an Avr gene in *Verticillium dahliae* responsible for vascular wilt of tomato (de Jonge et al., 2012).

Protein-coding polymorphisms of Avr genes should also be considered in searching for Avr genes using a comparative genomics approach. In the identification of *Avr5* from *Cladosporium fulvum* (causing tomato leaf mold), the gene was present in both virulent and avirulent strains (Mesarich et al., 2014). Comparison of transcriptome sequencing data of strain OWU (carrying *Avr5*) with that of strain IPO 1979 (lacking functional *Avr5*) revealed two polymorphic genes in a set of 44 candidates. One of the two genes, which contained a 2-bp insertion/deletion between the two strains, conferred avirulence in strain IPO 1979 after transformation into the strain, validating the gene to be *Avr5*.

1.6 Study aims and significance

Prior to the availability of NGS, most rust effectors were identified with sexual crossing and map-based cloning (Catanzariti et al., 2006; Dodds et al., 2004, 2006). However, some rust pathogens lack a known sexual cycle or sexual hosts are not available for study. Proteomics of fungal infection structures also showed potential in effector identification (Song et al., 2011). However, this method is not able to separate alleles of Avr genes and cannot validate Avr function. On the other hand, protein feature analysis combined with comparative genomics provide great advantages in Avr effector identification, especially given that sequencing costs keep dropping and more fungal pathogen genome assemblies are becoming available.

Globalization and industrialization of agroecosystems may increase the frequency of rust diseases, which have caused crop yield losses in the order of millions of dollars. The identification and characterization of Avr genes is fundamental for estimating resistance durability in the field, as changes in these genes can lead to resistance breakdown. Monitoring rust isolates in agroecosystems, especially with respect to their Avr genes, will provide critical information to assess propensity of virulence gain and inform procedures to stop emerging rust epidemics. In addition, it will provide a better understanding of pathogenesis mechanism in the rust-crop pathosystems, a necessity for novel resistance development.

This thesis aims to identify Avr effectors in *Pgt* and *Ph*, causal agents for wheat stem rust and barley leaf rust, respectively. Chapter 2 will compare the genomes of two *Pgt* isolates differing on virulence to one single R gene *Sr50*, targeting identification of the corresponding Avr gene *AvrSr50*. Chapter 3 will assemble the first reference genome for *Ph*, and annotated the gene repertoire with a focus on candidate effectors. Based on these results, Chapter 4 will compare the genomes of five *Ph* isolates derived from a clonal lineage, with a purpose to search for three Avr genes that were lost via single step mutations within the lineage.

References

Anderson, C., Khan, M. A., Catanzariti, A.-M., Jack, C. A., Nemri, A., Lawrence, G. J., et al. (2016). Genome analysis and avirulence gene cloning using a high-density RADseq linkage map of the flax rust fungus, *Melampsora lini*. *BMC Genomics* 17, 667 14

Anikster, Y. (1989). Host specificity versus plurivority in barley leaf rusts and their microcyclic relatives. *Mycological Research* 93, 175–181 6

Anikster, Y. et al. (1982). Alternate hosts of *Puccinia hordei*. *Phytopathology* 72, 733–735 5

Boasson, R. and Shaw, M. (1982). A stimulus for sporulation of axenically grown flax rust. *Experimental Mycology* 6, 1–6 4

Bozkurt, T. O., Schornack, S., Banfield, M. J., and Kamoun, S. (2012). Oomycetes, effectors, and all that jazz. *Current Opinion in Plant Biology* 15, 483–492 13

Burdon, J. and Silk, J. (1997). Sources and patterns of diversity in plant-pathogenic fungi. *Phytopathology* 87, 664–669 9

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., et al. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics* 14, 270 15, 17, 20

Catanzariti, A.-M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., and Ellis, J. G. (2006). Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *The Plant Cell* 18, 243–256 14, 23

Clark, H. F., Gurney, A. L., Abaya, E., Baker, K., Baldwin, D., Brush, J., et al. (2003). The secreted protein discovery initiative (SPDI), a large-scale effort to identify novel human secreted and transmembrane proteins: a bioinformatics assessment. *Genome Research* 13, 2265–2270 21

Clavijo, B. J., Venturini, L., Schudoma, C., Accinelli, G. G., Kaithakottil, G., Wright, J., et al. (2017). An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. *Genome Research* 885, 885–896 3

Colledge, S. and Conolly, J. (2007). *The origins and spread of domestic plants in Southwest Asia and Europe* (Left Coast Press: Walnut Creek, CA) 2

Cotterill, P., Rees, R., Platz, G., and Dill-Macky, R. (1992). Effects of leaf rust on selected Australian barleys. *Australian Journal of Experimental Agriculture* 32, 747–751 4

Dai, F., Nevo, E., Wu, D., Comadran, J., Zhou, M., Qiu, L., et al. (2012). Tibet is one of the centers of domestication of cultivated barley. *Proceedings of the National Academy of Sciences U.S.A.* 109, 16969–16973 3

de Jonge, R., van Esse, H. P., Maruthachalam, K., Bolton, M. D., Santhanam, P., Saber, M. K., et al. (2012). Tomato immune receptor *Ve1* recognizes effector of multiple fungal pathogens uncovered by genome and RNA sequencing. *Proceedings of the National Academy of Sciences U.S.A.* 109, 5110–5115 22

Deslandes, L. and Rivas, S. (2012). Catch me if you can: bacterial effectors and plant targets. *Nature Reviews* 17, 644–655 13

Dill-Macky, R., Rees, R., and Platz, G. (1990). Stem rust epidemics and their effects on grain yield and quality in Australian barley cultivars. *Australian Journal of Agricultural Research* 41, 1057–1063 4

Djamei, A., Schipper, K., Rabe, F., Ghosh, A., Vincon, V., Kahnt, J., et al. (2011). Metabolic priming by a secreted fungal effector. *Nature* 478, 395–398 14

Dodds, P. N., Lawrence, G. J., Catanzariti, A.-M., Ayliffe, M. A., and Ellis, J. G. (2004). The *Melampsora lini AvrL567* avirulence genes are expressed in haustoria and their products are recognized inside plant cells. *The Plant Cell* 16, 755–768 14, 15, 19, 23

Dodds, P. N., Lawrence, G. J., Catanzariti, A.-M., Teh, T., Wang, C.-I., Ayliffe, M. A., et al. (2006). Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proceedings of the National Academy of Sciences U.S.A.* 103, 8888–8893 15, 23

Dodds, P. N. and Rathjen, J. P. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. *Nature Reviews. Genetics* 11, 539 11, 13

Duplessis, S., Cuomo, C. A., Lin, Y.-C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proceedings of the National Academy of Sciences U.S.A.* 108, 9166–9171. doi:10.1073/pnas.1019315108 7, 18

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 xiii, 15, 16

FAOUN (2017). World food situation. [http://www.fao.org; accessed 9-August-2017] 2

Flor, H. (1946). Genetics of pathogenicity in *Melampsora lini*. *Journal of Agricultural Research* 73, 335–357 10, 11

Flor, H. (1947). Inheritance of reaction to rust in flax. *Journal of Agricultural Research* 74, 241–262 10

Flor, H. H. (1971). Current status of the gene-for-gene concept. *Annual Review of Phytopathology* 9, 275–296 10, 11, 22

Garnica, D. P., Nemri, A., Upadhyaya, N. M., Rathjen, J. P., and Dodds, P. N. (2014). The ins and outs of rust haustoria. *PLoS Pathogens* 10, e1004329 7

Garnica, D. P., Upadhyaya, N. M., Dodds, P. N., and Rathjen, J. P. (2013). Strategies for wheat stripe rust pathogenicity identified by transcriptome sequencing. *PloS One* 8, e67150. doi:10.1371/journal.pone.0067150 15

Gassmann, W. and Bhattacharjee, S. (2012). Effector-triggered immunity signaling: from gene-for-gene pathways to protein-protein interaction networks. *Molecular Plant-Microbe Interactions* 25, 862–868 12

Godfrey, D., Böhlenius, H., Pedersen, C., Zhang, Z., Emmersen, J., and Thordal-Christensen, H. (2010). Powdery mildew fungal effector candidates share n-terminal y/f/wxc-motif. *BMC Genomics* 11, 317 21

Hahn, M., Neef, U., Struck, C., Göttfert, M., and Mendgen, K. (1997). A putative amino acid transporter is specifically expressed in haustoria of the rust fungus *Uromyces fabae*. *Molecular Plant-Microbe Interactions* 10, 438–445 7

Hemetsberger, C., Herrberger, C., Zechmann, B., Hillmer, M., and Doehlemann, G. (2012). The *Ustilago maydis* effector pep1 suppresses plant immunity by inhibition of host peroxidase activity. *PLoS Pathogens* 8, e1002684 14

Heun, M., Schäfer-Pregl, R., Klawan, D., Castagna, R., Accerbi, M., Borghi, B., et al. (1997). Site of einkorn wheat domestication identified by DNA fingerprinting. *Science* 278, 1312–1314 2

Hovmøller, M. S. and Justesen, A. F. (2007). Rates of evolution of avirulence phenotypes and DNA markers in a northwest European population of *Puccinia striiformis* f. sp. *tritici*. *Molecular Ecology* 16, 4637–4647 10, 17

Hubbard, A., Lewis, C. M., Yoshida, K., Ramirez-Gonzalez, R. H., de Vallavieille-Pope, C., Thomas, J., et al. (2015). Field pathogenomics reveals the emergence of a diverse wheat yellow rust population. *Genome Biology* 16, 23 17

International Wheat Genome Sequencing Consortium (2014). A chromosome-based draft sequence of the hexaploid bread wheat (*Triticum aestivum*) genome. *Science* 345, 1251788 3

Jelenska, J., Van Hal, J. A., and Greenberg, J. T. (2010). *Pseudomonas syringae* hijacks plant stress chaperone machinery for virulence. *Proceedings of the National Academy of Sciences U.S.A.* 107, 13177–13182 13

Jiang, R. H., Tripathy, S., Govers, F., and Tyler, B. M. (2008). RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. *Proceedings of the National Academy of Sciences U.S.A.* 105, 4874–4879 21

Joly, D., Langor, D., and Hamelin, R. (2006). Molecular and morphological evidence for interspecific hybridization between *Cronartium ribicola* and *C. comandrae* on *Pinus flexilis* in southwestern Alberta. *Plant Disease* 90, 1552–1552 8

Jones, J. D. and Dangl, J. L. (2006). The plant immune system. *Nature* 444, 323 xiii, 10, 11, 12

Käll, L., Krogh, A., and Sonnhammer, E. L. (2004). A combined transmembrane topology and signal peptide prediction method. *Journal of Molecular Biology* 338, 1027–1036 21

Kamoun, S. (2006). A catalogue of the effector secretome of plant pathogenic oomycetes. *Annual Review of Phytopathology* 44, 41–46 21

Karaoglu, H. and Park, R. (2014). Isolation and characterization of microsatellite markers for the causal agent of barley leaf rust, *Puccinia hordei*. *Australasian Plant Pathology* 43, 47–52 17

Kay, S., Hahn, S., Marois, E., Hause, G., and Bonas, U. (2007). A bacterial effector acts as a plant transcription factor and induces a cell size regulator. *Science* 318, 648–651 13

Keiper, F. J., Hayden, M. J., Park, R. F., and Wellings, C. R. (2003). Molecular genetic variability of Australian isolates of five cereal rust pathogens. *Mycological Research* 107, 545–556 17

Kemen, E., Kemen, A. C., Rafiqi, M., Hempel, U., Mendgen, K., Hahn, M., et al. (2005). Identification of a protein from rust fungi transferred from haustoria into infected plant cells. *Molecular Plant-Microbe Interactions* 18, 1130–1139 14

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology* 305, 567–580 20

Leonard, K. (2001). *Stem rust-future enemy? In 'Stem rust of wheat: from ancient enemy to modern foe'. (Ed. Peterson P.) pp. 119–146* (APS Press: St. Paul, MN) 4

Leonard, K. J. and Szabo, L. J. (2005). Stem rust of small grains and grasses caused by *Puccinia graminis*. *Molecular Plant Pathology* 6, 99–111 4

López-Solanilla, E., Bronstein, P. A., Schneider, A. R., and Collmer, A. (2004). HopPtoN is a *Pseudomonas syringae* Hrp (type III secretion system) cysteine protease effector that suppresses pathogen-induced necrosis associated with both compatible and incompatible plant interactions. *Molecular Microbiology* 54, 353–365 13

Mascher, M., Gundlach, H., Himmelbach, A., Beier, S., Twardziok, S. O., Wicker, T., et al. (2017). A chromosome conformation capture ordered sequence of the barley genome. *Nature* 544, 427–433 3

Mayer, K., Waugh, R., Brown, J., Schulman, A., Langridge, P., Platzer, M., et al. (2012). A physical, genetic and functional sequence assembly of the barley genome. *Nature* 491, 711–716 3

Mendgen, K. and Hahn, M. (2002). Plant infection and the establishment of fungal biotrophy. *Trends in Plant Science* 7, 352–356 7

Mendgen, K., Hahn, M., and Deising, H. (1996). Morphogenesis and mechanisms of penetration by plant pathogenic fungi. *Annual Review of Phytopathology* 34, 367–386 6

Menne, K. M., Hermjakob, H., and Apweiler, R. (2000). A comparison of signal sequence prediction methods using a test set of signal peptides. *Bioinformatics* 16, 741–742 20

Mesarich, C. H., Griffiths, S. A., van der Burgt, A., Ökmen, B., Beenen, H. G., Etalo, D. W., et al. (2014). Transcriptome sequencing uncovers the *Avr5* avirulence gene of the tomato leaf mold pathogen *Cladosporium fulvum*. *Molecular Plant-Microbe Interactions* 27, 846–857 23

Mueller, A. N., Ziemann, S., Treitschke, S., Aßmann, D., and Doehlemann, G. (2013). Compatibility in the *Ustilago maydis*–maize interaction requires inhibition of host cysteine proteases by the fungal effector pit2. *PLoS Pathogens* 9, e1003177 14

Nelson, R., Wilcoxson, R. D., and Christensen, J. (1955). Heterocaryosis as a basis for variation in *Puccinia graminis* var. *tritici*. *Phytopathology* 45, 639–643 8, 9

Nevo, E. (1992). *Origin, evolution, population genetics and resources for breeding of wild barley, Hordeum spontaneum, in the Fertile Crescent. In 'Barley: genetics, biochemistry, molecular biology and biotechnology' (Ed. Shewry, P. R.) pp. 19-43* (the Alden Press: Wallingford, UK) 3

Özkan, H., Brandolini, A., Schäfer-Pregl, R., and Salamini, F. (2002). AFLP analysis of a collection of tetraploid wheats indicates the origin of emmer and hard wheat domestication in southeast Turkey. *Molecular Biology and Evolution* 19, 1797–1801 2

Park, R. (2008). Breeding cereals for rust resistance in Australia. *Plant Pathology* 57, 591–602 4

Park, R., Burdon, J., and Jahoor, A. (1999). Evidence for somatic hybridization in nature in *Puccinia recondita* f. sp. *tritici*, the leaf rust pathogen of wheat. *Mycological Research* 103, 715–723 8

Park, R., Jahoor, A., and Felsenstein, F. (2000). Population structure of *Puccinia recondita* in western Europe during 1995, as assessed by variability in pathogenicity and molecular markers. *Journal of Phytopathology* 148, 169–179 17

Park, R. F. and Wellings, C. R. (2012). Somatic hybridization in the Uredinales. *Annual Review of Phytopathology* 50, 219–239 8

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* 8, 785–786 21

Petre, B., Joly, D. L., and Duplessis, S. (2014). Effector proteins of rust fungi. *Frontiers in Plant Science* 5 14

Rodríguez-Herva, J. J., González-Melendi, P., Cuartas-Lanza, R., Antúnez-Lamas, M., Río-Alvarez, I., Li, Z., et al. (2012). A bacterial cysteine protease effector protein interferes with photosynthesis to suppress plant innate immune responses. *Cellular Microbiology* 14, 669–681 13

Römer, P., Hahn, S., Jordan, T., Strauß, T., Bonas, U., and Lahaye, T. (2007). Plant pathogen recognition mediated by promoter activation of the pepper *Bs3* resistance gene. *Science* 318, 645–648 13

Rose, J. K., Ham, K.-S., Darvill, A. G., and Albersheim, P. (2002). Molecular cloning and characterization of glucanase inhibitor proteins coevolution of a counterdefense mechanism by plant pathogens. *The Plant Cell* 14, 1329–1345 13

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. doi:10.1371/journal.pone.0029847 20, 21

Schmidt, S. M., Lukasiewicz, J., Farrer, R., Dam, P., Bertoldo, C., and Rep, M. (2016). Comparative genomics of *Fusarium oxysporum* f. sp. *melonis* reveals the secreted protein recognized by the *Fom-2* resistance gene in melon. *New Phytologist* 209, 307–318 22

Shewry, P. R. (2009). Wheat. *Journal of Experimental Botany* 60, 1537–1553 3

Sonah, H., Deshmukh, R. K., and Bélanger, R. R. (2016). Computational prediction of effector proteins in fungi: opportunities and challenges. *Frontiers in Plant Science* 7 20

Song, X., Rampitsch, C., Soltani, B., Mauthe, W., Linning, R., Banks, T., et al. (2011). Proteome analysis of wheat leaf rust fungus, *Puccinia triticina*, infection structures enriched for haustoria. *Proteomics* 11, 944–963 23

Spanu, P. D., Abbott, J. C., Amselem, J., Burgis, T. A., Soanes, D. M., Stüber, K., et al. (2010). Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science* 330, 1543–1546 18

Sperschneider, J., Dodds, P. N., Gardiner, D. M., Manners, J. M., Singh, K. B., and Taylor, J. M. (2015). Advances and challenges in computational prediction of effectors from plant pathogenic fungi. *PLoS Pathogens* 11, e1004806 22

Sperschneider, J., Ying, H., Dodds, P. N., Gardiner, D. M., Upadhyaya, N. M., Singh, K. B., et al. (2014). Diversifying selection in the wheat stem rust fungus acts predominantly on pathogen-associated gene families and reveals candidate effectors. *Frontiers in Plant Science* 5 21

Steele, K. A., Humphreys, E., Wellings, C., and Dickinson, M. (2001). Support for a stepwise mutation model for pathogen evolution in Australasian *Puccinia striiformis* f. sp. *tritici* by use of molecular markers. *Plant Pathology* 50, 174–180 9

Struck, C., Ernst, M., and Hahn, M. (2002). Characterization of a developmentally regulated amino acid transporter (AAT1p) of the rust fungus *Uromyces fabae*. *Molecular Plant Pathology* 3, 23–30 7

Struck, C., Mueller, E., Martin, H., and Lohaus, G. (2004). The *Uromyces fabae UfAAT3* gene encodes a general amino acid permease that prefers uptake of in planta scarce amino acids. *Molecular Plant Pathology* 5, 183–189 7

Tian, M., Huitema, E., da Cunha, L., Torto-Alalibo, T., and Kamoun, S. (2004). A kazal-like extracellular serine protease inhibitor from *Phytophthora infestans* targets the tomato pathogenesis-related protease p69b. *Journal of Biological Chemistry* 279, 26370–26377 13

Tian, M., Win, J., Song, J., van der Hoorn, R., van der Knaap, E., and Kamoun, S. (2007). A *Phytophthora infestans* cystatin-like protein targets a novel tomato papain-like apoplastic protease. *Plant physiology* 143, 364–377 13

Upadhyaya, N. M., Garnica, D. P., Karaoglu, H., Sperschneider, J., Nemri, A., Xu, B., et al. (2015). Comparative genomics of Australian isolates of the wheat stem rust pathogen *Puccinia graminis* f. sp. *tritici* reveals extensive polymorphism in candidate effector genes. *Frontiers in Plant Science* 5, 759 15, 17

Upadhyaya, N. M., Mago, R., Staskawicz, B. J., Ayliffe, M. A., Ellis, J. G., and Dodds, P. N. (2014). A bacterial type III secretion assay for delivery of fungal effector proteins into wheat. *Molecular Plant-Microbe Interactions* 27, 255–264 14, 21

USDA (2017). USDA Food Composition Databases. [https://www.usda.gov/; accessed 9-August-2017] 2

Van Der Biezen, E. A. and Jones, J. D. (1998). Plant disease-resistance proteins and the gene-for-gene concept. *Trends in Biochemical Sciences* 23, 454–456 11

Viklund, H., Bernsel, A., Skwark, M., and Elofsson, A. (2008). SPOCTOPUS: a combined predictor of signal peptides and membrane protein topology. *Bioinformatics* 24, 2928–2929 21

Voegele, R. T., Struck, C., Hahn, M., and Mendgen, K. (2001). The role of haustoria in sugar supply during infection of broad bean by the rust fungus *Uromyces fabae*. *Proceedings of the National Academy of Sciences U.S.A.* 98, 8133–8138 7

Wallwork, H., Preece, P., and Cotterill, P. (1992). *Puccinia hordei* on barley and *Omithogalum umbellatum* in South Australia. *Australasian Plant Pathology* 21, 95–97 6

Wang, C.-I. A., Gunčar, G., Forwood, J. K., Teh, T., Catanzariti, A.-M., Lawrence, G. J., et al. (2007). Crystal structures of flax rust avirulence proteins *AvrL567-A* and *-D* reveal details of the structural basis for flax disease resistance specificity. *The Plant Cell* 19, 2898–2912 15

Wang, X. and McCallum, B. (2009). Fusion body formation, germ tube anastomosis, and nuclear migration during the germination of urediniospores of the wheat leaf rust fungus, *Puccinia triticina*. *Phytopathology* 99, 1355–1364 9

Watson, I. and Luig, N. (1959). Somatic hybridization between *Puccinia graminis* var. *tritici* and *Puccinia graminis* var. *secalis*. *Proceedings of the Linnean Society of New South Wales* 84, 207–208 9

Webb, C. A. and Fellers, J. P. (2006). Cereal rust fungi genomics and the pursuit of virulence and avirulence factors. *FEMS Microbiology Letters* 264, 1–7 6

Wellings, C. (2007). *Puccinia striiformis* in Australia: a review of the incursion, evolution, and adaptation of stripe rust in the period 1979–2006. *Australian Journal of Agricultural Research* 58, 567–575 8

Whisson, S. C., Boevink, P. C., Moleleki, L., Avrova, A. O., Morales, J. G., Gilroy, E. M., et al. (2007). A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature* 450, 115 21

Wicker, T., Oberhaensli, S., Parlange, F., Buchmann, J. P., Shatalina, M., Roffler, S., et al. (2013). The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. *Nature Genetics* 45, 1092–1096 18

Wiethölter, N., Horn, S., Reisige, K., Beike, U., and Moerschbacher, B. M. (2003). *In vitro* differentiation of haustorial mother cells of the wheat stem rust fungus, *Puccinia graminis* f. sp. *tritici*, triggered by the synergistic action of chemical and physical signals. *Fungal Genetics and Biology* 38, 320–326 6

Wu, J. Q., Sakthikumar, S., Dong, C., Zhang, P., Cuomo, C. A., and Park, R. F. (2017). Comparative genomics integrated with association analysis identifies candidate effector genes corresponding to *Lr20* in phenotype-paired *Puccinia triticina* isolates from Australia. *Frontiers in Plant Science* 8 17

Zheng, W., Huang, L., Huang, J., Wang, X., Chen, X., Zhao, J., et al. (2013). High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nature Communications* 4, 2673. doi:10.1038/ncomms3673 17

Chapter 2

A spontaneous mutation in *Puccinia graminis* f. sp. *tritici* to virulence on wheat resistance gene *Sr50* is associated with an asexual chromosomal recombination event

2.1 Introduction

Wheat is one of the most important staple foods in the world, but its production continues to be threatened by rust pathogens, especially the devastating *Puccinia graminis* f. sp. *tritici* (*Pgt*) (Zwer et al., 1992). Resistance breeding is one of the most cost-effective approaches to control rust diseases in cereal crops (Ellis et al., 2014; Park, 2008). However, rust pathogens constantly evolve, often acquiring virulence to rust resistant commercial wheat varieties and rendering them susceptible. Genetic resistance to many rust pathogens is commonly under the genetic control of corresponding pairs of genes in the host and pathogen, which govern resistance and avirulence, respectively. This gene-for-gene relationship was first formulated by Flor, who studied genetic interactions between flax and the flax rust pathogen *Melampsora lini* (Flor, 1955).

Rust pathogens are obligate biotrophs that are reliant upon their host plants to complete their life cycles. They have evolved specialized feeding structures known as haustoria that are produced between the host cell wall and plasma membrane. These feeding structures serve as sites for nutrient and water uptake from infected plant tissue (Garnica et al., 2014). Also, effector proteins are secreted from this structure into the host cytoplasm to target a variety of components of the host immune response (Garnica et al., 2014).

Most avirulence (Avr) genes that have been characterized in plant pathogens to date encode effector proteins that are recognized by the corresponding resistance gene products inside host plants cells. The response is often manifested as local cell death around the infection site, which stops pathogen growth, and is known as the hypersensitive response (Dodds and Rathjen, 2010). The cloning of Avr genes is important in permitting an understanding of the mechanisms that govern virulence in rust pathogens and the processes involved in recognition of Avr proteins by the resistant host genotype. It also enables studies of the diversity and evolution of the Avr genes in both natural ecosystems and farming systems, informing how rust pathogens evolve new virulence and in doing so overcome the disease resistance of important crop species. The first Avr gene cloned from a rust pathogen was *AvrL567*, which was isolated from *M. lini* using a map-based cloning approach in a population segregating for 10 avirulence loci (Dodds et al., 2004). Catanzariti et al. (2006) subsequently identified three additional Avr genes from *M. lini* by sequencing a cDNA library constructed from isolated haustoria and mapping clones encoding secreted proteins. Two further *M. lini* Avr genes were isolated by Anderson et al. (2016) using a genome mapping approach with the same family. All the flax rust Avr genes encode secreted proteins that are preferentially expressed in haustoria, indicating that these could be useful criteria for identifying new Avr gene candidates in rust fungi. Genome sequencing and transcriptional profiling of several cereal rust species have identified many such effector candidates based on these criteria.

However, associating these candidates with avirulence phenotypes has been difficult due to a lack of genetic and functional analysis tools. Functional confirmation of the *M. lini* Avr genes involved Agrobacterium-mediated delivery of Avr genes to leaves of flax genotypes with and without the corresponding resistance genes, followed by observation of resistance gene dependent host cell death. Although there are no reports of successful Agrobacterium-mediated transient expression assays in wheat leaves, bacterial type three secretion systems can be used to deliver proteins into wheat cells as an alternative (Thomas et al., 2009; Upadhyaya et al., 2014; Yin and Hulbert, 2011). Maia et al. (2017) recently used a bacterial delivery system to screen a number of infection-induced secreted effectors of the coffee leaf rust pathogen (*Hemilieia vastatrix*) identified by transcriptome analysis and found one, *HvEC-016*, that is recognized by the *SH1* resistance gene in coffee.

The first genomic study of *Pgt* was based on an American isolate, CDL 75-36-700-3 (7a), for which a genome assembly spanning 88.6 Mb was constructed by Duplessis et al. (2011). Approximately 1,000 small secreted proteins were predicted in the *Pgt* 7a genome. Upadhyaya et al. (2015) sequenced five genetically diverse Australian *Pgt* isolates and assembled a pan-genome (designated as PGTAus-pan henceforth) of 95 Mbp. Transcriptional profiling of isolated haustoria and germinated spores identified 520 haustorially-expressed secreted proteins (now updated to 586, Upadhyaya et al., unpublished).

The stem rust resistance gene *Sr50* is present in wheat on a translocated chromosome segment derived from cereal rye chromosome 1RS (Shepherd et al., 1973). *Sr50* has not been deployed in commercial wheat varieties, but it is broadly effective against *Pgt* and confers resistance to highly virulent pathotypes within the *Pgt* lineage Ug99. *Sr50* was cloned recently, and found to encode a coiled-coil nucleotide-binding leucine-rich repeat protein homologous to the barley gene *Mla* (Mago et al., 2015). However, the *Sr50*-matching Avr gene *AvrSr50* remains unknown. Previously, we identified a spontaneous *Sr50*-virulent mutant after inoculation of *Sr50* containing wheat with an avirulent *Pgt* culture (Mago et al., 2015). Here I have sequenced the genomes of both the parental and *Sr50*-virulent mutant isolates in order to identify candidate genes for *AvrSr50* with a comparative genomics approach.

2.2 Results

2.2.1 Confirmation of the *Sr50* virulence-gain in *Pgt*632

The spontaneous *Sr50*-virulent mutant *Pgt*632 was collected from a single virulent pustule on the wheat translocation line Gabo1DL•1RS (carrying *Sr50*) that had been inoculated with the *Sr50*-aviruent isolate *Pgt*279 (Mago et al., 2015). Pathotype analysis indicated the origin of *Pgt*632 from *Pgt*279 by mutation as both isolates showed the same virulence pattern on a set of 39 wheat differential lines (Mago et al., 2015). To confirm the virulence gain to the *Sr50* in *Pgt*632, both isolates were inoculated onto the wheat variety Gabo, the Gabo derivatives Gabo 1DL•1RS, and the tertiary recombinant T6-1 (Sr50+Sec-1; (Anugrahwati et al., 2008)). While Gabo lacks *Sr50* as a control, the latter two lines carry the resistance gene *Sr50*. Figure 2.1 shows that the parental isolate *Pgt*279 is fully virulent on Gabo, but its growth on the *Sr50* lines was restricted, resulting in an incompatible infection type (phenotype) (McIntosh et al., 1995). The mutant isolate *Pgt*632 showed full virulence to both Gabo 1DL•1RS and T6-1, confirming it has lost the functional allele *AvrSr50*. Like many other rust pathogens, the uredinial phase of *Pgt* has a dikaryotic genome separated in two haploid nuclei. Assuming dominance of avirulence, the avirulent parent most likely is heterozygous for *AvrSr50* (genotype *AvrSr50avrSr50*), with a spontaneous mutation resulting in alteration or loss of the dominant allele in the virulent mutant *Pgt*632.

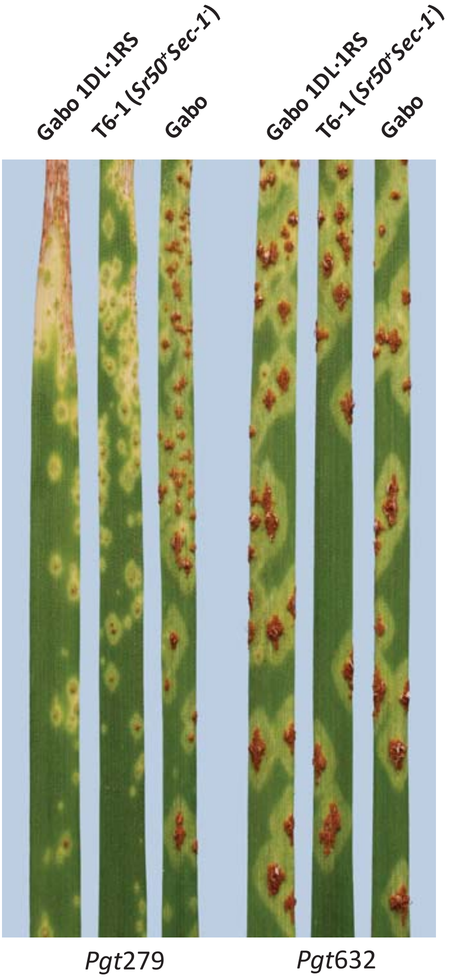


Figure 2.1: The infection types of *Pgt*279 and *Pgt*632 inoculated on three wheat lines Gabo, Gabo 1DL●1RS, and T6-1.

2.2.2 Comparative genomics of *Pgt*279 and *Pgt*632 identified 18 HSP genes with non-synonymous variations

In order to identify *AvrSr50* candidates, I sequenced the genomes of *Pgt*279 and *Pgt*632 using Illumina technology (250 bp paired end reads). About 5.7 Gbp and 5.4 Gbp of sequence data were obtained for *Pgt*279 and *Pgt*632, respectively, after quality trimming. Because the DNA library insert size was ∼550 bp and paired-end sequencing reads were 250 bp, some read pairs (∼14%) could overlap and they were merged to one sequence. Both the merged and unmerged reads were mapped to the reference genome PGTAus-pan, which was based on five Australian *Pgt* isolates that included pathotype 326-1,2,3,5,6 (Upadhyaya et al., 2015), from which *Pgt*279 is believed to have arisen via single step-wise mutation to virulence towards resistance gene *Sr9g* (Park, 2007). In total, 1,244,744 sequence variants (including single/multiple nucleotide variants, small insertions and deletions [InDel]) were detected in comparing *Pgt*279 to the reference, and 1,182,213 variants were detected for *Pgt*632 compared to the reference. To reduce this complexity, 586 genes that encoded haustorial secreted proteins (HSP) were targeted, in which a total of 4,183 and 4,177 amino acid changing variants were found in *Pgt*279 and *Pgt*632, respectively, relative to the reference genome. The two variant sets were scanned manually for read count support in the read mapping to identify false calling. The manual curation finally reported 4,376 and 4,316 variants for *Pgt*279 and *Pgt*632 (Supplementary Table 2.1). All 4,316 variants present in *Pgt*632 were also present in *Pgt*279, but there were a total of 59 variants from 18 HSP genes that were unique to *Pgt*279.

