School of Molecular and Life Sciences

Fighting fungal pathogens with big data: new computational approaches for effector discovery and crop disease management

Darcy Adam Bain Jones
0000-0002-6459-6259

This thesis is presented for the Degree of
Doctor of Philosophy
of
Curtin University

September 6, 2021

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made. This thesis contains no material which has been accepted for the award of any other degree or diploma in any university.

Signed: _____

Date: _____

Darcy Adam Bain Jones

Principal Supervisor Statement

It was my great pleasure to supervise Darcy Jones during his Ph. D. candidature, and I endorse his collected works be submitted for examination.

Signed: _____

Date: _____

Dr James Hane

Senior Research Fellow
Centre for Crop & Disease Management
School of Molecular & Life Sciences
Curtin University

Abstract

Fungal plant-pathogens are a major contributor to crop yield losses, and understanding their modes of population dynamics, adaptation, and molecular determinants of pathogenicity is critical to effective agricultural disease management. For example, knowledge of effector proteins—being critical molecular determinants of pathogen host-specificity—and their cognate host resistance or susceptibility genes has been a central focus in crop disease resistance breeding, and has led to tangible economic benefits in several crops. With the continually increasing accessibility of whole genome sequencing, sequencing entire populations of a species has become feasible. The potential insights gained from such a rich dataset may have a broad impact on the field, including information of effector gene content, genome diversification, and patterns of population genetic structure or adaptation. However, there are several technical barriers to extracting the full extent of information available in such a resource. For example, effector genes and proteins remain difficult and time-consuming to identify and the necessary combinations of such methods are unstandardised and difficult to apply to a large number of genomes. Similarly, consistently predicting genes and identifying hyper-variable genomic regions in a large number of genomes is rarely explored in large fungal pathogen genomics studies. In this thesis we bridge some of these technical barriers, developing methods to predict and describe effector proteins, fungal lifestyle characteristics, population consistent gene and transposable element sets, and genome variability. The thesis is organised under three main themes which are discussed below.

To address the need for tools to describe pathogen-host interactions we developed a suite of tools to easily identify high-quality effector candidates and general pathogen characteristics, under the first theme "host-pathogen interactions". In chapter 3 we developed an effector protein prediction pipeline–Predector–which combines numerous existing analyses with newly developed effector prediction resources and tools to generate efficient and standardised ranked effector candidate lists. In chapter 4 we investigated the emerging poorly conserved "families" of effector proteins identified using tertiary structural comparisons and explored the potential of highly sensitive sequence comparison and clustering techniques to identify novel families, which resulted in the identification of a new family containing 5 known effector proteins. In chapter 5 we explored the concept of pathogen "trophy" (i.e. saprotrophy, biotrophy, hemibiotrophy, or necrotrophy), and proposed a new empirical lifestyle categorisation scheme and pipeline—CATAStrophy—based on carbohydrate active enzyme frequencies.

Spatial patterns of fungal genetic and community composition are investigated in the second theme, "spatial survey and pattern detection", in which we consider Western Australian (WA) populations of agriculturally relevant fungal species, with a particular focus on the wheat necrotrophic pathogen *Parastagonospora nodorum*. In chapter 7 we investigated the spatial trends of fungal taxonomic diversity of agricultural weed leaf surfaces—which may serve as alternative hosts for crop pathogens—using targeted metagenomics. We found that latitude (or the associated climatic changes) rather than host weed species differentiates the composition of pathogenic species or fungal lifestyle in Western Australian agricultural weed phyllospheres. In chapters 8 and 11 we investigated the spatial, temporal, and potential adaptive changes in the WA population of *P. nodorum*, using both simple sequence repeat (SSR) genetic markers (chapter 8) and whole genome sequencing (chap-

ter 11). We observed high levels of genetic diversity and evidence of clonally expanded clusters of isolates, which may represent local expansions of well adapted individuals. In chapter 8 we also observed a potential shift in the *P. nodorum* population, which coincided with the adoption of a partially *P. nodorum* resistant wheat cultivar; but failed to reproduce this result in chapter 11.

We also performed an in-depth pan-genomic analysis of the WA *P. nodorum* population in the third theme "Deep population analyses". In chapter 11 we assembled, comprehensively annotated, and compared the genomes of WA *P. nodorum* isolates and applied the effector prediction solutions developed in theme 1 to these genomes. In doing this, we have also developed several methods to enhance consistency of gene and transposable element annotation across multiple genomes, and to characterise genome diversity and compartmentalisation. We observed a large number of gene presence-absence variations (PAVs) across the population, some of which correspond to losses or gains of large genomic regions. We also observed evidence of recent repeat-induce point mutation (RIP) activity, and compartmentalisation of mutations in sub-telomeric regions of the genome. Finally, numerous effector candidates were identified from combined analysis of PAV, population cluster specificity, positive selection, effector homology searches, and output from the Predector pipeline presented in chapter 3.

In this thesis we developed state of the art computational methods to predict effectors and used those tools to contribute to the growing understanding of *P. nodorum* molecular determinants of virulence. We have also contributed the first in-depth population genetic and genomic analyses of the Western Australian *P. nodorum* population, and demonstrated the potential for population level sequencing to inform pathogen disease management. Finally, we conducted one of the first large scale pan-genomic analyses of a major wheat pathogen, and identified genomic features that may contribute to pathogen adaptive capacity. Together, this thesis has provided a valuable suite of methods and insights into plant pathogen genetics, from the molecular level to the population level.

Acknowledgements

First and foremost I would like to thank my primary supervisor James Hane, for his tireless support, patience, and guidance. I would also like to thank my co-supervisors Paula Moolhuijzen and Fran Lopez-Ruiz, for their technical and academic advice. The Centre for Crop and Disease Management (CCDM) has been a fantastic place to work, and I thank everyone for their encouragement, support, and companionship. In particular, I would like to acknowledge: my desk friends Catherine Rawlinson and Stefania Bertazzoni; my former *P. nodorum* bosses Huyen Phan and Kar-Chun Tan; the bioinformatics meeting group Johannes Debler, Chala Turo, Lina Rozano, Fredrick Mobegi, Rob Syme, and Mark Derbyshire; and the current and past CCDM directors Mark Gibberd, Richard Oliver, Karam Singh, and Joshua Milne.

Thankyou to the thesis examiners and all of the reviewers of the published material presented in this thesis. Review and examination has been invaluable, and has substantially improved the quality of my writing and analyses.

Thankyou to the Curtin Institute for Computation for providing an interim scholarship for the first year of my candidacy, and to the Australian federal government for providing an Australian Government Research Training Program (RTP) Scholarship for the remainder of my candidacy. Thanks also to the CCDM for providing additional financial support for the entirety of my Ph. D.

Much of this research was undertaken using resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia. Special thanks to the Pawsey Supercomputing Centre for providing training and technical support.

Finally, I would like to thank my family Alison Bain, Graeme Woods, Eliza Bain, Stephanie Rogers, Bert Kun, and Freja Rogers-Kun for their support and love.

Copyright statement

I acknowledge and pay respect to the past, present and future traditional custodians and elders of the Nation's First People and the continuation of their cultural, spiritual and educational practices. I pay particular respect to the traditional owners of the land on which this thesis and research was conducted, the Wadjuk people of the Noongar nation.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Introduction

In 2019, nearly 10% of the global human population was exposed to severe levels of food insecurity (FAO et al., 2020), which is a major contributing factor to malnutrition and has numerous long term effects on economic stability and quality of life. Several countries already experiencing high levels of food insecurity and poor access to nutritious food are expected to contribute the bulk of future human population growth, which is projected to reach 9.7 billion people in 2050 (United Nations, 2019). Equitable access to high quality food is imperative to the future well-being of the global human population. Being the first step in the food production pipeline, the sustainable production of grains, fruits and vegetables (collectively referred to here as crops) is of critical importance toward ensuring food availability.

Numerous factors combine to reduce overall crop production from theoretically possible levels, including abiotic factors (e.g. nutrient availability, temperature), plant quality (e.g. seed germination rates, genetic factors), microbial pathogens, insect pests, weed competition, post-harvest losses, and complex interactions of each of these individual factors (Savary et al., 2017). Fungal pathogens are a major cause of pre-harvest crop losses. After weeds they are often the leading biotic cause of potential yield loss (Oerke, 2006). In staple crops they are estimated to account for up to 30% of global yield loss (Savary et al., 2019), but these levels vary highly depending on the crop and location, with higher losses often found in food insecure regions. Furthermore, the low genetic diversity of cropping plants, increasing mobility of pathogens facilitated by human movement, and changing potential niches caused by climate change means that disease emergence (i.e. significant incidences of new diseases or movement of an existing disease to a new region or host) will likely remain a constant challenge for food security (Fones et al., 2020). Understanding how to manage disease requires knowledge of how the disease infects hosts, obtains nutrients from the host, reproduces, spreads, and persists. The broad field of plant pathology deals with obtaining this knowledge, deploying management strategies (e.g. disease resistant crops), and developing the capability to rapidly respond to new challenges.

This thesis investigates how pathogens evolve in agricultural ecosystems and how molecular determinants of fungal pathogenicity can be predicted using genetics, genomics and computational methods. We develop methods to predict and describe virulence promoting "effector" proteins which are secreted from pathogens into their host. We also develop methods to support the analysis of large numbers of fungal genomes, overcoming important technical barriers to using genomics for disease monitoring and deep evolutionary analyses. Finally, we apply these methods to the whole genome sequences of a population of the important cereal pathogen *Parastagonospora nodorum*, with a view towards demonstrating the utility of genomics and computational methods for fungal disease monitoring and management. Note that this thesis and introduction focusses on fungal plant pathogens, but several other lineages of important plant pathogens exist such as bacteria (Mansfield et al., 2012), oomycetes (Kamoun et al., 2015), and viruses (Scholthof et al., 2011). Each of these groups have some distinct infection

strategies and molecular-plant interactions which are not discussed in this review.

1.2 Molecular plant-pathogen interactions

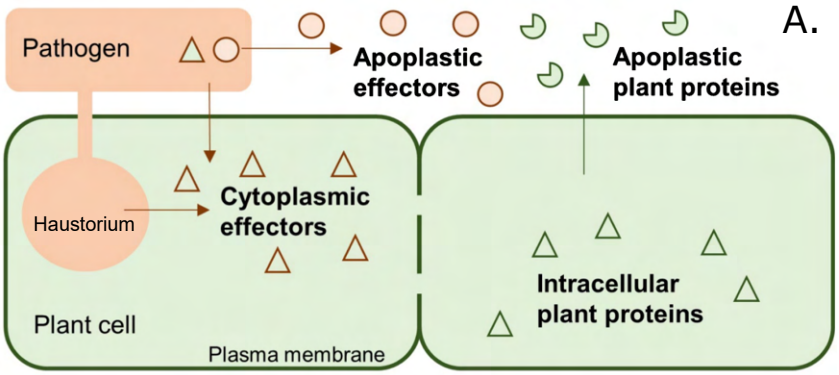
Fungal pathogens and their natural hosts have long co-evolved to form complex interactions resulting in infection or resistance to infection. Over time, many cycles of pathogens acquiring new genes or alleles to promote infection and hosts acquiring new means to resist infection, has resulted in complex and highly specific molecular interactions between the hosts and pathogens. Broadly speaking, plant defences against pathogens can be divided into two mechanisms: innate and acquired. Innate resistance confers immunity toward the majority of non-specialist pathogens that a plant might normally encounter. Structural features such as the cuticle and cell walls constitute the first barrier encountered by pathogens, which are comprised of several waxes (in the case of the cuticle) and polysaccharides (e.g. cellulose, pectin), and may be reinforced by lignin, callose, or suberin (Chassot & Métraux, 2005; Malinovsky et al., 2014). Cell walls and organelles may also harbor preformed anti-microbial and phenolic compounds which can further hinder disease establishment (Osbourn, 1996; Röpenack et al., 1998; Wittstock & Gershenzon, 2002).

To overcome these basal barriers, fungi secrete a range of enzymes which degrade the host cell walls and other structural defences (Kubicek et al., 2014). However, the degradation products of these secreted enzymes can then be recognised by the plant (termed damage associated molecular patterns (DAMPs)) to induce a second layer of basal resistance called pattern-triggered immunity (PTI). PTI is a non-specific antimicrobial response, which usually begins with a rapid accumulation of Ca^{2+} ions and reactive oxygen species (ROS) in the extracellular space (Yu et al., 2017), which in turn can signal a suite of responses including cell wall strengthening and callose deposition (Luna et al., 2010), and the localised accumulation of non-specific microbe inhibitory phytoalexins (Lin et al., 2014) and ROS (Lehmann et al., 2015). Additionally, plants constitutively express several enzymes, such as chitinases, that may degrade pathogen cell walls or secreted molecules, and the resulting pathogen or microbe associated molecular patterns (PAMPs or MAMPs) such as free chitin or flagellin can also trigger PTI (Tsuda & Katagiri, 2010).

Some pathogens specialise to overcome these basal and pattern-triggered defence responses by secreting molecules, called effectors, into the host to subvert host defence responses or otherwise facilitate their own disease establishment and progression. The activity of these effectors is usually highly specific to a particular host, and may act to prevent MAMP recognition (de Jonge et al., 2010), inhibit defensive enzymes (Rose et al., 2002; Shabab et al., 2008), alter transcription patterns (Shen et al., 2007; Weiberg et al., 2013), alter signaling pathways (Zeng et al., 2012), detoxify defensive compounds (Schäfer et al., 1989), or kill the host in a way that is compatible with the pathogens infection strategy (Z. Liu et al., 2009; Z. Liu et al., 2006; Z. Liu et al., 2012; J. P. Martinez et al., 2001). Effectors may act outside the host cell in the apoplast, or be internalised into the host cell (Figure 1.1A). Effectors acting inside the host

cell are commonly referred to as "cytoplasmic", but may be localised within other subcellular localisations such as the chloroplast (Manning et al., 2007; Petre et al., 2015) or nucleus (Qi et al., 2019; L. Zhang et al., 2017). The mechanisms of fungal effector internalisation into host cells are not currently well understood (Lo Presti & Kahmann, 2017). Some effectors are proposed to be internalised autonomously by endocytosis, requiring a surface exposed membrane interacting motif such as the RGD motif in ToxA (Manning et al., 2008) or the degenerate RxLR-like motif found in AvrL567 (Kale et al., 2010) and MiSSP7 (Plett, Kemppainen, Kale, Kohler, Legué, et al., 2011a). Other effectors depend on some trigger from the pathogen, as is the case with the *Fusarium oxysporum* effector Avr2, which is first secreted into the apoplast and then internalised by some unknown host-dependent mechanism (Di et al., 2016). Some pathogens form specialised infection structures such as intracellular hyphae and haustoria (Figure 1.1A) which manipulate the host cells and membranes to form an interfacial complex which mediates molecular interactions between the pathogen and host (O'Connell & Panstruga, 2006). These interfacial complexes are thought to be key sites of effector delivery in fungi and oomycetes, and are commonly thought to be associated with cytoplasmic effector delivery into the host either by receptor mediated endocytosis (i.e. RxLR proteins), internalisation of extracellular vesicles containing effectors, or some other unknown process (Bozkurt & Kamoun, 2020). Extracellular vesicles in particular are increasingly recognised as a major vector of molecular interactions between pathogens and hosts, both through interfacial complexes and the broader apoplast (Samuel et al., 2015; Zamith-Miranda et al., 2018). For example, the rice pathogen *Magnaporthe grisea* (syn. *Magnaporthe oryzae*) forms a specialised biotrophic interfacial complex (BIC) from intracellular hyphae, through which cytoplasmic effectors are delivered into host cells via extracellular vesicles (Giraldo et al., 2013). Filamentous pathogen effector host internalisation is complex and has likely arisen multiple times independently to support their many specialised modes of infection.

Effectors causing host cell death to promote infection, called necrotrophic effectors (NEs) or host selective toxins (HSTs), target a specific host S-protein which themselves cause the actual cell death, often by triggering a subset of pathways involved in PTI. Loss or modification of the host S-genes causing non-recognition of the NE renders the NE ineffective, and confers some degree of resistance against the pathogen. Additionally, plant hosts may acquire new receptors or R-proteins, which recognise effectors and initiate a resistance response called effector triggered immunity (ETI) (Cui et al., 2015). ETI commonly involves a highly localised form of cell death called hypersensitive response (HR), followed by accumulation of non-specific antimicrobial compounds such as phytoalexins in surrounding tissue to prevent the spread of the pathogen from the infection site (Balint-Kurti, 2019). Effectors that elicit ETI by interaction with R-proteins are called avirulence (Avr) effectors, as they confer an avirulence phenotype to the pathogen (when the host has the R-gene). Both PTI and ETI work in concert to defend plants against pathogens, and the distinction between the two processes is not always clear (Thomma et al., 2011). Note that S-genes may also refer to a broader class of host-susceptibility factors that might be required for pathogen infection, and may also be the



B.	R-	rr
Avr	✔	✘
avr	✘	✘

C.	S-	ss
NE	✘	✔
ne	✔	✔

Figure 1.1: Effectors are secreted from fungi into the host apoplast or cells to facilitate their infection (A. Reproduced and modified with permission (Appendix A) from Sperschneider, Dodds, Singh, et al. (2018)). Cytoplasmic effectors may be first secreted into the apoplast and then internalised, or delivered via specialised infection structures only present in some fungi such as haustoria. Avirulence effectors (Avrs) will usually target some host protein to suppress host defense or redirect resources. Both avirulence and necrotrophic effectors (NEs) may be recognised by host resistance (R) or susceptibility (S) proteins, which can trigger resistance responses (in the case of Avrs), or cell death (in the case of NEs). Mutation of pathgen effectors, or host R- or S-proteins may result in non-recognition of the effector, and no resistance (B) or susceptibility response (C). Green ticks in B and C indicate the interaction we want from a breeding perspective. The goal of plant breeding for resistance is to introduce functional resistance (R)-genes and remove susceptibility (S)-genes, to make undesirable interactions (red crosses) less likely.

molecular targets of Avr effector proteins (van Schie & Takken, 2014). The broadly conserved MLO gene conferring susceptibility to powdery mildew is a well known example of such a susceptibility gene (Appiano et al., 2015; Büschges et al., 1997). While these susceptibility factors are undoubtedly important, references to S-genes and proteins in this thesis concern the molecular targets of necrotrophic effectors.

Broadly speaking, there are two extremes of pathogen lifestyle: those that live and feed on living hosts (biotrophs), and those that kill the host and then feed on the dead and dying tissue (necrotrophs) (Lewis, 1973; Thrower, 1966), with the majority of pathogens falling on some continuum between the two (Newton et al., 2010). Effector function tends to split across these lines of biotrophy and necrotrophy, where effectors of necrotrophs are usually toxins (i.e. NEs or HSTs) (Z. Liu et al., 2009; Z. Liu et al., 2006; Z. Liu et al., 2012; J. P. Martinez et al., 2001) and their cognate host proteins are susceptibility factors, and effectors of biotrophs are more

often related to subverting host defence response (Stergiopoulos & de Wit, 2009) and host resistance proteins may recognise these to induce ETI (i.e. Avr effectors) (de Jonge et al., 2012; Van de Wouw et al., 2014; van den Burg et al., 2006; van Esse et al., 2007). Most necrotrophic pathogens have a short phase of biotrophy (or at least of non-necrotrophic behaviour) during and after initial penetration of the host, termed the latent phase. Pathogens with longer latent phases are called hemibiotrophs and commonly possess both Avr and necrotrophic effectors.

The model of Avr effectors and their cognate R-genes resulting in incompatibility (avirulence) upon recognition of the effector is referred to as the gene-for-gene (GFG) hypothesis (Flor, 1971), where avirulence is a dominant mendelian phenotype (Figure 1.1B). Conversely, in the case of NEs, where interaction of the effector with its cognate S-protein results in infection (rather than avirulence), interaction of an effector and S-protein is instead described by the inverse gene-for-gene (IGFG) model, where susceptibility is a dominant mendelian phenotype (Figure 1.1C) (Fenton et al., 2009). GFG and IGFG serve as foundational assumptions in much plant pathology research (Thrall et al., 2016), however, molecular interactions are often more complex in reality. Some effectors, such as *Fusarium oxysporum* Avr2 (Cao et al., 2018), require functional copies of another effector protein (SIX5 in the case of Avr2) for their phenotypic expression. In other cases phenotypic expression of an effector-receptor interaction may be masked by functional copies of other effectors such as *Parastagonospora nodorum* Tox3, which is epistatic to Tox5 (Phan et al., 2016), or *Blumeria graminis* AvrPm3, where avirulence is suppressed by SvrPm3 (Bourras et al., 2015). R-genes also show more complex interactions with pathogens than the original GFG model might suggest. For example, some effectors are recognised by multiple R-genes, as is the case for *Leptosphaeria maculans* AvrLm1, which can trigger resistance response via interaction with either LepR3 and LM1 R-proteins (Larkan et al., 2012). Other R-genes may depend on the concerted activity of other R-genes for normal function, as is the case for *Magnaporthe oryzae* effectors AVR-Pia and AVR1-CO39, where for both effectors the recognition and elicitation of resistance response requires both the RGA4 and RGA5 R-proteins (Cesari et al., 2013; Okuyama et al., 2011). Alternatively, many R-proteins display an "indirect" mode of effector recognition, where the interaction of an effector with its target is recognised by a "guard" protein, which can then initiate resistance response by interacting with a separate R-protein (Dangl & Jones, 2001). Alternatively, an effector targeted protein itself might be duplicated and modified to form a "decoy" target protein which can then interact with a separate R-protein upon recognition of an effector (van der Hoorn & Kamoun, 2008). Another direct decoy mode of recognition is also observed where some effector targetted domains or proteins themselves may become integrated directly into R-genes (Kroj et al., 2016). These complex molecular interactions reflect the long co-evolutionary histories of adapted pathogens and their hosts, and highlight some of the difficulties encountered by researchers attempting to determine the genetic identities of effectors, and R-, or S-genes.

The overall goal of plant disease resistance-breeding is to introduce genes resulting in resistance to infection, and to remove genes resulting in susceptibility, often targeting R- and S-genes involved in effector interactions. However, as the introduction of R-genes represents a

significant selective pressure on pathogen populations, fungi may lose or modify Avr effectors to avoid recognition, rendering the host resistance ineffective (Farman et al., 2002; Rouxel et al., 2003; Stergiopoulos et al., 2007). To avoid the loss of R-genes, breeders aspire toward incorporating multiple R-genes into crops (often called stacking or pyramiding) which would require pathogens to simultaneously lose multiple avirulence genes, which is far less likely (Ellis et al., 2014; Schaart et al., 2016). Engineered integrated decoy R-genes may also offer novel sources of durable resistance in the future (Kourelis et al., 2016). Although ultimately the goal of this research is to introduce host resistance or remove susceptibility, knowledge of the effectors present in a pathogen or population is critical to evaluating which host-resistances will be effective (and for how long), and can be strong selectable markers for breeding strategies, particularly when attempting to select for multiple resistance genes (S. Zhu et al., 2012). Effectors are therefore a critical aspect of host-pathogen interactions, and their interactions with cognate R- or S-genes represent the best major resistance effects as targets for breeding. Plant pathologists aspire to discover the full complement of a pathogen's effectors, and their cognate R- or S-genes. However, because effectors are a highly diverse group of molecules (usually proteins), which result from highly specific interactions with hosts, they can be difficult to discover.

1.3 Fungal effector discovery

Effectors are a diverse group of proteins which are traditionally defined in genetic rather than biochemical terms. Commonly, effector discovery efforts target a known pathogen-host phenotypic interaction that could be exploited for breeding, which are usually initially assumed to follow the pattern of a gene-for-gene (or inverse gene-for-gene) interaction (Thrall et al., 2016). A variety of experimental techniques have been employed towards their discovery, but the most common include proteomics (Z. Liu et al., 2009; Mesarich et al., 2018), forward genetics (Fudal et al., 2007), and reverse genetics (Molina & Kahmann, 2007). Finally, when a small number of candidate effectors have been discovered by association with the trait, they are screened for specific activity by isolating the responsible gene or protein and exposing the host plant to that protein to see if they cause the same effect as the trait that was originally targeted. Such effectors are said to have been "cloned".

The last 20 years have seen an explosion of bioinformatics driven research in plant pathology, as the costs of sequencing genomes and transcriptomes have declined (Kanja & Hammond-Kosack, 2020). Such techniques have greatly increased the resolution and volume of effector candidate discovery (see chapter 2; D. A. Jones et al., 2018). Although experiment driven research remains the most reliable method toward effector discovery, and bioinformatics has a clear place in that process, it can be prohibitively expensive and time consuming, so the potential to mine genomes directly for effector candidates is an attractive one. However, one of the key tools used in *in-silico* functional prediction, sequence similarity or homology, has been of limited utility as the majority of known fungal effectors have not been reported to

share sequence similarity. Furthermore, even for avirulence (Avr) effectors with conserved function and sequence (e.g. carbohydrate-active enzymes (CAZymes) (Pollet et al., 2009) and LysM (Bolton et al., 2008) effectors), sequence similarity is not necessarily indicative of the phenotype of interest (i.e. avirulence) as this is largely determined by the host resistance (R)-genes. This has resulted in an broader use of the term "effector" in recent years, where it may now refer to a more general group of proteins that promote virulence and comes into direct contact with the host (i.e. secreted into the apoplast or into host cells). Bioinformatics driven effector discovery usually adopts this more general definition when identifying effector candidates; where common sequence features such as cysteine-richness, secretion, and low molecular weight, or the genomic context are used as indicators (as discussed in chapter 2).

1.3.1 Fungal effector and virulence factor "families"

Recently, a number of protein families with at least some members having effector-like phenotypes have been described, with some families containing multiple known effectors displaying tertiary structure similarity rather than sequence similarity (de Guillen et al., 2015; Lu et al., 2015; Praz et al., 2017; S. M. Schmidt et al., 2016; Spanu, 2017). There is remarkable diversity across these families, both between families and within them, yet common themes are emerging. We summarise some effector families and their characteristics below. The other major families ToxA-like, MAX, RALPH, and AvrLm6-like are described in detail in chapter 4.

RxLR

The RXLR-like family of fungal effectors were first identified by their weak similarity to the RXLR effectors of the oomycetes. First described as RXLR-X{5,21}-ddEER in *Hyaloperonospora parasitica* (Rehmany et al., 2005), the oomycete RXLR motif tends to reside within 40AA of the N-terminal signal peptide, is highly conserved, and is required for host cell translocation via interaction of the RXLR (or a second similarly positively-charged region) with PI3P phospholipids of the host cell membrane (Dou et al., 2008; Kale et al., 2010; Whisson et al., 2007), or more rarely may be involved in secretion from the pathogen cell (Wawra et al., 2017). The oomycete RXLR effectors contain WY structural domains, an alpha-helical repeat usually repeated up to 11 times (Boutemy et al., 2011), which is proposed to be the core structural backbone upon which functional diversity may emerge through mutations in loop sequences or in WY repeat copy number (Win et al., 2012). Mutagenesis of the RxLR-dEER motif were used to define an RxLR-like motif [RHK]X[LMIFYW], which was used to identify potential RxLR-like motifs in fungal proteins (Kale et al., 2010). Although this relaxed motif pattern may match hundreds of non-effector proteins in most fungal species (D. A. Jones et al., 2018), host cell uptake has been experimentally demonstrated for several fungal effectors with RXLR-like motifs, including: *Melampsoria lini* AvrL567 and AvrM, *Fusarium oxysporum* f. sp *lycopersicon* Avr2, *Leptosphaeria maculans* AvrLm6, *Magnaporthe oryzae* AVR-Pita (Kale et al., 2010). Notably, these confirmed fungal RXLR-like motifs do not exhibit the conserved sequence homology

that is typical of oomycete RXLRs. However, the findings presented by Kale et al. (2010) remain controversial, as several subsequent studies have failed to reproduce their results (Petre & Kamoun, 2014).

Crinkler (CRN)

Like RXLR effectors, CRN effectors are well conserved and best studied in the oomycota (L. Liu et al., 2019; Schornack et al., 2010; Stam et al., 2013; D. Zhang et al., 2016), primarily in *Phytophthora* spp., *Pythium* spp. and *Plasmopara* spp.. They appear to be broadly conserved across kingdoms, with reports of CRN proteins in multiple plant species including *Arabidopsis thaliana*, *Vitis vinifera*, *Solanaceum lycopersicum* and *Theobroma cacao* (D. Zhang et al., 2016). In the fungi, CRN-like effectors have been reported in mycorrhizal and endophytic fungi including *Piriformospora indica* (Zuccaro et al., 2011), *Rhizophagus irregularis* (Voß et al., 2018), and *Laccaria bicolor* (Plett, Kemppainen, Kale, Kohler, Legué, et al., 2011b). These include RiCRN1 of *R. irregularis* which appears to be important for symbiotic establishment but does not induce necrosis (Voß et al., 2018) and the *Laccaria bicolor* protein MiSSP7 which is imported into the host nucleus and represses host defences (Plett et al., 2014; Plett, Kemppainen, Kale, Kohler, Legué, et al., 2011b).

NLP/NEP

The necrosis and ethylene production (NEP)-like proteins (NLP) are a large evolutionarily conserved family found in Fungi, Bacteria, and Oomycetes (Fellbrich et al., 2002). They were first discovered in *Fusarium oxysporum* causing necrosis and ethylene production (NEP) in *Erythroxylum coca* (Bailey, 1995). All proteins in the NLP family have a conserved central motif 'GHRHDWE' (Fellbrich et al., 2002; Gijzen & Nürnberger, 2006) (PFAM: PF05630), and are found in 3 main lineages (Gijzen & Nürnberger, 2006): type 1) with 2 conserved cysteines, found in Fungi, Oomycetes, and Bacteria, type 2) with 4 conserved cysteines, found in Fungi, and Bacteria, and type 3) with 6 conserved cysteines and have the motif but the N and C terminals do not display obvious similarity with types 1 or 2 (Oome & Van den Ackerveken, 2014). NLPs appear to be active only with eudicot hosts, acting as an elicitor of host resistance or causing host cell death from the apoplast (Qutob et al., 2006).

LysM

The Lys-M domain [PFAM: PF01476] is an evolutionarily conserved chitin binding domain broadly conserved across all kingdoms (Buist et al., 2008). Lys-M effectors do not directly cause disease symptoms, but instead prevent host-mediated hydrolysis of pattern-triggered immunity (PTI) by masking chitin pathogen associated molecular patterns (PAMPs) of the pathogen cell wall or sequestering free chitin (e.g. damage associated molecular patterns (DAMPs)). The exemplar for this class is Ecp6 of *Cladosporium fulvum* (syn. *Passalora fulva*) (Bolton et al., 2008; de Jonge et al., 2010). LysM domains are usually present in multiple copies

and can be associated with other functional domains (Marshall et al., 2011; D. A. Martinez et al., 2012). The LysM effectors can be overlooked during effector discovery efforts, but are vital to biotrophic and symbiotic interactions involving evasion of host defences (Romero-Contreras et al., 2019).

Protease inhibitors

Plants secrete proteases into the extracellular space, which can generate PAMPs and trigger PTI. Fungi and Oomycetes may secrete proteins that inhibit these proteases, thereby preventing recognition. Two main groups of protease inhibitors have been described, the Cystatin-like, and Kazal-type. The *Phytophthora infestans* effectors EPIC1 and EPIC2b (Tian et al., 2007), and *Passalora fulva* Avr2 (Esse et al., 2008; Rooney et al., 2005) are cystatin-like protease inhibitors. *P. fulva* Avr2 inhibits the tomato extracellular cysteine proteases Rcr3 and Pip1, which prevents Cf-2 mediated HR induction (Esse et al., 2008; Rooney et al., 2005). *P. infestans* EPIC1 and EPIC2b also target the Rcr3 protein in Tomato, but do not trigger Cf-2 HR (Song et al., 2009). The effectors EPI1 and EPI10 of *Phytophthora infestans* are both Kazal-type protease inhibitors (Tian et al., 2005; Tian et al., 2004), which inhibit extracellular serine proteases in tomato hosts. Similarly, the *P. palmivora* kazal-type effector PpEPI10 inhibits an extracellular serine protease in rubber trees (Ekchaweng et al., 2017).

Hce2

The Hce2 (Homologues of *Cladosporium* ECP2) effector family [Pfam: PF14856] is a relatively well conserved group that was originally identified via homology to the Ecp2 effector of *Cladosporium fulvum* (Stergiopoulos et al., 2012). It is observed broadly in many classes of the Pezizomycotina (e.g. Sordariomycetes, Dothideomycetes, Eurotiomycetes and Leotiomycetes) and in the Agaricomycotina. Hce2 loci are reported to undergo high rates of duplication, loss, and diversification, suggesting roles in adaptation to new niches or host co-evolution. Hce2 effectors are categorised into 3 classes. Class I is the most common, being of short length (80-400 aa) and containing an N-terminal signal peptide (SP) domain and the Ecp2 domain with conserved 4 cysteine positions. Class II is longer (400-800 aa) and contains the same domains as Class I, with an additional N-terminal region after the SP domain with conserved positioning of 7 cys. Class III is the longest (>1000 aa) and appears to be a membrane-anchored or chitin-bound version of class I, with the highly conserved Ecp2 domain and the addition of a large modular region between the SP and Ecp2 domains of 3 chitin binding domains (2 Lysm, 1 ChtBD1) followed by glycoside hydrolase family 18 (GH18, chitinase) domain. Class III Hce2 proteins are structurally similar to the alpha subunit of zymocin killer toxin produced by the dairy yeast *Kluyveromyces lactis* (Stergiopoulos et al., 2012), which is an exochitinase and does not confer the cytotoxic activity of zymocin.

Five class I Hce2 effectors (VmHEP1 - VmHEP5) were recently reported to fully account for virulence of *Valsa mali* on apple (valsa canker) (M. Zhang et al., 2019). *VmHEP1* and *VmHEP2*

are tandemly duplicated paralogous loci which together accounted for the majority of the virulence symptoms, and were both upregulated in early infection. There have been multiple recent reports of class I Hce2 effector candidates across multiple species, including: *P. nodorum* (Syme et al., 2016), *F. graminearum* (Lu & Edwards, 2016), *Verticillium nonalfalfae* (Marton et al., 2018), and *Cochliobolus miyabeanus* (Castell-Miller et al., 2016).

Cerato-platanins

Cerato-platanins are a conserved group of secreted proteins unique to fungi, and are present in many Ascomycetes and Basidiomycetes (H. Chen et al., 2013). They were originally reported as structural homologs of Hydrophobin proteins (Pazzagli et al., 2006), however, subsequent structural determination indicated that cerato-platanins are more-similar to the expansin protein group, possessing a single fold domain forming a double $\Psi\beta$-barrel [Pfam: PF07249] (de Oliveira et al., 2011). Despite the common fold, there appear to be diverse functions associated with cerato-platanins paralogues, including: hyphal development (Baccelli et al., 2012), aggregation and amyloid-like structure formation (H. Chen et al., 2013; de O. Barsottini et al., 2013), cytotoxicity (H. Chen et al., 2015; Frías et al., 2011; Jeong et al., 2007; Yang et al., 2018), host defence priming (Gaderer et al., 2015; Gomes et al., 2015), expansin-like activity (Baccelli, Luti, et al., 2014; Luti et al., 2017), and host resistance elicitation (Ashwin et al., 2017; Baccelli, Lombardi, et al., 2014; Luti et al., 2016; Luti et al., 2017). The necrosis inducing activity of some cerato-platanins may be associated with two surface peptides, which retain activity independently of the whole protein and each other (Frías et al., 2014). A specific aspartic acid residue may also be required for the expansin-like and PAMP activities of some cerato-platanins (Luti et al., 2017).

Knottins

Knottins are a family of proteins that contain the inhibitor cysteine knot (ICK) motif, which forms a ring and knot structure with 6 cysteines forming 3 disulphide bonds (Postic et al., 2018). Prominent examples of non-fungal cytotoxins belong to this class, including scorpion and spider venoms (Fry et al., 2009) and antifungal plant defensins (Tam et al., 2015). The knottins share this structural feature but appear to have little to no sequence homology or common modes of action. In the fungi, the best example of this family is Avr9 of *C. fulvum*, which requires all 6 of its cysteines to retain its structure and necrosis phenotype (Pallaghy et al., 1994). Other knottin effectors that have been reported in Fungi so far include MLP124266 and MLP124017 of *Melampsora larici-populina* and AvrP4 of *M. lini* (Catanzariti et al., 2006). Extensive databases of knottins have been compiled which extend to several fungal species (Gelly et al., 2004), however their presence in plant pathogens as putative effectors has not yet been widely explored beyond these examples.

AvrLm4-7

AvrLm4-7 is a well characterised Avr effector of *L. maculans* (Blondeau et al., 2015), which has an avirulent phenotype in hosts with the *Rlm4* or *Rlm7* loci (Parlange et al., 2009). Low sequence identity homologs of AvrLm4-7 were found in *Pyrenophora tritici-repentis*, *Pyrenophora teres*, and *Macrophomina phaseolina*, with a well conserved (R/N)(Y/F)(R/S)E(F/W) motif flanking a β-strand and loop region, several conserved cysteine, glycine, arginine, and glutamic acid residues (Blondeau et al., 2015).

Ave1

Ave1 is an avirulence effector and pathogenicity factor of *Verticillium dahliae*, and governs avirulence for tomato hosts possessing the Ve1 resistance receptor (de Jonge et al., 2012). Ave1 displays some sequence similarity to a large group of plant-natriuretic peptides, with numerous homologs in plants, the bacterium *Xanthomonas axonopodis* pv. *citri*, and other Pezizomycetes including: *Fusarium oxysporum* f. sp. *lycopersici*, *Colletotrichum higginsianum*, and *Cercospora beticola* (de Jonge et al., 2012). The Ave1-like family appears to have undergone a large clonal expansion in *Venturia spp.* (Deng et al., 2017).

CAP domain proteins

The Cysteine-rich secretory proteins, Antigen 5, and Pathogenesis-related 1 (CAP) domain [Pfam: PF00188] is broadly conserved across all kingdoms, and its best studied representatives are mammalian proteins with a diverse range of roles including protease inhibition, ion chelation and ion channel regulation, cell adhesion, morphogenesis and oncogenesis (Gibbs et al., 2008). The structure of the CAP domain is well conserved and appears to confer calcium chelating serine protease activity (Milne et al., 2003). There are numerous reports of fungal CAP domain proteins among the predicted effectors of various fungal pathogens, including: *Fusarium oxysporum* (Prados-Rosales et al., 2012), *Melampsora larici-populina*, *Puccinia graminis* f. sp. *tritici* (Saunders et al., 2012), *Passalora fulva* (Mesarich et al., 2018), and *Microbotryum lychnidis-dioicae* (Perlin et al., 2015).

CFEM proteins

CFEM proteins contain a 'common in fungal extracellular matrix' (CFEM) domain [Pfam: PF05730] with 8 conserved cysteines, that is unique to but widely conserved across the Fungi (Kulkarni et al., 2003; Z.-N. Zhang et al., 2015) and is involved in haem-iron uptake and virulence (Nasser et al., 2016). There are numerous examples of this family with confirmed roles in host-pathogen interactions. The CFEM protein Pth11 of *M. oryzae* is a membrane-associated G-protein-coupled receptor required for appressorium formation and virulence (Di et al., 2017). Similarly, the glycosylphosphatidylinositol (GPI)-anchored BcCFEM1 of *B. cinerea* is involved in pathogenicity, conidiation and stress tolerance (W. Zhu et al., 2017). The

expanded family of CFEM proteins of *C. graminicola* group into 2 major clades, based on the presence or absence of TM domains, but also exhibited a diverse range of host sub-cellular localisations, 3D structure and disease phenotypes (Gong et al., 2020).

Finding patterns of functional commonality and detection of sequence or structural homology is becoming more robust as more effectors are discovered and more pathogen genomes are sequenced. The goal of identifying all of a pathosystems' effector and R- or susceptibility (S)-gene partners is more feasible than ever. However, the functional and sequence diversity of the currently known effectors highlights how much we don't yet know, and how imprecise the language in molecular pathology can be. The long held genetic definition of effectors lacks semantic clarity when applied to *in-silico* or non-targeted molecular "effector" discovery. Researchers sometimes use terms like "effector candidate" or "virulence factor" to skirt this issue, but the commonly and inconsistently used umbrella term "effector" is often confusing to researchers outside pathology. The field may benefit from further effort describing effector "families" (i.e those with sequence or structural homology as described here), or using qualifying functional groupings (e.g. CAZyme, Protease, or PAMP masking) to lower the barrier of understanding.

1.4 Pathogen genomics

Genomes are a foundational resource in modern pathology, serving as a common reference for many comparative and population based studies such as genome-wide association studies (GWAS) (Hartmann et al., 2017; Zhong et al., 2017) or evolutionary analyses (de Vries et al., 2020; Plissonneau et al., 2017), and for high throughput expression studies such as proteomics (Karimi Jashni et al., 2020; Mesarich et al., 2018) and transcriptomics (D. A. B. Jones et al., 2019). Bioinformatics and the broader *-omics (i.e. genomic, transcriptomics, proteomics, epigenomics) disciplines have used these increasingly comprehensive resources to enhance our understanding of how specific (including non-model) pathogens cause infection, evolve, and respond to numerous stimuli. However, fungi possess many novel genomic features and patterns of mutation that pose unique challenges to researchers working in pathogen bioinformatics (Priest et al., 2020), such as repeat-induced point mutation (RIP) (Gladyshev & Kleckner, 2017b), accessory genomic regions (Bertazzoni et al., 2018), mesosynteny (Hane et al., 2011), lateral gene transfer (Fitzpatrick, 2012), multiple nuclei (heterokaryosis) (Saupe, 2000; Strom & Bushley, 2016), and parasexuality (Drenth et al., 2019; Noguchi et al., 2006; Sherwood & Bennett, 2009).

Like many organisms, fungi are often said to possess compartmentalised or multi-speed genomes, where essential genes tend to be conserved in gene-rich, mutation-poor regions of the genome, and non-essential genes may be in gene-poor regions enriched in repeats, transposable elements (TEs), mutations, and the weaker binding A/T nucleotides (Bertazzoni et al., 2018; Testa et al., 2016). In the case of pathogen genomes, it is often suggested that

genes contained in genomic regions that are enriched in mutations or features that promote mutation (such as TEs) might contribute to virulence or adaptive function as they can mutate more rapidly (Rouxel et al., 2011; Williams et al., 2016), and haplotypes can become more or less common by selection and drift events. The genomic regions near the telomeres (subtelomeres) are often highly diverse, exhibiting frequent rearrangements, mutations, and a high repeat content (Hocher & Taddei, 2020). Subtelomeres often harbor genes with potentially adaptive functions, such as effector candidates and virulence factors (Wyatt et al., 2020), niche specialisation genes (van Wyk et al., 2018), and secondary metabolite gene clusters (Graham-Taylor et al., 2020). Similarly, fungal centromeres also exhibit high levels of diversity and rearrangement, but are not commonly thought of as sites of adaptive potential. Their structure varies considerably between species (Smith et al., 2012; Smith et al., 2011; Yadav et al., 2018), but some species have genes relatively close to centromeric regions, which can resemble the AT-rich regions commonly found in fungal pathogens (Schotanus et al., 2015; Yadav et al., 2019).

1.4.1 RIP

Repeat-induced point mutation (RIP) is a mutagenesis mechanism that selectively mutates a specific nucleotide pattern at and around repeated regions, and is generally thought to have evolved as a way of deactivating transposable elements. RIP was first described in *Neurospora crassa* (Cambareri et al., 1989; Selker, 1990), where RIP results in the mutation of CpA dinucleotides to TpA in genome duplications, and "leaking" into regions surrounding the duplications. In *N. crassa*, RIP occurs during pre-meiosis (Wang et al., 2020), but few other organisms have been studied so extensively and the process has also been observed in typically asexual fungi (Braumann et al., 2008). Recognition of duplications appears to be recombination independent (Gladyshev & Kleckner, 2017b), and detectable levels of RIP activity can be induced by homologous stretches as small as 150 bp, with relatively different sequences so long as they have identical stretches of at least three bp in an 11 or 12 bp periodicity (Gladyshev & Kleckner, 2014). Additionally, the presence of a homologous trinucleotide (GpApC) in the duplicated region has been observed to promote RIP activity more than other trinucleotides (Gladyshev & Kleckner, 2016). Two methyltransferases, 'RIP defective' (RID) (Freitag et al., 2002; Malagnac et al., 1997) and DIM-2 (Gladyshev & Kleckner, 2017a), are known to methylate cytosines during RIP following homology recognition, and the methylated cytosines are then thought to separately deaminate to thymine either spontaneously or by the activity of another unknown enzyme. Interestingly, DIM-2 mediated RIP requires the action of DIM-5, HP1 and other heterochromatin factors (Gladyshev & Kleckner, 2017a), suggesting a connection between RIP and heterochromatin formation.

RIP is only observed in some fungi and is most commonly described in the Dothideomycetes class (Galagan & Selker, 2004; Hane et al., 2015; John Clutterbuck, 2011; Testa et al., 2016), where the prevalence of canonical CpA $\rightarrow$ TpA conversions results in large, distinct, gene-poor regions of the genome with high AT-content. However, RIP and RIP-like mutations of differ-

ent dinucleotides and tri-nucleotides are observed in several distinct taxon, including some Basidiomycetes where the TpCpG trinucleotide motif may serve as a site for similar RIP-like mutations (Hood et al., 2005; Horns et al., 2012). The presence of RIP-like activity across both Ascomycetes and Basidiomycetes suggests an ancient common origin, and the fungal lineages which have retained RIP (such as the Pezizomycotina and Pucciniales) contain the majority of important plant pathogenic species (Hane et al., 2015).

1.4.2 Accessory genomes

Individual fungi within a species may possess additional 'dispensable' chromosomes or large regions of the genome that are not required for basal function, but may confer some adaptive advantage in certain environments (Covert, 1998; Zolan, 1995). Discovery of these regions, now commonly referred to as accessory chromosomes (ACs) and accessory regions (ARs) in line with bacterial pangenomics literature (Rouli et al., 2015), has been a key topic in fungal pathogen genomics over the last decade as the community has moved towards sequencing multiple individuals of a species (Bertazzoni et al., 2018). Numerous fungal ACs and ARs have been described, some of which confer virulence (or avirulence) phenotypes against different host plants (Bertazzoni et al., 2018). Perhaps the most notable examples are the ACs of *Fusarium oxysporum formae speciales*, which carry several 'secreted-in-xylem' (SIX) proteins and the presence or absence of an AC largely determines an individual's host specificity. Other fungal pathogens possessing ACs and ARs include *Fusarium solani* (teleomorph *Nectria haematococca*) (Coleman et al., 2009; Miao et al., 1991), *Zymoseptoria tritici* (Goodwin et al., 2011), *Alternaria spp.* (Akamatsu et al., 1999), *Colletotrichum spp.* (He et al., 1998; O'Connell et al., 2012), *Leptosphaeria maculans* (Balesdent et al., 2013), *Magnaporthe oryzae* (syn. *Magnaporthe grisea*) (Talbot et al., 1993), *Verticillium spp.* (de Jonge et al., 2013; Klosterman et al., 2011; Pantou & Typas, 2005), and *Parastagonospora nodorum* (Neil Cooley & Caten, 1991; Richards et al., 2019).

ACs and ARs likely originate from erroneous recombination events, but the specific mechanism has not yet been determined (Bertazzoni et al., 2018). One prevailing hypothesis is that breakage-fusion-bridge (BFB) cycles might initiate ACs and ARs as well as rapidly rearranging genomes (Croll et al., 2013). BFB cycles are initiated when a chromosome loses a telomere and enters mitosis, where deoxyribonucleic acid (DNA) repair mechanisms mistakenly fuse the homologous chromosome ends resulting in a single chromosome with two centromeres (McClintock, 1941). During cell division the two centromeres are pulled apart and the fused chromatids are broken at a weak point in the DNA, resulting in each child cell having a chromosome without a telomere, continuing the cycle. BFB-cycles are an active area of research and the frequency of formation in nature and different species is unknown.

1.4.3 Chromosomal rearrangement

The genomes of closely related species often share similar genetic content. Synteny describes the conservation of genetic loci on homologous chromosomes of two genomes (R. Schmidt,

2000). In the genomics age, we are often more concerned with collinearity, where the order and orientation of genetic loci are conserved, either at small scales (<10 loci; microsynteny) or at near chromosome level scales (macrosynteny). Over time, random chromosomal inversions, translocations, and gene loss or gain results in the loss of synteny, and increasing species chromosomal divergence. Genomes of the fungal Dothideomycete class show a peculiar syntenic pattern, called mesosynteny, where the overall gene content is preserved within syntenic chromosomes (or large parts of chromosomes) of closely related species, but the order of genes is completely rearranged (i.e. they are not collinear) (Hane et al., 2011). This is in contrast to the usually observed pattern, where homologous gene content is increasingly randomly assorted to different chromosomes (loss of synteny) as genomes lose micro- and macro- collinearity. This pattern of mesosynteny can be explained by a series of successive chromosomal inversions, with few or no translocation events (i.e. movement of genetic material between chromosomes) (Hane et al., 2011; Ohm et al., 2012). However, the mechanisms causing this series of inversions is not yet known. As in the case of ARs and ACs, BFB cycles have been proposed as a possible mechanism leading to this pattern (Moolhuijzen et al., 2018), but this is yet to be demonstrated experimentally. The restriction of this phenomenon to the dothideomycetes is also currently unexplained; however, the recent discovery that RIP-like homology recognition might facilitate heterochromatin formation (Gladyshev & Kleckner, 2017a), and that functional neocentromeres can form spontaneously at heterochromatic regions in fungi (Burrack et al., 2016), suggests a possible way that these fungi might enter BFB-cycles more frequently.

Overall, the propensity for rapid accumulation of mutations in fungal pathogens is indicative of the importance of adaptive potential in pathogenic lifestyle. Fungicides and host disease resistance exert strong selective pressures on populations of organisms, and maintaining extant diversity or the ability to rapidly acquire new alleles enhances a species' capacity to survive in agricultural ecosystems. For example, the activity of RIP or loss of ACs or ARs may result in the non-recognition of an avirulence (Avr) effector (either by loss or modification) by an introduced host resistance (R)-gene, leading to a pathogen regaining virulence on a host. Similarly, the common occurrence of mesosynteny between relatively closely related pathogenic species suggests that processes such as BFBs may play an important role in speciation events and adaptation to new hosts. The tendency of pathogens to maintain compartmentalised gene-rich core regions with few mutations and hypervariable regions reflects the competing needs for diversification and the maintenance of core functions (e.g. metabolism, cell division etc). Together, the numerous mutagenesis mechanisms within fungi and fungal pathogens appear to have converged on common patterns of genomic architecture that favour a balance between basal function and adaptability, which is ideal for the pathogenic lifestyle.

1.5 Population and pan-genomics

The cost of sequencing whole genomes has been decreasing faster than Moore's law since 2007 ("The Cost of Sequencing a Human Genome", 2020), which has enabled numerous non-model organisms to be sequenced. By late 2016, the genomes of more than 1090 fungal species were publically available, including many plant pathogens (Aylward et al., 2017), but relatively few species had more than 10 genomes available in the GOLD database, including only two plant pathogens: *Fusarium oxysporum* (17) (Ma et al., 2010; S. M. Schmidt et al., 2016) and *Magnaporthe oryzae* (48) (C. Chen et al., 2013; Chiapello et al., 2015; Dean et al., 2005; Dong et al., 2015; Wu et al., 2015; Xue et al., 2012; Yoshida et al., 2009). The past five years have seen a transition in fungal plant pathology from sequencing one or few reference isolates, to sequencing entire populations of a species and performing large comparative studies including in *Parastagonospora nodorum* (Richards et al., 2019; Syme et al., 2018), *Zymoseptoria tritici* (Badet et al., 2020; Hartmann et al., 2017; Plissonneau et al., 2018), *Venturia spp*. (Le Cam et al., 2019), *Puccinia hordei* (J. Chen et al., 2019), *Pyrenophora teres f. sp. teres* (Wyatt et al., 2020), *Fusarium graminearum* (Kelly & Ward, 2018; Talas & McDonald, 2015; Walkowiak et al., 2016), *Microbotryum spp.* (Badouin et al., 2017), *Verticillium dahliae* (Milgroom et al., 2016), *Blumeria graminis* (Menardo et al., 2016), *Heterobasidion annosum* (Dalman et al., 2013), and *Magnaporthe oryzae* (Islam et al., 2016).

The increased use of genotyping by sequencing (GBS) for short genetic polymorphism discovery has enabled researchers to rapidly conduct population genomics, phylogenetic and epidemiological studies (Grünwald et al., 2016). For example, Holt et al. (2013) used genomics and phylogeography to track the origins of a drug resistant population of *Shigella sonnei*, and found that convergent horizontal transfer events and multiple selective sweeps had resulted in the dominance of the resistant strain. Similarly, in 2016 following an emergent epidemic of wheat blast in Bangladesh, Islam et al. (2016) sequenced 20 strains of the *Magnaporthe oryzae* species complex and determined that the emergent pathogen likely originated from wheat infecting *M. oryzae* isolates in South America. Population genetics and phylogenetics can also enable disease monitoring efforts, which has clear applications in crop cultivar or fungicide selection. For example, Hubbard et al. (2015) found that the UK population of the wheat pathogen *Puccinia striiformis* f. sp. *tritici* had rapidly shifted from a genetically diverse population to having a dominant new virulent "warrior" genotype, and suggested that such surveillance programs could inform regional management decisions.

1.6 *Parastagonospora nodorum*

A significant focus of this thesis is to investigate the genomic features influencing pathogenicity, and to develop bioinformatic tools that facilitate these investigations, including identifying effector proteins. In order to demonstrate the use and effectiveness of these methods, the model cereal necrotroph *Parastagonospora nodorum* (teleomorph: *Phaeosphaeria nodorum*) was

selected for detailed genetic analysis and the methods developed as part of this thesis were applied to it.

P. nodorum is a pathogen of wheat, causing necrotic patches on leaves and glumes preceded by a chlorotic front (Solomon et al., 2006). It is the third most economically devastating pathogen of wheat in Australia, causing an estimated average annual loss of $108 M (Murray & Brennan, 2009). Taxonomically, *P. nodorum* is placed alongside many other plant pathogens in the Dothideomycetes class (Quaedvlieg et al., 2013). The primary inocula are the sexual ascospores, which typically establish initial infection, followed by multiple cycles of water splash mediated reinfection by asexual pycnidiospores (Solomon et al., 2006). The fungus survives on stubble as pseudothecium or pycnidia (producing ascospores and pycnidiospores, respectively), or on infected seeds (Solomon et al., 2006; Sommerhalder et al., 2011). Long range dispersal of *P. nodorum* is by wind dispersal of ascospores, and transport of infected seed. Several studies have indicated that *P. nodorum* is highly genetically diverse in the wild, and does regularly undergo sexual reproduction (Caten & Newton, 2000; Keller et al., 1997; Murphy et al., 2000; Sommerhalder et al., 2006; Stukenbrock et al., 2006). However, small local populations of clones have also been observed (Caten & Newton, 2000), and the capacity for asexual spread combined with a highly diverse background population poses a high risk for disease management (B. A. McDonald & Linde, 2002). A multi-locus sequence phylogeny found that *P. nodorum* likely originated alongside wheat in the fertile crescent, near Iran (M. C. McDonald et al., 2012).

Currently there are eight described necrotrophic effectors and nine known susceptibility genes in the *P. nodorum*-wheat pathosystem (Table 1.1); however, recent analysis of a diverse wheat panel (Phan et al., 2018) identified several other quantitative trait loci (QTL) which might form previously undescribed combinations. Of the known inverse gene-for-gene (IGFG) interactions, three necrotrophic effectors (NEs) have been genetically determined: ToxA (Ciuffetti et al., 1997; Friesen et al., 2006), Tox1 (Z. Liu et al., 2012), and Tox3 (Z. Liu et al., 2009). *Tox1* encodes a small (117 AA), cysteine rich (16 Cysteines) protein with a signal peptide, which exhibits light-dependent necrosis on hosts possessing the *Snn1* gene (Z. Liu et al., 2012). It was identified using an effector candidate ranking approach, where molecular weight, number of cysteines, and appropriate expression profiles were combined using a manual linear equation. *Tox3* also encodes a small (230 AA) cysteine rich (three predicted disulphide bonds) protein with a signal peptide, which causes necrosis in wheat possessing the *Snn3* susceptibility (S)-gene (Z. Liu et al., 2009). ToxA was originally discovered in another wheat pathogen *Pyrenophora tritici-repentis* (Ciuffetti et al., 1997), and a nearly identical copy was subsequently discovered in *P. nodorum* when the genome was sequenced (Friesen et al., 2006). ToxA is a small (178 AA) protein containing two cysteines which causes light dependent necrosis in wheat possessing the *Tsn1* S-gene. Friesen et al. (2006) proposed that *ToxA*, along with an 11 kb surrounding genomic region, was horizontally acquired by *P. tritici-repentis* from *P. nodorum*. However, the recent discovery of *ToxA* and the surrounding genomic region in another wheat pathogen, *Bipolaris sorokiniana*, suggested a more complex horizontal transfer event (M. C.

McDonald et al., 2018). M. C. McDonald et al. (2019) performed long-read sequencing of of several isolates, and after discovering an extended region (~80 kb) of genomic similarity surrounding *ToxA* between *P. nodorum* and *P. tritici-repentis*, proposed two competing models of transfer: 1) that *ToxA* originated in *P. nodorum* in which it is more sequence diverse, and was transferred with the larger region first to *P. tritici-repentis*, and then to *B. sorokiniana* where it is still mobile in the genome due to flanking active transposable elements, or 2) that *ToxA* originated in *B. sorokiniana* and was transferred to *P. tritici-repentis*, then *P. nodorum* where it has rapidly diversified since ~1940. The origins and events leading to the presence of *ToxA* in these three species remains unresolved.

Globally, the sequences of the three known effectors of *P. nodorum* are relatively diverse, with 9, 14, and 6 distinct protein isoforms of ToxA, Tox1, and Tox3, respectively (M. C. McDonald et al., 2013). In the case of ToxA, these protein level differences can have quantitative effects on necrosis inducing activity on wheat (Tan et al., 2012). Syme et al. (2018) identified 13 of the 21 isolates they sequenced were missing at least one of *ToxA*, *Tox1*, or *Tox3*, indicating that they are not individually necessary for infection. Indeed, differences in virulence of *P. nodorum* and susceptibility of wheat in the presence of different combinations of effectors appears to be quantitative in nature (Tan et al., 2015). The contributions of effectors to *P. nodorum* virulence is further complicated by the observations of epistasis of Tox1 and Tox2 over Tox3 (Friesen et al., 2007; Phan et al., 2016).

P. nodorum has a long history of study and has served as a model necrotrophic pathogen (Table 1.2). *P. nodorum* was the first fungal plant-necrotroph and the first Dothideomycete to undergo whole genome sequencing (Hane et al., 2007). The SN15 reference isolate has since undergone extensive resequencing and data curation (Syme et al., 2016), culminating in a telomere-to-telomere complete genome supported by optical mapping and long read sequencing, with extensive manual gene annotation (Bertazzoni et al., 2021, chapter 10). Numerous other isolates have also had their whole genomes sequenced (Table 1.2), including three other near complete genome sequences of the U.S.A. isolates SN79-1087, SN4, and SN2000 (Richards et al., 2018), and two pangenomics studies (Richards et al., 2019; Syme et al., 2018). Several other high throughput studies have also investigated the transcriptomes (Ipcho et al., 2012; D. A. B. Jones et al., 2019) and proteomes of *P. nodorum*, which have been used to identify effector candidates and dissect the infection strategy of *P. nodorum*.

Table 1.1: Known inverse gene-for-gene interactions involved in the *Parastagonospora nodorum*-wheat pathosystem. Published cloned genes are indicated in bold.

Effector	S-gene	Description	Reference(s)
ToxA	***Tsn1***	*ToxA* identified from *P. tritici-repentis* by Ciuffetti et al. (1997). *Tsn1* identified by Faris et al. (2010). *ToxA* identified in *P. nodorum* by Friesen et al. (2006).	Tomas et al. (1990). Tomas and Bockus (1987)
Tox1	***Snn1***	*Tox1* identified by Z. Liu et al. (2012). *Snn1* identified by Shi et al. (2016). Epistatic to Tox3-Snn3 (Friesen et al., 2007).	Z. H. Liu et al. (2004)
Tox2[b]	*Snn2*	Predicted 7-10 kDa. Protease abolished activity. Light dependent necrosis. Epistatic to Tox3-Snn3. *Snn2* QTL[a] mapped to wheat chromosome 2DS.	*Tox267*[b] and *Snn2* were recently identified by Richards et al. (2021)[c]
Tox3	*Snn3-B1*, *Snn3-D1*	*Tox3* identified by Z. Liu et al. (2009). Z. Zhang et al. (2011) describe two major QTL involved, *Snn3-B1* confers higher sensitivity to Tox3 than Snn3-D1. Snn3-B1 QTL mapped to wheat chromosome 5BS (Downie et al., 2018; Friesen et al., 2012; Z. Zhang et al., 2011). Snn3-D1 QTL mapped to wheat chromosome 5DS (Z. Zhang et al., 2011).	Friesen et al. (2008)
Tox4	*Snn4*	Predicted 10-30 kDa protein. Light dependent necrosis. *Snn4* QTL mapped to wheat chromosome 1AS.	Abeysekara et al. (2009)
Tox5	*Snn5*	Predicted size 10-30 kDA protein. Light dependent necrosis. Protease abolished activity. Resistant to heat treatment. *Snn5* QTL mapped to wheat chromosome 4BL.	Friesen et al. (2012). *Tox5* was recently identified by Kariyawasam et al. (2021)[c]
Tox6[b]	*Snn6*	Light dependent necrosis.	Gao et al. (2015). *Tox267*[b] and *Snn6* were recently identified by Richards et al. (2021)[c]
Tox7[b]	*Snn7*	Predicted size 10-30 kDa protein. Resistant to heat treatment. Loss of function with SDS (surfactant) and Dithiothreitol (reduces disulphide bonds) treatment.	Shi et al. (2015). *Tox267*[b] was recently identified by Richards et al. (2021)[c]

[a] quantitative trait loci [b] Richards et al. (2021) reported that a single effector protein is responsible for the described activities of *Tox2*, *Tox6*, and *Tox7*, which they refer to as *SnTox267*. [c] These references were added during final revisions, but are not discussed further in text as they were not available at the time of writing or examination.

Table 1.2: Published data resources for *Parastagonospora nodorum*.

Type	Isolate(s)	Description	Reference	URL
genome	SN15	The first genome of the species. Sanger	Hane et al. (2007)	https://www.ncbi.nlm.nih.gov/assembly/GCF_000146915.1/
genome	SN79-1087	Illumina sequenced genomes of two isolates. Only SN79 was made available.	Syme et al. (2013)	https://www.ncbi.nlm.nih.gov/assembly/GCA_002216185.1/
genome	SN15	Updated illumina sequenced genome and gene annotations.	Syme et al. (2016)	https://github.com/robsyme/Parastagonospora_nodorum_SN15
genome	21 *P. nodorum* isolates	Illumina	Syme et al. (2018)	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA476481/
genome	SN79-1087, SN2000, SN4	Pacbio sequenced genomes of 3 isolates.	Richards et al. (2018)	https://www.ncbi.nlm.nih.gov/bioproject/PRJNA398070/
genome	SN15	PacBio sequenced genome.	M. C. McDonald et al. (2019)	https://www.ncbi.nlm.nih.gov/assembly/GCA_008452785.1
genome	197 *P. nodorum* isolates	Illumina genome-wide association study (GWAS) and population genetics analysis	Richards et al. (2019)	https://www.ncbi.nlm.nih.gov/bioproject/398070
genome	159 *P. nodorum* isolates	Illumina GWAS and population genetics analysis.	Pereira et al. (2020)	https://www.ncbi.nlm.nih.gov/bioproject/?term=PRJNA606320
genome	SN15	PacBio sequenced genome with optical map.	Bertazzoni et al. (2021)[a]	
transcriptome	SN15	EST sequencing used to accompany gene prediction.	Hane et al. (2007)	
transcriptome	SN15, SN79-1087	Microarray analysis of *P. nodorum* at multiple time points during infection.	Ipcho et al. (2012)	
transcriptome	SN15, *pf2-69*	Mixed illumina RNA-seq of *P. nodorum* and wheat. Isolate *pf2-69* is an SN15 mutant lacking the transcription factor PnPf2.	D. A. B. Jones et al. (2019)[b]	https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE150493
proteome	SN15, *gna1-35*	MSMS qualitative comparison of SN15 and an SN15 mutant lacking the Gα signal transduction protein.	Tan et al. (2009)	
proteome	SN15, *gna1-35*	MSMS quantitative comparison of SN15 and an SN15 mutant lacking the Gα signal transduction protein.	Casey et al. (2010)	
proteome	SN15	MSMS qualitative analysis of SN15 in culture. Used to improve gene predictions.	Syme et al. (2016)	

[a] included as supplementary chapter 10 [b] included as supplementary chapter 9

1.7 Objectives

Plant pathology is an important field, helping to ensure future food security and economic stability of farmers and nations. Effector discovery and characterisation, and evolution of pathogen populations are primary research foci in the field, and bioinformatics plays a major role in driving those investigations. However, numerous technical barriers remain to the widespread application of bioinformatics and genomics techniques in these investigations. Effector proteins are difficult to predict as they generally lack sequence conservation. Similarly, efficiently processing large numbers of genomes while ensuring consistent quality requires the development of new pipelines. This project aims to overcome these technical barriers and apply the techniques developed to the genomes of a *Parastagonospora nodorum* population. Additionally, we aim to understand the forces driving evolution of fungal plant-pathogens in agricultural ecosystems and predict plant pathogen interactions using genetic data. We divide these broad goals into three separate but complementary themes described in the following sections and Table 1.3.

1.7.1 Theme 1. Host-pathogen interactions

In bioinformatics research and when working with non-model organisms, it is common to encounter aspects of biology for no which adequate tools to enable their analysis have been developed. Broadly applicable bioinformatics tools or pipelines such as genome assemblers, read aligners, variant callers, and orthology clustering tools frequently require modification or optimisation when applied to non-model organisms which may have specific characteristics that deviate from model assumptions. The analysis of genomes and other data often requires the development of tools and pipelines to deal with these peculiarities. Within the field of fungal plant pathology, important examples include the EffectorP (Sperschneider, Dodds, Gardiner, et al., 2018; Sperschneider et al., 2016), ApoplastP (Sperschneider, Dodds, Singh, et al., 2018), and LOCALIZER (Sperschneider et al., 2017) tools which have enabled prediction of effector proteins, RIPcal and OculterCut (Hane & Oliver, 2008; Testa et al., 2016) which have enabled detection and quantitation of fungal-specific repeat-induced point mutations (RIP), and the specialised fungal gene prediction tools CodingQuarry (Testa et al., 2015) and SnowyOwl (Reid et al., 2014). To supplement the applied research themes detailed below, a suite of tools to describe and predict characteristics of fungal pathogens and host-pathogen interactions will be developed. These tools will primarily focus on effector protein prediction using machine learning (chapter 3) and remote homology matching (chapter 4). We will also use genomic data to describe trophic functional groups of pathogens (chapter 5. These tools will provide valuable metadata to fungal genomes and proteomes, and help prioritise future research efforts (e.g. functional characterisation and effector discovery).