2.2.3 Heterozygosity and read depth analysis revealed a chromosome recombination associated to the loss of *AvrSr50*

Manual inspection of the 18 HSP genes containing non-synonymous differences between the wildtype and mutant strains showed that in each case *Pgt*279 was heterozygous for two gene sequence variants, while *Pgt*632 contained only a single variant (for example Figure 2.2). Thus, the virulence mutation in *Pgt*632 is associated with a loss-of-heterozygosity (LOH) across this set of HSP candidates, suggesting this may have resulted from a single large mutational event. Therefore, an attempt was made to map the extent of this LOH event in the genome by comparing heterozygosity between *Pgt*279 and *Pgt*632 at all variant positions on the reference genome PGTAus-pan. The average heterozygosity rate in the two isolates was calculated for each contig in the reference, which revealed five large contiguous blocks of LOH in *Pgt*632. The first of these included contigs 364 to 402 (Figure 2.3A), which cover approximately half of scaffold number 4 in the reference (Supplementary Table 2.2). Contigs in this region showed a much lower heterozygosity rate in *Pgt*632 than *Pgt*279. However, the remainder of the

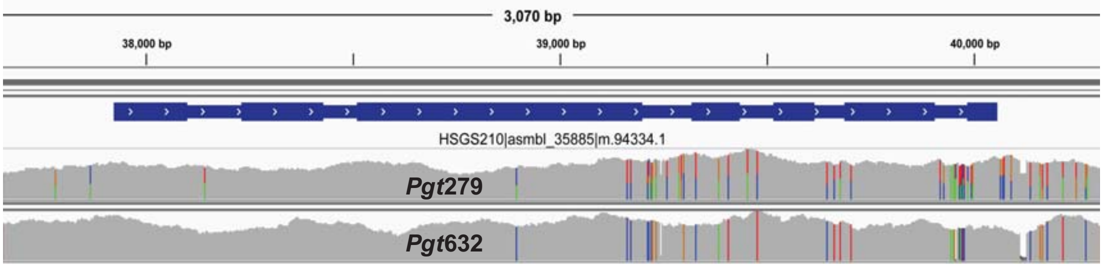


Figure 2.2: Illustration of loss-of-heterozygosity for the HSP gene *m.94334.1*. Graphs indicate the sequence coverage across this gene (coding sequence and introns indicated by blue bar above the graphs) for *Pgt*279 (upper panel) and *Pgt*632 (lower panel). Nucleotide polymorphisms relative to the reference *Pgt* genome assembly sequence are indicated by color coded bars showing the proportion of reads containing each SNV. Green, red, blue and orange represent the nucleotides A, T, C and G respectively.

contigs in scaffold 4 (Contigs 323-363) showed equivalent levels of heterozygosity in both isolates, suggesting that one end of the LOH event occurred between contigs 363 and 364 in this scaffold. In addition, LOH was also observed in four other large blocks including contigs 2662-2683 (Figure 2.3B), 1490-1545 (Figure 2.3C) and 3257-3271 (Figure 2.3D), and 4052-4058 (not shown) which represent the entire scaffolds 21, 51 and 75, and 131 respectively. Overall, the LOH in *Pgt*632 spanned a total of 2.5 Mbp over four full scaffolds and part of a fifth, suggesting that these scaffolds represent a single contiguous chromosomal region.

Two possible causes were considered for the observed LOH in isolate *Pgt*632: firstly, a large deletion in one of the two haploid nuclei; and secondly, a somatic chromosome recombination that replaced part of one chromosome in one nucleus with the corresponding region from the other nucleus. A deletion event would result in a single copy of the LOH region being present in *Pgt*632, while a somatic recombination event would retain two identical copies of this region. Thus, I compared the depth of read coverage for all contigs to distinguish between these possibilities. Twenty million reads were selected randomly for both isolates to normalise the sequencing read depth for the two isolates. The randomly selected reads were aligned to the reference genome and coverage depth was calculated. In *Pgt*632, average coverage depth of non-LOH contigs (including Contigs 1-363, 403-1489, 1546-2661, 2684-3256, 3272-4051, and

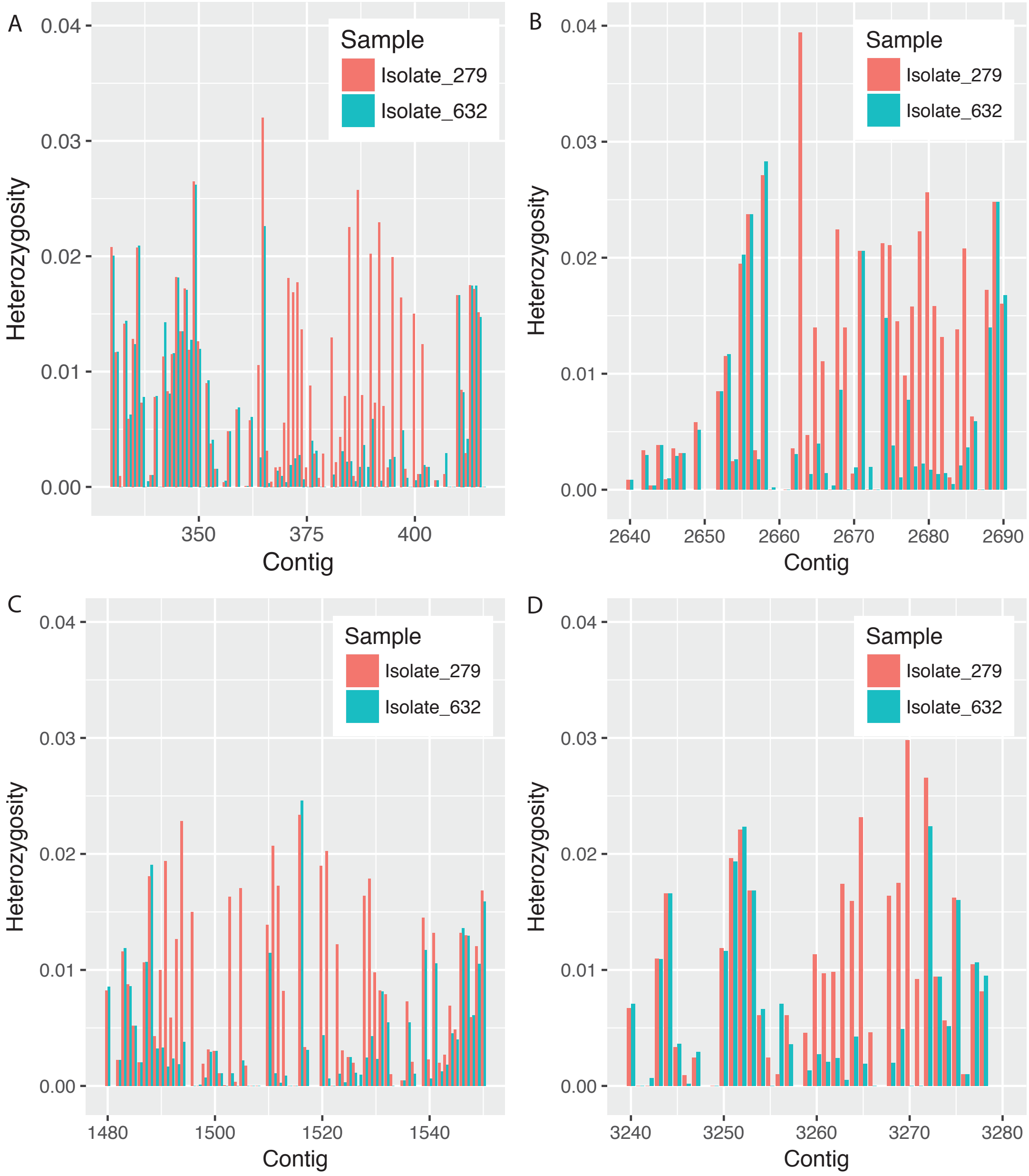


Figure 2.3: Heterozygosity of *Pgt*279 and *Pgt*632 on Contigs 330-416 (A), 1480-1550 (B), 2640-2690 (C), and 3240-3280 (D).

4059-4557) was 38.3X, while the average depth of the LOH Contigs (364-402, 1490-1545, 2662-2683, 3257-3271 and 4052-4058) was 23.2X, 33.3X, 31.2X, 32.1X, and

27.8X, respectively (Supplementary Table 2.3). Although these read depths are slightly lower than the rest of the genome average, a similar difference was also observed in *Pgt*279, where the average depth for these regions was 32.0X, 36.8X, 31.5X, 32.6X and 30.9X compared to the genome wide average of 38.7X. The similarity suggests that these differences are related to the sequence content of this region (e.g. the proportion of repetitive DNA) rather than to a loss of DNA copy number as expected for a deletion event.

To reduce possible bias in the read depth coverage estimates associated with intergenic repetitive elements, I also calculated coverage depth for the 724 individual gene loci in the LOH region and compared this to the read depth for the 278 genes at the 5 prime end of Scaffold 4 outside the LOH region, and also 1,000 genes in Contigs 1-206 as a control (shown in red, blue and green dots in Figures 2.4 A, B and C). For both *Pgt*279 (Figure 2.4 A) and *Pgt*632 (Figure 2.4 B), the read depth coverage was uniform at about 40X for genes in both the control and LOH regions, and there was no significant difference in the read depth distribution for these regions (Figure 2.4 D). The ratio of read depth in *Pgt*632 to *Pgt*279 was also calculated for each individual gene (Figure 2.4 C), which also showed no reduction in genomic mass of the LOH region compared to the control region in *Pgt*632 or relative to *Pgt*279. These data indicate that the loss of one haplotype in the LOH region in *Pgt*632 was accompanied by the duplication of the homologous chromosome segment, suggesting that a somatic recombination event occurred between the two chromosomes in this homologous pair rather than a deletion.

In another similar study, three *Pgt* mutants that gained virulence to R gene *Sr27* showed LOH on specific regions on Contigs 1873 to 1979 (Upadhyaya et al., unpublished), and also a reduction by half of read depth in these regions (Figure 2.5 A, B, and C) . This result suggests that the LOH events in these *Pgt Sr27*-virulent mutants were caused by deletions of DNA in one of the nuclei. It also shows that read depth is a highly sensitive means of measuring DNA copy number, and that this measurement can reliably detect DNA deletions in the dikaryotic genomes of rust fungi. In addition, a comparison of Figure 2.4B and Figure 2.5 further confirms no reduction of copy number for the LOH region in *Pgt*632.

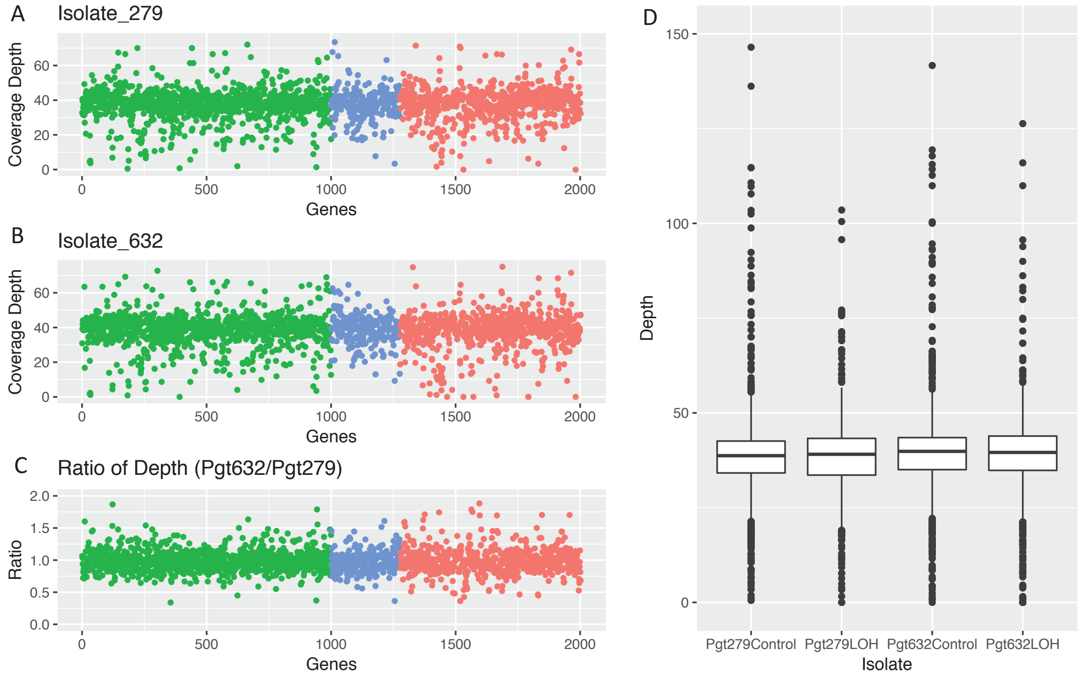


Figure 2.4: A loss-of-heterozygosity (LOH) event associated with virulence of *Pgt*632 on *Sr50*. (A) and (B) Coverage depth of 1000 genes from scaffold 1 (green), 278 genes from scaffold 4 but not included in the LOH region (blue), and 724 genes from the LOH region (red). (C) Ratio of coverage depth *Pgt*632 to *Pgt*279. (D) Distribution of coverage depth for genes in control and LOH regions of Pgt279 and Pgt632. Box plot graph illustrates the average frequency with the box indicating +/- 1 standard deviation, bars +/- 2 standard deviations and dots representing outlying data points.

2.2.4 Haplotype phasing of candidate genes for *AvrSr50*

The LOH event in *Pgt*632 is the potential cause of the mutation to virulence on *Sr50*, through loss of a dominant *AvrSr50* allele. Indeed, the previous variant imputation identified 18 out of the 586 HSP genes with non-synonymous SNVs, all of which were located in the LOH region. A total of 25 HSP genes were annotated in the reference genome assembly in this region. Therefore, each of these genes was examined in detail.

Firstly, to determine whether the 25 HSP genes identified represent single copy genes with two allelic variants, the allele frequency at individual SNV sites for each gene

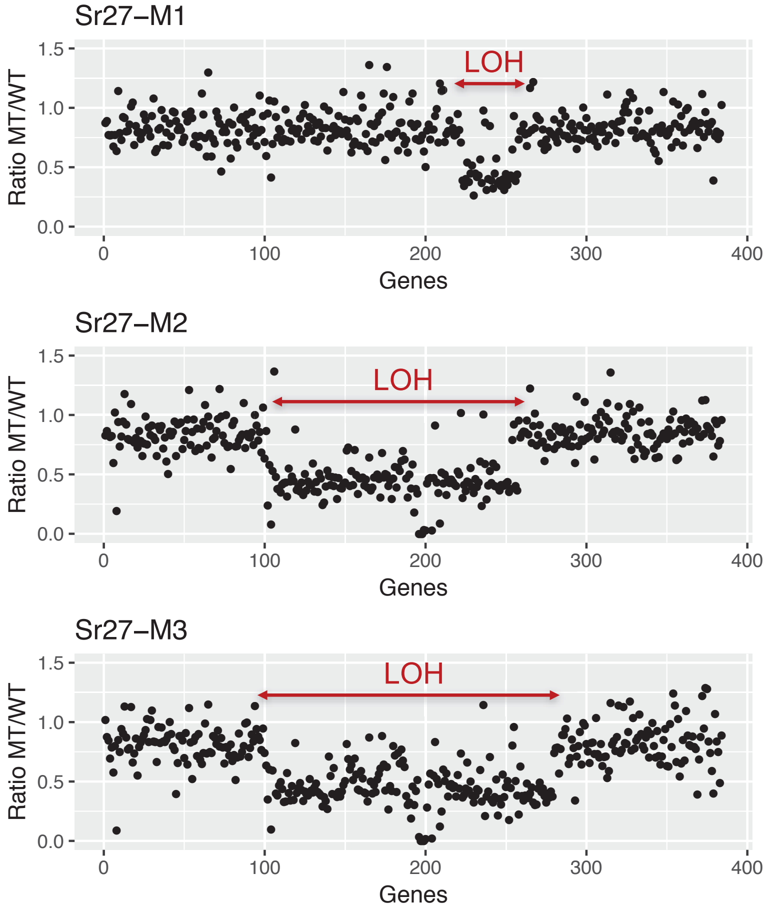


Figure 2.5: Read depth of three *Pgt* mutants (*Sr27*-M1, *Sr27*-M2, and *Sr27*-M3) on gene regions from Contig 1873 to Contig 1979 (Upadhyaya et al., unpublished).

was calculated using the *Pgt*279 sequence data (Figure 2.6). Twenty HSP genes showed confident read mapping and their SNV frequency (synonymous to allele frequency) distribution was close to the expected 50% derived from diploid alleles, suggesting that the reads mapping to these genes derived solely from a single genomic copy of the gene. Five genes showed wider ranges of SNV frequency distribution, suggesting the possibility of more complex genomic structures.

Two genes with aberrant SNV allele frequencies, *PGTAUSPE_213* and *m.100841*[1], are both located on contig 402 and are also closely related in sequence (95% nucleotide

[1] Full name: *HSGS210|asmbl_37899|m.100841*

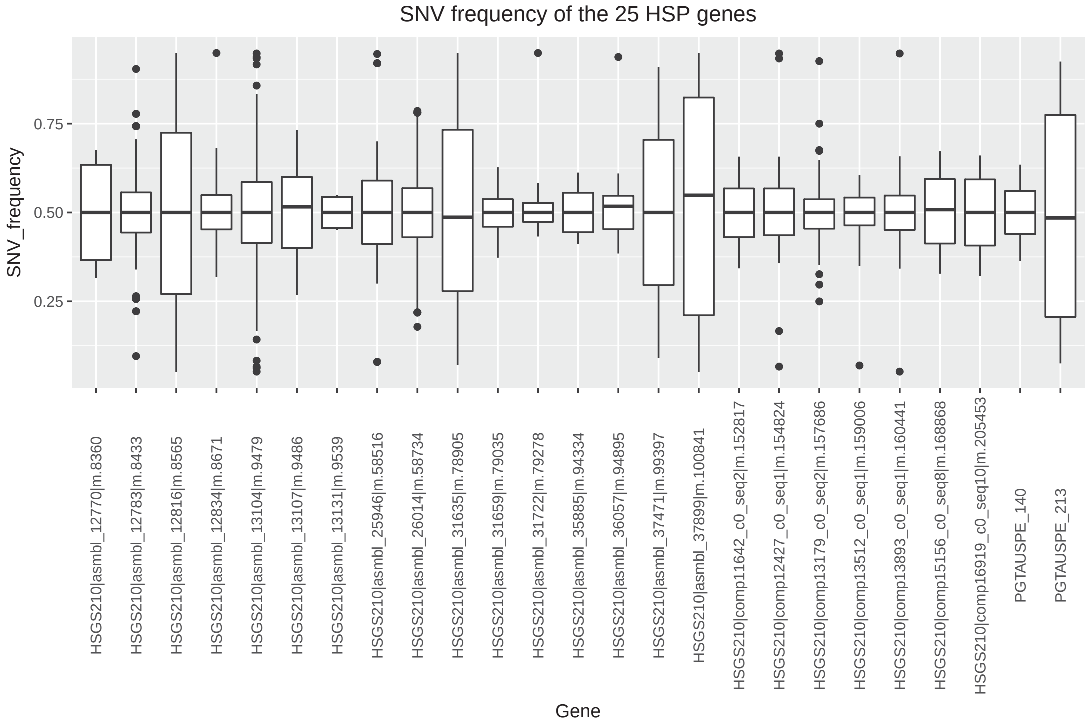


Figure 2.6: Boxplots showing allele frequency distribution at SNV sites of the 25 HSP genes.

identity), indicating that they are paralogous genes within a tandemly repeated cluster. The average read depth mapping to the two genes in *Pgt*279 were 60.56x and 49.93x, higher than the average read depth of Contig 402 (47.60x). Even in *Pgt*632, some non-homozygous SNVs were detected mapping to the genes, suggesting that additional copies of this gene family are present in both pathotypes but are not assembled separately in the reference genome. One other gene, *HSGS210|asmbl_12816|m.8565*, is also multi-copy with three other closely related genes annotated at unlinked positions in the reference and many reads mapping to multiple locations, skewing the SNV frequencies. Another gene, *HSGS210|asmbl_37471|m.99397*, showed very low read depth suggesting that it is not present in *Pgt*279. The fifth gene *m.78905* appeared to only show two sequence variants, and *Pgt*632 contained a single sequence with no heterozygous positions, indicating a single copy in this haplotype. It is possible that two identical copies occur at the other haplotype in *Pgt*279.

The alleles of the HSP genes that are present in *Pgt*279 but lost in *Pgt*632 could potentially encode *AvrSr50*, while the alleles common to both isolates could include the virulence allele. Therefore, the sequence data from each isolate was used to resolve the two allelic variants for each gene. The *Pgt*632 reads were used to assemble the allele from the virulence-associated haplotype and these were subtracted from the *Pgt*279 reads containing genetic information of both alleles. Two allelic variants could be confidently resolved in this way for 20 genes, and their coding sequences were extracted and are deposited in Supplementary Table 2.4. It was not possible to resolve alleles for the three multi-copy genes, nor for the single gene that was absent in these isolates. A fifth gene, *HSGS210|comp16919_c0_seq10|m.205453*, is incomplete in the reference assembly, with only the 3 prime end present at the edge of Contig 10678. In two cases, *HSGS210|asmbl_13107|m.9486* and *HSGS210|asmbl_35885|m.94334*, the two allelic variants encode no amino acid differences. Thus the original identification of 18 HSP genes by SNV analysis detected all of the single copy genes with amino acid changes, but could not detect the multi-copy genes in this region.

2.3 Materials and methods

2.3.1 Characterisation of rust isolates

The Australian Wheat Stem Rust Differential Set (Park, 2007), plus three additional lines (Gabo, Gabo1BL•1RS and T6-1), were used to confirm the pathogenicity of two *Pgt* isolates, *Pgt*279 (pathotype 98-1,2,3,5,6; accession number 781219) and *Pgt*632 (pathotype 98-1,2,3,5,6 +Sr50; accession number 130176) maintained at the University of Sydney Plant Breeding Institute. Urediniospores of the two rust isolates were generated by inoculation to uncontaminated seedlings of the wheat genotype Morocco, followed by hand collection once sporulation had commenced. DNA was extracted from urediniospores with a CTAB method with minor modifications (Rogers et al., 1989). After quantity and quality assessment with a Nanodrop Spectrophotometer (Thermo Scientific, Wilmington, DE), the DNA was sent to the Australian Genome Research Facility Ltd for library preparation and Illumina sequencing (250 bp paired-end, insert size 550 bp) on the Hiseq2500 platform.

2.3.2 Imputation of Genomic variation

The raw sequencing reads were first imported to CLC Genomics Workbench (CLCGW) version 9.0.1, and then filtered and trimmed to remove low quality ends, sequencing adapters and low quality reads (Trim using quality score 0.01, maximum number of ambiguities allowed is 2). The short insert size ($\sim$550 bp) generated some overlapping pairs of reads, which were merged with the tool "Merge Overlapping Pairs" in CLCGW. Reads of both *Pgt*279 and *Pgt*632 were mapped to a previously built pan genome for *Pgt*, PGTAus-pan[1] (Upadhyaya et al., 2015) using the default stringency (similarity fraction 0.8 and length fraction 0.5) followed by local realignment. Variant calling was performed using the "Probabilistic Variant Detection" program on locally realigned read mapping with parameters "ignore non-specific matches, minimum coverage 10, variant probability 90, minimum variant count 2, include broken pairs". The variants called are genomic nucleotide differences between the individual samples and the reference. The program "Compare variants within group" in CLCGW was used to compare the two samples viewed as a group to identify variants specific to each sample as well as shared variants. Functional consequences of such polymorphic variants in the 586 HSPs annotated in the PGTAus-pan genome were predicted using the CLCGW tool "Amino Acid Changes" and curated manually by visual inspection of read mapping tracks in CLCGW.

2.3.3 Heterozygosity rate calculation

This task was performed in the Linux command environment. The sequencing reads were filtered and trimmed again with Trim Galore v0.3.7[2] with parameters "-quality 20 -phred33 -fastqc -gzip -length 35". Sequencing reads of *Pgt*632 and *Pgt*279 were mapped to the reference genome PGTAus_pan with Bowtie2 v2.2.5 (Langmead and

[1] http://webapollo.bioinformatics.csiro.au/puccinia_graminis_tritici_PGTAus-pan/index.html

[2] http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

Salzberg, 2012) with parameters "–very-sensitive -minins 0 -maxins 900". Alignments around InDels were processed with RealignerTargetCreator and IndelRealigner in the GATK package v3.3.0 (McKenna et al., 2010). SNV calling was performed with GATK UnifiedGenotyper with parameters "–output_mode EMIT_VARIANTS_ONLY -glm SNP".

Heterozygosity rates were calculated for each contig in the reference genome assembly based on the GATK SNV calling result with following formula:

$$CH = NSNV/CL$$

in which CH represents contig heterozygosity, NSNV is the number of heterozygous SNV in each contig, and CL is contig length. The heterozygosity rates of both samples were calculated with custom Perl script and visualized with R ggplot2 package.

2.3.4 Read depth analysis

To investigate the genomic mass in the LOH region, sequencing reads of both isolates were first normalized by randomly sampling an equal amount of data of 20 million reads with the tool seqtk v.1.0-r82[1] using random seed 100. Then the sub-sampled reads were mapped to the reference with Bowtie2 using the same parameter settings as in Heterozygosity rate calculation. ReSeqtools v0.21 (He et al., 2013) was used to calculate read depth on the contigs and also within gene loci (defined by the genomic sequence between the start and stop codons of the predicted coding sequences and including introns).

2.3.5 Haplotype phasing

To evaluate the read mapping quality of the 25 HSP genes for two allele separation, allele frequency at variant sites in gene coding sequences was calculated. The read mapping of *Pgt*279 in CLCGW was exported as a BAM format file, which was then inputted

[1] https://github.com/lh3/seqtk

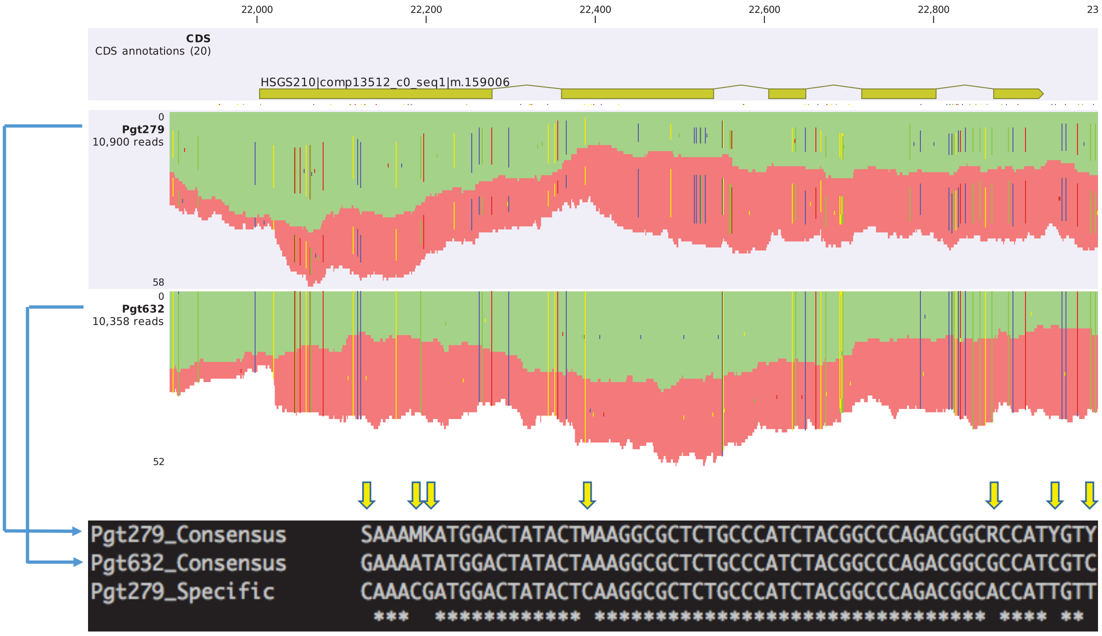
to software bam-readcount v0.5.0[1] to calculate read counts representing A/G/C/T genotypes at heterozygous positions. To remove the noise introduced by sequencing errors, variants with a frequency lower than 0.05 were discarded. The distribution of resulting allele frequencies was calculated and plotted with geom_boxplot package in R.

To obtain full length allelic sequences for the HSP genes, the allele of each gene from the *Sr50*-virulence haplotype was first assembled using the *Pgt*632 reads mapping to the gene position on the reference with tool "Extract Consensus Sequence" in CLCGW (parameters "N for 0 coverage, conflict resolution with ambiguity code, noise threshold 0.2, minimum NT 2"). As the tool does not consider InDel in the assembly, InDels were manually added to the assembly based on visual scan of read mapping and variant calling. The read mapping of *Pgt*279 was then also assembled into consensus sequences for these genes, with genotypes at heterozygous sites represented with IUPAC ambiguity codes. The assembled *Pgt*279 and *Pgt*632 consensus sequences were then aligned end-by-end and the IUPAC codes in *Pgt*279 consensus sequences were reduced to single base genotypes by subtracting the *Pgt*632 genotypes. One instance of allele separation is shown in Figure 2.7.

2.4 Discussion

By comparing whole genome sequencing data of two *Pgt* isolates differing in pathogenicity to wheat stem rust resistance gene *Sr50*, a 2.5 Mbp region with LOH was identified in the virulent mutant isolate, within which 25 HSP genes were considered as candidates for the avirulence gene corresponding to *Sr50* (Figures 2.2 and 2.3). Several Avr genes have been identified in fungal plant pathogens via a comparative genomics approach similar to that used here. de Jonge et al. (2012) sequenced four virulent and seven avirulent *Verticillium dahliae* isolates to tomato immune receptor gene *Ve1*, and identified a 50 Kbp DNA segment deleted in the virulent isolates. One gene in the deleted region was highly expressed in *V. dahlia*-infected tobacco plants and was further

[1] https://github.com/genome/bam-readcount



```
Pgt279_Consensus     SAAAMKATGGACTATACTMAAGGCGCTCTGCCCATCTACGGCCCAGACGGCRCCATYGTY
Pgt632_Consensus     GAAAATATGGACTATACTAAAGGCGCTCTGCCCATCTACGGCCCAGACGGCGCCATCGTC
Pgt279_Specific      CAAACGATGGACTATACTCAAGGCGCTCTGCCCATCTACGGCCCAGACGGCACCATTGTT
                      ***  ************ ******************************** **** **
```

Figure 2.7: Allele sequence separation from read mapping in *Pgt*279 and *Pgt*632 for $HSGS210|comp13512\_c0\_seq1|m.159006$. The upper panel shows read coverage, with nucleotide polymorphisms with respect to the reference genome shown as color coded bars representing genotypes: green = T, red = A, blue = C and yellow=G. The single allele of the gene in *Pgt*632 was assembled from read mapping to the locus, and consensus of both alleles was assembled from *Pgt*279 with IUPAC codes to represent heterozygous genotypes. In the lower panel, the two consensus sequences were aligned and *Pgt*279 specific allele was obtained by subtracting the IUPAC codes with *Pgt*632 genotypes.

confirmed to be *Ave1* with functional assays. Similarly, Schmidt et al. (2016) performed whole genome sequencing of *Fusarium oxysporum* f. sp. *melonis* isolates with either avirulence or virulence on host resistance gene *Fom-2*, and identified *AvrFom2* based on its presence/absence polymorphism in the avirulent/virulent isolates. Similar to these two Avr gene discoveries, the identification of 25 candidates associated with the LOH event implicates loss of the *Sr50*-avirulence via deletion of the Avr allele. In separate work, three other *Pgt* isolates with spontaneous virulence gains to *Sr27* showed overlapping deletions in one karyon of ∼200 Kbp (Figure 2.5; Upadhyaya et al., unpublished). The commonly deleted region harbors one HSP gene as a highly confident candidate for *AvrSr27*. These data indicate an association of Avr gene loss with genomic DNA deletion events in *Pgt*. However, Avr gene function can also be lost via simple sequence mutation rather than a large DNA deletion. The study by Mesarich et al. (2014) showed the loss of *Avr5* in *Cladosporium fulvum* strain IPO 1979 was caused by a 2 bp deletion in the gene that resulted in coding frame modification. No non-synonymous mutations were identified in the 561 HSP genes outside the LOH regions, suggesting that *AvrSr50* was not very likely to be lost via simple protein sequence modification.

Haplotype phasing for 20 of the genes was performed, while the other five genes could not be resolved due to their complex genomic structures (Figures 2.6 and 2.7). Among the genes, *PGTAUSPE_213* and *HSGS210|asmbl_37899|m.100841* appeared to be paralogous genes with only two copies present in the reference genome, but apparently have higher copy number in *Pgt*279. Thus, the reference based bioinformatics analysis could not resolve genetic information for all paralogs of the two genes. *De novo* assembly is not limited to a reference genome, but it could still collapse paralogous genes into a single "mosaic" copy when their assembly involves repetitive elements. For instance, six paralogs of *AvrM* family of flax rust were assembled as a single "mosaic" gene with 75 bp Illumina reads, whereas their flanking repetitive regions were assembled as separate contigs (Nemri et al., 2014). It is essential to ensure accuracy of multi-copy gene assembly, because paralogous variants of Avr genes can be associated with recognition specificities by their host R genes. For instance, Dodds et al. (2006) described 12 variants of the *AvrL567* gene family that occur as part of a tandem locus with copy number variation between 1 to 4 in different haplotypes. Seven variants were differentially recognized by flax *L5, L6* and *L7* resistance gene alleles.

As sequencing technologies keep advancing, new platforms like single molecule real-time (SMRT) sequencing will provide much longer read length (1,000s bp; Chin et al., 2013) for *de novo* assembly, overcoming many of the difficulties associated with current shorter-read platforms. In a recent study of the bacterial plant pathogen *Xanthomonas translucens* (causal agent of bacterial leaf streak), SMRT sequencing resolved Transcription Activator-Like effector genes, which have been difficult to study because their repetitive nature has complicated assemblies based on short read technology (Peng et al., 2016). SMRT sequencing has also proven powerful in phasing diploid genome assemblies to separate two haplotypes for fungal genomes (Chin et al., 2016). Future work on *Pgt* genomics will include the improvement of genome assemblies with the new longer read length sequencing, providing a haplotype phased reference genome with correct copy number representation of candidate effector genes.

The 2.5 Mbp LOH region in *Pgt*632 covered part of scaffold 4 and all of four other scaffolds in the *Pgt*7a and PGTAus-pan genome assemblies, suggesting that these scaffolds represent a contiguous region on a single chromosome in *Pgt*. Further analysis showed no significant loss of depth of sequencing read coverage across the 2.5 Mb region (Figure 2.4), suggesting that the LOH resulted from a recombination event in which this chromosomal region from one nucleus was replaced by the corresponding region from the other nucleus. The two nuclei present in the dikaryotic asexual stage of rust fungi replicate and divide independently (Heath, 2012). There have been reports that individual nuclei can be exchanged between *Pgt* isolates co-inoculated on wheat giving rise to novel pathotypes (Bridgmon and Wilcoxson, 1959; Park and Wellings, 2012; Watson and Luig, 1959). In some cases, the number of resulting pathotypes observed from such mixed inoculations exceeded those possible from simple nuclear exchange, suggesting somatic recombination between nuclei (Elligboe, 1961; Watson and Luig, 1958), although at that time it was not possible to verify the recombinant nature of the isolates using independent genetic markers. In a more recent study using RADseq markers, Anderson et al. (2016) observed LOH events in mutants of *M. lini*. In one case, mutation to virulence on the *M1* and *M4* resistance genes was accompanied by a LOH spanning 32.14 cM in the linkage group LG19 of *M. lini* (including the *AvrM14* locus). The average read depth of markers across this region was halved compared to the rest of the genome, suggesting a deletion event. However, in another case a LOH event spanning 208 cM at one end of the linkage group 5 (LG5) including the *AvrN* locus, which resulted in virulence on the *N* resistance gene, was associated with no loss of read coverage. In this case, markers on the remainder of LG5 showed a 50% increase in read depth and change to a 2:1 allelic ratio (rather than 1:1), suggesting that the apparent deletion (LOH) of a large chromosomal segment was accompanied by a duplication of the entire equivalent chromosome from the other nucleus. However, it was unknown whether there were chromosome exchanges between the two nuclei in this mutant. The data presented here suggest that a somatic chromosome recombination event occurred in *Pgt* isolate 279 to give rise to isolate 632, resulting in the loss of the *AvrSr50* gene. The mechanism by which such nuclear exchange may occur is unknown. One possibility is crossing over between the *AvrSr50*-containing chromosomes during mitotic division across the equatorial plane when both nuclei are undergoing mitosis simultaneously. This could result in a new cell with two copies of the virulence allele, and the other cell with two copies of the avirulence allele. Another possibility is that fusion between the two nuclei may have occurred allowing crossing over between chromosomes in a single nucleus. Williams and Mendgen (1975) described a monokaryotic *Pgt* strain and showed evidence for diploid DNA content in urediniospores of this strain, suggesting that it resulted from nuclear fusion of a dikaryon.

To summarize, genome sequencing of two *Pgt* isolates differing in virulence to wheat stem rust resistance gene *Sr50* revealed a genomic region with a LOH spanning 2.5 Mbp continuous on five scaffolds in *Pgt*632. Further analysis showed that the read coverage depth in the LOH region was not halved, indicating it did not result from a simple genomic deletion but rather a somatic recombination event. Under the strong assumption of the gene-for-gene hypothesis, the allelic variants of the 25 HSP genes present in *Pgt*279 but missing in *Pgt*632 are candidates to encode *AvrSr50*. In planta expression constructs of the *Pgt*279-specific alleles were generated following the completion of this work. One of these alleles triggered a cell death response in *Nicotiana benthamiana* when it was co-expressed with the resistance gene *Sr50* (Chen et al., unpublished). In addition, the protein encoded by this allele showed a direct interaction with Sr50 protein in a yeast-two-hybrid assay, confirming *AvrSr50* recognition by *Sr50*. This result shows the effectiveness of comparative genomics between rust clonal mutant isolates and their wildtypes for Avr gene identification, which will ultimately contribute to better understanding of rust pathogenesis in wheat.

References

Anderson, C., Khan, M. A., Catanzariti, A.-M., Jack, C. A., Nemri, A., Lawrence, G. J., et al. (2016). Genome analysis and avirulence gene cloning using a high-density RADseq linkage map of the flax rust fungus, *Melampsora lini*. *BMC Genomics* 17, 667 37, 53

Anugrahwati, D. R., Shepherd, K. W., Verlin, D. C., Zhang, P., Mirzaghaderi, G., Walker, E., et al. (2008). Isolation of wheat–rye 1RS recombinants that break the linkage between the stem rust resistance gene *SrR* and secalin. *Genome* 51, 341–349 38

Bridgmon, G. and Wilcoxson, R. (1959). New races from mixtures of urediospores of varieties of *Puccinia graminis*. *Phytopathology* 49, 428–429 53

Catanzariti, A.-M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., and Ellis, J. G. (2006). Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *The Plant Cell* 18, 243–256 37

Chin, C.-S., Alexander, D. H., Marks, P., Klammer, A. A., Drake, J., Heiner, C., et al. (2013). Nonhybrid, finished microbial genome assemblies from long-read smrt sequencing data. *Nature Methods* 10, 563–569 52

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased diploid genome assembly with single molecule real-time sequencing. *Nature Methods* 13, 1050 52

de Jonge, R., van Esse, H. P., Maruthachalam, K., Bolton, M. D., Santhanam, P., Saber, M. K., et al. (2012). Tomato immune receptor *Ve1* recognizes effector of multiple fungal pathogens uncovered by genome and RNA sequencing. *Proceedings of the National Academy of Sciences U.S.A.* 109, 5110–5115 50

Dodds, P. N., Lawrence, G. J., Catanzariti, A.-M., Ayliffe, M. A., and Ellis, J. G. (2004). The *Melampsora lini AvrL567* avirulence genes are expressed in haustoria and their products are recognized inside plant cells. *The Plant Cell* 16, 755–768 37

Dodds, P. N., Lawrence, G. J., Catanzariti, A.-M., Teh, T., Wang, C.-I., Ayliffe, M. A., et al. (2006). Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proceedings of the National Academy of Sciences U.S.A.* 103, 8888–8893 52

Dodds, P. N. and Rathjen, J. P. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. *Nature Reviews. Genetics* 11, 539 36

Duplessis, S., Cuomo, C. A., Lin, Y.-C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proceedings of the National Academy of Sciences U.S.A.* 108, 9166–9171. doi:10.1073/pnas.1019315108 37

Elligboe, A. (1961). Somatic recombination in *Puccinia graminis* var *tritici*. *Phytopathology* 51, 13 53

Ellis, J. G., Lagudah, E. S., Spielmeyer, W., and Dodds, P. N. (2014). The past, present and future of breeding rust resistant wheat. *Frontiers in Plant Science* 5 36

Flor, H. (1955). Host-parasite interactions in flax rust-its genetics and other implications. *Phytopathology* 45, 680–685 36

Garnica, D. P., Nemri, A., Upadhyaya, N. M., Rathjen, J. P., and Dodds, P. N. (2014). The ins and outs of rust haustoria. *PLoS Pathogens* 10, e1004329 36

He, W., Zhao, S., Liu, X., Dong, S., Lv, J., Liu, D., et al. (2013). ReSeqTools: an integrated toolkit for large-scale next-generation sequencing based resequencing analysis. *Genetics and Molecular Research* 12, 6275–6283 49

Heath, M. (2012). *Ultrastructure of rust fungi* (Elsevier) 53

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9, 357–359 48

Mago, R., Zhang, P., Vautrin, S., Šimková, H., Bansal, U., Luo, M.-C., et al. (2015). The wheat *Sr50* gene reveals rich diversity at a cereal disease resistance locus. *Nature Plants* 1, 15186 38

Maia, T., Badel, J. L., Marin-Ramirez, G., Rocha, C. d. M., Fernandes, M. B., Silva, J. C., et al. (2017). The *Hemileia vastatrix* effector HvEC-016 suppresses bacterial blight symptoms in coffee genotypes with the *SH1* rust resistance gene. *New Phytologist* 213, 1315–1329 37

McIntosh, R. A., Wellings, C. R., and Park, R. F. (1995). *Wheat rusts: an atlas of resistance genes* (Melbourne Australia: CSIRO) 38

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20, 1297–1303 49

Mesarich, C. H., Griffiths, S. A., van der Burgt, A., Ökmen, B., Beenen, H. G., Etalo, D. W., et al. (2014). Transcriptome sequencing uncovers the *Avr5* avirulence gene of the tomato leaf mold pathogen *Cladosporium fulvum*. *Molecular Plant-Microbe Interactions* 27, 846–857 52

Nemri, A., Saunders, D. G. O., Anderson, C., Upadhyaya, N. M., Win, J., Lawrence, G. J., et al. (2014). The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. *Frontiers in Plant Science* 5, 98. doi:10.3389/fpls.2014.00098 52

Park, R. (2007). Stem rust of wheat in Australia. *Australian Journal of Agricultural Research* 58, 558–566 40, 47

Park, R. (2008). Breeding cereals for rust resistance in Australia. *Plant Pathology* 57, 591–602 36

Park, R. F. and Wellings, C. R. (2012). Somatic hybridization in the Uredinales. *Annual Review of Phytopathology* 50, 219–239 53

Peng, Z., Hu, Y., Xie, J., Potnis, N., Akhunova, A., Jones, J., et al. (2016). Long read and single molecule DNA sequencing simplifies genome assembly and TAL effector gene analysis of *Xanthomonas translucens*. *BMC Genomics* 17, 21 52

Rogers, S. O., Rehner, S., Bledsoe, C., Mueller, G. J., and Ammirati, J. F. (1989). Extraction of DNA from *Basidiomycetes* for ribosomal DNA hybridizations. *Canadian Journal of Botany* 67, 1235–1243 47

Schmidt, S. M., Lukasiewicz, J., Farrer, R., Dam, P., Bertoldo, C., and Rep, M. (2016). Comparative genomics of *Fusarium oxysporum* f. sp. *melonis* reveals the secreted protein recognized by the *Fom-2* resistance gene in melon. *New Phytologist* 209, 307–318 51

Shepherd, K. et al. (1973). Homoeology of wheat and alien chromosomes controlling endosperm protein phenotypes. In *Proceedings of the Fourth International Wheat Genetics Symposium. Cytogenetics.* (University of Missouri.), 745–760 37

Thomas, W. J., Thireault, C. A., Kimbrel, J. A., and Chang, J. H. (2009). Recombineering and stable integration of the *Pseudomonas syringae* pv. *syringae* 61 hrp/hrc cluster into the genome of the soil bacterium *Pseudomonas fluorescens* Pf0-1. *The Plant Journal* 60, 919–928 37

Upadhyaya, N. M., Garnica, D. P., Karaoglu, H., Sperschneider, J., Nemri, A., Xu, B., et al. (2015). Comparative genomics of Australian isolates of the wheat stem rust pathogen *Puccinia graminis* f. sp. *tritici* reveals extensive polymorphism in candidate effector genes. *Frontiers in Plant Science* 5, 759 37, 40, 48

Upadhyaya, N. M., Mago, R., Staskawicz, B. J., Ayliffe, M. A., Ellis, J. G., and Dodds, P. N. (2014). A bacterial type III secretion assay for delivery of fungal effector proteins into wheat. *Molecular Plant-Microbe Interactions* 27, 255–264 37

Watson, I. and Luig, N. (1958). Somatic hybridization in *Puccinia graminis* var. *tritici*. In *Proceedings of the Linnean Society of New South Wales*. vol. 83, 190–195 53

Watson, I. and Luig, N. (1959). Somatic hybridization between *Puccinia graminis* var. *tritici* and *Puccinia graminis* var. *secalis*. *Proceedings of the Linnean Society of New South Wales* 84, 207–208 53

Williams, P. and Mendgen, K. (1975). Cytofluorometry of DNA in uredospores of *Puccinia graminis* f. sp. *tritici*. *Transactions of the British Mycological Society* 64, 23IN3–28 54

Yin, C. and Hulbert, S. (2011). Prospects for functional analysis of effectors from cereal rust fungi. *Euphytica* 179, 57–67 37

Zwer, P., Park, R., and McIntosh, R. (1992). Wheat stem rust in Australia dash 1969-1985. *Australian Journal of Agricultural Research* 43, 399–431 36

Chapter 3

Genome assembly and characterization for the barley leaf rust fungus, *Puccinia hordei*

3.1 Introduction

Barley leaf rust, caused by the fungus *Puccinia hordei* (*Ph*), is a severe disease in all barley-growing regions of Australia (Park, 2003). Substantial yield losses of up to 62% due to the disease have been reported (Cotterill et al., 1995, 1992; Waterhouse, 1927). Economic losses in Australia were estimated to be AU$9 million per year (Murray and Brennan, 2009). The continuing emergence of new *Ph* pathotypes with virulence to important resistance genes pose a great threat to sustainable barley production. Understanding virulence mechanisms and evolution of this pathogen is important in ensuring that genetic approaches to control are effective. Progress in molecular studies of this pathogen have been limited by a lack of genome sequence information in the *Ph* community, with less than 100 nucleotide or protein sequences available in NCBI at the time this study was commenced. In contrast, the genomes of three other cereal rust pathogens *Puccinia graminis* f. sp. *tritici* (*Pgt*), *Puccinia triticina* (*Pt*) and *Puccinia striiformis* f. sp. *tritici* (*Pst*) have been fully sequenced and annotated (Cantu et al., 2011; Cuomo et al., 2017; Duplessis et al., 2011). Genomes of the flax rust fungus *Melampsora lini* (*Mli*), a model species for the gene-for-gene hypothesis, and the poplar rust fungus *Melampsora larici-populina* (*Mlp*), a devastating pathogen of poplar, have also been sequenced (Duplessis et al., 2011; Nemri et al., 2014).

The obligate biotrophic nature of *Ph* has made genetic research on this pathogen very difficult. This can be partially circumvented by recent advances in genome sequencing. By comparing the genomes of *Pgt* and *Mli* to pathogenic fungi with necrotrophic lifestyles, a core set of genes were identified as being related to their obligate biotrophic lifestyles, along with a reduction in nutrient assimilation ability and an expansion of transporter families for increased host nutrient uptake (Duplessis et al., 2011). For instance, the observed expansion of proteases and oligopeptide transporter families in the two rust fungi indicated that they had evolved to degrade host extracellular proteins and assimilating the resulting peptides.

To maintain continuous infection and reproduction inside their hosts, pathogens including *Ph* need to overcome plant defenses. In the first stage of pathogen-host interactions, plants recognize pathogen associated molecular patterns (PAMPs; e.g.

chitin or flagellin) with cell surface receptors and subsequently induce PAMP-triggered immunity (PTI) (Dodds and Rathjen, 2010; Jones and Dangl, 2006). Under this selection pressure, some pathogens are adapted to overcome PTI by delivering effector proteins into the host to interfere with PTI. However, in the second stage, host cytoplasmic receptors can recognize many of these effectors and activate a defense termed effector-triggered immunity (ETI), which launches a hypersensitive response to kill host cells around the infection site to starve the pathogen. Pathogens may evade this recognition by mutating or deleting effector genes. Identification of these avirulence effectors as "Achilles heel" for these pathogens will provide valuable information for resistance strategies.

Genome assemblies of several rust pathogens have enabled initial cataloguing of putative effector genes. Usually, rust proteins are considered as candidate effectors if they contain a signal peptide for secretion, and no other targeting sequence or transmembrane domains (Saunders et al., 2012). Duplessis et al. (2011) implemented this selection method to predict 1,184 and 1,106 effector complements from *Mli* and *Pgt* genome sequences, respectively. Similarly, Cuomo et al. (2017) assembled a draft genome of *Pt*, annotated 15,685 protein-coding genes and characterized 1,358 of them as candidate effectors. When candidate effector genes are expressed during infection in haustoria, the structure considered as a major site for effector secretion, the expression is further evidence that they are likely effectors genes. In a second genome assembly project for *Pgt*, 520 genes were classified as candidate effectors based on their up-regulation in haustoria over germinated spores (Upadhyaya et al., 2015).

The life cycle of *Ph* is very similar to that of *Pgt* and *Pt*. Asexual urediniospores are produced, allowing the pathogen to attain epidemic levels on cultivated barley. Towards the end of the cropping cycle, the pathogen forms teliospores in which karyogamy occurs. Subsequently, promycelia are formed out of teliospore germination. The promycelia generate haploid basidiospores in which mating types segregate. Infection of basidiospores on alternate hosts leads to hybridization between monokaryons of different mating types. *Ph* can complete this sexual cycle on any one of five known alternate host species (*Ornithogalum brachystachys, O. trichophyllum, O. umbellatum, Dipcadi erythraeum* or *Leopoldia eburnea*) (Anikster et al., 1982; Wallwork et al., 1992). In Australia, five new pathotypes of *Ph* were isolated from the alternate host *O. umbellatum*, suggesting new virulence combinations in this pathogen can arise due to sexual recombination (Wallwork et al., 1992).

In the phylum Basidiomycota to which the rust fungi belong, mating loci are composed of at least two parts. One encodes pheromones and pheromone receptors, and the other encodes homeodomain (HD)-containing transcription factors (Raudaskoski and Kothe, 2010). The interaction of pheromones and its receptors enables formation of heterodimeric HD transcription factors coded by two HD-encoding genes HD1 and HD2 (Kües, 2015). Cuomo et al. (2017) identified two allelic pairs of the HD1 and HD2 genes, and three pheromone receptor genes in each of the *Pgt*, *Pt* and *Pst* species. As a close relative of the three wheat infecting *Puccinia* species, *Ph* would be expected to have the same copy number for the mating loci.

To obtain initial molecular information for *Ph*, a high quality draft genome sequence of this pathogen was generated. Gene model prediction was performed with the guide of the *Ph* transcriptome and homologous protein sequences from *Pgt*. Genomic characterizations including hallmarks of the obligate biotrophic lifestyle, mating type loci, and secreted proteins as candidate effectors were examined.

3.2 Results

3.2.1 *De novo* Assembly of *Ph*612

P. hordei isolate 612 was selected for the *de novo* genome assembly, as it contained fewer Avr genes and easier to amplify. Two genomic libraries were constructed from DNA extracted from urediniospores that had been serially increased from a single pustule. One was paired-end library with insert size 500 bp for deep sequencing (Table 3.7), the other was a mate-pair library with a long insert size 6,000 bp designed for resolving highly repetitive genomic regions. The sequencing reads of the paired-end library were assembled into 26,833 contigs adding up to 116 Mbp. The sequencing reads from the mate-pair library were then mapped to the contigs and connected into longer scaffolds. A total of 15,913 scaffolds were produced totaling 127 Mbp (Table 3.1), which is very close to a previously reported genome size of 122 Mbp obtained using flow cytometry (Kullman et al., 2005).

Table 3.1: Statistics of *Puccinia hordei* genome assembly

	Contig	Scaffold
Total Number (#)	26,833	15,913
Total Length (bp)	116,550K	127,347K
N50 (bp)	10,262	21,945
N90 (bp)	1,660	2,456
Min Length (bp)	200	1,000
Ave Length (bp)	4,343.5	8,002.7
Max Length (bp)	100,373	242,651
No. Sequence > 10Kbp	3,156	3,556
No. Sequence > 100Kbp	2	41
Gap region (bp)	0	10,797K
GC Content (%)	41.11	41.11

3.2.1.1 Continuity and Completeness

A commonly used measurement of assembly continuity is N50, which describes the average length of a set of sequences. It is defined as the length N that sequences longer than such length account for 50% of all bases. The N50 of the *Ph* scaffolds was 21,945 bp, comparable to another *Puccinia* species assembly of *Puccinia sorghi* (19,081 bp; Rochi et al., 2016). To examine gene space completeness in the *Ph* assembly, a pipeline called CEGMA (core eukaryotic gene mapping approach; Parra et al., 2007) was used to map 248 highly conserved eukaryotic genes to the assembly. This analysis showed 96.4% (239/248) of the CEGMA gene set was present in the assembly. This completeness is better than the CEGMA statistics for a *Pgt* genome assembly (94%;

Duplessis et al., 2011), and similar to other rust genome assemblies (97-98%; Table 3.8). In addition, a subset of the sequencing reads was selected randomly and mapped back to the scaffolds with an alignment rate of 78.06%, demonstrating a majority of the DNA library was assembled in the scaffolds.

3.2.1.2 Genome heterozygosity

The sequencing reads were derived from urediniospores, which comprise two heterozygous nuclei. Although the assembly algorithm was designed to assemble a consensus haploid genome, high heterozygosity in some genomic regions between the two nuclei may have prevented haploid assembly and resulted in two separate allelic scaffold sequences.

To investigate the amount of allelic scaffolds originating from a homologous chromosome pair, reads were mapped to all scaffolds with the software program Bowtie2. For each read, Bowtie2 can search up to N mapping positions in the scaffolds where N is an adjustable parameter. A setting of "N=1" reports only the best mapping for each read. With such a setting, allelic scaffolds showed half read coverage depth on merged scaffolds, because their coverage depth was equivalent to haploid genomic mass (55, Figure 3.1 A; Supplementary Table 3.1). In contrast, consensus or merged scaffolds showed read coverage depth of diploid genomic mass (110, Figure 3.1 A; Supplementary Table 3.1). When N was set to two, up to two scaffold mappings were allowed for each read. In another mapping with such a setting, allelic scaffolds gained diploid read depth as other merged scaffolds (Figure 3.1 C). As repetitive regions or gaps in scaffolds might have skewed average read depth for the scaffolds, the distribution of read depth over individual positions throughout the whole assembly was calculated for the two mappings (Figures 3.1 B and D). The "smoother" distribution curves (Figure 3.1 B versus A; D versus C) showed that the method was effective in filtering noise resulting from repeat and gap regions.

The two-peak distribution in the "N=1" mapping (Figure 3.1 A and B) indicated that a substantial number of allelic scaffolds resulted from high heterozygosity between the two dikaryotic nuclei. On the other hand, the transformation of two peaks to one

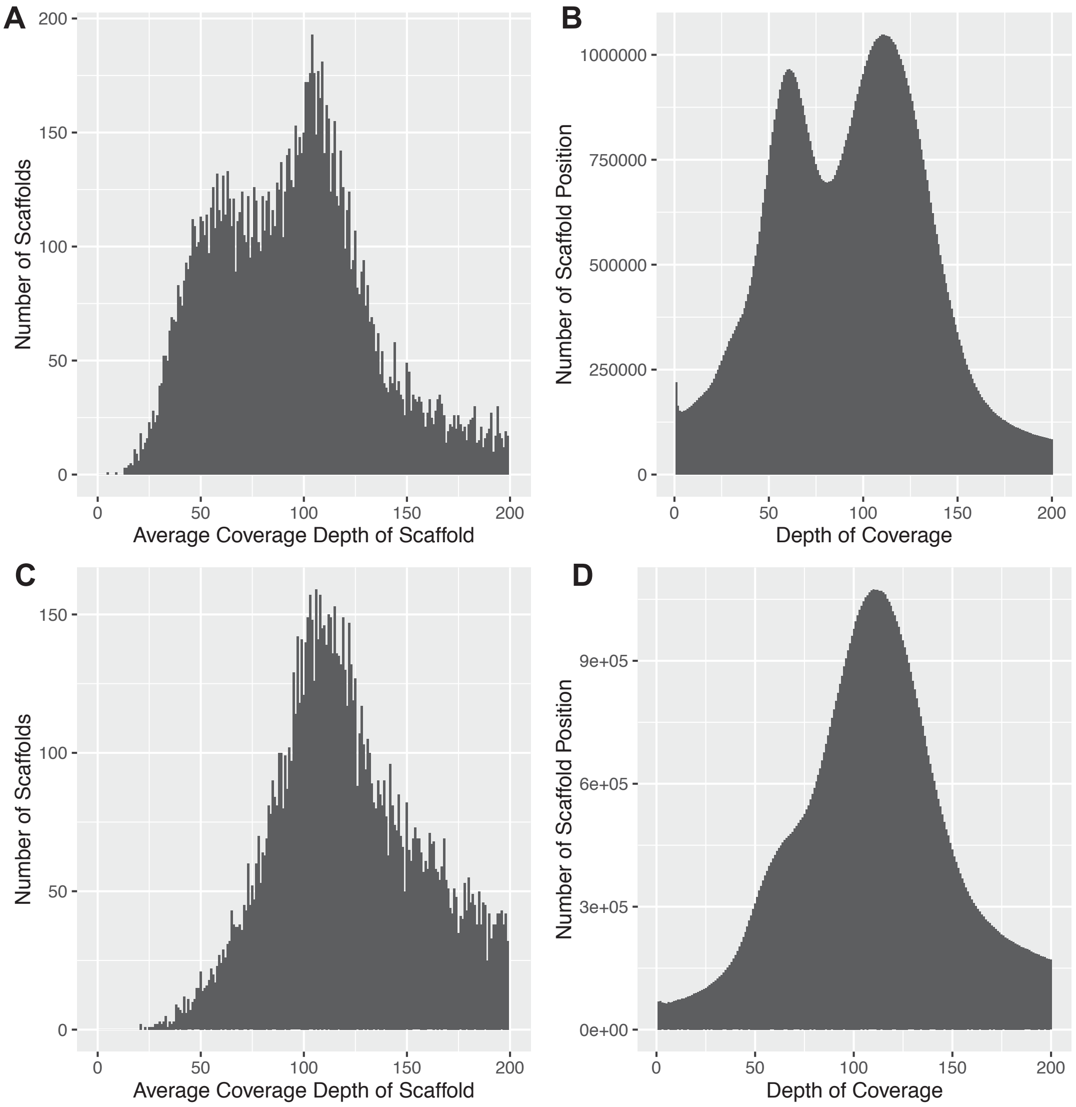


Figure 3.1: Distribution of read coverage depth as an indicator of allelic scaffolds. Reads were mapped to scaffolds with Bowtie2 using parameters allowing up to one or two optimal mapping positions. (A) shows average read depth over individual scaffolds in the one position only mapping, whereas (B) demonstrates read depth over individual positions in all scaffolds for this mapping. (C) shows average read depth over individual scaffolds when up to two positions allowed for read mapping, and (D) displays read depth over individual positions under the same condition.

with the “N=2” mapping (Figure 3.1 C and D) indicated that those allelic scaffolds still shared a certain degree of sequence similarity, in a manner that reads derived from one scaffold could be mapped to both allelic scaffolds. In addition to the observed sequence conservation between the allelic scaffolds, some allelic scaffolds did differ significantly, to an extent that their reads were not “mappable” to their corresponding allelic scaffold. This feature was shown as a minor yet still noticeable arch at depth 55 in Figure 3.1 D.

One instance of allelic scaffolds was Scaffold_738 and Scaffold_2365. Each of the two scaffolds contained one pair of genes encoding homeodomain-containing transcription factors HD1 and HD2 (identification details in the section Protein function survey). The two genes in Scaffold_738 were designated as *P1-HD1* and *P1-HD2*, and the other pair in Scaffold_2365 was referred to as *P2-HD1* and *P2-HD2*. Average read depths of gene regions in the two scaffolds were 51.4 fold and 53.1 fold respectively, whereas the depth of a consensus Scaffold_3, a randomly selected control, was about two fold of the putative allelic scaffolds (102.5 fold; Supplementary Table 3.1). In addition, the read mapping on the homeodomain encoding genes showed a haploid allele lacking heterozygous variation, whereas a randomly selected control segment in Scaffold_3 on positions 23,500-26,000 bp showed diploid heterozygosity (Figure 3.2).

To estimate heterozygosity between the two nuclei of the dikaryotic *Ph* urediniospores, single nucleotide polymorphisms (SNP) were detected based on reads mapping back to the scaffolds (N=2). In total, 602,188 SNPs were identified across all scaffolds, out of which 548,638 SNPs were in a heterozygous condition. The frequencies of total SNPs and heterozygous SNPs were 4.7 and 4.3 SNPs/kb, respectively.

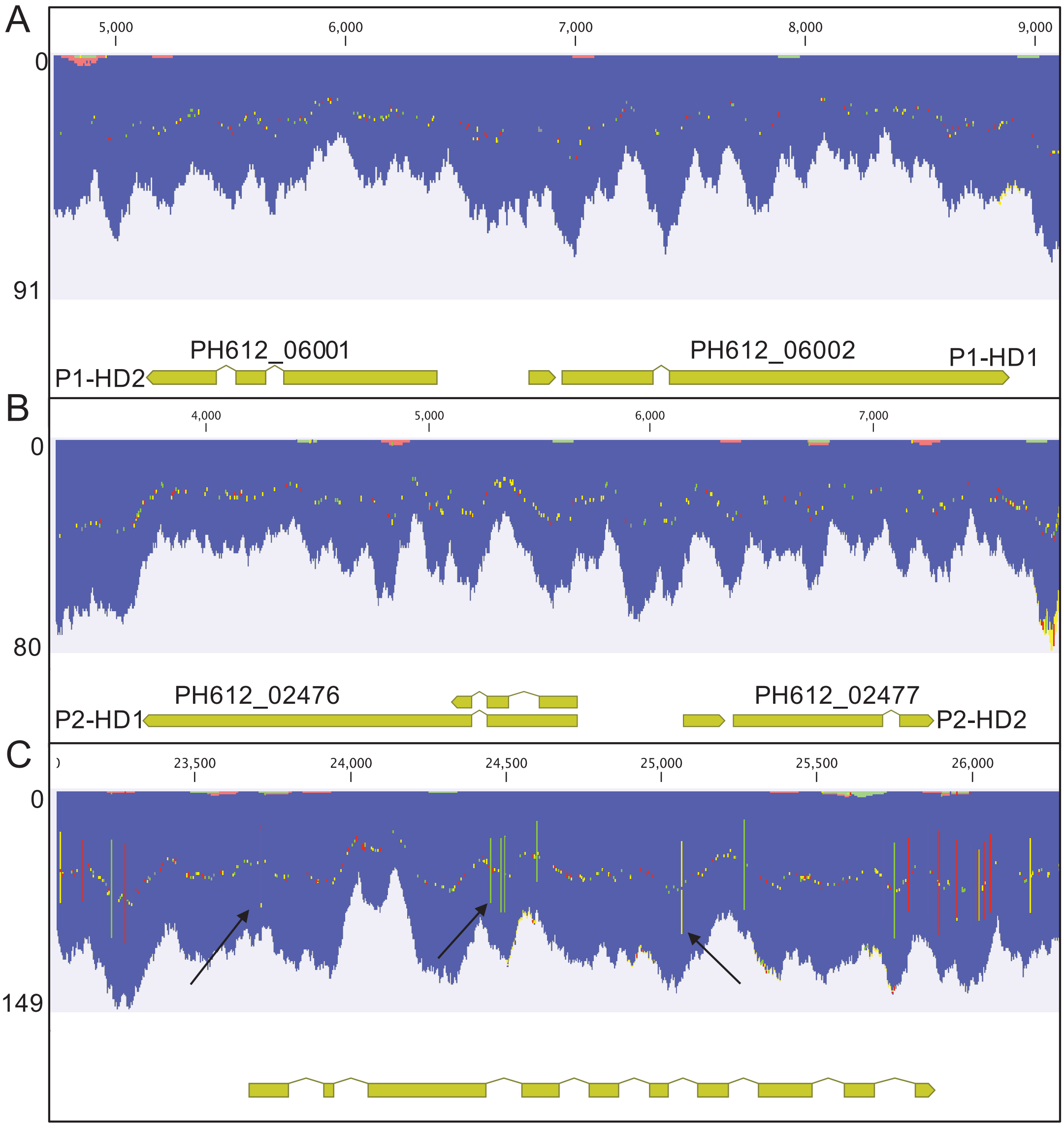


Figure 3.2: Coverage graph of reads mapping to local regions of three scaffolds 738 , 2365 ,and 3. (A) shows positions 5000-9000 bp of Scaffold_738 which harbor a pair HD-containing transcription factors genes *PH612_06001* and *PH612_06002*; (B) shows Scaffold_2365 that harbors the other pair of HD encoding genes *PH612_02476* and *PH612_02477*; (C) shows an example of heterozygous positions on Scaffold_3. The colorful bars (three examples pointed with arrows) indicate nucleotide variation of reads to the reference scaffold, in contrast to blue background indicating residue matching the reference. The bar colors green, yellow, red, light blue represent T, G, A and C variants to the reference, respectively.