Table 1.3: The themes of this thesis and the chapters that address them.

Theme	Name	Description
1	Host-pathogen interactions	Develop tools and methods to use genomic data to characterise fungal pathogen behaviour and identify effector protein candidates.
2	Spatial survey and pattern detection	Investigate spatial and genetic variability of local Western Australian pathogen populations.
3	Deep population analysis	Perform an in-depth analysis of the *P. nodorum* pangenome, describe patterns of mutation, and identify effector content.
Chapter	**Themes addressed**	**Title**
3	Theme 1	Predector: an automated and combinative method for the predictive ranking of candidate effector proteins of fungal plant-pathogens.
4	Theme 1	Remote homology clustering identifies lowly conserved families of effector proteins in plant-pathogenic fungi.
5	Theme 1	"CATAStrophy", a Genome-Informed Trophic Classification of Filamentous Plant Pathogens — How Many Different Types of Filamentous Plant Pathogens Are There?
6[a]	Theme 1	Hierarchical clustering of MS/MS spectra from the firefly metabolome identifies new lucibufagin compounds.
7	Theme 2	Crop-Zone Weed Mycobiomes of the South-Western Australian Grain Belt.
8	Theme 2	Low Amplitude Boom-and-Bust Cycles Define the Septoria Nodorum Blotch Interaction.
9[a]	Theme 3	A specific fungal transcription factor controls effector gene expression and orchestrates the establishment of the necrotrophic pathogen lifestyle on wheat.
10[a]	Theme 3	Chromosome-level genome assembly and manually-curated proteome of model necrotroph *Parastagonospora nodorum* Sn15 reveals a genome-wide trove of effector-like homologs, and redundancy of virulence-related functions within an accessory chromosome.
11	Themes 2 and 3	Novel effector candidates and large accessory genome revealed by population genomic analysis of *Parastagonospora nodorum*.

[a] These chapters are submitted as supplementary material and should not contribute to thesis assessment

1.7.2 Theme 2. Spatial survey and pattern detection

We will describe spatio-temporal distributions of pathogen species and agriculturally relevant loci using geographic information from biological samples. Using an existing collection of Western Australian (WA) *P. nodorum* isolates as a primary use case, we will investigate the genetic structure and phylogenetic relationships between isolates and analyse potential covariates that might explain these relationships such as sampling time and location. In chapter 8 we will investigate this population using short sequence repeat (SSR) markers, and in chapter 11 we will perform a detailed analysis using thousands of single nucleotide polymorphism (SNP) markers generated from whole genome sequencing (WGS). We hope to demonstrate the utility of population analysis and large-scale sequencing as a tool not only for scientific discovery, but also for disease monitoring and data-driven disease management. The presence of distinct sub-populations with differing pathogenicity profiles might, for example, inform cultivar selection. But there are also more fundamental questions that this project can answer, including how diverse is the population and whether it's likely that a population is sexually reproducing in the wild.

In addition to the population level analysis of *P. nodorum*, in chapter 7 we will also investigate the fungal species diversity of the leaf microbiomes of a range of common agricultural weeds, which might serve as secondary hosts to fungal pathogens. Using ribosomal DNA (rDNA) internal transcribed spacer (ITS) and large subunit D2 amplicon metabarcoding sequencing, we will assess species presence and absence across Western Australia, and compare the microbiome profiles between sampled locations and weed species. Weeds and vegetation surrounding agricultural crops could harbor pathogens and enable their persistence over cropping seasons or rotations, leading to potential future outbreaks. This research will identify whether fungal pathogens do persist on weed phyllospheres and whether additional control measures are required to prevent pathogens with complex lifecycles (e.g. heteroecious rust pathogens) or secondary hosts from reinfecting crops.

1.7.3 Theme 3. Deep population analyses

Building upon concepts from theme 2, to determine genetic features and evolutionary signatures associated with agriculturally relevant traits, the genomes of a *P. nodorum* population will be sequenced and compared. Variations between the genomes such as SNPs and gene presence-absence variations (PAVs) will be determined and used to estimate phylogenies and dissect the population structure. These variants will also be used to describe the possible mechanisms of mutation generating diversity in these populations, which may have practical applications in assessing adaptive potential and the long-term viability of fungicide targets or crop-resistance. Gene predictions for each genome will be used to construct a pangenome of protein orthologous groups, from which effector candidates can be derived using the tools developed as part of theme 1 and from signatures of positive selection.
Together these complementary themes will provide a suite of useful tools and methods for

fungal pathogen genomics, and deliver one of the first insights into the evolution and structure of the WA *Parastagonospora nodorum* population.

1.8 References

Abeysekara, N. S., Friesen, T. L., Keller, B., & Faris, J. D. (2009). Identification and characterization of a novel host–toxin interaction in the wheat–*Stagonospora nodorum* pathosystem. *Theoretical and Applied Genetics*, *120*(1), 117–126. https://doi.org/10.1007/s00122-009-1163-6

Akamatsu, H., Taga, M., Kodama, M., Johnson, R., Otani, H., & Kohmoto, K. (1999). Molecular karyotypes for Alternaria plant pathogens known to produce host-specific toxins. *Current Genetics*, *35*(6), 647–656. https://doi.org/10.1007/s002940050464

Appiano, M., Catalano, D., Martínez, M. S., Lotti, C., Zheng, Z., Visser, R. G. F., Ricciardi, L., Bai, Y., & Pavan, S. (2015). Monocot and dicot MLO powdery mildew susceptibility factors are functionally conserved in spite of the evolution of class-specific molecular features. *BMC Plant Biology*, *15*(1). https://doi.org/10.1186/s12870-015-0639-6

Ashwin, N. M. R., Barnabas, L., Ramesh Sundar, A., Malathi, P., Viswanathan, R., Masi, A., Agrawal, G. K., & Rakwal, R. (2017). Comparative secretome analysis of *Colletotrichum falcatum* identifies a cerato-platanin protein (EPL1) as a potential pathogen-associated molecular pattern (PAMP) inducing systemic resistance in sugarcane. *Journal of Proteomics*, *169*, 2–20. https://doi.org/10.1016/j.jprot.2017.05.020

Aylward, J., Steenkamp, E. T., Dreyer, L. L., Roets, F., Wingfield, B. D., & Wingfield, M. J. (2017). A plant pathology perspective of fungal genome sequencing. *IMA Fungus*, *8*(1), 1–15. https://doi.org/10.5598/imafungus.2017.08.01.01

Baccelli, I., Comparini, C., Bettini, P. P., Martellini, F., Ruocco, M., Pazzagli, L., Bernardi, R., & Scala, A. (2012). The expression of the cerato-platanin gene is related to hyphal growth and chlamydospores formation in Ceratocystis platani. *FEMS Microbiology Letters*, *327*(2), 155–163. https://doi.org/10.1111/j.1574-6968.2011.02475.x

Baccelli, I., Lombardi, L., Luti, S., Bernardi, R., Picciarelli, P., Scala, A., & Pazzagli, L. (2014). Cerato-Platanin Induces Resistance in Arabidopsis Leaves through Stomatal Perception, Overexpression of Salicylic Acid- and Ethylene-Signalling Genes and Camalexin Biosynthesis. *PLOS ONE*, *9*(6). https://doi.org/10.1371/journal.pone.0100959

Baccelli, I., Luti, S., Bernardi, R., Scala, A., & Pazzagli, L. (2014). Cerato-platanin shows expansin-like activity on cellulosic materials. *Applied Microbiology and Biotechnology*, *98*(1), 175–184. https://doi.org/10.1007/s00253-013-4822-0

Badet, T., Oggenfuss, U., Abraham, L., McDonald, B. A., & Croll, D. (2020). A 19-isolate reference-quality global pangenome for the fungal wheat pathogen *Zymoseptoria tritici*. *BMC Biology*, *18*(1), 12. https://doi.org/10.1186/s12915-020-0744-3

Badouin, H., Gladieux, P., Gouzy, J., Siguenza, S., Aguileta, G., Snirc, A., Prieur, S. L., Jeziorski, C., Branca, A., & Giraud, T. (2017). Widespread selective sweeps throughout the genome of model plant pathogenic fungi and identification of effector candidates. *Molecular Ecology*, *26*(7), 2041–2062. https://doi.org/10.1111/mec.13976

Bailey, B. A. (1995). Purification of a Protein from Culture Filtrates of *Fusarium oxysporum* that Induces Ethylene and Necrosis in Leaves of *Erythroxylum coca*. *Phytopathology*, *85*(10), 1250–1255. https://doi.org/10.1094/Phyto-85-1250

Balesdent, M.-H., Fudal, I., Ollivier, B., Bally, P., Grandaubert, J., Eber, F., Chèvre, A.-M., Leflon, M., & Rouxel, T. (2013). The dispensable chromosome of *Leptosphaeria maculans* shelters an effector gene conferring avirulence towards *Brassica rapa*. *New Phytologist*, *198*(3), 887–898. https://doi.org/10.1111/nph.12178

Balint-Kurti, P. (2019). The plant hypersensitive response: Concepts, control and consequences. *Molecular Plant Pathology*, *20*(8), 1163–1178. https://doi.org/10.1111/mpp.12821

Bertazzoni, S., Jones, D. A. B., Phan, H. T., Tan, K.-C., & Hane, J. K. (2021). Chromosome-level genome assembly and manually-curated proteome of model necrotroph *Parastagonospora nodorum* sn15 reveals a genome-wide trove of candidate effector homologs, and redundancy of virulence-related functions within an accessory chromosome. *BMC Genomics*, *22*(382). https://doi.org/10.1186/s12864-021-07699-8

Bertazzoni, S., Williams, A. H., Jones, D. A., Syme, R. A., Tan, K.-C., & Hane, J. K. (2018). Accessories Make the Outfit: Accessory Chromosomes and Other Dispensable DNA Regions in Plant-Pathogenic Fungi. *Molecular Plant-Microbe Interactions*, *31*(8), 779–788. https://doi.org/10.1094/MPMI-06-17-0135-FI

Blondeau, K., Blaise, F., Graille, M., Kale, S. D., Linglin, J., Ollivier, B., Labarde, A., Lazar, N., Daverdin, G., Balesdent, M.-H., Choi, D. H. Y., Tyler, B. M., Rouxel, T., van Tilbeurgh, H., & Fudal, I. (2015). Crystal structure of the effector AvrLm4–7 of *Leptosphaeria maculans* reveals insights into its translocation into plant cells and recognition by resistance proteins. *The Plant Journal*, *83*(4), 610–624. https://doi.org/10.1111/tpj.12913

Bolton, M. D., van Esse, H. P., Vossen, J. H., de Jonge, R., Stergiopoulos, I., Stulemeijer, I. J. E., Berg, G. C. M. V. D., Borrás-Hidalgo, O., Dekker, H. L., de Koster, C. G., de Wit, P. J. G. M., Joosten, M. H. A. J., & Thomma, B. P. H. J. (2008). The novel *Cladosporium fulvum* lysin motif effector Ecp6 is a virulence factor with orthologues in other fungal species. *Molecular Microbiology*, *69*(1), 119–136. https://doi.org/10.1111/j.1365-2958.2008.06270.x

Bourras, S., McNally, K. E., Ben-David, R., Parlange, F., Roffler, S., Praz, C. R., Oberhaensli, S., Menardo, F., Stirnweis, D., Frenkel, Z., Schaefer, L. K., Flückiger, S., Treier, G., Herren, G., Korol, A. B., Wicker, T., & Keller, B. (2015). Multiple Avirulence Loci and Allele-Specific Effector Recognition Control the Pm3 Race-Specific Resistance of Wheat to Powdery Mildew. *The Plant Cell*, *27*(10), 2991–3012. https://doi.org/10.1105/tpc.15.00171

Boutemy, L. S., King, S. R. F., Win, J., Hughes, R. K., Clarke, T. A., Blumenschein, T. M. A., Kamoun, S., & Banfield, M. J. (2011). Structures of *Phytophthora* RXLR Effector Proteins: A CONSERVED BUT ADAPTABLE FOLD UNDERPINS FUNCTIONAL DIVERSITY. *Journal of Biological Chemistry*, *286*(41), 35834–35842. https://doi.org/10.1074/jbc.M111.262303

Bozkurt, T. O., & Kamoun, S. (2020). The plant–pathogen haustorial interface at a glance (A.-M. Lennon-Duménil, Ed.). *Journal of Cell Science*, *133*(5). https://doi.org/10.1242/jcs.237958

Braumann, I., van den Berg, M., & Kempken, F. (2008). Repeat induced point mutation in two asexual fungi, *Aspergillus niger* and *Penicillium chrysogenum*. *Current Genetics*, *53*(5), 287–297. https://doi.org/10.1007/s00294-008-0185-y

Buist, G., Steen, A., Kok, J., & Kuipers, O. P. (2008). LysM, a widely distributed protein motif for binding to (peptido)glycans. *Molecular Microbiology*, *68*(4), 838–847. https://doi.org/10.1111/j.1365-2958.2008.06211.x

Burrack, L. S., Hutton, H. F., Matter, K. J., Clancey, S. A., Liachko, I., Plemmons, A. E., Saha, A., Power, E. A., Turman, B., Thevandavakkam, M. A., Ay, F., Dunham, M. J., & Berman, J. (2016). Neocentromeres Provide Chromosome Segregation Accuracy and Centromere Clustering to

Multiple Loci along a Candida albicans Chromosome. *PLOS Genetics*, *12*(9), e1006317. https://doi.org/10.1371/journal.pgen.1006317

Büschges, R., Hollricher, K., Panstruga, R., Simons, G., Wolter, M., Frijters, A., van Daelen, R., van der Lee, T., Diergaarde, P., Groenendijk, J., Töpsch, S., Vos, P., Salamini, F., & Schulze-Lefert, P. (1997). The barley mlo gene: A novel control element of plant pathogen resistance. *Cell*, *88*(5), 695–705. https://doi.org/10.1016/S0092-8674(00)81912-1

Cambareri, E. B., Jensen, B. C., Schabtach, E., & Selker, E. U. (1989). Repeat-induced G-C to A-T mutations in Neurospora. *Science*, *244*(4912), 1571–1575. https://doi.org/10.1126/science.2544994

Cao, L., Blekemolen, M. C., Tintor, N., Cornelissen, B. J. C., & Takken, F. L. W. (2018). The *Fusarium oxysporum* Avr2-Six5 Effector Pair Alters Plasmodesmatal Exclusion Selectivity to Facilitate Cell-to-Cell Movement of Avr2. *Molecular Plant*, *11*(5), 691–705. https://doi.org/10.1016/j.molp.2018.02.011

Casey, T., Solomon, P. S., Bringans, S., Tan, K.-C., Oliver, R. P., & Lipscombe, R. (2010). Quantitative proteomic analysis of G-protein signalling in *Stagonospora nodorum* using isobaric tags for relative and absolute quantification. *PROTEOMICS*, *10*(1), 38–47. https://doi.org/10.1002/pmic.200900474

Castell-Miller, C. V., Gutierrez-Gonzalez, J. J., Tu, Z. J., Bushley, K. E., Hainaut, M., Henrissat, B., & Samac, D. A. (2016). Genome Assembly of the Fungus *Cochliobolus miyabeanus*, and Transcriptome Analysis during Early Stages of Infection on American Wildrice (*Zizania palustris* L.) (J.-H. Yu, Ed.). *PLOS ONE*, *11*(6), e0154122. https://doi.org/10.1371/journal.pone.0154122

Catanzariti, A.-M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., & Ellis, J. G. (2006). Haustorially Expressed Secreted Proteins from Flax Rust Are Highly Enriched for Avirulence Elicitors. *The Plant Cell*, *18*(1), 243–256. https://doi.org/10.1105/tpc.105.035980

Caten, C. E., & Newton, A. C. (2000). Variation in cultural characteristics, pathogenicity, vegetative compatibility and electrophoretic karyotype within field populations of *Stagonospora nodorum*. *Plant Pathology*, *49*(2), 219–226. https://doi.org/10.1046/j.1365-3059.2000.00441.x

Cesari, S., Thilliez, G., Ribot, C., Chalvon, V., Michel, C., Jauneau, A., Rivas, S., Alaux, L., Kanzaki, H., Okuyama, Y., Morel, J.-B., Fournier, E., Tharreau, D., Terauchi, R., & Kroj, T. (2013). The rice resistance protein pair RGA4/RGA5 recognizes the *Magnaporthe oryzae* effectors AVR-pia and AVR1-CO39 by direct binding. *The Plant Cell*, *25*(4), 1463–1481. https://doi.org/10.1105/tpc.112.107201

Chassot, C., & Métraux, J.-P. (2005). The cuticle as source of signals for plant defense. *Plant Biosystems*, *139*(1), 28–31. https://doi.org/10.1080/11263500500056344

Chen, C., Lian, B., Hu, J., Zhai, H., Wang, X., Venu, R., Liu, E., Wang, Z., Chen, M., Wang, B., Wang, G.-L., Wang, Z., & Mitchell, T. K. (2013). Genome comparison of two *Magnaporthe oryzae* field isolates reveals genome variations and potential virulence effectors. *BMC Genomics*, *14*(1), 887. https://doi.org/10.1186/1471-2164-14-887

Chen, H., Kovalchuk, A., Keriö, S., & Asiegbu, F. O. (2013). Distribution and bioinformatic analysis of the cerato-platanin protein family in Dikarya. *Mycologia*, *105*(6), 1479–1488. https://doi.org/10.3852/13-115

Chen, H., Quintana, J., Kovalchuk, A., Ubhayasekera, W., & Asiegbu, F. O. (2015). A cerato-platanin-like protein HaCPL2 from *Heterobasidion annosum* sensu stricto induces cell death in *Nicotiana tabacum* and *Pinus sylvestris*. *Fungal Genetics and Biology*, *84*, 41–51. https://doi.org/10.1016/j.fgb.2015.09.007

Chen, J., Wu, J., Zhang, P., Dong, C., Upadhyaya, N. M., Zhou, Q., Dodds, P., & Park, R. F. (2019). De Novo Genome Assembly and Comparative Genomics of the Barley Leaf Rust Pathogen *Puccinia hordei* Identifies Candidates for Three Avirulence Genes. *G3: Genes, Genomes, Genetics*, *9*(10), 3263–3271. https://doi.org/10.1534/g3.119.400450

Chiapello, H., Mallet, L., Guérin, C., Aguileta, G., Amselem, J., Kroj, T., Ortega-Abboud, E., Lebrun, M.-H., Henrissat, B., Gendrault, A., Rodolphe, F., Tharreau, D., & Fournier, E. (2015). Deciphering Genome Content and Evolutionary Relationships of Isolates from the Fungus *Magnaporthe oryzae* Attacking Different Host Plants. *Genome Biology and Evolution*, *7*(10), 2896–2912. https://doi.org/10.1093/gbe/evv187

Ciuffetti, L. M., Tuori, R. P., & Gaventa, J. M. (1997). A single gene encodes a selective toxin causal to the development of tan spot of wheat. *The Plant Cell*, *9*(2), 135–144. https://doi.org/10.1105/tpc.9.2.135

Coleman, J. J., Rounsley, S. D., Rodriguez-Carres, M., Kuo, A., Wasmann, C. C., Grimwood, J., Schmutz, J., Taga, M., White, G. J., Zhou, S., Schwartz, D. C., Freitag, M., Ma, L.-j., Danchin, E. G. J., Henrissat, B., Coutinho, P. M., Nelson, D. R., Straney, D., Napoli, C. A., … VanEtten, H. D. (2009). The Genome of *Nectria haematococca*: Contribution of Supernumerary Chromosomes to Gene Expansion. *PLOS Genetics*, *5*(8), e1000618. https://doi.org/10.1371/journal.pgen.1000618

The Cost of Sequencing a Human Genome. (2020). Retrieved September 4, 2020, from https://www.genome.gov/about-genomics/fact-sheets/Sequencing-Human-Genome-cost

Covert, S. F. (1998). Supernumerary chromosomes in filamentous fungi. *Current Genetics*, *33*(5), 311–319. https://doi.org/10.1007/s002940050342

Croll, D., Zala, M., & McDonald, B. A. (2013). Breakage-fusion-bridge Cycles and Large Insertions Contribute to the Rapid Evolution of Accessory Chromosomes in a Fungal Pathogen. *PLOS Genetics*, *9*(6), e1003567. https://doi.org/10.1371/journal.pgen.1003567

Cui, H., Tsuda, K., & Parker, J. E. (2015). Effector-Triggered Immunity: From Pathogen Perception to Robust Defense. *Annual Review of Plant Biology*, *66*(1), 487–511. https://doi.org/10.1146/annurev-arplant-050213-040012

Dalman, K., Himmelstrand, K., Olson, Å., Lind, M., Brandström-Durling, M., & Stenlid, J. (2013). A Genome-Wide Association Study Identifies Genomic Regions for Virulence in the Non-Model Organism *Heterobasidion annosum s.s*. *PLOS ONE*, *8*(1), e53525. https://doi.org/10.1371/journal.pone.0053525

Dangl, J. L., & Jones, J. D. G. (2001). Plant pathogens and integrated defence responses to infection. *Nature*, *411*(6839), 826–833. https://doi.org/10.1038/35081161

Dean, R. A., Talbot, N. J., Ebbole, D. J., Farman, M. L., Mitchell, T. K., Orbach, M. J., Thon, M., Kulkarni, R., Xu, J.-R., Pan, H., Read, N. D., Lee, Y.-H., Carbone, I., Brown, D., Oh, Y. Y., Donofrio, N., Jeong, J. S., Soanes, D. M., Djonovic, S., … Birren, B. W. (2005). The genome sequence of the rice blast fungus *Magnaporthe grisea*. *Nature*, *434*(7036), 980–986. https://doi.org/10.1038/nature03449

de Guillen, K. d., Ortiz-Vallejo, D., Gracy, J., Fournier, E., Kroj, T., & Padilla, A. (2015). Structure Analysis Uncovers a Highly Diverse but Structurally Conserved Effector Family in Phytopathogenic Fungi. *PLOS Pathogens*, *11*(10), e1005228. https://doi.org/10.1371/journal.ppat.1005228

de Jonge, R., Bolton, M. D., Kombrink, A., van den Berg, G. C. M., Yadeta, K. A., & Thomma, B. P. H. J. (2013). Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. *Genome Research*, *23*(8), 1271–1282. https://doi.org/10.1101/gr.152660.112

de Jonge, R., van Esse, H. P., Kombrink, A., Shinya, T., Desaki, Y., Bours, R., van der Krol, S., Shibuya, N., Joosten, M. H. A. J., & Thomma, B. P. H. J. (2010). Conserved Fungal LysM Effector Ecp6

Prevents Chitin-Triggered Immunity in Plants. *Science*, *329*(5994), 953–955. https://doi.org/10.1126/science.1190859

de Jonge, R., van Esse, H. P., Maruthachalam, K., Bolton, M. D., Santhanam, P., Saber, M. K., Zhang, Z., Usami, T., Lievens, B., Subbarao, K. V., & Thomma, B. P. H. J. (2012). Tomato immune receptor Ve1 recognizes effector of multiple fungal pathogens uncovered by genome and RNA sequencing. *Proceedings of the National Academy of Sciences*, *109*(13), 5110–5115. https://doi.org/10.1073/pnas.1119623109

Deng, C. H., Plummer, K. M., Jones, D. A. B., Mesarich, C. H., Shiller, J., Taranto, A. P., Robinson, A. J., Kastner, P., Hall, N. E., Templeton, M. D., & Bowen, J. K. (2017). Comparative analysis of the predicted secretomes of Rosaceae scab pathogens *Venturia inaequalis* and *V. pirina* reveals expanded effector families and putative determinants of host range. *BMC Genomics*, *18*(1), 339. https://doi.org/10.1186/s12864-017-3699-1

de O. Barsottini, M. R., de Oliveira, J. F., Adamoski, D., Teixeira, P. J. P. L., do Prado, P. F. V., Tiezzi, H. O., Sforça, M. L., Cassago, A., Portugal, R. V., de Oliveira, P. S. L., de M. Zeri, A. C., Dias, S. M. G., Pereira, G. A. G., & Ambrosio, A. L. B. (2013). Functional Diversification of Cerato-Platanins in *Moniliophthora perniciosa* as Seen by Differential Expression and Protein Function Specialization. *Molecular Plant-Microbe Interactions*, *26*(11), 1281–1293. https://doi.org/10.1094/MPMI-05-13-0148-R

de Oliveira, A. L., Gallo, M., Pazzagli, L., Benedetti, C. E., Cappugi, G., Scala, A., Pantera, B., Spisni, A., Pertinhez, T. A., & Cicero, D. O. (2011). The structure of the elicitor cerato-platanin (cp), the first member of the cp fungal protein family, reveals a double $\Psi\beta$-barrel fold and carbohydrate binding. *Journal of Biological Chemistry*, *286*(20), http://www.jbc.org/content/286/20/17560.full.pdf+html, 17560–17568. https://doi.org/10.1074/jbc.M111.223644

de Vries, S., Stukenbrock, E. H., & Rose, L. E. (2020). Rapid evolution in plant–microbe interactions – an evolutionary genomics perspective. *New Phytologist*, *226*(5), 1256–1262. https://doi.org/10.1111/nph.16458

Di, X., Cao, L., Hughes, R. K., Tintor, N., Banfield, M. J., & Takken, F. L. W. (2017). Structure–function analysis of the *Fusarium oxysporum* Avr2 effector allows uncoupling of its immune-suppressing activity from recognition. *New Phytologist*, *216*(3), 897–914. https://doi.org/10.1111/nph.14733

Di, X., Gomila, J., Ma, L., van den Burg, H. A., & Takken, F. L. W. (2016). Uptake of the fusarium effector avr2 by tomato is not a cell autonomous event. *Frontiers in Plant Science*, *7*. https://doi.org/10.3389/fpls.2016.01915

Dong, Y., Li, Y., Zhao, M., Jing, M., Liu, X., Liu, M., Guo, X., Zhang, X., Chen, Y., Liu, Y., Liu, Y., Ye, W., Zhang, H., Wang, Y., Zheng, X., Wang, P., & Zhang, Z. (2015). Global Genome and Transcriptome Analyses of *Magnaporthe oryzae* Epidemic Isolate 98-06 Uncover Novel Effectors and Pathogenicity-Related Genes, Revealing Gene Gain and Lose Dynamics in Genome Evolution. *PLOS Pathogens*, *11*(4), e1004801. https://doi.org/10.1371/journal.ppat.1004801

Dou, D., Kale, S. D., Wang, X., Jiang, R. H. Y., Bruce, N. A., Arredondo, F. D., Zhang, X., & Tyler, B. M. (2008). RXLR-Mediated Entry of *Phytophthora sojae* Effector Avr1b into Soybean Cells Does Not Require Pathogen-Encoded Machinery. *The Plant Cell*, *20*(7), 1930–1947. https://doi.org/10.1105/tpc.107.056093

Downie, R. C., Bouvet, L., Furuki, E., Gosman, N., Gardner, K. A., Mackay, I. J., Campos Mantello, C., Mellers, G., Phan, H. T. T., Rose, G. A., Tan, K.-C., Oliver, R. P., & Cockram, J. (2018). Assessing European Wheat Sensitivities to *Parastagonospora nodorum* Necrotrophic Effectors and Fine-

Mapping the *Snn3*-B1 Locus Conferring Sensitivity to the Effector SnTox3. *Frontiers in Plant Science*, *9*. https://doi.org/10.3389/fpls.2018.00881

Drenth, A., McTaggart, A. R., & Wingfield, B. D. (2019). Fungal clones win the battle, but recombination wins the war. *IMA Fungus*, *10*(1), 18. https://doi.org/10.1186/s43008-019-0020-8

Ekchaweng, K., Evangelisti, E., Schornack, S., Tian, M., & Churngchow, N. (2017). The plant defense and pathogen counterdefense mediated by *Hevea brasiliensis* serine protease HbSPA and *Phytophthora palmivora* extracellular protease inhibitor PpEPI10. *PLOS ONE*, *12*(5), e0175795. https://doi.org/10.1371/journal.pone.0175795

Ellis, J. G., Lagudah, E. S., Spielmeyer, W., & Dodds, P. N. (2014). The past, present and future of breeding rust resistant wheat. *Frontiers in Plant Science*, *5*. https://doi.org/10.3389/fpls.2014.00641

Esse, H. P. v., Klooster, J. W. v., Bolton, M. D., Yadeta, K. A., Baarlen, P. v., Boeren, S., Vervoort, J., Wit, P. J. G. M. d., & Thomma, B. P. H. J. (2008). The Cladosporium fulvum Virulence Protein Avr2 Inhibits Host Proteases Required for Basal Defense. *The Plant Cell*, *20*(7), 1948–1963. https://doi.org/10.1105/tpc.108.059394

FAO, IFAD, UNICEF, WFP, & WHO. (2020). *The State of Food Security and Nutrition in the World 2020. Transforming food systems for affordable healthy diets.* Retrieved August 26, 2020, from https://doi.org/10.4060/ca9692en

Faris, J. D., Zhang, Z., Lu, H., Lu, S., Reddy, L., Cloutier, S., Fellers, J. P., Meinhardt, S. W., Rasmussen, J. B., Xu, S. S., Oliver, R. P., Simons, K. J., & Friesen, T. L. (2010). A unique wheat disease resistance-like gene governs effector-triggered susceptibility to necrotrophic pathogens. *Proceedings of the National Academy of Sciences*, *107*(30), 13544–13549. https://doi.org/10.1073/pnas.1004090107

Farman, M. L., Eto, Y., Nakao, T., Tosa, Y., Nakayashiki, H., Mayama, S., & Leong, S. A. (2002). Analysis of the Structure of the AVR1-CO39 Avirulence Locus in Virulent Rice-Infecting Isolates of *Magnaporthe grisea*. *Molecular Plant-Microbe Interactions*, *15*(1), 6–16. https://doi.org/10.1094/MPMI.2002.15.1.6

Fellbrich, G., Romanski, A., Varet, A., Blume, B., Brunner, F., Engelhardt, S., Felix, G., Kemmerling, B., Krzymowska, M., & Nürnberger, T. (2002). NPP1, a Phytophthora-associated trigger of plant defense in parsley and Arabidopsis. *The Plant Journal*, *32*(3), 375–390. https://doi.org/10.1046/j.1365-313X.2002.01454.x

Fenton, A., Antonovics, J., & Brockhurst, M. A. (2009). Inverse-Gene-for-Gene Infection Genetics and Coevolutionary Dynamics. *The American Naturalist*, *174*(6), E230–E242. https://doi.org/10.1086/645087

Fitzpatrick, D. A. (2012). Horizontal gene transfer in fungi. *FEMS Microbiology Letters*, *329*(1), 1–8. https://doi.org/10.1111/j.1574-6968.2011.02465.x

Flor, H. H. (1971). Current Status of the Gene-For-Gene Concept. *Annual Review of Phytopathology*, *9*(1), 275–296. https://doi.org/10.1146/annurev.py.09.090171.001423

Fones, H. N., Bebber, D. P., Chaloner, T. M., Kay, W. T., Steinberg, G., & Gurr, S. J. (2020). Threats to global food security from emerging fungal and oomycete crop pathogens. *Nature Food*, *1*(6), 332–342. https://doi.org/10.1038/s43016-020-0075-0

Freitag, M., Williams, R. L., Kothe, G. O., & Selker, E. U. (2002). A cytosine methyltransferase homologue is essential for repeat-induced point mutation in *Neurospora crassa*. *Proceedings of the National Academy of Sciences*, *99*(13), 8802–8807. https://doi.org/10.1073/pnas.132212899

Frías, M., Brito, N., González, M., & González, C. (2014). The phytotoxic activity of the cerato-platanin BcSpl1 resides in a two-peptide motif on the protein surface. *Molecular Plant Pathology*, *15*(4), 342–351. https://doi.org/10.1111/mpp.12097

Frías, M., González, C., & Brito, N. (2011). BcSpl1, a cerato-platanin family protein, contributes to *Botrytis cinerea* virulence and elicits the hypersensitive response in the host. *New Phytologist*, *192*(2), 483–495. https://doi.org/10.1111/j.1469-8137.2011.03802.x

Friesen, T. L., Stukenbrock, E. H., Liu, Z., Meinhardt, S., Ling, H., Faris, J. D., Rasmussen, J. B., Solomon, P. S., McDonald, B. A., & Oliver, R. P. (2006). Emergence of a new disease as a result of interspecific virulence gene transfer. *Nature Genetics*, *38*(8), 953–956. https://doi.org/10.1038/ng1839

Friesen, T. L., Chu, C., Xu, S. S., & Faris, J. D. (2012). *SnTox5-Snn5*: A novel *Stagonospora nodorum* effector-wheat gene interaction and its relationship with the SnToxA- *Tsn1* and SnTox3- *Snn3* - *B1* interactions: Characterization of the SnTox*5-Snn5*interaction. *Molecular Plant Pathology*, *13*(9), 1101–1109. https://doi.org/10.1111/j.1364-3703.2012.00819.x

Friesen, T. L., Meinhardt, S. W., & Faris, J. D. (2007). The *Stagonospora nodorum*-wheat pathosystem involves multiple proteinaceous host-selective toxins and corresponding host sensitivity genes that interact in an inverse gene-for-gene manner. *The Plant Journal*, *51*(4), 681–692. https://doi.org/10.1111/j.1365-313X.2007.03166.x

Friesen, T. L., Zhang, Z., Solomon, P. S., Oliver, R. P., & Faris, J. D. (2008). Characterization of the Interaction of a Novel *Stagonospora nodorum* Host-Selective Toxin with a Wheat Susceptibility Gene. *Plant Physiology*, *146*(2), 682–693. https://doi.org/10.1104/pp.107.108761

Fry, B. G., Roelants, K., Champagne, D. E., Scheib, H., Tyndall, J. D., King, G. F., Nevalainen, T. J., Norman, J. A., Lewis, R. J., Norton, R. S., Renjifo, C., & de la Vega, R. C. R. (2009). The Toxicogenomic Multiverse: Convergent Recruitment of Proteins Into Animal Venoms. *Annual Review of Genomics and Human Genetics*, *10*(1), 483–511. https://doi.org/10.1146/annurev.genom.9.081307.164356

Fudal, I., Ross, S., Gout, L., Blaise, F., Kuhn, M. L., Eckert, M. R., Cattolico, L., Bernard-Samain, S., Balesdent, M. H., & Rouxel, T. (2007). Heterochromatin-Like Regions as Ecological Niches for Avirulence Genes in the *Leptosphaeria maculans* Genome: Map-Based Cloning of *AvrLm6*. *Molecular Plant-Microbe Interactions*, *20*(4), 459–470. https://doi.org/10.1094/MPMI-20-4-0459

Gaderer, R., Lamdan, N. L., Frischmann, A., Sulyok, M., Krska, R., Horwitz, B. A., & Seidl-Seiboth, V. (2015). Sm2, a paralog of the Trichoderma cerato-platanin elicitor Sm1, is also highly important for plant protection conferred by the fungal-root interaction of Trichoderma with maize. *BMC Microbiology*, *15*(1), 2. https://doi.org/10.1186/s12866-014-0333-0

Galagan, J. E., & Selker, E. U. (2004). RIP: The evolutionary cost of genome defense. *Trends in Genetics*, *20*(9), 417–423. https://doi.org/10.1016/j.tig.2004.07.007

Gao, Y., Faris, J. D., Liu, Z., Kim, Y. M., Syme, R. A., Oliver, R. P., Xu, S. S., & Friesen, T. L. (2015). Identification and Characterization of the SnTox6-Snn6 Interaction in the *Parastagonospora nodorum*–Wheat Pathosystem. *Molecular Plant-Microbe Interactions*, *28*(5), 615–625. https://doi.org/10.1094/MPMI-12-14-0396-R

Gelly, J.-C., Gracy, J., Kaas, Q., Le-Nguyen, D., Heitz, A., & Chiche, L. (2004). The KNOTTIN website and database: A new information system dedicated to the knottin scaffold. *Nucleic Acids Research*, *32*(Database issue), D156–D159. https://doi.org/10.1093/nar/gkh015

Gibbs, G. M., Roelants, K., & O'Bryan, M. K. (2008). The CAP superfamily: Cysteine-rich secretory proteins, antigen 5, and pathogenesis-related 1 proteins–roles in reproduction, cancer, and immune defense. *Endocrine Reviews*, *29*(7), 865–897. https://doi.org/10.1210/er.2008-0032

Gijzen, M., & Nürnberger, T. (2006). Nep1-like proteins from plant pathogens: Recruitment and diversification of the NPP1 domain across taxa. *Phytochemistry*, *67*(16), 1800–1807. https://doi.org/10.1016/j.phytochem.2005.12.008

Giraldo, M. C., Dagdas, Y. F., Gupta, Y. K., Mentlak, T. A., Yi, M., Martinez-Rocha, A. L., Saitoh, H., Terauchi, R., Talbot, N. J., & Valent, B. (2013). Two distinct secretion systems facilitate tissue invasion by the rice blast fungus *Magnaporthe oryzae*. *Nature Communications*, *4*(1). https://doi.org/10.1038/ncomms2996

Gladyshev, E., & Kleckner, N. (2014). Direct recognition of homology between double helices of DNA in *Neurospora crassa*. *Nature Communications*, *5*(1), 3509. https://doi.org/10.1038/ncomms4509

Gladyshev, E., & Kleckner, N. (2016). Recombination-Independent Recognition of DNA Homology for Repeat-Induced Point Mutation (RIP) Is Modulated by the Underlying Nucleotide Sequence. *PLOS Genetics*, *12*(5), e1006015. https://doi.org/10.1371/journal.pgen.1006015

Gladyshev, E., & Kleckner, N. (2017a). DNA sequence homology induces cytosine-to-thymine mutation by a heterochromatin-related pathway in Neurospora. *Nature Genetics*, *49*(6), 887–894. https://doi.org/10.1038/ng.3857

Gladyshev, E., & Kleckner, N. (2017b). Recombination-independent recognition of DNA homology for repeat-induced point mutation. *Current Genetics*, *63*(3), 389–400. https://doi.org/10.1007/s00294-016-0649-4

Gomes, E. V., Costa, M. d. N., de Paula, R. G., Ricci de Azevedo, R., da Silva, F. L., Noronha, E. F., José Ulhoa, C., Neves Monteiro, V., Elena Cardoza, R., Gutiérrez, S., & Nascimento Silva, R. (2015). The Cerato-Platanin protein Epl-1 from *Trichoderma harzianum* is involved in mycoparasitism, plant resistance induction and self cell wall protection. *Scientific Reports*, *5*(1), 17998. https://doi.org/10.1038/srep17998

Gong, A.-d., Jing, Z.-y., Zhang, K., Tan, Q.-q., Wang, G.-l., & Liu, W.-d. (2020). Bioinformatic analysis and functional characterization of the CFEM proteins in maize anthracnose fungus *Colletotrichum graminicola*. *Journal of Integrative Agriculture*, *19*(2), 541–550. https://doi.org/10.1016/S2095-3119(19)62675-4

Goodwin, S. B., M'Barek, S. B., Dhillon, B., Wittenberg, A. H. J., Crane, C. F., Hane, J. K., Foster, A. J., van der Lee, T. A. J., Grimwood, J., Aerts, A., Antoniw, J., Bailey, A., Bluhm, B., Bowler, J., Bristow, J., Burgt, A. v. d., Canto-Canché, B., Churchill, A. C. L., Conde-Ferràez, L., … Kema, G. H. J. (2011). Finished Genome of the Fungal Wheat Pathogen *Mycosphaerella graminicola* Reveals Dispensome Structure, Chromosome Plasticity, and Stealth Pathogenesis. *PLOS Genetics*, *7*(6), e1002070. https://doi.org/10.1371/journal.pgen.1002070

Graham-Taylor, C., Kamphuis, L. G., & Derbyshire, M. C. (2020). A detailed in silico analysis of secondary metabolite biosynthesis clusters in the genome of the broad host range plant pathogenic fungus *Sclerotinia sclerotiorum*. *BMC Genomics*, *21*(1), 7. https://doi.org/10.1186/s12864-019-6424-4

Grünwald, N. J., McDonald, B. A., & Milgroom, M. G. (2016). Population Genomics of Fungal and Oomycete Pathogens. *Annual Review of Phytopathology*, *54*(1), 323–346. https://doi.org/10.1146/annurev-phyto-080614-115913

Hane, J. K., Lowe, R. G. T., Solomon, P. S., Tan, K.-C., Schoch, C. L., Spatafora, J. W., Crous, P. W., Kodira, C., Birren, B. W., Galagan, J. E., Torriani, S. F. F., McDonald, B. A., & Oliver, R. P. (2007). Dothideomycete–Plant Interactions Illuminated by Genome Sequencing and EST Analysis of the Wheat Pathogen Stagonospora nodorum. *The Plant Cell*, *19*(11), 3347–3368. https://doi.org/10.1105/tpc.107.052829

Hane, J. K., & Oliver, R. P. (2008). RIPCAL: A tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. *BMC Bioinformatics*, *9*(1), 478. https://doi.org/10.1186/1471-2105-9-478

Hane, J. K., Rouxel, T., Howlett, B. J., Kema, G. H., Goodwin, S. B., & Oliver, R. P. (2011). A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome Biology*, *12*(5), R45. https://doi.org/10.1186/gb-2011-12-5-r45

Hane, J. K., Williams, A. H., Taranto, A. P., Solomon, P. S., & Oliver, R. P. (2015). Repeat-Induced Point Mutation: A Fungal-Specific, Endogenous Mutagenesis Process. In M. A. van den Berg & K. Maruthachalam (Eds.), *Genetic Transformation Systems in Fungi* (pp. 55–68). Cham, Springer International Publishing. https://doi.org/10.1007/978-3-319-10503-1_4

Hartmann, F. E., Sánchez-Vallet, A., McDonald, B. A., & Croll, D. (2017). A fungal wheat pathogen evolved host specialization by extensive chromosomal rearrangements. *The ISME Journal*, *11*(5), 1189–1204. https://doi.org/10.1038/ismej.2016.196

He, C., Rusu, A. G., Poplawski, A. M., Irwin, J. A. G., & Manners, J. M. (1998). Transfer of a Supernumerary Chromosome Between Vegetatively Incompatible Biotypes of the Fungus *Colletotrichum gloeosporioides*. *Genetics*, *150*(4), 1459–1466. Retrieved August 31, 2020, from https://www.genetics.org/content/150/4/1459

Hocher, A., & Taddei, A. (2020). Subtelomeres as Specialized Chromatin Domains. *BioEssays*, *42*(5), 1900205. https://doi.org/10.1002/bies.201900205

Holt, K. E., Nga, T. V. T., Thanh, D. P., Vinh, H., Kim, D. W., Tra, M. P. V., Campbell, J. I., Hoang, N. V. M., Vinh, N. T., Minh, P. V., Thuy, C. T., Nga, T. T. T., Thompson, C., Dung, T. T. N., Nhu, N. T. K., Vinh, P. V., Tuyet, P. T. N., Phuc, H. L., Lien, N. T. N., … Baker, S. (2013). Tracking the establishment of local endemic populations of an emergent enteric pathogen. *Proceedings of the National Academy of Sciences*, *110*(43), 17522–17527. https://doi.org/10.1073/pnas.1308632110

Hood, M. E., Katawczik, M., & Giraud, T. (2005). Repeat-Induced Point Mutation and the Population Structure of Transposable Elements in *Microbotryum violaceum*. *Genetics*, *170*(3), 1081–1089. https://doi.org/10.1534/genetics.105.042564

Horns, F., Petit, E., Yockteng, R., & Hood, M. E. (2012). Patterns of Repeat-Induced Point Mutation in Transposable Elements of Basidiomycete Fungi. *Genome Biology and Evolution*, *4*(3), 240–247. https://doi.org/10.1093/gbe/evs005

Hubbard, A., Lewis, C. M., Yoshida, K., Ramirez-Gonzalez, R. H., de Vallavieille-Pope, C., Thomas, J., Kamoun, S., Bayles, R., Uauy, C., & Saunders, D. G. (2015). Field pathogenomics reveals the emergence of a diverse wheat yellow rust population. *Genome Biology*, *16*(1), 23. https://doi.org/10.1186/s13059-015-0590-8

Ipcho, S. V. S., Hane, J. K., Antoni, E. A., Ahren, D., Henrissat, B., Friesen, T. L., Solomon, P. S., & Oliver, R. P. (2012). Transcriptome analysis of *Stagonospora nodorum*: Gene models, effectors, metabolism and pantothenate dispensability. *Molecular Plant Pathology*, *13*(6), 531–545. https://doi.org/10.1111/j.1364-3703.2011.00770.x

Islam, M. T., Croll, D., Gladieux, P., Soanes, D. M., Persoons, A., Bhattacharjee, P., Hossain, M. S., Gupta, D. R., Rahman, M. M., Mahboob, M. G., Cook, N., Salam, M. U., Surovy, M. Z., Sancho, V. B., Maciel, J. L. N., NhaniJúnior, A., Castroagudín, V. L., Reges, J. T. d. A., Ceresini, P. C., … Kamoun, S. (2016). Emergence of wheat blast in Bangladesh was caused by a South American lineage of *Magnaporthe oryzae*. *BMC Biology*, *14*(1), 84. https://doi.org/10.1186/s12915-016-0309-7

Jeong, J. S., Mitchell, T. K., & Dean, R. A. (2007). The *Magnaporthe grisea* snodprot1 homolog, MSP1, is required for virulence. *FEMS Microbiology Letters*, *273*(2), 157–165. https://doi.org/10.1111/j.1574-6968.2007.00796.x

John Clutterbuck, A. (2011). Genomic evidence of repeat-induced point mutation (RIP) in filamentous ascomycetes. *Fungal Genetics and Biology*, *48*(3), 306–326. https://doi.org/10.1016/j.fgb.2010.09.002

Jones, D. A. B., John, E., Rybak, K., Phan, H. T. T., Singh, K. B., Lin, S.-Y., Solomon, P. S., Oliver, R. P., & Tan, K.-C. (2019). A specific fungal transcription factor controls effector gene expression and orchestrates the establishment of the necrotrophic pathogen lifestyle on wheat. *Scientific Reports*, *9*(1), 1–13. https://doi.org/10.1038/s41598-019-52444-7

Jones, D. A., Bertazzoni, S., Turo, C. J., Syme, R. A., & Hane, J. K. (2018). Bioinformatic prediction of plant–pathogenicity effector proteins of fungi. *Current Opinion in Microbiology*, *46*, 43–49. https://doi.org/10.1016/j.mib.2018.01.017

Kale, S. D., Gu, B., Capelluto, D. G. S., Dou, D., Feldman, E., Rumore, A., Arredondo, F. D., Hanlon, R., Fudal, I., Rouxel, T., Lawrence, C. B., Shan, W., & Tyler, B. M. (2010). External Lipid PI3P Mediates Entry of Eukaryotic Pathogen Effectors into Plant and Animal Host Cells. *Cell*, *142*(2), 284–295. https://doi.org/10.1016/j.cell.2010.06.008

Kamoun, S., Furzer, O., Jones, J. D. G., Judelson, H. S., Ali, G. S., Dalio, R. J. D., Roy, S. G., Schena, L., Zambounis, A., Panabières, F., Cahill, D., Ruocco, M., Figueiredo, A., Chen, X.-R., Hulvey, J., Stam, R., Lamour, K., Gijzen, M., Tyler, B. M., … Govers, F. (2015). The top 10 oomycete pathogens in molecular plant pathology. *Molecular Plant Pathology*, *16*(4), 413–434. https://doi.org/10.1111/mpp.12190

Kanja, C., & Hammond-Kosack, K. E. (2020). Proteinaceous effector discovery and characterization in filamentous plant pathogens. *Molecular Plant Pathology*. https://doi.org/10.1111/mpp.12980

Karimi Jashni, M., van der Burgt, A., Battaglia, E., Mehrabi, R., Collemare, J., & de Wit, P. J. G. M. (2020). Transcriptome and proteome analyses of proteases in biotroph fungal pathogen *Cladosporium fulvum*. *Journal of Plant Pathology*, *102*(2), 377–386. https://doi.org/10.1007/s42161-019-00433-0

Kariyawasam, G. K., Richards, J. K., Wyatt, N. A., Running, K., Xu, S. S., Liu, Z., Borowicz, P., Faris, J. D., & Friesen, T. L. (2021). The *Parastagonospora nodorum* necrotrophic effector SnTox5 targets the wheat gene Snn5 and facilitates entry into the leaf mesophyll. *BioRxiV*. https://doi.org/10.1101/2021.02.26.433117

Keller, S. M., McDermott, J. M., Pettway, R. E., Wolfe, M. S., & McDonald, B. A. (1997). Gene Flow and Sexual Reproduction in the Wheat Glume Blotch Pathogen *Phaeosphaeria nodorum* (Anamorph *Stagonospora nodorum*). *Phytopathology*, *87*(3), 353–358. https://doi.org/10.1094/PHYTO.1997.87.3.353

Kelly, A. C., & Ward, T. J. (2018). Population genomics of *Fusarium graminearum* reveals signatures of divergent evolution within a major cereal pathogen. *PLOS ONE*, *13*(3), e0194616. https://doi.org/10.1371/journal.pone.0194616

Klosterman, S. J., Subbarao, K. V., Kang, S., Veronese, P., Gold, S. E., Thomma, B. P. H. J., Chen, Z., Henrissat, B., Lee, Y.-H., Park, J., Garcia-Pedrajas, M. D., Barbara, D. J., Anchieta, A., de Jonge, R., Santhanam, P., Maruthachalam, K., Atallah, Z., Amyotte, S. G., Paz, Z., … Ma, L.-J. (2011). Comparative Genomics Yields Insights into Niche Adaptation of Plant Vascular Wilt Pathogens. *PLOS Pathogens*, *7*(7), e1002137. https://doi.org/10.1371/journal.ppat.1002137

Kourelis, J., van der Hoorn, R., & Sueldo, D. J. (2016). Decoy engineering: The next step in resistance breeding. *Trends in Plant Science*, *21*(5), 371–373. https://doi.org/10.1016/j.tplants.2016.04.001

Kroj, T., Chanclud, E., Michel-Romiti, C., Grand, X., & Morel, J.-B. (2016). Integration of decoy domains derived from protein targets of pathogen effectors into plant immune receptors is widespread. *New Phytologist*, *210*(2), 618–626. https://doi.org/10.1111/nph.13869

Kubicek, C. P., Starr, T. L., & Glass, N. L. (2014). Plant Cell Wall–Degrading Enzymes and Their Secretion in Plant-Pathogenic Fungi. *Annual Review of Phytopathology*, *52*(1), 427–451. https://doi.org/10.1146/annurev-phyto-102313-045831

Kulkarni, R. D., Kelkar, H. S., & Dean, R. A. (2003). An eight-cysteine-containing CFEM domain unique to a group of fungal membrane proteins. *Trends in Biochemical Sciences*, *28*(3), 118–121. https://doi.org/10.1016/S0968-0004(03)00025-2

Larkan, N. J., Lydiate, D. J., Parkin, I. A. P., Nelson, M. N., Epp, D. J., Cowling, W. A., Rimmer, S. R., & Borhan, M. H. (2012). The *Brassica napus* blackleg resistance gene LepR3 encodes a receptor-like protein triggered by the *Leptosphaeria maculans* effector AVRLM1. *New Phytologist*, *197*(2), 595–605. https://doi.org/10.1111/nph.12043

Le Cam, B., Sargent, D., Gouzy, J., Amselem, J., Bellanger, M.-N., Bouchez, O., Brown, S., Caffier, V., Gracia, M. D., Debuchy, R., Duvaux, L., Payen, T., Sannier, M., Shiller, J., Collemare, J., & Lemaire, C. (2019). Population Genome Sequencing of the Scab Fungal Species *Venturia inaequalis*, *Venturia pirina*, *Venturia aucupariae* and *Venturia asperata*. *G3: Genes, Genomes, Genetics*, *9*(8), 2405–2414. https://doi.org/10.1534/g3.119.400047

Lehmann, S., Serrano, M., L'Haridon, F., Tjamos, S. E., & Metraux, J.-P. (2015). Reactive oxygen species and plant resistance to fungal pathogens. *Phytochemistry*, *112*, 54–62. https://doi.org/10.1016/j.phytochem.2014.08.027

Lewis, D. H. (1973). Concepts in Fungal Nutrition and the Origin of Biotrophy. *Biological Reviews*, *48*(2), 261–277. https://doi.org/10.1111/j.1469-185X.1973.tb00982.x

Lin, Y.-M., Shih, S.-L., Lin, W.-C., Wu, J.-W., Chen, Y.-T., Hsieh, C.-Y., Guan, L.-C., Lin, L., & Cheng, C.-P. (2014). Phytoalexin biosynthesis genes are regulated and involved in plant response to *Ralstonia solanacearum* infection. *Plant Science*, *224*, 86–94. https://doi.org/10.1016/j.plantsci.2014.04.008

Liu, L., Xu, L., Jia, Q., Pan, R., Oelmüller, R., Zhang, W., & Wu, C. (2019). Arms race: Diverse effector proteins with conserved motifs. *Plant Signaling & Behavior*, *14*(2), 1557008. https://doi.org/10.1080/15592324.2018.1557008

Liu, Z. H., Faris, J. D., Meinhardt, S. W., Ali, S., Rasmussen, J. B., & Friesen, T. L. (2004). Genetic and Physical Mapping of a Gene Conditioning Sensitivity in Wheat to a Partially Purified Host-Selective Toxin Produced by *Stagonospora nodorum*. *Phytopathology*, *94*(10), 1056–1060. https://doi.org/10.1094/PHYTO.2004.94.10.1056

Liu, Z., Faris, J. D., Oliver, R. P., Tan, K.-C., Solomon, P. S., McDonald, M. C., McDonald, B. A., Nunez, A., Lu, S., Rasmussen, J. B., & Friesen, T. L. (2009). SnTox3 Acts in Effector Triggered Susceptibility to Induce Disease on Wheat Carrying the *Snn3* Gene. *PLOS Pathogens*, *5*(9), e1000581. https://doi.org/10.1371/journal.ppat.1000581

Liu, Z., Friesen, T. L., Ling, H., Meinhardt, S. W., Oliver, R. P., Rasmussen, J. B., & Faris, J. D. (2006). The Tsn1–ToxA interaction in the wheat–*Stagonospora nodorum* pathosystem parallels that of the wheat–tan spot system. *Genome*, *49*(10), 1265–1273. https://doi.org/10.1139/g06-088

Liu, Z., Zhang, Z., Faris, J. D., Oliver, R. P., Syme, R., McDonald, M. C., McDonald, B. A., Solomon, P. S., Lu, S., Shelver, W. L., Xu, S., & Friesen, T. L. (2012). The Cysteine Rich Necrotrophic Effector SnTox1 Produced by *Stagonospora nodorum* Triggers Susceptibility of Wheat Lines Harboring Snn1. *PLOS Pathogens*, *8*(1), e1002467. https://doi.org/10.1371/journal.ppat.1002467

Lo Presti, L., & Kahmann, R. (2017). How filamentous plant pathogen effectors are translocated to host cells. *Current Opinion in Plant Biology*, *38*, 19–24. https://doi.org/https://doi.org/10.1016/j.pbi.2017.04.005

Lu, S., & Edwards, M. C. (2016). Genome-Wide Analysis of Small Secreted Cysteine-Rich Proteins Identifies Candidate Effector Proteins Potentially Involved in *Fusarium graminearum*-Wheat Interactions. *Phytopathology*, *106*(2), 166–176. https://doi.org/10.1094/PHYTO-09-15-0215-R

Lu, S., Gillian Turgeon, B., & Edwards, M. C. (2015). A ToxA-like protein from *Cochliobolus heterostrophus* induces light-dependent leaf necrosis and acts as a virulence factor with host selectivity on maize. *Fungal Genetics and Biology*, *81*, 12–24. https://doi.org/10.1016/j.fgb.2015.05.013

Luna, E., Pastor, V., Robert, J., Flors, V., Mauch-Mani, B., & Ton, J. (2010). Callose Deposition: A Multifaceted Plant Defense Response. *Molecular Plant-Microbe Interactions*, *24*(2), 183–193. https://doi.org/10.1094/MPMI-07-10-0149

Luti, S., Caselli, A., Taiti, C., Bazihizina, N., Gonnelli, C., Mancuso, S., & Pazzagli, L. (2016). PAMP Activity of Cerato-Platanin during Plant Interaction: An -Omic Approach. *International Journal of Molecular Sciences*, *17*(6), 866. https://doi.org/10.3390/ijms17060866

Luti, S., Martellini, F., Bemporad, F., Mazzoli, L., Paoli, P., & Pazzagli, L. (2017). A single amino acid mutation affects elicitor and expansins-like activities of cerato-platanin, a non-catalytic fungal protein. *PLOS ONE*, *12*(5), e0178337. https://doi.org/10.1371/journal.pone.0178337

Ma, L.-J., van der Does, H. C., Borkovich, K. A., Coleman, J. J., Daboussi, M.-J., Di Pietro, A., Dufresne, M., Freitag, M., Grabherr, M., Henrissat, B., Houterman, P. M., Kang, S., Shim, W.-B., Woloshuk, C., Xie, X., Xu, J.-R., Antoniw, J., Baker, S. E., Bluhm, B. H., … Rep, M. (2010). Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. *Nature*, *464*(7287), 367–373. https://doi.org/10.1038/nature08850

Malagnac, F., Wendel, B., Goyon, C., Faugeron, G., Zickler, D., Rossignol, J.-L., Noyer-Weidner, M., Vollmayr, P., Trautner, T. A., & Walter, J. (1997). A Gene Essential for De Novo Methylation and Development in Ascobolus Reveals a Novel Type of Eukaryotic DNA Methyltransferase Structure. *Cell*, *91*(2), 281–290. https://doi.org/10.1016/S0092-8674(00)80410-9

Malinovsky, F. G., Fangel, J. U., & Willats, W. G. T. (2014). The role of the cell wall in plant immunity. *Frontiers in Plant Science*, *5*. https://doi.org/10.3389/fpls.2014.00178

Manning, V. A., Hamilton, S. M., Karplus, P. A., & Ciuffetti, L. M. (2008). The arg-gly-asp–containing, solvent-exposed loop of ptr ToxA is required for internalization. *Molecular Plant-Microbe Interactions*, *21*(3), 315–325. https://doi.org/10.1094/mpmi-21-3-0315

Manning, V. A., Hardison, L. K., & Ciuffetti, L. M. (2007). Ptr toxa interacts with a chloroplast-localized protein. *Molecular Plant-Microbe Interactions*, *20*(2), 168–177. https://doi.org/10.1094/MPMI-20-2-0168

Mansfield, J., Genin, S., Magori, S., Citovsky, V., Sriariyanum, M., Ronald, P., Dow, M., Verdier, V., Beer, S. V., Machado, M. A., Toth, I., Salmond, G., & Foster, G. D. (2012). Top 10 plant pathogenic bacteria in molecular plant pathology. *Molecular Plant Pathology*, *13*(6), 614–629. https://doi.org/10.1111/j.1364-3703.2012.00804.x

Marshall, R., Kombrink, A., Motteram, J., Loza-Reyes, E., Lucas, J., Hammond-Kosack, K. E., Thomma, B. P. H. J., & Rudd, J. J. (2011). Analysis of Two in Planta Expressed LysM Effector Homologs from the Fungus *Mycosphaerella graminicola* Reveals Novel Functional Properties and Varying Contributions to Virulence on Wheat. *Plant Physiology*, *156*(2), 756–769. https://doi.org/10.1104/pp.111.176347

Martinez, D. A., Oliver, B. G., Gräser, Y., Goldberg, J. M., Li, W., Martinez-Rossi, N. M., Monod, M., Shelest, E., Barton, R. C., Birch, E., Brakhage, A. A., Chen, Z., Gurr, S. J., Heiman, D., Heitman, J., Kosti, I., Rossi, A., Saif, S., Samalova, M., … White, T. C. (2012). Comparative Genome Analysis

of *Trichophyton rubrum* and Related Dermatophytes Reveals Candidate Genes Involved in Infection. *mBio*, *3*(5). https://doi.org/10.1128/mBio.00259-12

Martinez, J. P., Ottum, S. A., Ali, S., Francl, L. J., & Ciuffetti, L. M. (2001). Characterization of the *ToxB* Gene from *Pyrenophora tritici-repentis*. *Molecular Plant-Microbe Interactions*, *14*(5), 675–677. https://doi.org/10.1094/MPMI.2001.14.5.675

Marton, K., Flajšman, M., Radišek, S., Košmelj, K., Jakše, J., Javornik, B., & Berne, S. (2018). Comprehensive analysis of *Verticillium nonalfalfae* in silico secretome uncovers putative effector proteins expressed during hop invasion. *PLOS ONE*, *13*(6), e0198971. https://doi.org/10.1371/journal.pone.0198971

McClintock, B. (1941). The Stability of Broken Ends of Chromosomes in *Zea mays*. *Genetics*, *26*(2), 234–282.