3.2.1.3 Phylogenetic analysis of *P. hordei*

The phylogenetic relationship of several rust pathogens including *Ph* has been studied using ribosomal DNA sequence alignment (Zambino and Szabo, 1993). To confirm *Ph*'s evolutionary relationships with other rust fungi using the newly assembled genome information, an analysis was performed with the 18S rRNA biogenesis proteins from *Ph*, *Pt*, *Pst*, *Pgt*, *P. sorghi* and *Mli*. Initially, the position of the biogenesis gene in *Ph* scaffolds was identified by aligning the *Pgt* 18S rRNA biogenesis protein (NCBI accession PGTG_02406) to the scaffolds. Secondly, the locus identified (92,356-93,801 bp on Scaffold_29) was translated to protein *in silico* after removing intron sequences. Finally, the protein sequences of the six rust fungi were aligned and their sequence substitutions were used to infer a phylogenetic tree. The branch of *Mli* and other *Puccinia* species was midpoint to set *Mli* as an outgroup. The results from this comparison indicated that *Ph* was most closely related to *Pt* (Figure 3.3), and the topological relationships of *Ph* to other rust species were consistent with those reported by Zambino and Szabo (1993). With the whole genome assembly of *Ph*, future phylogenetic analysis can be expanded to a larger set of core genes to provide higher confidence.

To further confirm that *Ph* is genetically closer to *Pt* than to other fungi, the proteins of *Ph* (introduced in section Protein function survey) were aligned to the proteomes of *Pt*, *Pgt*, *Pst* and *Mli* individually. Consistently, *Pt* showed largest proteome conservation (61.25%) with *Ph*, whereas *Pgt*, *Pst* and *Mli* showed descending conservation with ratios of 58.73%, 46.63%, and 44.19%, respectively (Table 3.2). These results were consistent with the phylogenetic tree that showed the closest relative of *Ph* is *Pt*, followed by *Pgt*, *Pst* and *Mli* (Figure 3.3).

3.2.1.4 Repeat content

All currently published rust genomes were shown to contain high percentage of repeat elements (30-50%: (Cantu et al., 2011; Cuomo et al., 2017; Duplessis et al., 2011; Zheng et al., 2013)). An analysis using RepeatModler revealed that 55.28% of the *Ph*

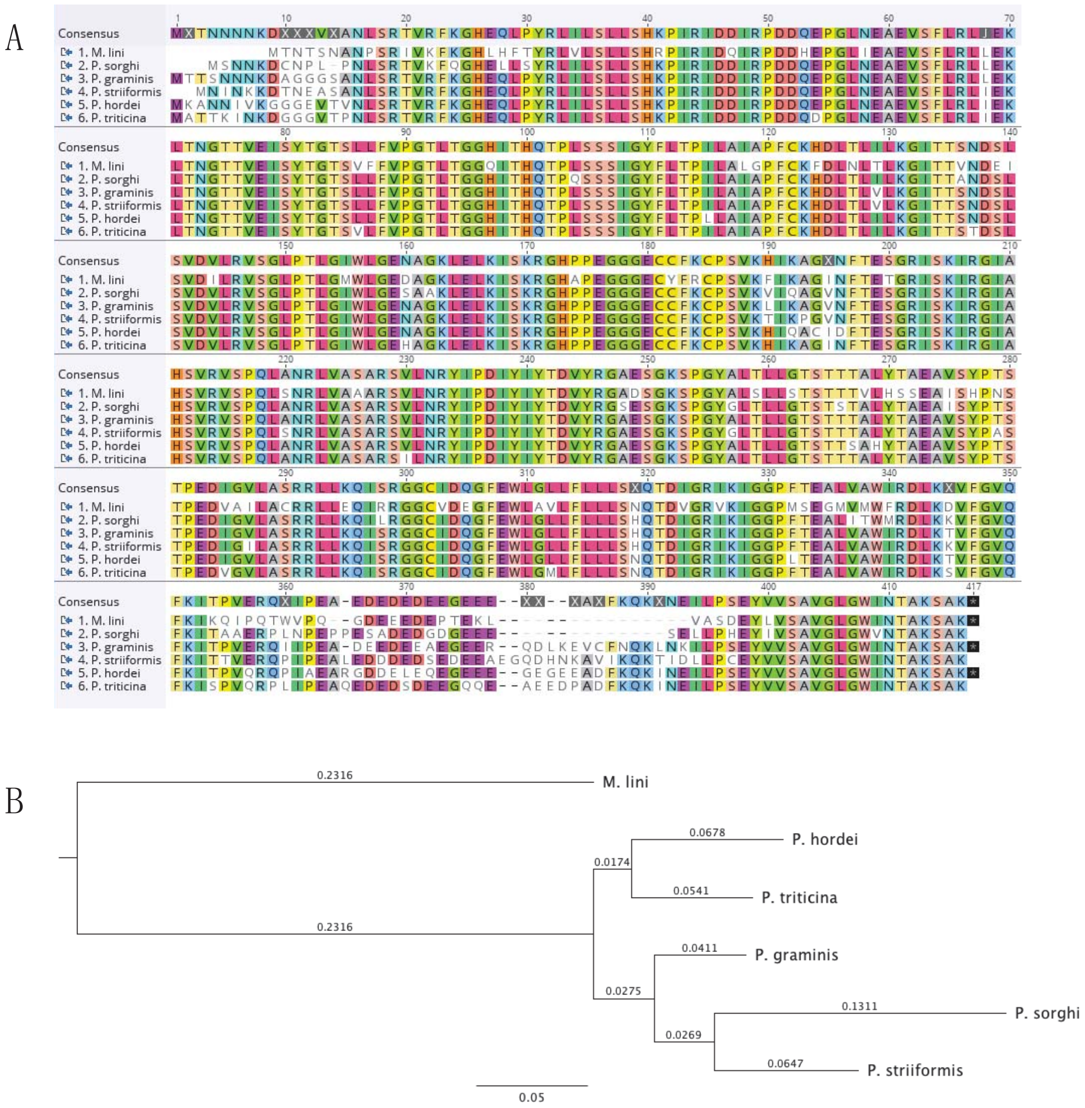


Figure 3.3: Multiple sequence alignment of 18s rRNA biogenesis proteins and phylogenetic inferences. (A) The protein sequences were from *M. lini* (selected as *Puccinia* outgroup; ID TU.MELLI_sc114.1), *P. hordei* (PH612_14511), *P. triticina* (OAV98697), *P. graminis* (XP_003321364), *P. sorghi* (KNZ50814), and *P. striiformis* (KNF00205). (B) Phylogenetic tree was calculated using PhyML based on the protein alignment. The branch lengths were proportional to number of amino acid substitutions in the alignment.

assembly was composed of repeat elements, ranking it highest among the published rust genomes in this regard (Table 3.3). The majority of the repeats are interspersed

Table 3.2: Protein conservation between *P. hordei* and other rust species

Rust species	Total proteins	Conserved proteins in *Ph*	Ratio
Puccinia triticina	15,685	9,623	61.35%
Puccinia graminis f. sp. *tritici*	15,979	9,384	58,73%
Puccinia striiformis f. sp. *tritici*	20,502	9,561	46,63%
Melampsora lini	16,339	7,221	44,19%

repeat elements that include transposons, which can replicate and integrate randomly into the genome and subsequently cause expansion.

Table 3.3: Classification of predicted repetitive elements

	# elements	Length	Ratio
Small interspersed:	72	26,972 bp	0.02%
Long interspersed:	3,214	2,159,880 bp	1.70%
Long terminal repeat:	56,765	29,885,627 bp	23.47%
DNA elements:	49,069	15,248,156 bp	11.97%
Unclassified:	69,427	21,295,947 bp	16.72%
Total interspersed repeats:		68,616,582 bp	53.88%
Small RNA:	54	18,476 bp	0.01%
Simple repeats:	19,006	1,817,855 bp	1.43%
Low complexity:	3,102	232,528 bp	0.18%
Total tandem repeats:		2,068,859 bp	1.62%
Total repeat content		70,403,062 bp	55.28%

3.2.2 Gene prediction guided by RNA-Seq and *Pgt* proteins

To guide gene annotation, RNA-Seq was performed for infected barley leaf tissues collected at two time points (4th and 7th day) post *Ph* inoculation. The RNA-Seq reads were aligned to *Ph* scaffolds with TopHat2. A total of 36% of the reads were successfully aligned and retained, whereas the remaining 64% of reads were discarded as derived from the barley leaf transcriptome. The read alignment was then assembled to *Ph* transcripts using software Cufflink, which were subsequently used to train a gene annotation tool CodingQuarry. The tool predicted an initial set of 29,520 genes. To minimize false positive predictions, these genes were filtered based on the RNA-Seq expression data. Specifically, genes with less than 70% coverage by the RNA-Seq read mapping were considered as lacking expression data support and were removed, leaving a set of 11,882 gene predictions. The RNA-Seq guided annotation should capture a majority of the genes expressed *in planta* at the two infection time points.

To obtain a comprehensive gene repertoire, various life stages should be covered (e.g. urediniospore germination). Therefore, *Pgt* proteins predicted in the Australian *Pgt* Pan-genome project (Upadhyaya et al., 2015) were used to further improve the *Ph* gene annotation. The prediction of the *Pgt* proteins was based on RNA-Seq of geminated urediniospores and purified haustoria. As *Pgt* and *Ph* are close relatives in the *Puccinia* lineage, mapping the *Pgt* protein sequences to the *Ph* genome should provide clues to gene locations and intron-exon boundaries for the *Ph* gene prediction. A pipeline called MAKER was used for the protein mapping and resulted in 8,881 gene predictions in the *Ph* genome. To remove the redundancy of gene sets produced by CodingQuarry and MAKER, the 8,881 MAKER-predicted genes were filtered based on their overlap with the CodingQuarry-predicted genes (Figure 3.7). After filtering, 4,472 MAKER genes remained and were concatenated with the 11,882 CodingQuarry-predicted genes to give a final set of 16,354 genes.

The gene regions, including exons and introns, accounted for 16.6% of the genome assembly (Table 3.4), with an average gene length of 1,278.3 bp. Average numbers of exon and intron per gene were 3.61 and 2.61, respectively, whereas average lengths of these two genic regions were 190.7 bp and 87.6 bp, respectively. The predicted genes were named with a common prefix “PH612” showing origin from *Puccinia hordei* isolate 612, and a unique suffix number derived from the prediction tools CodingQuarry and MAKER (e.g. *PH612_12290*).

Table 3.4: Statistics of predicted genes

Feature	Number
No. genes predicted with RNA-Seq	11,882
No. genes predicted with *Pgt* proteins	4,472
Gene regions (bp)	21,144,698
Average gene length[1]	1,278.3
Average coding sequence length	952.2
Average exon length	290.7
Average exon number per gene	3.61
Average intron length	87.6
Average intron number per gene	2.61
GC content in coding regions	48.0%

[1] Gene length is defined as number of nucleotides from start to end of coding sequence. For genes with multiple isoforms, the longest coding sequence was used.

3.2.3 Functional annotation for predicted genes

To annotate the predicted genes with functional information, their encoded proteins (longest protein was selected for multiple isoform genes) were searched against the NCBI Non-redundant (Nr) protein database using BLAST. When homologous proteins were identified, the corresponding homolog function descriptions were assigned to the *Ph* genes (Supplementary Table 3.2). For instance, *PH612_06830* was annotated to be a 3-isopropylmalate dehydrogenase, a catalysis enzyme for chemical reaction, based on its protein similarity with a *Pgt* protein PGTG_02188. For *Ph* proteins that had several homologs from the database, only the homolog with the most significant E-value was reported here. For most *Ph* genes, their most similar homologs were from *Puccinia* species, then from *Melampsora* species and then other plant pathogens including species of *Rhizoctonia* and *Moniliophthora* species (Figure 3.4).

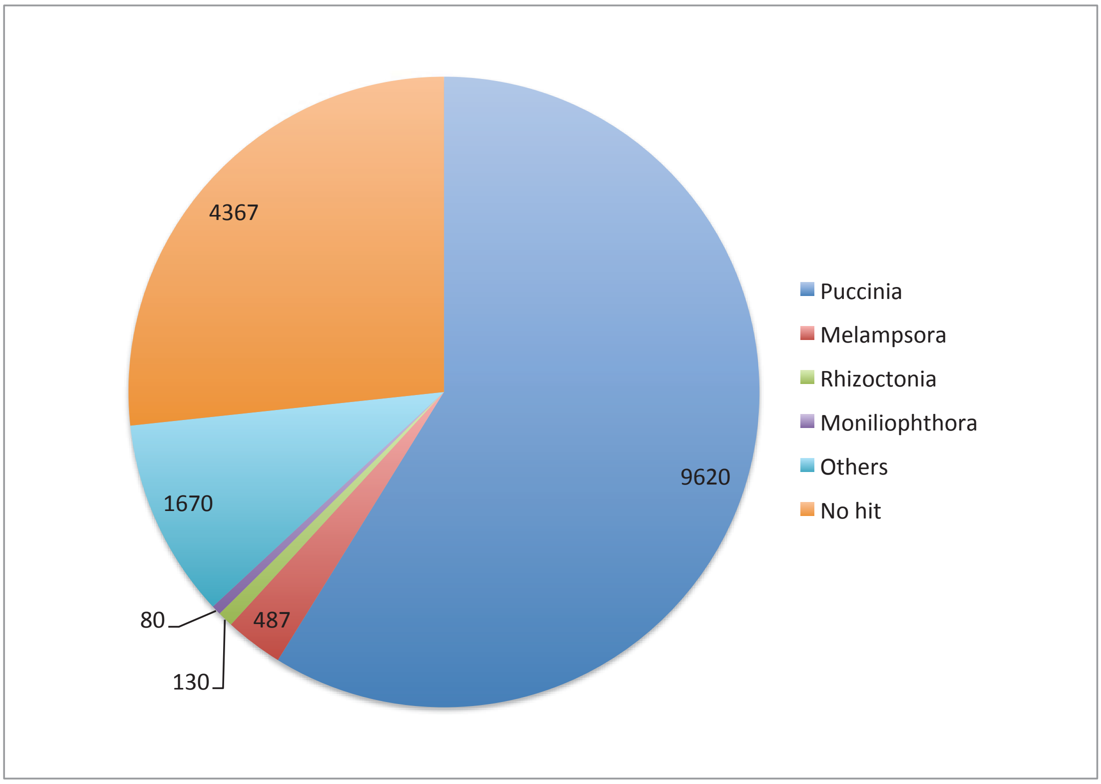


Figure 3.4: Distribution of NCBI Non-redundant database hits. The predicted proteins of *Puccinia hordei* were searched against NCBI non-redundant protein database for homologs, and the closest homolog for each *P. hordei* protein was returned. This pie chart shows organisms from which the homologs are derived.

The BLAST based annotation identified homologous proteins in Nr database, but it did not give any indication of protein families or functional domains for *Ph* proteins. To fill this gap, *Ph* proteins were scanned for matches in the InterPro protein database that curates protein classifications and families by using InterProScan (IPS). The descriptions of protein families revealed by IPS were added to the functional annotations (Supplementary Table 3.2). About 51.1% of the total genes could be classified into one or more protein families, or be associated with functional domains. This annotation approach provided a variety of functional information for *Ph* genes, including those relevant to the mechanisms of plant pathogenicity. For instance, *PH612_10432* is annotated with Pectinesterase activity (InterPro entry IPR011050) that catalyzes the hydrolysis of pectin families.

3.2.3.1 Protein clustering of six rust species

Family detection of proteins from different rust species can provide an overview of lineage specific proteins that implicate lineage functionality. The proteins of currently deep-sequenced rust genomes including two *Melampsora* species (*Mlp* 98AG31 and *Mli* CH5) and four *Puccinia* species (*Ph* isolate 612, *Pgt* isolate CDL75-36700-3, *Pst* isolate 2K41-Yr9, and *Pt* Race BBBD) were compared in a pair-wise manner for sequence similarity detection using BLASTP. The output of a similarity matrix was used to infer protein families using a Markov Model Clustering tool TRIBE-MCL (Enright et al., 2002). In total, 101,019 proteins from the six rust fungi were classified into 10,157 clusters (Supplementary Table 3.3). A total of 2,180 clusters contained only one member, and these proteins were referred to as orphans. Among them, there were 544 orphan proteins of *Ph* that are likely to play a key role in *Ph* speciation and host specialization because they were present exclusively in *Ph*. The majority of clusters (about 60%) had from 2-10 members from different taxa, showing conserved protein families across different species (Figure 3.5).

To evaluate the accuracy of the clustering, cluster assignment of 11 known pheromone receptors from four rust fungi (*Pgt*, *Pt*, *Pst*, and *Ph*; *Ph* pheromone receptors identification will be described in the subsequent section Protein function survey) was checked. All of them were allocated to a single cluster, 774, which contained 19 proteins from all six species. Figure 3.6 demonstrates 80 clusters with their protein species origin. The member allocation in different clusters reflects their evolutionary relationship. For example, clusters containing members from all six rust fungi were relatively large and contain a similar number of members from each species, suggesting no significant evolutionary shift to expansion or reduction in these protein families. Several clusters, such as C41, C46, and C55, contained proteins specific to the two *Melampsora* species, whereas some clusters (e.g. C31, C32, C47 and C65) contained only *Puccinia* species

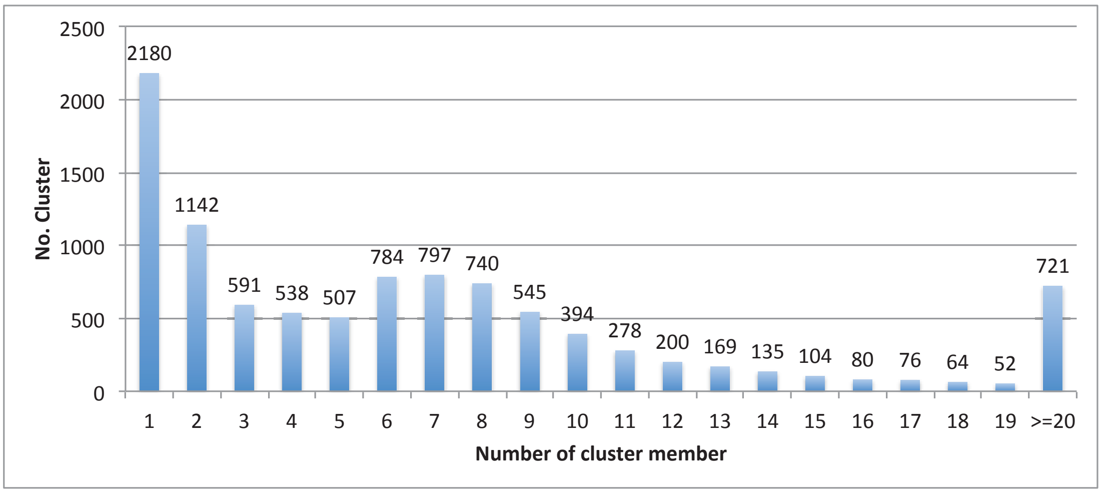


Figure 3.5: Distribution of members in protein clusters. A total of 101,019 protein from six rust species (*Puccinia graminis* f. sp. *tritici, P. triticina, P. striiformis* f. sp. *tritici, P. hordei, Melampsora lini* and *M. larici-populina*) were allocated into different 10,157 clusters based on their pair-wise sequence similarity. This figure shows number of clusters (Y-axis) that contain a certain amount of protein members (X-axis).

proteins. This result delineates lineage specific proteins and provides a useful guide for functional analysis.

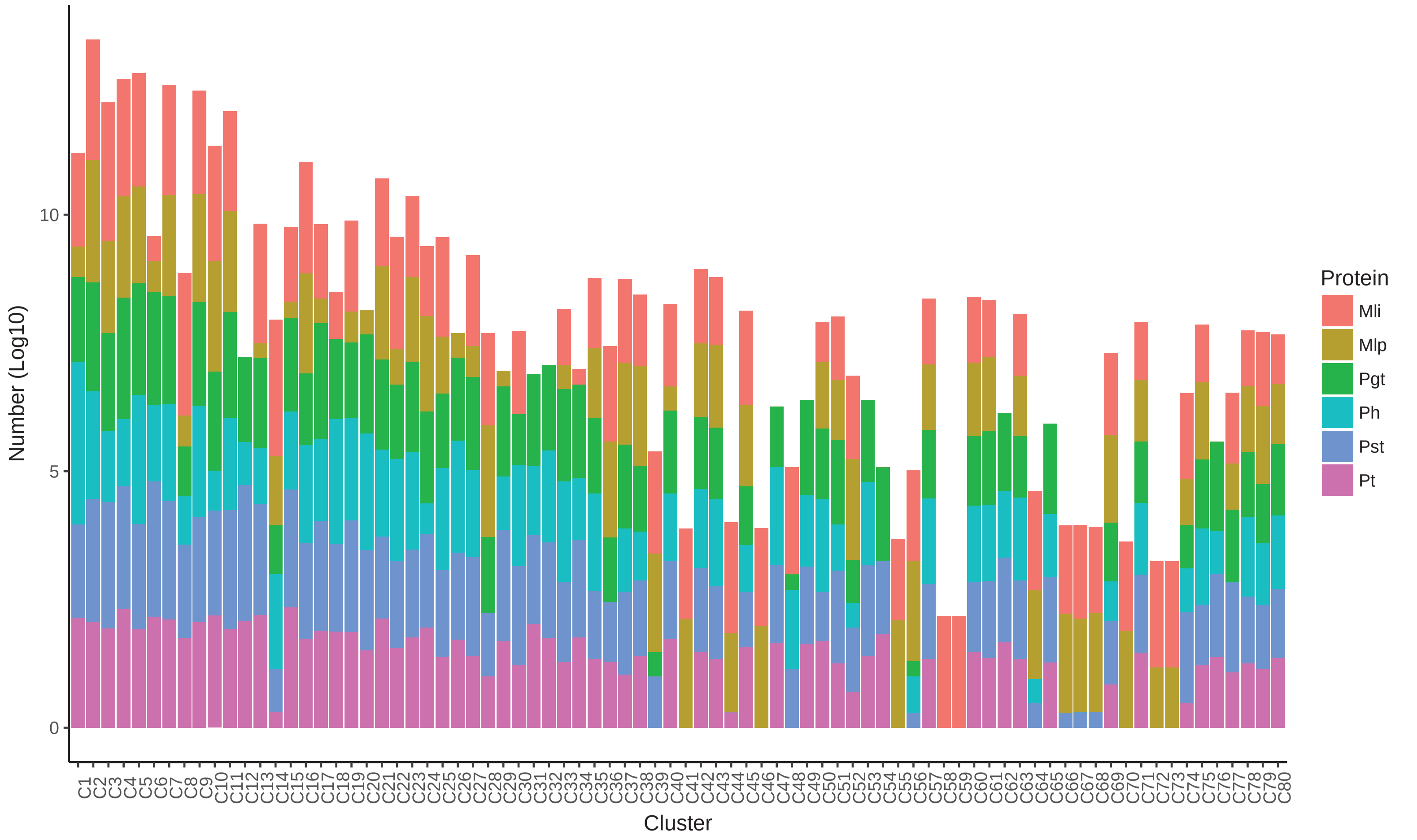


Figure 3.6: A demonstration of 80 protein clusters with different proportions of six species. The Y-axis shows number of proteins in a specific cluster (X-axis) with different colors displaying different rust species. The color legends *Mli, Mlp, Pgt, Ph, Pst,* and *Pt* are acronyms for *Melampsora lini, M. larici-populina, Puccinia graminis* f. sp. *tritici, P. hordei, P. striiformis* f. sp. *tritici and P. triticina*

3.2.4 Protein function survey

3.2.4.1 Mating-type associated genes

Ph is a heteroecious fungus that completes its asexual cycle on primary hosts and sexual cycle on alternate host species. During karyogamy in the sexual cycle, haploid basidiospores are generated within which mating types segregate. In other Basidiomycetes, mating type is mediated by two types of loci. One includes pheromone and pheromone receptor genes, and the other contains HD1 and HD2 homeodomain transcription factor genes (Coelho et al., 2017).

The interaction of pheromone and pheromone receptor from different mating types activates the pathway for hyphal fusions and septal dissolution. To identify the pheromone in *Ph*, the *Ph*612 putative proteins were searched against *Pst* mfa2 (Cuomo et al., 2017) using BLAST, however, no significant hits were found. The failure of BLAST identification was likely due to the short length of the *Pst* mfa2 (33 amino acids) that resulted in the inability to pass the E-value threshold of e-10. An alternative search with pheromone conserved motifs (CVLT, CILT and CIIC; Cuomo et al., 2017) found a predicted protein PH612_03487 with motif CIIC and 33 amino acids, the same length as *Pst* mfa2. Manual comparison of the two protein sequences PH612_03487 and *Pst* mfa2 revealed that they shared 20 amino acids with a sequence identity of 60.61%. Taken together, these results indicate that *PH612_03487* is highly likely to encode a pheromone.

Using a similar approach, two putative pheromone receptors, PH612_03488 and PH612_11123, were identified with close homology to pheromone receptors in other two rust fungi *Pgt* and *Pt* (NCBI Accession: PGTG_01392, PGTG_19559, PGTG_00333, PTTG_09751, PTTG_28830, PTTG_09693). InterProScan (IPS) showed that the two proteins belong to family IPR001499 with annotation of "GPCR (G protein-coupled receptors) fungal pheromone mating factor". Two classical members of this family IPR001499 are Uhpra1 and Uhpra2 from *Ustilago hordei*, which were shown to be involved in the initiation of the sexual cycle (Anderson et al., 1999). The pheromone gene *PH612_03487* and pheromone receptor *PH612_03488* were located in close proximity on Scaffold 3215. The adjacency of the two genes in the genome is a conserved structure observed in other Basidiomycete fungi (Kües, 2015).

When mating cells fuse, the HD1 and HD2 homeodomain containing proteins heterodimerize to produce a transcription factor that induces expression of genes required for sexual development (Spit et al., 1998). BLAST searches for *Ph* homologs to *Pt* HD1 and HD2 homeodomain containing proteins (PTTG_27730 and PTTG_03697 respectively; Cuomo et al., 2017) identified two HD1 proteins (PH612_06002 and PH612_02476) and two HD2 proteins (PH612_06001 and PH612_02477). All four proteins contained a homeobox domain annotated with InterPro entry IPR001356 that is involved in transcription regulation by binding to DNA via a helix-turn-helix structure. While *PH612_06001* and *PH612_06002* were located on Scaffold 738 at position 5,132-8,889 bp, *PH612_02476* and *PH612_02477* were located on Scaffold 2365 at position 3,713-7,117 bp. The two scaffolds were shown to be allelic sequences in the section Genome heterozygosity. On each scaffold, the HD1 and HD2 genes were transcribed in opposite directions (Figure 3.2 A and B), referred to as divergent transcribed structure in previous studies (Fraser and Heitman, 2004; Kües, 2015).

3.2.4.2 Obligate biotrophic lifestyle reflected by the putative proteins

It has been hypothesized that the obligate biotrophic lifestyle of rust fungi is associated with a reduced need/ability for inorganic nitrate and sulfate assimilation, along with transporter family expansion to enhance nitrogen and sulfur uptake from the host. (Duplessis et al., 2011). The genome assembly and gene prediction of *Ph*612 offer a reference to explore these features in *P. hordei*. A BLAST search was performed to query *Ph*612 predicted proteins against a custom database composed of metabolic components of *Mli* curated by Nemri et al. (2014). The search results provided evidence of homology of *Ph* proteins to *Mli* metabolites. In addition, the IPS annotations were also scanned to consider the presence of specific metabolic molecules.

Nitrate assimilation involves reduction of nitrate to ammonium, a process catalyzed by the enzymes nitrate reductase and nitrite reductase (Crawford and Arst Jr, 1993). The BLAST search showed that two *Ph* proteins PH612_04830 and PH612_06223 were homologous to the putative *Mli* nitrate reductase MELLI_sc3720.2 (E-value e-28). As no nitrite reductase was identified in the *Mli* genome, the BLAST search could not verify the presence or absence of this protein in *Ph*. Further inspection with IPS annotations did not identify any proteins annotated with nitrite reductase function, either. In addition, a tBLASTn search of the nitrite reductase from *Bacillus mycoides* (accession: OOG90654.1) in the *Ph* assembly did not find any significant hit. Therefore, the nitrate assimilation is probably deficient in *Ph* and hence the inorganic nitrogen assimilation pathway is likely non-functional.

The nitrate assimilation deficiency can be compensated by nitrogen uptake from the host mediated by ammonium and amino acid transporters. One protein PH612_02741 was homologous to *Mli* putative ammonium transporters[1]. IPS annotation of this protein showed that it contained an ammonium transporter AmtB-like domain (IPR024041), which is associated with ammonium transmembrane transporter activity. The RNA-Seq data suggested that this gene was expressed *in planta* during infection, indicating its potential role in ammonium transport. With a set of catalytic enzymes, the acquired ammonium can be assimilated to glutamate and glutamine used for amino acid biosynthesis.

To investigate the presence of key enzymes in *Ph* involved in this assimilation reaction, a BLAST search was checked to identify homology of *Ph* proteins to the *Mli* enzymatic components. Proteins with significant hits (E-value <e-10) were selected and then inspected manually for their IPS annotations to further confirm their enzymatic functions. All components were identified (Table 3.5). In summary, the identification of the putative ammonium transporter and enzymatic components for ammonium assimilation demonstrated that this pathway was very likely complete in *Ph*.

In addition to AA biosynthesis, *Ph* may obtain AA directly from the host via amino acid and peptide transporters. According to the IPS annotation, 39 proteins have the function of amino acid transportation (Supplementary Table 3.4). Among them, 31 proteins belong to the family Amino acid/polyamine transporter I (IPR002293), which are integral membrane proteins containing up to 12 transmembrane segments. In addition, there are 45 oligopeptide transporters annotated with IPR004813 (Supplementary Table

[1] MELLI_sc457.12, MELLI_sc152.7, and MELLI_sc152.8

Table 3.5: Putative enzymes for ammonium assimilation

Enzyme	*Mli* protein ID	*Ph* protein ID	E-value	IPS Annotation
Glutamate synthase	MELLI_sc11.10_sc11.11	PH612_02653	0.0	IPR002489
	MELLI_sc11.10_sc11.11	PH612_02654	0.0	IPR002489
Glutamine synthetase	MELLI_sc3079.2	PH612_05569	9.51e-156	IPR008146
Glutamate dehydrogenase	MELLI_sc1197.2	PH612_10966	2.82e-156	IPR006096
	MELLI_sc1197.2	PH612_06237	2.71e-132	NA
	MELLI_sc1197.2	PH612_06238	0.0	NA
Aspartate aminotransferase	MELLI_sc1978.2	PH612_13303	0.0	IPR000796
	MELLI_sc1978.2	PH612_10503	2.45e-128	IPR000796
Asparagine synthase	MELLI_sc1683.3	PH612_02840	1.09e-171	IPR001962
	MELLI_sc1683.3	PH612_09948	5.51e-35	IPR001962
	MELLI_sc1683.3	PH612_11343	8.87e-41	IPR017932
Asparaginase	MELLI_sc2460.3	PH612_13918	0.0	IPR002110

3.4).

Sulfur is an essential component for the synthesis of sulfur-containing AA, cysteine and methionine. Inorganic sulfate assimilation in filamentous fungi is catalyzed by four essential enzymes: ATP-sulfurylase, adenosine phosphosulfate kinase, phospho-adenosine phosphosulfate reductase and sulfite reductase (Marzluf, 1997). Because ATP-sulfurylase was not identified in *Mli*, a randomly selected ATP-sulfurylase from the mould fungus *Aspergillus parasiticus* (NCBI accession KJK67935.1) was used as a probe for a BLAST search of the *Ph* putative proteins. The search (Evalue threshold e-10) did not identify any homologs of ATP-sulfurylase, similar to the absence of this enzyme in *Mli* and *Mlp* (Duplessis et al., 2011; Nemri et al., 2014). This deficiency suggested that *Ph* may assimilate sulfur via unknown metabolic pathway(s). In support of this speculation, a scan of IPS annotations identified three proteins (PH612_10998, PH612_06285, and PH612_15860) with predicted lysosomal cystine transporter function (IPR005282). This finding indicated that *Ph* may take up cysteine directly from its host during infection.