McDonald, B. A., & Linde, C. (2002). The population genetics of plant pathogens and breeding strategies for durable resistance. *Euphytica*, *124*(2), 163–180. https://doi.org/10.1023/A:1015678432355

McDonald, M. C., Ahren, D., Simpfendorfer, S., Milgate, A., & Solomon, P. S. (2018). The discovery of the virulence gene *ToxA* in the wheat and barley pathogen *Bipolaris sorokiniana*. *Molecular Plant Pathology*, *19*(2), 432–439. https://doi.org/10.1111/mpp.12535

McDonald, M. C., Oliver, R. P., Friesen, T. L., Brunner, P. C., & McDonald, B. A. (2013). Global diversity and distribution of three necrotrophic effectors in *Phaeosphaeria nodorum* and related species. *New Phytologist*, *199*(1), 241–251. https://doi.org/10.1111/nph.12257

McDonald, M. C., Razavi, M., Friesen, T. L., Brunner, P. C., & McDonald, B. A. (2012). Phylogenetic and population genetic analyses of *Phaeosphaeria nodorum* and its close relatives indicate cryptic species and an origin in the Fertile Crescent. *Fungal Genetics and Biology*, *49*(11), 882–895. https://doi.org/10.1016/j.fgb.2012.08.001

McDonald, M. C., Taranto, A. P., Hill, E., Schwessinger, B., Liu, Z., Simpfendorfer, S., Milgate, A., & Solomon, P. S. (2019). Transposon-Mediated Horizontal Transfer of the Host-Specific Virulence Protein ToxA between Three Fungal Wheat Pathogens. *mBio*, *10*(5). https://doi.org/10.1128/mBio.01515-19

Menardo, F., Praz, C. R., Wyder, S., Ben-David, R., Bourras, S., Matsumae, H., McNally, K. E., Parlange, F., Riba, A., Roffler, S., Schaefer, L. K., Shimizu, K. K., Valenti, L., Zbinden, H., Wicker, T., & Keller, B. (2016). Hybridization of powdery mildew strains gives rise to pathogens on novel agricultural crop species. *Nature Genetics*, *48*(2), 201–205. https://doi.org/10.1038/ng.3485

Mesarich, C. H., Ökmen, B., Rovenich, H., Griffiths, S. A., Wang, C., Karimi Jashni, M., Mihajlovski, A., Collemare, J., Hunziker, L., Deng, C. H., van der Burgt, A., Beenen, H. G., Templeton, M. D., Bradshaw, R. E., & de Wit, P. J. G. M. (2018). Specific Hypersensitive Response–Associated Recognition of New Apoplastic Effectors from *Cladosporium fulvum* in Wild Tomato. *Molecular Plant-Microbe Interactions*, *31*(1), 145–162. https://doi.org/10.1094/MPMI-05-17-0114-FI

Miao, V. P., Covert, S. F., & VanEtten, H. D. (1991). A fungal gene for antibiotic resistance on a dispensable ("B") chromosome. *Science*, *254*(5039), 1773–1776. https://doi.org/10.1126/science.1763326

Milgroom, M. G., del Mar Jiménez-Gasco, M., Olivares-García, C., & Jiménez-Díaz, R. M. (2016). Clonal Expansion and Migration of a Highly Virulent, Defoliating Lineage of *Verticillium dahliae*. *Phytopathology*, *106*(9), 1038–1046. https://doi.org/10.1094/PHYTO-11-15-0300-R

Milne, T. J., Abbenante, G., Tyndall, J. D. A., Halliday, J., & Lewis, R. J. (2003). Isolation and characterization of a cone snail protease with homology to CRISP proteins of the pathogenesis-related protein superfamily. *The Journal of Biological Chemistry*, *278*(33), 31105–31110. https://doi.org/10.1074/jbc.M304843200

Molina, L., & Kahmann, R. (2007). An *Ustilago maydis* Gene Involved in H2O2 Detoxification Is Required for Virulence. *The Plant Cell*, *19*(7), 2293–2309. https://doi.org/10.1105/tpc.107.052332

Moolhuijzen, P., See, P. T., Hane, J. K., Shi, G., Liu, Z., Oliver, R. P., & Moffat, C. S. (2018). Comparative genomics of the wheat fungal pathogen *Pyrenophora tritici-repentis* reveals chromosomal variations and genome plasticity. *BMC Genomics*, *19*(1), 279. https://doi.org/10.1186/s12864-018-4680-3

Murphy, N. E., Loughman, R., Appels, R., Lagudah, E. S., & Jones, M. G. K. (2000). Genetic variability in a collection of *Stagonospora nodorum* isolates from Western Australia. *Australian Journal of Agricultural Research*, *51*(6), 679–684. https://doi.org/10.1071/ar99107

Murray, G. M., & Brennan, J. P. (2009). Estimating disease losses to the Australian wheat industry. *Australasian Plant Pathology*, *38*(6), 558–570. https://doi.org/10.1071/AP09053

Nasser, L., Weissman, Z., Pinsky, M., Amartely, H., Dvir, H., & Kornitzer, D. (2016). Structural basis of haem-iron acquisition by fungal pathogens. *Nature Microbiology*, *1*(11), 1–10. https://doi.org/10.1038/nmicrobiol.2016.156

Neil Cooley, R., & Caten, C. E. (1991). Variation in electrophoretic karyotype between strains of *Septoria nodorum*. *Molecular and General Genetics MGG*, *228*(1-2), 17–23. https://doi.org/10.1007/BF00282442

Newton, A. C., Fitt, B. D. L., Atkins, S. D., Walters, D. R., & Daniell, T. J. (2010). Pathogenesis, parasitism and mutualism in the trophic space of microbe–plant interactions. *Trends in Microbiology*, *18*(8), 365–373. https://doi.org/10.1016/j.tim.2010.06.002

Noguchi, M. T., Yasuda, N., & Fujita, Y. (2006). Evidence of Genetic Exchange by Parasexual Recombination and Genetic Analysis of Pathogenicity and Mating Type of Parasexual Recombinants in Rice Blast Fungus, *Magnaporthe oryzae*. *Phytopathology*, *96*(7), 746–750. https://doi.org/10.1094/PHYTO-96-0746

O'Connell, R. J., & Panstruga, R. (2006). Tête à tête inside a plant cell: Establishing compatibility between plants and biotrophic fungi and oomycetes. *New Phytologist*, *171*(4), 699–718. https://doi.org/10.1111/j.1469-8137.2006.01829.x

O'Connell, R. J., Thon, M. R., Hacquard, S., Amyotte, S. G., Kleemann, J., Torres, M. F., Damm, U., Buiate, E. A., Epstein, L., Alkan, N., Altmüller, J., Alvarado-Balderrama, L., Bauser, C. A., Becker, C., Birren, B. W., Chen, Z., Choi, J., Crouch, J. A., Duvick, J. P., … Vaillancourt, L. J. (2012). Lifestyle transitions in plant pathogenic Colletotrichum fungi deciphered by genome and transcriptome analyses. *Nature Genetics*, *44*(9), 1060–1065. https://doi.org/10.1038/ng.2372

Oerke, E.-C. (2006). Crop losses to pests. *The Journal of Agricultural Science*, *144*(1), 31–43. https://doi.org/10.1017/S0021859605005708

Ohm, R. A., Feau, N., Henrissat, B., Schoch, C. L., Horwitz, B. A., Barry, K. W., Condon, B. J., Copeland, A. C., Dhillon, B., Glaser, F., Hesse, C. N., Kosti, I., LaButti, K., Lindquist, E. A., Lucas, S., Salamov, A. A., Bradshaw, R. E., Ciuffetti, L., Hamelin, R. C., … Grigoriev, I. V. (2012). Diverse Lifestyles and Strategies of Plant Pathogenesis Encoded in the Genomes of Eighteen Dothideomycetes Fungi. *PLOS Pathogens*, *8*(12), e1003037. https://doi.org/10.1371/journal.ppat.1003037

Okuyama, Y., Kanzaki, H., Abe, A., Yoshida, K., Tamiru, M., Saitoh, H., Fujibe, T., Matsumura, H., Shenton, M., Galam, D. C., Undan, J., Ito, A., Sone, T., & Terauchi, R. (2011). A multifaceted genomics approach allows the isolation of the rice pia-blast resistance gene consisting of two adjacent NBS-LRR protein genes. *The Plant Journal*, *66*(3), 467–479. https://doi.org/10.1111/j.1365-313x.2011.04502.x

Oome, S., & Van den Ackerveken, G. (2014). Comparative and Functional Analysis of the Widely Occurring Family of Nep1-Like Proteins. *Molecular Plant-Microbe Interactions*, *27*(10), 1081–1094. https://doi.org/10.1094/MPMI-04-14-0118-R

Osbourn, A. E. (1996). Preformed Antimicrobial Compounds and Plant Defense against Fungal Attack. *The Plant Cell*, *8*(10), 1821–1831. https://doi.org/10.1105/tpc.8.10.1821

Pallaghy, P. K., Nielsen, K. J., Craik, D. J., & Norton, R. S. (1994). A common structural motif incorporating a cystine knot and a triple-stranded beta-sheet in toxic and inhibitory polypeptides. *Protein Science*, *3*(10), 1833–1839. https://doi.org/10.1002/pro.5560031022

Pantou, M. P., & Typas, M. A. (2005). Electrophoretic karyotype and gene mapping of the vascular wilt fungus *Verticillium dahliae*. *FEMS Microbiology Letters*, *245*(2), 213–220. https://doi.org/10.1016/j.femsle.2005.03.011

Parlange, F., Daverdin, G., Fudal, I., Kuhn, M.-L., Balesdent, M.-H., Blaise, F., Grezes-Besset, B., & Rouxel, T. (2009). *Leptosphaeria maculans* avirulence gene *AvrLm4-7* confers a dual recognition specificity by the *Rlm4* and *Rlm7* resistance genes of oilseed rape, and circumvents Rlm4-mediated recognition through a single amino acid change. *Molecular Microbiology*, *71*(4), 851–863. https://doi.org/10.1111/j.1365-2958.2008.06547.x

Pazzagli, L., Pantera, B., Carresi, L., Zoppi, C., Pertinhez, T. A., Spisni, A., Tegli, S., Scala, A., & Cappugi, G. (2006). Cerato-platanin, the first member of a new fungal protein family. *Cell Biochemistry and Biophysics*, *44*(3), 512–521. https://doi.org/10.1385/CBB:44:3:512

Pereira, D., McDonald, B. A., & Croll, D. (2020). The genetic architecture of emerging fungicide resistance in populations of a global wheat pathogen. *bioRxiv*, 2020.03.26.010199. https://doi.org/10.1101/2020.03.26.010199

Perlin, M. H., Amselem, J., Fontanillas, E., Toh, S. S., Chen, Z., Goldberg, J., Duplessis, S., Henrissat, B., Young, S., Zeng, Q., Aguileta, G., Petit, E., Badouin, H., Andrews, J., Razeeq, D., Gabaldón, T., Quesneville, H., Giraud, T., Hood, M. E., … Cuomo, C. A. (2015). Sex and parasites: Genomic and transcriptomic analysis of *Microbotryum lychnidis-dioicae*, the biotrophic and plant-castrating anther smut fungus. *BMC Genomics*, *16*(1), 461. https://doi.org/10.1186/s12864-015-1660-8

Petre, B., & Kamoun, S. (2014). How do filamentous pathogens deliver effector proteins into plant cells? *PLoS Biology*, *12*(2), e1001801. https://doi.org/10.1371/journal.pbio.1001801

Petre, B., Lorrain, C., Saunders, D. G., Win, J., Sklenar, J., Duplessis, S., & Kamoun, S. (2015). Rust fungal effectors mimic host transit peptides to translocate into chloroplasts. *Cellular Microbiology*, *18*(4), 453–465. https://doi.org/10.1111/cmi.12530

Phan, H. T. T., Rybak, K., Bertazzoni, S., Furuki, E., Dinglasan, E., Hickey, L. T., Oliver, R. P., & Tan, K.-C. (2018). Novel sources of resistance to Septoria nodorum blotch in the Vavilov wheat collection identified by genome-wide association studies. *Theoretical and Applied Genetics*, *131*(6), 1223–1238. https://doi.org/10.1007/s00122-018-3073-y

Phan, H. T. T., Rybak, K., Furuki, E., Breen, S., Solomon, P. S., Oliver, R. P., & Tan, K.-C. (2016). Differential effector gene expression underpins epistasis in a plant fungal disease. *The Plant Journal*, *87*(4), 343–354. https://doi.org/10.1111/tpj.13203

Plett, J. M., Daguerre, Y., Wittulsky, S., Vayssieres, A., Deveau, A., Melton, S. J., Kohler, A., Morrell-Falvey, J. L., Brun, A., Veneault-Fourrey, C., & Martin, F. (2014). Effector MiSSP7 of the mutualistic fungus *Laccaria bicolor* stabilizes the Populus JAZ6 protein and represses jasmonic acid (JA) responsive genes. *Proceedings of the National Academy of Sciences*, *111*(22), 8299–8304. https://doi.org/10.1073/pnas.1322671111

Plett, J. M., Kemppainen, M., Kale, S. D., Kohler, A., Legué, V., Brun, A., Tyler, B. M., Pardo, A. G., & Martin, F. (2011a). A secreted effector protein of *Laccaria bicolor* is required for symbiosis development. *Current Biology*, *21*(14), 1197–1203. https://doi.org/10.1016/j.cub.2011.05.033

Plett, J. M., Kemppainen, M., Kale, S. D., Kohler, A., Legué, V., Brun, A., Tyler, B. M., Pardo, A. G., & Martin, F. (2011b). A Secreted Effector Protein of Laccaria bicolor Is Required for Symbiosis Development. *Current Biology*, *21*(14), 1197–1203. https://doi.org/10.1016/j.cub.2011.05.033

Plissonneau, C., Benevenuto, J., Mohd-Assaad, N., Fouché, S., Hartmann, F. E., & Croll, D. (2017). Using Population and Comparative Genomics to Understand the Genetic Basis of Effector-Driven Fungal Pathogen Evolution. *Frontiers in Plant Science*, *8*. https://doi.org/10.3389/fpls.2017.00119

Plissonneau, C., Hartmann, F. E., & Croll, D. (2018). Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biology*, *16*(1), 5. https://doi.org/10.1186/s12915-017-0457-4

Pollet, A., Beliën, T., Fierens, K., Delcour, J. A., & Courtin, C. M. (2009). *Fusarium graminearum* xylanases show different functional stabilities, substrate specificities and inhibition sensitivities. *Enzyme and Microbial Technology*, *44*(4), 189–195. https://doi.org/10.1016/j.enzmictec.2008.12.005

Postic, G., Gracy, J., Périn, C., Chiche, L., & Gelly, J.-C. (2018). KNOTTIN: The database of inhibitor cystine knot scaffold after 10 years, toward a systematic structure modeling. *Nucleic Acids Research*, *46*(D1), D454–D458. https://doi.org/10.1093/nar/gkx1084

Prados-Rosales, R. C., Roldán-Rodríguez, R., Serena, C., López-Berges, M. S., Guarro, J., Martínez-del-Pozo, Á., & Di Pietro, A. (2012). A PR-1-like Protein of *Fusarium oxysporum* Functions in Virulence on Mammalian Hosts. *Journal of Biological Chemistry*, *287*(26), 21970–21979. https://doi.org/10.1074/jbc.M112.364034

Praz, C. R., Bourras, S., Zeng, F., Sánchez-Martín, J., Menardo, F., Xue, M., Yang, L., Roffler, S., Böni, R., Herren, G., McNally, K. E., Ben-David, R., Parlange, F., Oberhaensli, S., Flückiger, S., Schäfer, L. K., Wicker, T., Yu, D., & Keller, B. (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*(3), 1301–1314. https://doi.org/10.1111/nph.14372

Priest, S. J., Yadav, V., & Heitman, J. (2020). Advances in understanding the evolution of fungal genome architecture. *F1000Research*, *9*. https://doi.org/10.12688/f1000research.25424.1

Qi, T., Guo, J., Liu, P., He, F., Wan, C., Islam, M. A., Tyler, B. M., Kang, Z., & Guo, J. (2019). Stripe rust effector PstGSRE1 disrupts nuclear localization of ROS-promoting transcription factor TaLOL2 to defeat ROS-induced defense in wheat. *Molecular Plant*, *12*(12), 1624–1638. https://doi.org/10.1016/j.molp.2019.09.010

Quaedvlieg, W., Verkley, G. J. M., Shin, H. .-D., Barreto, R. W., Alfenas, A. C., Swart, W. J., Groenewald, J. Z., & Crous, P. W. (2013). Sizing up Septoria. *Studies in Mycology*, *75*, 307–390. https://doi.org/10.3114/sim0017

Qutob, D., Kemmerling, B., Brunner, F., Küfner, I., Engelhardt, S., Gust, A. A., Luberacki, B., Seitz, H. U., Stahl, D., Rauhut, T., Glawischnig, E., Schween, G., Lacombe, B., Watanabe, N., Lam, E., Schlichting, R., Scheel, D., Nau, K., Dodt, G., … Nürnberger, T. (2006). Phytotoxicity and Innate Immune Responses Induced by Nep1-Like Proteins. *The Plant Cell*, *18*(12), 3721–3744. https://doi.org/10.1105/tpc.106.044180

Rehmany, A. P., Gordon, A., Rose, L. E., Allen, R. L., Armstrong, M. R., Whisson, S. C., Kamoun, S., Tyler, B. M., Birch, P. R. J., & Beynon, J. L. (2005). Differential Recognition of Highly Divergent Downy Mildew Avirulence Gene Alleles by *RPP1* Resistance Genes from Two Arabidopsis Lines. *The Plant Cell*, *17*(6), 1839–1850. https://doi.org/10.1105/tpc.105.031807

Reid, I., O'Toole, N., Zabaneh, O., Nourzadeh, R., Dahdouli, M., Abdellateef, M., Gordon, P. M., Soh, J., Butler, G., Sensen, C. W., & Tsang, A. (2014). SnowyOwl: Accurate prediction of fungal genes by using RNA-Seq and homology information to select among ab initio models. *BMC Bioinformatics*, *15*(1), 229. https://doi.org/10.1186/1471-2105-15-229

Richards, J. K., Kariyawasam, G. K., Seneviratne, S., Wyatt, N. A., Xu, S. S., Liu, Z., Faris, J. D., & Friesen, T. L. (2021). A triple threat: The *Parastagonospora nodorum* SnTox267 effector exploits three distinct host genetic factors to cause disease in wheat. *New Phytologist*. https://doi.org/10.1111/nph.17601

Richards, J. K., Stukenbrock, E. H., Carpenter, J., Liu, Z., Cowger, C., Faris, J. D., & Friesen, T. L. (2019). Local adaptation drives the diversification of effectors in the fungal wheat pathogen *Parastagonospora nodorum* in the United States. *PLOS Genetics*, *15*(10), e1008223. https://doi.org/10.1371/journal.pgen.1008223

Richards, J. K., Wyatt, N. A., Liu, Z., Faris, J. D., & Friesen, T. L. (2018). Reference Quality Genome Assemblies of Three *Parastagonospora nodorum* Isolates Differing in Virulence on Wheat. *G3: Genes, Genomes, Genetics*, *8*(2), 393–399. https://doi.org/10.1534/g3.117.300462

Romero-Contreras, Y. J., Ramírez-Valdespino, C. A., Guzmán-Guzmán, P., Macías-Segoviano, J. I., Villagómez-Castro, J. C., & Olmedo-Monfil, V. (2019). Tal6 From *Trichoderma atroviride* Is a LysM Effector Involved in Mycoparasitism and Plant Association. *Frontiers in Microbiology*, *10*. https://doi.org/10.3389/fmicb.2019.02231

Rooney, H. C. E., van't Klooster, J. W., van der Hoorn, R. A. L., Joosten, M. H. A. J., Jones, J. D. G., & de Wit, P. J. G. M. (2005). Cladosporium Avr2 Inhibits Tomato Rcr3 Protease Required for Cf-2-Dependent Disease Resistance. *Science*, *308*(5729), 1783–1786. https://doi.org/10.1126/science.1111404

Röpenack, E. v., Parr, A., & Schulze-Lefert, P. (1998). Structural Analyses and Dynamics of Soluble and Cell Wall-bound Phenolics in a Broad Spectrum Resistance to the Powdery Mildew Fungus in Barley. *Journal of Biological Chemistry*, *273*(15), 9013–9022. https://doi.org/10.1074/jbc.273.15.9013

Rose, J. K. C., Ham, K.-S., Darvill, A. G., & Albersheim, P. (2002). Molecular cloning and characterization of glucanase inhibitor proteins: Coevolution of a counterdefense mechanism by plant pathogens. *The Plant Cell*, *14*(6), 1329–1345. https://doi.org/10.1105/tpc.002253

Rouli, L., Merhej, V., Fournier, P.-E., & Raoult, D. (2015). The bacterial pangenome as a new tool for analysing pathogenic bacteria. *New Microbes and New Infections*, *7*, 72–85. https://doi.org/10.1016/j.nmni.2015.06.005

Rouxel, T., Grandaubert, J., Hane, J. K., Hoede, C., van de Wouw, A. P., Couloux, A., Dominguez, V., Anthouard, V., Bally, P., Bourras, S., Cozijnsen, A. J., Ciuffetti, L. M., Degrave, A., Dilmaghani, A., Duret, L., Fudal, I., Goodwin, S. B., Gout, L., Glaser, N., … Howlett, B. J. (2011). Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. *Nature Communications*, *2*(1), 202. https://doi.org/10.1038/ncomms1189

Rouxel, T., Penaud, A., Pinochet, X., Brun, H., Gout, L., Delourme, R., Schmit, J., & Balesdent, M.-H. (2003). A 10-year Survey of Populations of *Leptosphaeria maculans* in France Indicates a Rapid Adaptation Towards the *Rlm1* Resistance Gene of Oilseed Rape. *European Journal of Plant Pathology*, *109*(8), 871–881. https://doi.org/10.1023/A:1026189225466

Samuel, M., Bleackley, M., Anderson, M., & Mathivanan, S. (2015). Extracellular vesicles including exosomes in cross kingdom regulation: A viewpoint from plant-fungal interactions. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.00766

Saunders, D. G. O., Win, J., Cano, L. M., Szabo, L. J., Kamoun, S., & Raffaele, S. (2012). Using Hierarchical Clustering of Secreted Protein Families to Classify and Rank Candidate Effectors of Rust Fungi. *PLOS ONE*, *7*(1), e29847. https://doi.org/10.1371/journal.pone.0029847

Saupe, S. J. (2000). Molecular Genetics of Heterokaryon Incompatibility in Filamentous Ascomycetes. *Microbiology and Molecular Biology Reviews*, *64*(3), 489–502. https://doi.org/10.1128/MMBR.64.3.489-502.2000

Savary, S., McRoberts, N., Esker, P. D., Willocquet, L., & Teng, P. S. (2017). Production situations as drivers of crop health: Evidence and implications. *Plant Pathology*, *66*(6), 867–876. https://doi.org/10.1111/ppa.12659

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., & Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nature Ecology & Evolution*, *3*(3), 430–439. https://doi.org/10.1038/s41559-018-0793-y

Schaart, J. G., van de Wiel, C. C. M., Lotz, L. A. P., & Smulders, M. J. M. (2016). Opportunities for Products of New Plant Breeding Techniques. *Trends in Plant Science*, *21*(5), 438–449. https://doi.org/10.1016/j.tplants.2015.11.006

Schäfer, W., Straney, D., Ciuffetti, L., van Etten, H. D., & Yoder, O. C. (1989). One Enzyme Makes a Fungal Pathogen, But Not a Saprophyte, Virulent on a New Host Plant. *Science*, *246*(4927), 247–249. https://doi.org/10.1126/science.246.4927.247

Schmidt, R. (2000). Synteny: Recent advances and future prospects. *Current Opinion in Plant Biology*, *3*(2), 97–102. https://doi.org/10.1016/S1369-5266(99)00048-5

Schmidt, S. M., Lukasiewicz, J., Farrer, R., Dam, P. v., Bertoldo, C., & 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*(1), 307–318. https://doi.org/10.1111/nph.13584

Scholthof, K.-B. G., Adkins, S., Czosnek, H., Palukaitis, P., Jacquot, E., Hohn, T., Hohn, B., Saunders, K., Candresse, T., Ahlquist, P., Hemenway, C., & Foster, G. D. (2011). Top 10 plant viruses in molecular plant pathology. *Molecular Plant Pathology*, *12*(9), 938–954. https://doi.org/10.1111/j.1364-3703.2011.00752.x

Schornack, S., Damme, M. v., Bozkurt, T. O., Cano, L. M., Smoker, M., Thines, M., Gaulin, E., Kamoun, S., & Huitema, E. (2010). Ancient class of translocated oomycete effectors targets the host nucleus. *Proceedings of the National Academy of Sciences*, *107*(40), 17421–17426. https://doi.org/10.1073/pnas.1008491107

Schotanus, K., Soyer, J. L., Connolly, L. R., Grandaubert, J., Happel, P., Smith, K. M., Freitag, M., & Stukenbrock, E. H. (2015). Histone modifications rather than the novel regional centromeres of *Zymoseptoria tritici* distinguish core and accessory chromosomes. *Epigenetics & Chromatin*, *8*(1), 41. https://doi.org/10.1186/s13072-015-0033-5

Selker, E. U. (1990). Premeiotic instability of repeated sequences in *Neurospora crassa*. *Annual Review of Genetics*, *24*(1), 579–613. https://doi.org/10.1146/annurev.ge.24.120190.003051

Shabab, M., Shindo, T., Gu, C., Kaschani, F., Pansuriya, T., Chintha, R., Harzen, A., Colby, T., Kamoun, S., & van der Hoorn, R. A. L. (2008). Fungal Effector Protein AVR2 Targets Diversifying Defense-Related Cys Proteases of Tomato. *The Plant Cell*, *20*(4), 1169–1183. https://doi.org/10.1105/tpc.107.056325

Shen, Q.-H., Saijo, Y., Mauch, S., Biskup, C., Bieri, S., Keller, B., Seki, H., Ulker, B., Somssich, I. E., & Schulze-Lefert, P. (2007). Nuclear activity of MLA immune receptors links isolate-specific and basal disease-resistance responses. *Science*, *315*(5815), 1098–1103. https://doi.org/10.1126/science.1136372

Sherwood, R. K., & Bennett, R. J. (2009). Fungal Meiosis and Parasexual Reproduction – Lessons from Pathogenic Yeast. *Current opinion in microbiology*, *12*(6), 599–607. https://doi.org/10.1016/j.mib.2009.09.005

Shi, G., Friesen, T. L., Saini, J., Xu, S. S., Rasmussen, J. B., & Faris, J. D. (2015). The Wheat *Snn7* Gene Confers Susceptibility on Recognition of the *Parastagonospora nodorum* Necrotrophic Effector SnTox7. *The Plant Genome*, *8*(2), plantgenome2015.02.0007. https://doi.org/10.3835/plantgenome2015.02.0007

Shi, G., Zhang, Z., Friesen, T. L., Raats, D., Fahima, T., Brueggeman, R. S., Lu, S., Trick, H. N., Liu, Z., Chao, W., Frenkel, Z., Xu, S. S., Rasmussen, J. B., & Faris, J. D. (2016). The hijacking of a receptor kinase–driven pathway by a wheat fungal pathogen leads to disease. *Science Advances*, *2*(10), e1600822. https://doi.org/10.1126/sciadv.1600822

Smith, K. M., Galazka, J. M., Phatale, P. A., Connolly, L. R., & Freitag, M. (2012). Centromeres of filamentous fungi. *Chromosome Research*, *20*(5), 635–656. https://doi.org/10.1007/s10577-012-9290-3

Smith, K. M., Phatale, P. A., Sullivan, C. M., Pomraning, K. R., & Freitag, M. (2011). Heterochromatin is required for normal distribution of *Neurospora crassa* CenH3. *Molecular and Cellular Biology*, *31*(12), 2528–2542. https://doi.org/10.1128/MCB.01285-10

Solomon, P. S., Lowe, R. G. T., Tan, K.-C., Waters, O. D. C., & Oliver, R. P. (2006). *Stagonospora nodorum*: Cause of stagonospora nodorum blotch of wheat. *Molecular Plant Pathology*, *7*(3), 147–156. https://doi.org/10.1111/j.1364-3703.2006.00326.x

Sommerhalder, R. J., McDonald, B. A., Mascher, F., & Zhan, J. (2011). Effect of hosts on competition among clones and evidence of differential selection between pathogenic and saprophytic phases in experimental populations of the wheat pathogen *Phaeosphaeria nodorum*. *BMC Evolutionary Biology*, *11*(1), 188. https://doi.org/10.1186/1471-2148-11-188

Sommerhalder, R. J., McDonald, B. A., & Zhan, J. (2006). The Frequencies and Spatial Distribution of Mating Types in *Stagonospora nodorum* Are Consistent with Recurring Sexual Reproduction. *Phytopathology*, *96*(3), 234–239. https://doi.org/10.1094/PHYTO-96-0234

Song, J., Win, J., Tian, M., Schornack, S., Kaschani, F., Ilyas, M., van der Hoorn, R. A. L., & Kamoun, S. (2009). Apoplastic effectors secreted by two unrelated eukaryotic plant pathogens target the tomato defense protease Rcr3. *Proceedings of the National Academy of Sciences*, *106*(5), 1654–1659. https://doi.org/10.1073/pnas.0809201106

Spanu, P. D. (2017). Cereal immunity against powdery mildews targets RNase-Like Proteins associated with Haustoria (RALPH) effectors evolved from a common ancestral gene. *New Phytologist*, *213*(3), 969–971. https://doi.org/10.1111/nph.14386

Sperschneider, J., Catanzariti, A.-M., DeBoer, K., Petre, B., Gardiner, D. M., Singh, K. B., Dodds, P. N., & Taylor, J. M. (2017). LOCALIZER: Subcellular localization prediction of both plant and effector proteins in the plant cell. *Scientific Reports*, *7*(1), 1–14. https://doi.org/10.1038/srep44598

Sperschneider, J., Dodds, P. N., Gardiner, D. M., Singh, K. B., & Taylor, J. M. (2018). Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. *Molecular Plant Pathology*, *19*(9), 2094–2110. https://doi.org/10.1111/mpp.12682

Sperschneider, J., Dodds, P. N., Singh, K. B., & Taylor, J. M. (2018). ApoplastP: Prediction of effectors and plant proteins in the apoplast using machine learning. *New Phytologist*, *217*(4), 1764–1778. https://doi.org/10.1111/nph.14946

Sperschneider, J., Gardiner, D. M., Dodds, P. N., Tini, F., Covarelli, L., Singh, K. B., Manners, J. M., & Taylor, J. M. (2016). EffectorP: Predicting fungal effector proteins from secretomes using machine learning. *New Phytologist*, *210*(2), 743–761. https://doi.org/10.1111/nph.13794

Stam, R., Jupe, J., Howden, A. J. M., Morris, J. A., Boevink, P. C., Hedley, P. E., & Huitema, E. (2013). Identification and Characterisation CRN Effectors in *Phytophthora capsici* Shows Modularity and Functional Diversity. *PLOS ONE*, *8*(3), e59517. https://doi.org/10.1371/journal.pone.0059517

Stergiopoulos, I., De Kock, M. J. D., Lindhout, P., & de Wit, P. J. G. M. (2007). Allelic Variation in the Effector Genes of the Tomato Pathogen *Cladosporium fulvum* Reveals Different Modes of Adaptive Evolution. *Molecular Plant-Microbe Interactions*, *20*(10), 1271–1283. https://doi.org/10.1094/MPMI-20-10-1271

Stergiopoulos, I., & de Wit, P. J. (2009). Fungal Effector Proteins. *Annual Review of Phytopathology*, *47*(1), 233–263. https://doi.org/10.1146/annurev.phyto.112408.132637

Stergiopoulos, I., Kourmpetis, Y. A. I., Slot, J. C., Bakker, F. T., De Wit, P. J. G. M., & Rokas, A. (2012). In Silico Characterization and Molecular Evolutionary Analysis of a Novel Superfamily of Fungal Effector Proteins. *Molecular Biology and Evolution*, *29*(11), 3371–3384. https://doi.org/10.1093/molbev/mss143

Strom, N. B., & Bushley, K. E. (2016). Two genomes are better than one: History, genetics, and biotechnological applications of fungal heterokaryons. *Fungal Biology and Biotechnology*, *3*(1), 4. https://doi.org/10.1186/s40694-016-0022-x

Stukenbrock, E. H., Banke, S., & McDonald, B. A. (2006). Global migration patterns in the fungal wheat pathogen *Phaeosphaeria nodorum*. *Molecular Ecology*, *15*(10), 2895–2904. https://doi.org/10.1111/j.1365-294X.2006.02986.x

Syme, R. A., Tan, K.-C., Hane, J. K., Dodhia, K., Stoll, T., Hastie, M., Furuki, E., Ellwood, S. R., Williams, A. H., Tan, Y.-F., Testa, A. C., Gorman, J. J., & Oliver, R. P. (2016). Comprehensive Annotation of the *Parastagonospora nodorum* Reference Genome Using Next-Generation Genomics, Transcriptomics and Proteogenomics. *PLOS ONE*, *11*(2), e0147221. https://doi.org/10.1371/journal.pone.0147221

Syme, R. A., Tan, K.-C., Rybak, K., Friesen, T. L., McDonald, B. A., Oliver, R. P., & Hane, J. K. (2018). Pan-Parastagonospora Comparative Genome Analysis—Effector Prediction and Genome Evolution. *Genome Biology and Evolution*, *10*(9), 2443–2457. https://doi.org/10.1093/gbe/evy192

Syme, R. A., Hane, J. K., Friesen, T. L., & Oliver, R. P. (2013). Resequencing and Comparative Genomics of *Stagonospora nodorum*: Sectional Gene Absence and Effector Discovery. *G3: Genes, Genomes, Genetics*, *3*(6), 959–969. https://doi.org/10.1534/g3.112.004994

Talas, F., & McDonald, B. A. (2015). Genome-wide analysis of *Fusarium graminearum* field populations reveals hotspots of recombination. *BMC Genomics*, *16*(1), 996. https://doi.org/10.1186/s12864-015-2166-0

Talbot, N. J., Salch, Y. P., Ma, M., & Hamer, J. E. (1993). Karyotypic Variation within Clonal Lineages of the Rice Blast Fungus, *Magnaporthe grisea*. *Applied and Environmental Microbiology*, *59*(2), 585–593. Retrieved August 31, 2020, from https://aem.asm.org/content/59/2/585

Tam, J., Wang, S., Wong, K., & Tan, W. (2015). Antimicrobial Peptides from Plants. *Pharmaceuticals*, *8*(4), 711–757. https://doi.org/10.3390/ph8040711

Tan, K.-C., Ferguson-Hunt, M., Rybak, K., Waters, O. D. C., Stanley, W. A., Bond, C. S., Stukenbrock, E. H., Friesen, T. L., Faris, J. D., McDonald, B. A., & Oliver, R. P. (2012). Quantitative Variation in Effector Activity of ToxA Isoforms from *Stagonospora nodorum* and *Pyrenophora tritici-repentis*. *Molecular Plant-Microbe Interactions*, *25*(4), 515–522. https://doi.org/10.1094/MPMI-10-11-0273

Tan, K.-C., Heazlewood, J. L., Millar, A. H., Oliver, R. P., & Solomon, P. S. (2009). Proteomic identification of extracellular proteins regulated by the Gna1 Gα subunit in *Stagonospora nodorum*. *Mycological Research*, *113*(5), 523–531. https://doi.org/10.1016/j.mycres.2009.01.004

Tan, K.-C., Phan, H. T. T., Rybak, K., John, E., Chooi, Y. H., Solomon, P. S., & Oliver, R. P. (2015). Functional redundancy of necrotrophic effectors – consequences for exploitation for breeding. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.00501

Testa, A. C., Hane, J. K., Ellwood, S. R., & Oliver, R. P. (2015). CodingQuarry: Highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics*, *16*(1), 170. https://doi.org/10.1186/s12864-015-1344-4

Testa, A. C., Oliver, R. P., & Hane, J. K. (2016). OcculterCut: A Comprehensive Survey of AT-Rich Regions in Fungal Genomes. *Genome Biology and Evolution*, *8*(6), 2044–2064. https://doi.org/10.1093/gbe/evw121

Thomma, B. P. H. J., Nürnberger, T., & Joosten, M. H. A. J. (2011). Of PAMPs and Effectors: The Blurred PTI-ETI Dichotomy. *The Plant Cell*, *23*(1), 4–15. https://doi.org/10.1105/tpc.110.082602

Thrall, P. H., Barrett, L. G., Dodds, P. N., & Burdon, J. J. (2016). Epidemiological and Evolutionary Outcomes in Gene-for-Gene and Matching Allele Models. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.01084

Thrower, L. B. (1966). Terminology for Plant Parasites. *Journal of Phytopathology*, *56*(3), 258–259. https://doi.org/10.1111/j.1439-0434.1966.tb02261.x

Tian, M., Benedetti, B., & Kamoun, S. (2005). A Second Kazal-Like Protease Inhibitor from *Phytophthora infestans* Inhibits and Interacts with the Apoplastic Pathogenesis-Related Protease P69B of Tomato. *Plant Physiology*, *138*(3), 1785–1793. https://doi.org/10.1104/pp.105.061226

Tian, M., Huitema, E., Da Cunha, L., Torto-Alalibo, T., & 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*(25), 26370–26377. https://doi.org/10.1074/jbc.M400941200

Tian, M., Win, J., Song, J., van der Hoorn, R., van der Knaap, E., & Kamoun, S. (2007). A *Phytophthora infestans* Cystatin-Like Protein Targets a Novel Tomato Papain-Like Apoplastic Protease. *Plant Physiology*, *143*(1), 364–377. https://doi.org/10.1104/pp.106.090050

Tomas, A., & Bockus, W. W. (1987). Cultivar-specific toxicity of culture filtrates of *Pyrenophora tritici-repentis*. *Phytopathology*, *77*(9), 1337–1340.

Tomas, A., Feng, G. H., Reeck, G. R., Bockus, W. W., & Leach, J. E. (1990). Purification of a cultivar-specific toxin from *Pyrenophora tritici-repentis*, causal agent of tan spot of wheat. *Molecular Plant-Microbe Interactions*, *3*(4), 221–224. Retrieved September 3, 2020, from https://www.cabdirect.org/cabdirect/abstract/19901146455

Tsuda, K., & Katagiri, F. (2010). Comparing signaling mechanisms engaged in pattern-triggered and effector-triggered immunity. *Current Opinion in Plant Biology*, *13*(4), 459–465. https://doi.org/10.1016/j.pbi.2010.04.006

United Nations, P. D., Department of Economic and Social Affairs. (2019). *World Population Prospects 2019: Highlights. ST/ESA/SER.A/423*. Retrieved August 26, 2020, from https://population.un.org/wpp/Publications/Files/WPP2019_Highlights.pdf

Van de Wouw, A. P., Lowe, R. G. T., Elliott, C. E., Dubois, D. J., & Howlett, B. J. (2014). An avirulence gene, *AvrLmJ1*, from the blackleg fungus, *Leptosphaeria maculans*, confers avirulence to *Brassica juncea* cultivars. *Molecular Plant Pathology*, *15*(5), 523–530. https://doi.org/10.1111/mpp.12105

van den Burg, H. A., Harrison, S. J., Joosten, M. H. A. J., Vervoort, J., & de Wit, P. J. G. M. (2006). *Cladosporium fulvum* Avr4 Protects Fungal Cell Walls Against Hydrolysis by Plant Chitinases Accumulating During Infection. *Molecular Plant-Microbe Interactions*, *19*(12), 1420–1430. https://doi.org/10.1094/MPMI-19-1420

van der Hoorn, R. A., & Kamoun, S. (2008). From guard to decoy: A new model for perception of plant pathogen effectors. *The Plant Cell*, *20*(8), 2009–2017. https://doi.org/10.1105/tpc.108.060194

van Esse, H. P., Bolton, M. D., Stergiopoulos, I., de Wit, P. J. G. M., & Thomma, B. P. H. J. (2007). The Chitin-Binding *Cladosporium fulvum* Effector Protein Avr4 Is a Virulence Factor. *Molecular Plant-Microbe Interactions*, *20*(9), 1092–1101. https://doi.org/10.1094/MPMI-20-9-1092

van Schie, C. C., & Takken, F. L. (2014). Susceptibility genes 101: How to be a good host. *Annual Review of Phytopathology*, *52*(1), 551–581. https://doi.org/10.1146/annurev-phyto-102313-045854

van Wyk, S., Wingfield, B. D., de Vos, L., Santana, Q. C., Van der Merwe, N. A., & Steenkamp, E. T. (2018). Multiple independent origins for a subtelomeric locus associated with growth rate in *Fusarium circinatum*. *IMA Fungus*, *9*(1), 27–36. https://doi.org/10.5598/imafungus.2018.09.01.03

Voß, S., Betz, R., Heidt, S., Corradi, N., & Requena, N. (2018). RiCRN1, a Crinkler Effector From the Arbuscular Mycorrhizal Fungus *Rhizophagus irregularis*, Functions in Arbuscule Development. *Frontiers in Microbiology*, *9*. https://doi.org/10.3389/fmicb.2018.02068

Walkowiak, S., Rowland, O., Rodrigue, N., & Subramaniam, R. (2016). Whole genome sequencing and comparative genomics of closely related Fusarium Head Blight fungi: *Fusarium graminearum*, *F. meridionale* and *F. asiaticum*. *BMC Genomics*, *17*(1), 1014. https://doi.org/10.1186/s12864-016-3371-1

Wang, L., Sun, Y., Sun, X., Yu, L., Xue, L., He, Z., Huang, J., Tian, D., Hurst, L. D., & Yang, S. (2020). Repeat-induced point mutation in *Neurospora crassa* causes the highest known mutation rate and mutational burden of any cellular life. *Genome Biology*, *21*(1), 142. https://doi.org/10.1186/s13059-020-02060-w

Wawra, S., Trusch, F., Matena, A., Apostolakis, K., Linne, U., Zhukov, I., Stanek, J., Koźmiński, W., Davidson, I., Secombes, C. J., Bayer, P., & West, P. v. (2017). The RxLR Motif of the Host Targeting Effector AVR3a of *Phytophthora infestans* Is Cleaved before Secretion. *The Plant Cell*, *29*(6), 1184–1195. https://doi.org/10.1105/tpc.16.00552

Weiberg, A., Wang, M., Lin, F.-M., Zhao, H., Zhang, Z., Kaloshian, I., Huang, H.-D., & Jin, H. (2013). Fungal Small RNAs Suppress Plant Immunity by Hijacking Host RNA Interference Pathways. *Science*, *342*(6154), 118–123. https://doi.org/10.1126/science.1239705

Whisson, S. C., Boevink, P. C., Moleleki, L., Avrova, A. O., Morales, J. G., Gilroy, E. M., Armstrong, M. R., Grouffaud, S., West, P. v., Chapman, S., Hein, I., Toth, I. K., Pritchard, L., & Birch, P. R. J. (2007). A translocation signal for delivery of oomycete effector proteins into host plant cells. *Nature*, *450*(7166), 115–118. https://doi.org/10.1038/nature06203

Williams, A. H., Sharma, M., Thatcher, L. F., Azam, S., Hane, J. K., Sperschneider, J., Kidd, B. N., Anderson, J. P., Ghosh, R., Garg, G., Lichtenzveig, J., Kistler, H. C., Shea, T., Young, S., Buck, S.-A. G., Kamphuis, L. G., Saxena, R., Pande, S., Ma, L.-J., … Singh, K. B. (2016). Comparative genomics and prediction of conditionally dispensable sequences in legume–infecting *Fusarium oxysporum* formae speciales facilitates identification of candidate effectors. *BMC Genomics*, *17*(1), 191. https://doi.org/10.1186/s12864-016-2486-8

Win, J., Krasileva, K. V., Kamoun, S., Shirasu, K., Staskawicz, B. J., & Banfield, M. J. (2012). Sequence Divergent RXLR Effectors Share a Structural Fold Conserved across Plant Pathogenic Oomycete Species. *PLOS Pathogens*, *8*(1), e1002400. https://doi.org/10.1371/journal.ppat.1002400

Wittstock, U., & Gershenzon, J. (2002). Constitutive plant toxins and their role in defense against herbivores and pathogens. *Current Opinion in Plant Biology*, *5*(4), 300–307. https://doi.org/10.1016/S1369-5266(02)00264-9

Wu, J., Kou, Y., Bao, J., Li, Y., Tang, M., Zhu, X., Ponaya, A., Xiao, G., Li, J., Li, C., Song, M.-Y., Cumagun, C. J. R., Deng, Q., Lu, G., Jeon, J.-S., Naqvi, N. I., & Zhou, B. (2015). Comparative genomics

identifies the *Magnaporthe oryzae* avirulence effector AvrPi9 that triggers Pi9-mediated blast resistance in rice. *New Phytologist*, *206*(4), 1463–1475. https://doi.org/10.1111/nph.13310

Wyatt, N. A., Richards, J. K., Brueggeman, R. S., & Friesen, T. L. (2020). A Comparative Genomic Analysis of the Barley Pathogen *Pyrenophora teres* f. *teres* Identifies Subtelomeric Regions as Drivers of Virulence. *Molecular Plant-Microbe Interactions*, *33*(2), 173–188. https://doi.org/10.1094/MPMI-05-19-0128-R

Xue, M., Yang, J., Li, Z., Hu, S., Yao, N., Dean, R. A., Zhao, W., Shen, M., Zhang, H., Li, C., Liu, L., Cao, L., Xu, X., Xing, Y., Hsiang, T., Zhang, Z., Xu, J.-R., & Peng, Y.-L. (2012). Comparative Analysis of the Genomes of Two Field Isolates of the Rice Blast Fungus *Magnaporthe oryzae*. *PLOS Genetics*, *8*(8), e1002869. https://doi.org/10.1371/journal.pgen.1002869

Yadav, V., Sreekumar, L., Guin, K., & Sanyal, K. (2018). Five pillars of centromeric chromatin in fungal pathogens. *PLoS Pathogens*, *14*(8). https://doi.org/10.1371/journal.ppat.1007150

Yadav, V., Yang, F., Reza, M. H., Liu, S., Valent, B., Sanyal, K., & Naqvi, N. I. (2019). Cellular Dynamics and Genomic Identity of Centromeres in Cereal Blast Fungus. *mBio*, *10*(4). https://doi.org/10.1128/mBio.01581-19

Yang, G., Tang, L., Gong, Y., Xie, J., Fu, Y., Jiang, D., Li, G., Collinge, D. B., Chen, W., & Cheng, J. (2018). A cerato-platanin protein SsCP1 targets plant PR1 and contributes to virulence of *Sclerotinia sclerotiorum*. *New Phytologist*, *217*(2), 739–755. https://doi.org/10.1111/nph.14842

Yoshida, K., Saitoh, H., Fujisawa, S., Kanzaki, H., Matsumura, H., Yoshida, K., Tosa, Y., Chuma, I., Takano, Y., Win, J., Kamoun, S., & Terauchi, R. (2009). Association Genetics Reveals Three Novel Avirulence Genes from the Rice Blast Fungal Pathogen *Magnaporthe oryzae*. *The Plant Cell*, *21*(5), 1573–1591. https://doi.org/10.1105/tpc.109.066324

Yu, X., Feng, B., He, P., & Shan, L. (2017). From Chaos to Harmony: Responses and Signaling upon Microbial Pattern Recognition. *Annual Review of Phytopathology*, *55*(1), 109–137. https://doi.org/10.1146/annurev-phyto-080516-035649

Zamith-Miranda, D., Nimrichter, L., Rodrigues, M., & Nosanchuk, J. (2018). Fungal extracellular vesicles: Modulating host–pathogen interactions by both the fungus and the host. *Microbes and Infection*, *20*(9-10), 501–504. https://doi.org/10.1016/j.micinf.2018.01.011

Zeng, L., Velásquez, A. C., Munkvold, K. R., Zhang, J., & Martin, G. B. (2012). A tomato LysM receptor-like kinase promotes immunity and its kinase activity is inhibited by AvrPtoB. *The Plant Journal*, *69*(1), 92–103. https://doi.org/10.1111/j.1365-313X.2011.04773.x

Zhang, D., Burroughs, A. M., Vidal, N. D., Iyer, L. M., & Aravind, L. (2016). Transposons to toxins: The provenance, architecture and diversification of a widespread class of eukaryotic effectors. *Nucleic Acids Research*, *44*(8), 3513–3533. https://doi.org/10.1093/nar/gkw221

Zhang, L., Ni, H., Du, X., Wang, S., Ma, X.-W., Nürnberger, T., Guo, H.-S., & Hua, C. (2017). The verticillium-specific protein VdSCP7 localizes to the plant nucleus and modulates immunity to fungal infections. *New Phytologist*, *215*(1), 368–381. https://doi.org/10.1111/nph.14537

Zhang, M., Xie, S., Zhao, Y., Meng, X., Song, L., Feng, H., & Huang, L. (2019). Hce2 domain-containing effectors contribute to the full virulence of *Valsa mali* in a redundant manner. *Molecular Plant Pathology*, *20*(6), 843–856. https://doi.org/10.1111/mpp.12796

Zhang, Z., Friesen, T. L., Xu, S. S., Shi, G., Liu, Z., Rasmussen, J. B., & Faris, J. D. (2011). Two putatively homoeologous wheat genes mediate recognition of SnTox3 to confer effector-triggered susceptibility to *Stagonospora nodorum*. *The Plant Journal*, *65*(1), 27–38. https://doi.org/10.1111/j.1365-313X.2010.04407.x

Zhang, Z.-N., Wu, Q.-Y., Zhang, G.-Z., Zhu, Y.-Y., Murphy, R. W., Liu, Z., & Zou, C.-G. (2015). Systematic analyses reveal uniqueness and origin of the CFEM domain in fungi. *Scientific Reports*, *5*(1), 13032. https://doi.org/10.1038/srep13032

Zhong, Z., Marcel, T. C., Hartmann, F. E., Ma, X., Plissonneau, C., Zala, M., Ducasse, A., Confais, J., Compain, J., Lapalu, N., Amselem, J., McDonald, B. A., Croll, D., & Palma-Guerrero, J. (2017). A small secreted protein in *Zymoseptoria tritici* is responsible for avirulence on wheat cultivars carrying the Stb6 resistance gene. *New Phytologist*, *214*(2), 619–631. https://doi.org/10.1111/nph.14434

Zhu, S., Li, Y., Vossen, J. H., Visser, R. G. F., & Jacobsen, E. (2012). Functional stacking of three resistance genes against *Phytophthora infestans* in potato. *Transgenic Research*, *21*(1), 89–99. https://doi.org/10.1007/s11248-011-9510-1

Zhu, W., Wei, W., Wu, Y., Zhou, Y., Peng, F., Zhang, S., Chen, P., & Xu, X. (2017). BcCFEM1, a CFEM Domain-Containing Protein with Putative GPI-Anchored Site, Is Involved in Pathogenicity, Conidial Production, and Stress Tolerance in *Botrytis cinerea*. *Frontiers in Microbiology*, *8*. https://doi.org/10.3389/fmicb.2017.01807

Zolan, M. E. (1995). Chromosome-length polymorphism in fungi. *Microbiological Reviews*, *59*(4), 686–698. Retrieved August 31, 2020, from https://mmbr.asm.org/content/59/4/686

Zuccaro, A., Lahrmann, U., Güldener, U., Langen, G., Pfiffi, S., Biedenkopf, D., Wong, P., Samans, B., Grimm, C., Basiewicz, M., Murat, C., Martin, F., & Kogel, K.-H. (2011). Endophytic Life Strategies Decoded by Genome and Transcriptome Analyses of the Mutualistic Root Symbiont *Piriformospora indica*. *PLOS Pathogens*, *7*(10), e1002290. https://doi.org/10.1371/journal.ppat.1002290

CHAPTER 2

Bioinformatic prediction of plant–pathogenicity effector proteins of fungi

This chapter is also published in:
Current Opinion in Microbiology, 2018, vol. 46, pp. 43–49
https://doi.org/10.1016/j.mib.2018.01.017

2.1 Declaration

Title	Bioinformatic prediction of plant–pathogenicity effector proteins of fungi.
Authors	**Darcy A. B. Jones**, Stefania Bertazzoni, Chala J. Turo, Robert A. Syme, and James K. Hane
Publication	2018. *Current Opinion in Microbiology*, 46, 43–49.
DOI	`https://doi.org/10.1016/j.mib.2018.01.017`

This thesis chapter is submitted in the form of a collaboratively-written and peer-reviewed manuscript. As such, not all work contained within this chapter can be attributed to the Ph. D. candidate.

The following contributions were made by the Ph.D candidate (DABJ) and co-authors:

- **DABJ** and JKH contributed to research for all sections.
- **DABJ** and JKH contributed to writing and editing all sections.
- JKH researched and wrote sections involving simple amino-acid motifs and Table 1.
- **DABJ** researched and wrote sections involving secretion prediction, machine learning, prioritisation of candidates, future prospects, Figure 1, and Table 2.
- SB researched and contributed to writing sections involving gene prediction.
- CJT researched and contributed to writing sections involving transcriptomics and genomic landscape.
- RAS researched and contributed to writing sections involving comparative genomics.
- **DABJ** and JKH edited the manuscript.
- All authors read and approved the manuscript.

I, Darcy Jones, hereby certify that this attribution statement is an accurate record of my contribution to the research presented in this chapter.

Darcy Jones

I certify that this attribution statement is an accurate record of Darcy Jones' contribution to the research presented in this chapter.

James K. Hane | Stefania Bertazzoni | Chala J. Turo

Robert A. Syme

Available online at www.sciencedirect.com

ScienceDirect

Current Opinion in Microbiology

Bioinformatic prediction of plant–pathogenicity effector proteins of fungi

Darcy AB Jones[1], Stefania Bertazzoni[1], Chala J Turo[1], Robert A Syme[1] and James K Hane[1,2]

Effector proteins are important virulence factors of fungal plant pathogens and their prediction largely relies on bioinformatic methods. In this review we outline the current methods for the prediction of fungal plant pathogenicity effector proteins. Some fungal effectors have been characterised and are represented by conserved motifs or in sequence repositories, however most fungal effectors do not generally exhibit high conservation of amino acid sequence. Therefore various predictive methods have been developed around: general properties, structure, position in the genomic landscape, and detection of mutations including repeat-induced point mutations and positive selection. A combinatorial approach incorporating several of these methods is often employed and candidates can be prioritised by either ranked scores or hierarchical clustering.

Addresses
[1] Centre for Crop and Disease Management, Curtin University, Kent Street, Bentley, WA 6102, Australia
[2] Curtin Institute for Computation, Curtin University, Kent Street, Bentley, WA 6102, Australia

Corresponding author: Hane, James K (james.hane@curtin.edu.au)

Current Opinion in Microbiology 2018, **46**:43–49

This review comes from a themed issue on **Host–microbe interactions: fungi**

Edited by **Ralph Dean** and **Thierry Rouxel**

For a complete overview see the Issue and the Editorial

Available online 22nd February 2018

https://doi.org/10.1016/j.mib.2018.01.017

Introduction

Fungal and oomycete plant pathogens are a major food security problem, with as few as five major species destroying stocks capable of feeding >600 million people [1]. Many plant pathogens possess a battery of 'effector' molecules, usually proteins, which initiate disease and circumvent host defences by either masking the presence of the pathogen or killing the host cell directly [2–4]. Effector identification is critical to developing crop resistance [5] and their prediction largely relies on bioinformatics [6]. Wide adoption and affordability of genome sequencing has enabled multiple pathogen genome studies predicting numerous effector candidates with limited capacity for experimental validation. This is indicative of an ongoing community need for improved knowledge around definitive effector properties and bioinformatic prediction methods to prioritise validation of candidate effectors. This review provides an overview of current and emerging methods for proteinaceous effector prediction in fungi.

The term 'effector' is used to describe multiple loosely conserved families of proteins that are cytotoxic or otherwise compromise cells of a host organism. In plant pathogenic fungal species, these may share basic properties including: low molecular weight, externalisation from the pathogen cell, and cysteine-richness [2–4], however there are several exceptions [7•,8]. Identification of conserved sequence motifs that correlate to pathogenicity-related functions has had mixed success. Two publicly available repositories of proteins with confirmed roles in pathogenicity exist (PHIbase [8] and DFVF [9]). PHIbase aggregates experimental reports validating virulence activities, predominantly of fungal–plant interactions. DFVF groups confirmed pathogenicity factors according to host range. General conserved domain databases also contain small but growing sets of plant–pathogenic functional domains (Table 1A).

A handful of conserved amino-acid motifs have been identified in plant pathogen effector proteins (Table 1B). These are primarily observed in oomycetes and tend to be commonly enriched in the secretomes of species from the same genus [4,10–13]. The crinkler motif is broader, and observed across many sequenced oomycetes [4]. For conserved effector families such as these (Table 1), it is possible to generate profile hidden Markov models (HMMs) to represent the class. Based on current examples from the oomycetes [4], it would appear that motif enrichment analysis within predicted secretomes may be sufficient to predict pathogenicity motifs in novel oomycete genomes.

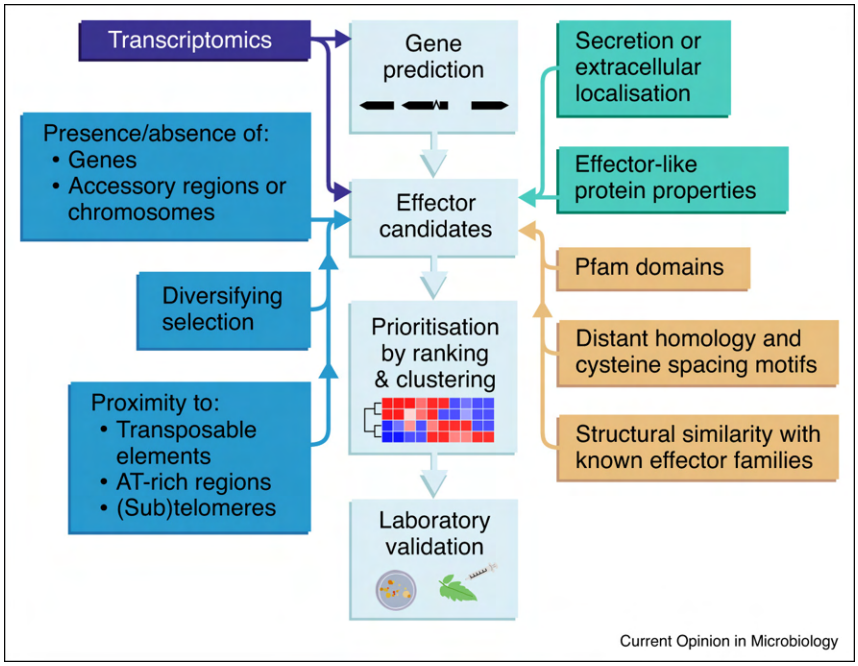
Prediction methods in fungi (Figure 1, Table 2) are more complex as their known effectors tend to lack sequence conservation, likely due to genome-wide mutagenesis processes that make most fungal genomes inherently plastic [14–16]. For example, while some fungi possess known 'RXLR-like' effectors with proposed similar membrane interaction functions [17] a simple pattern-based representation of known fungal RXLR-like motifs

Table 1

Summary of conserved domains (A) and conserved amino-acid motifs (B) observed in plant–pathogen effector proteins.

(A) Plant–pathogenic conserved domains	Pfam ID
ToxA	Toxin_ToxA (PF11584)
Phytotoxin PcF protein	PcF (PF09461)
Putative necrosis-inducing factor	Hce2 (PF14856)
RXLR phytopathogen effector protein	RXLR (PF16810)
Avirulence protein ATR13, RxLR effector	ATR13 (PF16829)
Elicitin	Elicitin (PF00964)
(B) Plant–pathogen effector conserved AA motifs	Primary taxa
RxLR . . . dEER	Phytophthora spp.
Crinklers: LxLFLAK . . . (DWL)n...HVLVxxP	Oomycetes, for example, *Phytophthora* spp., *H. arabidopsidis*, *B. lactucae*, *Pythium* spp.
CxHC	*Albugo laibachii*
YxSL[RK]	*Phythium* spp.
[YFW]xC	*Blumeria graminis*
ETVIC and HRxxH	*Blumeria graminis*

Figure 1



Suggested bioinformatic workflow for generating and prioritising fungal effector candidates.

yields a high false discovery rate in fungal genomes. Similarly, reports of fungal genes possessing the crinkler motif are rare and diverse in sequence [18,19]. Although fungal effectors tend to lack conservation, an exception are the conserved 'Secreted In Xylem' (SIX) genes of *Fusarium oxysporum formae speciales* (ff. spp.) [20], some of which are downstream of a highly conserved miniature impala transposable element that can be used as a predictive marker [21]. However, in most cases the overall lack of usable sequence conservation for effector prediction necessitates a composite approach using various other properties that have been observed for known effectors (Figure 1, Table 2).

Although effector sequences generally lack sequence similarity, common protein structural features have been recently observed within a handful of effector 'families' [3], including ToxA-like [22,23•], MAX [24•], AvrLm6-like [25], RXLR-like with WY-domains [13,26] and RALPH [27•]. Interestingly, both the ToxA-like and MAX families possess a β sandwich tertiary structure formed from 7 and 6 β sheets respectively. Structural homology searches based on position specific scoring matrices (PSSMs) or profile-HMMs (Table 1) have been used to predict new candidates based on matches to existing families [13,25] or for assumption-free prediction [28]. Outside of plant pathology other

Table 2

Commonly-used and recommended software for key tasks in fungal effector gene prediction.

Category	Software	Application	URL
Fungal gene annotation	CodingQuarry	Gene predictor specifically for Fungi. Additional 'pathogen mode' to detect aberrant genes including many effector-like genes	www.codingquarry.sourceforge.io
	SnowyOwl	A software pipeline tailored to fungal gene prediction	www.snowyowl.sourceforge.io
	BRAKER	General purpose gene prediction software which performs well with fungal genes	www.bioinf.uni-greifswald.de/bioinf/braker
	EvidenceModeller/ PASA	Combines results from multiple gene prediction methods	www.evidencemodeler.github.io
Genomic landscape and mutation hotspots	Occultercut	Detection of AT-rich regions in genomes, which may contain effectors	www.occultercut.sourceforge.io
	RIPCAL	Prediction of RIP activity in duplicated sequences	www.ripcal.sourceforge.io
	PAML	Maximum likelihood methods for detection of diversifying selection via CODEML	www.abacus.gene.ucl.ac.uk/software/paml
	SnpEff	General SNP annotation and methods for predicting SNP effects	www.snpeff.sourceforge.net
Comparative genomics	MUMmer	General purpose pairwise genome alignment. Useful for finding presence absence variations	www.mummer.sourceforge.net
	ProgressiveCactus	Multiple genome alignment appropriate for eukaryotic genomes. Useful for finding presence absence variations	www.github.com/glennhickey/progressiveCactus
	EffectorDB.com	Database of effector-like rare orthology or lateral gene transfer candidates	www.effectordb.com
Distant or structural homology	HMMER	General profile-HMM database search and alignment tools	www.hmmer.org
	PSIBLAST	Progressive BLAST algorithm employing PSSMs	www.blast.ncbi.nlm.nih.gov
	Hhblits	HMM–HMM search tool for detection of distant sequence homologues	www.github.com/soedinglab/hh-suite
Effector-like protein properties, secretion, localisation	EffectorP	Machine learning predictor for fungal effector proteins	www.effectorp.csiro.au
	SignalP	Prediction of secreted proteins by identification of N-terminal signal peptides. A common first step in effector prediction. See Sperschneider *et al.* [42•] for discussion of use in Fungi and Oomycetes	www.cbs.dtu.dk/services/SignalP
	TargetP	General prediction of protein subcellular or extracellular localisation	www.cbs.dtu.dk/services/TargetP
	TMHMM	Prediction of transmembrane domains in proteins. Presence of secretion signal and absence of transmembrane domains suggests secretion	www.cbs.dtu.dk/services/TMHMM
	LOCALIZER	Prediction of localisation within host plant cells	www.localizer.csiro.au
	ApoplastP	Prediction of proteins localised in the plant apoplast	www.apoplastp.csiro.au

cytotoxic cysteine-rich SSPs, for example, conotoxins [29] and defensins [30], have been classified based on conserved cysteine spacing patterns with profile sequence searches and machine learning [31]. The identification of conserved structural features and spacing of cysteine motifs across some effectors supports further exploration of the applicability of cysteine pattern searches to effector discovery in fungi and oomycetes [32–34].

General properties of effector genes and proteins

Classically, effector-candidate sets are identified from a genome by iteratively filtering a predicted proteome for small secreted proteins (SSPs) based on loosely generalised effector properties, including: predicted secretion or extracellular localisation, small size or low molecular weight (e.g. <300 AA or <30 kDa), and enriched cysteine content (e.g. ≥4 cysteine residues) [2,6]. These generalised thresholds risk excluding a small number of genuine effectors, as larger and cysteine-poor effectors have been observed [7•,8]. Candidates meeting these criteria may then be refined with comparative genomics or supplementary experimental data, such as *in planta* transcriptomics (Figure 1).

Gene prediction

Effector prediction methods are reductive, therefore their success relies on the presence of effectors within the whole proteome. Thus the prediction of genes is a critical first step in this process, but not always straightforward.

Current gene prediction methods are usually assisted by a combination of *in silico* predictions [35•] (Table 2) and supporting RNAseq and homology alignments [36]. Fungal genes have high density and short introns (50–100 bp) [37], and it is common for untranslated regions (UTRs) of adjacent loci to overlap the same region of the genome [38•]. This complicates use of genome-mapped RNAseq data to assist in the prediction of exon structure, and requires specialised methods such as CodingQuarry [35•] and/or manual curation to resolve.

The prediction of effector genes poses additional challenges. Effectors often lack recognisable homology and may exhibit uncommon amino acid or codon sequence biases [38•] that do not resemble the predictive model used by *in silico* gene predictors trained on the majority of genes. Gene prediction methods that aim to minimise false positives produce high confidence gene sets, but may miss effectors. The gene predictor CodingQuarry [35•] includes a secondary 'Pathogen Mode' (CQPM) which trains a secondary gene model using genes from the first pass that were predicted to be secreted, cysteine-rich and with atypical codon usage. The secretion-specific model is then used to annotate new loci between primary gene predictions. Many effectors are also located near or within regions of repetitive DNA, which are typically difficult to assemble and may be missing, incomplete or incorrect [39], however this is becoming less of an issue as 3rd generation sequencing methods are adopted [40,41].

Secretion prediction

Effector prediction is primarily focussed on extracellular and/or secreted proteins, thus a common initial step is the prediction of secretion signals within the predicted proteome and/or the exclusion of trans-membrane domains [6]. Comparison of several secretion prediction methods for a set of known effectors [42•] showed that neural-network versions of SignalP 2 and 3 perform best for fungal effectors, while the HMM model of SignalP 3 performs better for oomycete effectors. Three confirmed effectors (Avra10, Avrk1, Vdlsc1 [8]) lack predicted signal peptides and may be secreted by non-classical mechanisms. SecretomeP is designed to predict non-classically secreted proteins, but was trained on non-fungal sequences, raising questions about its applicability to fungi.

Machine learning

Modelling of general trends in protein properties (e.g. molecular weight or amino-acid frequencies) using machine learning (ML) has been applied both broadly to the prediction of functional protein families [43] as well as specifically to the prediction of effector-like cytotoxic peptides (small, charged, cysteine-rich, membrane interacting) such as antimicrobial peptides, defensins, and toxins [44]. The first ML classifier of fungal effectors — EffectorP [7•] — was trained using sequences of experimentally validated effectors, and offers a valuable alternative to simple cysteine or size thresholds. Sperschneider *et al.* [7•] reported that low molecular weight, cysteine-richness, overall protein charge, serine content and tryptophan content were important features distinguishing effectors from non-effectors. Subcellular-localisation upon entry into the host cell, as well as apoplastic localisation of effector candidates can be predicted using LOCALIZER [45•] and ApoplastP [46•] respectively (Table 2).

ML model performance is dependent on the training data and selection of features, and would not necessarily correctly classify new samples with characteristics infrequently seen during training. This poses challenges, as effectors are a diverse group with relatively few validated examples. However, model accuracy will likely improve as more effectors capturing greater diversity within the group become available. LOCALIZER [45•] and ApoplastP [46•] partially address this issue by supplementing the small number of effectors in the training set with plant proteins of known subcellular/extracellular localisations. EffectorP [7•], LOCALIZER [45•], and ApoplastP [46•] perform remarkably well given the data available for training but will occasionally yield incorrect classifications. Discrepancies with proteins of interest should be critically evaluated against model biases introduced by training data and features, and may indicate novel or uncommon effector protein properties.

Comparative genomics

Comparative genomics encompasses both comparisons between different pathogen species, and individuals of the same species in a population. Genomic comparisons between species have been significant in highlighting rare homology or orthology relationships between proteins from distantly related pathogen species, which may indicate lateral gene transfer [16] or important virulence function. This can also provide strong evidence for new effector candidates (see www.effectordb.com, Table 2).

Reduced sequencing cost has made comparisons between multiple pathogen isolates or populations of a single species increasingly feasible, which can be leveraged for effector prediction through detection of anomalous selection pressures, presence-absence variation (PAV) [36], or via genome-wide association where phenotypic data are available [47,48]. RNAseq is a viable alternative for population level sequence comparisons of pathogens that cannot be cultured [48]. The 'arms race' model of pathogen–host co-evolution implies that virulence and effector genes undergo diversifying selection [49•,50•] indicated by an elevated ratio of synonymous to non-synonymous SNPs (dN/dS). This can be calculated from codon-aligned nucleotide sequences of multiple isolates [49•], or through read alignment relative to a reference isolate [36]. PAV across a pathogen population can

support the prediction of candidates at either the gene-level [36,51] or regional-level, that is, genes located on accessory chromosomes [52,53]. It should be noted that the success of PAV analyses are dependent on one or more highly contiguous reference genome assemblies, particularly if effector loci are absent from the primary reference. Some fungal effectors also have been reported to be associated with sub-telomeric regions [16,54], which can be inferred with high-quality genome assemblies [40,41].

Genomic landscape

Many fungal pathogen genomes are plastic due to mutagenesis mechanisms or characteristics specific to — or enriched in — fungal taxa, such as: repeat-induced point mutations (RIP) [15], genome structural rearrangements (e.g. producing mesosyntenic patterns of conservation) [14], and potential for lateral gene transfer [16]. Many fungal effectors are also located in genome regions with hypervariable sequence, as such, physical proximity to these regions can be used to predict effector candidates [38•]. Discrete compartmentalisation of normal and hypervariable genome regions can be observed in *Leptosphaeria maculans* where most known effectors are located in RIP-mutated AT-rich regions [38•,54,55•]. However, these regions harbour effector candidates that are expressed early in infection during its endophytic stage, but exclude candidates expressed late in infection during its necrotrophic stage [55•]. This indicates that the genomic context should be considered alongside experimental data (e.g. transcriptomics) when used to support effector prediction.

Transcriptomics

In addition to assisting gene prediction, transcriptomics data indicating elevated expression during host infection is a useful predictor of virulence. Differential expression between differentially-pathogenic isolates or mutants may also indicate candidate effectors [36,56]. Comparison of multiple infection stages or time-points can provide a finer-grained view of transcriptional regulation during infection [55•,56]. These expression profiles can be clustered into groups of genes with common expression patterns, and those containing known effectors or virulence factors can be used to predict new candidates by association [57].

Prioritisation of effector candidates

Effector prediction often yields more candidates than can reasonably be validated. The simplest method of obtaining small high-priority sets of candidates is by applying stricter filters, though this has an intrinsic risk of excluding genuine candidates with unusual characteristics [7•]. Alternatives methods employ ranking strategies, assigning a cumulative score to candidates based on multiple characteristics such as: genomic context, amino-acid properties, evidence of positive selection, or other experimental data [36]. Alternatively, hierarchical clustering can be used to group candidates with similar characteristics [58] and has the advantage of reducing the candidate set through exclusion of candidate clusters containing known non-effectors or members with non-effector-like properties.

The future of bioinformatic prediction of fungal effector proteins

Bioinformatic prediction of effectors has significantly advanced in recent years due to the adoption of long-read sequencing, trends towards pan-genomics, increased accessibility of machine learning methods in non-specialist domains, and the discovery of several sequence diverse effector 'families'. The increased number of sequenced pathogen genomes benefits effector discovery efforts, but also necessitates more curation to prevent error propagation via public databases. Machine learning techniques show great promise in discovering new effector candidates and determining characteristics common to effector proteins, and their accuracy will improve as more experimentally confirmed effectors become available. Similarly, structural protein comparison methods (such as cysteine-spacing analysis, sequence profile comparisons, and 3D structural modelling) may facilitate further identification of sequence-diverse effector families and lead to improved predictive methods in the future. Bioinformatic effector prediction will continue to be an important element enabling fungal effector protein discovery, exploring their evolutionary origins, common characteristics and subfamilies thereof, which will in turn accelerate further effector discovery in the future.

Author contributions

DABJ, SB, CJT, RAS and JKH wrote the manuscript. DABJ and JKH edited the manuscript. All authors read and approved the manuscript.

Acknowledgements

The PhD candidature of DABJ was funded by the Centre for Crop and Disease Management and the Curtin Institute for Computation.

References and recommended reading

Papers of particular interest, published within the period of review, have been highlighted as:

- • of special interest

1. Fisher MC, Henk DA, Briggs CJ, Brownstein JS, Madoff LC, McCraw SL, Gurr SJ: **Emerging fungal threats to animal, plant and ecosystem health**. *Nature* 2012, **484**:186-194.

2. de Wit PJ, Mehrabi R, van den Burg HA, Stergiopoulos I: **Fungal effector proteins: past, present and future**. *Mol Plant Pathol* 2009, **10**:735-747.

3. Franceschetti M, Maqbool A, Jiménez-Dalmaroni MJ, Pennington HG, Kamoun S, Banfield MJ: **Effectors of filamentous plant pathogens: commonalities amid diversity**. *Microbiol Mol Biol Rev* 2017, **81** e00066-00016.

4. Jiang RH, Tyler BM: **Mechanisms and evolution of virulence in oomycetes**. *Ann Rev Phytopathol* 2012, **50**:295-318.

5. Vleeshouwers VG, Oliver RP: **Effectors as tools in disease resistance breeding against biotrophic, hemibiotrophic, and necrotrophic plant pathogens**. *Mol Plant–Microbe Interact* 2014, **27**:196-206.

6. Sonah H, Deshmukh RK, Bélanger RR: **Computational prediction of effector proteins in fungi: opportunities and challenges**. *Front Plant Sci* 2016, **7**:126.

7. • Sperschneider J, Gardiner DM, Dodds PN, Tini F, Covarelli L, Singh KB, Manners JM, Taylor JM: **EffectorP: predicting fungal effector proteins from secretomes using machine learning**. *New Phytol* 2016, **210**:743-761.
The first application of machine learning to fungal effector prediction, offers a strong alternative to classical cysteine content or size filters.

8. Urban M, Cuzick A, Rutherford K, Irvine A, Pedro H, Pant R, Sadanadan V, Khamari L, Billal S, Mohanty S *et al.*: **PHI-base: a new interface and further additions for the multi-species pathogen–host interactions database**. *Nucleic acids Res* 2017, **45**:D604-D610.

9. Lu T, Yao B, Zhang C: **DFVF: database of fungal virulence factors**. *Database* 2012, **2012**:bas032.

10. Leonelli L, Pelton J, Schoeffler A, Dahlbeck D, Berger J, Wemmer DE, Staskawicz B: **Structural elucidation and functional characterization of the *Hyaloperonospora arabidopsidis* effector protein ATR13**. *PLoS Pathog* 2011, **7**: e1002428.

11. Stassen JH, Seidl MF, Vergeer PW, Nijman IJ, Snel B, Cuppen E, Van den Ackerveken G: **Effector identification in the lettuce downy mildew *Bremia lactucae* by massively parallel transcriptome sequencing**. *Mol Plant Pathol* 2012, **13**:719-731.

12. Lévesque CA, Brouwer H, Cano L, Hamilton JP, Holt C, Huitema E, Raffaele S, Robideau GP, Thines M, Win J: **Genome sequence of the necrotrophic plant pathogen *Pythium ultimum* reveals original pathogenicity mechanisms and effector repertoire**. *Genome Biol* 2010, **11**:R73.

13. Jiang RH, Tripathy S, Govers F, Tyler BM: **RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members**. *Proc Natl Acad Sci* 2008, **105**:4874-4879.

14. Hane JK, Rouxel T, Howlett BJ, Kema GH, Goodwin SB, Oliver RP: **A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi**. *Genome Biol* 2011, **12**:R45.

15. Hane JK, Williams AH, Taranto AP, Solomon PS, Oliver RP: **Repeat-induced point mutation: a fungal-specific, endogenous mutagenesis process**. In *Genetic Transformation Systems in Fungi*, vol 2. Edited by van den Berg M, Maruthachalam K . Springer; 2015:55-68.

16. Schmidt SM, Panstruga R: **Pathogenomics of fungal plant parasites: what have we learnt about pathogenesis?** *Curr Opin Plant Biol* 2011, **14**:392-399.

17. Kale SD, Gu B, Capelluto DGS, Dou D, Feldman E, Rumore A, Arredondo FD, Hanlon R, Fudal I, Rouxel T *et al.*: **External lipid PI3P mediates entry of eukaryotic pathogen effectors into plant and animal host cells**. *Cell* 2010, **142**:284-295.

18. James TY, Pelin A, Bonen L, Ahrendt S, Sain D, Corradi N, Stajich JE: **Shared signatures of parasitism and phylogenomics unite cryptomycota and microsporidia**. *Curr Biol* 2013, **23**:1548-1553.

19. Sun G, Yang Z, Kosch T, Summers K, Huang J: **Evidence for acquisition of virulence effectors in pathogenic chytrids**. *BMC Evol Biol* 2011, **11**:195.

20. Lievens B, Houterman PM, Rep M: **Effector gene screening allows unambiguous identification of *Fusarium oxysporum* f. sp. *lycopersici* races and discrimination from other *formae speciales***. *FEMS Microbiol Lett* 2009, **300**:201-215.

21. Schmidt SM, Houterman PM, Schreiver I, Ma L, Amyotte S, Chellappan B, Boeren S, Takken FLW, Rep M: **MITEs in the promoters of effector genes allow prediction of novel virulence genes in *Fusarium oxysporum***. *BMC Genom* 2013, **14**:119.

22. Di X, Cao L, Hughes RK, Tintor N, Banfield MJ, Takken FLW: **Structure–function analysis of the *Fusarium oxysporum* Avr2 effector allows uncoupling of its immune-suppressing activity from recognition**. *New Phytol* 2017, **216**:897-914.

23. • Lu S, Turgeon BG, Edwards MC: **A ToxA-like protein from *Cochliobolus heterostrophus* induces light-dependent leaf necrosis and acts as a virulence factor with host selectivity on maize**. *Fungal Genet Biol* 2015, **81**:12-24.
Identification of functional but sequence-diverse ToxA homolog in *Cochliobolus heterostrophus* with structural similarity.

24. • de Guillen K, Ortiz-Vallejo D, Gracy J, Fournier E, Kroj T, Padilla A: **Structure analysis uncovers a highly diverse but structurally conserved effector family in phytopathogenic fungi**. *PLoS Pathog* 2015, **11**:e1005228.
Study proposing the MAX effector family, which contains structurally similar but sequence-diverse members from *Magnaporthe oryzae* and *Pyrenophora tritici-repentis*.

25. Shiller J, van de Wouw AP, Taranto AP, Bowen JK, Dubois D, Robinson A, Deng CH, Plummer KM: **A large family of AvrLm6-like genes in the apple and pear scab pathogens, *Venturia inaequalis* and *Venturia pirina***. *Front Plant Sci* 2015, **6**:980.

26. Boutemy LS, King SR, Win J, Hughes RK, Clarke TA, Blumenschein TM, Kamoun S, Banfield MJ: **Structures of Phytophthora RXLR effector proteins a conserved but adaptable fold underpins functional diversity**. *J Biol Chem* 2011, **286**:35834-35842.

27. • Praz CR, Bourras S, Zeng F, Sánchez-Martín J, Menardo F, Xue M, Yang L, Roffler S, Böni R, Herren G *et al.*: ***AvrPm2* encodes an RNase-like avirulence effector which is conserved in the two different specialized forms of wheat and rye powdery mildew fungus**. *New Phytol* 2017, **213**:1301-1314.
An example of GWAS applied to fungal effector prediction, prompting the proposal of the new 'RALPH' effector family.

28. Sperschneider J, Gardiner DM, Taylor JM, Hane JK, Singh KB, Manners JM: **A comparative hidden Markov model analysis pipeline identifies proteins characteristic of cereal-infecting fungi**. *BMC Genom* 2013, **14**:807.

29. Lavergne V, Harliwong I, Jones A, Miller D, Taft RJ, Alewood PF: **Optimized deep-targeted proteotranscriptomic profiling reveals unexplored Conus toxin diversity and novel cysteine frameworks**. *Proc Natl Acad Sci U S A* 2015, **112**: E3782-E3791.

30. Shafee TMA, Lay FT, Hulett MD, Anderson MA: **The defensins consist of two independent, convergent protein superfamilies**. *Mol Biol Evol* 2016, **33**:2345-2356.

31. Islam SMA, Sajed T, Kearney CM, Baker EJ: **PredSTP: a highly accurate SVM based model to predict sequential cystine stabilized peptides**. *BMC Bioinform* 2015, **16**:210.

32. Bhadauria V, MacLachlan R, Pozniak C, Banniza S: **Candidate effectors contribute to race differentiation and virulence of the lentil anthracnose pathogen *Colletotrichum lentis***. *BMC Genom* 2015, **16**:628.

33. Duplessis S, Cuomo CA, Lin Y-C, Aerts A, Tisserant E, Veneault-Fourrey C, Joly DL, Hacquard S, Amselem J, Cantarel BL *et al.*: **Obligate biotrophy features unraveled by the genomic analysis of rust fungi**. *Proc Natl Acad Sci U S A* 2011, **108**:9166-9171.

34. Kohler AC, Chen L-H, Hurlburt N, Salvucci A, Schwessinger B, Fisher AJ, Stergiopoulos I: **Structural analysis of an Avr4 effector ortholog offers insight into chitin binding and recognition by the Cf-4 receptor**. *Plant Cell* 2016, **28**:1945-1965.

35. • Testa AC, Hane JK, Ellwood SR, Oliver RP: **CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts**. *BMC Genom* 2015, **16**:170.
A gene predictor specialised for fungal genomes.

36. Syme RA, Hane JK, Friesen TL, Oliver RP: **Resequencing and comparative genomics of *Stagonospora nodorum*: sectional gene absence and effector discovery**. *G3: Genes Genom Genet* 2013, **3**:959-969.

37. Kupfer DM, Drabenstot SD, Buchanan KL, Lai H, Zhu H, Dyer DW, Roe BA, Murphy JW: **Introns and splicing elements of five diverse fungi**. *Eukaryot Cell* 2004, **3**:1088-1100.

38. • Testa AC, Oliver RP, Hane JK: **OcculterCut: a comprehensive survey of AT-rich regions in fungal genomes**. *Genome Biol Evol* 2016, **8**:2044-2064.
Software to predict AT-rich compartmentalisation in fungal genomes and AT-rich associated genes.

39. Thomma BPHJ, Seidl MF, Shi-Kunne X, Cook DE, Bolton MD, van Kan JAL, Faino L: **Mind the gap; seven reasons to close fragmented genome assemblies**. *Fungal Genet Biol* 2016, **90**:24-30.

40. Derbyshire M, Denton-Giles M, Hegedus D, Seifbarghy S, Rollins J, van Kan J, Seidl MF, Faino L, Mbengue M, Navaud O *et al.*: **The complete genome sequence of the phytopathogenic fungus *Sclerotinia sclerotiorum* reveals insights into the genome architecture of broad host range pathogens**. *Genome Biol Evol* 2017, **9**:593-618.

41. Faino L, Seidl MF, Datema E, van den Berg GCM, Janssen A, Wittenberg AHJ, Thomma BPHJ: **Single-molecule real-time sequencing combined with optical mapping yields completely finished fungal genome**. *mBio* 2015, **6** e00936-00915.

42. • Sperschneider J, Williams AH, Hane JK, Singh KB, Taylor JM: **Evaluation of secretion prediction highlights differing approaches needed for oomycete and fungal effectors**. *Front Plant Sci* 2015, **6**:1168.
A study benchmarking secretion prediction tools for filamentous plant pathogens, which is an integral step in effector prediction.

43. Li YH, Xu JY, Tao L, Li XF, Li S, Zeng X, Chen SY, Zhang P, Qin C, Zhang C *et al.*: **SVM-Prot 2016: a web-server for machine learning prediction of protein functional families from sequence irrespective of similarity**. *PLoS ONE* 2016, **11**: e0155290.

44. Waghu FH, Barai RS, Gurung P, Idicula-Thomas S: **CAMPR3: a database on sequences, structures and signatures of antimicrobial peptides**. *Nucleic Acids Res* 2016, **44**:D1094-D1097.

45. • Sperschneider J, Catanzariti A-M, DeBoer K, Petre B, Gardiner DM, Singh KB, Dodds PN, Taylor JM: **LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell**. *Sci Rep* 2017, **7**:44598.
A tool for predicting plant subcellular localisation of plant or pathogen proteins. Takes a novel approach of including plant proteins in machine learning training sets to improve predictions for effectors.

46. • Sperschneider J, Dodds PN, Singh KB, Taylor JM: **ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning**. *New Phytol* 2018, **217**:1764-1778.
A tool for predicting fungal proteins localised to the plant apoplast. Allows for separation of cytoplasmic and apoplastic candidate effectors.

47. Gao Y, Liu Z, Faris JD, Richards J, Brueggeman RS, Li X, Oliver RP, McDonald BA, Friesen TL: **Validation of genome-wide association studies as a tool to identify virulence factors in *Parastagonospora nodorum***. *Phytopathology* 2016, **106**:1177-1185.

48. Lu X, Kracher B, Saur IML, Bauer S, Ellwood SR, Wise R, Yaeno T, Maekawa T, Schulze-Lefert P: **Allelic barley MLA immune receptors recognize sequence-unrelated avirulence effectors of the powdery mildew pathogen**. *Proc Natl Acad Sci U S A* 2016, **113**:E6486-E6495.

49. • Poppe S, Dorsheimer L, Happel P, Stukenbrock EH: **Rapidly evolving genes are key players in host specialization and virulence of the fungal wheat pathogen *Zymoseptoria tritici* (*Mycosphaerella graminicola*)**. *PLoS Pathog* 2015, **11**: e1005055.
Structural modelling and analysis of diversifying selection in characterising evolution of effector candidates.

50. • Sperschneider J, Gardiner DM, Thatcher LF, Lyons R, Singh KB, Manners JM, Taylor JM: **Genome-wide analysis in three *Fusarium* pathogens identifies rapidly evolving chromosomes and genes associated with pathogenicity**. *Genome Biol Evol* 2015, **7**:1613-1627.
Application of diversifying selection searches to detect effector candidates.

51. Golicz AA, Martinez PA, Zander M, Patel DA, van de Wouw AP, Visendi P, Fitzgerald TL, Edwards D, Batley J: **Gene loss in the fungal canola pathogen *Leptosphaeria maculans***. *Funct Integr Genom* 2015, **15**:189-196.

52. Ma L-J, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-J, Di Pietro A, Dufresne M, Freitag M, Grabherr M, Henrissat B *et al.*: **Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium***. *Nature* 2010, **464**:367-373.

53. Plissonneau C, Stürchler A, Croll D: **The evolution of orphan regions in genomes of a fungal pathogen of wheat**. *mBio* 2016, **7** e01231-01216.

54. Soyer JL, El Ghalid M, Glaser N, Ollivier B, Linglin J, Grandaubert J, Balesdent M-H, Connolly LR, Freitag M, Rouxel T *et al.*: **Epigenetic control of effector gene expression in the plant pathogenic fungus *Leptosphaeria maculans***. *PLoS Genet* 2014, **10**:e1004227.

55. • Gervais J, Plissonneau C, Linglin J, Meyer M, Labadie K, Cruaud C, Fudal I, Rouxel T, Balesdent M-H: **Different waves of effector genes with contrasted genomic location are expressed by *Leptosphaeria maculans* during cotyledon and stem colonization of oilseed rape**. *Mol Plant Pathol* 2017, **18**:1113-1126.
Observation of differential temporal expression of *Leptosphaeria maculans* effectors. Effectors in hypervariable AT-rich regions were expressed during the early endophytic phase of infection, whereas effector candidates in gene-rich regions were associated with the late necrotrophic phase.

56. Palma-Guerrero J, Ma X, Torriani SFF, Zala M, Francisco CS, Hartmann FE, Croll D, McDonald BA: **Comparative transcriptome analyses in *Zymoseptoria tritici* reveal significant differences in gene expression among strains during plant infection**. *Mol Plant–Microbe Interact* 2017, **30**:231-244.

57. Dong Y, Li Y, Zhao M, Jing M, Liu X, Liu M, Guo X, Zhang X, Chen Y, Liu Y *et al.*: **Global genome and transcriptome analyses of *Magnaporthe oryzae* epidemic isolate 98-06 uncover novel effectors and pathogenicity-related genes, revealing gene gain and lose dynamics in genome evolution**. *PLoS Pathog* 2015, **11**:e1004801.

58. Saunders DGO, Win J, Cano LM, Szabo LJ, Kamoun S, Raffaele S: **Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi**. *PLoS ONE* 2012, **7**:e29847.

CHAPTER 3 — THEME 1

Predector: an automated and combinative method for the predictive ranking of candidate effector proteins of fungal plant-pathogens

A revised preprint of this chapter is available at:
Research Square https://doi.org/10.21203/rs.3.rs-379941/v1

3.1 Declaration

Title	Predector: an automated and combinative method for the predictive ranking of candidate effector proteins of fungal plant-pathogens.
Authors	**Darcy A.B. Jones**, Lina Rozano, Johannes Debler, Ricardo Mancera, Paula Moolhuijzen, and James K. Hane

This thesis chapter is submitted in the form of a collaboratively-written manuscript ready for journal submission. As such, not all work contained within this chapter can be attributed to the Ph. D. candidate.

The following contributions were made by the Ph.D candidate (DABJ) and co-authors:

- **DABJ**, JKH, LR, and JD conceived the study.
- **DABJ**, JKH, LR, and JD contributed to software development and documentation.
- **DABJ** performed all analyses and generated all figures and tables.
- **DABJ** and JKH wrote the manuscript.
- **DABJ**, JKH, LR, PM, and RM edited the manuscript.
- All authors read and approved the manuscript.

I, Darcy Jones, hereby certify that this attribution statement is an accurate record of my contribution to the research presented in this chapter. Note that Lina Rozano is a current Ph. D. student, but this work will not contribute to their thesis.

Darcy Jones

I certify that this attribution statement is an accurate record of Darcy Jones' contribution to the research presented in this chapter.

James K. Hane | Lina Rozano | Johannes Debler

Ricardo Mancera | Paula Moolhuijzen

Abstract

'Effectors' are a broad class of cytotoxic, virulence-promoting, or resistance eliciting molecules that are released from plant-pathogen cells to facilitate disease progression in their host. Fungal effectors are a core research area for improving host disease resistance; however, because they generally lack common distinguishing features or obvious sequence similarity, discovery of effectors remains a major challenge. This study presents a novel tool and pipeline for effector prediction — Predector — which interfaces with multiple software tools and methods, aggregates disparate features that are relevant to fungal effector proteins, and ranks effector candidate proteins using a pairwise learning to rank approach. Predector outperformed alternative effector prediction methods that were applied to a curated set of confirmed effectors derived from multiple species. We present Predector (https://github.com/ccdmb/predector) as a useful tool for the prediction and ranking of effector candidates, which aggregates and reports additional supporting information relevant to effector and secretome prediction in a simple, efficient, and reproducible manner.

3.2 Introduction

'Effectors' are a broad class of cytotoxic, virulence-promoting, or resistance eliciting molecules that are released from plant-pathogen cells to facilitate disease progression in their host. Fungal effectors are a core research area for improving host disease resistance; however, because they generally lack common features or obvious sequence similarity, discovery of effectors is non-trivial (He et al., 2020; D. A. Jones et al., 2018; L. Liu et al., 2019). Secreted effector proteins of plant pathogens have been studied more comprehensively in the Oomycetes (a separate lineage of filamentous microbes), in which *in silico* identification of effectors is more feasible compared to fungi as they exhibit highly conserved sequence motifs (e.g. RXLR, LXLFLAK) (Boutemy et al., 2011; Jiang et al., 2008). Fungal effectors in contrast are highly diverse in sequence and function which may be a result of their highly plastic genomes, diversified by a number of Fungal specific genome mutagenesis mechanisms such as repeat-induced point mutation (RIP) (Galagan & Selker, 2004; Ohm et al., 2012) and mesosynteny (Hane et al., 2011), as well as other more general genome characteristics present in other pathogen groups such as the presence of accessory genomes (Bertazzoni et al., 2018) and laterally acquired genetic material (Schmidt & Panstruga, 2011). As a consequence, effector candidate discovery is performed using experimental techniques such as phenotype association and comparative genomics (Beckerson et al., 2019; Mousavi-Derazmahalleh et al., 2019; Plissonneau et al., 2017; Williams et al., 2016), transcriptomics (Gervais et al., 2017; Human et al., 2020; D. A. B. Jones et al., 2019), proteomics (Gawehns et al., 2015; Mesarich et al., 2018) and genome-wide association studies (GWAS) (Richards et al., 2019; Sánchez-Vallet et al., 2018). There are, however, some protein characteristics — such as structural features (e.g. functional domains), signal peptides (SPs), amino-acid frequencies — that can be used as an alternative to simple homology searches. Several methods using these characteristics have been developed to

prioritise effector candidates for experimental validation (D. A. Jones et al., 2018).

In-silico effector prediction has typically involved *ad hoc* hard set criteria such as a SP, no transmembrane (TM) domains outside the SP, small overall size (often <300 aa), and a high number of cysteine amino-acids. These thresholds were based on the properties of early discovered effectors; however, numerous known effectors do not conform to this profile (Supplementary table S1). The use of simple hard filters risks excluding these proteins from candidacy. SP prediction is the most common *in-silico* technique used to refine effector candidates from proteomes (Sperschneider et al., 2015), with SignalP being the most common prediction tool (Armenteros, Tsirigos, et al., 2019; Bendtsen et al., 2004; Petersen et al., 2011), though other tools are frequently used in combination (Käll et al., 2004; Savojardo et al., 2018) and different tools can perform better or worse with different protein groups or organisms (Sperschneider et al., 2015). Subcellular localisation prediction tools such as TargetP (Armenteros, Salvatore, et al., 2019) or DeepLoc (Armenteros et al., 2017) are also frequently used to predict the location of proteins. Their reliability for predicting protein secretion is questionable (Sperschneider et al., 2015), but proteins predicted to be localised in organelles might reasonably be excluded. Because most effectors are expected to be free in the extracellular space or host cells, TM domains are also an important feature for excluding candidates, commonly predicted using TMHMM (Krogh et al., 2001) or Phobius (Käll et al., 2004).

Recently developed machine learning tools tailored to predicting effector-like properties have presented new opportunities for improving effector prediction pipelines. EffectorP (Sperschneider, Dodds, Gardiner, et al., 2018; Sperschneider et al., 2016) and FunEffector-Pred (C. Wang et al., 2020) use amino acid frequencies, molecular weight, charge, AA k-mers, and other protein characteristics to predict effector-like proteins directly. In combination with secretion prediction, tools like EffectorP and FunEffector-Pred may be a more robust alternative to simple hard filters. LOCALIZER (Sperschneider et al., 2017) and ApoplastP (Sperschneider, Dodds, Singh, et al., 2018), which predict host subcellular or apoplastic localisations, are useful for evaluating candidates but are not necessarily predictive of effector candidacy themselves.

While many fungal effectors have previously not had similar sequences in public databases, a small but increasing number of families based on conserved domains or structure are becoming known (D. A. Jones et al., 2018), including the ToxA-like (S. Lu et al., 2015), MAX (de Guillen et al., 2015), RALPH (Praz et al., 2017; Spanu, 2017), and RXLR-like (Kale et al., 2010) families. Homology with effector-like conserved domains (i.e. selected Pfam domains) or effector-like sequences within databases such as the Plant-Host Interactions database (PHI-base) (Urban et al., 2017; Urban et al., 2020) and the Database of Fungal Virulence Factors (DFVF) (T. Lu et al., 2012), are growing in their relevance. Secondary and tertiary structural modelling and similarity searches against known effectors are not commonly used for high-throughput effector discovery, but this could yet become an important component of future effector prediction pipelines (D. A. Jones et al., 2018).

Current effector prediction pipelines face two major challenges: 1) the necessity of reducing 10-20 thousand proteins per genome down to a set of effector candidates that is both reliable

and within a number that is feasible for experimental validation, and 2) the amalgamation of outputs from a large and diverse range of bioinformatics tools and methods, for both prediction and informative purposes. Fungal genome datasets typically contain thousands of predicted secreted proteins, of which hundreds of small secreted proteins (SSPs) may be predicted by standard methods (D. A. Jones et al., 2018). Further filtering or ranking based on supporting data from GWAS, RNAseq, positive selection, or comparative genomics can still generate hundreds of candidates (Anderson et al., 2017; Dutreux et al., 2018; Sonah et al., 2016; Syme, Tan, Rybak, et al., 2018). The prioritisation of effector candidates based on simple biochemical properties is, therefore, still relevant to effector prediction. However, there is little consensus on how to combine multiple analyses (Sperschneider et al., 2015), and the common use of multiple successive hard filters risks increasing the error with each step, causing good candidates to be excluded. Although such hard filters are useful for identifying sets of well defined classes of effectors (e.g. small cysteine rich), the requirements of identifying and prioritising candidates for experimental follow-up without making such assumptions or strong reliance on any one prediction method is subtly different.

Saunders et al. (2012) approached this problem by ranking clusters of homologous proteins using a number of e-value like scores based on the expected frequencies of each effector-indicating property of interest within a cluster, and used hierarchical clustering to combine information from the e-value scores and identify extended groups of effector candidates with common features. While this method addresses some of the issues described here, the highly rust-specific criteria used and heavy dependency on protein homology clustering potentially limits its wider use.

Rank-based methods in general are a simple way to avoid exclusion of candidates lacking clearly discriminative features, via assigning weighted scores to features that are presumed to be important in determining effector-likelihood, and summing these into a single score that is used to rank candidates (Syme, Tan, Rybak, et al., 2018). However, these simple combinations of manually assigned feature weights may still fail to place proteins with uncommon characteristics near the top of the list. More sophisticated ranking decisions may come from a group of machine learning techniques called "learn to rank". Rather than offering a binary classification (i.e. effector or non-effector), these methods attempt to order elements optimally so that relevant elements are nearer the beginning of the list. Although these algorithms are most often employed in search engine and e-commerce websites, they have been used successfully to combine diverse sources of information and rank protein structure predictions (Qiu et al., 2008), remote homology predictions (B. Liu et al., 2015), gene ontology term assignments (You et al., 2018), and predicting protein-phenotype associations in human disease (L. Liu et al., 2020).

In this study, we present a novel tool and pipeline for effector prediction — Predector — which interfaces with multiple software tools and methods, aggregates disparate features that are relevant to fungal effector proteins, and ranks effector candidate proteins using a pairwise learning to rank approach. Predector simplifies effector prediction workflows by providing

simplified software dependency installation, a standardised pipeline that can be run efficiently on both commodity hardware and supercomputers, and user friendly tabular formatted results. In this study, we compare the performance of Predector against a typical effector prediction method (i.e. SP prediction, TM domain prediction, and EffectorP), on a curated set of confirmed effectors derived from multiple species. While the small number of currently known effectors and relatively loose definition of the group precludes the possibility of perfectly precise effector prediction tools, we present Predector as a tool enabling useful effector candidate ranks alongside supporting information for effector and secretome prediction in a simple, efficient, and reproducible manner.

3.3 Methods

3.3.1 Pipeline implementation

The Predector pipeline runs a range of commonly used effector and secretome prediction bioinformatics tools for complete predicted proteome, accepted as input in FASTA formatted files (Table 3.1), and combines all raw and summarised outputs into newline-delimited JSON, tab-delimited text and GFF3 formats. The pipeline is implemented in Nextflow (version >20) (Di Tommaso et al., 2017), and a conda environment and Docker container are available for easy installation of dependencies, with scripts to integrate user-downloaded proprietary software into these environments. Predector is available from https://github.com/ccdmb/predector.

3.3.2 Datasets

The training and evaluation datasets consisted of: confirmed fungal effectors, fungal proteins with confirmed subcellular localisation, and an 'unlabelled' fungal protein set derived from whole proteomes of well-annotated, model fungal species. The experimentally-confirmed effector protein dataset was curated from literature, PHI-base (Urban et al., 2020), and EffectorP (Sperschneider, Dodds, Gardiner, et al., 2018; Sperschneider et al., 2016) training datasets (Supplementary table S2). Effector homologues were also identified from literature (Supplementary table S2) and by searching the UniRef-90 fungal proteins (UniProtKB query: taxonomy:"Fungi [4751]" AND identity:0.9, UniProt version 2020_01, Downloaded 2020-06-01) using MMSeqs2 version 11-e1a1c (Steinegger & Söding, 2017) requiring a minimum reciprocal coverage of 70% and a maximum e-value of 10^{-5} (`-e 0.00001 --start-sens 3 -s 7.0 --sens-steps 3 --cov-mode 0 -c 0.7`). Fungal proteins with experimentally annotated subcellular localisation were downloaded from UniProtKB/SwissProt (version 2020_01, Downloaded 2020-06-01), and were labelled "secreted" (non-transmembrane) or "non-secreted" (membrane associated, endoplasmic reticulum localised, golgi localised, and glycosylphosphatidylinositol (GPI) anchored). UniProtKB download queries are provided in supplementary table S2. The 'unlabelled' whole proteome dataset was derived from well studied pathogens, with at least one representative chosen from a range of trophic phenotypes (Hane et al., 2020) monomertrophs/biotrophs:

Table 3.1: Bioinformatics tools and methods integrated into the Predector pipeline. Non-default parameters are indicated where applicable.

Software	Description	Reference(s)
A) Localisation		
SignalP v3.0, 4.1g, 5.0b	Extracellular secretion via signal peptide prediction. Both NN and HMM methods are run for v3.0. Eukyaryotic types specified.	Armenteros, Tsirigos, et al. (2019), Bendtsen et al. (2004), Petersen et al. (2011)
Deepsig commit 69e01cb	Extracellular secretion via signal peptide prediction. `-k euk`	Savojardo et al. (2018)
Phobius v1.01	Extracellular secretion via signal peptide and transmembrane domain prediction.	Käll et al. (2004)
LOCALIZER v1.0.4	Host sub-cellular localisation prediction. Using predicted mature proteins from SignalP 5.0b. `-e -M`.	Sperschneider et al. (2017)
ApoplastP v1.0.1	Apoplast-specific localisation prediction.	Sperschneider, Dodds, Singh, et al. (2018)
DeepLoc v1.0	Sub-cellular localisation prediction.	Armenteros et al. (2017)
TargetP v2.0	Sub-cellular localisation prediction. `-org non-pl`.	Armenteros, Salvatore, et al. (2019)
TMHMM v2.0c	Membrane localisation via transmembrane domain prediction. `-d`.	Krogh et al. (2001)
B) Effector-like properties		
EffectorP v1.0, 2.0	Probabilistic prediction of effector likelihood.	Sperschneider, Dodds, Gardiner, et al. (2018), Sperschneider et al. (2016)
EMBOSS: pepstats v6.5.7	Amino acid properties and frequencies.	Rice et al. (2000)
C) Functional annotation		
HMMER v3.2.1 (vs dbCAN v8)	Used to identify putative CAZymes.	Eddy (2011), Zhang et al. (2018)
MMSeqs2 v10-6d92c (vs PHIBase v4.9)	Used to identify potential virulence factors. `--max-seqs 300 -e 0.01 -s 7 --num-iterations 3 -a`	Steinegger and Söding (2017), Urban et al. (2020)
MMSeqs2 v10-6d92c (vs known effectors in supplementary table S2)	Used to identify potential effector homologues. `--max-seqs 300 -e 0.01 -s 7 --num-iterations 3 -a`	Steinegger and Söding (2017)
PfamScan (vs Pfam v33.1)	To identify functional domains. With active site prediction. `-as`.	Finn et al. (2014)

Blumeria graminis f. sp. *hordei* (Frantzeskakis et al., 2018), *Blumeria graminis* f. sp. *tritici* (Müller et al., 2019), *Melampsora lini* (Nemri et al., 2014), *Melampsora larici-populina* (Duplessis et al., 2011), *Puccinia graminis* f. sp. *tritici* (Li et al., 2019); polymertrophs/necrotrophs - *Parastagonospora nodorum* (Syme et al., 2016), *Pyrenophora tritici-repentis* (Moolhuijzen et al., 2018), *Pyrenophora teres* f. *teres* (Syme, Martin, et al., 2018), and *Pyrenophora teres* f. *maculata* (Syme, Martin, et al., 2018); mesotrophs/hemibiotrophs - *Leptosphaeria maculans* (Dutreux et al., 2018), *Zymoseptoria tritici* (Goodwin et al., 2011; Plissonneau et al., 2018), *Passalora fulva* (de Wit et al., 2012), *Dothistroma septosporum* (de Wit et al., 2012); wilts/vasculartrophs - *Fusarium oxysporum* f. sp. *lycopersici* (DeIulio et al., 2018; Ma et al., 2010), *Fusarium oxysporum* f. sp. *melonis* (Ma et al., 2014); and saprotroph (or opportunistic monomertroph/biotroph) *Neurospora crassa* (MacCallum et al., 2009) (Supplementary table S2). Fourteen of the 24 proteomes above were retained as a separate dataset for final evaluation (Supplementary table S2). The remainder of the datasets were combined, and redundant sequences were removed to prevent the undue influence of conserved or well studied sequences with multiple records. Redundancy was reduced by clustering proteins with MMSeqs2 version 11-e1a1c (Steinegger & Söding, 2017) requiring a minimum reciprocal coverage of 70% and minimum sequence identity of 30% (`--min-seq-id 0.3 --cov-mode 0 -c 0.7 --cluster-mode 0`). A single sequence was chosen to represent a set of clustered, redundant sequences, which was prioritised based on supporting information (in order of preference): known effector, SwissProt secreted, SwissProt non-secreted, proteome/effector homologue, longest member of cluster. Clusters that corresponded to the known effectors from the EffectorP 2 (Sperschneider, Dodds, Gardiner, et al., 2018) training and test data sets were automatically assigned to training and test data sets in this study. A randomly selected subset of 20% of the remaining representative members of clusters were also assigned to the test dataset. Data and scripts for generating the datasets are available at https://github.com/ccdmb/predector-data.

3.3.3 Manual effector and secretion prediction scoring

Predicted proteins were ranked using the sum of several weight-adjusted scores derived from a range of software and methods (Table 3.1, Supplementary table S3). Proteins were annotated as "multiple_transmembrane" if it was assigned more than one transmembrane (TM) domain by either TMHMM or Phobius, and "single_transmembrane" if it was assigned one TM domain by TMHMM or Phobius (but neither had more than one). For TMHMM "single_transmembrane" we add the additional constraint that if there is a signal peptide (SP) prediction (by any method) and that the number of expected TM AAs in the first 60 residues is less than ten. A protein was annotated as "secreted" if it was predicted to have a SP by any method and was not annotated as a multiple TM protein.

Protein matches to PHI-base were summarised based on the experimental phenotypes of the matched proteins. Proteins were marked as a "phibase_effector_match" if they had any matches with the "Loss of pathogenicity", "Increased virulence (Hypervirulence)", or "Effector (plant avirulence determinant)" phenotypes; as a "phibase_virulence_match" if they had any

matches with the "Reduced virulence" phenotype and not any of the effector phenotypes; and as a "phibase_lethal_match" if they had any matches with the "Lethal" phenotype. Proteins were also labelled as "effector_match", "pfam_match", or "dbcan_match" if they had a significant match to a custom database of known effectors, selected virulence associated Pfam hidden markov models (HMMs), or selected virulence associated dbCAN HMMs, respectively (Supplementary table S2).

Each protein was given two manually designed scores to evaluate effector or secreted protein candidates based on the values and weights in supplementary table S3. The secretion score is the sum of the products of value and weight for TM, secreted, signalp3_hmm, signalp3_nn, phobius, signalp4, deepsig, targetp, and deeploc parameters. The effector score is the sum of the secretion score and the sum of the products of EffectorP, and the homology parameters (effector match, virulence match, and lethal match) values and weights.

3.3.4 Learning to rank model training

A "learning to rank" pairwise machine learning method based on LambdaMart (Wu et al., 2010) was developed using XGBoost (Chen & Guestrin, 2016) to prioritise effectors. Effector homologues in the training data set were held out as an informal validation set, known effector proteins were considered relevant (priority 2), and all other proteins in the train dataset were considered irrelevant (priority 1). To mitigate issues caused by unbalanced class sizes, training data were weighted for effectors as #irrelevant / #relevant and unlabelled proteins were given weight #relevant / #irrelevant. A subset of features output by the Predector pipeline and model constraints for the direction of effect (indicated in brackets as + or - when a constraint was applied; + indicating that increasing values of the feature can only contribute positively towards effector prediction) were selected based on the distributions of parameters in supplementary figures S3–S40: molecular weight, proportion of cysteines, proportion of tiny AAs (Gly, Ala, Ser and Pro), proportion of small AAs (Thr, Asp and Asn), proportion of non-polar AAs, proportion of basic AAs, EffectorP 1 probability (+), EffectorP 2 probability (+), ApoplastP probability (+), TMHMM TM count (-), TMHMM expected TM residues in first 60 AAs, Phobius TM count (-), DeepLoc membrane probability (-), DeepLoc extracellular probability (+), DeepSig SP prediction (+), Phobius SP prediction (+), SignalP 3 neural network D-score (+), SignalP 3 HMM S-score (+), SignalP 4 D-score (+), SignalP 5 SP probability (+), and TargetP secreted probability (+). The hyperparameters max_depth, min_child_weight, gamma, lambda (L2 regularisation), subsample (dropout), colsample_bytree, eta (learning rate), and num_boost_round (number of boosted trees) were optimised by maximising the normalised discounted cumulative gain (NDCG) (Y. Wang et al., 2013) for the highest 500 ranked proteins (NDCG@500) in 5-fold cross validated training. The final model was trained using the optimised hyper-parameters.

3.3.5 Model and score evaluation

The learning to rank model, manually designed scores, and EffectorP pseudo-probabilities were evaluated using rank summarisation statistics using the scikit-learn library (Pedregosa et al., 2011), which included the coverage error (the rank of the lowest scoring effector), label ranking average precision (LRAP)(average proportion of correctly labelled samples with a lower score than each position in the sorted results), the label ranking loss (the average number of results that are incorrectly ordered), and the normalised discounted cumulative gain (NDCG; the sum of all ranking priorities divided by the $\log_2$ of the rank position in the sorted list (DCG), normalised by the best theoretically possible DCG score) (Y. Wang et al., 2013). NDCG, LRAP, and label ranking loss were also evaluated for the top 50, 500, and 5000 proteins (indicated with the suffix @50, @500, or @5000). Additionally, to compare classification performance of the learn to rank model with the combined EffectorP and secretion prediction decisions, a decision threshold of 0 was set for the learn to rank model (with > 0 indicating an effector prediction), and the classification metrics precision (the proportion of predicted effectors that are labelled as true effectors), recall (the proportion of known effectors that are predicted to be effectors), accuracy (the fraction of correct predictions), balanced accuracy (the arithmetic mean of precision and recall for binary cases like this, and is less affected by unbalanced data-sets than accuracy), F1 score (the harmonic mean of precision and recall), and matthews correlation coefficient (MCC). For unbalanced datasets like the training set of effectors and non-effectors, MCC is considered a more reliable indicator of model performance than the other methods mentioned above (Chicco & Jurman, 2020). Additionally, to evaluate the performance at different decision thresholds, the precision, recall, and MCC were calculated for 100 score thresholds along the range of each score, and the receiver operating characteristic (ROC) curves were plotted.

For the effector ranking scores, only known effectors were used as the relevant (positive) set with the irrelevant (negative or unlabelled) set consisting of secreted, non-secreted, and proteomes. Because EffectorP is intended to be run on secreted datasets, ranking statistics were only calculated for the subset of proteins that were predicted to have a SP (by any method) and with fewer than two predicted TM domains (by either Phobius or TMHMM), and classification statistics were considered on both this secreted subset, and as a combined classifier (secretion and EffectorP prediction) on the whole datasets. For the secretion ranking score the positive set consisted of the known effectors and SwissProt secreted set, and the negative set was made of the SwissProt non-secreted proteins.

3.4 Results

To develop and evaluate the predector pipeline, a dataset of unprocessed fungal proteins was collected and split into train and test datasets (supplementary table S2). The datasets included redundancy reduced proteins of known fungal effectors (numbers of proteins in

train set: 125, and test set: 28), fungal proteins in the SwissProt database annotated as secreted (train: 256, test: 64) and non-secreted (train: 8676, test: 2169), and the whole proteomes from 10 well studied fungal genomes (train: 52224, test: 13056). The predector pipeline runs numerous tools related to effector and secretome prediction (Table 3.1). Benchmarking those tools against the set of confirmed effector proteins in the train dataset, it was observed that the secretion prediction tools were frequently correct with a small number of exceptions (Figure 3.1). Signal peptide (SP) prediction recall in the training dataset of known effectors ranged from 84% (DeepSig) to 92% (TargetP 2). SignalP 3, 4, 5, and Phobius generally predicted about 90% of effectors to have SPs (Figure 3.1). Transmembrane (TM) predictors were, as expected, generally not able to predict TM domains in confirmed effectors, with the few single TM predictions by TMHMM or Phobius likely to be mis-predictions within N-terminal SPs. In the case of TMHMM, all effectors with at least one TM domain had more than ten AAs predicted to be TM associated in the first 60 residues by TMHMM (Supplementary figure S39). Effector prediction tools (EffectorP 1 and 2) were also able to predict most, but not all, of the confirmed effector set. EffectorP correctly predicted 85.6% and 76.8% of effectors in the training dataset for versions 1 and 2, respectively. Evaluation of protein features that might allow for distinction between the different protein classes considered in this study (effectors, effector homologues, secreted proteins, non-secreted proteins, and unlabelled proteomes) identified twelve features that could be used effectively. These included: the proportion of cysteines, small, non-polar, charged, acidic, and basic amino acids; ApoplastP prediction; DeepLoc extracellular or membrane predicted localisations; molecular weight; EffectorP scores, and SP raw scores (supplementary figures S3–S40). Many of the protein properties identified here were also identified as important features in EffectorP (Sperschneider, Dodds, Gardiner, et al., 2018; Sperschneider et al., 2016).

To incorporate information from the selected features related to effector and secretion prediction, a pairwise learning to rank model was trained. The mean cross validated normalised discounted cumulative gain (NDCG) in the top 500 ranked predictions (NDCG@500) for the hyper-parameter optimised model was 0.925942 with standard deviation 0.009421, indicating high performance and little effect of substructure within the dataset. The mean NDCG@500 for the train sets within the cross validation was 0.886542 (std. dev. 0.015099), indicating that the model was not overfitting.

Benchmarked against a test set of confirmed effectors (Figure 3.2), the Predector model consistently gave higher scores to effector proteins, and also to homologues of confirmed effectors (those on which the model was not trained) than to non-secreted or unlabelled proteins. Secreted proteins from SwissProt tended to have intermediate scores centred around 0. Non-secreted and the unlabelled effectors were heavily skewed towards more negative scores, with a long tail that included some proteins with high scores (which in the case of proteomes was expected as this dataset was unlabelled). The test and train sets showed similar distributions of scores, though there tended to be slightly lower scores for known effectors in the test set.

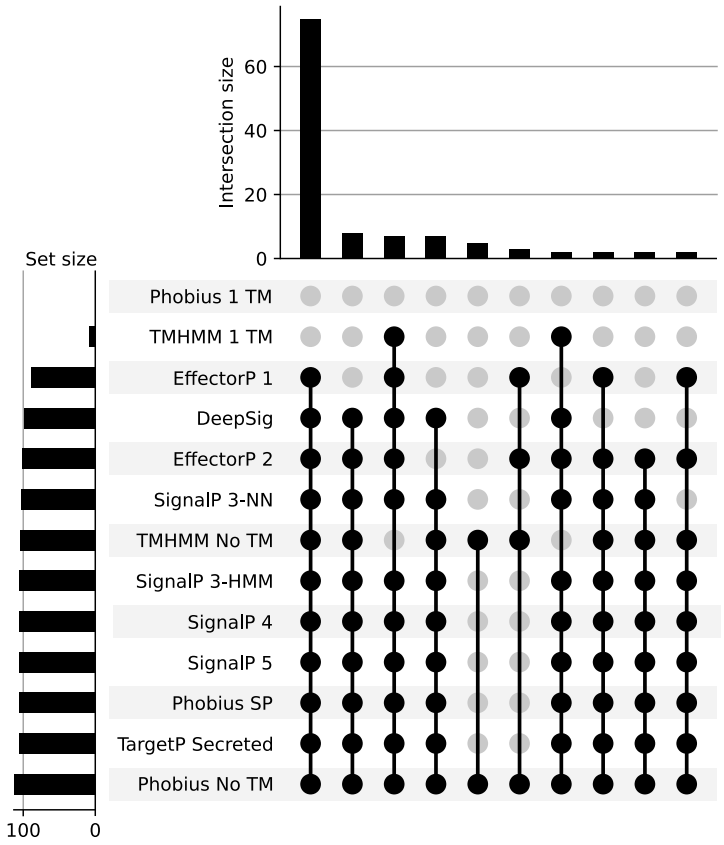


Figure 3.1: UpSet plot showing predictions of signal peptides, transmembrane domains, and effector-like properties for all known effectors in the training dataset (N=125). Rows indicate sets of proteins predicted to have a property related to effector prediction (e.g. a signal peptide), with the horizontal bar chart indicating set size. Columns indicate where the horizontal sets intersect with each other, where the vertical bar-chart indicates the number of proteins in that intersection. For clarity, intersections with only 1 member have been excluded.

The main features used for sorting effectors from non-effectors in the Predector model were TargetP secretion prediction, SignalP 3-HMM S-scores, SignalP4 D-scores, DeepLoc extracellular and membrane predictions, and EffectorP 1 and 2. TargetP secretion was overwhelmingly the most important feature according to the gain metric (the average increase in predictive score when the feature is used), which was consistent with the observation that it was the most sensitive of the SP prediction methods for effectors (Figure 3.1). The most commonly used predictors were EffectorP 2 pseudo-probabilities, molecular weight, and the proportions of cysteines, basic AAs, non-polar AAs and tiny AAs. Feature importances and boosted trees indicated overall that the Predector model first coarsely sorts proteins into the predicted secretome and non-secreted proteins, then proceeds to separate proteins with effector-like properties from the remainder of the secretome using more decision nodes each with smaller overall gain (Supplementary figure S43).

Predector separated some proteins predicted to be secreted (i.e. with a SP and fewer than two TM domains), from those that are not (Figure 3.3). Most “non-secreted” proteins have a

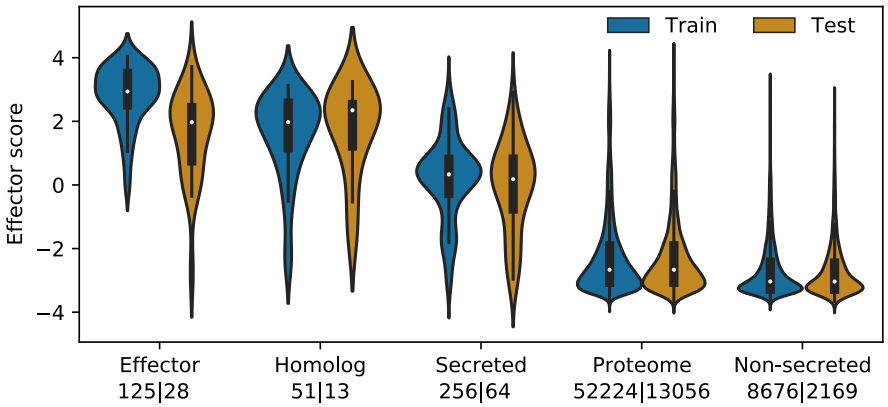


Figure 3.2: A violin plot showing the distributions of Predector effector ranking scores for each class in the test and training datasets. The effectors consist of experimentally validated fungal effector sequences. "Secreted" and "non-secreted" proteins are manually annotated proteins from the SwissProt database. Proteomes consist of the complete predicted proteomes from 10 well studied fungi (Supplementary table S2). The number of proteins represented by each violin are indicated on the x-axis.

score < 0, while a tri-modal distribution of "secreted proteins" was observed, which spanned the full range of scores and roughly coincided with the distributions of effectors/homologues, SwissProt secreted and the non-secreted/proteome datasets (Figure 3.2). This contrasted with EffectorP predictions (which was trained and is intended to be used on secretomes only), which gave poor separation of non-secreted and secreted proteins. EffectorP 1 showed a high bias to predicting proteins as either 0 or 1, indicating that it may be unsuited for ranking and should only be used as a decision classifier with a score threshold of 0.5. EffectorP 2 showed a more continuous separation of known effectors, and was moderately correlated with Predector scores for secreted proteins.

Predector consistently outperformed EffectorP 1 and 2 (restricted to the predicted secretome, as per intended usage) in classification recall and Matthews correlation coefficient, and in metrics assessing the ranked order of effector candidates (Table 3.2, Supplementary table S4). While EffectorP was not optimised for effector candidate ranking or intended to be used this way, we note that its probability score is likely to be used for this purpose. Conversely, although Predector was not intended to be used for effector classification, we also compared its predictive performance with EffectorP 1 and 2 on the secreted subset, and on the full dataset using the joint estimator of secretion and EffectorP score > 0.5. For the purpose of this comparison, a minimum Predector score of 0 was selected as a classification threshold based on the observation that the model assigns positive scores to effector associated branches in the trees (and negative scores to non-effector associated branches). EffectorP 1 and 2 performed identically in terms of effector classification on our test dataset, and gave highly similar results on the training dataset (Supplementary table S4, Supplementary figure S50), although fewer false positives were reported by EffectorP 2. Predector correctly predicted all but two effectors in the full test set, and all but one in the secreted test subset. In contrast, EffectorP 1 and 2

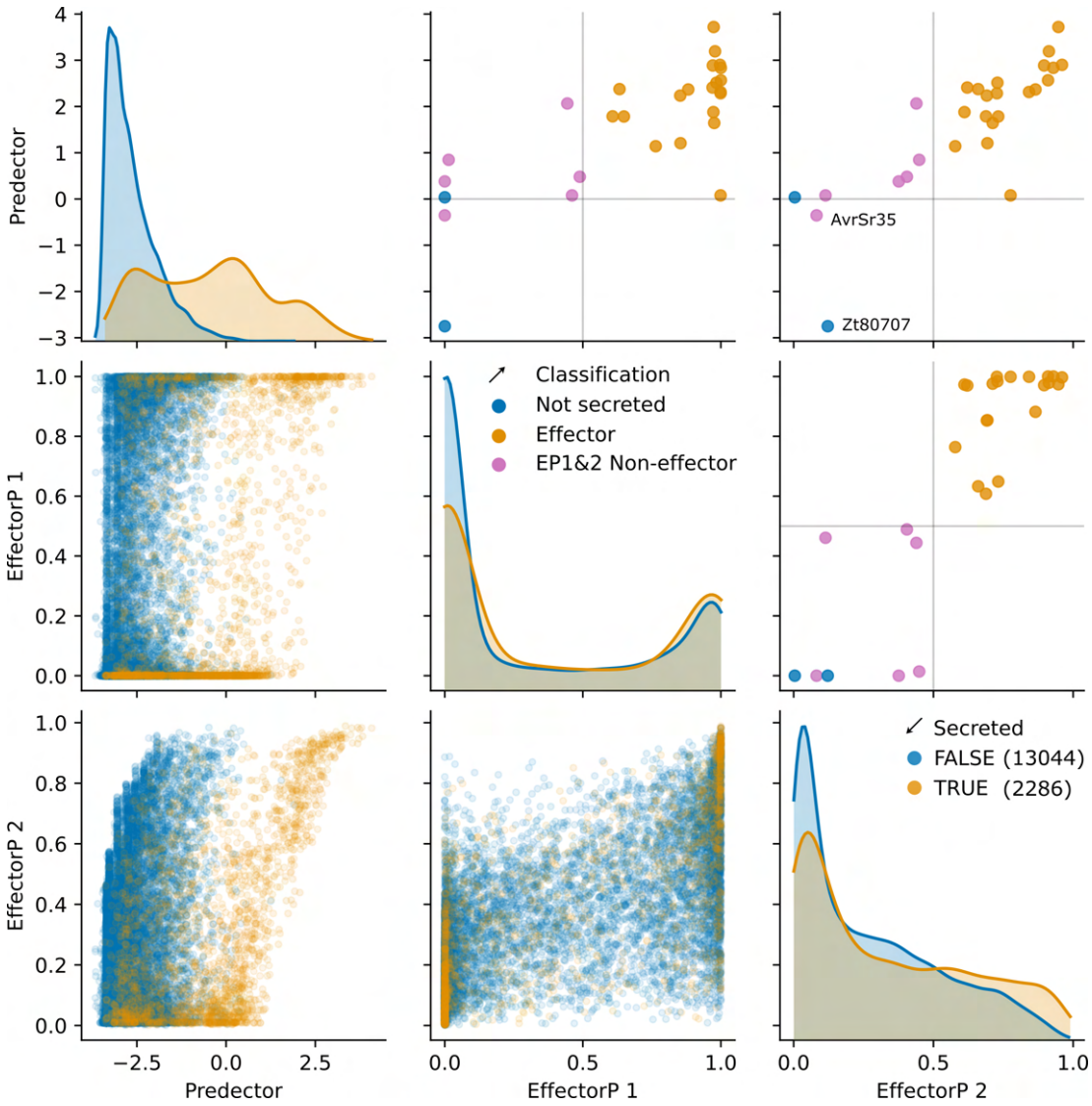


Figure 3.3: Comparing the scores of Predector with EffectorP versions 1 and 2. Scatter plots in the lower-left corner indicate comparisons of predictive scores between methods, with predicted secreted proteins (any signal peptide and fewer than two TM domains predicted) indicated in yellow, and non-secreted proteins indicated in blue. Density plots along the diagonal indicate distributions of the full test dataset versus predictive scores for each method (indicated along the x-axis), also coloured by secretion prediction as before (Note: there are far more non-secreted than secreted proteins in the dataset). Scatter plots in the top-right corner indicate score comparisons between methods for confirmed effectors, coloured by whether they have been predicted as secreted (criteria as above), or additionally predicted by EffectorP versions 1 or 2. Two proteins that are misclassified by a Predector score > 0 are labelled in the top-right subplot.

both mis-classified six effectors in the secreted subset, and two known effectors in the test dataset were not predicted to be secreted thus would have been excluded from prediction by an EffectorP pipeline. Predector also correctly predicted two confirmed effectors, FolSix12 and BghBEC3, that were not predicted to be secreted because they lacked a signal peptide prediction or had multiple TM domain predictions, respectively. Although Predector, not being optimised for classification, had a higher false positive rate than EffectorP 1 and 2, it compared favourably for the matthews correlation coefficient (MCC) metric which is considered more reliable for unbalanced datasets (Chicco & Jurman, 2020). It is worth noting that in this study secretion prediction incorporates multiple methods, whereas many studies rely on a single prediction tool, thus the proportion of potentially missed effector candidates may be higher than we report here.

Table 3.2: Effector prediction and ranking statistics for Predector and EffectorP on the test dataset. Note that EffectorP is not optimised for ranking tasks and Predector is not optimised for classification. These scores are shown merely for comparison and not necessarily as an endorsement of how they should be used. Coverage error is the index of the last known effector in the test dataset. NDCG is a measure of how often effectors are placed ahead of unlabelled samples in the list sorted by score, penalising incorrect orderings more highly near the top of the list. NDCG@N is the same statistic but only for the top N items in the sorted list. TP, TN, FP, FN are the number of true positives, true negatives, false positives, and false negatives for the classification task, respectively. Precision indicates how many of the predicted effectors are false positives (unlabelled in this case, so these could be real effectors), and recall indicates how many of the known effectors are correctly predicted as effectors. Balanced accuracy and MCC are better indicators of model predictive performance than precision for unbalanced data. The secreted test subset consists only of known effector proteins and proteins with a signal peptide (by any method) and fewer than two predicted TM domains (by either TMHMM or Phobius). Correct classification for EffectorP in the full dataset is conditional on secretion prediction by the same criterion as the secreted dataset (SP and < 2 TM). For the same reason, Predector and EffectorP cannot be fairly compared by ranking statistics in the full dataset.

		Full test dataset			Secreted test subset		
		EP1[b] & Sec[a]	EP2[c] & Sec	Predector	EP1	EP2	Predector
Ranking	Coverage error	-	-	8054	2275	1593	1115
	NDCG@50	-	-	0.640	0.615	0.629	0.652
	NDCG@500	-	-	0.928	0.916	0.926	0.933
	NDCG	-	-	0.447	0.365	0.402	0.448
Classification	TP	20	20	26	20	20	25
	TN	14450	14609	14317	1410	1569	1323
	FP	839	680	972	839	680	926
	FN	8	8	2	6	6	1
	Precision	0.023	0.028	0.026	0.023	0.028	0.026
	Recall	0.714	0.714	0.928	0.769	0.769	0.961
	Accuracy	0.944	0.955	0.936	0.628	0.698	0.592
	Balanced accuracy	0.829	0.834	0.932	0.698	0.733	0.774
	MCC	0.122	0.137	0.149	0.086	0.107	0.118

[a] Secreted [b] EffectorP 1 [c] EffectorP 2

For a set of 14 fungal proteomes retained separately for evaluation (Table 3.3, Supplementary table S4), Predector predicted on average 7.2% of proteins to have a score > 0, with an average of 6.4 effector homologues in the 50 highest scoring predictions. Predector processed whole fungal proteome datasets with an average rate of 1,814.67 proteins per hour on four CPUs, and 3,922.15 proteins per hour on 16 CPUs. DeepLoc is overwhelmingly the longest running task (~75 mins for 5000 proteins), so while the 16 CPUs take only half the time, much of it is idle waiting for DeepLoc to finish (dependent on when DeepLoc is scheduled to run) and run times can be improved with configuration.

Predector effector predictions and run time evaluation for results on proteomes held out of the training set. Runs with 4 CPUs are performed on a cloud instance running ubuntu 20.04 (4 AMD EPYC vcpus, 16Gb RAM), runs with 16 CPUs are performed on a partially occupied single HPC node (16 Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz, 48 Gb RAM). Both were carried out with default configuration (maximum within task parallelisation is 4 CPUs).

3.5 Discussion

The Predector pipeline unites, for the first time, numerous computational tasks commonly involved in effector and secretion prediction to determine a ranked set of candidate effectors from unprocessed (immature) proteins, simplifying complex data gathering steps. The effector ranking model run as part of Predector provides additional benefits over the standalone use of its composite tools, in combining their individual strengths while being less prone to their weaknesses. It was observed that while the most recently updated effector prediction tool available — EffectorP 2 (Sperschneider, Dodds, Gardiner, et al., 2018) — performed well as a very specific classifier, it still missed several confirmed effectors. The preliminary step of secretion prediction can also be error prone, and the combined false positives from both effector and secretion prediction methods, coupled with their common implementation as hard filters, may result in many genuine candidate effectors being discarded. For this reason, we propose that ranking and clustering methods should be preferred over hard filters for prioritising effector candidates.

In terms of effector candidate ranking, EffectorP 2 performed reasonably well for ordering confirmed effectors based on probability score, but was not designed to be used in this way. Predector maintained higher recall with higher scores (Table 3.2; Supplementary figure S46, S47) and achieved comparable or better precision than EffectorP 2 alone for higher effector scores. Thus, while Predector is not intended to be used as a classifier, we demonstrate its utility as a highly sensitive method for combined secretion and effector prediction, and suggest a decision threshold (score) of 0 for summarisation purposes alongside standard EffectorP and secretion classifiers (which can be obtained from Predector output). However, the appropriate threshold may change with future versions. Although the recall scores for Predector were very high, Predector also predicted 292 more false positives in the test dataset than the commonly used method of combining a predicted secretion hard filter with EffectorP

Table 3.3: Pedector effector predictions and run time evaluation for results on proteomes held out of the training set. Runs with 4 CPUs are performed on a cloud instance running ubuntu 20.04 (4 AMD EPYC vcpus, 16Gb RAM), runs with 16 CPUs are performed on a partially occupied single HPC node (16 Intel(R) Xeon(R) CPU E5-2680 v4 @ 2.40GHz, 48 Gb RAM). Both were carried out with default configuration (maximum within task parallelisation is 4 CPUs). The number of protein sequence similarity matches to known effectors and matches to Pfam domains with putative virulence functions are noted for the top 50 candidates.

Organism	CPUs	# proteins	Run time (h:m:s)	# >0[a]	#Effectors @ 50[b]	#Pfam @ 50[c]
Blumeria graminis f. sp. *hordei* RACE1	4	5317	3:12:54	366	12	0
Blumeria graminis f. sp. *tritici* 96224	4	8347	4:33:55	696	20	0
Dothistroma septosporum NZE10	4	12415	6:53:22	610	2	4
Fusarium oxysporum f. sp. *lycopersici* MN25	4	24733	12:41:24	1313	8	8
Fusarium oxysporum f. sp. *melonis* 26406	4	26719	13:47:00	1464	7	8
Leptosphaeria maculans G12-14	4	12678	6:56:57	821	5	10
Leptosphaeria maculans NzT4	4	14026	7:26:47	868	9	9
Melampsora larici-populina 98AG31	4	16372	8:39:07	1282	1	0
Puccinia graminis f. sp. *tritici* 21-0	4	37843	20:08:17	4169	6	0
Pyrenophora teres f. sp. *maculata* SG1	4	10571	6:50:25	981	3	2
Pyrenophora tritici-repentis M4	4	13795	7:49:00	850	6	8
Zymoseptoria tritici 1A5	16	12072	2:55:22	970	2	1
Zymoseptoria tritici 1E4	16	12023	3:33:47	981	5	1
Zymoseptoria tritici 3D1	16	11991	2:48:49	971	4	1

[a] Number of proteins with a Predector score greater than 0. [b] Number of effector homologs in the top-50 scoring proteins. [c] Number of proteins with virulence related PFam matches in the top-50 scoring proteins.

2 (Table 3.2). We argue that recall should be prioritised for effector prediction, as the unlabelled proteome datasets used here may contain genuine novel effectors, and the focus of Predector on ranking rather than classification mitigates some of the issues associated with lower precision. Encouragingly, we observed that Predector was capable of giving positive scores to known effectors which were not predicted to have a signal peptide (SP) by some methods or have transmembrane (TM) domain predictions (in both the train and test datasets) and thus would have failed to be predicted by alternate methods with a secretion prediction hard filter.

The predictive rankings provided by Predector are complemented with additional information that can be used to manually evaluate groups of effector candidates, and represents a comprehensive summary of various predicted types of proteins within a fungal proteome dataset, including candidate pathogenicity effectors, effector homologues, predicted secreted proteins, and carbohydrate-active enzymes (CAZymes) (Zhang et al., 2018). Predector reports the results of database searches against PHI-base, a curated set of known fungal effectors, Pfam domains, and dbCAN HMMs. We recommend that users examine the functionally annotated candidates closely, particularly with respect to homologues of confirmed effectors, prior to consideration of candidates ranked by Predector scores. Similarly, supplementation with experimental evidence or information derived from external tools and pipelines will further improve the utility of the Predector outputs, e.g. selection profiles derived from pan-genome comparisons (Schweizer et al., 2018; Syme, Tan, Rybak, et al., 2018), presence-absence profiles in comparative genomics, genome wide association studies, differential gene expression, or pathogenicity-relevant information relating to the genomic landscape: the distance to a DNA repeats, telomeres or distal regions of assembled sequences (Bertazzoni et al., 2018; Testa et al., 2016); or codon adaptation. By selecting indicators of general effector properties or molecular interactions of interest, and sorting these lists first by those functionally-guided features and then by Predector score(s), users gain a rich and clear guide for prioritising candidates before proceeding to more resource-expensive experiments (e.g. cloning or structure modelling).

Among known effectors there is considerable diversity of their molecular roles and functions. The modern plant pathology community has yet to come to firm agreement on the broad definition of an effector, or to refine a broader definition with effector sub-types. Effectors may promote virulence through directly targeting and disrupting host cell biological processes, including ribogenesis, photosynthesis or mitochondrial activity. In contrast various extracellular chitin-binding proteins have also long been described as effectors, yet promote virulence through passively protecting the pathogen cell from host pathogen and damage associated molecular pattern (PAMP and DAMP) recognition. CAZymes are not typically considered to act as effectors, yet there are several examples of secreted CAZymes that are reported as virulence factors or may be recognised by host major resistance (R)-genes (Urban et al., 2020). The focus of many effector prediction methods, including Predector, on biochemical or functional aspects of effector proteins also neglects the crucial contribution of host R- and susceptibility (S)-proteins in gene-for-gene interactions (and inverse GFG), which is best determined using well designed experiments. An inclusive predictive model

spanning diverse effector types may not offer a reliable pathway to rapid effector identification, rather they are likely to focus on general biochemical properties unrelated to necrotrophic or avirulence activities, e.g. that would enable the majority to interact with membranes and translocate into a host cell or to function in the apoplast. We present Predector as a reasonable compromise between functional diversity and common purpose, accounting for this inherent diversity through incorporation of multiple predictive methods. Additionally, with rapidly decreasing costs of genome sequencing and improvements to the automation of genome analysis and gene feature annotation, the availability and utility of fungal pathogen genomes is steadily increasing (Aylward et al., 2017). There is a growing need for tools which will minimise the effects of poor data quality control and ensure reproducibility and comparability across multiple genome resources. The Predector pipeline is an important time-saving tool which applies a standardised and reproducible set of tests for effector prediction.

3.6 Acknowledgements

This study was supported by the Centre for Crop and Disease Management, a joint initiative of Curtin University and the Grains Research and Development Corporation (Research Grant CUR00023). This research was undertaken with the assistance of resources and services from the Pawsey Supercomputing Centre and the National Computational Infrastructure (NCI), which is supported by the Australian Government. This research is supported by an Australian Government Research Training Program (RTP) Scholarship.

3.7 Data availability

Sequences and scripts used to generate training data are available online at https://github.com/ccdmb/predector-data. The Predector pipeline is available online at https://github.com/ccdmb/predector. The machine learning model used in this study is available online at https://github.com/ccdmb/predector-utils.

3.8 Supplementary material

All supplementary material is available online at https://doi.org/10.6084/m9.figshare.13325213.

Supplementary table S1. Examples of confirmed fungal plant pathogenicity effector proteins that do not exhibit the commonly targeted protein properties of low-molecular weight, cysteine-richness and presence of classical N-terminal secretion signal peptide.

Supplementary table S2. Datasets used for training and evaluation.

Supplementary table S3. Weights assigned for manual scores. Description of medskipameters used to calculate combined Predector scores, based on weight-adjusted values. Individual

scores were determined by multiplying the value by weight, and the combined Predector score was calculated from the sum of all individual scores.

Supplementary table S4. Extended model evaluation and statistics.

Supplementary figures. Captions for supplementary figures are contained in the separate document on figshare.