As an obligate biotrophic parasite, *Ph* needs to assimilate nutrients from hosts with efficient transport systems. A total of 342 proteins were annotated with transporter functions by IPS (Supplementary Table 3.4), including the 39 amino acid transporters and 48 oligopeptide transporters previously mentioned. At least 105 *Ph* proteins were from the Major Facilitator Superfamily (MFS), a family of membrane transporters that contribute to movement of small solutes across membranes under chemiosmotic gradients. For instance, PH612_12781 from MFS had been identified as a homologue to hexose transporter HXT1p from *Uromyces fabae*, which was preferentially expressed in haustoria for sugar transportation (Voegele et al., 2001). In addition, 42 P-type ATPases were identified. These enzymes were involved in ATP hydrolysis to provide energy for ion transport across membranes. Thirty-three out of 42 proteins had an annotated HAD-like domain (IPR023214), a conserved alpha/beta-domain functioning as a hydrolase fold. In addition, a cytoplasmic domain N (IPR023299) that comprises the nucleotide binding site, and a conserved N-terminal domain (IPR004014) responsible for sequential phosphorylation inducing conformational changes with unknown ATPase regulation function have been detected in the 42 ATPases. Thirty-five ATP-binding cassette (ABC) transporters were also identified, which belong to an omnipresent transporter family using the energy generated from ATP hydrolysis for export or import of a wide variety of substrates ranging from small ions to macromolecules. The other transporters include phosphate, magnesium, and copper transporters as well as members from small transporter family.

3.2.4.3 Effector prediction and functional enrichment analysis

Rust fungi secrete effector proteins into the plant apoplast or cytoplasm via membrane translocation. The signal peptides for secretion at the N termini of proteins can be used to derive the *Ph* secretome. A scan of the *Ph* proteome with SignalP4.1 identified 1,115 proteins with a signal peptide. Of these proteins, 43 contained a predicted transmembrane segment, which was associated with membrane integration function. Therefore, 1,072 proteins were predicted as candidate secreted proteins (Supplementary Table 3.5).

Enrichment analysis based on IPS annotations

To highlight biological functions significantly enriched in the SP genes, IPS annotations were compared between secreted and non-secreted proteins. A Chi-squared test was applied to evaluate the significance of enrichment of each IPS annotation in secreted versus non-secreted proteins, resulting 123 InterPro entries ($p<0.05$; Supplementary Table 3.6). Out of these 123 entries, 100 IPS annotations corresponded to typical enzymes with pathogenicity functions. These included enzymes likely to be involved in plant cell wall degradation during infection, such as glycoside hydrolase (IPR017853), cutinase (IPR000675), glucoamylase (IPR000165), and carboxylesterase (IPR019826). Additionally, 15 annotations implicated peptide catalytic enzymes, which might have a potential role in targeting host proteins either to suppress immune response or to degrade host protein for peptide uptake. Such entries include serine carboxypeptidase (IPR001563 and IPR018202) and subtilases (IPR023827).

The remaining 23 entries are non-enzymatic and some of them have annotated functions highly relevant to rust pathogenicity (Table 3.6). For instance, the enrichment of cysteine (IPR001283) has been reported as an important criterion for fungal effector selection (Saunders et al., 2012). Also related to cysteine residue, the CFEM domain (Common in Fungal Extracellular Membrane; IPR008427) is an eight cysteine motif exclusive to fungal extracellular membrane proteins with a proposed pathogenesis function for appressorium development in *Magnaporthe grisea* (Kulkarni et al., 2003). In addition, the domain has also been identified in the secretome of another rust species *Mlp* (Joly et al., 2010). It is plausible to hypothesize that the enrichment of this domain in *Ph* secreted proteins could be related to haustorial (a specialized tissue similar to appressorium) development during infection. The phosphatidylethanolamine-binding protein family (IPR008914), highly conserved across a wide variety of prokaryotic and eukaryotic species, has also been noted. Proteins from this family have documented functions of lipid binding, serine protease inhibition, and the participation in crucial signalizing pathway, such as Ras-Raf-MEK-ERK pathway which relays cell signaling from membrane receptor to nuclear DNA. The MD-2-related lipid-recognition domain (IPR003172) has been implicated in the recognition of pathogen related products, such as lipopolysaccharide, which have been detected in the haustorial neck region of mint rust pathogen *P. menthae* (Larous et al., 2008). In summary, this enrichment analysis of the *Ph* secretome based on IPS annotations not only draws a big picture of secreted proteins with biological functions relevant to *Ph* pathogenesis, but also provides crucial clues for future functional experiments.

Table 3.6: Non-enzymetic InterPro entries enriched in the predicted secretome

InterPro ID	Description	Not secreted	Secreted	P-value	Enrichment
IPR001283	Cysteine-rich secretory protein	0	6	2.24E-20	NA
IPR008427	CFEM domain	0	5	3.06E-17	NA
IPR003172	MD-2-related lipid-recognition domain	1	4	3.20E-11	57.2
IPR008914	Phosphatidylethanolamine-binding protein	1	3	3.17E-08	42.9
IPR001938	Thaumatin	0	3	6.14E-11	NA
IPR032514	Domain of unknown function DUF4965	1	5	2.95E-14	71.6
IPR018466	Kre9/Knh1 family	2	5	4.01E-12	35.8
IPR009020	Proteinase inhibitor, propeptide	2	5	4.01E-12	35.8
IPR029070	Chitinase insertion domain	4	6	8.43E-12	21.4
IPR006153	Cation/H+ exchanger	5	6	1.24E-10	17.1
IPR012946	X8 domain	1	3	3.17E-08	42.8
IPR000782	FAS1 domain	1	3	3.17E-08	42.8
IPR021476	Protein of unknown function DUF3129	4	4	6.82E-07	14.3
IPR009078	Ferritin-like superfamily	2	3	1.36E-06	21.4
IPR001902	SLC26A/SulP transporter	1	2	2.58E-05	28.5
IPR002645	STAS domain	1	2	2.58E-05	28.5
IPR011511	Variant SH3 domain	1	2	2.58E-05	28.5
IPR015202	Domain of unknown function DUF1929	1	2	2.58E-05	28.5
IPR009011	Mannose-6-phosphate receptor binding domain	1	2	2.58E-05	28.5
IPR013766	Thioredoxin domain	6	3	1.16E-3	7.1
IPR001762	Disintegrin domain	3	2	2.51E-3	9.5
IPR013126	Heat shock protein 70 family	7	3	2.73E-3	6.1

3.3 Materials and Methods

3.3.1 Genomic DNA extraction

A *Ph* isolate with culture number 612 and accession number 090017 (curation of the Plant Breeding Institute Rust Collection) was selected for the *de novo* assembly of a high quality *Ph* reference genome. The CTAB (cetyl triethylammonium bromide) extraction method (Rogers et al., 1989) was used to extract nucleic acids from desiccated urediniospores of the isolate *Ph*612. Specifically, about 25 mg spore was weighed in a Lysing Matrix C tube (MP Biomedicals, Australia). 1 ml of 2xCTAB extraction buffer (2% w/v CTAB, 100mM Tris-HCl (PH8.0), 20mM EDTA, 1.4 M NaCl, 1% w/v polyvinylpyrrolidone) was added to the tube, mixed by inversion and placed on ice for 2 minutes. The tube was then transferred to a FastPrep homogenizer (FP100, Thermo Savant, Bio101, Australia) and shaken at speed 6 for 15 seconds. It was immediately returned to ice incubation for 3 minutes, and again placed in the FastPrep homogenizer and shaken for 20 seconds at speed 6 and again returned to ice. This process aimed to disrupt the spore cell wall and release the nucleic acids and proteins. After that, the tube was incubated in a $65\,^{\circ}\mathrm{C}$ water bath for 30 minutes to soften the phospholipids in the cell membranes, and to denature the DNAse enzymes to prevent DNA digestion. After the tube was returned to room temperature, DNA extraction was performed by adding $\sim$250 ml of cold phenol:chloroform:isoamyl alcohol (25:24:1) to the tube and mixed thoroughly by gentle inversion (about 100 times). The tube was then centrifuged at maximum speed (13,200 rpm) for 30 minutes, and the resulting supernatant was transferred to a new 1.5 ml Eppendorf tube, in which the same phenol:chloroform:isoamyl alcohol procedure was repeated. Afterwards, cold chloroform:isoamyl alcohol (24:1) was added to the tube and mixed by gentle inversion, centrifuged at maximum speed for 20 minutes and the top aqueous phase was transferred to a new 1.5 ml Eppendorf tube. 50 ml of 3 M NaOAc and 500 ml of cold isopropanol was added to the tube which was then stored overnight at $-20\,^{\circ}\mathrm{C}$. The following day, the tube was centrifuged at 13,000 rpm for 30 minutes and the DNA pellet was drained carefully. The pellet was then washed with $500\,\mu\mathrm{L}$ of 70% ethanol, centrifuged at 13,000 rpm for 15 minutes, drained carefully and air dried. The dried pellet was re-suspended in $50\,\mu\mathrm{L}$ of double distilled water and stored overnight at $4\,^{\circ}\mathrm{C}$. The following day, $5\,\mu\mathrm{L}$ of Rnase-A was added to the tube and the tube was incubated at $37\,^{\circ}\mathrm{C}$ for 2 hours. The DNA sample was quantified using a Nanodrop ND-1000 spectrophotometer (Nanodrop Inc., America).

3.3.2 Genome sequencing, assembly and chracterization

The genomic DNA was sent to BGI Tech Solutions (HongKong) Co., Ltd for two library constructions, one was typical paired-end library with 500 bp insert size, the second was 6,000 bp mate pair library (Table 3.7). Both libraries were sequenced with 90 bp reading on Illumina Hiseq2000 platform. The sequencing reads were assembled into contigs and scaffolds in BGI Tech Solutions with SOAPdenovo v1.05[1], using parameters "-kmer 81". To assess genome heterozygosity, one lane of paired-end sequencing reads (totally 121,000,000 pairs) were aligned to the scaffolds by using Bowtie2 v2.2.5 (Langmead and Salzberg, 2012) with parameter setting "-D 20 -R 3 -N 0 -L 20 -i S,1,0.50, -minins 0, -maxins 900". To minimize false positive SNP calls around insertion/deletion (InDel) regions, poorly aligned reads around InDel were identified and realigned locally using RealignerTargetCreator and IndelRealigner in GATK package (McKenna et al., 2010). SNPs were called using freebayes v1.0.2 with parameters "-no-indels -no-mnps -ploidy 2" (Garrison and Marth, 2012). The number of heterozygous SNPs (548,638) was divided by the scaffold length (127.35Mb) to give an estimate of heterozygosity between the two dikaryotic nuclei.

Table 3.7: Sequencing Strategies for *P. hordei* 612

Library	Sample	Insert Size	Data	Sequencing Institute
Paired-end DNA	Urediniospores	500 bp	42G	BGI
Mate-pair DNA	Urediniospores	6,000 bp	4G	BGI
RNA-Seq	Infected leaf on 4,7 dpi	500 bp	50G	AGRF

CEGMA v2.4.0 (Parra et al., 2007) was used to map 248 core eukaryotic genes to the scaffolds for evaluating genome completeness. To identify and classify repeat content of the genome, RepeatModeler version open-1.0.8 (Smit and Hubley, 2008) was used to scan the scaffold sequences using default papramaters.

[1] http://soap.genomics.org.cn/soapdenovo.html

3.3.3 Phylogenetic analyses

The 18S rRNA biogenesis proteins of *P. sorghi, P. graminis, P. striiformis,* and *P. triticina* were downloaded from NCBI with accession KNZ50814, XP_003321364, KNF00205, and OAV98697. The *M. lini* 18S rRNA protein (ID TU.MELLI_sc114.1) was downloaded from the Genome Browser of CSIRO[1]. Along with the putative 18S rRNA biogenesis protein of *Ph*, the protein sequences were loaded in Geneious v10.0.5, a NGS analysis tool by Biomatters Ltd., for global sequence alignment with tool "Geneious Alignment" (key parameters: Gap open penalty 12, Gap extension penalty 3). Based on the alignment result, a Geneious plugin PhyML (Guindon and Gascuel, 2003) was used to calculate phylogenetic relationship of the proteins with a maximum likelihood algorithm.

3.3.4 RNA-seq of infected plant tissue

Barley seedlings (cultivar Morex) were infected with *Ph* strain 612. Infected leaf material was harvested at 0, 1, 2, 4, 7 and 11 days of post-inoculation (dpi) and stored at $-80\,^{\circ}\mathrm{C}$. Total RNA was extracted from infected leaf tissue using TRIzol reagent (Life Technologies Australia Pty Ltd.) according to the manufacturers instructions. The total RNA was then treated with RNase-free DNase I (New England BioLabs Inc.), and the column purified using ISOLATE II RNA Mini Kit (Bioline Australia) according to the manufacturers instruction. The quantity and quality of the total RNA were examined by Nanodrop (Thermo Scientific) and Agilent 2100 Bioanalyzer (Agilent Technologies). Two micrograms of RNA from 4 dpi and 7 dpi were combined and then processed to make a library using TruSeq Stranded mRNA Library Prep kit. The library was then sent to Australian Genome Research Facility Ltd (AGRF) for sequencing.

[1] http://webapollo.bioinformatics.csiro.au:8080/melampsora_lini/

3.3.5 Gene prediction guided by RNA-Seq and homologous proteins from Pgt

The raw RNA-seq data delivered by AGRF totaled 250,141,134 pairs of reads with length 101 bp, summing up to ∼50 Gbp. The data set was processed with Trim Galore v0.3.7[1], a software package integrating Cutadapt and FastQC for quality trimming, adapter trimming and quality control (Martin, 2011). The Trim Galore was run with parameters "-quality 20 -phred33 -length 35" commanding trimming low quality ends (Phred score: 20) from reads using ASCII+33 quality scores as Phred score, and a minimum read length of 35 bp to retain after trimming. The trimmed reads were aligned to *Ph612* scaffolds to decide transcript structures. The alignment was performed using TopHat v2.0.14 (Kim et al., 2013) with parameter setting "min-intron-length 10 -max-intron-length 5000 -mate-inner-dist 100 -mate-std-dev 100 -min-segment-intron 10 -max-segment-intron 5000". The mapped reads were assembled into transcripts and isoforms using Cufflinks v2.2.1 (Trapnell et al., 2012) with parameters "min-intron-length 10 -max-intron-length 5000 -minisoform-fraction 0.1". CodingQuarry v1.2 (Testa et al., 2015) was then used to infer gene models from the Cufflinks output and 29,520 genes were predicted. The prediction was performed in two stages, the first of which is predictions directly from transcript sequences and thus these predictions were well supported by the RNA-Seq experiment. In the second stage, Generalised hidden Markov models trained with the transcripts were used to *ab initio* predict genes that are not expressed or captured in the RNA-Seq experiment. To ensure high quality of gene models well supported by the RNA-Seq data, the 29,520 genes were inspected with the RNA-Seq reads alignments produced by TopHat. First, tool genomecov of Bedtools v2.25.0 (Quinlan and Hall, 2010) was used to calculate RNA-Seq read coverage for each base on the genomic scaffolds. Second, a custom Perl script was developed to calculate ratio of the predicted gene regions covered by the RNA-Seq reads. Genes with >70% region covered were retained (a total of 11,403 genes meet the criterion).

Conserved protein sequences generally diverge slowly over large taxonomic distance. Therefore, proteins from other *Puccinia* species can be mapped to the *Ph* genome to identify regions of homology. This approach should be useful at predicting genes not

[1] http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

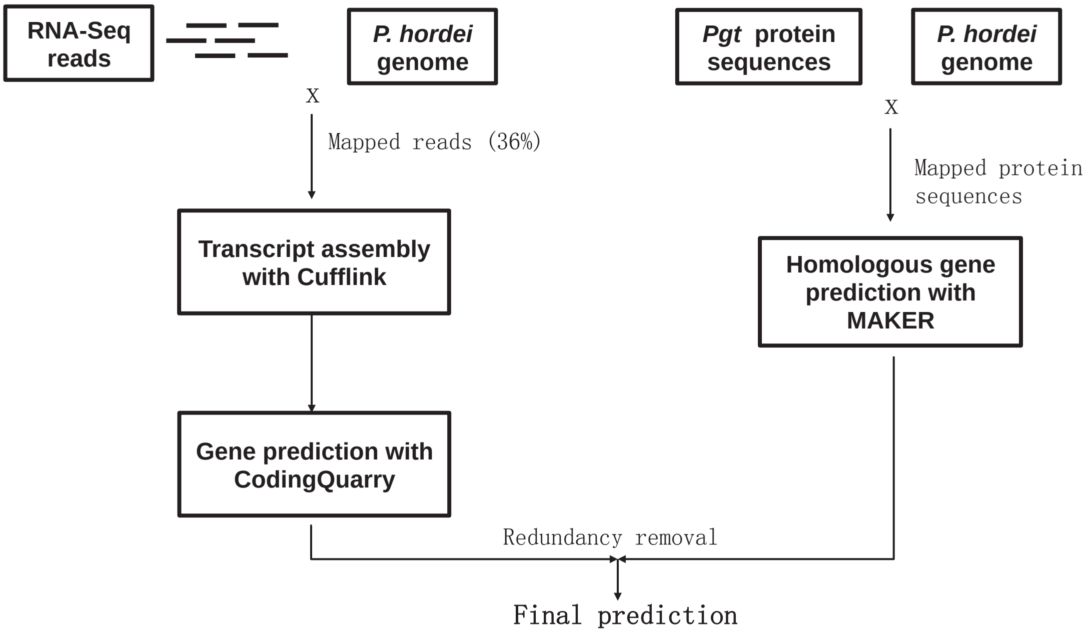


Figure 3.7: Illustration of gene prediction pipeline.

captured by the RNA-Seq experiment. In the Australian *Pgt* pan-genome project, Upadhyaya et al. (2015) predicted 22,391 protein sequences for Australian *Pgt* pathotype 21-0 from sequencing of RNA isolated from purified haustoria and germinated spores. 7,921 of the 22,391 protein sequences were fragmented with only partial sequences published. Here, only complete proteins (total number 14,470) were selected and used. The complete *Pgt* protein sequences were inputted to a genome annotation pipeline called MAKER v2.31.8 (Holt and Yandell, 2011), which mapped the proteins to the *Ph* genome with BLASTX and Exonerate. It then predicted 9,580 *Ph* gene models based on the protein mappings.

Some of the MAKER predicted genes were already predicted by CodingQuarry. To integrate both prediction sets, the redundancy in MAKER predictions was removed by discarding 4,639 MAKER predicted genes that had >60% genomic regions overlapping with the CodingQuarry predicted genes (Figure 3.7). The resulting 4,941 MAKER genes were merged with the 11,403 CodingQuarry genes together as a final confident gene set totaling 15,984 genes well supported by the RNA-Seq experiment and protein homology with *Pgt*.

3.3.6 Protein clustering

Protein sequences of *Mli* CH5 were downloaded from online Supplementary Material of Nemri et al. (2014), and *Pgt* CDL75-36700-3 proteins were downloaded from the website of *Puccinia* Comparative Genomic Projects[1]. The protein sequences of *Mlp* 98AG31, *Pst* 2K41-Yr9 and *Pt* BBBD were downloaded from NCBI with genome accessions GCA 000204055.1, GCA 001191645.1 and GCA 000151525.2 respectively. Pair-wise sequence similarities for all the proteins of the six rust species including the *Ph* putative proteins and the five downloaded protein sets were calculated with BLASTP (default parameters). The Markov cluster (MCL) algorithm developed by Enright et al. (2002) was used to perform protein clustering based on the BLASTP result.

3.3.7 Protein functional annotation and survey

Protein sequences were translated from the predicted genes based on their CDS structures. The putative protein sequences were searched against the Nr protein database v2012-02-29 using BLASTP (Altschul et al., 1997) with parameters "-e 1e-5 -F F -a 4" and the BLASTP hits were filtered requiring minimum 40% of the *Ph* query proteins matched Nr proteins and the sequence identities were not less than 30%. For *Ph* proteins with several BLASTP hits, the best hit (with the smallest E-value) was reported. The description of the matched proteins are listed in Supplementary Table 3.2. The 16,354 protein sequences were also functionally annotated with InterProScan v5 (Jones et al., 2014) and results were also reported in Supplementary Table 3.2.

In the surveys of mating type loci and obligate biotrophic lifestyle, the BLASTP hits of *Ph* proteins against *Mli* proteome were taken from those produced in Protein clustering section. To examine whether a specific gene was expressed *in planta*, the RNA-Seq read mapping produced by Tophat2 and gene annotation by CodingQuarry were imported and displayed in Integrated Genome Browser. A gene was considered expressed if its CDS region was covered RNA-Seq read mapping.

[1] https://www.broadinstitute.org/scientific-community/science/projects/fungal-genome-initiative/puccinia-comparative-genomic-projects

SignalP v4.1 (Petersen et al., 2011) was used to predict signal peptide and transmembrane domains in the putative proteins with default software parameters. Chi-square test was used to check enrichment of individual IPS annotations in secreted proteins. The null hypothesis was that each protein's IPS annotation was independent of whether it was secreted or not. The Chi-square for each IPS annotation was calculated with a Perl script, and imported into an Excel table for calculating its P-value in Chi-square distribution.

3.4 Discussion

3.4.1 Assembly

This research reported the first genome assembly of an economically important cereal rust pathogen *Puccinia hordei*, which will provide a valuable genomic resource for the rust research community. As a reference, the assembly enables genome and transcriptome comparison of isolates within and across different *Ph* lineages in Australia (Park et al., 2015), especially in studies targeting diversity in candidate effector genes. Genomes of the currently sequenced cereal rust fungi differ significantly in size, ranging from 88 Mbp in *Pgt* to 135 Mbp in *Pt* (Table 3.8). Gaps in the *Ph* scaffolds amounted to 8.48%, which is much less than 32% in the *Pt* assembly (Cuomo et al., 2017). Thus, the actual genome size of *Ph* may be larger than *Pt*, as a proportion of small contigs in the *Pt* assembly may sink into the scaffold gaps. The relatively large estimated genome size of *Ph* (127 Mbp) may be partially attributed to the large amount of interspersed repeats (∼54%; most associated with transposition activities), as its predicted gene content (16,354 protein coding genes) is not larger than other sequenced species (Table 3.8). The enrichment of transposable elements in the genome contributes to genetic polymorphisms, especially in the case that most known *Ph* lineages in Australia are under clonal propagation without sexual recombination (Park et al., 2015). It is interesting to study the cost associated with the replication of this excessive repetitive DNA as fitness penalty in the future. The genome size variations among rust fungi may be also related to their hosts and lifecycles (Tavares et al., 2014). Even within one species,

Table 3.8: Genome characteristics of five rust pathogens

Species	Assembly size	Scaffolds	Scaffold N50	%GC	Genes	CEGMA
P. hordei	127.35 Mbp	15,913	22.95 Kbp	41.11	15,984	96%
P. triticina[1]	135.34 Mbp	14,818	544.26 Kbp	46.72	14,880	97%
P. striiformis tritici[2]	117.31 Mbp	9,715	519.86 Kbp	44.43	19,542	97%
P. graminis tritici[3]	88.64 Mbp	392	964.97 Kbp	42.35	15,800	94%
P. sorghi[4]	99.64 Mbp	15,722	19.08 Kbp	43.14	21,087	98%

[1] Cuomo et al. (2017)
[2] Zheng et al. (2013)
[3] Duplessis et al. (2011)
[4] Rochi et al. (2016)

host-dependent variations in nuclear content have been observed for *Ph* (Eilam et al., 1994).

The N50, a measurement of assembly continuity, for the *Ph* scaffolds generated in the present study is 21,945 bp, comparable to the *P. sorghi* assembly (19,081 bp; Rochi et al., 2016), but is notably lower than the other three rust genome assemblies, *Pgt*, *Pt* and *Pst* (965 Kbp, 544 Kbp and 520 Kbp respectively; Table 3.8). This performance difference on assembly continuity is partially due to the differences of sequencing technologies and DNA libraries used for the genome projects. The *Ph* and *P. sorghi* were sequenced on Illumina short read (about 100 bp) platforms with DNA libraries of insert sizes up to 6 Kbp, while the other three rust sequencing were performed on Roche 454 (read length ∼330 bp) or Sanger technology (∼800 bp), with libraries of insert size up to 100 Kbp (Cuomo et al., 2017; Duplessis et al., 2011; Metzker, 2010). Longer read length provides better-overlapped reads for assembly, and large insert size fragments help to resolve repeat structures in the genome. The future work of *Ph* genomics study will include a new assembly with third generation sequencing technologies such as PacBio SMRT and Oxford Nanopore that produce read length (averaging 1,500 bp; (Weirather et al., 2017)). These technologies offer opportunities to improve draft genome assemblies generated with Illumina short read technologies, which can be easily confounded by complex genome structures with large amounts of repetitive elements and high heterozygosity between the two nuclei in the rust fungi.

3.4.2 Heterozygosity

To assess genome heterozygosity, the reads were mapped back to the *Ph* genome assembly, similar to the approach used in previous studies (Cantu et al., 2011; Wu et al., 2017; Zheng et al., 2013). When reads were allowed for only one optimal mapping in the reference, the distribution of read depth shows a two-peak distribution (Figure 3.1 A and B). Similarly, the distribution of k-mer depth of the *Pst* isolate CY32 genome sequencing also showed two major peaks, based on which the authors suggested high heterozygosity for the genome (Zheng et al., 2013). In contrast, Rochi et al. (2016) observed only one major peak in the distribution of k-mer depth of the *P. sorghi* race RO10H11247 sequencing, and accordingly the authors suggested low genome heterozygosity. The reliability of peaks in the distribution of sequencing amount for judging heterozygosity was further supported by the mapping of *Ph* reads allowing up to two optimal positions that showed the first peak was a result of inter-nucleus heterozygous differences (Figure 3.1 C and D).

To remove the effect of allelic sequences in the *Ph* genome assembly on the heterozygosity calculation, the up-to-two-position mapping was used. This mapping revealed 602,188 SNPs (4.7/kb) across the scaffolds, with a majority of SNPs (548,638; 4.3/kb) occurring in a heterozygous condition. The frequency of heterozygous positions showed a high level of divergence between the two independent nuclei in *Ph*'s dikaryotic genome. Similarly high genome heterozygosity has been observed in several sequenced isolates of *Pt*, *Pst*, *Pgt* with a range from 2.6 to 11.3 SNP/kb (Cantu et al., 2013; Cuomo et al., 2017; Upadhyaya et al., 2015; Wu et al., 2017).

Although the reported heterozygosity rates from different rust species were consistent in the same order of magnitude (several SNPs per thousand bp), comparisons of this metric across different genomics projects is still affected by differences in the methods used, such as reference sequence ploidy, sequencing technology, read mapping algorithm, and stringency of SNP calling in theory. The future development and standardization of sequencing technologies may reduce such noise to facilitate the comparisons of bioinformatics results from various projects.

3.4.3 Gene prediction

To predict gene models for the newly assembled genome, transcriptome sequencing of infected leaf tissues collected on 4 and 7 days post infection was used to guide prediction of *Ph* genes expressed in hyphae and haustoria during those two time points. In addition, the alignment of *Pgt* proteins to the *Ph* genome enabled the identification of genes missing in the RNA-Seq data due to their exclusive expression at other life stages or in other tissues. The combination of these two predictions provided a comprehensive gene space for *Ph*. Future work in gene annotation shall focus more on biological functions, such as secreted protein-encoding genes preferentially expressed in haustoria, the presumed sites for effector delivery into host, and the specific tissue that has been shown to be enriched for effector genes in *Mli* (Catanzariti et al., 2006).

Ph is macrocyclic and heteroecious, with uredinia and telia occurring on the primary hosts *Hordeum vulgare*, *Hordeum vulgare* ssp. *spontaneum*, *Hordeum bulbosum*, and *Hordeum murinum*, and aecia occurring on several species of the genera *Ornithogalum, Leopoldia*, and *Dipcadi* (Park et al., 2015). The uredinial stages showed distinct specialization to primary hosts, with isolates collected from *H. vulgare* and *H. vulgare* ssp. *spontaneum* infecting only the two host species, and isolates recovered from *H. vulgare* and *H. vulgare* ssp. *spontaneum* infecting these two host species only (Anikster, 1989; Anikster et al., 1982). With the reference genome generated here, it will be very interesting to compare gene expression profiles of *Ph* after inoculation to different hosts at different life-cycle stages.

3.4.4 Mating type loci

Like other Basidiomycetes, the heterothallic nature of *Ph* is probably mediated by two types of mating type loci. One locus, the HD1 and HD2 homeodomain transcription factor genes, was identified in the *Ph* genome assembly with a conserved divergent transcription structure (Figure 3.2). This structure has been observed in genome assemblies of other rust fungi including *M. larici-populina, P. graminis* f. sp. *tritici* (*Pgt*) and *Pt* (Cuomo et al., 2017; Duplessis et al., 2011). Here, both alleles of the locus were assembled and separated in Scaffold 738 and Scaffold 2365 (Figure 2A and 2B), whereas in the *Pgt* assembly the allelic homologs were merged in one copy in Supercontig 2.13. In contrast, the identified copy number of pheromone and pheromone receptor genes for *Ph*612 (1 and 2, respectively) is less than that in *Mlp* (11 pheromones and 4 receptors). This is partly due to rich repeats surrounding the pheromone loci in the *Ph* genome, hindering assembly of the regions. It may be also because of the duplication and divergence of some of the allelic variants in *Mlp*. The close proximity of the pheromone gene *PH612_03487* and pheromone receptor gene *PH612_03488* is a common P/R organization observed in several other basidiomycetes (Kües, 2015). The identification of *Ph* mating loci will provide support for future studies of organization, composition and of sexual recombination in this pathogen.

3.4.5 Obligate biotrophy hallmarks

It has been suggested that the evolution of obligate biotrophy is associated with loss of some metabolic pathways (Kemen and Jones, 2012; Spanu et al., 2010). The results of several rust genomics studies illustrate that the degree of deletion in specific pathways can vary across different species. The sulfate metabolism pathway seems to be complete in *Mli* and *Mlp*, but appears to be at least partially deleted in *Pgt*, *Pst* (Duplessis et al., 2011; Garnica et al., 2013; Nemri et al., 2014) and in *Ph* studied here. Consistent with the observations for other rust pathogens, the apparent loss of ability in *Ph* to metabolize nitrite is associated with amplification of peptide and oligopeptide transporters. Determining the direction of the transporters and the molecules translocated will require substantial further study.

3.4.6 Candidate effectors

A large number of candidate effectors were identified for *Ph*, providing a fundamental starting point for screening for avirulence genes. Functional annotation highlighted several pathogenesis aspects for these candidate effectors: (1) plant tissue degradation; (2) host peptide cleavage; (3) infection tissue development. Further significant work is needed to relate these effector candidates to specific metabolic pathways at individual time points in different fungal and plant organs.

3.4.7 Summary

This chapter assembled the first genome sequence for *Ph*, and performed a thorough gene annotation with a focus on mating type loci, obligate biotrophic lifestyle, and candidate effectors. These results are a first step for genome-wide molecular investigations of virulence mechanism of the economically important fungus.