3.9 References

Anderson, J. P., Sperschneider, J., Win, J., Kidd, B., Yoshida, K., Hane, J., Saunders, D. G. O., & Singh, K. B. (2017). Comparative secretome analysis of *Rhizoctonia solani* isolates with different host ranges reveals unique secretomes and cell death inducing effectors. *Scientific Reports*, *7*(1), 10410. https://doi.org/10.1038/s41598-017-10405-y

Armenteros, J. J. A., Salvatore, M., Emanuelsson, O., Winther, O., Heijne, G. v., Elofsson, A., & Nielsen, H. (2019). Detecting sequence signals in targeting peptides using deep learning. *Life Science Alliance*, *2*(5). https://doi.org/10.26508/lsa.201900429

Armenteros, J. J. A., Sønderby, C. K., Sønderby, S. K., Nielsen, H., & Winther, O. (2017). DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics*, *33*(21), 3387–3395. https://doi.org/10.1093/bioinformatics/btx431

Armenteros, J. J. A., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., Heijne, G. v., & Nielsen, H. (2019). SignalP *5*.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology*, *37*(4), 420–423. https://doi.org/10.1038/s41587-019-0036-z

Aylward, J., Steenkamp, E. T., Dreyer, L. L., Roets, F., Wingfield, B. D., & Wingfield, M. J. (2017). A plant pathology perspective of fungal genome sequencing. *IMA Fungus*, *8*(1), 1–15. https://doi.org/10.5598/imafungus.2017.08.01.01

Beckerson, W. C., Vega, R. C. R. d. l., Hartmann, F. E., Duhamel, M., Giraud, T., & Perlin, M. H. (2019). Cause and Effectors: Whole-Genome Comparisons Reveal Shared but Rapidly Evolving Effector Sets among Host-Specific Plant-Castrating Fungi. *mBio*, *10*(6). https://doi.org/10.1128/mBio.02391-19

Bendtsen, J. D., Nielsen, H., von Heijne, G., & Brunak, S. (2004). Improved Prediction of Signal Peptides: SignalP 3.0. *Journal of Molecular Biology*, *340*(4), 783–795. https://doi.org/10.1016/j.jmb.2004.05.028

Bertazzoni, S., Williams, A. H., Jones, D. A., Syme, R. A., Tan, K.-C., & Hane, J. K. (2018). Accessories Make the Outfit: Accessory Chromosomes and Other Dispensable DNA Regions in Plant-Pathogenic Fungi. *Molecular Plant-Microbe Interactions*, *31*(8), 779–788. https://doi.org/10.1094/MPMI-06-17-0135-FI

Boutemy, L. S., King, S. R. F., Win, J., Hughes, R. K., Clarke, T. A., Blumenschein, T. M. A., Kamoun, S., & Banfield, M. J. (2011). Structures of *Phytophthora* RXLR Effector Proteins: A CONSERVED BUT ADAPTABLE FOLD UNDERPINS FUNCTIONAL DIVERSITY. *Journal of Biological Chemistry*, *286*(41), 35834–35842. https://doi.org/10.1074/jbc.M111.262303

Chen, T., & Guestrin, C. (2016). XGBoost: A Scalable Tree Boosting System, In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, New York, NY, USA, ACM. https://doi.org/10.1145/2939672.2939785

Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 6. https://doi.org/10.1186/s12864-019-6413-7

de Guillen, K. d., Ortiz-Vallejo, D., Gracy, J., Fournier, E., Kroj, T., & Padilla, A. (2015). Structure Analysis Uncovers a Highly Diverse but Structurally Conserved Effector Family in Phytopathogenic Fungi. *PLOS Pathogens*, *11*(10), e1005228. https://doi.org/10.1371/journal.ppat.1005228

DeIulio, G. A., Guo, L., Zhang, Y., Goldberg, J. M., Kistler, H. C., & Ma, L.-J. (2018). Kinome Expansion in the *Fusarium oxysporum* Species Complex Driven by Accessory Chromosomes. *mSphere*, *3*(3). https://doi.org/10.1128/mSphere.00231-18

de Wit, P. J. G. M., van der Burgt, A., Ökmen, B., Stergiopoulos, I., Abd-Elsalam, K. A., Aerts, A. L., Bahkali, A. H., Beenen, H. G., Chettri, P., Cox, M. P., Datema, E., de Vries, R. P., Dhillon, B., Ganley, A. R., Griffiths, S. A., Guo, Y., Hamelin, R. C., Henrissat, B., Kabir, M. S., … Bradshaw, R. E. (2012). The Genomes of the Fungal Plant Pathogens *Cladosporium fulvum* and *Dothistroma septosporum* Reveal Adaptation to Different Hosts and Lifestyles But Also Signatures of Common Ancestry. *PLOS Genetics*, *8*(11), e1003088. https://doi.org/10.1371/journal.pgen.1003088

Di Tommaso, P., Chatzou, M., Floden, E. W., Barja, P. P., Palumbo, E., & Notredame, C. (2017). Nextflow enables reproducible computational workflows. *Nature Biotechnology*, *35*(4), 316–319. https://doi.org/10.1038/nbt.3820

Duplessis, S., Cuomo, C. A., Lin, Y.-C., Aerts, A., Tisserant, E., Veneault-Fourrey, C., Joly, D. L., Hacquard, S., Amselem, J., Cantarel, B. L., Chiu, R., Coutinho, P. M., Feau, N., Field, M., Frey, P., Gelhaye, E., Goldberg, J., Grabherr, M. G., Kodira, C. D., … Martin, F. (2011). Obligate biotrophy features unraveled by the genomic analysis of rust fungi. *Proceedings of the National Academy of Sciences*, *108*(22), 9166. https://doi.org/10.1073/pnas.1019315108

Dutreux, F., Da Silva, C., d'Agata, L., Couloux, A., Gay, E. J., Istace, B., Lapalu, N., Lemainque, A., Linglin, J., Noel, B., Wincker, P., Cruaud, C., Rouxel, T., Balesdent, M.-H., & Aury, J.-M. (2018). De novo assembly and annotation of three Leptosphaeria genomes using Oxford Nanopore MinION sequencing. *Scientific Data*, *5*(1), 180235. https://doi.org/10.1038/sdata.2018.235

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLOS Computational Biology*, *7*(10), e1002195. https://doi.org/10.1371/journal.pcbi.1002195

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2014). Pfam: The protein families database. *Nucleic Acids Research*, *42*(Database issue), D222–D230. https://doi.org/10.1093/nar/gkt1223

Frantzeskakis, L., Kracher, B., Kusch, S., Yoshikawa-Maekawa, M., Bauer, S., Pedersen, C., Spanu, P. D., Maekawa, T., Schulze-Lefert, P., & Panstruga, R. (2018). Signatures of host specialization and a recent transposable element burst in the dynamic one-speed genome of the fungal barley powdery mildew pathogen. *BMC Genomics*, *19*(1), 381. https://doi.org/10.1186/s12864-018-4750-6

Galagan, J. E., & Selker, E. U. (2004). RIP: The evolutionary cost of genome defense. *Trends in Genetics*, *20*(9), 417–423. https://doi.org/10.1016/j.tig.2004.07.007

Gawehns, F., Ma, L., Bruning, O., Houterman, P. M., Boeren, S., Cornelissen, B. J. C., Rep, M., & Takken, F. L. W. (2015). The effector repertoire of *Fusarium oxysporum* determines the tomato xylem proteome composition following infection. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.00967

Gervais, J., Plissonneau, C., Linglin, J., Meyer, M., Labadie, K., Cruaud, C., Fudal, I., Rouxel, T., & Balesdent, M.-H. (2017). Different waves of effector genes with contrasted genomic location

are expressed by *Leptosphaeria maculans* during cotyledon and stem colonization of oilseed rape. *Molecular Plant Pathology*, *18*(8), 1113–1126. https://doi.org/10.1111/mpp.12464

Goodwin, S. B., M'Barek, S. B., Dhillon, B., Wittenberg, A. H. J., Crane, C. F., Hane, J. K., Foster, A. J., van der Lee, T. A. J., Grimwood, J., Aerts, A., Antoniw, J., Bailey, A., Bluhm, B., Bowler, J., Bristow, J., Burgt, A. v. d., Canto-Canché, B., Churchill, A. C. L., Conde-Ferràez, L., … Kema, G. H. J. (2011). Finished Genome of the Fungal Wheat Pathogen *Mycosphaerella graminicola* Reveals Dispensome Structure, Chromosome Plasticity, and Stealth Pathogenesis. *PLOS Genetics*, *7*(6), e1002070. https://doi.org/10.1371/journal.pgen.1002070

Hane, J. K., Paxman, J., Jones, D. A. B., Oliver, R. P., & de Wit, P. (2020). "CATAStrophy", a Genome-Informed Trophic Classification of Filamentous Plant Pathogens – How Many Different Types of Filamentous Plant Pathogens Are There? *Frontiers in Microbiology*, *10*, 3088. https://doi.org/10.3389/fmicb.2019.03088

Hane, J. K., Rouxel, T., Howlett, B. J., Kema, G. H., Goodwin, S. B., & Oliver, R. P. (2011). A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome Biology*, *12*(5), R45. https://doi.org/10.1186/gb-2011-12-5-r45

He, Q., McLellan, H., Boevink, P. C., & Birch, P. R. (2020). All roads lead to susceptibility: The many modes-of-action of fungal and oomycete intracellular effectors. *Plant Communications*, 100050. https://doi.org/10.1016/j.xplc.2020.100050

Human, M. P., Berger, D. K., & Crampton, B. G. (2020). Time-Course RNAseq Reveals *Exserohilum turcicum* Effectors and Pathogenicity Determinants. *Frontiers in Microbiology*, *11*. https://doi.org/10.3389/fmicb.2020.00360

Jiang, R. H. Y., Tripathy, S., Govers, F., & 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*, *105*(12), 4874–4879. https://doi.org/10.1073/pnas.0709303105

Jones, D. A. B., John, E., Rybak, K., Phan, H. T. T., Singh, K. B., Lin, S.-Y., Solomon, P. S., Oliver, R. P., & Tan, K.-C. (2019). A specific fungal transcription factor controls effector gene expression and orchestrates the establishment of the necrotrophic pathogen lifestyle on wheat. *Scientific Reports*, *9*(1), 1–13. https://doi.org/10.1038/s41598-019-52444-7

Jones, D. A., Bertazzoni, S., Turo, C. J., Syme, R. A., & Hane, J. K. (2018). Bioinformatic prediction of plant–pathogenicity effector proteins of fungi. *Current Opinion in Microbiology*, *46*, 43–49. https://doi.org/10.1016/j.mib.2018.01.017

Kale, S. D., Gu, B., Capelluto, D. G. S., Dou, D., Feldman, E., Rumore, A., Arredondo, F. D., Hanlon, R., Fudal, I., Rouxel, T., Lawrence, C. B., Shan, W., & Tyler, B. M. (2010). External Lipid PI3P Mediates Entry of Eukaryotic Pathogen Effectors into Plant and Animal Host Cells. *Cell*, *142*(2), 284–295. https://doi.org/10.1016/j.cell.2010.06.008

Käll, L., Krogh, A., & Sonnhammer, E. L. L. (2004). A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology*, *338*(5), 1027–1036. https://doi.org/10.1016/j.jmb.2004.03.016

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, *305*(3), 567–580. https://doi.org/10.1006/jmbi.2000.4315

Li, F., Upadhyaya, N. M., Sperschneider, J., Matny, O., Nguyen-Phuc, H., Mago, R., Raley, C., Miller, M. E., Silverstein, K. A. T., Henningsen, E., Hirsch, C. D., Visser, B., Pretorius, Z. A., Steffenson, B. J., Schwessinger, B., Dodds, P. N., & Figueroa, M. (2019). Emergence of the Ug99 lineage of the

wheat stem rust pathogen through somatic hybridisation. *Nature Communications*, *10*(1), 5068. https://doi.org/10.1038/s41467-019-12927-7

Liu, B., Chen, J., & Wang, X. (2015). Application of learning to rank to protein remote homology detection. *Bioinformatics*, *31*(21), 3492–3498. https://doi.org/10.1093/bioinformatics/btv413

Liu, L., Xu, L., Jia, Q., Pan, R., Oelmüller, R., Zhang, W., & Wu, C. (2019). Arms race: Diverse effector proteins with conserved motifs. *Plant Signaling & Behavior*, *14*(2), 1557008. https://doi.org/10.1080/15592324.2018.1557008

Liu, L., Huang, X., Mamitsuka, H., & Zhu, S. (2020). HPOLabeler: Improving prediction of human protein–phenotype associations by learning to rank. *Bioinformatics*. https://doi.org/10.1093/bioinformatics/btaa284

Lu, S., Gillian Turgeon, B., & Edwards, M. C. (2015). A ToxA-like protein from *Cochliobolus heterostrophus* induces light-dependent leaf necrosis and acts as a virulence factor with host selectivity on maize. *Fungal Genetics and Biology*, *81*, 12–24. https://doi.org/10.1016/j.fgb.2015.05.013

Lu, T., Yao, B., & Zhang, C. (2012). DFVF: Database of fungal virulence factors. *Database*, *2012*(0), bas032–bas032. https://doi.org/10.1093/database/bas032

Ma, L.-J., Shea, T., Young, S., Zeng, Q., & Kistler, H. C. (2014). Genome Sequence of *Fusarium oxysporum* f. sp. *melonis* Strain NRRL 26406, a Fungus Causing Wilt Disease on Melon. *Genome Announcements*, *2*(4). https://doi.org/10.1128/genomeA.00730-14

Ma, L.-J., van der Does, H. C., Borkovich, K. A., Coleman, J. J., Daboussi, M.-J., Di Pietro, A., Dufresne, M., Freitag, M., Grabherr, M., Henrissat, B., Houterman, P. M., Kang, S., Shim, W.-B., Woloshuk, C., Xie, X., Xu, J.-R., Antoniw, J., Baker, S. E., Bluhm, B. H., … Rep, M. (2010). Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. *Nature*, *464*(7287), 367–373. https://doi.org/10.1038/nature08850

MacCallum, I., Przybylski, D., Gnerre, S., Burton, J., Shlyakhter, I., Gnirke, A., Malek, J., McKernan, K., Ranade, S., Shea, T. P., Williams, L., Young, S., Nusbaum, C., & Jaffe, D. B. (2009). ALLPATHS 2: Small genomes assembled accurately and with high continuity from short paired reads. *Genome Biology*, *10*(10), R103. https://doi.org/10.1186/gb-2009-10-10-r103

Mesarich, C. H., Ökmen, B., Rovenich, H., Griffiths, S. A., Wang, C., Karimi Jashni, M., Mihajlovski, A., Collemare, J., Hunziker, L., Deng, C. H., van der Burgt, A., Beenen, H. G., Templeton, M. D., Bradshaw, R. E., & de Wit, P. J. G. M. (2018). Specific Hypersensitive Response–Associated Recognition of New Apoplastic Effectors from *Cladosporium fulvum* in Wild Tomato. *Molecular Plant-Microbe Interactions*, *31*(1), 145–162. https://doi.org/10.1094/MPMI-05-17-0114-FI

Moolhuijzen, P., See, P. T., Hane, J. K., Shi, G., Liu, Z., Oliver, R. P., & Moffat, C. S. (2018). Comparative genomics of the wheat fungal pathogen *Pyrenophora tritici-repentis* reveals chromosomal variations and genome plasticity. *BMC Genomics*, *19*(1), 279. https://doi.org/10.1186/s12864-018-4680-3

Mousavi-Derazmahalleh, M., Chang, S., Thomas, G., Derbyshire, M., Bayer, P. E., Edwards, D., Nelson, M. N., Erskine, W., Lopez-Ruiz, F. J., Clements, J., & Hane, J. K. (2019). Prediction of pathogenicity genes involved in adaptation to a lupin host in the fungal pathogens *Botrytis cinerea* and *Sclerotinia sclerotiorum* via comparative genomics. *BMC Genomics*, *20*(1), 385. https://doi.org/10.1186/s12864-019-5774-2

Müller, M. C., Praz, C. R., Sotiropoulos, A. G., Menardo, F., Kunz, L., Schudel, S., Oberhänsli, S., Poretti, M., Wehrli, A., Bourras, S., Keller, B., & Wicker, T. (2019). A chromosome-scale genome assembly reveals a highly dynamic effector repertoire of wheat powdery mildew. *New Phytologist*, *221*(4), 2176–2189. https://doi.org/10.1111/nph.15529

Nemri, A., Saunders, D. G. O., Anderson, C., Upadhyaya, N. M., Win, J., Lawrence, G., Jones, D., Kamoun, S., Ellis, J., & Dodds, P. (2014). The genome sequence and effector complement of the flax rust pathogen *Melampsora lini*. *Frontiers in Plant Science*, *5*. https://doi.org/10.3389/fpls.2014.00098

Ohm, R. A., Feau, N., Henrissat, B., Schoch, C. L., Horwitz, B. A., Barry, K. W., Condon, B. J., Copeland, A. C., Dhillon, B., Glaser, F., Hesse, C. N., Kosti, I., LaButti, K., Lindquist, E. A., Lucas, S., Salamov, A. A., Bradshaw, R. E., Ciuffetti, L., Hamelin, R. C., … Grigoriev, I. V. (2012). Diverse Lifestyles and Strategies of Plant Pathogenesis Encoded in the Genomes of Eighteen Dothideomycetes Fungi. *PLOS Pathogens*, *8*(12), e1003037. https://doi.org/10.1371/journal.ppat.1003037

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Petersen, T. N., Brunak, S., Heijne, G. v., & Nielsen, H. (2011). SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nature Methods*, *8*(10), 785–786. https://doi.org/10.1038/nmeth.1701

Plissonneau, C., Benevenuto, J., Mohd-Assaad, N., Fouché, S., Hartmann, F. E., & Croll, D. (2017). Using Population and Comparative Genomics to Understand the Genetic Basis of Effector-Driven Fungal Pathogen Evolution. *Frontiers in Plant Science*, *8*. https://doi.org/10.3389/fpls.2017.00119

Plissonneau, C., Hartmann, F. E., & Croll, D. (2018). Pangenome analyses of the wheat pathogen *Zymoseptoria tritici* reveal the structural basis of a highly plastic eukaryotic genome. *BMC Biology*, *16*(1), 5. https://doi.org/10.1186/s12915-017-0457-4

Praz, C. R., Bourras, S., Zeng, F., Sánchez-Martín, J., Menardo, F., Xue, M., Yang, L., Roffler, S., Böni, R., Herren, G., McNally, K. E., Ben-David, R., Parlange, F., Oberhaensli, S., Flückiger, S., Schäfer, L. K., Wicker, T., Yu, D., & Keller, B. (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*(3), 1301–1314. https://doi.org/10.1111/nph.14372

Qiu, J., Sheffler, W., Baker, D., & Noble, W. S. (2008). Ranking predicted protein structures with support vector regression. *Proteins: Structure, Function, and Bioinformatics*, *71*(3), 1175–1182. https://doi.org/10.1002/prot.21809

Rice, P., Longden, I., & Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. *Trends in Genetics*, *16*(6), 276–277. https://doi.org/10.1016/S0168-9525(00)02024-2

Richards, J. K., Stukenbrock, E. H., Carpenter, J., Liu, Z., Cowger, C., Faris, J. D., & Friesen, T. L. (2019). Local adaptation drives the diversification of effectors in the fungal wheat pathogen *Parastagonospora nodorum* in the United States. *PLOS Genetics*, *15*(10), e1008223. https://doi.org/10.1371/journal.pgen.1008223

Sánchez-Vallet, A., Hartmann, F. E., Marcel, T. C., & Croll, D. (2018). Nature's genetic screens: Using genome-wide association studies for effector discovery. *Molecular Plant Pathology*, *19*(1), 3–6. https://doi.org/10.1111/mpp.12592

Saunders, D. G. O., Win, J., Cano, L. M., Szabo, L. J., Kamoun, S., & Raffaele, S. (2012). Using Hierarchical Clustering of Secreted Protein Families to Classify and Rank Candidate Effectors of Rust Fungi. *PLOS ONE*, *7*(1), e29847. https://doi.org/10.1371/journal.pone.0029847

Savojardo, C., Martelli, P. L., Fariselli, P., & Casadio, R. (2018). DeepSig: Deep learning improves signal peptide detection in proteins. *Bioinformatics*, *34*(10), 1690–1696. https://doi.org/10.1093/bioinformatics/btx818

Schmidt, S. M., & Panstruga, R. (2011). Pathogenomics of fungal plant parasites: What have we learnt about pathogenesis? *Current Opinion in Plant Biology*, *14*(4), 392–399. https://doi.org/10.1016/j.pbi.2011.03.006

Schweizer, G., Münch, K., Mannhaupt, G., Schirawski, J., Kahmann, R., & Dutheil, J. Y. (2018). Positively Selected Effector Genes and Their Contribution to Virulence in the Smut Fungus *Sporisorium reilianum*. *Genome Biology and Evolution*, *10*(2), 629–645. https://doi.org/10.1093/gbe/evy023

Sonah, H., Zhang, X., Deshmukh, R. K., Borhan, M. H., Fernando, W. G. D., & Bélanger, R. R. (2016). Comparative Transcriptomic Analysis of Virulence Factors in *Leptosphaeria maculans* during Compatible and Incompatible Interactions with Canola. *Frontiers in Plant Science*, *7*. https://doi.org/10.3389/fpls.2016.01784

Spanu, P. D. (2017). Cereal immunity against powdery mildews targets RNase-Like Proteins associated with Haustoria (RALPH) effectors evolved from a common ancestral gene. *New Phytologist*, *213*(3), 969–971. https://doi.org/10.1111/nph.14386

Sperschneider, J., Catanzariti, A.-M., DeBoer, K., Petre, B., Gardiner, D. M., Singh, K. B., Dodds, P. N., & Taylor, J. M. (2017). LOCALIZER: Subcellular localization prediction of both plant and effector proteins in the plant cell. *Scientific Reports*, *7*(1), 1–14. https://doi.org/10.1038/srep44598

Sperschneider, J., Dodds, P. N., Gardiner, D. M., Singh, K. B., & Taylor, J. M. (2018). Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. *Molecular Plant Pathology*, *19*(9), 2094–2110. https://doi.org/10.1111/mpp.12682

Sperschneider, J., Dodds, P. N., Singh, K. B., & Taylor, J. M. (2018). ApoplastP: Prediction of effectors and plant proteins in the apoplast using machine learning. *New Phytologist*, *217*(4), 1764–1778. https://doi.org/10.1111/nph.14946

Sperschneider, J., Gardiner, D. M., Dodds, P. N., Tini, F., Covarelli, L., Singh, K. B., Manners, J. M., & Taylor, J. M. (2016). EffectorP: Predicting fungal effector proteins from secretomes using machine learning. *New Phytologist*, *210*(2), 743–761. https://doi.org/10.1111/nph.13794

Sperschneider, J., Williams, A. H., Hane, J. K., Singh, K. B., & Taylor, J. M. (2015). Evaluation of Secretion Prediction Highlights Differing Approaches Needed for Oomycete and Fungal Effectors. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.01168

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35*, 1026–1028. https://doi.org/10.1038/nbt.3988

Syme, R. A., Martin, A., Wyatt, N. A., Lawrence, J. A., Muria-Gonzalez, M. J., Friesen, T. L., & Ellwood, S. R. (2018). Transposable Element Genomic Fissuring in *Pyrenophora teres* Is Associated With Genome Expansion and Dynamics of Host–Pathogen Genetic Interactions. *Frontiers in Genetics*, *9*. https://doi.org/10.3389/fgene.2018.00130

Syme, R. A., Tan, K.-C., Hane, J. K., Dodhia, K., Stoll, T., Hastie, M., Furuki, E., Ellwood, S. R., Williams, A. H., Tan, Y.-F., Testa, A. C., Gorman, J. J., & Oliver, R. P. (2016). Comprehensive Annotation of the *Parastagonospora nodorum* Reference Genome Using Next-Generation Genomics, Transcriptomics and Proteogenomics. *PLOS ONE*, *11*(2), e0147221. https://doi.org/10.1371/journal.pone.0147221

Syme, R. A., Tan, K.-C., Rybak, K., Friesen, T. L., McDonald, B. A., Oliver, R. P., & Hane, J. K. (2018). Pan-Parastagonospora Comparative Genome Analysis—Effector Prediction and Genome Evolution. *Genome Biology and Evolution*, *10*(9), 2443–2457. https://doi.org/10.1093/gbe/evy192

Testa, A. C., Oliver, R. P., & Hane, J. K. (2016). OcculterCut: A Comprehensive Survey of AT-Rich Regions in Fungal Genomes. *Genome Biology and Evolution*, *8*(6), 2044–2064. https://doi.org/10.1093/gbe/evw121

Urban, M., Cuzick, A., Rutherford, K., Irvine, A., Pedro, H., Pant, R., Sadanadan, V., Khamari, L., Billal, S., Mohanty, S., & Hammond-Kosack, K. E. (2017). PHI-base: A new interface and further additions for the multi-species pathogen–host interactions database. *Nucleic Acids Research*, *45*(D1), D604–D610. https://doi.org/10.1093/nar/gkw1089

Urban, M., Cuzick, A., Seager, J., Wood, V., Rutherford, K., Venkatesh, S. Y., De Silva, N., Martinez, M. C., Pedro, H., Yates, A. D., Hassani-Pak, K., & Hammond-Kosack, K. E. (2020). PHI-base: The pathogen–host interactions database. *Nucleic Acids Research*, *48*(D1), D613–D620. https://doi.org/10.1093/nar/gkz904

Wang, C., Wang, P., Han, S., Wang, L., Zhao, Y., & Juan, L. (2020). FunEffector-Pred: Identification of Fungi Effector by Activate Learning and Genetic Algorithm Sampling of Imbalanced Data. *IEEE Access*, *8*, 57674–57683. https://doi.org/10.1109/ACCESS.2020.2982410

Wang, Y., Wang, L., Li, Y., He, D., Liu, T.-Y., & Chen, W. (2013). A Theoretical Analysis of NDCG Type Ranking Measures. *arXiv:1304.6480 [cs, stat]*. Retrieved July 2, 2020, from http://arxiv.org/abs/1304.6480

Williams, A. H., Sharma, M., Thatcher, L. F., Azam, S., Hane, J. K., Sperschneider, J., Kidd, B. N., Anderson, J. P., Ghosh, R., Garg, G., Lichtenzveig, J., Kistler, H. C., Shea, T., Young, S., Buck, S.-A. G., Kamphuis, L. G., Saxena, R., Pande, S., Ma, L.-J., … Singh, K. B. (2016). Comparative genomics and prediction of conditionally dispensable sequences in legume–infecting *Fusarium oxysporum* formae speciales facilitates identification of candidate effectors. *BMC Genomics*, *17*(1), 191. https://doi.org/10.1186/s12864-016-2486-8

Wu, Q., Burges, C. J. C., Svore, K. M., & Gao, J. (2010). Adapting boosting for information retrieval measures. *Information Retrieval*, *13*(3), 254–270. https://doi.org/10.1007/s10791-009-9112-1

You, R., Zhang, Z., Xiong, Y., Sun, F., Mamitsuka, H., & Zhu, S. (2018). GOLabeler: Improving sequence-based large-scale protein function prediction by learning to rank. *Bioinformatics*, *34*(14), 2465–2473. https://doi.org/10.1093/bioinformatics/bty130

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P. K., Xu, Y., & Yin, Y. (2018). dbCAN2: A meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, *46*(W1), W95–W101. https://doi.org/10.1093/nar/gky418

CHAPTER 4 — THEME 1

Remote homology clustering identifies lowly conserved families of effector proteins in plant-pathogenic fungi

This chapter has been revised after thesis submission.
The definitive peer reviewed, edited version of this article is published in:
Microbial Genomics, 2021, vol. 7, issue 9
https://doi.org/10.1099/mgen.0.000637

4.1 Declaration

Title	Remote homology clustering identifies lowly conserved families of effector proteins in plant-pathogenic fungi.
Authors	**Darcy A. B. Jones**, Paula Moolhuijzen, James K. Hane

This thesis chapter is submitted in the form of a collaboratively-written manuscript ready for journal submission. As such, not all work contained within this chapter can be attributed to the Ph. D. candidate.

The following contributions were made by the Ph.D candidate (DABJ) and co-authors:

- **DABJ** and JKH conceived the study.
- **DABJ** performed all software development, analyses, figure and table generation.
- **DABJ** and JKH wrote the manuscript.
- **DABJ**, JKH, and PM edited the manuscript.
- All authors read and approved the manuscript.

I, Darcy Jones, hereby certify that this attribution statement is an accurate record of my contribution to the research presented in this chapter.

_____

Darcy Jones

I certify that this attribution statement is an accurate record of Darcy Jones' contribution to the research presented in this chapter.

_____ _____

James K. Hane Paula Moolhuijzen

Abstract

Plant disease is initiated by molecular interactions between 'effector' molecules released by a pathogen and receptor molecules on or within the plant host cell. In many cases these effector-receptor interactions directly determine host resistance or susceptibility. The search for fungal effector proteins is a developing area in fungal plant pathology, with more than 165 distinct confirmed fungal effector proteins in the public domain. For a small number of these, novel effectors can be rapidly discovered across multiple fungal species through the identification of known effector homologs. However many have no detectable homology by standard sequence based search methods. This study employs a novel comparison method (RemEff) that is capable of identifying protein families with greater sensitivity than traditional blast-based orthology methods, thus converting a previously unusable pool of confirmed fungal effector data into a powerful tool for the prediction of novel fungal effector candidates through protein family associations. Resources relating to the RemEff method and data used in this study are available from https://figshare.com/projects/Effector_protein_remote_homology/87965.

4.2 Introduction

Fungal-plant pathogens expose or secrete molecules called 'effectors' into the extracellular environment, which may interact with or be internalised by their host, to promote infection. Hosts in turn may recognise pathogen associated molecular patterns (PAMPs) and initiate defense responses, which for the majority of pathogens confers innate immunity, termed pattern-triggered immunity (PTI) (Thomma et al., 2011). While some generalist pathogens employ a range of non-specific mechanisms to overcome these basal defenses, other pathogens adapt and specialise to a infect a narrower range of hosts by secreting proteinaceous or secondary metabolite effectors which usually bypass host defences or cause host cell death (de Wit et al., 2009). Necrotic effector activity has been observed to rely on the presence of a cognate susceptibility (S)-gene in the host genome (de Wit et al., 2009). However a second layer of defence, termed effector triggered immunity (ETI) may also be employed by the host, if its genome possesses a cognate resistance (R)-gene that confers the ability to activate host defences in the presence of an effector. Effectors are sometimes divided into subclasses based on their known interactions with host S and R genes, with necrotrophic effectors (NEs) interacting with S genes but having no known R genes, and avirulence effectors (AVRs) interacting with known R genes (de Wit et al., 2009; Thrall et al., 2016). Crop disease resistance breeding is usually conducted on the basis of introducing beneficial R genes and removing deleterious S genes. The study and discovery of fungal effectors among the growing pool of fungal genome data is vital for ongoing resistance breeding efforts (Vleeshouwers & Oliver, 2014), however there are a number of challenges that need to be overcome.

Proteinaceous fungal effectors have long been considered to lack sequence conservation, and in many cases have been presumed to have arisen independently. The collective term 'effector' is most frequently used to describe a highly diverse group of proteins with a common

but broadly defined role in virulence on a narrow range of hosts, but is sometimes also used to describe highly conserved families of pathogenicity proteins with broad host specificity, such as the NEPs (Oome & Van den Ackerveken, 2014), cerato-platanins (H. Chen et al., 2013), and Ribotoxins (Olombrada et al., 2017). Very little sequence homology has been observed between host-specific fungal effectors, potentially due to relatively high levels of genome plasticity in fungi (Bertazzoni et al., 2018; Hane et al., 2011; Testa et al., 2016). This is in direct contrast to effectors of a separate microbial lineage with less plastic genomes, the oomycetes, for which conserved effector motifs including RxLR-dEER (Anderson et al., 2015), and LXLFLAK (Crinklers/CRN) (Amaro et al., 2017) are commonly reported. Traditional biochemical and structural analyses are the gold standard for the functional characterisation of effector candidates (K. de Guillen et al., 2019) but are unsuitable for high-throughput analyses. Moreover, existing high-throughput experimental methods, such as proteomics and genome wide association studies, routinely return numerous genes or proteins that may be associated with the phenotype of interest, necessitating some additional information to prioritise future experimental validation.

High-throughput bioinformatic identification of fungal effector candidates remains a significant challenge due to the lack of homology among most fungal effectors (Jones et al., 2018). The vast majority of fungal proteins have no experimentally determined function and the accurate annotation of fungal genes is impeded by the narrow taxonomic range of fungal species with high quality gene annotation and by technical issues in deconvoluting transcriptome data (Testa et al., 2015). Nevertheless, a small but growing number of fungal effector families have been described with members in taxonomically distinct pathogens including: ToxA-like (Lu et al., 2015; Schmidt et al., 2016), MAX (K. d. de Guillen et al., 2015), RALPH (Praz et al., 2017; Spanu, 2017), and RXLR-like (Kale et al., 2010). In line with elevated fungal genome plasticity, these effector "families" share conserved structures but lack significant primary sequence similarity. This raises the possibility that at least some effectors — rather than arising independently or via lateral transfer — may have been vertically inherited from ancestor effector genes that were subsequently heavily mutated by fungal genome mutagenesis mechanisms such as repeat-induced point mutation (RIP) (Galagan & Selker, 2004). Among the currently identified effector families, conserved structural folds with similar functions can be observed, which are typically missed by simple sequence alignments. Effector family relationships with high sequence divergence are difficult to predict with traditional methods (e.g. BLASTp), but more sophisticated structural prediction and comparison methods (e.g. protein threading and structural alignment) are not yet computationally feasible to include in a high throughput analysis of a whole fungal proteome. Suitable alternatives come in the form of search methods that incorporate protein redundancy, such as profile-hidden markov models (HMMs) or position site specific models (PSSMs), which offer viable methods for finding remote homologues of confirmed effector proteins. Also of note are the cysteine-spacing classification systems that have been successfully applied to non-fungal cytotoxic venoms, which appear to have similar basic protein properties to fungal effectors (Kaas et al., 2012; Saucedo et al.,

2012). As our understanding of fungal effector biology improves, it may also become possible to apply similarly simple pattern-based heuristics for fungal effector classification.

4.2.1 Fungal effector protein families

Fungal pathogenicity effector proteins can be divided into those which: 1) form family groupings using simple bioinformatics methods, i.e. conserved motifs/patterns identified via simple sequence-based alignment (e.g. RxLR proteins), and 2) those which cannot be grouped by the above methods. In the case of the latter, there have been several studies to date piecing together a growing set of small cysteine-rich, secreted, low molecular weight, protein families with at least some members having effector-like phenotypes. There is remarkable diversity across these families, both between families and within them, yet common themes are emerging. Structural homology and in some cases similar modes of action (Lu et al., 2015) are observed between proteins with very low sequence identity, and some conserved or functional motifs appear to comprise surface-exposed, positively charged residues. These observations have led to the proposal of several families including multiple long-established effector proteins over the past decade. This study focuses on a set of well described fungal effector protein families for evaluation of methods, including the ToxA-like (Lu et al., 2015), MAX (K. d. de Guillen et al., 2015), RALPH and extracellular ribonucleases (Kettles et al., 2018; Praz et al., 2017; Spanu, 2017), and the AvrLm6-like proteins (Shiller et al., 2015), which are introduced in greater detail below.

ToxA

The ToxA-like family is named after the ToxA effector originally characterised in the wheat pathogen Pyrenophora tritici-repentis (Ballance et al., 1989; Tuori et al., 1995), and for which putatively horizontally-transferred loci were later identified by varying degrees of sequence homology of the locus and a ~14 Kbp flanking region (Friesen et al., 2006; McDonald et al., 2018; McDonald et al., 2019; Moolhuijzen et al., 2018), to genomes of other cereal-pathogenic fungi *Parastagonospora nodorum* (Friesen et al., 2006), *Bipolaris maydis* (Lu et al., 2015) and *B. sorokiniana* (Friesen et al., 2018; McDonald et al., 2018). The full PtrToxA pre-pro-protein is 178 aa in length, with a signal peptide (SP) cut site at position 22-23, and an N-terminal pro-peptide with a conserved 'LXXR' motif (Lu et al., 2015) which is cleaved during secretion at positions 60-61, producing the mature ToxA protein that corresponds to position 61 to 178 (Ballance et al., 1989; Ciuffetti et al., 1997; Sarma et al., 2005). PtrToxA (and the identical PnToxA) interact with a NBS-LRR domain membrane protein Tsn1, which confers host sensitivity to ToxA (Z. Liu et al., 2006). Within wheat host cells, mature PtrToxA is reported to bind two chloroplast-localised proteins: ToxA binding protein 1 (ToxABP1, syn. *Triticum aestivum* thylakoid formation protein TaThf1), plastocyanin protein TaPCN (Tai et al., 2007) and TaPR-1-5 PR-1-5 (Lu et al., 2014). ToxA-mediated disruption of chloroplast function leads to host cell necrosis, which requires light (Manning et al., 2009) and conservation of a structural loop possessing an 'RGD' motif.

The ToxA homolog of *B. maydis* (syn. *Cochliobolus heterostrophus*), ChToxA, has poor sequence similarity (64%) with Ptr/PnToxA, but has highly conserved structural homology (Lu et al., 2015) and a similar light-dependent necrosis phenotype on maize. Despite the similar structure, the 'RGD' motif required by Ptr/PnToxA for necrosis of wheat is substituted with a 'SGN' motif (Lu et al., 2015). Broadened similarity searches using HMM-based methods have predicted many other ToxA-like proteins across the classes Dothideomycetes and Sordariomycetes (Lu et al., 2015), including Avr2 of *Fusarium oxysporum*. Like ChToxA, Avr2 has a virulence-promoting phenotype, poor sequence identity with Ptr/PnToxA (~5%), and high structural similarity (Di et al., 2017). There are however, a few motifs that are conserved across the currently predicted members of the ToxA-like effector family, including: the 'LXXR' motif within the pro-domain, three motifs located in beta sheets 4, 5 and 8 (LXVXIXN, LILTXY, WXXQ respectively), and an asparagine-rich WXXN(S)NXIXVXI motif (Lu et al., 2015).

4.2.2 MAX

The Magnaporthe Avrs and ToxB-like (MAX) effector family comprises another set of fungal proteins that are structurally conserved but divergent at the sequence level. The MAX family was originally derived from effectors of *Magnaporthe oryzae* (K. d. de Guillen et al., 2015). Similarity of NMR structures containing two anti-parallel 3-stranded beta sheets with a single disulfide bond has been demonstrated between *M. oryzae* AVR-Pia, AVR1-CO39, AvrPiz-t and *Pyrenophora tritici-repentis* ToxB (K. d. de Guillen et al., 2015; Nyarko et al., 2014). Sequence alignment, position-specific score matrix (PSSM) and profile-HMM searches against these structural homologs had subsequently revealed numerous homologs in other species, including: *P. bromi* (Andrie et al., 2008), *Bipolaris oryzae*, *Colletotrichum* spp., *Zymoseptoria tritici*, *Leptosphaeria maculans* and even weak homology to plant-associated bacteria *Pseudomonas* sp. *StFLB209* (K. d. de Guillen et al., 2015). Multiple paralogs of members of this family have also been reported for some species, including *Pyrenophora* spp. (Martinez et al., 2001; Moolhuijzen et al., 2018), *C. fioriniae*, *C. orbiculare* and *C. gloeosporioides* (K. d. de Guillen et al., 2015), suggesting the potential for duplication and diversification of the relatively broadly-conserved MAX effector family.

4.2.3 AvrLm6

AvrLm6 is a well characterised AVR effector of the brassica pathogen *Leptosphaeria maculans*, which causes necrosis but has an avirulent phenotype in *Brassica napus* and *B. juncea* hosts (M. H. Balesdent et al., 2002) possessing the R locus *Rlm6*. Several AvrLm6-like homologs have been reported in other fungal pathogen species, including: *Colletotrichum spp.*, *Fusarium oxysporum*, *L. biglobosa*, and *Venturia spp.* (Grandaubert et al., 2014; Shiller et al., 2015). Notably in *V. inaequalis* and *V. pirina*, this family has undergone extensive clonal expansion (Shiller et al., 2015).

4.2.4 Ribotoxins, RIPs and RALPHs

Fungi secrete a broad variety of toxic and non-toxic ribonucleases (RNases) into the extracellular space and host (Lacadena et al., 2007). One set of cytotoxic RNases, the ribotoxins, are a group of fungal proteins that target the sarcin-ricin loop (SRL) of the host ribosome. This cleaves a single phosphodiester bond of the ribosomal ribonucleic acid (RNA), rendering it catalytically inactive and ultimately causing cell death (Glück & Wool, 1996; Olombrada et al., 2017). Fungal secreted RNases tend to share a common α-helix β-sheet fold topology, but differ in their terminal and loop domains (Lacadena et al., 2007; Olombrada et al., 2017). Ribotoxins possess an extended positively charged loop that non-cytotoxic secreted RNases lack, which is presumed to be important for interaction with the host-SRL (Glück & Wool, 1996; Lacadena et al., 2007). Ribotoxins are well documented in entomopathogens of the Ascomycetes (e.g. *Aspergillus giganteus* α-sarcin and *Aspergillus restrictus* restrictocin (Martínez-Ruiz et al., 1999)) and are also found in Basidiomycetes (e.g. white-rot *Agrocybe aegerita* (Citores et al., 2019; Landi et al., 2017)). Aspergillus ribotoxins are relatively well conserved (Martínez-Ruiz et al., 1999), however members from other genera (e.g. *Hirsutella thompsonii* - Hirsutellin and *Agrocybe aegerita* - Ageritin) share low sequence identity with the Aspergillus varieties but retain similar structures and activities (Herrero-Galán et al., 2008; Landi et al., 2017). Recently a cytotoxic secreted RNase protein, Zt6, was reported in the wheat pathogen *Zymoseptoria tritici* (Kettles et al., 2018). Although SRL binding has not yet been demonstrated for Zt6, it structurally resembles canonical ribotoxins and has RNase catalytic activity and exhibits toxicity to plants, some fungi and bacteria, but not to *Z. tritici* (Kettles et al., 2018).

Another group of non-toxic RNases have been reported in the *Blumeria* genus of biotrophic plant pathogens. *Blumeria* possesses several large families of effector candidates, with one of the largest groups containing RNase-associated domains with predicted structural similarity to RNase proteins - the RNase-like proteins associated with haustoria (RALPHs) (Pedersen et al., 2012; Spanu, 2017). RALPH effectors include *AvrPm2* (*BgtE-5845*) (Praz et al., 2017), AVRa13 (aka CSEP0372), BEC1011 (aka CSEP0264), and BEC1054 (aka CSEP0064) (Pennington et al., 2019; Spanu, 2017). Like many other mildew effectors the RALPHs possess a conserved 'Y(x)xC' motif after the signal peptide (Pedersen et al., 2012) as well as a RALPH-specific 'RxFP' motif, which may have roles in protein localisation or virulence (Praz et al., 2017). Like Ribotoxins, some RALPHs appear to bind the ribosomal SRL but lack a catalytic site for mRNA cleavage. They have been proposed to have a protective function against host ribosome inactivating ribonuclease (RIP) which may be induced as part of a resistance response (Pennington et al., 2019).

4.2.5 Prior efforts in remote homology

Fungal effectors do not frequently exhibit detectable sequence similarity to other known sequences, thus finding novel effector candidates in the form of distant homologues is challenging, and may involve relaxing BLAST e-values beyond recommended limits (K. d. de

Guillen et al., 2015; Deng et al., 2017). However, a range of more sensitive sequence-based search techniques are available which can exploit sequence features that may indicate conserved tertiary structures. Classification systems using cysteine spacing are well established for antimicrobial peptides and some venoms/toxins, where the number of and distance between cysteine residues indicates a possible shared topology of disulphide bonds (Islam et al., 2018; Linial et al., 2017; Robinson & Norton, 2014). However, conserved cysteine patterns are not guaranteed to indicate common structure or function (Saucedo et al., 2012), and known functional domains or discriminative motif analysis may also be necessary to separate active from non-functional forms (Asgari et al., 2019; Negi et al., 2017). Although they have been useful heuristic in other applications, cysteine spacing classification generally requires prior knowledge of a well-defined family, which would limit their application to effector family discovery.

Generally, *de novo* remote homology detection falls into two camps: iterative searches and alignments utilising sequence information from similar proteins (profile search methods) (Park et al., 1997), and machine learning methods which map the sequence into a multidimensional space (called an embedding, sequence space, or feature space) and perform a classification or ranking task. The latter form may use relatively simple sequence features such as kmers and sequence auto-correlation/covariance features (J. Chen et al., 2016), or may themselves use profile search results to construct a redundant representation of the sequence (Li et al., 2017; Rangwala & Karypis, 2005). Although these methods can achieve excellent results, they lack some of the interpretability of classical sequence search methods, and are still not in general use. Sequence based searches are much more commonly used, and the profile sequence based methods like PSSMs (e.g. PSI-BLAST Altschul et al., 1997) or profile-HMMs (e.g. HMMER Mistry et al., 2013) can find protein homologues with less than 30% sequence identity. Even more divergent homologues can be found using profile HMM-HMM comparisons (Steinegger et al., 2019) or Markov random fields (MRFs) (Ma et al., 2014). These more sensitive methods can be relatively computationally intensive and some pipelines for remote homology detection will first run PSSM based methods to reduce run-time (Szklarczyk et al., 2012; Wagner et al., 2014).

Many of these remote homology detection methods are designed to find homologs of a single protein, commonly as a precursor for *in silico* structural modelling, and are not always applicable to the task of protein family identification. Identification of protein families typically involves an all-vs-all comparison between proteins and the construction of a graph (aka network) from significant alignments, from which families can be identified as subgraphs (Tatusov et al., 1997). The best known and still most commonly used algorithm for finding subgraphs corresponding to protein families is by Markov clustering (MCL), used in TRIBE-MCL (Enright et al., 2002). More recent heuristic algorithms that don't require all-vs-all comparisons have been investigated (Petegrosso et al., 2019), but are yet to gain widespread use or a stable toolset.

In this study we apply a combination of protein clustering methods to investigate the

possibility of extending fungal effector protein families from the currently known set of fungal effectors. We use an agglomerative clustering approach with iteratively increasing sensitivity to find clusters of protein groups that show differing levels of homology, which we have termed RemEff. These groups highlight previously unreported relationships between several known effectors, the presence of large effector families, and will support future studies of fungal effector function and evolution. RemEff and its data from this study will also serve as an important resource in the field of molecular plant pathology for reliable effector candidate prediction, with relevance to multiple fungal plant pathogen species.

4.3 Methods

4.3.1 Data sets

Non-redundant fungal protein datasets were downloaded from the UniParc database (https://www.uniprot.org/uniparc/, filter: 'taxonomy:"Fungi (9FUNG) [4751]"', downloaded 2020-01-24) and the NCBI Identical Protein Groups database (https://www.ncbi.nlm.nih.gov/ipg/, filter: '"Fungi"[Organism] OR fungi[All Fields]', downloaded 2020-01-28) totalling 10,946,400 and 11,351,342 proteins, respectively. Data were supplemented using published genomes from JGI MycoCosm (https://genome.jgi.doe.gov/mycocosm/home), an Endophyte genome database (http://www.endophyte.uky.edu/) (L. Chen et al., 2015; Gao et al., 2011; Pan, 2014; Schardl et al., 2013; Schardl et al., 2014), the Alternaria genome database (http://alternaria.vbi.vt.edu/) (Dang et al., 2015), and the "Gemo" database (http://genome.jouy.inra.fr/gemo/) (Chiapello et al., 2015). Additional genomes, proteomes, and effector sequences collected from selected papers were included if they were not represented in the databases (Supplementary table S1).

Datasets were combined to give a single non-redundant dataset using "seguid" checksums (Babnigg & Giometti, 2006) implemented in BioPython (Cock et al., 2009). Proteins were filtered by length, including only proteins longer than 30 aa and shorter than 6000 aa. Unique sequences corresponding to published effectors and PHI-base (Urban et al., 2020) entries were identified by searching the initial dataset using MMSeqs2 (version 10-6d92c) (Steinegger & Söding, 2017), requiring a minimum sequence identity of 90% and at least 90% reciprocal coverage, selecting the match with the highest bit-score.

4.3.2 Clustering

The non-redundant fungal protein set was clustered in multiple stages using "MMSeqs2" (version 10-6d92c) (Steinegger & Söding, 2017). Protein sequences were initially clustered using the "cascade" clustering pipeline in three steps to a minimum of 30% sequence identity and 80% coverage of all members. To group more distant sequences, a second stage of clustering was performed using sequence profiles. Clusters were converted to sequence profiles and the profiles were enriched using the original input dataset of fungal proteins (including those sequences$\leq$30 aa or $\geq$6000 aa) to include information from sequences that didn't pass the

coverage threshold. The enriched profiles were searched against consensus sequences from the cluster profiles, and were clustered to have a minimum of 10% identical AAs and 70% reciprocal coverage. In further analyses in this study, these resulting clusters are referred to as "cluster level 1".

Multiple sequence alignments (MSAs) for each cluster's sequences were constructed using DECIPHER version 2.10 (Wright, 2015) using the PFASUM15 substitution matrix (Keul et al., 2017), 2 iterations, 2 refinement iterations, and alignment adjustments with staggering. A consensus sequence was added to the MSAs using DECIPHER, where columns with more than 50% gaps were considered gaps in the consensus. Code used for clustering sequences and constructing MSAs is available at https://github.com/darcyabjones/pclust.

4.3.3 Remote homology comparison

To find "low-level" sequence similarity between level 1 clusters, profile hidden markov model (HMM)-HMM searches were performed. MSAs with consensus sequences were first converted to MMSeqs2 profiles (`--match-mode 1 --match-ratio 1`) and enriched by searching against a database consisting of all fungal sequences of UniRef-90 (downloaded 2020-01-30. Query: 'taxonomy:"Fungi [4751]" AND identity:0.9') and the entire UniRef-50 database (downloaded 2020-01-30), selecting matches with a maximum e-value of 10^{-5}. Cluster MSAs and MSAs constructed from the profile matches were combined and converted into an HH-Suite database (version 3.2.0) (Steinegger et al., 2019). "Match" states in the HMMs and A3M alignments were determined by the consensus sequences of the cluster MSAs prior to enrichment (`--match-ratio=first`), where gaps in the consensus represent an insertion in the model.

To reduce computational requirements and to focus on fungal effectors, a subset of clusters were found by searching selected sequences of known effectors and virulence factors from numerous pathogens included in PHI-base version 4.8 (Urban et al., 2020), and a custom database of known effector sequences and homologues (Supplementary table 3). PHI-base entries to use for subsetting were selected based on annotated phenotypes, functional descriptions, and secretion prediction by SignalP versions 3, 4.1g, and 5.0b (Armenteros et al., 2019; Bendtsen et al., 2004; Petersen et al., 2011), DeepSig version 1 (Savojardo et al., 2018), Phobius version 1.01 (Käll et al., 2004), and TMHMM version 2.0c (Krogh et al., 2001). The sequences were first enriched into MSAs using the cluster HMMs, using two HHBlits search iterations. The enriched sequence MSAs were then searched against the cluster HMMs allowing a maximum e-value of 0.01, minimum probability of 0.20, and realigning up to 20000 matches (`-n 1 -e 0.01 -p 20 -Z 20000 -z 0 -B 20000 -b 0`). Code used for constructing HMMs, subsetting the database, and performing all-vs-all comparison is available at https://github.com/darcyabjones/pclust.

To identify remotely homologous clusters (referred to here as level 2 clusters or superclusters), the subset of HMMs matching selected phibase or effector sequences were searched against themselves (all-vs-all) using HHblits (`-n 1 -e 0.01 -E 0.01 -z 0 -Z 20000 -b 0 -B 20000 -pre_evalue_thresh 10 -min_prefilter_hits 10 -realign_max 20000`). Pairwise matches were considered significant if they had an e-value <= 10^{-5}, probability >= 0.9,

alignment length >= 30 AAs, and where the alignment covered at least 70% of at least one HMM in the pair. Where there were multiple alignments between the same pair of proteins, the alignment with the highest score was selected. Alignments were then filtered so that only reciprocally significant matches were retained. To reduce any score bias in alignments caused by HMM lengths, we adopted the normalisation approach used by OrthoFinder (Emms & Kelly, 2015) with modifications. Briefly, HMM search self matches were selected from the alignments, the HMM length was squared, the selected alignments were sorted by the squared HMM length, and the top 5% of alignments (by score) were selected from non-overlapping 1000 element sized bins in the sorted list. The $\log_2$ transformed alignment scores were regressed on the $\log_2$ transformed squared HMM lengths, and the slope and intercept were taken to transform scores using the same formula described in Emms and Kelly (2015). Conceptually, this transformation normalises the scores by the average maximum possible score for an alignment of two proteins with those lengths. Alignments were then further filtered to require alignments between both HMMs to be covered at least 70% of their respective lengths, and reciprocal matches were selected again. Each pair of alignments were grouped and the arithmetic mean of the two normalised scores was used as a single score for each pair, and the scores were converted to a value between 0 and 1 by dividing by the highest score in the full set of pairwise matches.

The filtered, score-normalised alignments were used to construct a weighted, undirected graph (AKA network) using the Python libraries, networkx (Hagberg et al., 2008), Pandas (McKinney, 2010), and SciPy (Virtanen et al., 2020). Clusters (superclusters) in the graph were found using a reimplementation of the greedy set cover algorithm (Hauser et al., 2016) and with Markov clustering (MCL) (Enright et al., 2002) (https://github.com/GuyAllard/markov_clustering), which due to their relatively similar stringencies were designated "cluster level 2A" and "cluster level 2B", respectively. Connected components were also found to summarise higher order relationships, which were also designated as "cluster level 3". The MCL inflation parameter was selected by running MCL on 10 randomly selected connected components containing 300-600 nodes, for a range of inflation parameters between 1.1 and 2.0. The inflation parameter that gave consistently higher modularity scores (Malliaros & Vazirgiannis, 2013) was selected for overall clustering. Graphs and subgraphs were visualised using the Graph Tool Python library (Peixoto, 2014). All code used to normalise alignment scores and perform higher-level clustering is provided in Supplementary data S6.

4.3.4 Supercluster comparison

To interrogate clusters within and between cluster levels 2 and 3, composite multiple sequence alignments were constructed and visualised as sequence logos (Schneider & Stephens, 1990). Enriched multiple sequence alignments from level 1 clusters (used to form HMMs) were combined using a progressive algorithm, guided by the maximum spanning tree of the subgraph containing the clusters of interest, where MSAs were pairwise aligned using hhalign (Steinegger et al., 2019). For each pairwise alignment, two alignments were computed using each MSA as the template. A pairwise alignment was considered to have succeeded if the resulting MSA

contained sequences from both input MSAs, and if both pairwise MSAs succeeded the result containing more sequences was selected for further iterations. If an alignment failed to merge two MSAs, the alignment was scheduled to be retried after all other alignments had been completed, stopping if no more MSAs could be merged. Un-connected components from this progressive method were then pairwise aligned in random order, shuffling the list if no MSAs had been merged in a full pass through the list, and stopping the process if no MSAs had been merged in 10 passes through the list, resulting in one or more MSA. Each combined MSA was converted into an HMMER profile-HMM (Mistry et al., 2013), and used as a template profile to align all sequences from the level 1 clusters (not including the enrichment sequences) using Clustal omega (Sievers & Higgins, 2018).

Sequence logos were computed by filtering out sequences from the final MSA with more than 90% pairwise identity using hhfilter (Steinegger et al., 2019), and the resulting MSAs were plotted using Logomaker (Tareen & Kinney, 2020) and matplotlib (Hunter, 2007) using the information content as the logo heights. All code used for supercluster comparison and sequence logo generation is provided in Supplementary data S6

4.4 Results

4.4.1 Protein dataset and initial sequence clustering

Nearly fourteen and a half million unique sequences spanning 69,724 distinct NCBI taxonomic ids were collected from public databases for clustering and remote homology comparison (Table 4.1). A first pass (level 1) clustering of these proteins with MMSeqs2 (Steinegger & Söding, 2017) yielded 3,111,468 clusters, which were designated as "level 1" clusters (Figure 4.1). Within this first pass, position-specific score matrix (PSSM) profile clustering did not merge any clusters from the standard MMSeqs2 "cascaded clustering" pipeline, but was observed to merge clusters in datasets with fewer sequences and when the coverage criteria were relaxed from 80% reciprocal as required here. The majority (1,918,741) of level 1 clusters consisted of one or two (the median) unique sequences (Figure 4.2F). A smaller number of large clusters were observed, with the largest 10% of clusters containing 8 or more unique sequences and a maximum cluster size of 10,244. From these level 1 clusters, enriched profile-hidden markov models (HMMs) were constructed for use with HH-Suite version 3.2.0 (Steinegger et al., 2019) (available online at https://doi.org/10.6084/m9.figshare.13289655.v1). To focus on finding potentially grouped families of effectors, HMMs for remote homology comparison were selected based on HHBlits matches to 6598 selected PHI-base (Urban et al., 2020) and effector sequences (Supplementary table S3). Of these sequences, 1078 sequences matched 286,512 level 1 cluster HMMs with a maximum e-value of 0.01 and minimum probability of 20%, which were selected for HMM-HMM comparisons. Of the subset of clusters selected for remote homology comparison, 2856 clusters contained unique sequences corresponding to 310 sequences from the effector and effector homologue dataset, and 3571 PHI-base entries.

4.4.2 Clustering of profile HMM-HMM matches to identify remote homology relationships between effector-like sequence clusters

To identify more distant relationships, all-vs-all profile HMM-HMM comparisons were performed on the 286,512 selected (effector-like) level 1 clusters (Figure 4.1). A total of 224,230 level 1 clusters were connected by 30,472,762 reciprocally significant alignments (e-value <= 10^{-5}, probability > 90%, reciprocal coverage > 70%). The remaining clusters had no matches at this significance threshold, and so could not be grouped into more remote clusters. Of these clusters without significant matches, 71 clusters contained unique sequences corresponding to 104 effectors and effector homologues. A strong correlation between alignment scores and sequence lengths was observed, which was effectively removed by normalisation (Supplementary figure S1). A graph was constructed of the level 1 effector-like clusters and their connecting alignments, using the mean of the pair of normalised scores as edge weights. The graph consisted of 6538 connected components (sub-graphs of directly or indirectly connected level 1 clusters), which were designated 'level 3' clusters. A single large component containing 171,346 nodes/level 1 clusters (representing 1.9 million unique proteins) was observed, with numerous small components typically with fewer than 1000 nodes also present (Table 4.1B; Figure 4.2A). Despite the presence of one large connected component, most level 1 clusters which corresponded to known effectors were found in smaller components, with only 19 out of 310 known effectors and effector homologues found in the largest component. These were typically highly conserved protein families including LysM-domain containing proteins, CRN, Tom1, Ave1 (expansins/PNPs), MoMSP (a cerato-platanin), though the NEPs and Ribotoxins each formed a separate component. In order to sub-divide larger connected components into more stringent remote-homology groupings, 'communities' or 'superclusters' of level 1 clusters were found within connected components using the greedy set cover (Hauser et al., 2016) and Markov clustering (Enright et al., 2002) algorithms. A markov clustering inflation value of 1.35 was selected for clustering, which gave the highest average modularity scores (Malliaros & Vazirgiannis, 2013) for a random selection of smaller connected components. Greedy and markov clustering (referred to as 'level 2A' and 'level 2B' clusters, respectively) generally yielded comparable groupings, but greedy clustering tended to produce more clusters with only one member (Figure 4.1D,E; Supplementary table S4, S5). Although each clustering method is different, conceptually the cluster levels from 1 to 3 represent progressively more distant homology relationships.

4.4.3 Level 2 and 3 clusters grouped multiple known effectors and predicted an expanded set of effector candidates across multiple pathogen species

We found 80 clusters at level 3, 103 and 104 clusters at level 2A and 2B, respectively, and 200 clusters at level 1, which contained known effectors and published effector homologues (Table 4.1). Of these, 20, 24 and 26, and 10, respectively, contained two or more known effectors

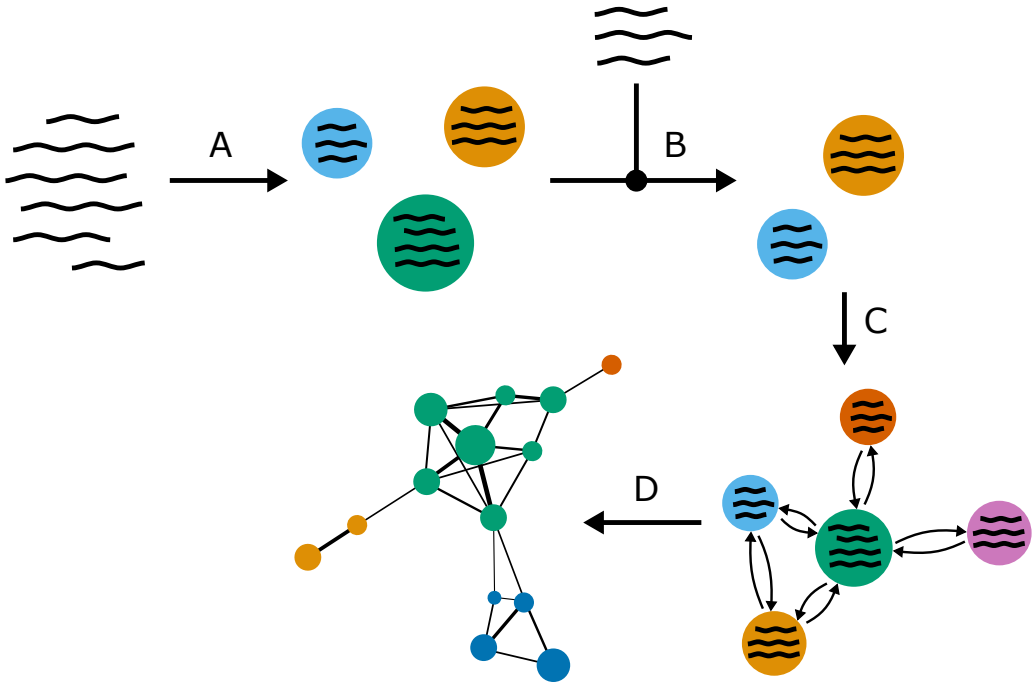


Figure 4.1: The clustering workflow employed in this study. A) Sequences are initially clustered using MMSeqs2 resulting in 3,111,468 level 1 clusters. B) A subset of 286,512 these clusters with any similarity to known effectors are found using HHBlits. C) All sequences from this subset are searched against themselves and reciprocally significant alignments are selected to form a graph. D) Clusters of the initial clusters are found within the resulting graph to form more distant sequence families. In the final graph, each point represents a level 1 cluster resulting from step A, the colours indicate level 2 clusters (markov or greedy clustering), and the whole graph forms a single connected component (level 3 cluster).

(Supplementary table S4; Figure 4.2D,E). To demonstrate how known effectors have been grouped into novel 'families' in this study, we present three examples in detail. The first example consists of a level 3 cluster (connected component) that contains *Leptosphaeria maculans* AvrLm6 (Fudal et al., 2007), *Magnaporthe oryzae* BAS4 and SPD5(Mosquera et al., 2009; Sharpee et al., 2017), *Fusarium oxysporum* f. sp. *lycopersici* SIX5 (Lievens et al., 2009), and *Cercospora beticola* NIP1 (Ebert, 2018)(Figure 4.3A,B). At cluster level 2, this group is further divided into sub-groups, with AvrLm6 and some published *Venturia inaequalis* AvrLm6 homologues (Shiller et al., 2015) forming a distinct sub-group, *M. oryzae* SPD5 and BAS4 forming another sub-group, and *C. beticola* NIP1 and *F. oxysporum* SIX5 both forming their own distinct sub-groups, with additional sub-groups that did not match a known effector. Sequence logos generated from a multiple sequence alignment of all level 1 clusters contained in these subgroups (Figure 4.3C, Supplementary figure S2, Supplementary data S1) indicated conservation of specific cysteine, threonine, and glycine residues, as well as distinct motifs that were specific to each sub-group. Level 2A clusters PC_02VR38 (containing AvrLm6), PC_01204B (containing Cb-Nip1), and PC_03MDGJ (containing SPD5 and BAS4) are found in numerous species from the

Table 4.1: Summary of the number of unique sequences in the input dataset (A) and the number of clusters obtained using various methods for remote homology clustering (B).

A) Initial clustering of input data for removal of sequence redundancies				
	Total	Uniparc	NCBI IPG	Custom
Unique Sequences	14,425,844	11,987,341	12,293,758	3,130,080
Unique sequences per taxid	23,351,787	19,081,482	14,297,803	3,302,707
B) Remote homology clustering				
Level	Total dataset		Containing known effectors	
1 (profile)	286,512		200	
2A (greedy)	45,363		103	
2B (markov)	27,851		104	
3 (connected components)	6538		80	

leotiomyceta clade including *Bipolaris spp*., *Colletotrichum spp*., *Leptosphaeria spp*., *Venturia spp*., and *Fusarium spp*. (Supplementary table S4, S5). The cluster containing *Fol* SIX5 (PC_05PCSX) was found in a broader range of taxa including the basidiomycetes *Jaapia argillacea* and *Plicaturopsis crispa*, but most observed sequences were from species in the Pezizomycotina, including other significant plant pathogens such as *Zymoseptoria tritici* and *Pyrenophora teres* f. sp. *teres*.

The second example consists of two separate connected components (level 3) clusters that correspond to the conserved ToxA effectors of *Pyrenophora tritici-repentis*, *Parastagonospora nodorum* and *Bipolaris spp.* (Friesen et al., 2018; Lu et al., 2015), and a set of loosely conserved 'ToxA-like' proteins which had been previously identified in other studies using PSI-BLAST searches, including ChEC13 (Lu et al., 2015) and AvrFOM2 (Schmidt et al., 2016) (Figure 4.4A,B). Our method did not link these two reportedly related groups within a single level 3 cluster. Alignments between the two connected components were observed, but failed the e-value significance threshold (data not shown). Multiple sequence alignment combining all sequences from both level 3 clusters showed only low level similarity between sequences of these two level 3 clusters (Figure 4.4C, Supplementary figure S3, Supplementary data S2). Between the ToxA and ToxA-like/AvrFOM2 clusters, there are several broadly conserved residues, most notably two cysteines, two aromatic [W|Y] residues, several aliphatic [L,I] residues, and an LxxRQ...C motif. The AvrFOM2 cluster(s) are more diverse than ToxA, with conserved residues separated by hypervariable regions, and also possess a phenylalanine rich region in the signal peptide. The AvrFOM2 cluster also lacks a recognisable 'RGD' (Meinhardt et al., 2002) or 'SGN' (Lu et al., 2015) motif (positions 138-140 in Figure 4.4C), which is absent in the level 2B (Markov) cluster PC_08EP4N and replaced by a poorly conserved 'TTP' consensus in level 2B cluster PC_07OLPP. The component containing ToxA sequences (PC_03B2DN) was observed to have members in several species that have been previously described: *Pyrenophora teres formae speciales*, *Pyrenophora tritici-repentis*, *Parastagonospora nodorum*, *Parastagonospora avenae*,

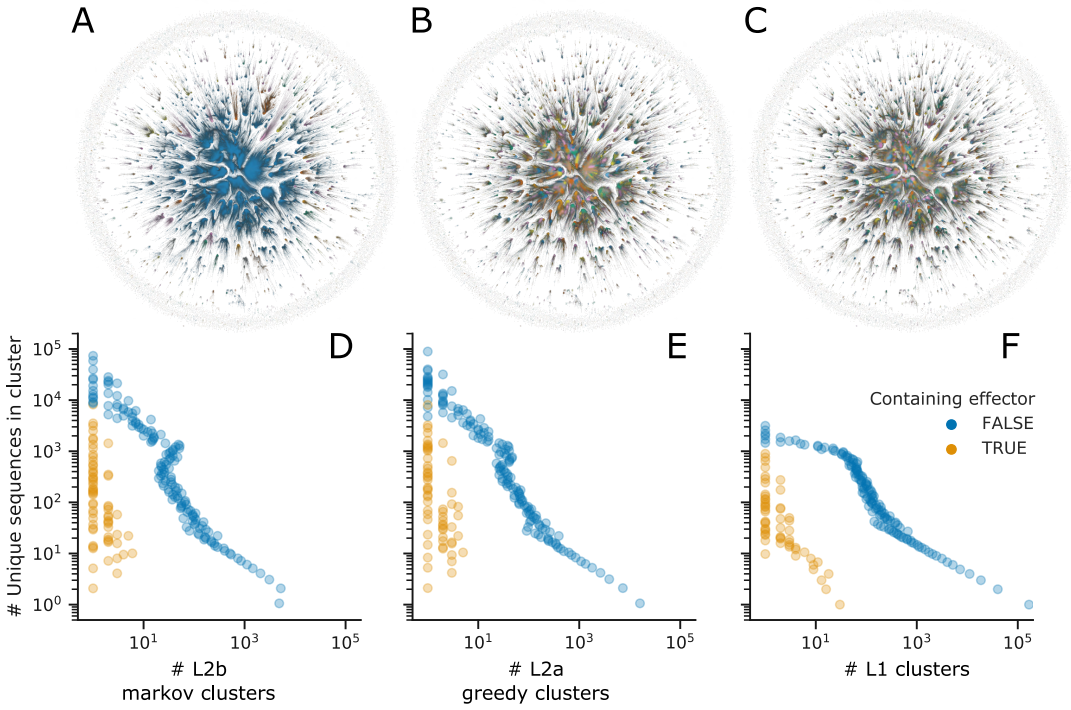


Figure 4.2: An overview of clustering of protein dataset. Top row, plot of graph coloured by connected components (A), and markov (B) and greedy (C) clusters. Bottom row: the number of unique sequences compared with the number of clusters with that size, within markov (D), greedy (E), and profile clusters (F). For the bottom row, Y-axis values are in binned into 100 evenly sized ranges taken from a 10-based exponential space ($10^{0..\max(\#seqs)}$).

Bipolaris maydis, and *Bipolaris sorokiniana* (Supplementary table S4, S5). The level 2B cluster containing AvrFOM2 and ChEC13 was observed in numerous species within the leotiomyceta clade, including *Epichloe spp.*, *Fusarium spp.*, *Pyrenophora spp.*, *Colletotrichum spp.*, and *Bipolaris spp.*. Other level 2 clusters within the component containing AvrFOM2 and ChEC13 also contain members from the leotiomyceta, but are specific to genus (*Epichloe*), or a strain (*Zymoseptoria ardabiliae* STIR04_1.1.1, *Balansia obtecta* B249).

In a similar manner to the ToxA and ToxA-like clusters, members of the MAX effector family and the homologues published by K. d. de Guillen et al. (2015) were found in 12 separate connected components in this study. None of these components contained more than one of the experimentally validated MAX members (ToxB, Avr1_CO39, AVR_Pik, AvrPiz_t, and AvrPib) (Supplementary figure S5, Supplementary data S4).

The third example is a level 3 cluster of RNase-like effectors that grouped level 1 clusters which were sufficiently divergent that profile alignments between all sequences in the multiple sequence alignment (MSA) was not possible. Consequently, further presentation of this example focuses on a sub-graph which includes level 2A and 2B clusters containing known RALPH (Pennington et al., 2019; Praz et al., 2017; Spanu, 2017) and ribotoxin (Kettles et al., 2018) effectors (Figure 4.5A). The ribotoxins (including Zt6) formed a large and densely connected

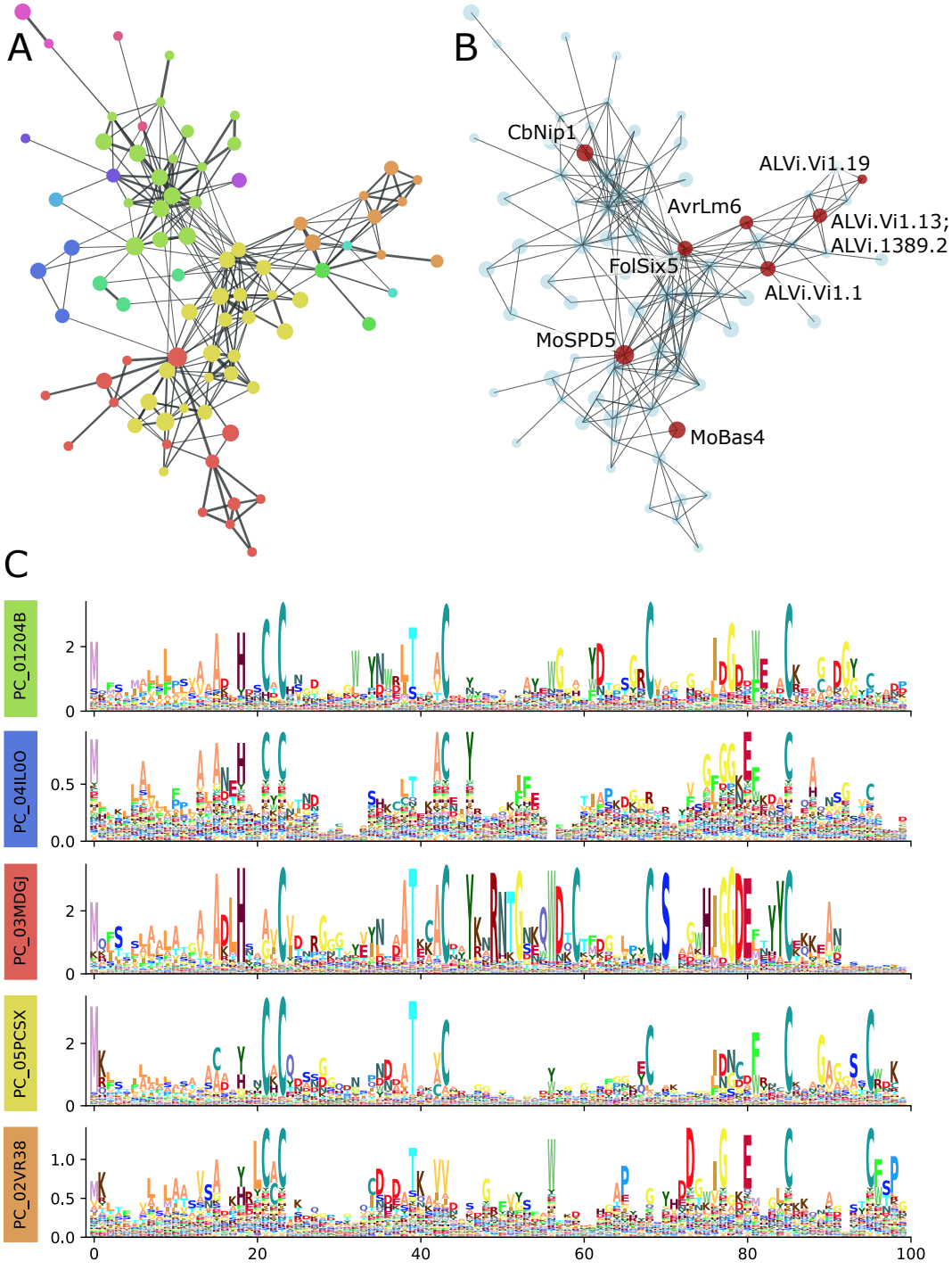


Figure 4.3: A family of SIX5-like effector sequences. A) The connected component containing the effectors AvrLm6, Bas4, SPD5, and SIX5, coloured by Markov cluster membership (level 2B). B) The same graph, but highlighting the level 1 clusters containing effector sequences and published effector homologues (ALVI*). C) Sequence logos resulting from multiple sequence alignment of all sequences in the connected component (level 3 clusters). Logos for markov clusters with more than 10 members are shown separately. Columns in the multiple sequence alignment with more than 50% gaps are excluded.

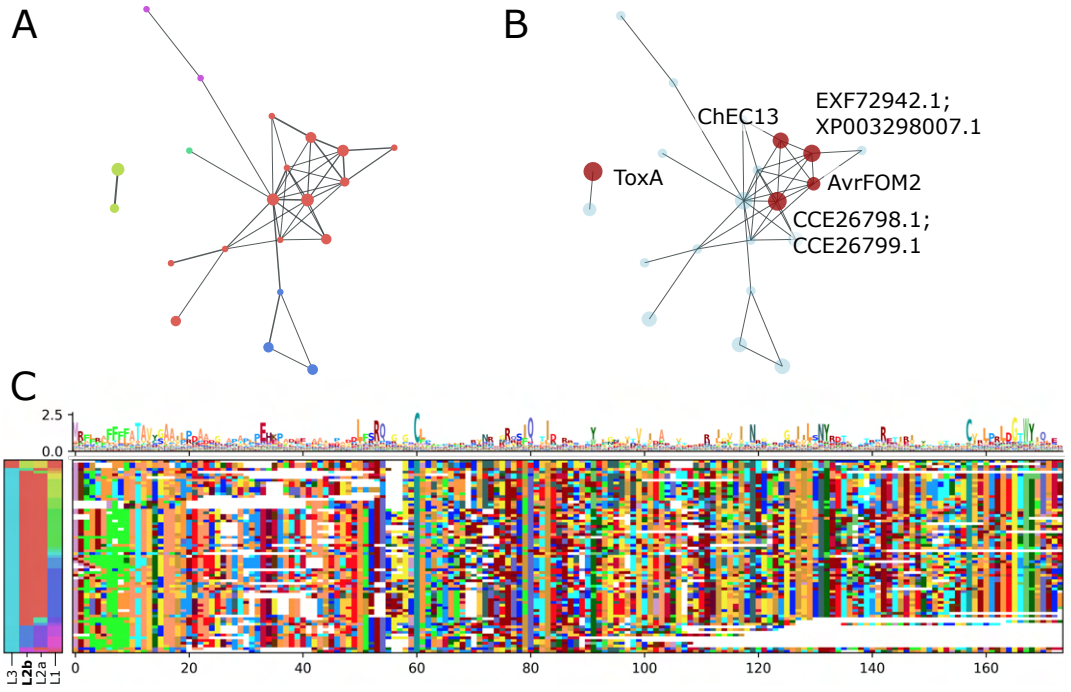


Figure 4.4: ToxA-like fungal effector groups. A) The connected components (level 3 clusters) containing ToxA-like and AvrFOM2-like sequences, coloured by Markov cluster membership (level 2B). B) The same graph shown in A, but highlighting level 1 clusters containing known effectors and published effector homologues. C) A multiple sequence alignment constructed from all sequences in the ToxA-like and AvrFOM2-like connected components. Columns in the multiple sequence alignment with more than 50% gaps are excluded. Colours on the y-axis indicate the level 1, 2, and 3 clusters that members belong to, with level 2B (markov) cluster colours matching those in A.

cluster, which was distinct from all RALPH effectors (Figure 4.5B,C). The RALPH effectors consist of three main groups: AvrPm3$^{a2/f2}$, AvrPm2/BEC1054/AVR$_{al3}$, and SvrPm3$^{a1/f1}$, and are sparsely connected. Multiple sequence alignment of all sequences in the selected clusters indicate 2 or 4 conserved cysteine positions in the RALPH and Ribotoxin logos, respectively (Figure 4.5D, Supplementary figure S4, Supplementary data S3). Additional conserved proline, aromatic [Y|F], and aliphatic [V|I] residues were observed. The clusters containing AvrPm2-like RALPH proteins (level 2B cluster PC_04SK9M) were more similar to the Ribotoxin/Zt6-like cluster (level 2B cluster PC_032CKH), than the clusters containing AvrPm3$^{a2/f2}$/SvrPm3$^{a1/f1}$ sequences (level 2B clusters PC_01D3OM and PC_0278ZT, respectively). The Y(x)xC motif commonly found after the signal peptide in *Blumeria* effectors (Pedersen et al., 2012; Praz et al., 2017) appeared to be enriched in AvrPm2-like and AvrPm3$^{a2/f2}$-like RALPH sequences, but may be replaced by an F(x)xC motif in SvrPm3$^{a1/f1}$-like sequences. The level 2B cluster containing the known Ribotoxin effector Zt6 (PC_032CKH) was broadly conserved in the Fungal kingdom (Supplementary table S4, S5). Sequences belonging to level 2B clusters corresponding to RALPH effectors (AvrPm2/PC_04SK9M, SvrPm3$^{a1/f1}$/PC_0278ZT, and

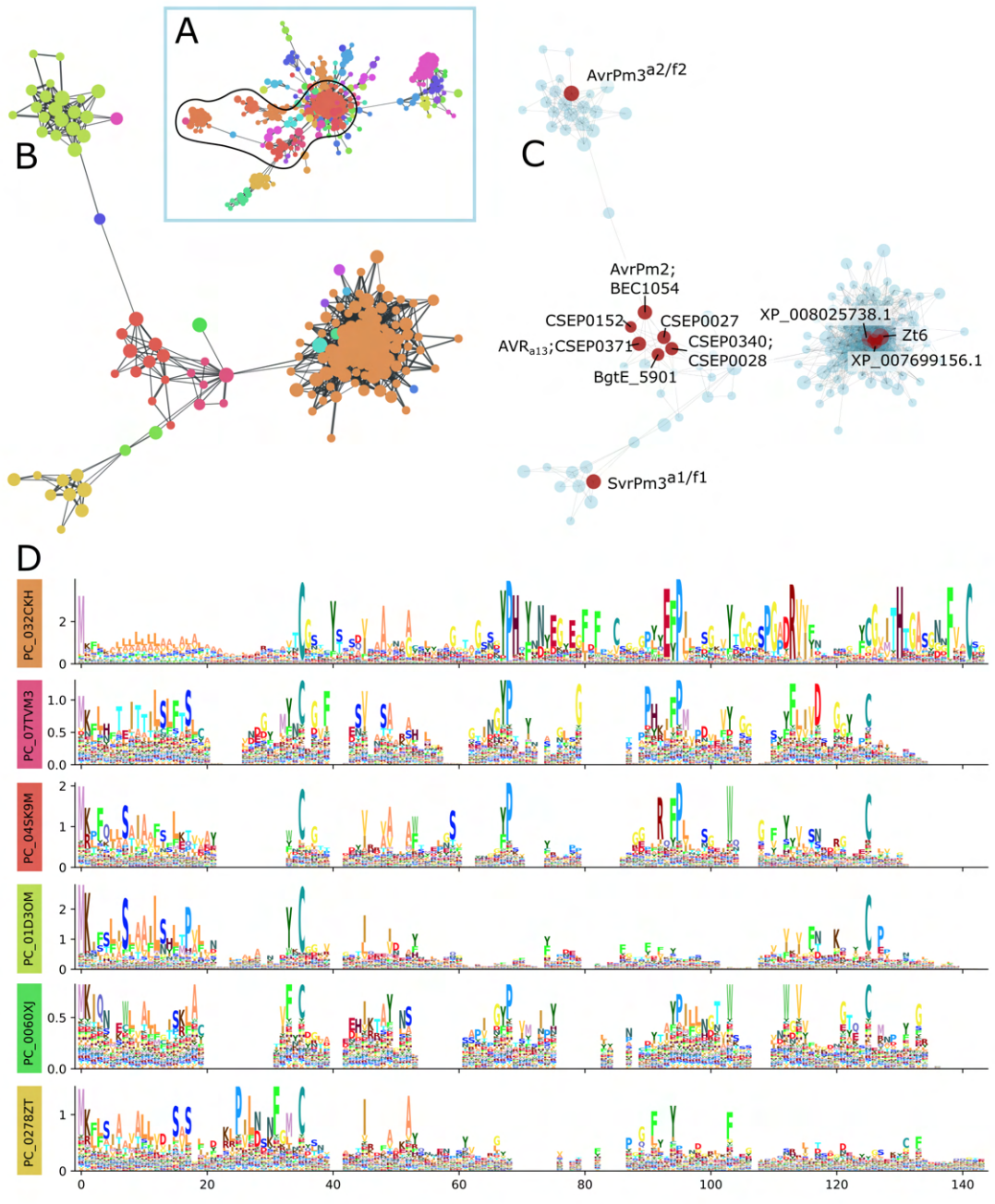


Figure 4.5: A connected component containing RNase-like effectors. A single connected component containing the Ribotoxins and RALPH effectors was observed (A). B) Shows a subset of the connected component containing all level 2 clusters containing effector sequences (C). D) Sequence logos for each level 2B (Markov) cluster from a multiple sequence alignment of all sequences in (B). Colours in the left boxes corresponding to colours in (B). Logos with fewer than 10 members are not shown. Columns in the MSA with greater than 50% gaps are excluded from the visualisation.

AvrPm3$^{a2/f2}$/PC_01D3OM) were only found in *Blumeria graminis formae speciales*. However, several other lineage specific level 2 clusters were observed within the same connected component, which were most often associated with the Pezizomycotina.

Ten other level 2 clusters that grouped two or more known effectors were identified (Supplementary table S4), grouping: *Leptosphaeria maculans* AvrLm2 and *Fusarium oxysporum f. sp. lycopersici* SIX1 (Ghanbarnia et al., 2015; Rep et al., 2004); *Zymoseptoria tritici* NIP1 and *Passalora fulva* Ecp2 (Laugé et al., 1997; M'Barek et al., 2015); *Blumeria graminis f. sp. hordei* BEC2 and *Golovinomyces orontii* GoEC2 (Schmidt et al., 2014); *Passalora fulva* Ecp6 and Zymoseptoria tritici Mg3LysM (Bolton et al., 2008; Marshall et al., 2011); NIS1 effectors (Irieda et al., 2019; Yoshino et al., 2012); *Zymoseptoria tritici* MgxLysM and Mg1LysM (Marshall et al., 2011); Magnaporthe oryzae AVR-Pita and AVR-Pita2 (Chuma et al., 2011; Dai et al., 2010); *Puccinia striiformis* Shr4 and Shr6 (Ramachandran et al., 2016); Pit2 effectors (Mueller et al., 2013; Schweizer et al., 2018); and the NEP virulence factors (Bailey et al., 2002; Garcia et al., 2007; Staats et al., 2007; J.-Y. Wang et al., 2004).

In other cases, clusters containing a single known effector were assigned functional annotations of high relevance to potential effector functions. For example, a large cluster (level 3: PC_07OBLJ, level 2: PC_058FSP, Supplementary table S4) corresponding to known effector BAS3 (biotrophy-associated secreted protein 3) of *Magnaporthe oryzae* (Mosquera et al., 2009), was functionally annotated as similar to scorpion knottin toxins [InterPro: IPR036574]. This group contained unconfirmed candidates from other *Pyricularia spp.*, as well as *Colletotrichum spp.*, *Macrophomina phaseolina*, *Neofusicoccum parvum* and *Monosporascus spp.*.

Some known effectors were not able to be grouped beyond cluster level 1 (Supplementary table S4). Of these, 21 were within clusters that contained a single unique sequence, including: AvrLm11 and AvrLmJ1 of *Leptosphaeria maculans* (M.-H. Balesdent et al., 2013; Van de Wouw et al., 2014), Avr5 of *Passalora fulva* (Mesarich et al., 2014), AVR_{al0} of *Blumeria graminis f. sp. tritici* (Ridout et al., 2006), PIIN_08944 and FGB1 of *Piriformospora indica* (Akum et al., 2015; Wawra et al., 2016), CDIP3 and Slp1 of *Magnaporthe oryzae* (S. Chen et al., 2012; Mentlak et al., 2012), lsc1 of *Verticillium dahliae* (T. Liu et al., 2014), Zt80707 of *Zymoseptoria tritici* (Poppe et al., 2015), and the putative effector CSEP-07 of *Phakopsora pachyrhizi* (Kunjeti et al., 2016). Another 50 level 1 clusters containing a single effector possessed two or more unique sequences, for which most were restricted to isolates of the same species or genus. These included Tox1 of *Parastagonospora* (Z. Liu et al., 2012); SIX2, SIX4, and SIX8 of *Fusarium oxysporum* (Houterman et al., 2008; Lievens et al., 2009); DN3 and EP1 of *Colletotrichum* (Stephenson et al., 2000; Vargas et al., 2015); AvrP4 and AvrL567 of Melampsora (Catanzariti et al., 2006; Dodds et al., 2004); SSVP1 of Sclerotiniaceae (Lyu et al., 2016); Shr5 and Shr7 of *Puccinia* (Ramachandran et al., 2016); NIP3 of *Rhynchosporium* (Kirsten et al., 2012); and SCP7 of *Verticillium* (Zhang et al., 2017). Notable exceptions of level 1 clusters which spanned genera were Ecp1 of *Passalora fulva* (Laugé et al., 1997) which had a homolog in *Pseudocercospora eumusae*, AVR4 of *P. fulva* (Joosten et al., 1994) which had a homolog in *Dothistroma septosporum*, AvrLm3 of *L. maculans* (Plissonneau et al., 2016) which had homologs in *P. fulva* and *Fusarium oxysporum* f. sp. *narcissi*,

and SSVP1 of *Sclerotinia sclerotiorum* (Lyu et al., 2016) which had homologs in *Botrytis spp.* and *Monilinia laxa*.

4.5 Discussion

With a growing number of experimentally-confirmed fungal "effector" proteins in the public domain (Supplementary table S3) (Urban et al., 2020), there are emerging opportunities to mine this data and develop improved methods for effector and virulence factor discovery. However, basic homology-based methods cannot necessarily be applied, as many known effector proteins are either sufficiently divergent or of independent origin to prevent their grouping into larger 'effector families'. Comparisons between effector proteins and candidates at the structural level have indicated recognisable structural homology between many emerging groupings, including the ToxA-like (Lu et al., 2015; Schmidt et al., 2016), MAX (K. d. de Guillen et al., 2015), RALPH (Pennington et al., 2019; Praz et al., 2017; Spanu, 2017), and Hce2 (Stergiopoulos et al., 2012) families. Structural homology may become the basis for reliable effector prediction in future studies; however, the application of protein structure prediction to large sets of effector candidates is not currently computationally feasible. This study applied a highly sensitive sequence clustering approach — termed 'RemEff' — to a large protein dataset to form novel protein clusters, leveraging known effectors to identify effector 'family' clusters and predict high-confidence effector candidates within them by association.

While the RemEff method has taken a 'top-down' approach that has identified a large number of 'effector families' (Supplementary table S4), we focus here on selected examples. In our first detailed example (Figure 4.3), we presented a previously undescribed expanded family of effectors containing the effectors *Leptosphaeria maculans* AvrLm6 (Fudal et al., 2007), *Magnaporthe oryzae* BAS4 and SPD5 (Mosquera et al., 2009; Sharpee et al., 2017), *Fusarium oxysporum* f. sp. *lycopersici* SIX5 (Lievens et al., 2009), and *Cercospora beticola* NIP1 (Ebert, 2018). Each study describing the effectors has noted the presence of homologues of these effectors in multiple species. Numerous homologues of AvrLm6 have been previously observed in *Venturia*, *Colletotrichum*, and *Fusarium* species (Grandaubert et al., 2014; Shiller et al., 2015). However, not all *Venturia* AvrLm6 homologues published by Shiller et al. (2015) were identified as members of this superfamily. In that study the only restriction on matches was a PSI-BLAST e-value of 10^{-2}, so it is likely that the focus here on finding full length homologues might have excluded these potential matches. Each of the five level 2 clusters within the level 3 cluster had a different cysteine spacing pattern, with four or six conserved cysteine positions each. Some cysteine residues were conserved across multiple groups, and two positions were conserved in all subgroups suggesting their functional relevance.

None of these AvrLm6-like effectors have yet been structurally determined, however SIX-5 and BAS4 may operate in similar environments and may have similar functions. SIX5 appears to interact with plasmodesmata and mediates the intercellular translocation of another effector Avr2 (where it can then exert virulence promoting and avirulence function) (Cao et al., 2018).

Intriguingly, another pair of *Leptosphaeria maculans* effectors not present in this study, one of which is a SIX5 homologue, appear to show a similar interaction in that pathosystem (Petit-Houdenot et al., 2019). Magnaporthe biotrophy-associated secreted protein 4 (BAS4) elicits a host defence response late in the biotrophic phase, which promotes cell death during the necrotrophic phase (C. Wang et al., 2019). Fluorescently labelled BAS4 was found to uniformly outline the invasive hyphae (IH) of *Magnaporthe oryzae* during compatible infection (Mosquera et al., 2009), and does not enter the host cell under normal circumstances. Cytoplasmic effectors PWL2 and BAS1, but not BAS4, move from cell to cell preceding the IH, possibly through plasmodesmata (Khang et al., 2010), but it's unclear whether BAS4 has any role in this. Suppressor of cell death 5 (SPD5) was a known homologue of BAS4, and has known homologues in numerous ascomycete fungi. It suppresses BAX- and NEP1-induced cell death (Sharpee et al., 2017). CbNIP1 induces light-independent necrosis (Ebert, 2018), but its specific activity and subcellular (or intercellular) location is unknown. The relatively broad taxonomic distribution of this group is interesting, and the common association with membranes in the host apoplast or biotrophic interfacial complex (BIC) of SIX-5 and BAS4 is intriguing.

In our second detailed example (Figure 4.4), we compared two other clusters containing the effectors ToxA and AvrFOM2, which were previously reported as similar (Schmidt et al., 2016). The cluster containing AvrFOM2 is much larger and more sequence diverse compared to the one containing ToxA. Within the ToxA level three cluster are only the canonical ToxA-like effectors of *Parastagonospora spp.*, *Pyrenophora spp.*, and *Bipolaris sorokiniana* and maydis, many of which are identical and are thought to have arisen by a complex horizontal transfer event (McDonald et al., 2018). The level three cluster containing AvrFOM2 and ChEC13 overlaps considerably with the candidate homologues identified in Lu et al. (2015). Although a *Fusarium oxysporum f. sp. melonis* homologue was described in that paper, it does not appear to have been AvrFOM2. The multiple sequence alignment does show the conservation of some of the motifs previously described (Lu et al., 2015), including the LXXR pro-peptide cleavage site, and the three motifs found in beta sheets 4 (LXVXIXN, here replaced by IXVXIXN in PC_07OLPP containing AvrFOM2), 5 (LILTXY, replaced by I[VI]LSNY in PC_07OLPP) and 8 (WXXQ). However, neither the asparagine rich motif (WXXN(S)NXIXVXI) nor the RGD/SGN motif were observed. The level 2 cluster containing AvrFOM2 and ChEC13 (PC_07OLPP) exhibits a number of phenylalanine residues in the signal peptide (SP) at the junction of the N-region and the hydrophobic core. Hydrophobic amino acids near the N-region tend to decrease secretion efficiency (Owji et al., 2018) and although phenylalanine residues are found in the hydrophobic core regions of human signal peptides, it is not generally known in yeasts (Duffy et al., 2010). However, the amino acid composition of efficient SPs can vary between species, and the hydrophobic and N-terminal regions of the SP may be involved in directing proteins through different secretion pathways (Owji et al., 2018).

In our last detailed example (Figure 4.5) including the ribotoxins and RALPH effectors, the clusters containing ribotoxins (Zt6) and RNase-like protein associated with haustoria (RALPH) effectors formed distinct clusters, but a clear similarity existed at specific regions between

clusters PC_032CHK containing Zt6 and PC_04SK9M containing AvrPm2 and BEC1054. In the ribotoxin sequence α-sarcin the active sites are Histidine in the YPH motif, Glutamine in EFP motif, and a Histidine between the last two cysteine residues, all of which are missing in RALPHs though they possess the conserved surrounding sequence of the former two (Pérez-Cañadillas et al., 2000; Viegas et al., 2009; Yang & Moffat, 1996). Additionally, all RALPH sequences lacked the extended N-terminal loop that has previously been thought to be necessary for ribotoxin activity, though it was also poorly conserved in the cluster containing Zt6 (Olombrada et al., 2017). Overall, the profiles of RALPH effectors, with only two conserved cysteine positions, is more like RNase T1 than the Ribotoxins, are missing many of the previously described active sites, and have shorter loop sequences than the canonical ribotoxins. This is consistent with previous structural prediction analysis (Praz et al., 2017), and makes sense given that *Blumeria graminis*, to which this group appears to be restricted, are obligate biotrophs which would not benefit from effectors with cytotoxic activity. This also supports speculation that BEC1014 acts as a pseudoenzyme, binding host ribosomes but not cleaving the SRL (Pennington et al., 2019).

Both AvrPm3$^{a2/f2}$ and the suppressor SvrPm3$^{a1/f1}$ form distinct level 2 clusters branching from the main group of RALPH effectors (PC_01D3OM and PC_0278ZT, respectively). SvrPm3$^{a1/f1}$ was originally described as being a member of the RALPH group, but AvrPm3$^{a2/f2}$ was not (Spanu, 2017). It has previously been demonstrated that high expression of SvrPm3$^{a1/f1}$ suppresses the recognition of AvrPm3$^{a2/f2}$ by Pm3 receptors (Bourras et al., 2015), and that positive selection in the *avrpm3$^{a2/f2}$* gene does not appear to be related to evasion of recognition by Pm3 (McNally et al., 2018). Although the clusters are quite distinct, their association may suggest a possible mode of SvrPm3$^{a1/f1}$ suppression, where it may act as a 'bodyguard decoy' to AvrPm3$^{a2/f2}$ (Paulus & van der Hoorn, 2018). However, we note that the level 2 clusters containing AvrPm3$^{a2/f2}$ and SvrPm3$^{a1/f1}$ may be poorly aligned here, and AvrPm3$^{a2/f2}$ shares little conserved sequence similarity with the other RALPH effectors beyond the signal peptide and the cysteine positions.

Several other effectors formed groups of more than one effector, including two that have not previously been reported: AvrLm2 and SIX1 (Ghanbarnia et al., 2015; Rep et al., 2004), and *Puccinia striiformis* Shr4 and Shr6 (Ramachandran et al., 2016). However in addition to the groupings that the RemEff method presented here has formed between known and candidate effector proteins, the absence of predicted groupings may also offer biological insights. The presence of effectors in "orphan" clusters might be an indicator of their evolutionary histories involving either high sequence divergence or independent origin. The AvrFOM2 level 3 cluster described above, which contained sequences that were previously reported to be ToxA-like (Lu et al., 2015; Schmidt et al., 2016), failed to group with the ToxA (level 3) cluster despite weak overall sequence similarity (Figure 4). Similarly we failed to group the MAX proteins into a single component (Supplementary figure S5). Previously this ToxA-like group, and other prominent examples such as the MAX effector family (K. d. de Guillen et al., 2015), had been reported to be grouped by progressive hidden markov model (HMM) or position-specific score matrix (PSSM) searches and manual interrogation of the alignments. An important

distinction between those methods and this study is that the former searches for any sequence homologues of a single query sequence, whereas our method seeks to identify reciprocally-matching families at a higher level of stringency. Additionally, detecting low sequence identity matches to a single query sequence often involves manually evaluating matches, whereas clustering and HMM-HMM comparisons makes this an automated (and unbiased) procedure. Past studies using progressive searches have focussed on a single or reduced set of effectors (e.g. AvrLm6, Zt6, MAX), whereas this study mines a larger dataset spanning multiple fungal species, greatly increasing its predictive potential. As long as a confirmed effector is present within a level 1 or 2 cluster, the predictive outcome is much the same regardless of higher level groupings, and the two processes of effector prediction and effector family grouping should be considered to be somewhat independent. In fact, ensembles of HMMs have previously been demonstrated to represent families better than single large HMMs, and can more accurately classify sequences within diverse families (Nguyen et al., 2016). We have made the underlying profile data available for further analysis (https://figshare.com/projects/Effector_protein_remote_homology/87965), which can serve as a useful resource for future plant pathology studies.

4.6 Conclusion

The RemEff method predicts remote homology relationships between known effectors and candidate effector proteins, allowing for the prediction of distantly related effector 'families' in plant-pathogenic fungi. We have presented case studies of novel effector family groupings that both demonstrate the utility of this method to enhance effector discovery research, and highlight important similarities and differences between effector family sub-groups. This is illustrated well by sequence logos generated from the AvrLm6-like (Figure 4.3) and ribonuclease (Figure 4.5) level 2 clusters, which show a combination of overall conservation and motif diversity. We observe cysteine spacing to be a major conserved feature, sometimes in the absence of other defining sequence features. Given the potential overlap in modes of action of some fungal effectors and other non-fungal cytotoxic peptides, we speculate that it may be useful to further explore conservation of cysteine-spacing as a heuristic classification system for some groups of fungal effectors, similar to those that have been well established for arthropod venoms (Fry et al., 2009; Saucedo et al., 2012) and for snail conotoxins (Robinson & Norton, 2014). Alternatively, as protein structure prediction methods become more feasible to apply at large scale, it may become possible to predict effector candidates solely on the basis of structure. However, given the number and diversity of 'effector' proteins and their functions, we anticipate that neither method would be broadly applicable and maintaining an ensemble of profile-HMMs will be preferable for the foreseeable future.

4.7 Acknowledgements

This study was supported by the Centre for Crop and Disease Management, a joint initiative of Curtin University and the Grains Research and Development Corporation (Research Grant CUR00023). This research was undertaken with the assistance of resources and services from the Pawsey Supercomputing Centre and the National Computational Infrastructure (NCI), which is supported by the Australian Government. This research is supported by an Australian Government Research Training Program (RTP) Scholarship.

4.8 Data availability

Sequence data and metadata used for clustering is available online at https://doi.org/10.6084/m9.figshare.12833678.v1. Multiple sequence alignments (MSAs) and HMMs for all level 1 clusters is available online at https://doi.org/10.6084/m9.figshare.13289655.v1. MSAs and HMMER3 formatted HMMs for level 2 and 3 clusters containing known effectors is available online at https://doi.org/10.6084/m9.figshare.12838253.v1.

4.9 Supplementary material

All supplementary material is available online at https://doi.org/10.6084/m9.figshare.13325246.

Supplementary table S1. Additional genomes and proteomes used for clustering in addition to non-redundant sets from NCBI IPG and Unimedskipc. Genomes collected from the JGI mycocosm database have a corresponding JGI id.

Supplementary table S2. Taxonomic summary of input protein dataset. The sheet "summary" contains the numbers of distinct phyla, classes, orders, families, genera, and species. The sheets "superkingdom", "phylum", "class", and "order" indicate the number of clusters that contain a member of each taxon at the sheet names rank.

Supplementary table S3. The query sequences and summarised search results used to subset the clusters prior to pairwise comparison. Sheet "phibase_effector_selection" contains summarised PHI-base entries, which were curated to identify extracellular proteins in the "selected" column. Sheet "custom_effector_sequences" contains details of validated and hypothetical effector sequences collected from literature, which fills some gaps in the PHIBase dataset. Both the PHIBase sequences and custom effector sequences were used to search for matches against cluster HMMs. Sheet "match_totals" contains the numbers of search matches overall of the selected PHIBase sequences and the custom effector set searched against the database of cluster HMMs. Sheets "matches_unfiltered" and "matches_filtered" contain summaries of evalues, HMM alignment probabilities, alignment lengths, and sequence identity for the

raw matches (unfiltered) and for matches filtered to have a maximum e-value of 1e-5 and a minimum alignment length of 15 residues. The "include" column indicates whether the query sequence should be considered to be an effector (either "selected" in the phibase dataset or in the custom dataset).

Supplementary table S4. Full list of level 1 clusters, and membership in levels 2a/b and 3.

Supplementary table S5. Detailed lowest common ancestor summary for clusters level 1-3.

Supplementary figure S1. HHBlits alignment score length normalisation. A) Alignment scores show a strong correlation with the product of HMM lengths in a log-log space for the top 10 matches of each query. B) The scores after normalisation show little dependence on the product of HMM lengths. C) The two normalised scores for each pair of a significant match (i.e. A vs B and B vs A) are highly correlated, indicating that the arithmetic mean is a reasonable combination of the scores.

Supplementary figure S2. Unfiltered logos for Six5-like group.

Supplementary figure S3. Unfiltered logos for ToxA-like group.

Supplementary figure S4. Unfiltered logos for RNase-like effector group.

Supplementary figure S5. MAX graph and alignment (unfiltered).

Supplementary data S1. Multiple sequence alignment of Six5-like group.

Supplementary data S2. MSA of ToxA like group.

Supplementary data S3. MSA of RNase effector group.

Supplementary data S4. MSA of MAX effector group.

Supplementary data S5. Logos for all effector clusters.

Supplementary data S6. Code used for clustering profile HMM-HMM alignments.

4.10 References

Akum, F. N., Steinbrenner, J., Biedenkopf, D., Imani, J., & Kogel, K.-H. (2015). The *Piriformospora indica* effector PIIN_08944 promotes the mutualistic Sebacinalean symbiosis. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.00906

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*, *25*(17), 3389–3402. https://doi.org/10.1093/nar/25.17.3389

Amaro, T. M. M. M., Thilliez, G. J. A., Motion, G. B., & Huitema, E. (2017). A Perspective on CRN Proteins in the Genomics Age: Evolution, Classification, Delivery and Function Revisited. *Frontiers in Plant Science*, *8*. https://doi.org/10.3389/fpls.2017.00099

Anderson, R. G., Deb, D., Fedkenheuer, K., & McDowell, J. M. (2015). Recent Progress in RXLR Effector Research. *Molecular Plant-Microbe Interactions*, *28*(10), 1063–1072. https://doi.org/10.1094/MPMI-01-15-0022-CR

Andrie, R. M., Schoch, C. L., Hedges, R., Spatafora, J. W., & Ciuffetti, L. M. (2008). Homologs of ToxB, a host-selective toxin gene from *Pyrenophora tritici-repentis*, are present in the genome of sister-species *Pyrenophora bromi* and other members of the Ascomycota. *Fungal Genetics and Biology*, *45*(3), 363–377. https://doi.org/10.1016/j.fgb.2007.10.014

Armenteros, J. J. A., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., Heijne, G. v., & Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology*, *37*(4), 420–423. https://doi.org/10.1038/s41587-019-0036-z

Asgari, E., McHardy, A. C., & Mofrad, M. R. K. (2019). Probabilistic variable-length segmentation of protein sequences for discriminative motif discovery (DiMotif) and sequence embedding (ProtVecX). *Scientific Reports*, *9*(1), 3577. https://doi.org/10.1038/s41598-019-38746-w

Babnigg, G., & Giometti, C. S. (2006). A database of unique protein sequence identifiers for proteome studies. *PROTEOMICS*, *6*(16), 4514–4522. https://doi.org/10.1002/pmic.200600032

Bailey, B. A., Apel-Birkhold, P. C., & Luster, D. G. (2002). Expression of NEP1 by *Fusarium oxysporum* f. sp. *erythroxyli* After Gene Replacement and Overexpression Using Polyethylene Glycol-Mediated Transformation. *Phytopathology*, *92*(8), 833–841. https://doi.org/10.1094/PHYTO.2002.92.8.833

Balesdent, M. H., Attard, A., Kühn, M. L., & Rouxel, T. (2002). New Avirulence Genes in the Phytopathogenic Fungus *Leptosphaeria maculans*. *Phytopathology*, *92*(10), 1122–1133. https://doi.org/10.1094/PHYTO.2002.92.10.1122

Balesdent, M.-H., Fudal, I., Ollivier, B., Bally, P., Grandaubert, J., Eber, F., Chèvre, A.-M., Leflon, M., & Rouxel, T. (2013). The dispensable chromosome of *Leptosphaeria maculans* shelters an effector gene conferring avirulence towards *Brassica rapa*. *New Phytologist*, *198*(3), 887–898. https://doi.org/10.1111/nph.12178

Ballance, G. M., Lamari, L., & Bernier, C. C. (1989). Purification and characterization of a host-selective necrosis toxin from *Pyrenophora tritici-repentis*. *Physiological and Molecular Plant Pathology*, *35*(3), 203–213. https://doi.org/10.1016/0885-5765(89)90051-9

Bendtsen, J. D., Nielsen, H., von Heijne, G., & Brunak, S. (2004). Improved Prediction of Signal Peptides: SignalP 3.0. *Journal of Molecular Biology*, *340*(4), 783–795. https://doi.org/10.1016/j.jmb.2004.05.028

Bertazzoni, S., Williams, A. H., Jones, D. A., Syme, R. A., Tan, K.-C., & Hane, J. K. (2018). Accessories Make the Outfit: Accessory Chromosomes and Other Dispensable DNA Regions in Plant-Pathogenic Fungi. *Molecular Plant-Microbe Interactions*, *31*(8), 779–788. https://doi.org/10.1094/MPMI-06-17-0135-FI

Bolton, M. D., van Esse, H. P., Vossen, J. H., de Jonge, R., Stergiopoulos, I., Stulemeijer, I. J. E., Berg, G. C. M. V. D., Borrás-Hidalgo, O., Dekker, H. L., de Koster, C. G., de Wit, P. J. G. M., Joosten, M. H. A. J., & Thomma, B. P. H. J. (2008). The novel *Cladosporium fulvum* lysin motif effector Ecp6 is a virulence factor with orthologues in other fungal species. *Molecular Microbiology*, *69*(1), 119–136. https://doi.org/10.1111/j.1365-2958.2008.06270.x

Bourras, S., McNally, K. E., Ben-David, R., Parlange, F., Roffler, S., Praz, C. R., Oberhaensli, S., Menardo, F., Stirnweis, D., Frenkel, Z., Schaefer, L. K., Flückiger, S., Treier, G., Herren, G., Korol, A. B., Wicker, T., & Keller, B. (2015). Multiple Avirulence Loci and Allele-Specific Effector Recognition Control the Pm3 Race-Specific Resistance of Wheat to Powdery Mildew. *The Plant Cell*, *27*(10), 2991–3012. https://doi.org/10.1105/tpc.15.00171

Cao, L., Blekemolen, M. C., Tintor, N., Cornelissen, B. J. C., & Takken, F. L. W. (2018). The *Fusarium oxysporum* Avr2-Six5 Effector Pair Alters Plasmodesmatal Exclusion Selectivity to Facilitate

Cell-to-Cell Movement of Avr2. *Molecular Plant*, *11*(5), 691–705. https://doi.org/10.1016/j.molp.2018.02.011

Catanzariti, A.-M., Dodds, P. N., Lawrence, G. J., Ayliffe, M. A., & Ellis, J. G. (2006). Haustorially Expressed Secreted Proteins from Flax Rust Are Highly Enriched for Avirulence Elicitors. *The Plant Cell*, *18*(1), 243–256. https://doi.org/10.1105/tpc.105.035980

Chen, H., Kovalchuk, A., Keriö, S., & Asiegbu, F. O. (2013). Distribution and bioinformatic analysis of the cerato-platanin protein family in Dikarya. *Mycologia*, *105*(6), 1479–1488. https://doi.org/10.3852/13-115

Chen, J., Liu, B., & Huang, D. (2016). Protein Remote Homology Detection Based on an Ensemble Learning Approach. *BioMed Research International*, *2016*, 1–11. https://doi.org/10.1155/2016/5813645

Chen, L., Li, X., Li, C., Swoboda, G. A., Young, C. A., Sugawara, K., Leuchtmann, A., & Schardl, C. L. (2015). Two distinct Epichloë species symbiotic with *Achnatherum inebrians*, drunken horse grass. *Mycologia*, *107*(4), 863–873. https://doi.org/10.3852/15-019

Chen, S., Songkumarn, P., Venu, R. C., Gowda, M., Bellizzi, M., Hu, J., Liu, W., Ebbole, D., Meyers, B., Mitchell, T., & Wang, G.-L. (2012). Identification and Characterization of In planta–Expressed Secreted Effector Proteins from *Magnaporthe oryzae* That Induce Cell Death in Rice. *Molecular Plant-Microbe Interactions*, *26*(2), 191–202. https://doi.org/10.1094/MPMI-05-12-0117-R

Chiapello, H., Mallet, L., Guérin, C., Aguileta, G., Amselem, J., Kroj, T., Ortega-Abboud, E., Lebrun, M.-H., Henrissat, B., Gendrault, A., Rodolphe, F., Tharreau, D., & Fournier, E. (2015). Deciphering Genome Content and Evolutionary Relationships of Isolates from the Fungus *Magnaporthe oryzae* Attacking Different Host Plants. *Genome Biology and Evolution*, *7*(10), 2896–2912. https://doi.org/10.1093/gbe/evv187

Chuma, I., Isobe, C., Hotta, Y., Ibaragi, K., Futamata, N., Kusaba, M., Yoshida, K., Terauchi, R., Fujita, Y., Nakayashiki, H., Valent, B., & Tosa, Y. (2011). Multiple Translocation of the *AVR-Pita* Effector Gene among Chromosomes of the Rice Blast Fungus *Magnaporthe oryzae* and Related Species. *PLOS Pathogens*, *7*(7), e1002147. https://doi.org/10.1371/journal.ppat.1002147

Citores, L., Ragucci, S., Ferreras, J. M., Di Maro, A., & Iglesias, R. (2019). Ageritin, a Ribotoxin from Poplar Mushroom (*Agrocybe aegerita*) with Defensive and Antiproliferative Activities. *ACS Chemical Biology*, *14*(6), 1319–1327. https://doi.org/10.1021/acschembio.9b00291

Ciuffetti, L. M., Tuori, R. P., & Gaventa, J. M. (1997). A single gene encodes a selective toxin causal to the development of tan spot of wheat. *The Plant Cell*, *9*(2), 135–144. https://doi.org/10.1105/tpc.9.2.135

Cock, P. J. A., Antao, T., Chang, J. T., Chapman, B. A., Cox, C. J., Dalke, A., Friedberg, I., Hamelryck, T., Kauff, F., Wilczynski, B., & de Hoon, M. J. L. (2009). Biopython: Freely available Python tools for computational molecular biology and bioinformatics. *Bioinformatics*, *25*(11), 1422–1423. https://doi.org/10.1093/bioinformatics/btp163

Dai, Y., Jia, Y., Correll, J., Wang, X., & Wang, Y. (2010). Diversification and evolution of the avirulence gene *AVR-Pita1* in field isolates of *Magnaporthe oryzae*. *Fungal Genetics and Biology*, *47*(12), 973–980. https://doi.org/10.1016/j.fgb.2010.08.003

Dang, H. X., Pryor, B., Peever, T., & Lawrence, C. B. (2015). The Alternaria genomes database: A comprehensive resource for a fungal genus comprised of saprophytes, plant pathogens, and allergenic species. *BMC Genomics*, *16*(1), 239. https://doi.org/10.1186/s12864-015-1430-7

de Guillen, K., Lorrain, C., Tsan, P., Barthe, P., Petre, B., Saveleva, N., Rouhier, N., Duplessis, S., Padilla, A., & Hecker, A. (2019). Structural genomics applied to the rust fungus *Melampsora larici-populina*

reveals two candidate effector proteins adopting cystine knot and NTF2-like protein folds. *Scientific Reports*, *9*(1), 1–12. https://doi.org/10.1038/s41598-019-53816-9

de Guillen, K. d., Ortiz-Vallejo, D., Gracy, J., Fournier, E., Kroj, T., & Padilla, A. (2015). Structure Analysis Uncovers a Highly Diverse but Structurally Conserved Effector Family in Phytopathogenic Fungi. *PLOS Pathogens*, *11*(10), e1005228. https://doi.org/10.1371/journal.ppat.1005228

Deng, C. H., Plummer, K. M., Jones, D. A. B., Mesarich, C. H., Shiller, J., Taranto, A. P., Robinson, A. J., Kastner, P., Hall, N. E., Templeton, M. D., & Bowen, J. K. (2017). Comparative analysis of the predicted secretomes of Rosaceae scab pathogens *Venturia inaequalis* and *V. pirina* reveals expanded effector families and putative determinants of host range. *BMC Genomics*, *18*(1), 339. https://doi.org/10.1186/s12864-017-3699-1

de Wit, P. J. G. M., Mehrabi, R., Van den Burg, H. A., & Stergiopoulos, I. (2009). Fungal effector proteins: Past, present and future. *Molecular Plant Pathology*, *10*(6), 735–747. https://doi.org/10.1111/j.1364-3703.2009.00591.x

Di, X., Cao, L., Hughes, R. K., Tintor, N., Banfield, M. J., & Takken, F. L. W. (2017). Structure–function analysis of the *Fusarium oxysporum* Avr2 effector allows uncoupling of its immune-suppressing activity from recognition. *New Phytologist*, *216*(3), 897–914. https://doi.org/10.1111/nph.14733

Dodds, P. N., Lawrence, G. J., Catanzariti, A.-M., Ayliffe, M. A., & 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*(3), 755–768. https://doi.org/10.1105/tpc.020040

Duffy, J., Patham, B., & Mensa-Wilmot, K. (2010). Discovery of functional motifs in h-regions of trypanosome signal sequences. *Biochemical Journal*, *426*(2), 135–145. https://doi.org/10.1042/BJ20091277

Ebert, M. K. (2018). *Effector biology of the sugar beet pathogen Cercospora beticola* (Doctoral dissertation). Wageningen University. https://doi.org/10.18174/453825

Emms, D. M., & Kelly, S. (2015). OrthoFinder: Solving fundamental biases in whole genome comparisons dramatically improves orthogroup inference accuracy. *Genome Biology*, *16*(1), 157. https://doi.org/10.1186/s13059-015-0721-2

Enright, A. J., Van Dongen, S., & Ouzounis, C. A. (2002). An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Research*, *30*(7), 1575–1584. https://doi.org/10.1093/nar/30.7.1575

Friesen, T. L., Holmes, D. J., Bowden, R. L., & Faris, J. D. (2018). ToxA Is Present in the U.S. *Bipolaris sorokiniana* Population and Is a Significant Virulence Factor on Wheat Harboring Tsn1. *Plant Disease*, *102*(12), 2446–2452. https://doi.org/10.1094/PDIS-03-18-0521-RE

Friesen, T. L., Stukenbrock, E. H., Liu, Z., Meinhardt, S., Ling, H., Faris, J. D., Rasmussen, J. B., Solomon, P. S., McDonald, B. A., & Oliver, R. P. (2006). Emergence of a new disease as a result of interspecific virulence gene transfer. *Nature Genetics*, *38*(8), 953–956. https://doi.org/10.1038/ng1839

Fry, B. G., Roelants, K., Champagne, D. E., Scheib, H., Tyndall, J. D., King, G. F., Nevalainen, T. J., Norman, J. A., Lewis, R. J., Norton, R. S., Renjifo, C., & de la Vega, R. C. R. (2009). The Toxicogenomic Multiverse: Convergent Recruitment of Proteins Into Animal Venoms. *Annual Review of Genomics and Human Genetics*, *10*(1), 483–511. https://doi.org/10.1146/annurev.genom.9.081307.164356

Fudal, I., Ross, S., Gout, L., Blaise, F., Kuhn, M. L., Eckert, M. R., Cattolico, L., Bernard-Samain, S., Balesdent, M. H., & Rouxel, T. (2007). Heterochromatin-Like Regions as Ecological Niches for Avirulence Genes in the *Leptosphaeria maculans* Genome: Map-Based Cloning of *AvrLm6*. *Molecular Plant-Microbe Interactions*, *20*(4), 459–470. https://doi.org/10.1094/MPMI-20-4-0459

Galagan, J. E., & Selker, E. U. (2004). RIP: The evolutionary cost of genome defense. *Trends in Genetics*, *20*(9), 417–423. https://doi.org/10.1016/j.tig.2004.07.007

Gao, Q., Jin, K., Ying, S.-H., Zhang, Y., Xiao, G., Shang, Y., Duan, Z., Hu, X., Xie, X.-Q., Zhou, G., Peng, G., Luo, Z., Huang, W., Wang, B., Fang, W., Wang, S., Zhong, Y., Ma, L.-J., Leger, R. J. S., … Wang, C. (2011). Genome Sequencing and Comparative Transcriptomics of the Model Entomopathogenic Fungi *Metarhizium anisopliae* and *M. acridum*. *PLOS Genetics*, *7*(1), e1001264. https://doi.org/10.1371/journal.pgen.1001264

Garcia, O., Macedo, J. A. N., Tibúrcio, R., Zaparoli, G., Rincones, J., Bittencourt, L. M. C., Ceita, G. O., Micheli, F., Gesteira, A., Mariano, A. C., Schiavinato, M. A., Medrano, F. J., Meinhardt, L. W., Pereira, G. A. G., & Cascardo, J. C. M. (2007). Characterization of necrosis and ethylene-inducing proteins (NEP) in the basidiomycete *Moniliophthora perniciosa*, the causal agent of witches' broom in *Theobroma cacao*. *Mycological Research*, *111*(4), 443–455. https://doi.org/10.1016/j.mycres.2007.01.017

Ghanbarnia, K., Fudal, I., Larkan, N. J., Links, M. G., Balesdent, M.-H., Profotova, B., Fernando, W. G. D., Rouxel, T., & Borhan, M. H. (2015). Rapid identification of the *Leptosphaeria maculans* avirulence gene *AvrLm2* using an intraspecific comparative genomics approach. *Molecular Plant Pathology*, *16*(7), 699–709. https://doi.org/10.1111/mpp.12228

Glück, A., & Wool, I. G. (1996). Determination of the 28 S Ribosomal RNA Identity Element (G4319) for Alpha-sarcin and the Relationship of Recognition to the Selection of the Catalytic Site. *Journal of Molecular Biology*, *256*(5), 838–848. https://doi.org/10.1006/jmbi.1996.0130

Grandaubert, J., Lowe, R. G., Soyer, J. L., Schoch, C. L., Van de Wouw, A. P., Fudal, I., Robbertse, B., Lapalu, N., Links, M. G., Ollivier, B., Linglin, J., Barbe, V., Mangenot, S., Cruaud, C., Borhan, H., Howlett, B. J., Balesdent, M.-H., & Rouxel, T. (2014). Transposable element-assisted evolution and adaptation to host plant within the *Leptosphaeria maculans-Leptosphaeria biglobosa* species complex of fungal pathogens. *BMC Genomics*, *15*(1), 891. https://doi.org/10.1186/1471-2164-15-891

Hagberg, A. A., Schult, D. A., & Swart, P. J. (2008). Exploring Network Structure, Dynamics, and Function using NetworkX (G. Varoquaux, T. Vaught, & J. Millman, Eds.). In G. Varoquaux, T. Vaught, & J. Millman (Eds.), *Proceedings of the 7th Python in Science Conference*, Pasadena, CA USA.

Hane, J. K., Rouxel, T., Howlett, B. J., Kema, G. H., Goodwin, S. B., & Oliver, R. P. (2011). A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome Biology*, *12*(5), R45. https://doi.org/10.1186/gb-2011-12-5-r45

Hauser, M., Steinegger, M., & Söding, J. (2016). MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. *Bioinformatics*, *32*(9), 1323–1330. https://doi.org/10.1093/bioinformatics/btw006

Herrero-Galán, E., Lacadena, J., Pozo, Á. M. d., Boucias, D. G., Olmo, N., Oñaderra, M., & Gavilanes, J. G. (2008). The insecticidal protein hirsutellin A from the mite fungal pathogen *Hirsutella thompsonii* is a ribotoxin. *Proteins: Structure, Function, and Bioinformatics*, *72*(1), 217–228. https://doi.org/10.1002/prot.21910

Houterman, P. M., Cornelissen, B. J. C., & Rep, M. (2008). Suppression of Plant Resistance Gene-Based Immunity by a Fungal Effector. *PLOS Pathogens*, *4*(5), e1000061. https://doi.org/10.1371/journal.ppat.1000061

Hunter, J. D. (2007). Matplotlib: A 2D graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Irieda, H., Inoue, Y., Mori, M., Yamada, K., Oshikawa, Y., Saitoh, H., Uemura, A., Terauchi, R., Kitakura, S., Kosaka, A., Singkaravanit-Ogawa, S., & Takano, Y. (2019). Conserved fungal effector suppresses

PAMP-triggered immunity by targeting plant immune kinases. *Proceedings of the National Academy of Sciences*, *116*(2), 496–505. https://doi.org/10.1073/pnas.1807297116

Islam, S. M. A., Kearney, C. M., & Baker, E. (2018). Classes, Databases, and Prediction Methods of Pharmaceutically and Commercially Important Cystine-Stabilized Peptides. *Toxins*, *10*(6), 251. https://doi.org/10.3390/toxins10060251

Jones, D. A., Bertazzoni, S., Turo, C. J., Syme, R. A., & Hane, J. K. (2018). Bioinformatic prediction of plant–pathogenicity effector proteins of fungi. *Current Opinion in Microbiology*, *46*, 43–49. https://doi.org/10.1016/j.mib.2018.01.017

Joosten, M. H. A. J., Cozijnsen, T. J., & De Wit, P. J. G. M. (1994). Host resistance to a fungal tomato pathogen lost by a single base-pair change in an avirulence gene. *Nature*, *367*(6461), 384–386. https://doi.org/10.1038/367384a0

Kaas, Q., Yu, R., Jin, A.-H., Dutertre, S., & Craik, D. J. (2012). ConoServer: Updated content, knowledge, and discovery tools in the conopeptide database. *Nucleic Acids Research*, *40*(D1), D325–D330. https://doi.org/10.1093/nar/gkr886

Kale, S. D., Gu, B., Capelluto, D. G. S., Dou, D., Feldman, E., Rumore, A., Arredondo, F. D., Hanlon, R., Fudal, I., Rouxel, T., Lawrence, C. B., Shan, W., & Tyler, B. M. (2010). External Lipid PI3P Mediates Entry of Eukaryotic Pathogen Effectors into Plant and Animal Host Cells. *Cell*, *142*(2), 284–295. https://doi.org/10.1016/j.cell.2010.06.008

Käll, L., Krogh, A., & Sonnhammer, E. L. L. (2004). A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology*, *338*(5), 1027–1036. https://doi.org/10.1016/j.jmb.2004.03.016

Kettles, G. J., Bayon, C., Sparks, C. A., Canning, G., Kanyuka, K., & Rudd, J. J. (2018). Characterization of an antimicrobial and phytotoxic ribonuclease secreted by the fungal wheat pathogen *Zymoseptoria tritici*. *The New Phytologist*, *217*(1), 320–331. https://doi.org/10.1111/nph.14786

Keul, F., Hess, M., Goesele, M., & Hamacher, K. (2017). PFASUM: A substitution matrix from Pfam structural alignments. *BMC Bioinformatics*, *18*(1), 293. https://doi.org/10.1186/s12859-017-1703-z

Khang, C. H., Berruyer, R., Giraldo, M. C., Kankanala, P., Park, S.-Y., Czymmek, K., Kang, S., & Valent, B. (2010). Translocation of *Magnaporthe oryzae* Effectors into Rice Cells and Their Subsequent Cell-to-Cell Movement. *The Plant Cell*, *22*(4), 1388–1403. https://doi.org/10.1105/tpc.109.069666

Kirsten, S., Navarro-Quezada, A., Penselin, D., Wenzel, C., Matern, A., Leitner, A., Baum, T., Seiffert, U., & Knogge, W. (2012). Necrosis-Inducing Proteins of *Rhynchosporium commune*, Effectors in Quantitative Disease Resistance. *Molecular Plant-Microbe Interactions*, *25*(10), 1314–1325. https://doi.org/10.1094/MPMI-03-12-0065-R

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, *305*(3), 567–580. https://doi.org/10.1006/jmbi.2000.4315

Kunjeti, S. G., Iyer, G., Johnson, E., Li, E., Broglie, K. E., Rauscher, G., & Rairdan, G. J. (2016). Identification of *Phakopsora pachyrhizi* Candidate Effectors with Virulence Activity in a Distantly Related Pathosystem. *Frontiers in Plant Science*, *7*. https://doi.org/10.3389/fpls.2016.00269

Lacadena, J., Álvarez-García, E., Carreras-Sangrà, N., Herrero-Galán, E., Alegre-Cebollada, J., García-Ortega, L., Oñaderra, M., Gavilanes, J. G., & Martínez del Pozo, Á. (2007). Fungal ribotoxins: Molecular dissection of a family of natural killers. *FEMS Microbiology Reviews*, *31*(2), 212–237. https://doi.org/10.1111/j.1574-6976.2006.00063.x

Landi, N., Pacifico, S., Ragucci, S., Iglesias, R., Piccolella, S., Amici, A., Di Giuseppe, A. M. A., & Di Maro, A. (2017). Purification, characterization and cytotoxicity assessment of Ageritin: The first

ribotoxin from the basidiomycete mushroom *Agrocybe aegerita*. *Biochimica et Biophysica Acta (BBA) - General Subjects*, *1861*(5, Part A), 1113–1121. https://doi.org/10.1016/j.bbagen.2017.02.023

Laugé, R., Joosten, M. H. A. J., Van den Ackerveken, G. F. J. M., Van den Broek, H. W. J., & de Wit, P. J. G. M. (1997). The In Planta-Produced Extracellular Proteins ECP1 and ECP2 of *Cladosporium fulvum* Are Virulence Factors. *Molecular Plant-Microbe Interactions*, *10*(6), 725–734. https://doi.org/10.1094/MPMI.1997.10.6.725

Li, S., Chen, J., & Liu, B. (2017). Protein remote homology detection based on bidirectional long short-term memory. *BMC Bioinformatics*, *18*(1), 443. https://doi.org/10.1186/s12859-017-1842-2

Lievens, B., Houterman, P. M., & Rep, M. (2009). Effector gene screening allows unambiguous identification of *Fusarium oxysporum* f. sp. *lycopersici* races and discrimination from other formae speciales. *FEMS Microbiology Letters*, *300*(2), 201–215. https://doi.org/10.1111/j.1574-6968.2009.01783.x

Linial, M., Rappoport, N., & Ofer, D. (2017). Overlooked Short Toxin-Like Proteins: A Shortcut to Drug Design. *Toxins*, *9*(11). https://doi.org/10.3390/toxins9110350

Liu, T., Song, T., Zhang, X., Yuan, H., Su, L., Li, W., Xu, J., Liu, S., Chen, L., Chen, T., Zhang, M., Gu, L., Zhang, B., & Dou, D. (2014). Unconventionally secreted effectors of two filamentous pathogens target plant salicylate biosynthesis. *Nature Communications*, *5*(1), 4686. https://doi.org/10.1038/ncomms5686

Liu, Z., Friesen, T. L., Ling, H., Meinhardt, S. W., Oliver, R. P., Rasmussen, J. B., & Faris, J. D. (2006). The Tsn1–ToxA interaction in the wheat–*Stagonospora nodorum* pathosystem parallels that of the wheat–tan spot system. *Genome*, *49*(10), 1265–1273. https://doi.org/10.1139/g06-088

Liu, Z., Zhang, Z., Faris, J. D., Oliver, R. P., Syme, R., McDonald, M. C., McDonald, B. A., Solomon, P. S., Lu, S., Shelver, W. L., Xu, S., & Friesen, T. L. (2012). The Cysteine Rich Necrotrophic Effector SnTox1 Produced by *Stagonospora nodorum* Triggers Susceptibility of Wheat Lines Harboring Snn1. *PLOS Pathogens*, *8*(1), e1002467. https://doi.org/10.1371/journal.ppat.1002467

Lu, S., Faris, J. D., Sherwood, R., Friesen, T. L., & Edwards, M. C. (2014). A dimeric PR-1-type pathogenesis-related protein interacts with ToxA and potentially mediates ToxA-induced necrosis in sensitive wheat: PR-1 potentially mediates ToxA-induced necrosis. *Molecular Plant Pathology*, *15*(7), 650–663. https://doi.org/10.1111/mpp.12122

Lu, S., Gillian Turgeon, B., & Edwards, M. C. (2015). A ToxA-like protein from *Cochliobolus heterostrophus* induces light-dependent leaf necrosis and acts as a virulence factor with host selectivity on maize. *Fungal Genetics and Biology*, *81*, 12–24. https://doi.org/10.1016/j.fgb.2015.05.013

Lyu, X., Shen, C., Fu, Y., Xie, J., Jiang, D., Li, G., & Cheng, J. (2016). A Small Secreted Virulence-Related Protein Is Essential for the Necrotrophic Interactions of *Sclerotinia sclerotiorum* with Its Host Plants. *PLOS Pathogens*, *12*(2), e1005435. https://doi.org/10.1371/journal.ppat.1005435

Ma, J., Wang, S., Wang, Z., & Xu, J. (2014). MRFalign: Protein Homology Detection through Alignment of Markov Random Fields. *PLOS Computational Biology*, *10*(3), e1003500. https://doi.org/10.1371/journal.pcbi.1003500

Malliaros, F. D., & Vazirgiannis, M. (2013). Clustering and community detection in directed networks: A survey. *Physics Reports*, *533*(4), 95–142. https://doi.org/10.1016/j.physrep.2013.08.002

Manning, V. A., Chu, A. L., Steeves, J. E., Wolpert, T. J., & Ciuffetti, L. M. (2009). A host-selective toxin of *Pyrenophora tritici-repentis*, Ptr ToxA, induces photosystem changes and reactive oxygen species accumulation in sensitive wheat. *Molecular Plant-Microbe Interactions*, *22*(6), 665–676. https://doi.org/10.1094/MPMI-22-6-0665

Marshall, R., Kombrink, A., Motteram, J., Loza-Reyes, E., Lucas, J., Hammond-Kosack, K. E., Thomma, B. P. H. J., & Rudd, J. J. (2011). Analysis of Two in Planta Expressed LysM Effector Homologs

from the Fungus *Mycosphaerella graminicola* Reveals Novel Functional Properties and Varying Contributions to Virulence on Wheat. *Plant Physiology*, *156*(2), 756–769. https://doi.org/10.1104/pp.111.176347

Martinez, J. P., Ottum, S. A., Ali, S., Francl, L. J., & Ciuffetti, L. M. (2001). Characterization of the *ToxB* Gene from *Pyrenophora tritici-repentis*. *Molecular Plant-Microbe Interactions*, *14*(5), 675–677. https://doi.org/10.1094/MPMI.2001.14.5.675

Martínez-Ruiz, A., Kao, R., Davies, J., & Martínez del Pozo, Á. (1999). Ribotoxins are a more widespread group of proteins within the filamentous fungi than previously believed. *Toxicon*, *37*(11), 1549–1563. https://doi.org/10.1016/S0041-0101(99)00103-8

M'Barek, S. B., Cordewener, J. H. G., Ghaffary, S. M. T., van der Lee, T. A. J., Liu, Z., Mirzadi Gohari, A., Mehrabi, R., America, A. H. P., Robert, O., Friesen, T. L., Hamza, S., Stergiopoulos, I., de Wit, P. J. G. M., & Kema, G. H. J. (2015). FPLC and liquid-chromatography mass spectrometry identify candidate necrosis-inducing proteins from culture filtrates of the fungal wheat pathogen *Zymoseptoria tritici*. *Fungal Genetics and Biology*, *79*, 54–62. https://doi.org/10.1016/j.fgb.2015.03.015

McDonald, M. C., Ahren, D., Simpfendorfer, S., Milgate, A., & Solomon, P. S. (2018). The discovery of the virulence gene *ToxA* in the wheat and barley pathogen *Bipolaris sorokiniana*. *Molecular Plant Pathology*, *19*(2), 432–439. https://doi.org/10.1111/mpp.12535