References

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., et al. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Research* 25, 3389–3402 89

Anderson, C. M., Willits, D. A., Kosted, P. J., Ford, E. J., Martinez-Espinoza, A. D., and Sherwood, J. E. (1999). Molecular analysis of the pheromone and pheromone receptor genes of *Ustilago hordei*. *Gene* 240, 89–97 77

Anikster, Y. (1989). Host specificity versus plurivority in barley leaf rusts and their microcyclic relatives. *Mycological Research* 93, 175–181 93

Anikster, Y. et al. (1982). Alternate hosts of *Puccinia hordei*. *Phytopathology* 72, 733–735 61, 93

Cantu, D., Govindarajulu, M., Kozik, A., Wang, M., Chen, X., Kojima, K. K., et al. (2011). Next generation sequencing provides rapid access to the genome of *Puccinia striiformis* f. sp. *tritici*, the causal agent of wheat stripe rust. *PloS One* 6, e24230. doi:10.1371/journal.pone.0024230 60, 68, 92

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., et al. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics* 14, 270 92

Catanzariti, A.-M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., and Ellis, J. G. (2006). Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *The Plant Cell* 18, 243–256 93

Coelho, M. A., Bakkeren, G., Sun, S., Hood, M. E., and Giraud, T. (2017). Fungal sex: the basidiomycota. *Microbiology Spectrum* 5 77

Cotterill, P., Park, R., and Rees, R. (1995). Pathogenic specialization of *Puccinia hordei* Otth. in Australia, 1966-1990. *Crop and Pasture Science* 46, 127–134 60

Cotterill, P., Rees, R., Platz, G., and Dill-Macky, R. (1992). Effects of leaf rust on selected Australian barleys. *Australian Journal of Experimental Agriculture* 32, 747–751 60

Crawford, N. M. and Arst Jr, H. N. (1993). The molecular genetics of nitrate assimilation in fungi and plants. *Annual Review of Genetics* 27, 115–146 78

Cuomo, C. A., Bakkeren, G., Khalil, H. B., Panwar, V., Joly, D., Linning, R., et al. (2017). Comparative analysis highlights variable genome content of wheat rusts and divergence of the mating loci. *G3: Genes, Genomes, Genetics* 7, 361–376 60, 61, 62, 68, 77, 78, 90, 91, 92, 94

Dodds, P. N. and Rathjen, J. P. (2010). Plant immunity: towards an integrated view of plant-pathogen interactions. *Nature Reviews. Genetics* 11, 539 61

Duplessis, S., Cuomo, C. A., Lin, Y.-C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proceedings of the National Academy of Sciences U.S.A.* 108, 9166–9171. doi:10.1073/pnas.1019315108 60, 61, 64, 68, 78, 80, 91, 94

Eilam, T., Bushnell, W., and Anikster, Y. (1994). Relative nuclear DNA content of rust fungi estimated by flow cytometry of propidium iodide-stained pycniospores. *Phytopathology* 84, 728–734 91

Enright, A. J., Van Dongen, S., and Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research* 30, 1575–1584 89

Fraser, J. A. and Heitman, J. (2004). Evolution of fungal sex chromosomes. *Molecular Microbiology* 51, 299–306 78

Garnica, D. P., Upadhyaya, N. M., Dodds, P. N., and Rathjen, J. P. (2013). Strategies for wheat stripe rust pathogenicity identified by transcriptome sequencing. *PloS One* 8, e67150. doi:10.1371/journal.pone.0067150 94

Garrison, E. and Marth, G. (2012). Haplotype-based variant detection from short-read sequencing. *arXiv preprint arXiv:1207.3907* 85

Guindon, S. and Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology* 52, 696–704 86

Holt, C. and Yandell, M. (2011). MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* 12, 491 88

Joly, D. L., Feau, N., Tanguay, P., and Hamelin, R. C. (2010). Comparative analysis of secreted protein evolution using expressed sequence tags from four poplar leaf rusts (*Melampsora* spp.). *BMC Genomics* 11, 422 82

Jones, J. D. and Dangl, J. L. (2006). The plant immune system. *Nature* 444, 323 61

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., et al. (2014). Interproscan 5: genome-scale protein function classification. *Bioinformatics* 30, 1236–1240 89

Kemen, E. and Jones, J. D. (2012). Obligate biotroph parasitism: can we link genomes to lifestyles? *Trends in Plant Science* 17, 448–457 94

Kim, D., Pertea, G., Trapnell, C., Pimentel, H., Kelley, R., and Salzberg, S. L. (2013). Tophat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. *Genome Biology* 14, R36 87

Kües, U. (2015). From two to many: Multiple mating types in Basidiomycetes. *Fungal Biology Reviews* 29, 126–166 62, 78, 94

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 in Biochemical Sciences* 28, 118–121 82

Kullman, B., Tamm, H., and Kullman, K. (2005). Fungal genome size database 63

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9, 357–359 85

Larous, L., Kameli, A., and Lösel, D. (2008). Ultrastructural observations on *Puccinia menthae* infections. *Journal of Plant Pathology* , 185–190 82

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet. journal* 17, pp–10 87

Marzluf, G. A. (1997). Molecular genetics of sulfur assimilation in filamentous fungi and yeast. *Annual Reviews in Microbiology* 51, 73–96 80

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20, 1297–1303 85

Metzker, M. L. (2010). Sequencing technologies - the next generation. *Nature Reviews. Genetics* 11, 31–46. doi:10.1038/nrg2626 91

Murray, G. M. and Brennan, J. P. (2009). *The current and potential costs from diseases of barley in Australia* (Grains Research and Development Corporation: Canberra Australia) 60

Nemri, A., Saunders, D. G. O., Anderson, C., Upadhyaya, N. M., Win, J., Lawrence, G. J., et al. (2014). The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. *Frontiers in Plant Science* 5, 98. doi:10.3389/fpls.2014.00098 60, 78, 80, 89, 94

Park, R. (2003). Pathogenic specialization and pathotype distribution of *Puccinia hordei* in Australia, 1992 to 2001. *Plant Disease* 87, 1311–1316 60

Park, R. F., Golegaonkar, P. G., Derevnina, L., Sandhu, K. S., Karaoglu, H., Elmansour, H. M., et al. (2015). Leaf rust of cultivated barley: pathology and control. *Annual Review of Phytopathology* 53, 565–589 90, 93

Parra, G., Bradnam, K., and Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 23, 1061–1067 63, 85

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* 8, 785–786 90

Quinlan, A. R. and Hall, I. M. (2010). Bedtools: a flexible suite of utilities for comparing genomic features. *Bioinformatics* 26, 841–842 87

Raudaskoski, M. and Kothe, E. (2010). Basidiomycete mating type genes and pheromone signaling. *Eukaryotic Cell* 9, 847–859 62

Rochi, L., Diéguez, M. J., Burguener, G., Darino, M. A., Pergolesi, M. F., Ingala, L. R., et al. (2016). Characterization and comparative analysis of the genome of *Puccinia sorghi* schwein, the causal agent of maize common rust. *Fungal Genetics and Biology* 63, 91, 92

Rogers, S. O., Rehner, S., Bledsoe, C., Mueller, G. J., and Ammirati, J. F. (1989). Extraction of DNA from *Basidiomycetes* for ribosomal DNA hybridizations. *Canadian Journal of Botany* 67, 1235–1243 84

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. doi:10.1371/journal.pone.0029847 61, 82

Smit, A. and Hubley, R. (2008). Repeatmodeler open-1.0. *Available from: http://www.repeatmasker.org* 85

Spanu, P. D., Abbott, J. C., Amselem, J., Burgis, T. A., Soanes, D. M., Stüber, K., et al. (2010). Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science* 330, 1543–1546 94

Spit, A., Hyland, R. H., Mellor, E. J. C., and Casselton, L. A. (1998). A role for heterodimerization in nuclear localization of a homeodomain protein. *Proceedings of the National Academy of Sciences U.S.A.* 95, 6228–6233 78

Tavares, S., Ramos, A. P., Pires, A. S., Azinheira, H. G., Caldeirinha, P., Link, T., et al. (2014). Genome size analyses of Pucciniales reveal the largest fungal genomes. *Frontiers in Plant Science* 5, 422 90

Testa, A. C., Hane, J. K., Ellwood, S. R., and Oliver, R. P. (2015). CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics* 16, 170 87

Trapnell, C., Roberts, A., Goff, L., Pertea, G., Kim, D., Kelley, D. R., et al. (2012). Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nature Protocols* 7, 562–578 87

Upadhyaya, N. M., Garnica, D. P., Karaoglu, H., Sperschneider, J., Nemri, A., Xu, B., et al. (2015). Comparative genomics of Australian isolates of the wheat stem rust pathogen *Puccinia graminis* f. sp. *tritici* reveals extensive polymorphism in candidate effector genes. *Frontiers in Plant Science* 5, 759 61, 71, 88, 92

Voegele, R. T., Struck, C., Hahn, M., and Mendgen, K. (2001). The role of haustoria in sugar supply during infection of broad bean by the rust fungus *Uromyces fabae*. *Proceedings of the National Academy of Sciences U.S.A.* 98, 8133–8138 81

Wallwork, H., Preece, P., and Cotterill, P. (1992). *Puccinia hordei* on barley and *Omithogalum umbellatum* in South Australia. *Australasian Plant Pathology* 21, 95–97 61, 62

Waterhouse, W. L. (1927). *Studies in the Inheritance of Resistance to Leaf Rust, Puccinia Anomola Rostr., in Crosses of Barley*, vol. 61 (J. Proc. Royal Soc: New South Wales) 60

Weirather, J. L., de Cesare, M., Wang, Y., Piazza, P., Sebastiano, V., Wang, X.-J., et al. (2017). Comprehensive comparison of pacific biosciences and oxford nanopore technologies and their applications to transcriptome analysis. *F1000Research* 6 91

Wu, J. Q., Sakthikumar, S., Dong, C., Zhang, P., Cuomo, C. A., and Park, R. F. (2017). Comparative genomics integrated with association analysis identifies candidate effector genes corresponding to *Lr20* in phenotype-paired *Puccinia triticina* isolates from Australia. *Frontiers in Plant Science* 8 92

Zambino, P. J. and Szabo, L. J. (1993). Phylogenetic relationships of selected cereal and grass rusts based on rDNA sequence analysis. *Mycologia* 85, 401–414 68

Zheng, W., Huang, L., Huang, J., Wang, X., Chen, X., Zhao, J., et al. (2013). High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nature Communications* 4, 2673. doi:10.1038/ncomms3673 68, 92

Chapter 4

Identification of candidates for avirulence genes corresponding to resistance genes *Rph3*, *Rph13*, and *Rph19* using pathotypes from a clonal lineage of *Puccinia hordei*

4.1 Introduction

One of the most cost effective methods for barley leaf rust (causal agent *Puccinia hordei*, *Ph*) management is the use of resistance genes (Golegaonkar et al., 2010; Park, 2008; Park et al., 2015). Resistance genes employed in crop breeding typically encode immunoreceptors that recognize avirulence (Avr) gene products from infecting rust pathogens. This recognition subsequently leads to the initiation of the host immune response to arrest the development of the rust pathogen around the infection sites, and generally a localized hypersensitive reaction is observed. However, the evolution of new rust pathotypes can modify or delete the Avr genes to avoid host recognition, breaking down cultivar resistance and resulting in rust epidemics. Identification of these Avr genes in the pathogen and their corresponding resistance genes in the host is the first step to understand genetic interactions in the barley-*Ph* pathosystem, which is essential for developing fundamental solutions to reduce the threat posed by this pathogen.

Traditionally, the study of rust genetics was limited by their obligate biotrophic nature. The development of Next-generation sequencing (NGS) technology, with rapidly decreasing cost, has enabled the whole genome sequencing of several rust species and revolutionized the field of rust pathogen genetics. With the availability of reference genomes, genome re-sequencing followed by variation characterization has proven to be an effective tool for Avr gene identification in plant pathogens. A typical workflow of re-sequencing analysis is to map millions of sequencing reads of individual isolates of different pathotypes to a reference genome, and call the genotypes based on mapped reads, and then identify genetic variations in association with the pathotypes across the isolates. For instance, re-sequencing of 11 strains of the vascular wilt fungus *Verticillium dahliae* identified a 50 Kbp stretch of DNA present in 4 avirulent strains but absent in the remaining 7 virulent strains, which led to the successful cloning of the Avr gene *Ave1* responsible for the avirulence (de Jonge et al., 2012). In another study, sequencing and comparative genomics of two strains of *Cladosporium fulvum* (causal agent of tomato leaf mold) differing in virulence to resistance gene *Cf-5* identified the corresponding Avr gene *Avr5*, which has a 2 bp deletion in the virulent strain (Mesarich et al., 2014). Taken together, these studies demonstrated that gain of virulence could be caused by either deletion or simple mutation of the Avr gene.

Re-sequencing projects have also been performed for rust pathogens, which have revealed a number of candidates for various Avr genes. Persoons et al. (2014) sequenced 15 isolates of poplar leaf rust fungus *Melampsora larici-populina* (*Mlp*) and highlighted 30 candidate effector genes based on SNP polymorphisms across the isolates. Cantu et al. (2013) re-sequenced four *Pst* (*Puccinia striiformis* f. sp. *tritici*) isolates and reported five candidate effector genes that may distinguish pathotypes of isolate PST-87/7 and PST-08/21. In Chapter 2 of the present study, two *Pgt* (*Puccinia graminis* f. sp. *tritici*) isolates differing in virulence to stem rust resistance gene *Sr50* were sequenced and comparative genomics revealed a 25 Mbp loss-of-heterozygosity in the mutant isolate associated to the loss of Avr allele of *AvrSr50*, and the mutation event highlighted 25 genes residing in the 25 Mbp region as candidates for the Avr gene. Other published re-sequencing studies include the work by Zheng et al. (2013), Upadhyaya et al. (2015) and Wu et al. (2017) on *Pst*, *Pgt* and *Pt* (*Puccinia triticina*), respectively.

The emergence of new virulent pathotypes of *Ph* has been a significant challenge to stable barley yield in Australia (Park et al., 2015). The origins of new virulence include at least three mechanisms: sexual recombination, simple mutation, and exotic incursion. The characterisation of genetic diversity within the *Ph* population is essential in evaluating the evolutionary potential and the development of new pathotypes with an ability to break down resistance in current cultivars. Karaoglu and Park (2014) designed 600 microsatellite markers for *Ph*, 76 of which were able to identify polymorphism among 19 Australian *Ph* isolates. Later, Sandhu et al. (2016) applied PCR-fingerprinting technology to test genetic variability of 22 *Ph* pathotypes collected from several geographical locations in Australia. These two studies showed evidence of simple mutation as a factor driving pathogenic diversity in Australian *Ph* populations. But the genomic regions examined in these studies are limited to the targets by primers in polymerase chain reaction experiments.

In Chapter 3, a whole genome assembly of *Ph* isolate 612 was obtained and 1,072 genes encoding secreted proteins (SP) were predicted. In this Chapter, genome re-sequencing was performed for four additional isolates, which along with *Ph*612 belong to a clonal lineage of pathotypes that differ in virulence to barley resistance genes *Rph3, Rph13,* and *Rph19* (Figure 4.1). The presumed progenitor pathotype 5453P- (isolate

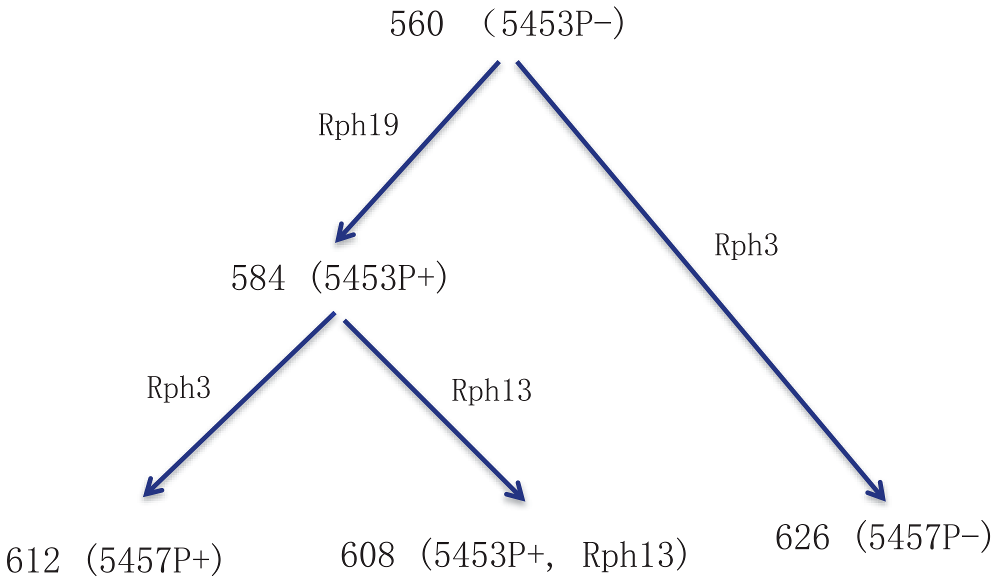


Figure 4.1: Five *Puccinia hordei* isolates and their virulence gain for three resistance genes. Nodes in the tree show isolate number and pathotype in bracket, and the labels on tree branches show resistance genes overcome by the pathotype mutation.

560) was first discovered in Western Australia in 2001, being pathogenically distinctive from all *Ph* pathotypes previously detected in Australia (Park RF, unpublished). This pathotype acquired virulence to *Rph19* in 2003, to generate the derivative pathotype 5453P+ (isolate 584) (Park et al., 2015). Later in 2008 and 2009, further independent virulence gains to *Rph13* and *Rph3* from pathotype 5453P+ were detected in isolates 608 and 612, respectively. In parallel, pathotype 5457P- (isolate 626) was detected in 2013, which likely arose from the progenitor pathotype 5453P- by gaining virulence to *Rph3*. This lineage relationship was supported by pilot studies based on five microsatellite markers, which showed that isolates 560, 626, 584 and 612 shared the same genotype on the amplified genomic regions, indicating that they were closely related. Further analysis of the microsatellite markers also suggested that isolates 626, 584, and 612 were possibly derived from the isolate 560 via simple mutations (Karaoglu and Park, unpublished). Here, sequencing data of the five isolates were mapped to the reference genome to examine genetic variations that may account for their virulence differences on resistance genes *Rph3*, *Rph13* and *Rph19*. According to the gene-for-gene hypothesis

Table 4.1: Five *Puccinia hordei* isolates and their virulence profile

Isolate	Pathotype[1]	Virulence to *Rph* genes	Collection time	Collection location
560	5453P-	*Rph1*, *Rph2*, *Rph4*, *Rph6*, *Rph9*, *Rph10*, *Rph12*	2001	Esperance, WA[2]
626	5457P-	*Rph1*, *Rph2*, *Rph3*, *Rph4*, *Rph6*, *Rph9*, *Rph10*, *Rph12*	2013	Boxwood Hill, WA
584	5453P+	*Rph1*, *Rph2*, *Rph4*, *Rph6*, *Rph9*, *Rph10*, *Rph12*, *Rph19*	2003	Wongan Hills, WA
608	5453P+, +Rph13	*Rph1*, *Rph2*, *Rph4*, *Rph6*, *Rph9*, *Rph10*, *Rph12*, *Rph13*, *Rph19*	2008	Aratula, QLD[3]
612	5457P+	*Rph1*, *Rph2*, *Rph3*, *Rph4*, *Rph6*, *Rph9*, *Rph10*, *Rph12*, *Rph19*	2009	Legume, QLD

[1] An octal system described by Gilmour (1973) was used to designate pathotypes. The suffix P+ or P- added to the octal designation indicate virulence or avirulence on the resistance gene *Rph19*.

[2] Western Australia

[3] Queensland

(Flor, 1971), each dominant Avr gene is corresponding to each of the three resistance genes, thus the three Avr genes are designated as *AvrRph3*, *AvrRph13* and *AvrRph19*, respectively. As the Avr genes may have been lost via genomic deletion or functional allele mutation, searching for them was focused on SP genes that contain copy number reduction or functional mutations.

4.2 Results

4.2.1 Assembly of 560 specific DNA

The genome assembly of *Ph*612 likely does not contain the two Avr genes *AvrRph3* and *AvrRph19*, as they might have been lost in the virulence gains to R genes *Rph3* and *Rph19* via deletion of the Avr components. To complement the assembly, DNA sequencing reads specific to *Ph*560 but absent in the scaffolds were assembled. This was achieved by first mapping *Ph*560 reads to the *Ph*612 reference and then *de novo* assembling of 6 million unmapped reads with SOAPdenovo (Li et al., 2010). This assembly produced 1,540 contigs with a total length of 596 Kbp (average length 387 bp). Genes in the contigs were predicted using homology guidance of *Pgt* proteins generated by Upadhyaya et al. (2015), similar to the MAKER annotation in Chapter 3. A total of 282 genes were predicted, one of which (*PH612_13953*) contained a signal peptide but no predicted transmembrane domain, suggesting it was a SP gene.

The *Ph*560 contigs were added to the *Ph*612 assembly (127,347 Kbp), generating a new, more comprehensive reference, designated Gn612_560 (127,943 Kbp). Combining gene annotations for the *Ph*612 assembly (Chapter 3 results) and those for the *Ph*560 contigs here, the new reference contained a total of 16,578 annotated genes with 1,073 genes encoding SPs. The SP genes were subjected to comparative genomics analysis for the identification of Avr gene candidates.

4.2.2 Read mapping

Whole genome sequencing (125 bp paired-end reads) was performed for the four *Ph* pathotypes in the 5453P- lineage as aforementioned. This yielded about 11 Gbp data for each isolate after quality trimming (Table 4.2). To genotype the isolates, the sequencing reads of each isolate were mapped to the reference genome Gn612_560 individually. For each isolate, about 85% of reads could be mapped to the reference (Table 4.2) and the percentage of the reference coverage was over 99%, suggesting that almost all genes annotated in the reference could be genotyped. The remaining 15% of reads that could not be mapped are likely due to assembly gaps and/or the absence of mitochondrial DNA in the reference. To compare genotypes across the five isolates including 612, a subset of *Ph*612 reads sequenced in Chapter 3 was sampled and mapped to the reference genome Gn612_560. The mapping rate of isolate 612 was slightly lower (77.54%).

Table 4.2: Sequencing and reference mapping of the five *Puccinia hordei* isolates

Isolate	Sequencing Institute	Read Length	Insert Size	Total Reads	Data Size	Mapping Rate	Coverage on Reference
560	Novogene	125bp	300bp	92,191,726	11.6Gbp	85.32%	78X
626	Novogene	125bp	600bp	89,044,138	11.1Gbp	86.26%	75X
584	AGRF	125bp	600bp	94,489,670	11.8Gbp	86.15%	80X
608	AGRF	125bp	600bp	88,277,542	11.0Gbp	85.36%	74X
612	BGI	90bp	500bp	242,000,000	21.8Gbp	77.54%	132X

4.2.3 Analysis of copy number variation for SP genes

Rust Avr genes characterized to date encode secreted proteins that are recognized by host immunoreceptors *in planta*, therefore, the 1,073 predicted SP genes are most likely to include the Avr genes *AvrRph3*, *AvrRph13*, and *AvrRph19*. As the loss of Avr functions in the derivative pathotypes could be caused by either allele deletions or simple sequence mutations, subsequent analyses focused on these two aspects.

To examine copy number variations (CNVs) of the 1,073 SP genes in the five isolates, read coverage depth for each individual SP gene was calculated as an indicator of copy number (Column B-F, Supplementary Table 4.1). For each gene, the ratio of the read depth from derivative pathotypes versus the progenitor isolate *Ph*560 was obtained (Column G-J Supplementary Table 4.1). During this analysis, it became apparent that the depth ratio was skewed by the different amount of sequence data available for each of the isolates. For example, a the majority of genes showed depth ratios larger than 1.5 in *Ph*612 versus *Ph*560 due to the larger amount of sequencing data for *Ph*612. Thus, a recalibration for the depth ratio was carried out to enable cross-isolate comparison. First, the average depth ratios of 1,073 SP genes in each derivative/progenitor comparison was calculated (e.g. 1.51 for *Ph*612/*Ph*560). Second, the depth ratio of individual genes was divided by the average value, removing bias effects caused by different amounts of sequence data (Column K-N, Supplementary Table 4.1). The recalibrated depth ratios of all SP genes are shown in scatter plots in Figure 4.2.

The normalized depth ratios of the SP genes in the four progenies were all close to one except for two genes, *PH612_08952* and *PH612_13953*, indicating no copy number reductions across the majority of the genes (Column K-N, Supplementary Table 4.1; Figure 4.2). For *PH612_08952*, isolates 584, 608 and 626 had the a normalized depth ratio close to 0.5 (the 211th point in Figure 4.2 A, C and D; Supplementary Table 4.1), suggesting a copy number reduction or allele deletion. However, the read mapping of the gene revealed a heterozygous genotype in the three isolates (Figure 4.3). In the three isolates, two positions (37,780 and 37,840 on Scaffold 272) show genotypes T/C, and another position on 37,830 displayed a 2 bp deletion in one allele. This deletion was most likely to be linked to the genotype T on the polymorphic site 37,840, as they were present together in several reads. Future PCR analysis is warranted to confirm this haplotype. The heterozygous structure on the three positions (37,780, 37,830 and 37,840) in the derivative pathotypes 608 and 626 was also observed in the progenitor 560, indicating no deletion of the parental allele in the two derivative pathotypes. Another gene *PH612_13953* also showed an apparent loss of copy number in pathotypes 584 and 608 (the 1,073th point in Figure 4.3 A and C), but a manual examination of the read mapping also suggested that the two parental alleles were retained in both isolates. Taken together, no allele deletions were detected across the 4 derivative pathotypes.

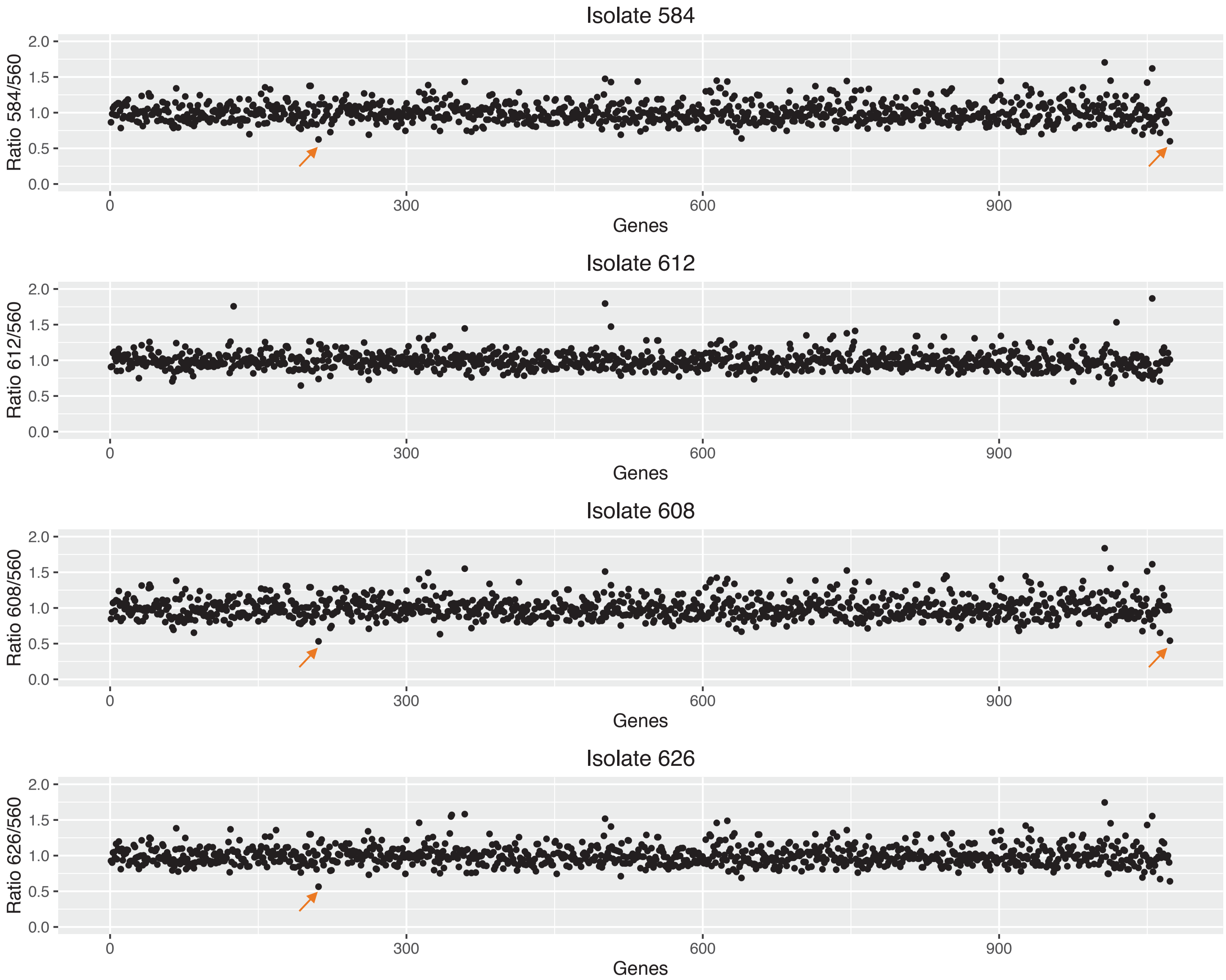


Figure 4.2: Read coverage depth on the secreted protein gene loci

Table 4.3: Genomic variants calling in the five *Puccinia hordei* isolates

Isolate	Total variants	SNP	Insertion	Deletion	Homozygous[1]	Heterozygous[2]
560	613,158	549,257	39,117	24,784	91,867	513,337
626	841,133	764,879	46,756	29,498	114,437	716,810
584	845,129	768,836	46,779	29,514	114,474	720,776
608	842,719	766,590	46,715	29,414	114,678	718,153
612	706,806	638,275	42,163	26,368	74,759	623,159

[1] Identical alleles for the SNP, Insertion or Deletion polymorphism from the two dikaryotic nuclei
[2] Different alleles for the SNP, Insertion or Deletion polymorphism from the two dikaryotic nuclei

4.2.4 Genome wide polymorphism of the five isolates

As no allele loss was identified for the SP genes, subsequent analysis focused on the inspection of sequence variations. Small variation events including single nucleotide polymorphisms (SNPs), small insertions and deletions (InDels) between individual isolates were examined based on mapping reads to the reference using GATK UnifiedGenotyper (McKenna et al., 2010). To consider a read base for variant calling, the UnifiedGenotyper required a minimum sequencing score of 17 (equivalent to error rate 1/17; 5.9%) for the base, reducing false positive calls due to sequencing errors. The calling of InDels was more stringent than SNPs, with a specialized requirement of a minimum five reads containing a consensus InDel and a minimum 25% of reads at the locus carrying the InDel.

The variant calling results for the five isolates are summarized in Table 4.3. For the putative progenitor pathotype 560, 613,158 variants were identified, occurring with a genome-wide frequency of 4.8/Kbp. Among the variants, SNPs and InDels showed a ratio of 8.6:1. A majority of the variants (88.3%) were in a heterozygous form, indicating large genetic divergence between the dikaryotic nuclei. The total number of sequence variants identified for the remaining four isolates ranged from 706,806 to 845,129 and showed similar ratios of SNP/InDel and homozygous/heterozygous SNPs.

SNPs occurred more frequently in transitions (changes A <->G and C <->T) than

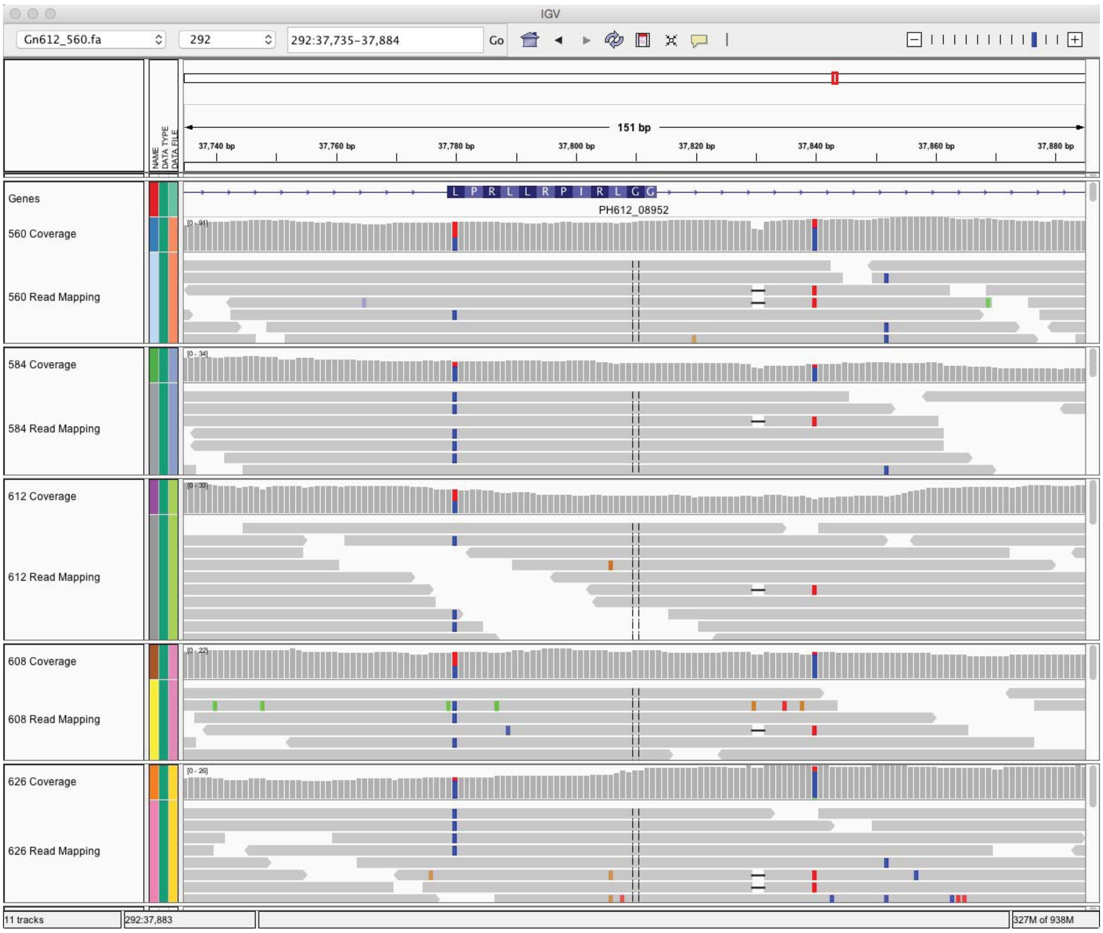


Figure 4.3: Read mapping of the five *Puccinia hordei* pathotypes to the reference genome Gn612_560. This window shows a local region of the gene *PH612_08952* from 37,735bp to 37,884bp on Scaffold 292. The left panel shows track names, and the right panel displays track details. On the tracks of Read Mapping and Coverage, polymorphism is shown with color bars: ref=T and blue=C.