McDonald, M. C., Taranto, A. P., Hill, E., Schwessinger, B., Liu, Z., Simpfendorfer, S., Milgate, A., & Solomon, P. S. (2019). Transposon-Mediated Horizontal Transfer of the Host-Specific Virulence Protein ToxA between Three Fungal Wheat Pathogens. *mBio*, *10*(5). https://doi.org/10.1128/mBio.01515-19

McKinney, W. (2010). Data Structures for Statistical Computing in Python (S. J. van der Walt & J. Millman, Eds.). In S. J. van der Walt & J. Millman (Eds.), *Proceedings of the 9th Python in Science Conference*. https://doi.org/10.25080/Majora-92bf1922-00a

McNally, K. E., Menardo, F., Lüthi, L., Praz, C. R., Müller, M. C., Kunz, L., Ben-David, R., Chandrasekhar, K., Dinoor, A., Cowger, C., Meyers, E., Xue, M., Zeng, F., Gong, S., Yu, D., Bourras, S., & Keller, B. (2018). Distinct domains of the AVRPM3A2/F2 avirulence protein from wheat powdery mildew are involved in immune receptor recognition and putative effector function. *New Phytologist*, *218*(2), 681–695. https://doi.org/10.1111/nph.15026

Meinhardt, S. W., Cheng, W., Kwon, C. Y., Donohue, C. M., & Rasmussen, J. B. (2002). Role of the Arginyl-Glycyl-Aspartic Motif in the Action of Ptr ToxA Produced by *Pyrenophora tritici-repentis*. *Plant Physiology*, *130*(3), 1545–1551. https://doi.org/10.1104/pp.006684

Mentlak, T. A., Kombrink, A., Shinya, T., Ryder, L. S., Otomo, I., Saitoh, H., Terauchi, R., Nishizawa, Y., Shibuya, N., Thomma, B. P. H. J., & Talbot, N. J. (2012). Effector-Mediated Suppression of Chitin-Triggered Immunity by *Magnaporthe oryzae* Is Necessary for Rice Blast Disease. *The Plant Cell*, *24*(1), 322–335. https://doi.org/10.1105/tpc.111.092957

Mesarich, C. H., Griffiths, S. A., van der Burgt, A., Ökmen, B., Beenen, H. G., Etalo, D. W., Joosten, M. H. A. J., & de Wit, P. J. G. M. (2014). Transcriptome Sequencing Uncovers the *Avr5* Avirulence Gene of the Tomato Leaf Mold Pathogen *Cladosporium fulvum*. *Molecular Plant-Microbe Interactions*, *27*(8), 846–857. https://doi.org/10.1094/MPMI-02-14-0050-R

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, *41*(12), e121–e121. https://doi.org/10.1093/nar/gkt263

Moolhuijzen, P., See, P. T., Hane, J. K., Shi, G., Liu, Z., Oliver, R. P., & Moffat, C. S. (2018). Comparative genomics of the wheat fungal pathogen *Pyrenophora tritici-repentis* reveals chromosomal variations and genome plasticity. *BMC Genomics*, *19*(1), 279. https://doi.org/10.1186/s12864-018-4680-3

Mosquera, G., Giraldo, M. C., Khang, C. H., Coughlan, S., & Valent, B. (2009). Interaction Transcriptome Analysis Identifies *Magnaporthe oryzae* BAS1-4 as Biotrophy-Associated Secreted Proteins in Rice Blast Disease. *The Plant Cell*, *21*(4), 1273–1290. https://doi.org/10.1105/tpc.107.055228

Mueller, A. N., Ziemann, S., Treitschke, S., Aßmann, D., & Doehlemann, G. (2013). Compatibility in the *Ustilago maydis*–Maize Interaction Requires Inhibition of Host Cysteine Proteases by the Fungal Effector Pit2. *PLOS Pathogens*, *9*(2), e1003177. https://doi.org/10.1371/journal.ppat.1003177

Negi, S. S., Schein, C. H., Ladics, G. S., Mirsky, H., Chang, P., Rascle, J.-B., Kough, J., Sterck, L., Papineni, S., Jez, J. M., Pereira Mouriès, L., & Braun, W. (2017). Functional classification of protein toxins as a basis for bioinformatic screening. *Scientific Reports*, *7*(1), 13940. https://doi.org/10.1038/s41598-017-13957-1

Nguyen, N.-p., Nute, M., Mirarab, S., & Warnow, T. (2016). HIPPI: Highly accurate protein family classification with ensembles of HMMs. *BMC Genomics*, *17*(Suppl 10). https://doi.org/10.1186/s12864-016-3097-0

Nyarko, A., Singarapu, K. K., Figueroa, M., Manning, V. A., Pandelova, I., Wolpert, T. J., Ciuffetti, L. M., & Barbar, E. (2014). Solution NMR Structures of *Pyrenophora tritici-repentis* ToxB and Its Inactive Homolog Reveal Potential Determinants of Toxin Activity. *Journal of Biological Chemistry*, *289*(37), 25946–25956. https://doi.org/10.1074/jbc.M114.569103

Olombrada, M., Lázaro-Gorines, R., López-Rodríguez, J. C., Martínez-del-Pozo, Á., Oñaderra, M., Maestro-López, M., Lacadena, J., Gavilanes, J. G., & García-Ortega, L. (2017). Fungal Ribotoxins: A Review of Potential Biotechnological Applications. *Toxins*, *9*(2). https://doi.org/10.3390/toxins9020071

Oome, S., & Van den Ackerveken, G. (2014). Comparative and Functional Analysis of the Widely Occurring Family of Nep1-Like Proteins. *Molecular Plant-Microbe Interactions*, *27*(10), 1081–1094. https://doi.org/10.1094/MPMI-04-14-0118-R

Owji, H., Nezafat, N., Negahdaripour, M., Hajiebrahimi, A., & Ghasemi, Y. (2018). A comprehensive review of signal peptides: Structure, roles, and applications. *European Journal of Cell Biology*, *97*(6), 422–441. https://doi.org/10.1016/j.ejcb.2018.06.003

Pan, J. (2014). *Ether Bridge Formation and Chemical Diversification in Loline Alkaloid Biosynthesis* (PhD). University of Kentucky. Plant Pathology. Retrieved June 12, 2019, from https://uknowledge.uky.edu/plantpath_etds/14/

Park, J., Teichmann, S. A., Hubbard, T., & Chothia, C. (1997). Intermediate sequences increase the detection of homology between sequences. *Journal of Molecular Biology*, *273*(1), 349–354. https://doi.org/10.1006/jmbi.1997.1288

Paulus, J. K., & van der Hoorn, R. A. L. (2018). Tricked or trapped—two decoy mechanisms in host–pathogen interactions. *PLOS Pathogens*, *14*(2), 1–6. https://doi.org/10.1371/journal.ppat.1006761

Pedersen, C., van Themaat, E. V. L., McGuffin, L. J., Abbott, J. C., Burgis, T. A., Barton, G., Bindschedler, L. V., Lu, X., Maekawa, T., Weßling, R., Cramer, R., Thordal-Christensen, H., Panstruga, R., & Spanu, P. D. (2012). Structure and evolution of barley powdery mildew effector candidates. *BMC Genomics*, *13*, 694. https://doi.org/10.1186/1471-2164-13-694

Peixoto, T. P. (2014). The graph-tool python library. *figshare*. https://doi.org/10.6084/m9.figshare.1164194

Pennington, H. G., Jones, R., Kwon, S., Bonciani, G., Thieron, H., Chandler, T., Luong, P., Morgan, S. N., Przydacz, M., Bozkurt, T., Bowden, S., Craze, M., Wallington, E. J., Garnett, J., Kwaaitaal,

M., Panstruga, R., Cota, E., & Spanu, P. D. (2019). The fungal ribonuclease-like effector protein CSEP0064/BEC1054 represses plant immunity and interferes with degradation of host ribosomal RNA. *PLOS Pathogens*, *15*(3), e1007620. https://doi.org/10.1371/journal.ppat.1007620

Pérez-Cañadillas, J. M., Santoro, J., Campos-Olivas, R., Lacadena, J., Martínez del Pozo, A., Gavilanes, J. G., Rico, M., & Bruix, M. (2000). The highly refined solution structure of the cytotoxic ribonuclease alpha-sarcin reveals the structural requirements for substrate recognition and ribonucleolytic activity. *Journal of Molecular Biology*, *299*(4), 1061–1073. https://doi.org/10.1006/jmbi.2000.3813

Petegrosso, R., Li, Z., Srour, M. A., Saad, Y., Zhang, W., & Kuang, R. (2019). Scalable remote homology detection and fold recognition in massive protein networks. *Proteins: Structure, Function, and Bioinformatics*, *87*(6), 478–491. https://doi.org/10.1002/prot.25669

Petersen, T. N., Brunak, S., Heijne, G. v., & Nielsen, H. (2011). SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nature Methods*, *8*(10), 785–786. https://doi.org/10.1038/nmeth.1701

Petit-Houdenot, Y., Degrave, A., Meyer, M., Blaise, F., Ollivier, B., Marais, C.-L., Jauneau, A., Audran, C., Rivas, S., Veneault-Fourrey, C., Brun, H., Rouxel, T., Fudal, I., & Balesdent, M.-H. (2019). A two genes – for – one gene interaction between *Leptosphaeria maculans* and *Brassica napus*. *New Phytologist*, *223*(1), 397–411. https://doi.org/10.1111/nph.15762

Plissonneau, C., Daverdin, G., Ollivier, B., Blaise, F., Degrave, A., Fudal, I., Rouxel, T., & Balesdent, M.-H. (2016). A game of hide and seek between avirulence genes *AvrLm4-7* and *AvrLm3* in *Leptosphaeria maculans*. *New Phytologist*, *209*(4), 1613–1624. https://doi.org/10.1111/nph.13736

Poppe, S., Dorsheimer, L., Happel, P., & Stukenbrock, E. H. (2015). Rapidly Evolving Genes Are Key Players in Host Specialization and Virulence of the Fungal Wheat Pathogen *Zymoseptoria tritici* (*Mycosphaerella graminicola*). *PLOS Pathogens*, *11*(7), e1005055. https://doi.org/10.1371/journal.ppat.1005055

Praz, C. R., Bourras, S., Zeng, F., Sánchez-Martín, J., Menardo, F., Xue, M., Yang, L., Roffler, S., Böni, R., Herren, G., McNally, K. E., Ben-David, R., Parlange, F., Oberhaensli, S., Flückiger, S., Schäfer, L. K., Wicker, T., Yu, D., & Keller, B. (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*(3), 1301–1314. https://doi.org/10.1111/nph.14372

Ramachandran, S. R., Yin, C., Kud, J., Tanaka, K., Mahoney, A. K., Xiao, F., & Hulbert, S. H. (2016). Effectors from Wheat Rust Fungi Suppress Multiple Plant Defense Responses. *Phytopathology*, *107*(1), 75–83. https://doi.org/10.1094/PHYTO-02-16-0083-R

Rangwala, H., & Karypis, G. (2005). Profile-based direct kernels for remote homology detection and fold recognition. *Bioinformatics*, *21*(23), 4239–4247. https://doi.org/10.1093/bioinformatics/bti687

Rep, M., van der Does, H. C., Meijer, M., Wijk, R. V., Houterman, P. M., Dekker, H. L., de Koster, C. G., & Cornelissen, B. J. C. (2004). A small, cysteine-rich protein secreted by Fusarium oxysporum during colonization of xylem vessels is required for I-3-mediated resistance in tomato. *Molecular Microbiology*, *53*(5), 1373–1383. https://doi.org/10.1111/j.1365-2958.2004.04177.x

Ridout, C. J., Skamnioti, P., Porritt, O., Sacristan, S., Jones, J. D. G., & Brown, J. K. M. (2006). Multiple Avirulence Paralogues in Cereal Powdery Mildew Fungi May Contribute to Parasite Fitness and Defeat of Plant Resistance. *The Plant Cell*, *18*(9), 2402–2414. https://doi.org/10.1105/tpc.106.043307

Robinson, S. D., & Norton, R. S. (2014). Conotoxin Gene Superfamilies. *Marine Drugs*, *12*(12), 6058–6101. https://doi.org/10.3390/md12126058

Sarma, G. N., Manning, V. A., Ciuffetti, L. M., & Karplus, P. A. (2005). Structure of Ptr ToxA: An RGD-Containing Host-Selective Toxin from *Pyrenophora tritici-repentis*. *The Plant Cell*, *17*(11), 3190–3202. https://doi.org/10.1105/tpc.105.034918

Saucedo, A. L., Flores-Solis, D., Vega, R. C. R. d. l., Ramírez-Cordero, B., Hernández-López, R., Cano-Sánchez, P., Navarro, R. N., García-Valdés, J., Coronas-Valderrama, F., Roodt, A. d., Brieba, L. G., Possani, L. D., & Río-Portilla, F. d. (2012). New Tricks of an Old Pattern: STRUCTURAL VERSATILITY OF SCORPION TOXINS WITH COMMON CYSTEINE SPACING. *Journal of Biological Chemistry*, *287*(15), 12321–12330. https://doi.org/10.1074/jbc.M111.329607

Savojardo, C., Martelli, P. L., Fariselli, P., & Casadio, R. (2018). DeepSig: Deep learning improves signal peptide detection in proteins. *Bioinformatics*, *34*(10), 1690–1696. https://doi.org/10.1093/bioinformatics/btx818

Schardl, C. L., Young, C. A., Hesse, U., Amyotte, S. G., Andreeva, K., Calie, P. J., Fleetwood, D. J., Haws, D. C., Moore, N., Oeser, B., Panaccione, D. G., Schweri, K. K., Voisey, C. R., Farman, M. L., Jaromczyk, J. W., Roe, B. A., O'Sullivan, D. M., Scott, B., Tudzynski, P., … Zeng, Z. (2013). Plant-Symbiotic Fungi as Chemical Engineers: Multi-Genome Analysis of the Clavicipitaceae Reveals Dynamics of Alkaloid Loci. *PLOS Genetics*, *9*(2), e1003323. https://doi.org/10.1371/journal.pgen.1003323

Schardl, C. L., Young, C. A., Moore, N., Krom, N., Dupont, P.-Y., Pan, J., Florea, S., Webb, J. S., Jaromczyk, J., Jaromczyk, J. W., Cox, M. P., & Farman, M. L. (2014). Chapter Ten Genomes of Plant-Associated Clavicipitaceae. In *Fungi* (pp. 291–327). https://doi.org/10.1016/b978-0-12-397940-7.00010-0

Schmidt, S. M., Kuhn, H., Micali, C., Liller, C., Kwaaitaal, M., & Panstruga, R. (2014). Interaction of a *Blumeria graminis* f. sp. *hordei* effector candidate with a barley ARF-GAP suggests that host vesicle trafficking is a fungal pathogenicity target: *Blumeria graminis* effector candidates. *Molecular Plant Pathology*, *15*(6), 535–549. https://doi.org/10.1111/mpp.12110

Schmidt, S. M., Lukasiewicz, J., Farrer, R., Dam, P. v., Bertoldo, C., & 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*(1), 307–318. https://doi.org/10.1111/nph.13584

Schneider, T. D., & Stephens, R. M. (1990). Sequence logos: A new way to display consensus sequences. *Nucleic Acids Research*, *18*(20), 6097–6100. Retrieved July 23, 2020, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC332411/

Schweizer, G., Münch, K., Mannhaupt, G., Schirawski, J., Kahmann, R., & Dutheil, J. Y. (2018). Positively Selected Effector Genes and Their Contribution to Virulence in the Smut Fungus *Sporisorium reilianum*. *Genome Biology and Evolution*, *10*(2), 629–645. https://doi.org/10.1093/gbe/evy023

Sharpee, W., Oh, Y., Yi, M., Franck, W., Eyre, A., Okagaki, L. H., Valent, B., & Dean, R. A. (2017). Identification and characterization of suppressors of plant cell death (SPD) effectors from *Magnaporthe oryzae*. *Molecular Plant Pathology*, *18*(6), 850–863. https://doi.org/10.1111/mpp.12449

Shiller, J., Van de Wouw, A. P., Taranto, A. P., Bowen, J. K., Dubois, D., Robinson, A., Deng, C. H., & Plummer, K. M. (2015). A Large Family of *AvrLm6*-like Genes in the Apple and Pear Scab Pathogens, *Venturia inaequalis* and *Venturia pirina*. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.00980

Sievers, F., & Higgins, D. G. (2018). Clustal Omega for making accurate alignments of many protein sequences. *Protein Science*, *27*(1), 135–145. https://doi.org/10.1002/pro.3290

Spanu, P. D. (2017). Cereal immunity against powdery mildews targets RNase-Like Proteins associated with Haustoria (RALPH) effectors evolved from a common ancestral gene. *New Phytologist*, *213*(3), 969–971. https://doi.org/10.1111/nph.14386

Staats, M., van Baarlen, P., Schouten, A., van Kan, J. A. L., & Bakker, F. T. (2007). Positive selection in phytotoxic protein-encoding genes of Botrytis species. *Fungal Genetics and Biology*, *44*(1), 52–63. https://doi.org/10.1016/j.fgb.2006.07.003

Steinegger, M., Meier, M., Mirdita, M., Vöhringer, H., Haunsberger, S. J., & Söding, J. (2019). HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinformatics*, *20*(1), 473. https://doi.org/10.1186/s12859-019-3019-7

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35*, 1026–1028. https://doi.org/10.1038/nbt.3988

Stephenson, S.-A., Hatfield, J., Rusu, A. G., Maclean, D. J., & Manners, J. M. (2000). *CgDN3*: An Essential Pathogenicity Gene of *Colletotrichum gloeosporioides* Necessary to Avert a Hypersensitive-Like Response in the Host *Stylosanthes guianensis*. *Molecular Plant-Microbe Interactions*, *13*(9), 929–941. https://doi.org/10.1094/MPMI.2000.13.9.929

Stergiopoulos, I., Kourmpetis, Y. A. I., Slot, J. C., Bakker, F. T., De Wit, P. J. G. M., & Rokas, A. (2012). In Silico Characterization and Molecular Evolutionary Analysis of a Novel Superfamily of Fungal Effector Proteins. *Molecular Biology and Evolution*, *29*(11), 3371–3384. https://doi.org/10.1093/molbev/mss143

Szklarczyk, R., Wanschers, B. F., Cuypers, T. D., Esseling, J. J., Riemersma, M., van den Brand, M. A., Gloerich, J., Lasonder, E., van den Heuvel, L. P., Nijtmans, L. G., & Huynen, M. A. (2012). Iterative orthology prediction uncovers new mitochondrial proteins and identifies C12orf62 as the human ortholog of COX14, a protein involved in the assembly of cytochrome coxidase. *Genome Biology*, *13*(2), R12. https://doi.org/10.1186/gb-2012-13-2-r12

Tai, Y.-S., Bragg, J., & Meinhardt, S. W. (2007). Functional Characterization of ToxA and Molecular Identification of its Intracellular Targeting Protein in Wheat. *American Journal of Plant Physiology*, *2*(2), 76–89. https://doi.org/10.3923/ajpp.2007.76.89

Tareen, A., & Kinney, J. B. (2020). Logomaker: Beautiful sequence logos in Python. *Bioinformatics*, *36*(7), 2272–2274. https://doi.org/10.1093/bioinformatics/btz921

Tatusov, R. L., Koonin, E. V., & Lipman, D. J. (1997). A Genomic Perspective on Protein Families. *Science*, *278*(5338), 631–637. https://doi.org/10.1126/science.278.5338.631

Testa, A. C., Hane, J. K., Ellwood, S. R., & Oliver, R. P. (2015). CodingQuarry: Highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics*, *16*(1), 170. https://doi.org/10.1186/s12864-015-1344-4

Testa, A. C., Oliver, R. P., & Hane, J. K. (2016). OcculterCut: A Comprehensive Survey of AT-Rich Regions in Fungal Genomes. *Genome Biology and Evolution*, *8*(6), 2044–2064. https://doi.org/10.1093/gbe/evw121

Thomma, B. P. H. J., Nürnberger, T., & Joosten, M. H. A. J. (2011). Of PAMPs and Effectors: The Blurred PTI-ETI Dichotomy. *The Plant Cell*, *23*(1), 4–15. https://doi.org/10.1105/tpc.110.082602

Thrall, P. H., Barrett, L. G., Dodds, P. N., & Burdon, J. J. (2016). Epidemiological and Evolutionary Outcomes in Gene-for-Gene and Matching Allele Models. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.01084

Tuori, R. P., Wolpert, T. J., & Ciuffetti, L. M. (1995). Purification and immunological characterization of toxic components from cultures of *Pyrenophora tritici-repentis*. *Molecular Plant-Microbe Interactions*, *8*(1), 41–48. https://doi.org/10.1094/mpmi-8-0041

Urban, M., Cuzick, A., Seager, J., Wood, V., Rutherford, K., Venkatesh, S. Y., De Silva, N., Martinez, M. C., Pedro, H., Yates, A. D., Hassani-Pak, K., & Hammond-Kosack, K. E. (2020). PHI-base:

The pathogen–host interactions database. *Nucleic Acids Research*, *48*(D1), D613–D620. https://doi.org/10.1093/nar/gkz904

Van de Wouw, A. P., Lowe, R. G. T., Elliott, C. E., Dubois, D. J., & Howlett, B. J. (2014). An avirulence gene, *AvrLmJ1*, from the blackleg fungus, *Leptosphaeria maculans*, confers avirulence to *Brassica juncea* cultivars. *Molecular Plant Pathology*, *15*(5), 523–530. https://doi.org/10.1111/mpp.12105

Vargas, W. A., Sanz-Martín, J. M., Rech, G. E., Armijos-Jaramillo, V. D., Rivera, L. P., Echeverria, M. M., Díaz-Mínguez, J. M., Thon, M. R., & Sukno, S. A. (2015). A Fungal Effector With Host Nuclear Localization and DNA-Binding Properties Is Required for Maize Anthracnose Development. *Molecular Plant-Microbe Interactions*, *29*(2), 83–95. https://doi.org/10.1094/MPMI-09-15-0209-R

Viegas, A., Herrero-Galán, E., Oñaderra, M., Macedo, A. L., & Bruix, M. (2009). Solution structure of hirsutellin A–new insights into the active site and interacting interfaces of ribotoxins. *The FEBS journal*, *276*(8), 2381–2390. https://doi.org/10.1111/j.1742-4658.2009.06970.x

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Contributors, S. 1. 0. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/https://doi.org/10.1038/s41592-019-0686-2

Vleeshouwers, V. G. A. A., & Oliver, R. P. (2014). Effectors as Tools in Disease Resistance Breeding Against Biotrophic, Hemibiotrophic, and Necrotrophic Plant Pathogens. *Molecular Plant-Microbe Interactions*, *27*(3), 196–206. https://doi.org/10.1094/MPMI-10-13-0313-IA

Wagner, I., Volkmer, M., Sharan, M., Villaveces, J. M., Oswald, F., Surendranath, V., & Habermann, B. H. (2014). morFeus: A web-based program to detect remotely conserved orthologs using symmetrical best hits and orthology network scoring. *BMC Bioinformatics*, *15*(1), 263. https://doi.org/10.1186/1471-2105-15-263

Wang, C., Liu, Y., Liu, L., Wang, Y., Yan, J., Wang, C., Li, C., & Yang, J. (2019). The biotrophy-associated secreted protein 4 (BAS4) participates in the transition of *Magnaporthe oryzae* from the biotrophic to the necrotrophic phase. *Saudi Journal of Biological Sciences*, *26*(4), 795–807. https://doi.org/10.1016/j.sjbs.2019.01.003

Wang, J.-Y., Cai, Y., Gou, J.-Y., Mao, Y.-B., Xu, Y.-H., Jiang, W.-H., & Chen, X.-Y. (2004). VdNEP, an Elicitor from *Verticillium dahliae*, Induces Cotton Plant Wilting. *Applied and Environmental Microbiology*, *70*(8), 4989–4995. https://doi.org/10.1128/AEM.70.8.4989-4995.2004

Wawra, S., Fesel, P., Widmer, H., Timm, M., Seibel, J., Leson, L., Kesseler, L., Nostadt, R., Hilbert, M., Langen, G., & Zuccaro, A. (2016). The fungal-specific β-glucan-binding lectin FGB1 alters cell-wall composition and suppresses glucan-triggered immunity in plants. *Nature Communications*, *7*(1), 13188. https://doi.org/10.1038/ncomms13188

Wright, E. S. (2015). DECIPHER: Harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics*, *16*(1), 322. https://doi.org/10.1186/s12859-015-0749-z

Yang, X., & Moffat, K. (1996). Insights into specificity of cleavage and mechanism of cell entry from the crystal structure of the highly specific Aspergillus ribotoxin, restrictocin. *Structure*, *4*(7), 837–852. https://doi.org/10.1016/s0969-2126(96)00090-1

Yoshino, K., Irieda, H., Sugimoto, F., Yoshioka, H., Okuno, T., & Takano, Y. (2012). Cell Death of *Nicotiana benthamiana* Is Induced by Secreted Protein NIS1 of *Colletotrichum orbiculare* and Is Suppressed by a Homologue of CgDN3. *Molecular Plant-Microbe Interactions*, *25*(5), 625–636. https://doi.org/10.1094/MPMI-12-11-0316

Zhang, L., Ni, H., Du, X., Wang, S., Ma, X.-W., Nürnberger, T., Guo, H.-S., & Hua, C. (2017). The Verticillium-specific protein VdSCP7 localizes to the plant nucleus and modulates immunity to fungal infections. *New Phytologist*, *215*(1), 368–381. https://doi.org/10.1111/nph.14537

CHAPTER 5 — THEME 1

“CATAStrophy”, a Genome-Informed Trophic Classification of Filamentous Plant Pathogens — How Many Different Types of Filamentous Plant Pathogens Are There?

This chapter is also published in:
Frontiers in Microbiology, 2020, vol. 10, p. 3088
https://doi.org/10.3389/fmicb.2019.03088

5.1 Declaration

Title	"CATAStrophy", a Genome-Informed Trophic Classification of Filamentous Plant Pathogens – How Many Different Types of Filamentous Plant Pathogens Are There?
Authors	James K. Hane, Jonathan Paxman, **Darcy A. B. Jones**, Richard P. Oliver, and Pierre de Wit
Publication	2020. *Frontiers in Microbiology*, 10, 3088.
DOI	`https://doi.org/10.3389/fmicb.2019.03088`

This thesis chapter is submitted in the form of a collaboratively-written and peer-reviewed manuscript. As such, not all work contained within this chapter can be attributed to the Ph. D. candidate.

The following contributions were made by the Ph.D candidate (DABJ) and co-authors:

- JKH, RPO, PDW conceived the study.
- JKH and **DABJ** performed the bioinformatics analysis.
- **DABJ** and **JP** performed the multivariate analysis.
- **DABJ** developed CATAStrophy software and pipeline.
- JKH and RPO wrote the manuscript.
- JKH, RPO, and PW edited the manuscript.
- All authors read and approved the manuscript.

I, Darcy Jones, hereby certify that this attribution statement is an accurate record of my contribution to the research presented in this chapter.

Darcy Jones

I certify that this attribution statement is an accurate record of Darcy Jones' contribution to the research presented in this chapter.

James K. Hane | Jonathan Paxman | Richard P. Oliver

Pierre de Wit

frontiers in Microbiology

METHODS
published: 21 January 2020
doi: 10.3389/fmicb.2019.03088

"CATAStrophy," a Genome-Informed Trophic Classification of Filamentous Plant Pathogens – How Many Different Types of Filamentous Plant Pathogens Are There?

James K. Hane[1,2], Jonathan Paxman[3], Darcy A. B. Jones[1], Richard P. Oliver[1]* and Pierre de Wit[4]*

[1] Centre for Crop and Disease Management, School of Molecular and Life Sciences, Curtin University, Perth, WA, Australia, [2] Curtin Institute for Computation, Faculty of Science and Engineering, Curtin University, Perth, WA, Australia, [3] Department of Mechanical Engineering, Curtin University, Perth, WA, Australia, [4] Laboratory of Phytopathology, Department of Plant Sciences, Wageningen University & Research, Wageningen, Netherlands

OPEN ACCESS

Edited by:
Baolei Jia,
Chung-Ang University, South Korea

Reviewed by:
Brett Merrick Tyler,
Oregon State University, United States
Krishna V. Subbarao,
University of California, Davis, United States
Remco Stam,
Technical University of Munich, Germany
Jinliang Liu,
Jilin University, China

***Correspondence:**
James K. Hane
James.Hane@curtin.edu.au
Richard P. Oliver
Richard.Oliver@curtin.edu.au

Specialty section:
This article was submitted to Evolutionary and Genomic Microbiology, a section of the journal Frontiers in Microbiology

Received: *29 August 2019*
Accepted: *20 December 2019*
Published: *21 January 2020*

Citation:
Hane JK, Paxman J, Jones DAB, Oliver RP and de Wit P (2020) "CATAStrophy," a Genome-Informed Trophic Classification of Filamentous Plant Pathogens – How Many Different Types of Filamentous Plant Pathogens Are There? Front. Microbiol. 10:3088. doi: 10.3389/fmicb.2019.03088

The traditional classification of fungal and oomycete phytopathogens into three classes – biotrophs, hemibiotrophs, or necrotrophs – is unsustainable. This study highlights multiple phytopathogen species for which these labels have been inappropriately applied. We propose a novel and reproducible classification based solely on genome-derived analysis of carbohydrate-active enzyme (CAZyme) gene content called CAZyme-Assisted Training And Sorting of -trophy (CATAStrophy). CATAStrophy defines four major divisions for species associated with living plants. These are monomertrophs (Mo) (corresponding to biotrophs), polymertrophs (P) (corresponding to necrotrophs), mesotrophs (Me) (corresponding to hemibiotrophs), and vasculartrophs (including species commonly described as wilts, rots, or anthracnoses). The Mo class encompasses symbiont, haustorial, and non-haustorial species. Me are divided into the subclasses intracellular and extracellular Me, and the P into broad and narrow host subclasses. This gives a total of seven discrete plant-pathogenic classes. The classification provides insight into the properties of these species and offers a facile route to develop control measures for newly recognized diseases. Software for CATAStrophy is available online at https://github.com/ccdmb/catastrophy. We present the CATAStrophy method for the prediction of trophic phenotypes based on CAZyme gene content, as a complementary method to the traditional tripartite "biotroph–hemibiotroph–necrotroph" classifications that may encourage renewed investigation and revision within the fungal biology community.

Keywords: fungi, plant pathogen, biotroph, necrotroph, hemibiotroph, CAZymes, metabolism

INTRODUCTION

Fungal and oomycete plant pathogens cause crop losses of ∼15–25% of yield potential (Fisher et al., 2018; Savary et al., 2019) and just five diseases destroy crops that could feed >600 million people (Fisher et al., 2012; Bebber and Gurr, 2015; Gurr et al., 2015). Combating such diseases is an ongoing challenge requiring good understanding of interactions between pathogens and hosts. Fungal and oomycete pathogens have been classified by modes of nutrition for over 130 years

(de Bary and Garnsey, 1887), but in the last 50 years the dominant model has been a division into three "trophic" classes, biotrophs, hemibiotrophs, and necrotrophs (Thrower, 1966; Lewis, 1973). Non-pathogen species are described as symbionts (or commensals) when living on or within a living host without causing significant damage, or as saprotrophs (S) (or in older literature as saprophytes) when they extract nutrients solely from decaying biomaterials. The suffix "-trophic" emphasizes that this model refers to the feeding mode of the pathogens. Biotrophs feed on living host tissues and necrotrophs on dead tissues. Hemibiotrophs start infection as a biotroph and subsequently switch to necrotrophy [see **Box 1** for a conventional statement of the definitions]. The biotrophic, hemibiotrophic, and necrotrophic classes have become associated with a number of other properties (**Table 1**).

It is widely acknowledged that this model of plant pathogen classification leaves much to be desired. Many pathogens are placed by different authors in two and, in a few cases, all three classes (Oliver and Ip-Cho, 2004; Stotz et al., 2014). None of the features listed in **Table 1** are diagnostic, with the possible exception that all obligate pathogens are biotrophic, but the converse is not true. There are substantial differences in the hemibiotrophic lifestyle with some species having a clear temporal division between biotrophic and necrotrophic phase, while in others the trophic phase can coincide in time but in differentiated tissues of the infected host. Classifications based on host-range or type of defense mechanism are not supported by well-established data. Furthermore nearly all resistance genes are, in some circumstances, quantitative (Poland et al., 2009).

The fundamental basis of the difference between biotrophy and necrotrophy – feeding on living and dead cells – is difficult to apply. Firstly, it is unclear precisely when a host cell dies and secondly, as all fungi and oomycetes feed by extracellular osmotrophic adsorption (Richards and Talbot, 2013), it is unclear which host cells can be said to be feeding the pathogen. Infected tissue might contain both living and dead host cells, both of which are releasing nutrients. Other groupings of plant pathogens have been proposed. For example, wilt pathogens are defined as colonizers of xylem vessels and surrounding parenchyma tissues and cause characteristic symptoms associated with water stress. It is unclear whether these pathogens have more in common with biotrophs or necrotrophs (Klosterman et al., 2011).

The first completed genome sequence was brewer's yeast in 1996 (Goffeau et al., 1996) and fungal plant pathogen genomes followed from 2005. In this report, we studied 158 plant pathogen genomes including those of 143 fungal and 15 oomycete species or isolates (Pedro et al., 2016; **Supplementary Data Sheet S1**). The motivation was to determine whether an unbiased examination of this wealth of genome sequence data would reveal an objective and robust classification system that had predictive power. We sought a method that would exclusively utilize genome-derived sequences and not require expression analyses or any other *in vivo* assessments to predict the trophic phenotype of a novel pathogen species.

In this study, we used counts of carbohydrate-active enzyme (CAZyme)-encoding genes (Lombard et al., 2014) to generate a novel classification of plant pathogens. Our analysis suggests the existing tripartite trophic classification system is unsustainable, highlights longstanding anomalies, and permits the objective prediction of trophic phenotype based on data common to all genome projects. The process grouped species with similar trophic phenotypes regardless of their phylogenetic history. We identified novel groups comprising four major plant pathogen classes [monomertrophs (Mo), polymertrophs (P), mesotrophs (Me), and vasculartrophs (V)], two of which could be further divided into two sub-classes (**Figure 1**). The Mo primarily metabolize simple sugars, P metabolize complex sugars, and Me have characteristics of both. These novel classes are roughly analogous to biotrophs, necrotrophs, and hemibiotrophs, respectively. The data included in this study were used to develop and train a predictive tool for CAZyme-Assisted Training And Sorting of -trophy (CATAStrophy), available online at https://github.com/ccdmb/catastrophy. We present the CATAStrophy method for the prediction of trophic phenotypes basedn on CAZyme gene content, as a complementary method to the traditional tripartite "biotroph–hemibiotroph–necrotroph" classifications that may encourage renewed investigation and revision within the fungal biology community.

BOX 1 | Conventional terms for describing plant pathogen trophic phenotypes.
Biotroph – feeding from within living host cells throughout its lifecycle.
Necrotroph – feeding from dead (or dying) host cells.
Hemibiotroph – initially feeding as a biotroph and then switching to necrotrophy.
Saprotroph – a fungus that only lives on dead organic material.

RESULTS

Our goal was to use only genome sequences to determine whether existing or new classifications of filamentous plant pathogens were objectively supported, as gene transcript data or cell-biological observations would eliminate the universality of the approach. Initial investigations revealed that a small set of gene functions was necessary to reduce noise. We focused on genes encoding CAZymes (Cantarel et al., 2009; Lombard et al., 2014), a ubiquitous, large, and well-defined set that can be auto-annotated in a consistent manner. Furthermore, CAZyme genes typically reside in genome regions less prone to *de novo* assembly errors (Soanes et al., 2008). The CAZyme gene contents of 133 fungal and 15 oomycete species/*formae speciales*, and CAZyme annotations were assigned for 136–1314 genes in fungi and 255–793 genes in oomycetes (**Supplementary Data Sheet S2**).

Principal component analysis (PCA) of CAZyme contents across a training set of 85 fungal and oomycete species (**Supplementary Data Sheet S1**) allowed the separation of most of the species with the first two principal components (PCs) (**Figure 2**, Step 1), containing 56.5 and 10.7% of variation, respectively. PC2 separated species predominantly based on phylogeny, with the Oomycota generally having high values, Ascomycota low values, and Basidiomycota low to intermediate

TABLE 1 | Alleged typical properties of pathogenic trophic classes.

Property	Biotroph	Hemibiotroph	Necrotroph
Feeding (Scott, 1972; Parbery, 1996)	On living host cells	Initially on living and later on dying/dead host cells	On dead or dying host cells
Obligate or facultative (Scott, 1972; Parbery, 1996)	Obligate	Facultative	Facultative
Feeding structures (Gay, 1984; Mendgen et al., 2000; Laluk and Mengiste, 2010)	Haustoria	Haustorium-like structures (appressoria/hyphopodia) in some cases	No haustoria
Host range (Lewis, 1973; Lucas, 1998; Zeilinger et al., 2016)	Narrow	Narrow	Broad
Hormones involved in defense (Hammond-Kosack and Parker, 2003; Glazebrook, 2005)	Salicylic acid	Salicylic/Jasmonic acid	Jasmonic acid
Effectors (Stergiopoulos and de Wit, 2009; Tan et al., 2010; Koeck et al., 2011)	Avirulence effectors; gene-for-gene interactions	Avirulence effectors; gene-for-gene interactions	Host-specific toxins; necrotrophic effectors
Resistance genes (Glazebrook, 2005; Wang et al., 2014)	Qualitative	Qualitative	Quantitative

values. PC1 separated trophic classes into an approximate spectrum progressing from the traditionally classified S to biotrophs, hemibiotrophs, and necrotrophs. While a trend was apparent, using the trophic terms assigned based on commonly usage in literature (**Figure 3** and **Supplementary Data Sheet S1**), these terms were not consistently clustered within the same regions of PCA space. We also used novel trophic classifications proposed in this study consisting of five major classes (**Figure 2**, Step 1), two of which were each sub-dividable into two sub-classes. Species commonly described as wilts formed a distinct group with high PC1 values and low PC2 values (**Figure 2**, Step 1), suggesting the need for the creation of a new class.

We propose a novel trophic nomenclature that contains five major classes (**Figure 1**, section "Materials and Methods") and introduces new class names derived from our CAZyme-based approach. The S class remains unchanged, while the traditional biotroph and necrotroph classes are replaced by Mo and P, respectively, reflecting a preference for either monomeric or oligomeric/polymeric primary nutrient sources. Two novel classes are proposed which broadly replace the hemibiotrophs; these are Me (from "meso" meaning intermediate) and V, which comprises pathogens commonly described as wilts, anthracnoses, and rots. The P are divided into two sub-classes that correspond to polyphagy [broad host range (PB) or host-specificity (narrow host range {PN}). The Me class divided into two sub-classes corresponding to intracellular (MeI)] or extracellular (MeE) interactions. Hence, there are a total of four major classes of fungi and oomycetes that all interact with living plants (Mo, P, Me, and V) alongside the non-pathogenic S, and four informative sub-classes (MeI, MeE, PN, and PB).

After applying our novel nomenclatures to the PCA data (**Figure 2**, Step 1) we observed improvements in how species of the same trophic classification grouped into homogeneous clusters (**Figures 3B,C**). Our method for testing and predicting trophic phenotypes had to deal with cases where species were roughly equidistant to two or more clusters within the PCA space (**Figure 2**, Step 2). We therefore calculated centroids in the PCA space and developed metrics for the relative distances to the centroids of each trophic class, which we refer to herein as "relative centroid distance" (RCD) (**Figure 2**, Step 3).

We predicted each species as a member of one of the five major classes (S, Mo, Me, P, and V), and also assigned one or more secondary "affinities," for sub-classes of the Me and P classes (MeI, MeE, PN, and PB) or alternate major classes that differed from the primary class prediction. We observed the RCD method (see the section "Materials and Methods") using our novel trophic classes to be generally consistent with our overall biological expectations of trophic phenotypes (**Figures 3B,C**) and report our predictions for the 158 isolates included in this study (**Table 2** and **Supplementary Data Sheet S1**). We observed several examples of distantly related taxa being predicted in the same trophic class and conversely species of the same genus accurately placed into different trophic classes. Rate of successful prediction (**Supplementary Data Sheet S1**) was 77% compared to terms derived from common usage in the literature; however, the curated success rate was 90% after taking into account recent literature revisions and other caveats outlined in the discussion and noted in **Supplementary Data Sheet S1**.

DISCUSSION

The Five-Trophic Classes: Saprotrophs, Monomertrophs, Polymertrophs, Mesotrophs, and Vasculartrophs

Since the inception of plant pathology, classification of filamentous fungal and oomycete plant pathogens into subgroups has been attempted based on nutritive phenotypes (de Bary and Garnsey, 1887). A tripartite division into biotrophs, necrotrophs, or hemibiotrophs has dominated the field for 50 years (Thrower, 1966; Lewis, 1973). It is striking that even with advancements in microscopy, allowing observations of host–microbe interactions

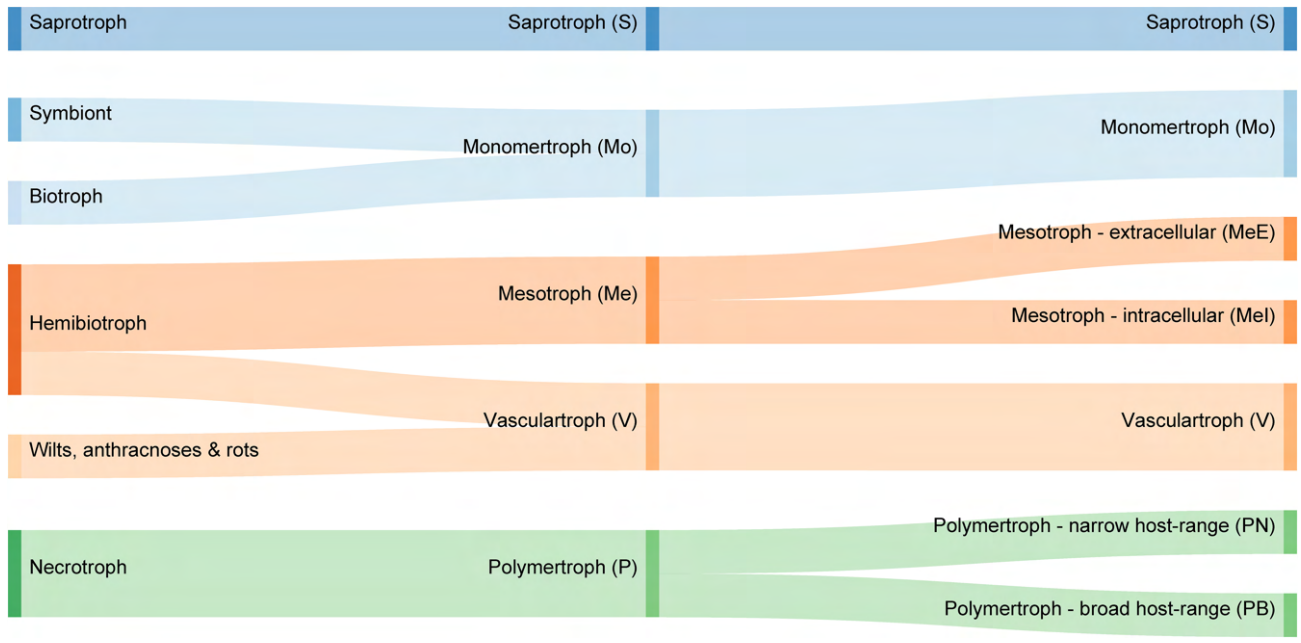


FIGURE 1 | Comparison of common trophic terms used in plant pathology literature **(left)**, with our proposed novel classification system of five major trophic classes **(middle)** and nine sub-classes **(right)**.

at the cellular level, these divisions have persisted despite many obvious anomalies (Kuo et al., 2014; Stotz et al., 2014; Sánchez-Vallet et al., 2015; Videira et al., 2017). These divisions have been causally linked to broader features of their host interactions (Glazebrook, 2005) and thence directed strategies for disease control (Oliver, 2009; Burdon et al., 2014).

The genomics era has given us a plethora of data with which to generate an objective classification system that would aid development of sustainable control strategies for both familiar and emergent plant pathogens (Fisher et al., 2012). The CATAStrophy method provides a non-biased way to predict the trophic (sub-)class of filamentous plant pathogens solely based on their CAZyme gene content. The discussion below focuses on key species – we invite readers to view comprehensive reports of species and their trophic predictions in **Supplementary Data Sheet S1** and **Supplementary Text S1**.

Monomertrophs

Perhaps the most distinctive of the traditionally defined pathogens classes are the biotrophs. Archetypal biotrophs complete their lifecycles only on their specific hosts and typically exhibit clear-cut gene-for-gene host interactions involving biotrophic effectors (syn. avirulence determinants) (Tanaka et al., 2015). Their extreme host specialization is linked to the absence of several primary biosynthetic pathways (**Supplementary Text S1**). Archetypal biotrophs feed via specific structures, haustoria, which invaginate the host cell membranes and permit the adsorption of nutrients directly from the host cytoplasm (Staples, 2001). Haustoria have evolved multiple times and are found in Ascomycota (powdery mildews), Basidiomycota (rusts), and the Oomycota (downy mildews and *Phytophthora* species). They are also found in true symbionts, including the mycorrhizal Glomeromycota.

The CATAStrophy method linked phylogenetically disparate groups of traditional biotrophs and symbionts into the Mo class, including oomycetes (e.g., *Phytophthora* spp., *Albugo* spp., and *Hyaloperonospora arabidopsidis*), rust and smut fungi (e.g., *Puccinia* spp., *Ustilago* spp., and *Melampsora laricis-populina*), the powdery mildews (e.g., *Erysiphe necator* and *Blumeria graminis*), and known symbionts and mycorrhiza (e.g., *Epichloë* spp., *Pisolithus* spp., *Laccaria bicolour*, and *Tuber melanosporum*). Excluded from this class were traditionally defined biotrophs such as *Fulvia fulva* (syn. *Cladosporium fulvum*, *Passalora fulva*) and *Venturia* spp., which lack haustoria. Indeed, *F. fulva* was long regarded as a model for the biotrophs (de Wit, 2016). However, recent studies have concluded that both *F. fulva* and *Venturia* spp. are hemibiotrophic (Stotz et al., 2014).

The Mo class was the least well-predicted by CATAStrophy, in that haustorial and non-haustorial sub-classes could not be adequately distinguished, nor could the symbionts. Biotrophs and symbionts have low CAZYme (**Supplementary Data Sheet S2**) and secondary metabolite gene contents (**Supplementary Text S1**). This is consistent with a common strategy of causing minimal damage to host cells, i.e., producing fewer PAMPs or DAMPs. Free-living yeast species were also cryptically predicted in this class, likely due to their preference for unpolymerized sugars (Rodrigues et al., 2006) that parallel haustorial biotrophic metabolism (Hahn and Mendgen, 1997, 2001; Voegele et al., 2001). Yeasts and species like *N. crassa* are the first colonizers of rich sources of sugars and amino

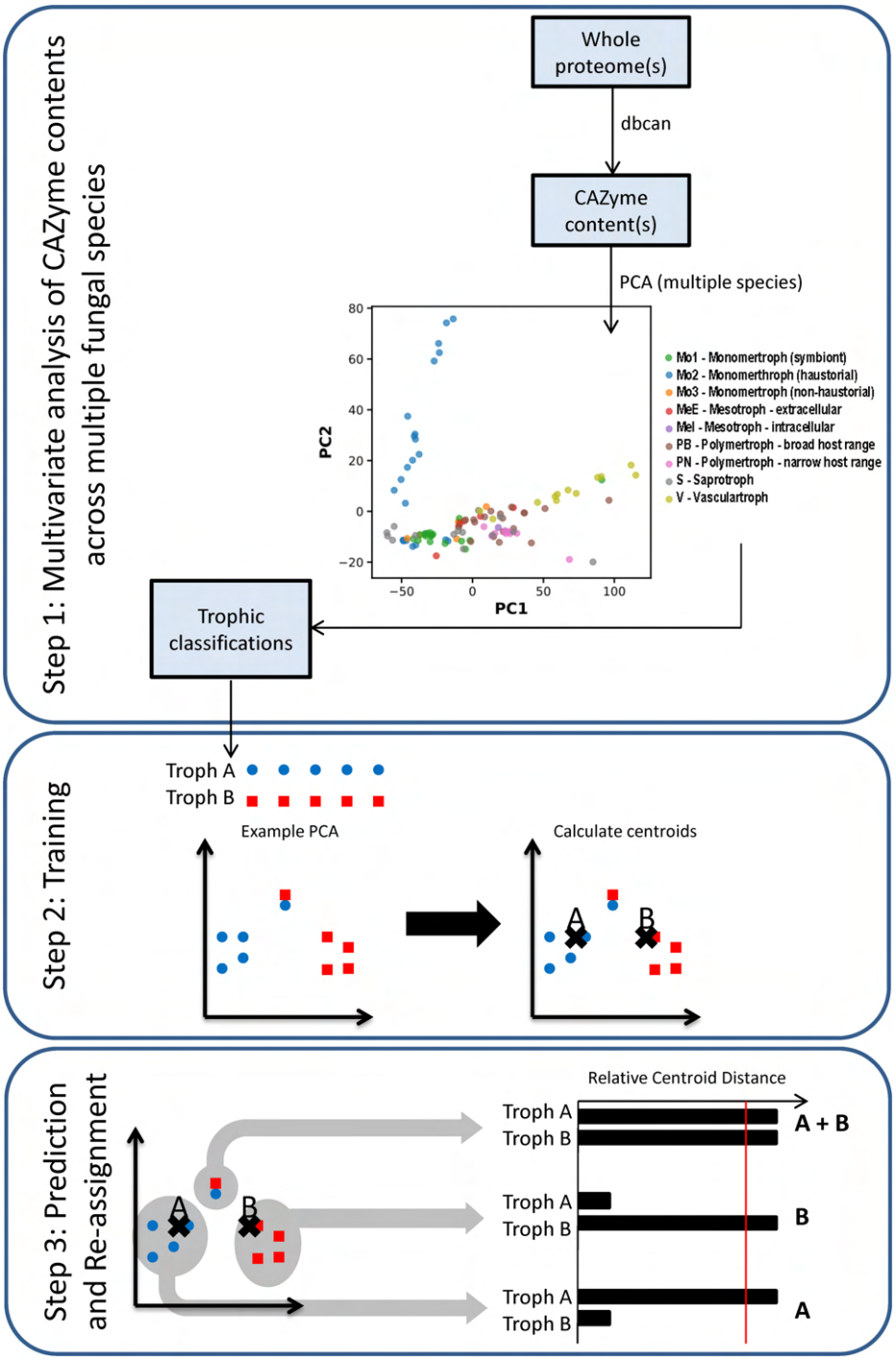


FIGURE 2 | Workflow of the trophic prediction method. Step 1: CAZyme gene contents are compared across species using multivariate analysis. Step 2: trophic classifications are assigned to each species (**Supplementary Data Sheet S1**) and centroids are calculated for each trophic class. Step 3: relative centroid distances (RCDs) are calculated for each species, with the closest centroid assigned an RCD score of 1, the furthest as 0, and other centroid distances expressed as a relative proportion. Species were predicted as members of a major trophic class where RCD = 1 and assigned additional affinities for other classes or sub-classes where RCD ≥ 0.95. Importantly, species may be predicted after RCD calculation into a different class than was initially assigned.

acids, and some strains lack enzymes needed even for modestly polymerized substrates (e.g., sucrose). Species in the Mo class generally have the lowest number of CAZymes, consistent with this explanation (Hahn and Mendgen, 2001). An improved method that might be able to resolve these issues, such as through use of and expanded set of appropriate

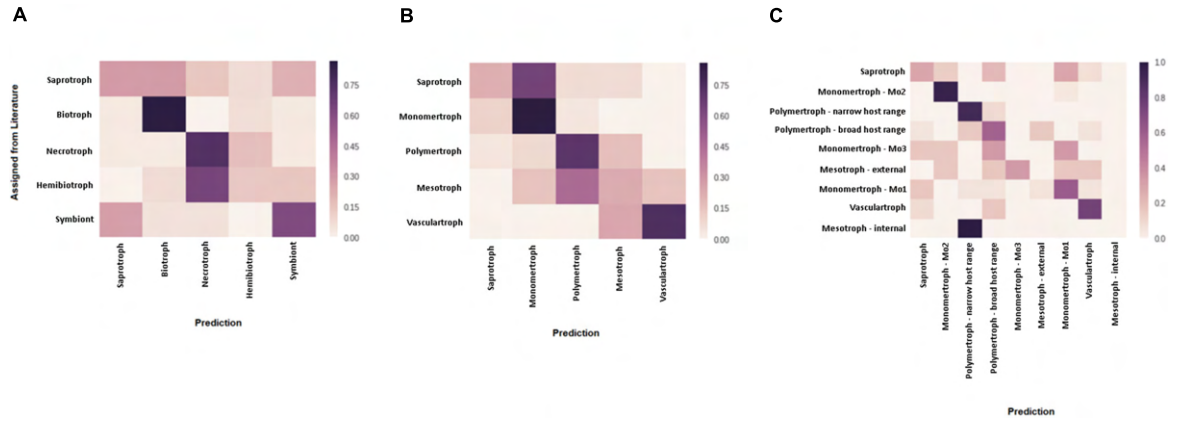


FIGURE 3 | Assessment of predicted CATAStrophy classifications of fungal trophic phenotypes. Confusion matrices were used to assess the relative suitability of each nomenclature [literature-derived terms **(A)** compared, novel major classes **(B)**, and novel sub-classes **(C)**] for use in the CATAStrophy RCD method. CATAStrophy predictions were assessed for accuracy compared to the assigned terms listed in **Supplementary Data Sheet S1**.

functional annotations, may be possible to address in a follow-up study.

Polymertrophs

Methods to classify facultative plant pathogens are less widely accepted. The term necrotroph has been applied to pathogens that cause rapid necrosis when inoculated onto hosts and whose culture filtrates also cause necrosis when applied to host tissue (Solomon et al., 2006). CATAStrophy grouped genera or species already widely accepted as necrotrophic into the P class, including: *Alternaria* spp., *Botrytis cinerea* (syn. *Botryotinia fuckeliana*), *Cochliobolus* (syn. *Bipolaris*) spp., *Pyrenophora* spp., *Parastagonospora nodorum*, *Ascochyta rabiei*, *Rhizoctonia solani*, *Gaeumannomyces graminis*, and *Sclerotinia* spp. *Fusarium graminearum* is commonly reported as a hemibiotroph, but polymertrophy is consistent with its broad host range and reliance on mycotoxins. *Verticillium* spp. were predicted as P despite initial assignment as V prior to RCD prediction (see below). *Magnaporthe oryzae* was also predicted as a P, and although commonly described as hemibiotrophic, it is capable of causing rapid necrosis. In contrast, the closely related *M. poae* was predicted as a Mo consistent with it known properties.

Broad Host-Range Polymertrophs

Botrytis cinerea quintessentially represents this sub-class. Others included *Sclerotinia* spp., *Verticillium* spp., *Aspergillus* spp., *Alternaria brassicicola*, *A. rabiei*, and *F. graminearum*. *R. solani* is divided into sexually incompatible anastomosis groups (AGs) exhibiting variable breadths in host ranges. The AG1-IA isolate (infecting rice) was predicted as PB but the AG8 isolate (infecting multiple legume and cereal species) was predicted across the S, Me(MeE), and PB classes. Both *R. solani* AG8 and *Leptosphaeria maculans* were predicted across three primary classes (S/Me/P, with affinities for MeE and PB sub-classes). Both exhibit wide host-ranges and complex and elongated life cycles that may indicate prolonged saprotrophic or biotrophic phases prior to necrotrophy.

Narrow Host-Range Polymertrophs

Broad host-range polymertrophs and PN pathogens can be distinguished by CAZyme content (Choquer et al., 2007; Andrew et al., 2012; Baroncelli et al., 2016), the former having expanded CAZyme contents ensuring activity across multiple hosts (Baroncelli et al., 2016), which may permit reduced reliance on effectors. Conversely, PN pathogens require less CAZyme diversity relative to the PB sub-class and are commonly reported to use host-specific necrotrophic effectors (Stergiopoulos and de Wit, 2009). The PN sub-class conformed well to conventional expectations, and included *Pyrenophora* spp., *P. nodorum*, *Cochliobolus* spp., and *Alternaria* spp. (except *Alt. brassicicola*) (see also **Supplementary Text S1**).

Mesotrophs

Hemibiotrophs are the most problematic traditional classification and some species described in this division were not predicted as Me. Instead our analysis grouped facultative biotrophic species that have longer latent periods than necrotrophs and do not use toxins as a primary virulence determinant into the Me class. They include most (but not all) *Colletotrichum* spp., *Venturia* spp., *Zymoseptoria* spp., *F. poae*, *Pseudocercospora fijiensis*, *F. fulva*, *L. maculans*, and *R. solani* AG8. Our analysis supported a further division into two sub-classes similar to that proposed earlier (Perfect et al., 1999) based on invasion of either intracellular or extracellular host tissues.

Extracellular (Non-appressorial) Mesotrophs

Hemibiotrophs including *L. maculans*, *Zymoseptoria* spp., and *P. fijiensis* exhibit an elongated latent phase prior to necrotrophy and were appropriately predicted with MeE affinity. *Venturia* spp. and *C. fulvum* were also predicted as MeE, in agreement with their recent re-classifications as hemibiotrophs (Stotz et al., 2014). *C. fulvum* – long regarded as a model biotroph – grows biotrophically under controlled greenhouse conditions with optimal temperature and relative humidity (de Wit, 2016), but under variable conditions or natural infection can cause noticeable necrosis.

TABLE 2 | Summary of predicted CATAStrophy classifications for selected fungal and oomycete species (full version in **Supplementary Data Sheet S1**).

Species	Strain/isolate	Phylum/sub-phylum (-mycota)	Class (-mycetes)	Common literature-based description (-troph)	Assigned sub-class for training (pre-prediction)	Saprotroph	Monomertroph	Mesotroph – inttracellular	Mesotroph – extracellular	Polymertroph – narrow host range	Polymertroph – broad host range	Vasculartroph	Predicted major class	Predicted sub-class affinities
Alternaria alternata	ATCC66891	Asco	Dothideo	Necro-	U	0.37	0.28	**0.98**	0.51	**1**	0.71	0.68	P	Mel, PN
Alternaria brassicicola	BMP1950	Asco	Dothideo	Necro-	PN	0.75	0.67	0.52	**0.97**	**1**	**0.97**	0	P	MeE, PN/PB
Ascochyta rabiei	ArDii	Asco	Dothideo	Necro-	U	0.60	0.51	0.46	0.93	0.85	**1**	0	P	PB
Cochliobolus heterostrophus	C5	Asco	Dothideo	Necro-	PN	0.37	0.31	0.73	0.51	**1**	0.65	0.37	P	PN
Cochliobolus sativus (syn. *Bipolaris sorokiniana*)	ND90Pr	Asco	Dothideo	Necro-	PN	0.41	0.35	0.64	0.57	**1**	0.68	0.22	P	PN
Dothistroma septosporum	NZE10	Asco	Dothideo	Hemibio-	MeE	0.89	0.98	0.26	**1**	0.57	0.76	0	Me	Mo, MeE
Fulvia fulva (syn *Cladosporium fulvum; Passalora fulva*)	CBS131901	Asco	Dothideo	Hemibio-	Mo	0.67	0.74	0.32	**1**	0.68	0.91	0	Me	MeE
Leptosphaeria maculans	v23.1.3	Asco	Dothideo	Hemibio-	MeE	**1**	0.93	0.47	**1**	0.68	**1**	0	S/Me/P	S, MeE, PB
Parastagonospora nodorum	SN15	Asco	Dothideo	Necro-	PN	0.41	0.32	0.59	0.56	**1**	0.67	0.24	P	PN
Pseudocercospora fijiensis (syn. *Mycosphaerella fijiensis*)	CIRAD86	Asco	Dothideo	Hemibio-	MeE	0.73	0.83	0.26	**1**	0.65	0.82	0	Me	MeE
Pyrenophora teres f. *teres*	0-1	Asco	Dothideo	Necro-	PN	0.47	0.4	0.48	0.67	**1**	0.75	0.05	P	PN
Pyrenophora tritici-repentis	Pt-1C-BFP	Asco	Dothideo	Necro-	PN	0.53	0.44	0.47	0.73	**1**	0.81	0	P	PN
Ramularia collo-cygni	DK05 Rcc001	Asco	Dothideo	Hemibio-	U	0.79	0.87	0.28	**1**	0.62	0.83	0	Me	MeE
Venturia inaequalis	20141010	Asco	Dothideo	Hemibio-	U	0.71	0.68	0.41	**1**	0.84	**1**	0	Me/P	MeE, PB
Venturia pirina	20150407	Asco	Dothideo	Hemibio-	U	0.74	0.75	0.44	**1**	0.84	**0.98**	0	Me	MeE, PB
Zymoseptoria tritici	IPO323	Asco	Dothideo	Hemibio-	MeE	0.86	0.97	0.24	**1**	0.57	0.73	0	Me	Mo, MeE
Blumeria graminis	DH14	Asco	Leotio	Bio-	Mo	0.90	**1**	0.21	0.81	0.44	0.60	0	Mo	–
Botrytis cinerea	B05	Asco	Leotio	Necro-	PB	0.61	0.5	0.40	0.75	0.52	**1**	0	P	PB
Erysiphe necator	C	Asco	Leotio	Bio-	Mo	0.90	**1**	0.22	0.81	0.45	0.60	0	Mo	–
Hymenoscyphus fraxineus (syn. *Chalara fraxinea*)	KW1	Asco	Leotio	Necro-	U	0.47	0.38	0.73	0.63	**1**	0.86	0.28	P	PN
Sclerotinia borealis	F-4128	Asco	Leotio	Necro-	PB	0.74	0.61	0.39	0.82	0.58	**1**	0	P	PB
Sclerotinia sclerotiorum	1980 UF-70	Asco	Leotio	Necro-	PB	0.91	0.82	0.39	0.94	0.59	**1**	0	P	PB
Colletotrichum gloeosporioides	Cg-14	Asco	Sordario	Hemibio-	Mel	0.27	0.19	0.90	0.37	0.65	0.55	**1**	V	–
Colletotrichum graminicola	M1.001	Asco	Sordario	Hemibio-	Mel	0.52	0.38	**1**	0.64	**1**	0.91	0.43	Me/P	Mel, PN
Colletotrichum higginsianum	IMI349063	Asco	Sordario	Hemibio-	Mel	0.39	0.3	**1**	0.52	0.75	0.75	0.54	Me	Mel
Epichloë festucae	E2368	Asco	Sordario	Symbiont	Mo	0.84	1	0.21	0.78	0.42	0.58	0	Mo	Mo
Epichloë glyceriae	E277	Asco	Sordario	Symbiont	Mo	0.84	1	0.30	0.93	0.56	0.69	0	Mo	Mo

(Continued)

TABLE 2 | Continued

Species	Strain/isolate	Phylum/sub-phylum (-mycota)	Class (-mycetes)	Common literature-based description (-troph)	Assigned sub-class for training (pre-prediction)	Saprotroph	Monomertroph	Mesotroph – intracellular	Mesotroph – extracellular	Polymertroph – narrow host range	Polymertroph – broad host range	Vasculartroph	Predicted major class	Predicted sub-class affinities
Fusarium graminearum	PH-1	Asco	Sordario	Hemibio-	MeE	0.58	0.5	**0.97**	0.81	**0.98**	**1**	0.77	P	Mel, PN/PB
Fusarium oxysporum f. sp. *lycopersici*	4287	Asco	Sordario	Wilt	V	0.24	0.19	0.69	0.33	0.51	0.46	**1**	V	–
Fusarium solani	mpVI	Asco	Sordario	Wilt	V	0.24	0.18	0.69	0.33	0.50	0.45	**1**	V	–
Gaeumannomyces graminis	R3-111a-1	Asco	Sordario	Root	PB	0.65	0.56	0.74	0.75	**1**	0.88	0.04	P	PN
Magnaporthe oryzae	70-15	Asco	Sordario	Hemibio-	Mel	0.56	0.48	0.73	0.71	**1**	0.83	0.15	P	PN
Magnaporthe poae	ATCC64411	Asco	Sordario	Root	Mo	**0.99**	1	0.44	**0.95**	0.71	0.82	0	Mo	S, MeE
Verticillium albo-atrum	VaMs.102	Asco	Sordario	Necro-	V	0.84	0.72	0.58	0.94	0.77	**1**	0	P	PB
Verticillium dahliae	VdSo316	Asco	Sordario	Hemibio-	V	0.67	0.48	0.70	0.79	0.85	**1**	0	P	PB
Rhizoctonia solani	AG1-IA	Basidio	Agarico	Necro-	PB	0.57	0.38	0.71	0.67	0.65	**1**	0.06	P	PB
Rhizoctonia solani	AG8 WAC10335	Basidio	Agarico	Necro-	PB	**1**	0.93	0.47	**1**	0.68	**1**	0	S/Me/P	S, MeE, PB
Melampsora laricis-populina	98AG31	Basidio	Puccinio	Bio-	Mo	**0.98**	**1**	0.33	**0.99**	0.59	0.79	0	Mo	S, MeE
Puccinia graminis	UG99	Basidio	Puccinio	Bio-	Mo	**0.97**	**1**	0.32	**0.97**	0.59	0.78	0	Mo	S, MeE
Puccinia striiformis	PST-130	Basidio	Puccinio	Bio-	Mo	0.92	**1**	0.27	0.86	0.51	0.67	0	Mo	–
Ustilago hordei	Uh4857_4	Basidio	Ustilagino	Bio-	Mo	0.90	**1**	0.21	0.82	0.43	0.61	0	Mo	–
Ustilago maydis	521	Basidio	Ustilagino	Bio-	Mo	0.75	**1**	0.22	0.78	0.45	0.64	0	Mo	–
Albugo candida	ASM107853v1	Oo	Oo	Bio-	Mo	0.74	**1**	0.19	0.66	0.37	0.51	0	Mo	–
Albugo laibachiic	ENA1	Oo	Oo	Bio-	Mo	0.84	**1**	0.23	0.72	0.42	0.56	0	Mo	–
Hyaloperonospora arabidopsidis	Emoy2	Oo	Oo	Bio-	Mo	0.71	**1**	0.18	0.64	0.36	0.49	0	Mo	–
Phytophthora ramorum	CDFA1418886	Oo	Oo	Bio-	Mo	0.58	**1**	0.21	0.60	0.33	0.51	0	Mo	–
Phytophthora sojae	P6497	Oo	Oo	Bio-	Mo	0.51	**1**	0.24	0.58	0.31	0.53	0	Mo	–

Relative centroid distance (RCD) scores from 0 to 1 are presented for each of the nine trophic sub-classes. An RCD value of 1 (bold and underlined) indicates membership in a major trophic class and a value ≥0.95 (bold) predicts affinity for one or more trophic sub-classes. Predicted trophic class and sub-classes are summarized in the right-hand columns. S, saprotroph; Mo, monomertroph; Me, mesotroph; Mel, mesotroph – intracellular; MeE, mesotroph – extracellular; P, polymertroph; PB, polymertroph – broad host range; PN, polymertroph – narrow host range; V, vasculartroph; U, unclassified (not included in training).