Table 4.4: Transition and transversion details for SNVs of the five isolates

SNV type	Base change	Count 560	Count 626	Count 584	Count 608	Count 612
Transition	A>G	72,566	100,533	100,874	100,680	84,073
Transition	G>A	75,609	103,254	103,695	103,663	86,138
Transition	C>T	77,044	105,131	105,785	105,413	87,572
Transition	T>C	72,474	100,163	100,356	100,148	83,803
Transversion	A>C	39,761	55,122	55,335	55,131	49,995
Transversion	C>A	41,781	57,628	57,946	57,868	47,883
Transversion	A>T	26,164	40,351	40,942	40,550	30,340
Transversion	T>A	25,954	39,909	40,533	40,295	29,991
Transversion	C>G	18,545	25,921	26,028	25,881	21,309
Transversion	G>C	18,444	25,763	25,834	25,714	21,279
Transversion	G>T	41,635	57,206	57,441	57,225	47,321
Transversion	T>G	39,280	53,898	54,067	54,022	48,571

in transversions (changes A <->C, A <->T, G <->C or G <->T; Table 4.4). Within a species, transition/transversion ratio (Ti/Tv) is generally stable, which provides a useful diagnostic check for SNP calling performance. The Ti/Tv ratios in the *Ph* pathotypes ranged from 1.15 to 1.18, demonstrating the reliability of the SNP calling analysis.

In order to relate genomic variants to gene structures, a bioinformatics tool SnpEff (Cingolani et al., 2012) was used to map all the SNPs and InDels to seven types of genic locations: 5 Kbp upstream and downstream of genes, untranslated regions (UTR, 500 bp upstream and downstream of genes), exons, introns, splice sites, and intergenic regions. The upstream and downstream regions of one gene can overlap with another gene, especially in gene dense locations. The genomic distributions of variants were similar across all isolates. While the intergenic regions displayed the highest number of

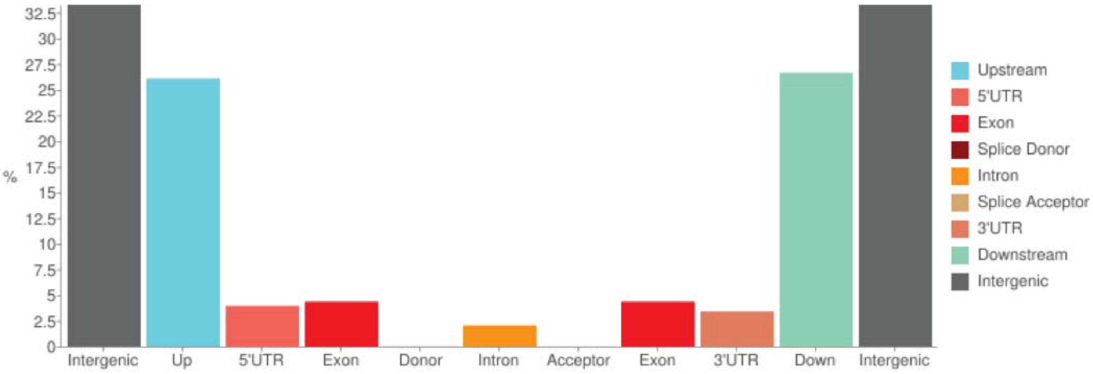


Figure 4.4: Percentage distribution of variants (SNV and InDel) over various genomic regions of *Ph*560. The genomic regions on X-axis are intergenic area, 5 Kbp upstream of gene, 5 prime UTR, exon, splice donor, intron, splice acceptor, exon, three prime UTR, 5 Kbp downstream of gene, and intergenic area.

variants (34 ± 1%), coding regions were less affected with only about 4% of variants (Table 4.5; Figure 4.4). The low variant density in coding sequences suggested selection pressure imposed on them to maintain stable proteins.

To characterize functional effects of sequence variants, SnpEff was used to determine the variants impact on protein coding, based on predicted coding regions of the 16,578 genes (Table 4.6). An InDel in a coding region could cause a coding frame shift if its size is not multiple of three, whereas an InDel of one or several codons induces a less severe effect on protein coding. InDels of these two types were referred to as frameshift and inframe mutations, respectively. The average counts of frameshift and inframe mutations derived from the five isolates were 1,456 and 1,109, respectively. In addition to codon frame change, sequence changes in start and stop codons may also have a high impact on gene function: loss of a start codon would abolish the gene; premature gained of a stop codon in coding regions would discard remaining amino acids (AA) at the 3' end; and a stop codon loss would incur more AAs that may result in a change in protein structure. These three types of mutation occurred with average amounts of 129, 1501 and 991 in the five pathotypes, respectively. Furthermore, in nonsynonymous (NSY) changes, the AA change caused by SNP, may also have direct functional implications, whereas synonymous (SYN) mutations referring to SNPs resulting in no AA change

Table 4.5: Variant distribution on genomic regions

Type (alphabetical order)	Count 560	Count 626	Count 584	Count 608	Count 612
Downstream	341,000	445,112	446,442	445,839	379,063
Exon	55,658	68,415	68,546	68,712	63,806
Intergenic	425,859	602,326	605,244	603,897	498,602
Intron	26,236	31,067	31,063	31,065	27,250
None	130	142	141	138	123
Splice_site_acceptor	115	145	153	145	138
Splice_site_donor	73	110	108	104	139
Splice_site_region	2,985	3,382	3,437	3,366	3,153
UTR_3_Prime	43,719	54,910	55,280	54,932	47,861
UTR_5_Prime	50,625	61,863	62,262	61,732	54,476

may not have such an effect. Across the five pathotypes, more NSY variants were detected as compared to the SYN variants, and the average counts of these two types were 36,084 and 24,019, respectively.

4.2.5 Genomic polymorphism in SP genes associated with avirulence/virulence

As effector proteins are most likely encoded by SP genes, our identification of the three Avr genes, *AvrRph3*, *AvrRph13*, and *AvrRph19* focused on detection of the mutations in SP genes between the 5 *Ph* isolates. From the presumed lineage relationship (Figure 4.1), it was believed that *AvrRph13* had mutated to virulence form in *Ph*608 derived from *Ph*584, and *AvrRph19* was modified to virulence form in *Ph*584 derived from *Ph*560. The modification of *AvrRph19* should be retained in the two decedents of *Ph*584, namely *Ph*612 and *Ph*608. *AvrRph3* had two independent mutations to virulence, from *Ph*560 to *Ph*626 and from *Ph*584 to *Ph*612.

Table 4.6: Modifications of genomic variants

Type (alphabetical order)	Count 560	Count 626	Count 584	Count 608	Count 612
Frameshift InDel	1,384	1,517	1,509	1,527	1,344
Inframe InDel	1,066	1,138	1,137	1,139	1,066
Non-Synonymous variant	30,675	38,058	38,203	38,179	35,304
Start codon lost	115	130	131	143	126
Stop codon gained	1,324	1,582	1,588	1,613	1,396
Stop codon lost	860	1,048	1,041	1,044	960
Synonymous variant	20,472	25,206	25,200	25,346	23,871

Candidates for *AvrRph13*

To identify candidates for *AvrRph13*, read mapping-based sequence comparison for each of the 1,073 SP genes was carried out between isolate 584 and 608. The two pathotypes displayed different genotypes at 544 positions in 160 SP genes in either a homozygous or heterozygous condition. The genetic polymorphism was mapped to the protein level based on CDS annotation, showing 185 intron variants, 124 SYN variants, and 235 protein changing variants in 99 genes (Figure 4.5A). Sequence variants that caused inframe InDel or AA modification were considered as moderate variants, whereas variants causing a coding frame shift or changing start/stop codons were regarded as having high impacts on protein functions. To further filter the 99 genes for a smaller list with top priority, genes with at least three moderate or one high impact variants were selected (Table 4.7), resulting in 21 genes.

Table 4.7: Top ranked candidates for *AvrRph13* and their functional annotation.

GeneID	Moderate impact variant	High impact variant	Nr	InterPro
PH612_05256	6	0	PGTG_13541	NA
PH612_13524	2	2	LOC100264919	IPR027806

Continued on next page

Table 4.7: Top ranked candidates for *AvrRph13* and their functional annotation. (continued)

GeneID	Moderate impact variant	High impact variant	Nr	InterPro
PH612_09003	5	0	PGTG_05567	IPR001623, IPR011990, IPR013026, IPR019734
PH612_04468	4	0	NA	NA
PH612_16351	4	1	PGTG_12287	NA
PH612_09800	10	0	PGTG_07617	NA
PH612_06301	5	0	NA	NA
PH612_08070	9	0	PGTG_15484	NA
PH612_08486	5	0	PGTG_01137	IPR008979, IPR013781, IPR017853
PH612_13926	5	1	PGTG_13447	NA
PH612_08215	5	0	PGTG_03187	NA
PH612_01773	3	3	PGTG_05774	NA
PH612_02528	5	0	NA	NA
PH612_01774	5	0	PGTG_07544	IPR017853, IPR022790
PH612_08941	4	1	PGTG_05929	NA
PH612_02193	8	2	NA	NA
PH612_02747	2	2	A306_02820	IPR009437
PH612_08029	4	0	NA	IPR010309
PH612_16275	4	0	PGTG_03481	IPR005198, IPR008928, IPR012341
PH612_05277	6	2	PGTG_10914	NA
PH612_08636	4	0	PGTG_08421	NA

Most currently identified Avr genes in rust fungi encode proteins that are species specific (Petre et al., 2014). To give conservation information for the 21 SP proteins,

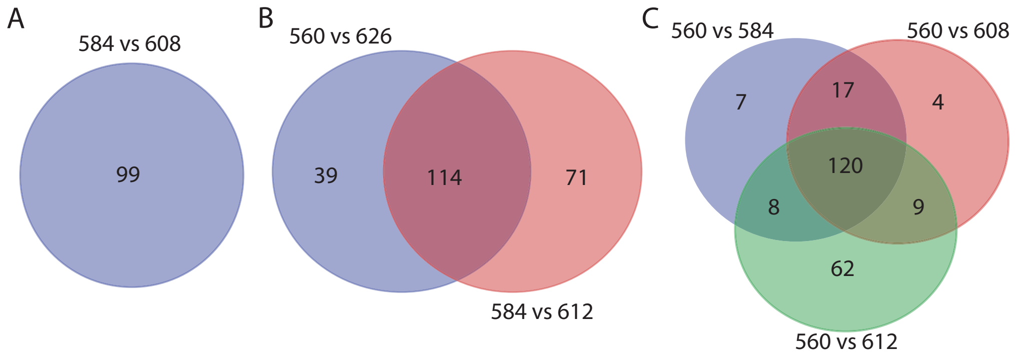


Figure 4.5: Venn diagrams for intersection and complement of secreted protein genes with non-synonymous mutations.

their NCBI Non-redoundant (Nr) and InterPro annotation performed in Chapter 3 was also displayed in Table 4.7. Sixteen of the 21 proteins had homologs in Nr database, and seven proteins contained protein signatures curated in InterPro database.

Candidates for *AvrRph3*

It was postulated that both *Ph*626 and *Ph*612 pathotypes gained the virulence to *Rph3* via independent single step mutations from their progenitor pathotypes, *Ph*560 and *Ph*584, respectively. Therefore, comparisons between avirulent and virulent isolates were carried out (*Ph*626 versus *Ph*560 and *Ph*612 versus *Ph*584) to detect the SP genes that showed functional variations. Because it was not known whether the gene had the same mutation for the two independent virulence gains, the SP genes that showed protein polymorphism in both comparisons were considered as potential candidates.

Within the SP genes, isolates 560 and 626 showed differences in 1,246 SNP sites. Of these SNPs, 275 were in CDS region with SYN effect; 424 were within intron regions, and the remaining 547 variants resulting in AA change were dispersed in 153 genes. Similar screening was used to compare pathotypes 584 and 612, which enabled the identification of 185 SP genes showing AA variations. The gene sets resulting from the two comparisons shared a common panel of 114 genes (Figure 4.5B), of which, 34 contained at least three moderate or one high impact variations in both comparisons. One of these 34 genes, *PH612_13377*, contained an IPS-annotated transposon domain from Transposase family tnp2, and was therefore removed from the candidate list. The details of the remaining 33 SP genes, including the total number of polymorphism and functional annotation are summarized in Table 4.8.

Table 4.8: Top ranked candidates for *AvrRph3* and their functional annotation.

GeneID	Moderate impact variant	High impact variant	Nr	InterPro
PH612_01790	5	0	PGTG_18959	NA
PH612_08215	10	0	PGTG_03187	NA
PH612_12987	6	1	PGTG_11120	IPR000560, IPR029033
PH612_01773	9	3	PGTG_05774	NA
PH612_04157	6	1	NA	NA
PH612_04959	5	0	NA	NA
PH612_05277	8	0	PGTG_10914	NA
PH612_11137	7	0	NA	NA
PH612_09091	5	2	NA	IPR015679
PH612_10437	5	0	NA	IPR002472, IPR030294
PH612_02747	5	1	A306_02820	IPR009437
PH612_04468	4	0	NA	NA
PH612_08636	6	0	PGTG_08421	NA
PH612_02148	7	0	NA	NA
PH612_10763	1	2	NA	NA
PH612_02193	12	3	NA	NA
PH612_16351	12	1	PGTG_12287	NA
PH612_08486	16	1	PGTG_01137	IPR008979, IPR013781, IPR017853
PH612_09800	11	1	PGTG_07617	NA
PH612_08070	40	0	PGTG_15484	NA
PH612_10385	10	0	PGTG_18748	NA

Continued on next page

Table 4.8: Top ranked candidates for *AvrRph3* and their functional annotation. (continued)

GeneID	Moderate impact variant	High impact variant	Nr	InterPro
PH612_09003	5	0	PGTG_05567	IPR001623, IPR011990, IPR013026, IPR019734
PH612_14034	16	2	PGTG_22484	IPR001584, IPR001878, IPR012337
PH612_12079	4	0	PGTG_02691	NA
PH612_10903	11	0	NA	NA
PH612_00759	8	2	PGTG_13292	NA
PH612_15319	6	0	PGTG_02691	NA
PH612_04864	8	0	NA	NA
PH612_00491	4	2	NA	IPR002509, IPR011330
PH612_13524	9	3	LOC100264919	IPR027806
PH612_12400	6	0	PGTG_12369	NA
PH612_01774	6	0	PGTG_07544	IPR017853, IPR022790
PH612_05256	5	0	PGTG_13541	NA

Candidates for *AvrRph19*

The pathotype 584 gained virulence for *Rph19* when it was derived from progenitor 560, and the virulence was retained in its two descendants 612 and 608. The two isolates 560 and 584 differed on 1,286 positions in 217 SP genes. These polymorphisms were composed of 458 intron variants, 287 SYN variants and 541 protein modifying variants that were distributed in 152 SP genes. With similar analyses framework, 150 and 199 SP genes displayed protein polymorphisms in comparing *Ph*560 to *Ph*608, and *Ph*560 to *Ph*612, respectively. As the *AvrRph19* needed to be changed in the three comparisons, the three sets of SP genes, which showed protein variations, were compared and they shared 120 common genes (Figure 4.5C). Of those genes, 23 were highlighted with more than three moderate or one high impact variants (Table 4.9).

Table 4.9: Top ranked candidates for *AvrRph19* and their functional annotation.

GeneID	Moderate impact variant	High impact variant	Nr	InterPro
PH612_00791	4	0	NA	NA
PH612_04923	8	0	PGTG_02965	IPR009437
PH612_09800	6	1	PGTG_07617	NA
PH612_15996	11	2	PGTG_11597	IPR027417
PH612_02193	7	3	NA	NA
PH612_13716	4	0	PGTG_10862	IPR010516
PH612_12582	7	0	PGTG_22720	NA
PH612_01773	6	3	PGTG_05774	NA
PH612_08029	9	0	NA	IPR010309
PH612_10385	13	0	PGTG_18748	NA
PH612_08486	16	1	PGTG_01137	IPR008979 IPR013781, IPR017853
PH612_08070	37	0	PGTG_15484	NA
PH612_02747	4	0	A306_02820	IPR009437
PH612_08215	10	0	PGTG_03187	NA
PH612_03444	5	0	PGTG_15716	NA
PH612_14034	16	2	PGTG_22484	IPR001584 IPR001878, IPR012337
PH612_00759	8	2	PGTG_13292	NA
PH612_00594	7	0	PGTG_13404	NA
PH612_05277	10	2	PGTG_10914	NA
PH612_00491	5	0	NA	IPR002509, IPR011330
PH612_10903	11	0	NA	NA

Continued on next page

Table 4.9: Top ranked candidates for *AvrRph19* and their functional annotation. (continued)

GeneID	Moderate impact variant	High impact variant	Nr	InterPro
PH612_16351	13	1	PGTG_12287	NA
PH612_13524	9	2	LOC100264919	IPR027806

4.2.6 Functional annotation

Effectors have been considered as highly adapted proteins for virulence functions specifically targeted at host immunity. The top-ranked candidates for the three Avr genes could have various *Ph* pathogenicity roles. Therefore, their biological functions were examined in more detail here. The homology of these proteins with those in the Nr protein database was examined and the results were listed in Tables 4.7, 4.8 and 4.9. A total of 46 genes were listed in the three tables, 30 of which had homolog hits in Nr. However, all of these homologs were described as hypothetical or uncharacterized proteins, providing functional information only on gene conservation between *Ph* and the species carrying the gene homolog. In contrast, IPS annotation did not rely on sequence conservation but on protein signatures, revealing domain functions in the SP proteins. These functions could be grouped broadly into two major categories: macromolecular modification (e.g. IPR010309 for ubiquitin ligase; IPR027806 for nuclease activity; IPR008979 for galactose-binding) and carbohydrate metabolism (e.g. IPR011330 and IPR013781 for glycoside hydrolase/deacetylase). These annotations, along with the sequence variant calling, provide information to set priority for future functional screens of Avr candidates.

4.3 Material and Methods

The four isolates (ID 560, 584, 608, 262; pathotypes 5453P-, 5453P+, 5453P+ +*Rph13*, 5457P-) selected for the re-sequencing study are archived in the Cereal Rust Collection maintained at the Plant Breeding Institute, University of Sydney. They were collected from 2001 to 2013 in Western Australia or Queensland (Table 4.1). For genomic DNA extraction, the slightly-modified CTAB method (Rogers et al., 1989) described in Chapter 3 was used. After quality checking with a Nanodrop Spectrophotometer (Thermo Scientific Ltd.), the DNA of *Ph*560 and *Ph*626 was sent to Novogene Ltd. for sequencing on Illumina Hiseq 2000 platform in the first batch, and the other two isolates (584 and 608) were sent to Australian Genome Research Facility Ltd. for sequencing on Illumina Hiseq 2500 in the second batch.

The raw sequencing reads were filtered and trimmed to remove low quality reads, sequencing adapters, and low quality ends by using Trim Galore v0.3.7[1]. The parameters used for this tool included "-quality 20" that demanded minimum base quality score of 20 to trim low quality nucleotides from 3' end, and "-length 35" setting minimum read length to keep after trimming. The amount of clean data after trimming for each isolate is shown in Table 4.2.

4.3.1 Assembly and annotation of *Ph*560-specific DNA

To assemble *Ph*560-specific DNA not present in *Ph*612 assembly, 46 million *Ph*560 reads were aligned to the assembly with Bowtie2 v 2.2.5 (Langmead and Salzberg, 2012) using parameters "-D 20 -R 3 -N 0 -L 20 -i S,1,0.50, -minins 0, -maxins 900", and the resulting 6.8 million unaligned reads were *de novo* assembled with SOAPdenovo2 vr240 (Luo et al., 2012) into 1,540 contigs with parameters "-R -K 25 -m 63 -L 100 -F -V -p 8". To characterize gene content in these contigs, the MAKER annotation pipeline used in Chapter 3 was employed to predict gene regions, by using *Pgt* proteins (Upadhyaya et al., 2015) as homology guidance. Signal peptides and transmembrane domains in the predicted genes were predicted using SignalP v4.1 (Petersen et al., 2011).

[1] http://www.bioinformatics.babraham.ac.uk/projects/trim_galore/

The newly assembled contigs were added to the scaffolds of isolate 612, resulting in a new genome referred to as Gn612_560.

4.3.2 Read mapping

To compare the genotype across all the isolates, their sequencing reads were mapped to Gn612_560 with Bowtie2 v2.2.5 (Langmead and Salzberg, 2012). For reference sequences longer than 300 bp, the mapping was run using parameters "-D 20 -R 3 -N 0 -L 20 -i S,1,0.50, -minins 0, -maxins 900 –end-to-end". For references shorter than 300 bp, the contig edge would not be able to recruit read mapping with the above parameters, and thus they were mapped using local alignment mode (parameters: -D 20 -R 3 -N 0 -L 20 -i S,1,0.50 -minins 0, -maxins 900, –local). After this initial mapping, reads with InDels at their edges where identified and remapped using RealignerTargetCreator and IndelRealigner in GATK package v3.3.0 (McKenna et al., 2010).

4.3.3 CNV analysis

Read depth analysis was performed to examine potential copy number deletion of SP genes in the derivative pathotypes. First, the locations of the 1,073 SP genes were compiled in a BED format file, which was a tab-delimited text file that contained three columns of scaffold ID, start and end positions for each gene. Second, the BED file was input to samtools (Li et al., 2009) to calculate total read base count for every SP gene in each pathotype. Third, base count for each gene was divided by its gene length. Finally, the read depth of each derivative pathotypes was normalized against the progenitor 560, and plotted using R ggplot2 package (Wickham, 2016) (Figure 4.2).

4.3.4 SNP and InDel calling

For genomic variants calling, the InDel realigned results were input in tool UnifiedGenotyper of GATK package v3.3.0 (McKenna et al., 2010), which produced a VCF format file containing calling results. Most variant callings based on UnifiedGenotyper have a filtering step to minimize false positive calls caused by repetitive elements. However, this study focused on variants in the 1,073 SP gene that showed uniform read depth (Figure 4.2) without fluctuation caused by repeats. Therefore a variant filtering was not necessary. To evaluate genome heterozygosity, heterozygous variants were selected from the GATK output, and variant number for each isolate was divided by effective genome size (assembly length 127.94 Mbp minus gap size 10.92 Kbp). In calculating variant frequency of genic regions, gene length was defined as region from start and end of coding sequence including introns, but not spanning UTR regions.

4.3.5 Variant annotation

The identified SNPs and InDels were annotated with SnpEff v4.1 (Cingolani et al., 2012). SnpEff needs a database for genomic annotations. The database of the *Ph* genome was built using a FASTA format file containing the reference genome, and a GFF format file representing genomic structures including intergenic regions, coding sequences and intron regions, UTR regions, and exon sequences. The UTR regions were approximated using 500 bp spaces before and after coding start and stop sites of genes, respectively. The two files were registered in an SnpEff configuration file named “snpEff.config”, which was further adjusted according to the software manual instructions[1]. After database construction, genomic variants produced by GATK UnifiedGenotyper in a VCF format file were assigned to genomic structures using SnpEff default command, which generated an annotated VCF file. Genotype differences across the isolates were identified using a custom Perl script that parse the annotated VCF file. The script also associated the genotypic differences to functional annotations in the VCF file. Candidate lists of AA polymorphisms derived from different isolate comparisons were delivered into an online tool[2] for identification of overlapping genes across different comparisons.

[1] http://snpeff.sourceforge.net/SnpEff_manual.html#databases

[2] http://bioinformatics.psb.ugent.be/webtools/Venn/

4.4 Discussion

To identify candidate Avr genes in *Ph*, comparisons were made between the whole genome sequences of isolates of five *Ph* pathotypes that comprised a progenitor and four mutational derivatives that differed in virulence for three barley leaf rust resistance genes (*viz. Rph3, Rph13,* and *Rph19*) (Figure 4.1; Table 4.1). The sequencing reads of the five isolates were mapped to the reference genome Gn612_560, a pan-genome constructed from the DNA of the isolates 612 and the presumed progenitor 560. The mapping rates ranged from 78%-86% with an average rate of 85% (Table 4.2). The remaining 15% of unmapped reads may have come from the gap regions or mitochondria, where DNA sequences were not present in our reference genome. Compared with previous studies of rust fungi, the mapping rate reported here is sound. For example, Upadhyaya et al. (2015) mapped ∼67% reads from four Australian *Pgt* isolates to a reference genome p7a assembled from an American isolate of *Pgt*. This mapping rate is lower than that of the present study, probably due to the genetic divergence between the Australian and American isolates. Persoons et al. (2014) mapped their re-sequenced isolates of *Mlp* with an average rate of 78%, also lower than in the present study, probably due to their application of high stringency parameters (100% of a read sequence map with at least 95% similarity to the reference). Furthermore, the quality of the reference genome may impact partially on the mapping rate. In general, the mapping rates of re-sequenced isolates using Sanger sequencing or NGS were lower than 90%, suggesting the absence of more than 10% genomic sequence in the reference genomes (Cantu et al., 2013; Wu et al., 2017; Zheng et al., 2013). With the advent of third generation sequencing (TGS) technologies (e.g. PacBio SMRT and Oxford Nanopore), higher mapping rates can be achieved as TGS enables more continuous and complete genome assemblies. For instance, mapping Illumina sequencing reads of a *Pst* isolate to a reference genome assembled with the latest PacBio sequencing protocol achieved a mapping rate as high as 96.8% (Schwessinger et al., SlideShare[1]).

When mapping the five *Ph* isolates, over 99% of the reference could be covered by reads from each isolate, suggesting that almost all genes annotated in the reference

[1] https://www.slideshare.net/BenjaminSchwessinger1/generating-haplotype-phased-reference-genomes-for-the-dikaryotic-wheat-stripe-rust-fungus-62757887

could be genotyped. While the distributions of the read depth mapped on the reference were similar across all isolates in this study, the study of *Mlp* by Persoons et al. (2014) showed that 6, 4 and 5 isolates were mapped to scaffold 90 with three distinct patterns, ranging from full, lower, and no coverage in certain regions. Compared with these differential patterns, the uniform coverage patterns observed for the *Ph* isolates here support our hypothesis that the five isolates examined are derived from a clonal lineage with limited genetic diversity.

To investigate possible Avr allele deletion in the derivative pathotypes, CNV analysis was performed. No reduction in copy number was detected for the 1,073 SP genes (Figure 4.2 and 4.3). A previous study of *Fusarium oxysporum* by Schmidt et al. (2016) reported that three isolates gained virulence to R gene *Fom2* by deleting the corresponding Avr gene *AvrFom2* in their genomes. The genomic deletions in these isolates spanned regions totaling 2.3 Mbp. The isolates examined were maintained under artificial greenhouse conditions, ensuring survival. In nature however, such large deletions would be expected to reduce pathogen fitness and possibly survival. The *Ph* isolates studied here were all collected from the field, where large genomic deletions would likely incur a high fitness penalty. Although this is a pilot study with limited sample size, our initial results seem to indicate that the virulence gains in the pathotypes derived from *Ph* 5453P- did not arise from gene deletion, but from allelic variation.

Here, the copy number reduction of genes was measured through inspection of read depth in a derivative virulent pathotype compared to the progenitor, whereas Schmidt et al. (2016) used read coverage for a gene to determine its absence/presence. A different method was used here because *Ph* has a diploid genome, whereas *F. oxysporum* is a haploid organism. With the advent of TGS, construction of a diploid reference genome with phasing of alleles will be expected to reduce the complexity of copy number analysis for many dikaryotic rust genomes such as *Ph* (Mostovoy et al., 2016).

Variant calling including SNPs and InDels was performed based on the read mapping, which predicted 613-845k variants for the five isolates with genome wide densities ranging from 4.8-6.6/kb (Table 4.3). The densities were notably lower than those observed for six *Pgt* isolates (14.2/kb; Upadhyaya et al., 2015), probably because the *Pgt* isolates belong to different clonal linages, and because they were collected over longer time spans (average 23.4 years) than the *Ph* isolates used in this study (average 7.3 years; Table 4.1). Indeed, the two clonal *Pgt* isolates studied in Chapter 2 were collected in the greenhouse in the same year and they showed as few as 59 NSY variants. The SNP and InDel ratio identified for the five isolates was about 8.6:1 (Table 4.3), consistent with the fact that InDels as length variants are often more deleterious to gene functions than are SNPs.

As a majority of the reference pan-genome Gn612_560 was assembled with *Ph*612 reads, mapping *Ph*612 reads back to the reference should have found variants only in heterozygous positions, because the assembly algorithm SOAPdenovo arbitrarily selects one of two dikaryotic alleles as a haploid consensus (Li et al., 2010). However, 74,759 homozygous variants were called for *Ph*612 (Table 4.3). Manual inspection of these variant positions found that most were caused by assembly errors in the reference. Given the reference genome size of 128 Mbp, the error rate of the assembly is very low (0.06%), which should have a negligible effect on the genotyping results based on nucleotides in the mapped reads. On the other hand, *Ph*612 did show the minimum number of homozygous variants as compared to the other isolates, which is consistent with the fact that a majority of the reference genome was assembled from this isolate.

It is not known whether the progenitor isolate 560 is homozygous or heterozygous at the three Avr loci *AvrRph3*, *AvrRph13* and *AvrRph19*. If homozygous, both alleles of the Avr gene need to mutate in order for virulence gain. In the latter situation, mutation of the one Avr allele only is required. In this study, both homozygous and heterozygous polymorphisms in the virulent isolates were analyzed to fit both scenarios. By focusing on 1,073 SP genes, this study identified 99, 114 and 120 genes that encode protein polymorphisms between isolates with differential virulence profiles for the R genes *Rph3*, *Rph13* and *Rph19*, respectively (Figure 4.1 and 4.5). A smaller candidate list of 15 secreted proteins with AA differences across six *Pt* isolates was obtained in a previous study by Bruce et al. (2013). However, this study was performed with consensus transcripts assembled from RNA-Seq data without considering dikaryotic allelic variations, and hence may have missed heterozygous polymorphisms.

Various methods have been used to screen gene repertoires in order to reduce the search scope for Avr genes. One proven method is to select genes that are preferentially expressed in haustoria, as this specialized structure had been shown to be enriched for effectors in a flax rust study (Catanzariti et al., 2006). In the Australian *Pgt* pan-genome project, 1,924 secreted proteins were predicted, of which 520 were considered primary candidates based on their upregulation in haustoria over germinated spores (Upadhyaya et al., 2015). The set of 520 genes served as a starting point for a comparative genomics analysis that led to the identification of *AvrSr50* in Chapter 2. Similar numbers of effector candidates were highlighted by haustoria transcriptomics in *Pst* (437; Garnica et al., 2013) and the common bean rust fungus *Uromyces appendiculatus* (395; Link et al., 2014). Future studies of *P. hordei* will include haustoria cDNA library sequencing and the identification of differentially expressed genes, which will further reduce the number of the candidate genes identified in this study.

Compared with other cereal rust species present in Australia, *Ph* is the only one that undergoes sexual recombination (Wallwork et al., 1992), making it a potential model pathogen for studying cereal-rust pathosystems. Sexual crosses of *Ph* isolates and following segregation of Avr genes in F_2 progenies should provide useful material for Avr gene identification and cloning using a map-based approach. This research method has provided substantial knowledge of the flax rust-flax interaction, leading to the isolation of several Avr genes (Anderson et al., 2016; Catanzariti et al., 2006; Dodds et al., 2004). Apart from traditional maker-based cloning, genome wide association studies (GWAS) by sequencing should also provide a powerful tool for Avr gene identification in isolates of *Ph* derived from sexual recombination. This technology has been shown to be efficient in several studies of plant pathogen populations with sexual recombination (Lu et al., 2016; Plissonneau et al., 2017; Praz et al., 2017).

The SP genes identified with polymorphisms associated with virulence profiles were prioritized based on their variant effects (Table 4.7, 4.8, and 4.9). Sequencing data from additional isolates will help to distinguish real Avr genes from background genetic diversity that is not related to virulence phenotype. With a more focused candidate list, functional validation of Avr genes using heterologous expression systems to deliver Avr gene products into barley lines containing *Rph3, Rph13,* and *Rph19* could be carried out as previous research suggested (Tingay et al., 1997). The functional study of these candidate genes will provide deeper understanding of the *Ph*-barley pathosystem.

The whole genome sequencing data of five isolates reported here provide a valuable resource for developing new DNA markers that will be useful for genetic studies and population in monitoring of the *Ph* 5453P- lineage in Australia and beyond. Several algorithms have been developed to design simple sequence repeat markers based on genome re-sequencing data (Du et al., 2013; Kitchen et al., 2012; Metz et al., 2016). Combined with these tools, the genome sequence data should prove useful to the rust research community after deposition in public database (e.g. NCBI sequencing reads archive). All this work will ultimately contribute to the development of more sustainable disease control and more effective surveillance and management strategies for this pathogen of barley.

References

Anderson, C., Khan, M. A., Catanzariti, A.-M., Jack, C. A., Nemri, A., Lawrence, G. J., et al. (2016). Genome analysis and avirulence gene cloning using a high-density RADseq linkage map of the flax rust fungus, *Melampsora lini*. *BMC Genomics* 17, 667 129

Bruce, M., Neugebauer, K. A., Joly, D. L., Migeon, P., Cuomo, C. A., Wang, S., et al. (2013). Using transcription of six *Puccinia triticina* races to identify the effective secretome during infection of wheat. *Frontiers in Plant Science* 4 128

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., et al. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics* 14, 270 104, 126

Catanzariti, A.-M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., and Ellis, J. G. (2006). Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *The Plant Cell* 18, 243–256 129

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., et al. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly* 6, 80–92 113, 125

de Jonge, R., van Esse, H. P., Maruthachalam, K., Bolton, M. D., Santhanam, P., Saber, M. K., et al. (2012). Tomato immune receptor *Ve1* recognizes effector of multiple fungal pathogens uncovered by genome and RNA sequencing. *Proceedings of the National Academy of Sciences U.S.A.* 109, 5110–5115 103

Dodds, P. N., Lawrence, G. J., Catanzariti, A.-M., Ayliffe, M. A., and Ellis, J. G. (2004). The *Melampsora lini AvrL567* avirulence genes are expressed in haustoria and their products are recognized inside plant cells. *The Plant Cell* 16, 755–768 129

Du, L., Li, Y., Zhang, X., and Yue, B. (2013). MSDB: a user-friendly program for reporting distribution and building databases of microsatellites from genome sequences. *Journal of Heredity* 104, 154–157 130

Flor, H. H. (1971). Current status of the gene-for-gene concept. *Annual Review of Phytopathology* 9, 275–296 106

Garnica, D. P., Upadhyaya, N. M., Dodds, P. N., and Rathjen, J. P. (2013). Strategies for wheat stripe rust pathogenicity identified by transcriptome sequencing. *PloS One* 8, e67150. doi:10.1371/journal.pone.0067150 129

Gilmour, J. (1973). Octal notation for designating physiologic races of plant pathogens. *Nature* 242 106

Golegaonkar, P., Park, R., and Singh, D. (2010). Genetic analysis of adult plant resistance to *Puccinia hordei* in barley. *Plant Breeding* 129, 162–166 103

Karaoglu, H. and Park, R. (2014). Isolation and characterization of microsatellite markers for the causal agent of barley leaf rust, *Puccinia hordei*. *Australasian Plant Pathology* 43, 47–52 104

Kitchen, J. L., Moore, J. D., Palmer, S. A., and Allaby, R. G. (2012). MCMC-ODPR: primer design optimization using Markov Chain Monte Carlo sampling. *BMC Bioinformatics* 13, 287 130

Langmead, B. and Salzberg, S. L. (2012). Fast gapped-read alignment with Bowtie 2. *Nature Methods* 9, 357–359 123, 124

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., et al. (2009). The sequence alignment/map format and SAMtools. *Bioinformatics* 25, 2078–2079 124

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., et al. (2010). *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research* 20, 265–72. doi:10.1101/gr.097261.109 107, 128

Link, T. I., Lang, P., Scheffler, B. E., Duke, M. V., Graham, M. A., Cooper, B., et al. (2014). The haustorial transcriptomes of *Uromyces appendiculatus* and *Phakopsora pachyrhizi* and their candidate effector families. *Molecular Plant Pathology* 15, 379–393 129

Lu, X., Kracher, B., Saur, I. M., Bauer, S., Ellwood, S. R., Wise, R., et al. (2016). Allelic barley MLA immune receptors recognize sequence-unrelated avirulence effectors of the powdery mildew pathogen. *Proceedings of the National Academy of Sciences U.S.A.* 113, E6486–E6495 129

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W., Yuan, J., et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience* 1, 18 123

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., et al. (2010). The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research* 20, 1297–1303 111, 124

Mesarich, C. H., Griffiths, S. A., van der Burgt, A., Ökmen, B., Beenen, H. G., Etalo, D. W., et al. (2014). Transcriptome sequencing uncovers the *Avr5* avirulence gene of the tomato leaf mold pathogen *Cladosporium fulvum*. *Molecular Plant-Microbe Interactions* 27, 846–857 103

Metz, S., Cabrera, J. M., Rueda, E., Giri, F., and Amavet, P. (2016). FullSSR: Microsatellite Finder and Primer Designer. *Advances in Bioinformatics* 2016 130

Mostovoy, Y., Levy-Sakin, M., Lam, J., Lam, E. T., Hastie, A. R., Marks, P., et al. (2016). A hybrid approach for *de novo* human genome sequence assembly and phasing. *Nature Methods* 13, 587–590 127

Park, R. (2008). Breeding cereals for rust resistance in Australia. *Plant Pathology* 57, 591–602 103

Park, R. F., Golegaonkar, P. G., Derevnina, L., Sandhu, K. S., Karaoglu, H., Elmansour, H. M., et al. (2015). Leaf rust of cultivated barley: pathology and control. *Annual Review of Phytopathology* 53, 565–589 103, 104, 105

Persoons, A., Morin, E., Delaruelle, C., Payen, T., Halkett, F., Frey, P., et al. (2014). Patterns of genomic variation in the poplar rust fungus *Melampsora larici-populina* identify pathogenesis-related factors. *Frontiers in Plant Science* 5, 450 104, 126, 127

Petersen, T. N., Brunak, S., von Heijne, G., and Nielsen, H. (2011). Signalp 4.0: discriminating signal peptides from transmembrane regions. *Nature Methods* 8, 785–786 123

Petre, B., Joly, D. L., and Duplessis, S. (2014). Effector proteins of rust fungi. *Frontiers in Plant Science* 5 117

Plissonneau, C., Benevenuto, J., Mohd-Assaad, N., Fouché, S., Hartmann, F. E., and Croll, D. (2017). Using population and comparative genomics to understand the genetic basis of effector-driven fungal pathogen evolution. *Frontiers in Plant Science* 8, 119 129

Praz, C. R., Bourras, S., Zeng, F., Sánchez-Martín, J., Menardo, F., Xue, M., et al. (2017). *AvrPm2* encodes an RNase-like avirulence effector which is conserved in the two different specialized forms of wheat and rye powdery mildew fungus. *New Phytologist* 213, 1301–1314 129

Rogers, S. O., Rehner, S., Bledsoe, C., Mueller, G. J., and Ammirati, J. F. (1989). Extraction of DNA from *Basidiomycetes* for ribosomal DNA hybridizations. *Canadian Journal of Botany* 67, 1235–1243 123

Sandhu, K., Karaoglu, H., and Park, R. (2016). Pathogenic and genetic diversity in *Puccinia hordei* Otth in Australasia. *Journal of Plant Breeding and Crop Science* 8, 197–205 104

Schmidt, S. M., Lukasiewicz, J., Farrer, R., Dam, P., Bertoldo, C., and Rep, M. (2016). Comparative genomics of *Fusarium oxysporum* f. sp. *melonis* reveals the secreted protein recognized by the *Fom-2* resistance gene in melon. *New Phytologist* 209, 307–318 127

Tingay, S., McElroy, D., Kalla, R., Fieg, S., Wang, M., Thornton, S., et al. (1997). *Agrobacterium tumefaciens*-mediated barley transformation. *The Plant Journal* 11, 1369–1376 129

Upadhyaya, N. M., Garnica, D. P., Karaoglu, H., Sperschneider, J., Nemri, A., Xu, B., et al. (2015). Comparative genomics of Australian isolates of the wheat stem rust pathogen *Puccinia graminis* f. sp. *tritici* reveals extensive polymorphism in candidate effector genes. *Frontiers in Plant Science* 5, 759 104, 107, 126, 127, 129

Wallwork, H., Preece, P., and Cotterill, P. (1992). *Puccinia hordei* on barley and *Omithogalum umbellatum* in South Australia. *Australasian Plant Pathology* 21, 95–97 129

Wickham, H. (2016). *ggplot2: elegant graphics for data analysis* (Springer) 124

Wu, J. Q., Sakthikumar, S., Dong, C., Zhang, P., Cuomo, C. A., and Park, R. F. (2017). Comparative genomics integrated with association analysis identifies candidate effector genes corresponding to *Lr20* in phenotype-paired *Puccinia triticina* isolates from Australia. *Frontiers in Plant Science* 8 104, 126

Zheng, W., Huang, L., Huang, J., Wang, X., Chen, X., Zhao, J., et al. (2013). High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nature Communications* 4, 2673. doi:10.1038/ncomms3673 104, 126

Chapter 5

General Discussion

To ensure successful infection and colonization in plants, pathogens secrete effector proteins into their hosts to modulate physiology and immunity. Therefore, the identification and functional study of effectors is an important topic in the research targeting plant-pathogen interactions. Many effector proteins were discovered based on their recognition by plant immunoreceptors in an Avirulence-Resistance gene (Avr and R gene) relationship, e.g. flax rust effectors *AvrM, AvrL567, AvrP4, AvrP2,* and *AvrM14* (Anderson et al., 2016; Catanzariti et al., 2006; Dodds et al., 2004, 2006). Some other effectors were found by insertional mutagenesis targeting specific genes followed by observation of virulence reduction as a result of loss of gene function (Jeon et al., 2007; Michielse et al., 2009; Saitoh et al., 2012).

5.1 Effector identification via comparative genomics

With the advent of NGS, reference genomes have been assembled for over 200 plant pathogens in the last few years. In a comparative genomics approach, genomes of individual isolates are sequenced and compared through alignment of sequencing reads to the reference genome. Such comparisons reveal genomic differences between the selected isolates or strains, and subsequently the genotypic variations detected are related to virulence differences in order to find effectors.

5.1.1 Inter-species comparison

The principle of effector identification in comparing different pathogens is to investigate species-specific genome content. In one such study that compared the smut fungi *Ustilago maydis* and *Sporisorium reilianum*, 43 genomic regions harboring genes with low sequence similarity were identified (Schirawski et al., 2010). A majority of the genes (71%) in these regions were conserved in both fungi, while 19% were specific to *S. reilianum* and the remaining 10% were *U. maydis*-specific. These regions were enriched for secreted proteins and included virulence clusters, indicating their important role in pathogenesis. Deletion experiments of four of these regions in *U. maydis* confirmed their role in pathogenicity by resulting in virulence reduction. In a later study, genome comparison between *U. maydis* and *U. hordei* revealed a higher divergence in effectors relative to the rest of the proteome, indicating the feasibility of a comparative genomics approach to identify these effector genes (Laurie et al., 2012).

Similarly, comparative genomics of three phylogenetically related species in the genus *Fusarium* identified lineage-specific (LS) genomic regions in *F. oxysporum*, which included four whole chromosomes. The effector genes *Six1* and *Six3* (Ma et al., 2013, 2010), along with other candidate effector genes, were all located in Chromosome 14, one of the LS chromosomes. Accordingly, it was proposed that the Chromosome 14 was a major determinant towards pathogenicity on tomato in *F. oxysporum*.

5.1.2 Intra-species comparison

Comparative genomics of isolates within species relate genotypic differences to their pathotype differences or Avr gene differences for effector identification. de Jonge et al. (2012) compared the genomes of four virulent and seven avirulent *Verticillium dahliae* isolates to immune receptor gene *Ve1* in tomato, and identified a 50 Kbp genomic region absent in all virulent isolates. One gene in this region was highly expressed in infected host tissues, and subsequent functional assays confirmed that the gene encoded for avirulence effector function. Similarly, the *AvrFom2* in *F. oxysporum* f. sp. *melonis*, a melon pathogen, was identified in a genomic comparison of isolates avirulent/virulent to the R gene *Fom-2*, based on presence/absence in the avirulent/virulent isolates, respectively (Schmidt et al., 2016). In these two studies, the avirulence genes showed presence/absence polymorphisms. In contrast, the *Avr5* effector from *Cladosporium fulvum* was found to have mutated, with a 2 bp exon deletion in the virulent isolate IPO 1979, and the gene was identified based on this simple deletion by comparing the virulent isolate to another avirulent isolate OWU (Mesarich et al., 2014). This study suggests that searches for Avr genes should also include genes that encode protein polymorphisms.

The rationales behind these three studies were implemented in this thesis. In Chapter 2, the *Puccinia graminis* f. sp. *tritici* (*Pgt*) Avr gene *AvrSr50* was narrowed down to 18 genes based on allele loss of these genes in the *Sr50*-virulent isolate *Pgt*632, which showed a loss-of-heterozygosity event when compared to its *Sr50*-avirulent parent *Pgt*279. In Chapter 4, the genomes of five *Puccinia hordei* (*Ph*) isolates with single virulence gains to three barley resistance genes *Rph3, Rph13,* and *Rph19* were sequenced and compared, yielding 99, 114, and 120 candidate genes for the three corresponding Avr genes, respectively.

5.1.3 Rust genome assembly

A high-quality genome assembly is an essential starting point for comparative genomics. For fungal species with small (less than 50 Mbp), haploid, and less repetitive genomes, NGS has produced large scaffolds that span long DNA sequences or even complete chromosomes (Amselem et al., 2011; Cuomo et al., 2007; Kämper et al., 2006; Rouxel et al., 2011). However, genome sizes of rust fungi are among some of the largest in eukaryotic plant pathogens. A flow cytometry study of 30 rust species identified elevated genomes sizes that ranged from 77 Mbp in *Puccinia triticina* to 893 Mbp in *Gymnosporangium confusum*, whereas the average genome size for Basidiomycota was only 50 Mbp (Tavares et al., 2014). More recently, Ramos et al. (2015) revealed that the rust fungus, *Uromyces bidentis*, contained an estimated genome size of 2.5 Gbp, the largest known for fungi.

In addition, individual dikaryon rust genomes are composed of two divergent nuclei that are highly heterozygous, with tens to hundreds of inter-nuclei SNPs in every 10 kilobases (Upadhyaya et al., 2015; Wu et al., 2017; Zheng et al., 2013). High heterozygosity causes more fragmentations in genome assembly, because highly heterozygous regions that have sufficient diversity are assembled as separate fragments (Huang et al., 2012; Li et al., 2012), although most NGS assemblers are designed to merge homologous regions as a haploid consensus copy (Li et al., 2010; Simpson et al., 2009; Zerbino and Birney, 2008). Assembly fragmentation can cause problems for downstream genomics analysis, including apparent paralogous genes (i.e. wrong gene copy number), duplicated genomic regions, and synteny analysis.

Rust genomes are notoriously repetitive, with about 50% of genomic regions comprising repetitive elements (Cuomo et al., 2017; Duplessis et al., 2011; Nemri et al., 2014; Zheng et al., 2013). These repeat ratios may have been underestimated, as repetitive copies with sufficient similarity could be collapsed to one copy by most NGS assemblers. For instance, the genome assemblies produced using Sanger and Illumina sequencing for *V. dahlae* estimated a 4% repeat content in the genome (de Jonge et al., 2013; Klosterman et al., 2011), significantly underestimated compared to the 12% based on a recently released gap-less genome assembly produced using PacBio sequencing (Faino et al., 2015).

From a computational perspective, repeats longer than read length cause ambiguous reads that cannot be positioned exclusively in the genome. Subsequently, genome assemblers output gap sequence or fragment contigs in the ambiguous regions (Treangen and Salzberg, 2012). Though *Pgt* (wheat stem rust) and *Mlp* (poplar rust) were both sequenced with Sanger technology that produced long reads (about 1 Kbp), the assemblies still contain 392 and 462 scaffolds, respectively (Duplessis et al., 2011). Compared to the first-generation Sanger technology, NGS lowers per-base sequencing cost at the expense of read length, and accordingly results in dramatically-fragmented genome assemblies. The published rust genome assemblies using NGS are considerably more fragmented, with a minimum of thousands of scaffolds (Cantu et al., 2013; Link et al., 2014; Zheng et al., 2013).

5.1.3.1 Effector genes in repeat regions

The read length limitation of NGS also causes potential assembly problems for effector genes located in repetitive genomic regions. Many effector genes have been found in repeat-rich genomic islands, where transposons facilitate effector duplication and diversification (Dong et al., 2015; Raffaele and Kamoun, 2012). For instance, the AT-rich regions in the blackleg fungus *Leptosphaeria maculans* comprise mainly transposons and effector genes (Rouxel et al., 2011). Similarly, an effector gene *AVR-Pita* in *Magnaporthe oryzae* was found in a transposon-rich, sub-telomeric region (Orbach et al., 2000).

5.1.3.2 Effector genes with multiple copies

For multi-copy Avr genes, the short reads of NGS may assemble inaccurate copy number, as paralogous copies share high sequence similarity. One instance is the “collapsed” assembly of six paralogs from the *AvrM* family in flax rust using 75 bp Illumina reads (Nemri et al., 2014). In Chapter 2, three multi-copy genes identified as *AvrSr50* candidates were previously assembled as one locus in the reference genome PGTAus-pan by Upadhyaya et al. (2015) using Illumina 75 bp reads. Despite the difficulty in assembling multi-copy genes, it is important to ensure accurate copy number, as paralogous variants of Avr genes can be associated with recognition specificities by their host R genes in a gene-for-gene manner (e.g. the *AvrL567* variants in flax rust differentially recognized by flax R genes *L5, L6,* and *L7* (Dodds et al., 2006)).

The difficulties in effector gene assembly with NGS in these scenarios hinder effector genomics including accurate identification of effector content and diversity analyses (Gibriel et al., 2016; Thomma et al., 2016). One solution is to identify reads that belong to candidate effector genes and then perform a targeted assembly of these reads. This idea has been developed in a bioinformatics pipeline for targeted assembly of the var gene family in the human pathogen *Plasmodium falciparum* (Assefa, 2013).

5.1.4 Third-generation sequencing and plant pathogens

The recent development of third-generation sequencing (TGS), mainly the SMRT platform by Pacific Biosciences and Oxford Nanopore technology, solves the short read limitation in NGS by producing reads of tens of kilobases (Goodwin et al., 2016; Levy and Myers, 2016). Ten kbp reads are longer than most repeat elements in microbial genomes, and accordingly they will improve assembly contiguity for rust fungi. Thus far, two gapless fungal plant pathogen genomes have been completed using TGS combined with optical mapping technology (Faino et al., 2015; Van Kan et al., 2017). In another study, TGS assembled the whole genome of the anther-smut fungus, *Microbotryum lychnidis-dioicae*, into gapless chromosomes or chromosome arms, even for the highly repetitive mating-type chromosomes a1 and a2 (Badouin et al., 2015).

Increased read lengths from TGS also allow haplotype phasing over long genomic distances in diploid or polyploid genomes. Rust fungi are dikaryotic with two copies of each chromosome. The two haplotypes with alleles at variant sites contain the complete genetic information in a rust isolate. Recovery of haplotypes from sequencing data is important for understanding rust genetic variation, and the interpretation of these variants in relation to isolate virulence variation. Current NGS assemblers produce a mosaic sequence by arbitrarily selecting allelic variants at heterozygous loci. This is mainly due to the short reads that make it infeasible to link distant heterozygous variants into haplotypes. In contrast, TGS reads are long enough to span multiple heterozygous variants. Recently, several bioinformatic tools have been developed to enable accurate haplotype assembly from long reads (Chin et al., 2016; Edge et al., 2017; Koren et al., 2017; Kuleshov, 2014). Along with TGS, these tools hold great promise in enabling the construction of haplotype phased reference genomes for rust fungi, which will be beneficial in identifying haplotype-specific information in the re-sequencing of isolates. Such information includes allele-specific expression of Avr genes related to isolate pathotypes (Gilroy et al., 2011; Pais et al., 2017).

Future work for the two cereal rust fungi studied here will include draft genome improvements with TGS, which will provide higher-resolution of genome landscapes than NGS. For instance, a more contiguous reference genome with fewer gaps may enable pinpointing the recombination breakpoint in *Pgt*632, which was not achieved here due to its probable location in the assembly gaps of Scaffold 4 on the reference PGTAus-Pan (Figure 2.4).

5.2 Concluding remarks

In this thesis, NGS techniques were used to perform comparative genomics, with an aim to identify avirulence genes in two cereal rust fungi, *Pgt* and *Ph*. The work in Chapter 2 led to the first identification of an avirulence protein in a cereal rust pathogen. In Chapter 3, a high quality draft genome was assembled and annotated for *Ph*, providing the first genomic resource of this pathogen for the rust research community. In Chapter 4, the genomes of five *Ph* isolates within a clonal lineage were sequenced and compared, yielding approximately 100 candidates for three avirulence genes that are associated with virulence differences in these isolates. With advancing sequencing technologies, the logical frame and bioinformatics methods developed in this thesis will contribute to future efforts in rust genomics.

The identification of avirulence genes allows detection of their alternate virulence alleles in natural rust populations, enabling molecular surveillance to determine virulence/avirulence and the heterozygosity/homozygosity of rust isolates for important resistance genes. Estimating the propensity of field isolates to evolve new virulences will allow further refinement in pre-emptive resistance breeding, reducing the probability of large-scale epidemics from developing. In addition, the future identification of more R-Avr gene pairs will provide opportunities to study the avirulence effector targets, and investigate molecular mechanisms of susceptibility and resistance in cereal crops. This knowledge will ultimately contribute to the development of more durable resistance in cereal cultivars.

References

Amselem, J., Cuomo, C. A., Van Kan, J. A., Viaud, M., Benito, E. P., Couloux, A., et al. (2011). Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. *PLoS Genetics* 7, e1002230 138

Anderson, C., Khan, M. A., Catanzariti, A.-M., Jack, C. A., Nemri, A., Lawrence, G. J., et al. (2016). Genome analysis and avirulence gene cloning using a high-density RADseq linkage map of the flax rust fungus, *Melampsora lini*. *BMC Genomics* 17, 667 136

Assefa, S. A. (2013). *De novo assembly of the var multi-gene family in clinical samples of Plasmodium falciparum*. Ph.D. thesis, University of Cambridge 140

Badouin, H., Hood, M. E., Gouzy, J., Aguileta, G., Siguenza, S., Perlin, M. H., et al. (2015). Chaos of rearrangements in the mating-type chromosomes of the anther-smut fungus *Microbotryum lychnidis-dioicae*. *Genetics* 200, 1275–1284 140

Cantu, D., Segovia, V., MacLean, D., Bayles, R., Chen, X., Kamoun, S., et al. (2013). Genome analyses of the wheat yellow (stripe) rust pathogen *Puccinia striiformis* f. sp. *tritici* reveal polymorphic and haustorial expressed secreted proteins as candidate effectors. *BMC Genomics* 14, 270 139

Catanzariti, A.-M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., and Ellis, J. G. (2006). Haustorially expressed secreted proteins from flax rust are highly enriched for avirulence elicitors. *The Plant Cell* 18, 243–256 136

Chin, C.-S., Peluso, P., Sedlazeck, F. J., Nattestad, M., Concepcion, G. T., Clum, A., et al. (2016). Phased diploid genome assembly with single molecule real-time sequencing. *Nature Methods* 13, 1050 141

Cuomo, C. A., Bakkeren, G., Khalil, H. B., Panwar, V., Joly, D., Linning, R., et al. (2017). Comparative analysis highlights variable genome content of wheat rusts and divergence of the mating loci. *G3: Genes, Genomes, Genetics* 7, 361–376 139

Cuomo, C. A., Güldener, U., Xu, J.-R., Trail, F., Turgeon, B. G., Di Pietro, A., et al. (2007). The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science* 317, 1400–1402 138

de Jonge, R., Bolton, M. D., Kombrink, A., van den Berg, G. C., Yadeta, K. A., and Thomma, B. P. (2013). Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. *Genome Research* 23, 1271–1282 139

de Jonge, R., van Esse, H. P., Maruthachalam, K., Bolton, M. D., Santhanam, P., Saber, M. K., et al. (2012). Tomato immune receptor *Ve1* recognizes effector of multiple fungal pathogens uncovered by genome and RNA sequencing. *Proceedings of the National Academy of Sciences U.S.A.* 109, 5110–5115 137

Dodds, P. N., Lawrence, G. J., Catanzariti, A.-M., Ayliffe, M. A., and Ellis, J. G. (2004). The *Melampsora lini AvrL567* avirulence genes are expressed in haustoria and their products are recognized inside plant cells. *The Plant Cell* 16, 755–768 136

Dodds, P. N., Lawrence, G. J., Catanzariti, A.-M., Teh, T., Wang, C.-I., Ayliffe, M. A., et al. (2006). Direct protein interaction underlies gene-for-gene specificity and coevolution of the flax resistance genes and flax rust avirulence genes. *Proceedings of the National Academy of Sciences U.S.A.* 103, 8888–8893 136, 140

Dong, S., Raffaele, S., and Kamoun, S. (2015). The two-speed genomes of filamentous pathogens: waltz with plants. *Current Opinion in Genetics & Development* 35, 57–65 139

Duplessis, S., Cuomo, C. A., Lin, Y.-C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., et al. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proceedings of the National Academy of Sciences U.S.A.* 108, 9166–9171. doi:10.1073/pnas.1019315108 139

Edge, P., Bafna, V., and Bansal, V. (2017). Hapcut2: robust and accurate haplotype assembly for diverse sequencing technologies. *Genome Research* 27, 801–812 141

Faino, L., Seidl, M. F., Datema, E., van den Berg, G. C., Janssen, A., Wittenberg, A. H., et al. (2015). Single-molecule real-time sequencing combined with optical mapping yields completely finished fungal genome. *MBio* 6, e00936–15 139, 140

Gibriel, H. A., Thomma, B. P., and Seidl, M. F. (2016). The age of effectors: genome-based discovery and applications. *Phytopathology* 106, 1206–1212 140

Gilroy, E. M., Breen, S., Whisson, S. C., Squires, J., Hein, I., Kaczmarek, M., et al. (2011). Presence/absence, differential expression and sequence polymorphisms between *PiAVR2* and *PiAVR2*-like in *Phytophthora infestans* determine virulence on *R2* plants. *New Phytologist* 191, 763–776 141

Goodwin, S., McPherson, J. D., and McCombie, W. R. (2016). Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics* 17, 333–351 140

Huang, S., Chen, Z., Huang, G., Yu, T., Yang, P., Li, J., et al. (2012). Haplomerger: reconstructing allelic relationships for polymorphic diploid genome assemblies. *Genome Research* 22, 1581–1588 138

Jeon, J., Sook-Young, P., Chi, M.-H., Choi, J., Park, J., Rho, H.-S., et al. (2007). Genome-wide functional analysis of pathogenicity genes in the rice blast fungus. *Nature Genetics* 39, 561 136

Kämper, J., Kahmann, R., Bölker, M., Li-Jun, M., Brefort, T., Saville, B. J., et al. (2006). Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature* 444, 97 138

Klosterman, S. J., Subbarao, K. V., Kang, S., Veronese, P., Gold, S. E., Thomma, B. P., et al. (2011). Comparative genomics yields insights into niche adaptation of plant vascular wilt pathogens. *PLoS Pathogens* 7, e1002137 139

Koren, S., Walenz, B. P., Berlin, K., Miller, J. R., Bergman, N. H., and Phillippy, A. M. (2017). Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Research* 27, 722–736 141

Kuleshov, V. (2014). Probabilistic single-individual haplotyping. *Bioinformatics* 30, i379–i385 141

Laurie, J. D., Ali, S., Linning, R., Mannhaupt, G., Wong, P., Güldener, U., et al. (2012). Genome comparison of barley and maize smut fungi reveals targeted loss of RNA silencing components and species-specific presence of transposable elements. *The Plant Cell* 24, 1733–1745 137

Levy, S. E. and Myers, R. M. (2016). Advancements in next-generation sequencing. *Annual Review of Genomics and Human Genetics* 17, 95–115 140

Li, R., Zhu, H., Ruan, J., Qian, W., Fang, X., Shi, Z., et al. (2010). *De novo* assembly of human genomes with massively parallel short read sequencing. *Genome Research* 20, 265–72. doi:10.1101/gr.097261.109 138

Li, Z., Chen, Y., Mu, D., Yuan, J., Shi, Y., Zhang, H., et al. (2012). Comparison of the two major classes of assembly algorithms: overlap–layout–consensus and *de-bruijn*-graph. *Briefings in Functional Genomics* 11, 25–37 138

Link, T., Seibel, C., and Voegele, R. T. (2014). Early insights into the genome sequence of *Uromyces fabae*. *Frontiers in Plant Science* 5 139

Ma, L., Cornelissen, B. J., and Takken, F. L. (2013). A nuclear localization for *Avr2* from *Fusarium oxysporum* is required to activate the tomato resistance protein *I-2*. *Frontiers in Plant Science* 4 137

Ma, L.-J., Van Der Does, H. C., Borkovich, K. A., Coleman, J. J., Daboussi, M.-J., Di Pietro, A., et al. (2010). Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. *Nature* 464, 367 137

Mesarich, C. H., Griffiths, S. A., van der Burgt, A., Ökmen, B., Beenen, H. G., Etalo, D. W., et al. (2014). Transcriptome sequencing uncovers the *Avr5* avirulence gene of the tomato leaf mold pathogen *Cladosporium fulvum*. *Molecular Plant-Microbe Interactions* 27, 846–857 137

Michielse, C. B., van Wijk, R., Reijnen, L., Cornelissen, B. J., and Rep, M. (2009). Insight into the molecular requirements for pathogenicity of *Fusarium oxysporum* f. sp. *lycopersici* through large-scale insertional mutagenesis. *Genome Biology* 10, R4 136

Nemri, A., Saunders, D. G. O., Anderson, C., Upadhyaya, N. M., Win, J., Lawrence, G. J., et al. (2014). The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. *Frontiers in Plant Science* 5, 98. doi:10.3389/fpls.2014.00098 139, 140

Orbach, M. J., Farrall, L., Sweigard, J. A., Chumley, F. G., and Valent, B. (2000). A telomeric avirulence gene determines efficacy for the rice blast resistance gene *Pi-ta*. *The Plant Cell* 12, 2019–2032 139

Pais, M., Yoshida, K., Giannakopoulou, A., Pel, M. A., Cano, L. M., Oliva, R. F., et al. (2017). Gene expression polymorphism underpins evasion of host immunity in an asexual lineage of the Irish potato famine pathogen. *bioRxiv* , 116012doi: https://doi.org/10.1101/116012 141

Raffaele, S. and Kamoun, S. (2012). Genome evolution in filamentous plant pathogens: why bigger can be better. *Nature Reviews. Microbiology* 10, 417 139

Ramos, A. P., Tavares, S., Tavares, D., Silva, M. D. C., Loureiro, J., and Talhinhas, P. (2015). Flow cytometry reveals that the rust fungus, *Uromyces bidentis* (pucciniales), possesses the largest fungal genome reported–2489 mbp. *Molecular Plant Pathology* 16, 1006–1010 138

Rouxel, T., Grandaubert, J., Hane, J. K., Hoede, C., Van de Wouw, A. P., Couloux, A., et al. (2011). Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. *Nature Communications* 2, 202 138, 139

Saitoh, H., Fujisawa, S., Mitsuoka, C., Ito, A., Hirabuchi, A., Ikeda, K., et al. (2012). Large-scale gene disruption in *Magnaporthe oryzae* identifies mc69, a secreted protein required for infection by monocot and dicot fungal pathogens. *PLoS Pathogens* 8, e1002711 136

Schirawski, J., Mannhaupt, G., Münch, K., Brefort, T., Schipper, K., Doehlemann, G., et al. (2010). Pathogenicity determinants in smut fungi revealed by genome comparison. *Science* 330, 1546–1548 136

Schmidt, S. M., Lukasiewicz, J., Farrer, R., Dam, P., Bertoldo, C., and Rep, M. (2016). Comparative genomics of *Fusarium oxysporum* f. sp. *melonis* reveals the secreted protein recognized by the *Fom-2* resistance gene in melon. *New Phytologist* 209, 307–318 137

Simpson, J. T., Wong, K., Jackman, S. D., Schein, J. E., Jones, S. J., and Birol, I. (2009). ABySS: a parallel assembler for short read sequence data. *Genome Research* 19, 1117–1123 138

Tavares, S., Ramos, A. P., Pires, A. S., Azinheira, H. G., Caldeirinha, P., Link, T., et al. (2014). Genome size analyses of Pucciniales reveal the largest fungal genomes. *Frontiers in Plant Science* 5, 422 138

Thomma, B. P., Seidl, M. F., Shi-Kunne, X., Cook, D. E., Bolton, M. D., van Kan, J. A., et al. (2016). Mind the gap; seven reasons to close fragmented genome assemblies. *Fungal Genetics and Biology* 90, 24–30 140

Treangen, T. J. and Salzberg, S. L. (2012). Repetitive DNA and next-generation sequencing: computational challenges and solutions. *Nature Reviews. Genetics* 13, 36 139

Upadhyaya, N. M., Garnica, D. P., Karaoglu, H., Sperschneider, J., Nemri, A., Xu, B., et al. (2015). Comparative genomics of Australian isolates of the wheat stem rust pathogen *Puccinia graminis* f. sp. *tritici* reveals extensive polymorphism in candidate effector genes. *Frontiers in Plant Science* 5, 759 138, 140

Van Kan, J. A., Stassen, J. H., Mosbach, A., Van Der Lee, T. A., Faino, L., Farmer, A. D., et al. (2017). A gapless genome sequence of the fungus *Botrytis cinerea*. *Molecular Plant Pathology* 18, 75–89 140

Wu, J. Q., Sakthikumar, S., Dong, C., Zhang, P., Cuomo, C. A., and Park, R. F. (2017). Comparative genomics integrated with association analysis identifies candidate effector genes corresponding to *Lr20* in phenotype-paired *Puccinia triticina* isolates from Australia. *Frontiers in Plant Science* 8 138

Zerbino, D. R. and Birney, E. (2008). Velvet: algorithms for *de novo* short read assembly using *de Bruijn* graphs. *Genome Research* 18, 821–829 138

Zheng, W., Huang, L., Huang, J., Wang, X., Chen, X., Zhao, J., et al. (2013). High genome heterozygosity and endemic genetic recombination in the wheat stripe rust fungus. *Nature Communications* 4, 2673. doi:10.1038/ncomms3673 138, 139