Intracellular (Appressorial) Mesotrophs

The MeI sub-class was initially assigned to species possessing appressoria-like feeding structures formed on the host surface prior to host penetration, exemplified by the *Colletotrichum* spp. Almost all *Colletotrichum* spp. were predicted as MeI, with the exception of *C. gloeosporioides* (V). Other appressorial species including *M. oryzae*, *G. graminis*, and *Alternaria* spp. were predicted instead as P (excepting *A. longipes*, MeI). *F. poae* and *F. graminearum* (P) were predicted with MeI affinity, which is supported in the latter by reports of mycotoxin-producing appressorium-like structures. While this class was initially assigned to appressorial hemibiotroph species prior to RCD prediction, the MeI sub-class appears not to be strictly linked to the presence of appressoria but still correlates to intracellular host interactions. This mirrors how reports of appressoria do not align consistently with the intracellular hemibiotrophic phenotype.

Vasculartrophs

We propose a novel V class which contains pathogens that are associated with wilt, anthracnose, and rot symptoms and grouped separately from the Mo, Me, or P classes. Several "wilt-like" species are not well-defined in terms of their mode of nutrition, but our analysis suggests that V are most similar in CAZyme content to the PB sub-class. This V class was initially assigned to the *Fusarium* spp. (excluding *F. graminearum*) and *Verticillium* spp. prior to RCD prediction. In final trophic predictions (Step 3) however, *Verticillium* spp., *F. poae*, and *F. lansethiae* were not predicted in this class. *Verticillium* spp. and *Fusarium* spp., despite both being commonly referred to as "wilts," do exhibit several differences including: host-range (*Verticillium* is broader), climate preference (*Verticillium* prefers cooler temperatures), and severity with less vascular browning and no cell death in *Verticillium* but more browning and necrosis in *Fusarium* wilt on tomato. Thus, the prediction of *Verticillium* outside this group (PB) may be due to genuine biological features that need to be further investigated. Although the *Colletotrichum* spp. are predominantly predicted as mesotrophic, *C. gloeosporioides*, *C. simmondsii*, and *C. nymphaeae* were predicted as primarily vasculartrophic.

CONCLUSION

The long history of the biotroph–hemibiotroph–necrotroph classification of plant pathogens (de Bary and Garnsey, 1887) is evidenced by its persistence in major textbooks and reviews (Horbach et al., 2011). Despite its ubiquity, the tripartite classification has long been regarded as problematic (Oliver and Ip-Cho, 2004; Glazebrook, 2005; Kuo et al., 2014; Stotz et al., 2014; Sánchez-Vallet et al., 2015; Videira et al., 2017). Increased availability of genomic data has allowed us to re-examine the suitability of this nomenclature. The CATAStrophy method allows for the prediction of trophic classes based solely on CAZYme gene content. In place of the three major classes of pathogen, we propose four novel pathogen classes: Mo, P, Me, and V.

Carbohydrate-active enzyme-Assisted Training And Sorting of -trophy focuses attention on the properties linking and separating these groups and provides a basis for a reproducible, objective, and unbiased classification of fungal trophic phenotypes. Current trends in whole-genome sequencing techniques and costs have led to a rapid increase in the number of fungal species sequenced. Correspondingly, the species studied by these techniques have rapidly spread from a few species with historically high economic and scientific relevance to species with local or recent impact. A good example is ash-dieback and Ramularia leaf spot (Saunders et al., 2014; Stam et al., 2018). There are clear differences in the strategies adopted to combat haustorial biotrophic and narrow-host range necrotrophic plant pathogens (Oliver, 2009; Burdon et al., 2014). Thus, the economic and societal impact of a rapid assessment of the causal organism of a novel disease could be significant. As microbial genomics data grow in volume, we anticipate an emerging need for bioinformatic techniques such as CATAStrophy that can predict agriculturally relevant phenotypes from genomic data, particularly as only a minor fraction of plant pathogenic fungi have been studied in detail. The CATAStrophy method suggests a novel and more detailed grouping of pathogens which we hope will stimulate the development and testing of hypotheses relating to pathogenicity, virulence, and control measures.

MATERIALS AND METHODS

Prediction of Carbohydrate-Active Enzyme Contents

Whole proteome (i.e., predicted gene translations) sequences were obtained in FASTA format as per **Supplementary Data Sheet S1** and **Supplementary Text S1**. The CAZyme (Cantarel et al., 2009; Lombard et al., 2014) functional annotations were utilized to represent *a priori* evidence reporting the "trophic type." CAZyme classes were annotated for all species via HMMER 3.0 (as per dbCAN recommendations, i.e., hmmscan with the -domtblout parameter, then dbcan hmmscan-parser.sh with 80 aa minimum alignment length, *e*-value $< 1e{-}5$ and $>30\%$ coverage of HMM) (Eddy, 2010) and the dbCAN (version 6) set of CAZyme HMMs (Yin et al., 2012), listed in full in **Supplementary Data Sheet S2**.

Organization of Reported Trophic Phenotypes Into Discrete Classes

We tested three discrete nomenclatures that describe the trophic phenotype. The first trophic nomenclature was assigned to species based on the terms – S, symbiont, biotroph, hemibiotroph, and necrotroph – commonly reported in published literature (**Table 2** and **Figure 3A**). The second nomenclature uses five major divisions (S, Mo, Me, P, and V) (**Figure 3B**). Nomenclature 3 uses the five major divisions (S, Mo, Me, P, and V), and included sub-divisions for MeI, MeE, PN, PB and three sub-divisions of the Mo (symbionts, haustorial, and non-haustorial; **Figure 1** top panel) that were later obsoleted (**Table 2** and **Figure 3C**). Due to difficulties in resolving the

sub-classes within the Mo, we assigned them numerical labels (Mo1, Mo2, and Mo3, respectively) where they appear in **Supplementary Material** (**Supplementary Data Sheet S1**), but for the purpose of summarizing CATAStrophy predictions have merged them into a single Mo class (Mo). Importantly, all three nomenclatures were initially based on reports derived from peer-reviewed literature (**Supplementary Data Sheet S1**). The three nomenclatures were tested for their relative efficacy (**Figure 3**) and nomenclature 2 (S, Mo, Me, P, V) is the primary one used for subsequent analyses presented in this study.

Prediction of Trophic Classes via Multivariate Analysis

The number of genes in each species assigned to each CAZyme class was used in PCA using singular value decomposition via scikit-learn v 0.18.1 (Pedregosa et al., 2011) to cluster species (**Figure 2**). Species were each assigned a trophic class based on the most commonly used term derived from literature reports, or the equivalent term from our novel proposed nomenclatures. Centroids corresponding to each trophic sub-class were calculated based on the positions in PCA space of the species assigned that class (**Supplementary Data Sheet S1**). Each species was then unassigned from its designated trophic class, its position in PCA space relative to centroids was calculated, and a RCD score was calculated for each species to assess the relative likelihood of its membership in each class. Centroids were re-calculated for the assessment of each species during RCD analysis, with the species currently being assessed being removed from centroid calculations so as not to influence the prediction. The centroid closest to a species in PCA space was assigned an RCD score of 1, with other centroids expressed as a normalized proportion of the closest centroid distance. RCD scores were rounded to two decimal places. Using data based on initial manual assignment of the novel classes and sub-classes (**Table 2** and **Supplementary Data Sheet S1**), species were predicted to belong to broad classes (**Table 2**) where RCD = 1 with high confidence, and also assigned additional "affinities" for sub-classes (**Table 2**) if RCD $\geq$ 0.95 at a lower confidence. RCD scores for the biotroph sub-divisions Mo1, Mo2, and Mo3 are reported individually in **Supplementary Data Sheet S1**, but only the maximum of these scores is reported for the Mo class in **Table 2**. Using this method it is possible for trophic classes to be revised, i.e., a species may be predicted in a different class than it was originally assigned to prior to RCD calculation. In order to demonstrate the efficacy of the newly proposed trophic nomenclatures for the CATAStrophy RCD method, each of the three nomenclatures (literature-derived, novel major classes, and novel sub-classes) was tested separately via the CATAStrophy method and the predictions were assessed using confusion matrices that the predictions to assigned terms (**Figure 3**). The PCA plot and principle component coordinates for each species included in the initial CATAStrophy analysis (i.e., not unassigned in **Supplementary Data Sheet S1**) are provided in **Supplementary Data Sheet S3**.

DATA AVAILABILITY STATEMENT

All datasets generated for this study are included in the article/**Supplementary Material**. CATAStrophy software available at https://github.com/ccdmb/catastrophy.

AUTHOR CONTRIBUTIONS

JH and DJ performed the bioinformatics analysis. DJ and JP performed the multivariate analysis. JH and RO wrote the manuscript. JH, RO, and PW edited the manuscript. All authors read and approved the manuscript.

ACKNOWLEDGMENTS

Thanks to Alison Testa for compilation of sequences and metadata prior to this study. This study was initiated as part of a Royal Dutch Academy of Sciences (KNAW) visiting professorship to R. Oliver to the Laboratory of Phytopathology of Wageningen University and benefitted from resources provided at the NCI National Facility systems and Pawsey Supercomputing Centre through State and National Computational Merit Allocation Schemes supported by the Australian Government.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2019.03088/full#supplementary-material

DATA SHEET S1 | Full list of fungal and oomycete species and isolates used in this study, their assigned trophic classifications based on literature consensus, and their predicted CATAStrophy classifications.

DATA SHEET S2 | CAZyme annotations assigned to species via dbCAN.

DATA SHEET S3 | Plot and PCA coordinates for the first 16 principle components derived from CAZyme gene contents for 110 fungal and oomycete species and isolates used to build the CATAStrophy RCD predictive method.

TEXT S1 | References for species and genome resources cited in this study, and additional discussion of CATAStrophy predictions for selected species.

REFERENCES

Andrew, M., Barua, R., Short, S. M., and Kohn, L. M. (2012). Evidence for a common toolbox based on necrotrophy in a fungal lineage spanning necrotrophs, biotrophs, endophytes, host generalists and specialists. *PLoS One* 7:e29943. doi: 10.1371/journal.pone.0029943

Baroncelli, R., Amby, D. B., Zapparata, A., Sarrocco, S., Vannacci, G., Le Floch, G., et al. (2016). Gene family expansions and contractions are associated with host range in plant pathogens of the genus *Colletotrichum*. *BMC Genomics* 17:555. doi: 10.1186/s12864-016-2917-6

Bebber, D. P., and Gurr, S. J. (2015). Crop-destroying fungal and oomycete pathogens challenge food security. *Fungal Genet. Biol.* 74, 62–64. doi: 10.1016/j.fgb.2014.10.012

Burdon, J. J., Barrett, L. G., Rebetzke, G., and Thrall, P. H. (2014). Guiding deployment of resistance in cereals using evolutionary principles. *Evol. Appl.* 7, 609–624. doi: 10.1111/eva.12175

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The carbohydrate-active enZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res.* 37, D233–D238. doi: 10.1093/nar/gkn663

Choquer, M., Fournier, E., Kunz, C., Levis, C., Pradier, J. M., Simon, A., et al. (2007). *Botrytis cinerea* virulence factors: new insights into a necrotrophic and polyphageous pathogen. *FEMS Microbiol. Lett.* 277, 1–10. doi: 10.1111/j.1574-6968.2007.00930.x

de Bary, A., and Garnsey, H. E. F. (1887). *Comparative Morphology and Biology of the Fungi, Mycetozoa and Bacteria*. Oxford: Clarendon Press.

de Wit, P. J. (2016). *Cladosporium fulvum* effectors: weapons in the arms race with tomato. *Annu. Rev. Phytopathol.* 54, 1–23. doi: 10.1146/annurev-phyto-011516-040249

Eddy, S. (2010). *HMMER3: A New Generation of Sequence Homology Search Software*. Available at: http://hmmer.janelia.org (accessed January 2019).

Fisher, M. C., Hawkins, N. J., Sanglard, D., and Gurr, S. J. (2018). Worldwide emergence of resistance to antifungal drugs challenges human health and food security. *Science* 360, 739–742. doi: 10.1126/science.aap7999

Fisher, M. C., Henk, D. A., Briggs, C. J., Brownstein, J. S., Madoff, L. C., McCraw, S. L., et al. (2012). Emerging fungal threats to animal, plant and ecosystem health. *Nature* 484, 186–194. doi: 10.1038/nature10947

Gay, J. L. (1984). *Plant, Disease: Infection Damage and Loss, Blackwell*, eds R. Wood, and G. J. Jellis. Oxford: Blackwell Scientific Publications.

Glazebrook, J. (2005). Contrasting mechanisms of defense against biotrophic and necrotrophic pathogens. *Annu. Rev. Phytopathol.* 43, 205–227. doi: 10.1146/annurev.phyto.43.040204.135923

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., et al. (1996). Life with 6000 genes. *Science* 274, 563–567.

Gurr, G. M., Zhu, Z.-R., and You, M. (2015). "The big picture: prospects for ecological engineering to guide the delivery of ecosystem services in global agriculture," in *Rice Planthoppers* eds K. L. Heong, J. Cheng, and M. M. Escalada (Dordrecht: Springer), 143–160.

Hahn, M., and Mendgen, K. (1997). Characterization of in planta-induced rust genes isolated from a haustorium-specific cDNA library. *Mol. Plant Microbe Interact.* 10, 427–437. doi: 10.1094/mpmi.1997.10.4.427

Hahn, M., and Mendgen, K. (2001). Signal and nutrient exchange at biotrophic plant–fungus interfaces. *Curr. Opin. Plant Biol.* 4, 322–327. doi: 10.1073/pnas.1308973110

Hammond-Kosack, K. E., and Parker, J. E. (2003). Deciphering plant–pathogen communication: fresh perspectives for molecular resistance breeding. *Curr. Opin. Biotechnol.* 14, 177–193. doi: 10.1016/s0958-1669(03)00035-1

Horbach, R., Navarro, A. R., Quesada Knogge, and Deising, H. B. (2011). When and how to kill a plant cell: infection strategies of plant pathogenic fungi. *J. Plant Physiol.* 168, 51–62. doi: 10.1016/j.jplph.2010.06.014

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. doi: 10.1371/journal.ppat.1002137

Koeck, M., Hardham, A. R., and Dodds, P. N. (2011). The role of effectors of biotrophic and hemibiotrophic fungi in infection. *Cell. Microbiol.* 13, 1849–1857. doi: 10.1111/j.1462-5822.2011.01665.x

Kuo, H. C., Hui, S., Choi, J., Asiegbu, F. O., Valkonen, J. P., and Lee, Y. H. (2014). Secret lifestyles of *Neurospora crassa*. *Sci. Rep.* 4:5135. doi: 10.1038/srep05135

Laluk, K., and Mengiste, T. (2010). Necrotroph attacks on plants: wanton destruction or covert extortion? *Arabidopsis Book* 8:e0136. doi: 10.1199/tab.0136

Lewis, D. (1973). Concepts in fungal nutrition and the origin of biotrophy. *Biol. Rev.* 48, 261–277. doi: 10.1111/j.1469-185x.1973.tb00982.x

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M., and Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res.* 42, D490–D495. doi: 10.1093/nar/gkt1178

Lucas, J. A. (1998). *Plant Pathology and Plant Pathogens*. Oxford: Blackwell.

Mendgen, K., Struck, C., Voegele, R. T., and Hahn, M. (2000). Biotrophy and rust haustoria. *Physiol. Mol. Plant Pathol.* 56, 141–145. doi: 10.1006/pmpp.2000.0264

Oliver, R. (2009). Plant breeding for disease resistance in the age of effectors. *Phytoparasitica* 37, 1–5. doi: 10.1007/s12600-008-0013-4

Oliver, R. P., and Ip-Cho, S. V. (2004). *Arabidopsis* pathology breathes new life into the necrotrophs—vs.—biotrophs classification of fungal pathogens. Mol. Plant Pathol. 5, 347–352. doi: 10.1111/j.1364-3703.2004.00228.x

Parbery, D. G. (1996). Trophism and the ecology of fungi associated with plants. *Biol. Rev.* 71, 473–527. doi: 10.1111/j.1469-185x.1996.tb01282.x

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Pedro, H., Maheswari, U., Urban, M., Irvine, A. G., Cuzick, A., McDowall, M. D., et al. (2016). PhytoPath: an integrative resource for plant pathogen genomics. *Nucleic Acids Res.* 44, D688–D693. doi: 10.1093/nar/gkv1052

Perfect, S. E., Hughes, H. B., O'Connell, R. J., and Green, J. R. (1999). *Colletotrichum*: a model genus for studies on pathology and fungal–plant interactions. *Fungal Genet. Biol.* 27, 186–198. doi: 10.1006/fgbi.1999.1143

Poland, J. A., Balint-Kurti, P. J., Wisser, R. J., Pratt, R. C., and Nelson, R. J. (2009). Shades of gray: the world of quantitative disease resistance. *Trends Plant Sci.* 14, 21–29. doi: 10.1016/j.tplants.2008.10.006

Richards, T. A., and Talbot, N. J. (2013). Horizontal gene transfer in osmotrophs: playing with public goods. *Nat. Rev. Microbiol.* 11, 720–727. doi: 10.1038/nrmicro3108

Rodrigues, F., Ludovico, P., and Leão, C. (2006). *Sugar Metabolism in Yeasts: An Overview of Aerobic and Anaerobic Glucose Catabolism, Biodiversity and Ecophysiology of Yeasts*. Berlin: Springer, 101–121.

Sánchez-Vallet, A., McDonald, M. C., Solomon, P. S., and McDonald, B. A. (2015). Is *Zymoseptoria tritici* a hemibiotroph? *Fungal Genet. Biol.* 79, 29–32. doi: 10.1016/j.fgb.2015.04.001

Saunders, D., Yoshida, K., Sambles, C., Glover, R., Clavijo, B., Corpas, M., et al. (2014). Crowdsourced analysis of ash and ash dieback through the open ash dieback project: a year 1 report on datasets and analyses contributed by a self-organising community. *bioRxiv* [Preprint]. doi: 10.1101/004564

Savary, S., Willocquet, L., Pethybridge, S. J., Esker, P., McRoberts, N., and Nelson, A. (2019). The global burden of pathogens and pests on major food crops. *Nat. Ecol. Evol.* 1, 430–439. doi: 10.1038/s41559-018-0793-y

Scott, K. J. (1972). Obligate parasitism by phytopathogenic fungi. *Biol. Rev.* 47, 537–572. doi: 10.1111/j.1469-185x.1972.tb01081.x

Soanes, D. M., Alam, I., Cornell, M., Wong, H. M., Hedeler, C., Paton, N. W., et al. (2008). Comparative genome analysis of filamentous fungi reveals gene family expansions associated with fungal pathogenesis. *PLoS One* 3:e2300. doi: 10.1371/journal.pone.0002300

Solomon, P. S., Wilson, T. J. G., Rybak, K., Parker, K., Lowe, R. G. T., and Oliver, R. P. (2006). Structural characterisation of the interaction between *Triticum aestivum* and the dothideomycete pathogen *Stagonospora nodorum*. *Eur. J. Plant Pathol.* 114, 275–282. doi: 10.1007/s10658-005-5768-6

Stam, R., Munsterkotter, M., Pophaly, S. D., Fokkens, L., Sghyer, H., Guldener, U., et al. (2018). A new reference genome shows the one-speed genome structure of the barley pathogen *Ramularia collo-cygni*. *Genome Biol. Evol.* 10, 3243–3249. doi: 10.1093/gbe/evy240

Staples, R. C. (2001). Nutrients for a rust fungus: the role of haustoria. *Trends Plant Sci.* 6, 496–498. doi: 10.1016/s1360-1385(01)02126-4

Stergiopoulos, I., and de Wit, P. J. (2009). Fungal effector proteins. *Annu. Rev. Phytopathol.* 47, 233–263. doi: 10.1146/annurev.phyto.112408.132637

Stotz, H. U., Mitrousia, G. K. P. J., de, G. M., Wit, and Fitt, B. D. (2014). Effector-triggered defence against apoplastic fungal pathogens. *Trends Plant Sci.* 19, 491–500. doi: 10.1016/j.tplants.2014.04.009

Tan, K. C., Oliver, R. P., Solomon, P. S., and Moffat, C. S. (2010). Proteinaceous necrotrophic effectors in fungal virulence. *Funct. Plant Biol.* 37, 907–912.

Tanaka, S., Han, X., and Kahmann, R. (2015). Microbial effectors target multiple steps in the salicylic acid production and signaling pathway. *Front. Plant Sci.* 6:349. doi: 10.3389/fpls.2015.00349

Thrower, L. B. (1966). Terminology for plant parasites. *J. Phytopathol.* 56, 258–259. doi: 10.1111/j.1439-0434.1966.tb02261.x

Videira, S., Groenewald, J., Nakashima, C., Braun, U., Barreto, R. W., de Wit, P. J., et al. (2017). Mycosphaerellaceae–chaos or clarity? *Stud. Mycol.* 87, 257–421. doi: 10.1016/j.simyco.2017.09.003

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*. *Proc. Natl. Acad. Sci. U.S.A.* 98, 8133–8138. doi: 10.1073/pnas.131186798

Wang, X., Jiang, N., Liu, J., Liu, W., and Wang, G.-L. (2014). The role of effectors and host immunity in plant–necrotrophic fungal interactions. *Virulence* 5, 722–732. doi: 10.4161/viru.29798

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., and Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* 40, W445–W451. doi: 10.1093/nar/gks479

Zeilinger, S., Gupta, V. K., Dahms, T. E., Silva, R. N., Singh, H. B., Upadhyay, R. S., et al. (2016). Friends or foes? *FEMS Microbiol. Rev.* 40, 182–207. doi: 10.1093/femsre/fuv045

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

CHAPTER 6 — THEME 1

Hierarchical clustering of MS/MS spectra from the firefly metabolome identifies new lucibufagin compounds

This chapter is also published in:
Scientific Reports, 2020, vol. 10, article 6043
https://doi.org/10.1038/s41598-020-63036-1

This chapter is submitted as supplementary material and should not contribute to assessment of this thesis. It is included here as an example of related research contributions made during the candidacy.

6.1 Declaration

Title	Hierarchical clustering of MS/MS spectra from the firefly metabolome identifies new lucibufagin compounds.
Authors	Catherine Rawlinson, **Darcy A. B. Jones**, Suman Rakshit, Shiv Meka, Caroline S. Moffat and Paula Moolhuijzen
Publication	2020. *Scientific Reports*, 10(1), 6043.
DOI	`https://doi.org/10.1038/s41598-020-63036-1`

This supplementary chapter has been submitted as part of another student's thesis (Catherine Rawlinson). As such, it is included here only as an example of contributions to related work conducted during the candidate's Ph. D. and should not contribute to assessment. This thesis chapter is submitted in the form of a collaboratively-written peer-reviewed manuscript.

The following contributions were made by the Ph.D candidate (DABJ) and co-authors:

- CR, PM and CSM conceived and designed the study.
- **DABJ**, SM, SR, and PM contributed to BioDendro code development.
- CR analysed the data.
- CR and PM wrote the manuscript.
- All authors read, edited, and approved the manuscript.

I, Darcy Jones, hereby certify that this attribution statement is an accurate record of my contribution to the research presented in this chapter.

Darcy Jones

I certify that this attribution statement is an accurate record of Darcy Jones' contribution to the research presented in this chapter.

Catherine Rawlinson | Suman Rakshit | Shiv Meka

Caroline S. Moffat | Paula Moolhuijzen

SCIENTIFIC REPORTS
nature research

OPEN

Hierarchical clustering of MS/MS spectra from the firefly metabolome identifies new lucibufagin compounds

Catherine Rawlinson[1*], Darcy Jones[1], Suman Rakshit[2], Shiv Meka[3], Caroline S. Moffat[1] & Paula Moolhuijzen[1*]

Metabolite identification is the greatest challenge when analysing metabolomics data, as only a small proportion of metabolite reference standards exist. Clustering MS/MS spectra is a common method to identify similar compounds, however interrogation of underlying signature fragmentation patterns within clusters can be problematic. Previously published high-resolution LC-MS/MS data from the bioluminescent beetle (*Photinus pyralis*) provided an opportunity to mine new specialized metabolites in the lucibufagin class, compounds important for defense against predation. We aimed to 1) provide a workflow for hierarchically clustering MS/MS spectra for metabolomics data enabling users to cluster, visualise and easily interrogate the identification of underlying cluster ion profiles, and 2) use the workflow to identify key fragmentation patterns for lucibufagins in the hemolymph of *P. pyralis*. Features were aligned to their respective MS/MS spectra, then product ions were dynamically binned and resulting spectra were hierarchically clustered and grouped based on a cutoff distance threshold. Using the simplified visualization and the interrogation of cluster ion tables the number of lucibufagins was expanded from 17 to a total of 29.

Metabolomics is the scientific study of the low molecular weight compounds (metabolites) within an organism, cell or tissue, which reflect underlying biochemical activities and cellular processes. A major challenge of metabolomic analysis is the identification of these compounds. If the reference MS/MS spectrum for a metabolite is not publicly or commercially available, identification is unlikely. However, compounds of similar structure often have similar MS fragmentation pathways leading to mass spectral patterns specific to a chemical class.

As an example, triticones are a class of specialized metabolites produced by the necrotrophic fungal pathogen, *Pyrenophora tritici-repentis* (Ptr), which have recently been functionally characterised[1]. Since triticones were first purified in 1988, several have been purified and characterized via NMR analyses. However, it is only recently that their MS/MS spectra have been explored which enabled the putative identification of a total of 38 triticones in the LC-MS/MS profile of Ptr. It is important to understand the complement of structurally similar bioactive molecules produced by an organism as activity across a class of compounds may offer customized responses to biological stressors.

With the introduction of acquisition techniques collecting MS/MS spectra with no prior knowledge of sample composition, thousands of unique spectra can be generated from a single sample. With substantially higher MS/MS coverage of features, information about chemical structure can be leveraged from repeated mass spectral patterns, aiding in classification of unknown metabolites. However, thorough exploration of MS/MS data generated with these new techniques is often cumbersome and impractical. Even after statistical analyses has derived a list of analytes of interest, correlating their mass spectra to other analytes can be laborious.

Several tools have been developed to assist with MS/MS pattern recognition. Molecular networking-based visualization is becoming increasingly popular in metabolomics and is used by tools such as Global Natural

[1]Centre for Crop and Disease Management, School of Molecular and Life Sciences, Curtin University, Bentley, Western Australia, Australia. [2]Statistics for the Australian Grains Industry-West, School of Molecular and Life Sciences, Curtin University, Bentley, Western Australia, Australia. [3]Curtin Institute for Computation, Curtin University, Bentley, Western Australia, Australia. *email: catherine.rawlinson@postgrad.curtin.edu.au; paula.moolhuijzen@curtin.edu.au

Products Social Molecular Networking (GNPS)[2–4]. Whilst use of such tools is becoming more prevalent, GNPS is web-based requiring upload of data to a server and is limited in parameter customization of workflow and little in exportable, easy to interrogate results. 'MetCirc" is an R based package[5] offering clustering of MS/MS spectra and uses a Circos plot for visualization[6]. However, because MetCirc uses a defined number of identically sized "fixed" bins, instrumental variability and precision may lead to incorrectly binned ions causing false or incomplete conclusions[7–9]. This motivated us to create a workflow that was easy to use and to interrogate fragmentation profiles using dynamic binning, which allows instrument resolution and precision to inform binning, and hierarchically clustering of MS/MS spectra.

In this study, we demonstrate the effectiveness of hierarchically clustering MS/MS spectra for the discovery of underlying ion profiles to identify or classify unknown metabolites. A previously published dataset of firefly predator defense lucibufagin compounds[10] was reanalysed using dynamic binning and hierarchical clustering, and results were compared to the gold standard web-based Feature Based Molecular Networking (FBMN) module of GNPS. We provide the techniques used as a simple but effective workflow, BioDendro (https://github.com/ccdmb/BioDendro), that users with minimal coding skills can easily use and customize to identify core fragmentation patterns.

Material and Methods

LC-MS/MS Firefly Metabolights data source. LC-MS/MS data was sourced from the MetaboLights repository, project ID MTBLS698 (https://www.ebi.ac.uk/metabolights/MTBLS698)[10] for the analysis of luminescent and non-luminescent tissue of beetle species[10]. Data were collected on a Thermo Q-Exactive Orbitrap using data dependent acquisition (DDA) with polarity switching using a C18 column[10]. A single file generated from the hemolymph of an adult male *Photinus Pyralis* beetle was selected (Ppyr_hemolymph_extract.mzML). The positive ion mode analysis was previously carried out using MZmine (v2.30) with MS^2 similarity search and published using the parameters described in section 4.6 of the supplementary information[10]. The positive ion data for Ppyr_hemolymph_extract.mzML was reanalyzed here with MZmine2 (v2.53) using the same settings. Parameters which have changed or added between versions were applied as was suitable for the data (Supplementary Table S1). To identify new lucibufagins, hierarchical clustering of MS/MS spectra with subsequent visualization and interrogation was applied using a newly created workflow application, BioDendro. Results from BioDendro were then compared to molecular networking of MS/MS spectra using FBMN[4] module of GNPS[2].

Firefly LC-MS/MS analysis using BioDendro workflow. BioDendro released under an Apache 2.0 license is available for download at https://github.com/ccdmb/BioDendro. BioDendro requires the use of Python3 and is run locally through the provided Jupyter Notebook (https://jupyter.org/) to execute the programs workflow. Both applications can be downloaded and installed through the package management tool, Anaconda (https://www.anaconda.com/distribution/). Detailed instructions on download, installation and usage of BioDendro can be found in GitHub (https://github.com/ccdmb/BioDendro). A detailed explanation of parameters and recommended settings can be found in Supplementary Information SI1. Two Jupyter notebooks have been supplied; "quick-start-example.ipynb" which contains all the settings applied herein and "longer-workflow.ipynb" which provides a set by step execution.

BioDendro requires two input text files to function; a file containing all the features within a data set (.txt) and MS/MS spectra in MGF format (.mgf) (Supplementary Information SI1). A single MS/MS spectrum was aligned to a feature based on a mass (m/z 0.005) and retention time (6 secs) user defined tolerances, where multiple matches exist, the closest in retention time was associated using pandas core package[11]. Two optional steps can be applied at this stage; an absolute/relative filtering of ions and application of neutral loss formatting to spectra. An absolute filtering of ions (minimum intensity of 5000) and no neutral loss was performed. Prior to comparison of spectra, all masses were binned to allow appropriate comparison of spectra using variable bin sizes and the numpy core package[12]. All product ions were ordered by m/z and a new bin was created when the difference between 2 consecutive masses exceeded a user defined threshold (defined here as m/z 0.0005). This value should reflect instrument precision. Pairwise distances were then calculated between all binned spectra using the Bray-Curtis metric implemented in scipy (Jaccard distance is also available; see S2 for description of these metrics)[13]. The distance matrix then hierarchically clustered using complete-linkage clustering implemented in scipy[13]. A user specifiable distance threshold can be used to select clusters from the hierarchically clustered data. Lastly, data were visualized as a tree using plotly[14] and the user defined distance threshold was set to 0.7. Resulting clusters were output as ion histograms using matplotlib[15] and in tabular format that represented clustered features and the associated MS/MS spectra. See Supplementary Table S2 for a summary of analysis parameters.

Firefly LC-MS/MS analysis using FBMN module of GNPS. Data were extracted from Ppyr_hemolymph_extract.mzML for analysis in the FBMN module of GNPS analysis platform[2,4,16]. Documentation for analysis using FBMN with MZmine2 can be found at https://ccms-ucsd.github.io/GNPSDocumentation/featurebasedmolecularnetworking-with-mzmine2/. Two text files are required for analysis with FBMN, a.txt file containing the sample features and the aligned MS/MS spectra in MGF. The files were generated using MZmine2 as per the parameters described by Fallon, *et al.*[10]. Molecular networking was carried out using BioDendro settings where applicable (Supplementary Table S2).

Comparison of BioDendro and FBMN. Comparisions of BioDendro and FBMN used the same feature list generated by MZmine2 as per the analysis settings in Supplementary Table S2. BioDendro used an MGF file produced from the freeware ProteoWizard[17], which exports all MS/MS spectra collected. For FBMN, an MGF of aligned MS/MS spectra were exported from MZmine2 as required.

Comparative analysis of clustering between BioDendro and FBMN was accomplished in a targeted manner by comparing the clustering of the putatively identified lucibufagin class of compounds. This involved a manual search using retention time and precursor mass to locate each feature within the respective pipelines.

Analyses using BioDendro and FBMN were carried out using a Windows 7 64-bit PC. The PC had an Intel i7 processor and 16GB of RAM.

Metabolite classification and identification. Putative identification of lucibufagins by Fallon, *et al.*[10] was by targeted search of masses in the LC-MS profile of known lucibufagin compounds and then expanded upon by MS^2 similarity searching in MZmine2. The putatively identified lucibufagins and their respective MS/MS spectra was employed herein. METLIN[18], MassBank[19] and NIST14[20] mass spectral databases were searched for reference spectra. Literature was also searched for mass spectral information pertaining to "lucibufagin MS/MS".

The molecular formula for fragment ions were predicted using the 'Elemental composition" function within the Qual Browser module of Thermo Xcalibur software. Details for prediction are outlined in Supplementary Information SI2. Fallon, *et al.*[10] reported greatest instrumental error as +9.9 ppm for tryptophan (*m/z* 205.09) and therefore a ±10 ppm tolerance was used. Elemental predictions were limited to formula containing only C, H and O as all reported lucibufagins by Fallon, *et al.*[10] contained only these elements.

CSI:FingerID[21] within the SIRIUS 4.0.1 GUI[22] was used to explore possible structures for fragmentation patterns of unknown lucibufagins.

Results

The BioDendro workflow (Supplementary Figure S1) was applied to the positive ion mode acquisition of a single sample, Ppyr_hemolymph_extract.mzML, within the project dataset representing the hemolymph of an adult male beetle, *Photinus pyralis* spp. Examination of the raw data for Ppyr_hemolymph_extract.mzML, showed the acquisition of 2,501 MS/MS spectra, of which 1,251 were collected in positive ion mode. Deconvolution of the sample in MZmine2 (v2.53) extracted 29,677 features for positive ion mode when identified isotopes were excluded.

Clustering the lucibufagins in P.pyralis hemolymph. To identify new lucibufagins (Supplementary Figure S2) in the hierarchically clustered MS/MS spectra we first focused on a comparison against the original analysis. Fallon, *et al.*[10] putatively identified 17 lucibufagin compounds in the *P. pyralis* hemolymph using MZmine2 (v2.30) that varied by the degree of substitution of hydroxyl groups with acetyl and propyl groups. These lucibufagins were putatively identified by a combination of accurate mass, retention time and MS/MS spectra. Analysis with MS^2 similarity search in MZmine2 (v2.30) aligned 9 of the 17 lucibufagins to an MS/MS spectrum and the remaining 8 were defined by precursor mass and retention time.

A targeted search for the putatively identified compounds from Fallon, *et al.*[10] using BioDendro revealed that 15 of the 17 targets had been assigned an MS/MS spectra during alignment and had been placed into 4 clusters identified as clusters 82, 83, 108, and 110 (Fig. 1a). There was a total of 27 features within these 4 clusters, of which 11 of the 12 additional features represented adducts or aggregate ions that had been manually removed from the original analysis and a single new dipropylated isomer that was found beyond the retention time analyzed by Fallon, *et al.*[10].

Clusters 82 and 83 represented the largest with 12 features each and contained 12 of the original lucibufagins identified by Fallon, *et al.*[10]. The remaining 2 clusters (108 and 110) both contain core lucibufagin isomers and a single monoacetylated isomer, clustered away from the main branch in cluster 108 and 110. Inspection of associated ion tables for cluster 82 and 83, showed that 4 ions were present in every feature of both clusters (m/z 105.0701, 121.0648, 147.0805, 185.0961) (Supplementary Dataset). In addition, there were 3 ions unique to and present in every feature of cluster 82 (m/z 135.0443, 205.0863 and 413.1965) and 2 ions for cluster 83 (m/z 151.0392 and 265.1592) (Table 1). Only a single molecular formula for each fragment mass was predicted using the parameters detailed in Supplementary Information S2. These ions were considered diagnostic of a lucbibufagin-like structure. METLIN[18], MassBank[19] and NIST 14[20] spectral databases were searched for "lucibufagin" in an attempt to corroborate putative identification of these compounds however, for all 3 searches, zero hits were returned. A search of PubChem[23] and ChemSpider[24] had entries for lucibufagin C and several compounds with molecular similarity but no MS/MS spectral data was found. A literature search for "lucibufagins MS/MS" found several papers containing mass spectral information of several purified and characterized lucibufagins[25–27]. A search for the highly represented ions in these publications showed the presence of these ions.

The incidence of these ions from cluster 82 and 83 was scrutinized in surrounding clusters. A total of 12 clusters (75–83, 108–110), comprising 44 features (inclusive of the original 27) were shown to have a complement of these ions (Tables 1 and S3). The hierarchically clustered tree revealed clusters 75–83 belonged to the same cluster when the tree distance threshold was set at 0.97 (Fig. 1a).

Retention time was used to identify features with likely multiple adducts and confirmed through comparison of the calculated mass for the proposed adducts. The 44 features represented 29 unique compounds, including 15 from the original analysis and 14 from the re-analysis using BioDendro (Table 2).

Mass spectral investigation of the unknowns. The additional 14 compounds include 10 lucibufagins of uncharacterized structure and 4 with precursor ions represented by the isomers identified by Fallon, *et al.*[10]. Unknown 1, 2 and 3 (of clusters 75 and 76) are possible new lucibufagins containing a nitrogen atom based on the single proposed molecular formula proposed within 2 ppm for each of these compounds. Further inspection identified unknown lucibufagins which may represent varying degrees of saturation or conversion of ketone groups to hydroxys (or vice versa) by mass differences of 2 Daltons. Unknown 7 (M + H *m/z*531.2230) in cluster 79 and unknown 10 (M + H *m/z*535.2543) in cluster 81 and vary by ±2 Daltons from that of diacetylated

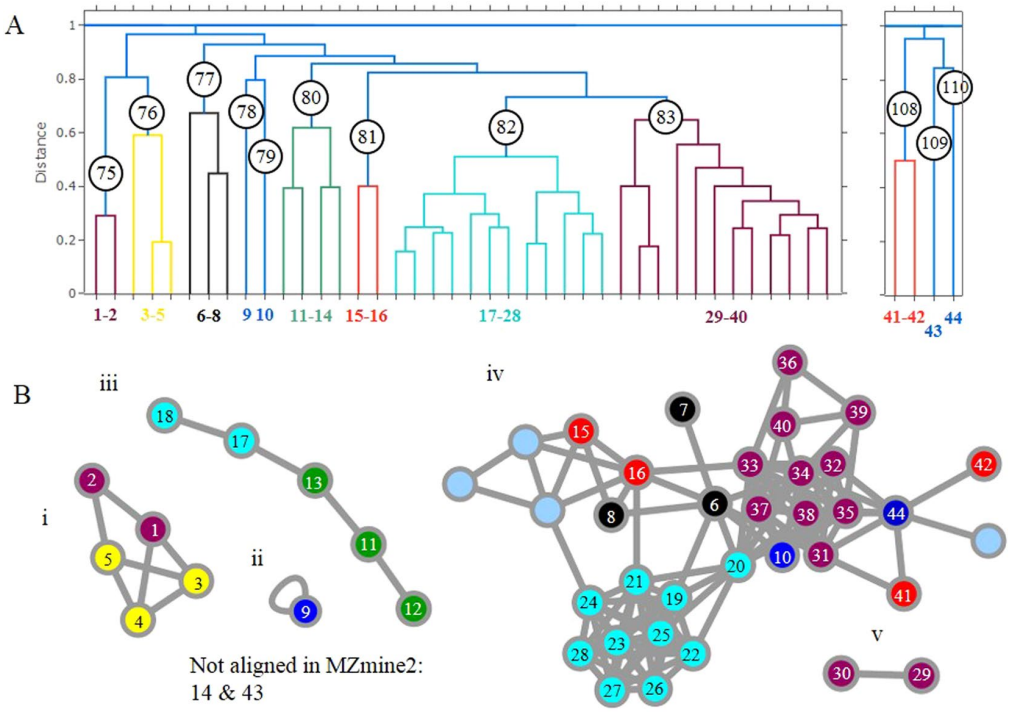


Figure 1. Clustering of Ppyr_hemolymph_extract.mzmL MS/MS spectra using (**a**) BioDendro complete linkage hierarchical clustering using a distance threshold of 0.7 and (**b**) FBMN molecular networking using a cosine score of 0.7. Features have been arbitrarily named from 1 to 44 to show the position in both tree and networks and represented in Table 1. The network nodes have been coloured to represent the same feature colour in the tree. Nodes without numbers are not clustered within the tree branches.

Fragment ion (m/z average_m/z min_m/z max)	% present		Predicted molecular formula	ppm error	Incidence in 44 features (%)
	cluster 82	cluster 83			
135.0443_135.0434_135.0448	100	0	C8H7O2	1.8	16 (36%)
205.0863_205.0840_205.0872	100	0	C12H13O3	1.8	18 (41%)
413.1965_413.1950_413.1970	100	0	C24H29O6	1.5	15 (34%)
105.0701_105.0698_105.0707	100	100	C8H9	2.1	36 (82%)
121.0648_121.0644_121.0653	100	100	C8H9O	0.1	41 (93%)
147.0805_147.0756_147.0812	100	100	C10H11O	1.1	35 (80%)
185.0961_185.0924_185.0975	100	100	C13H13O	−1.0	27 (61%)
151.0392_151.0388_151.0396	0	100	C8H7O3	1.5	14 (32%)
265.1592_265.1540_265.1670	0	100	C19H21O	1.9	17 (39%)

Table 1. Ions that show 100% representation in features of clusters 82 and 83 with their predicted molecular formula and ppm error compared to the average fragment ion mass.

lucibufagins (M + H *m/z*533.2385) and unknown 5 (M + H *m/z*549.2335) of cluster 77 is 2 Daltons higher than monoacetylated-monopropylated lucibufagins (M + H m/z 547.2541). Unknown 6 (m/z1114.5221) of cluster 78 and 8 (*m/z*1100.5061) of cluster 80 are postulated to be aggregate ions given size of the precursor ions comparative to the lucibufagins. Unknown 9 (m/z 517.2436, cluster 89) and 4 (m/z 507.2287, cluster 94) are M + H adducts of previously unidentified lucibufagins. There are 3 features proposed as aggregate ions of new diacetylated isomers eluting between 18.94 and 21.01 minutes in cluster 80. A new dipropylated isomer (m/z 561.2695) was also identified in cluster 83.

CSI:FingerID is a tool that predicts and ranks candidate structures of experimental MS/MS data. Structure prediction using the diacetylated lucibufagin 1 MS/MS (Feature 21, Table 2) elicited the structure with highest similarity as lucibufagin C with 71% similarity. However the next highest similarity was very close at 70% and was predicted to be 12-hydroxymoorastatin (Supplementary Figure S3). The best predicted structures for unknowns 7 (64% similarity), 9 (71% similarity) and 10 (68% similarity) closely resembled the 12-hydroxymoorastatin structure (Supplementary Figure S4). It is possible these features, in the absence of a lucibufagin-like structure with the

Feature[a]	Cluster #	Feature List ID	Fallon *et al.* (2018) putative identification	Putative classification	Adduct[b]	Aligned MS/MS in MZmine2[c]	FBMN cluster ID (as per Fig. 1)[d]	Molecular Formula[e]
1	75	Ppyr_hemolymph_extract_436.269439697265_13.538295		unknown 1	M + H		i	C24H37NO6
2	75	Ppyr_hemolymph_extract_453.296081542968_13.538295		adduct of unknown 1	M + NH4		i	C24H37NO6
3	76	Ppyr_hemolymph_extract_434.25390625_13.471111		unknown 2	M + H		i	C24H35NO6
4	76	Ppyr_hemolymph_extract_452.264434814453_8.0686129		unknown 3	M + H		i	C24H37NO7
5	76	Ppyr_hemolymph_extract_469.291137695312_8.0686129		adduct of unknown 3	M + NH4		i	C24H37NO7
6	77	Ppyr_hemolymph_extract_507.222839355468_13.718764		unknown 4	M + H		iv	C26H34O10
7	77	Ppyr_hemolymph_extract_524.249481201171_13.718764		adduct of unknown 4	M + NH4		iv	C26H34O10
8	77	Ppyr_hemolymph_extract_549.233520507812_15.795523		unknown 5	M + H		iv	C28H36O11
9	78	Ppyr_hemolymph_extract_1114.52197265625_19.262445		unknown 6	—		ii	—
10	79	Ppyr_hemolymph_extract_531.222991943359_16.840886		unknown 7	M + H		iv	C28H34O10
11	80	Ppyr_hemolymph_extract_1082.49517822265_19.965233		aggregate ion of an unknown diacetylated lucibufagin isomer	2 M + NH4		iii	C28H36O10
12	80	Ppyr_hemolymph_extract_1100.50646972656_18.35508		unknown 8			iii	
13	80	Ppyr_hemolymph_extract_1082.49487304687_21.307386		aggregate ion of an unknown diacetylated lucibufagin isomer	2 M + NH4		iii	C28H36O10
14	80	Ppyr_hemolymph_extract_1082.49530029296_18.991166		aggregate ion of an unknown diacetylated lucibufagin isomer	2 M + NH4		—	C28H36O10
15	81	Ppyr_hemolymph_extract_517.243591308593_15.954252		unknown 9	M + H		iv	C28H36O9
16	81	Ppyr_hemolymph_extract_535.254180908203_12.579138		unknown 10	M + H		iv	C28H38O10
17	82	Ppyr_hemolymph_extract_1082.49475097656_15.123002		aggregate ion of diacetylated lucibufagin isomer 1	2 M + NH4	—	iv	C28H36O10
18	82	Ppyr_hemolymph_extract_1110.52667236328_17.431269		aggregate ion of monoacetylated, monopropylated lucibufagin isomer 2	2 M + NH4	—	iv	C29H38O10
19	82	Ppyr_hemolymph_extract_491.227569580078_12.962671	monoacetylated lucibufagin isomer 4		M + H	yes	iv	C26H34O9
20	82	Ppyr_hemolymph_extract_491.227661132812_10.204906	nonoacetylated lucibufagin isomer 1		M + H	yes	iv	C26H34O9
21	82	Ppyr_hemolymph_extract_533.237884521484_15.123002	diacetylated lucibufagin isomer 1		M + H	yes	iv	C28H36O10
22	82	Ppyr_hemolymph_extract_547.254028320312_17.431269	monoacetylated, mono propylated lucibufagin isomer 2		M + H	yes	iv	C29H38O10
23	82	Ppyr_hemolymph_extract_550.264221191406_15.123002		adduct of diacetylated lucibufagin isomer 1	M + NH4	—	iv	C28H36O10
24	82	Ppyr_hemolymph_extract_561.269470214843_19.784291	dipropylated lucibufagin isomer 3		M + H	no	iv	C30H40O10
25	82	Ppyr_hemolymph_extract_561.26953125_19.535629	dipropylated lucibufagin isomer 2		M + H	yes	iv	C30H40O10
26	82	Ppyr_hemolymph_extract_564.280517578125_17.431269		adduct of monoacetylated, mono propylated lucibufagin isomer 2	M + NH4	—	iv	C29H38O10

Continued

Feature[a]	Cluster #	Feature List ID	Fallon *et al.* (2018) putative identification	Putative classification	Adduct[b]	Aligned MS/MS in MZmine2[c]	FBMN cluster ID (as per Fig. 1)[d]	Molecular Formula[e]
27	82	Ppyr_hemolymph_extract_574.264343261718_15.123002		adduct of diacetylated lucibufagin isomer 1	M + ACN + H	—	iv	C28H36O10
28	82	Ppyr_hemolymph_extract_578.296081542968_19.784291		adduct of dipropylated lucibufagin isomer 3	M + NH4	—	iv	C30H40O10
29	83	Ppyr_hemolymph_extract_1065.4677734375_15.366083		aggregate ion of diacetylated lucibufagin isomer 2	2 M + H	—	v	C28H36O10
30	83	Ppyr_hemolymph_extract_1082.49499511718_15.366083		aggregate ion of diacetylated lucibufagin isomer 2	2 M + NH4	—	v	C28H36O10
31	83	Ppyr_hemolymph_extract_491.227600097656_13.224755	monoacetylated lucibufagin isomer 5		M + H	no	iv	C26H34O9
32	83	Ppyr_hemolymph_extract_491.228057861328_11.943648	monoacetylated lucibufagin isomer 3		M + H	no	iv	C26H34O9
33	83	Ppyr_hemolymph_extract_533.238098144531_15.366083	diacetylated lucibufagin isomer 2		M + H	yes	iv	C28H36O10
34	83	Ppyr_hemolymph_extract_547.253814697265_17.719947	monoacetylated, mono propylated lucibufagin isomer 3		M + H	yes	iv	C29H38O10
35	83	Ppyr_hemolymph_extract_547.254211425781_17.046113	monoacetylated, mono propylated lucibufagin isomer 1		M + H	no	iv	C29H38O10
36	83	Ppyr_hemolymph_extract_550.264404296875_15.366083		adduct of diacetylated lucibufagin isomer 2	M + NH4	—	iv	C28H36O10
37	83	Ppyr_hemolymph_extract_561.269348144531_18.878158	dipropylated lucibufagin isomer 1		M + H	no	iv	C30H40O10
38	83	Ppyr_hemolymph_extract_561.269592285156_20.078779		unknown dipropylated lucibufagin isomer	M + H	—	iv	C30H40O10
39	83	Ppyr_hemolymph_extract_564.280212402343_17.719947		adduct of monoacetylated, monopropylated lucibufagin isomer 3	M + NH4	—	iv	C29H38O10
40	83	Ppyr_hemolymph_extract_578.295959472656_18.878158		adduct of dipropylated lucibufagin isomer 1	M + NH4	—	iv	C30H40O10
41	108	Ppyr_hemolymph_extract_449.217010498046_9.3229572	core lucibufagin isomer 2		M + H	yes	iv	C24H32O8
42	108	Ppyr_hemolymph_extract_491.227844238281_14.49087	monoacetylated lucibufagin isomer 6		M + H	no	iv	C26H34O9
43	109	Ppyr_hemolymph_extract_1615.72888183593_15.123002		aggregate ion of an unknown diacetylated lucibufagin isomer	3 M + NH4		—	C28H36O10
44	110	Ppyr_hemolymph_extract_449.217071533203_10.789573	core lucibufagin isomer 1		M + H	yes	iv	C24H32O8

Table 2. Features of the lucibufagin clusters 75–83 and 108–110. The feature list ID was aligned to the putative ID in Fallon, *et al.*[10]. Additional features were assigned a putative identity based on comparison to the original analysis and a calculated molecular formula. [a]Feature number is arbitrary and correlates to order in Fig. 1 tree. [b]Adducts identified for compounds with 2 or more co-eluting ions and accurate mass. Single ions are not identified to an adduct type. [c]Fallon *et al.* manually removed adducts from MS2 similarity search and are represented by a dash. [d]Dashes represent features that had no aligned MS/MS spectra by MZmine2. [e]Proposed formula based on accurate mass measurements. All formulas are within 2 ppm of the experimental measurement.

corresponding molecular formula were matched best to moorastatin-like compounds. Unknown 4 (65% similarity) and 5 (60% similarity) were predicted to be polycyclic compounds containing high numbers of acetyl groups, characteristics that are common to lucibufagins. The best matches for unknown 1, 2 and 3 contained single nitrogen atoms and multiple hydroxyl groups, however all 3 had matches below 60%.

Comparison of BioDendro to FBMN of GNPS. The clustering of lucibufagins using BioDendro was compared to the FBMN module within the GNPS infrastructure. Application of BioDendro here used the MZmine2 (v2.53) exported feature list and the MGF from ProteoWizard which encompassed the entire 1,251 MS/MS positive ion mode spectra and after alignment, 492 features had an associated MS/MS spectra. Alignment of features to spectra for use in FBMN occurs during analysis using MZmine2 and a.txt feature list and MGF file are exported containing aligned features only. MZmine2 exported a total of 402 MS/MS spectra. The 44 lucibufagin features interrogated with the BioDendro output were examined within the molecular networks of FBMN (Fig. 1b). 42 of the 44 features were aligned an MS/MS by MZmine2 and were located in 5 networks generated using a cosine score of 0.7. From the aligned spectra, there exists a visual similarity of the structure between molecular networking and hierarchically clustered tree. The 5 new lucibufagins of clusters 75 and 76 found by BioDendro share a single network and all of features of cluster 82 and 83, bar 2 features, are also networked. Notable differences are the additional 4 nodes (nodes that are without a feature number) that form part of network iv) (Fig. 1b), comparison of the MS/MS spectra that created edges to the new nodes exhibit similarity based on neutral losses, a comparison that is optionally carried out in separate analyses within BioDendro.

Mass spectral searching was enabled during analysis using all MS/MS libraries accessible through GNPS. To date, GNPS has access to over 2.4 million MS/MS reference spectra (when considering GC-MS and LC-MS) from 27 different libraries[3]. From the 402 networked spectra, 10 features were matched to a reference spectra which did not include lucibufagin compounds.

Application run time. Based on running on a Windows 7 PC, hierarchically clustering 492 spectra with BioDendro took 50 seconds. Molecular networking using FBMN is completed using a web-based server and analysis time will be dependent on the server load at that time. Several iterations of this analysis using identical settings at various times of the day took from 5 to 15 minutes to network 402 features.

Discussion

Metabolomics studies often produce several hundred to several thousand unique MS/MS spectra for which metabolite identification or classification can be facilitated. Here, a protocol is presented that hierarchically clusters MS/MS spectra from metabolomics data and outputs fragment ion tables in an easy to interrogate manner. Many studies have explored numerous ways in clustering MS/MS spectra for proteomics analysis and mass spectral dereplication[16,28–31] however, clustering MS/MS data for metabolomics has considerations not generally applicable to proteomics data such as distinguishing isobaric ions that are often dereplicated in many of these analysis pipelines[4]. The BioDendro protocol does not approach dereplication in the traditional manner of combining MS/MS data of same precursor mass and high similarity spectra regardless of retention time but rather a feature has a single alignment to an MS/MS spectrum within an m/z and retention time tolerance. In this way, all isobaric ions are represented individually.

Hierarchically clustering the MS/MS spectra of *P. pyralis* hemolymph facilitated the putative identification of 44 features containing a fragmentation pattern similar to that of the lucibufagins[25,26]. Using the tree to visualize clustering offers the opportunity to make easy, intuitive decisions regarding the structure of the clustered data. After review of the tree structure surrounding the analytes of interest it was particularly useful to adjust the distance threshold to adequately cluster the lucibufagins together.

Additionally, the ion tables and histograms output by BioDendro make it particularly easy to see the contribution of individual ions within the entirety of a single cluster, not presently available within FBMN. Identifying the ions which are heavily represented within a cluster can help to identify compounds which are not currently in mass spectral databases. Searching mass spectral databases[18–20,23,24] for lucibufagins returned no hits, however MS/MS spectra for several lucibufagins were found in the literature[25–27]. Many natural product mass spectra exist in individual publications that haven't been submitted to databases, especially those that were collected before the formation of mass spectral databases. Identification of those fragments which typify molecular structures could be used as search terms within literature to further leverage metabolite classification. Many natural product mass spectra exist in individual publications that haven't been submitted to databases, especially those that were collected before the formation of mass spectra databases. Identification of those fragments which typify molecular structures could be used as search terms within literature to further leverage metabolite classification.

Comparison of the tree created in BioDendro against the molecular networking of FBMN exhibited clustering that was similar across both platforms. GNPS and the newer module FBMN have become gold standards for clustering metabolomics MS/MS data. However, difficulties in interrogating complete unknown spectral patterns led us to develop an alternative pipeline that allowed simple application with easily interrogable clusters and inspection of the spectral information behind those clusters. FBMN is a web-based application requiring a) upload of data to a webserver, b) data inputs produced by a limited number of data processing software c) production of the network and limited visualization within FBMN and d) further visualization in Cytoscape, an additional platform external to GNPS. Whereas hierarchical clustering in BioDendro a) can be run locally, b) accepts a feature list from any processing software, and c) produces simplified visualisation outputs not dependent on external software. Additionally, processing time for hierarchically clustering and outputting results is 7 times quicker than FBMN meaning optimization or modification of analysis parameters can be done with minimal downtime.

Conclusion

Clustering MS/MS spectra for metabolomics data is often a way in which a user can interrogate feature classification or identification without the presence of authentic standards. It is a non-trivial undertaking that can often require a degree of technical knowledge regarding mass spectrometry and biological knowledge about the sample origin. Hierarchically clustering MS/MS spectra, visualized as a tree, presented an easily interrogable format, which coupled with cluster ion tables allowed users to make informed decisions for the classification of 29 unique compounds as lucibufagins. Accessing the ions which were highly represented within certain clusters improved the users' confidence and ability in assigning a metabolite class to compounds that are not present in current MS/MS databases.

Software availability statement. The BioDendro software developed in this study is available via https://github.com/ccdmb/BioDendro complete with example datasets and Jupyter Notebooks.

Data availability

The Firefly data is freely available MetaboLights – project number MTBLS698.

Received: 9 October 2019; Accepted: 24 March 2020;
Published online: 08 April 2020

References

1. Rawlinson, C. *et al*. The identification and deletion of the polyketide synthase-nonribosomal peptide synthase gene responsible for the production of the phytotoxic triticone A/B in the wheat fungal pathogen Pyrenophora tritici-repentis. *Environmental Microbiology* **21**, 4875–4886, https://doi.org/10.1111/1462-2920.14854 (2019).
2. Wang, M. *et al*. Sharing and community curation of mass spectrometry data with Global Natural Products Social Molecular Networking. *Nat Biotechnol* **34**, 828–837, https://doi.org/10.1038/nbt.3597 (2016).
3. Allegra T., A. *et al*. Reproducible Molecular Networking Of Untargeted Mass Spectrometry Data Using GNPS. https://doi.org/10.26434/chemrxiv.9333212.v1 (2019).
4. Nothias, L. F. *et al*. Feature-based Molecular Networking in the GNPS Analysis Environment. *bioRxiv*, 812404, https://doi.org/10.1101/812404 (2019).
5. R: a language and environment for statistical computing (R Foundation for Statistical Computing, Vienna, 2014).
6. Naake, T. & Gaquerel, E. MetCirc: navigating mass spectral similarity in high-resolution MS/MS metabolomics data. *Bioinformatics* **33**, 2419–2420, https://doi.org/10.1093/bioinformatics/btx159 (2017).
7. Åberg, K. M., Torgrip, R. J. O., Kolmert, J., Schuppe-Koistinen, I. & Lindberg, J. Feature detection and alignment of hyphenated chromatographic–mass spectrometric data: Extraction of pure ion chromatograms using Kalman tracking. *Journal of Chromatography A* **1192**, 139–146, https://doi.org/10.1016/j.chroma.2008.03.033 (2008).
8. Smith, C. A., Want, E. J., O'Maille, G., Abagyan, R. & Siuzdak, G. XCMS: Processing Mass Spectrometry Data for Metabolite Profiling Using Nonlinear Peak Alignment, Matching, and Identification. *Analytical Chemistry* **78**, 779–787, https://doi.org/10.1021/ac051437y (2006).
9. Grace, S. C., Embry, S. & Luo, H. Haystack, a web-based tool for metabolomics research. *BMC Bioinformatics* **15**, S12, https://doi.org/10.1186/1471-2105-15-S11-S12 (2014).
10. Fallon, T. R. *et al*. Firefly genomes illuminate parallel origins of bioluminescence in beetles. *eLife* **7**, e36495, https://doi.org/10.7554/eLife.36495 (2018).
11. Pedregosa, F. *et al*. Scikit-learn: Machine learning in Python. *Journal of machine learning research* **12**, 2825–2830 (2011).
12. Walt, Svd, Colbert, S. C. & Varoquaux, G. The NumPy Array: A Structure for Efficient Numerical Computation. *Computing in Science & Engineering* **13**, 22–30, https://doi.org/10.1109/mcse.2011.37 (2011).
13. Jones, E., Oliphant, E. & Peterson, P. *SciPy: Open Source Scientific Tools for Python*, <http://www.scipy.org/> (2001).
14. Plotly Technologies Inc. *Dendrograms in Python*, <https://plot.ly/python/dendrogram/>.
15. Hunter, J. D. Matplotlib: A 2D Graphics Environment. *Computing in Science & Engineering* **9**, 90–95, https://doi.org/10.1109/MCSE.2007.55 (2007).
16. Myers, O. D., Sumner, S. J., Li, S., Barnes, S. & Du, X. One Step Forward for Reducing False Positive and False Negative Compound Identifications from Mass Spectrometry Metabolomics Data: New Algorithms for Constructing Extracted Ion Chromatograms and Detecting Chromatographic Peaks. *Analytical Chemistry* **89**, 8696–8703, https://doi.org/10.1021/acs.analchem.7b00947 (2017).
17. Chambers, M. C. *et al*. A cross-platform toolkit for mass spectrometry and proteomics. *Nature Biotechnology* **30**, 918–920, https://doi.org/10.1038/nbt.2377 (2012).
18. Smith, C. A. *et al*. METLIN: A Metabolite Mass Spectral Database. *Therapeutic Drug Monitoring* **27**, 747–751, https://doi.org/10.1097/01.ftd.0000179845.53213.39 (2005).
19. Horai, H. *et al*. MassBank: a public repository for sharing mass spectral data for life sciences. *Journal of Mass Spectrometry* **45**, 703–714, https://doi.org/10.1002/jms.1777 (2010).
20. Stein, S. E. (National Institute of Standards and Technology, Gaithersburg, MD, 2014).
21. Dührkop, K., Shen, H., Meusel, M., Rousu, J. & Böcker, S. Searching molecular structure databases with tandem mass spectra using CSI:FingerID. *Proceedings of the National Academy of Sciences* **112**, 12580-12585, 10.1073/pnas.1509788112 (2015).
22. Dührkop, K. *et al*. SIRIUS 4: a rapid tool for turning tandem mass spectra into metabolite structure information. *Nature Methods* **16**, 299–302, https://doi.org/10.1038/s41592-019-0344-8 (2019).
23. Kim, S. *et al*. PubChem 2019 update: improved access to chemical data. *Nucleic Acids Research* **47**, D1102–D1109, https://doi.org/10.1093/nar/gky1033 (2018).
24. Pence, H. E. & Williams, A. ChemSpider: An Online Chemical Information Resource. *Journal of Chemical Education* **87**, 1123–1124, https://doi.org/10.1021/ed100697w (2010).
25. Eisner, T., Goetz, M. A., Hill, D. E., Smedley, S. R. & Meinwald, J. Firefly "femmes fatales" acquire defensive steroids (lucibufagins) from their firefly prey. *Proc Natl Acad Sci USA* **94**, 9723–9728, https://doi.org/10.1073/pnas.94.18.9723 (1997).
26. Smedley, S. R. *et al*. Bufadienolides (lucibufagins) from an ecologically aberrant firefly (Ellychnia corrusca). *Chemoecology* **27**, 141–153, https://doi.org/10.1007/s00049-017-0240-6 (2017).
27. Meinwald, J., Wiemer, D. F. & Eisner, T. Lucibufagins. 2. Esters of 12-oxo-2.beta.,5.beta.,11.alpha.-trihydroxybufalin, the major defensive steroids of the firefly Photinus pyralis (Coleoptera: Lampyridae). *Journal of the American Chemical Society* **101**, 3055–3060, https://doi.org/10.1021/ja00505a037 (1979).
28. Frank, A. M. *et al*. Clustering Millions of Tandem Mass Spectra. *Journal of Proteome Research* **7**, 113–122, https://doi.org/10.1021/pr070361e (2008).
29. Rasche, F. *et al*. Identifying the Unknowns by Aligning Fragmentation Trees. *Analytical Chemistry* **84**, 3417–3426, https://doi.org/10.1021/ac300304u (2012).

30. Broeckling, C. D., Afsar, F. A., Neumann, S., Ben-Hur, A. & Prenni, J. E. RAMClust: A Novel Feature Clustering Method Enables Spectral-Matching-Based Annotation for Metabolomics Data. *Analytical Chemistry* **86**, 6812–6817, https://doi.org/10.1021/ac501530d (2014).
31. Rieder, V. *et al.* Comparison and Evaluation of Clustering Algorithms for Tandem Mass Spectra. *Journal of Proteome Research* **16**, 4035–4044, https://doi.org/10.1021/acs.jproteome.7b00427 (2017).

Acknowledgements
This work was supported by the Curtin Institute for Computation (CIC), Curtin University, the Grains Research and Development Corporation (GRDC, grant CUR000023) and by resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia. CR was on a RTP scholarship provided by the Australian Government.

Author contributions
C.R. and C.S.M. conceived and designed research. D.A.J., S.M., S.R. and P.M. contributed to BioDendro code development. C.R. analyzed the data. C.R. and P.M. wrote the original manuscript. All authors contributed to the editing of the manuscript. All authors read and approved the manuscript.

Competing interests
The authors declare no competing interests.

Additional information
Supplementary information is available for this paper at https://doi.org/10.1038/s41598-020-63036-1.

Correspondence and requests for materials should be addressed to C.R. or P.M.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

CHAPTER 7 — THEME 2

Crop-Zone Weed Mycobiomes of the South-Western Australian Grain Belt

This chapter is also published in:
Frontiers in Microbiology, 2020, vol. 11, p. 2944
https://doi.org/10.3389/fmicb.2020.581592

7.1 Declaration

Title	Crop-Zone Weed Mycobiomes of the South-Western Australian Grain Belt.
Authors	Pippa J. Michael, **Darcy A. B. Jones**, Nicole White, James K. Hane, Michael Bunce, and Mark Gibberd
Publication	2020. *Frontiers in Microbiology*, 11.
DOI	`https://doi.org/10.3389/fmicb.2020.581592`

This thesis chapter is submitted in the form of a collaboratively-written and peer-reviewed manuscript. As such, not all work contained within this chapter can be attributed to the Ph. D. candidate.

The following contributions were made by the Ph.D candidate (DABJ) and co-authors:

- PM, JKH, and MG conceived the study.
- PM conducted field sampling.
- NW conducted DNA extraction and sequencing.
- **DABJ** and JKH contributed to DNA sequence analysis.
- **DABJ** and PM performed multivariate analyses.
- PM and JKH wrote the manuscript.
- All authors read and approved the manuscript.

I, Darcy Jones, hereby certify that this attribution statement is an accurate record of my contribution to the research presented in this chapter.

Darcy Jones

I certify that this attribution statement is an accurate record of Darcy Jones' contribution to the research presented in this chapter.

James K. Hane | Pippa Michael | Nicole White

Michael Bunce | Mark Gibberd

frontiers in Microbiology

ORIGINAL RESEARCH
published: 24 November 2020
doi: 10.3389/fmicb.2020.581592

Crop-Zone Weed Mycobiomes of the South-Western Australian Grain Belt

Pippa J. Michael[1], Darcy Jones[1], Nicole White[2], James K. Hane[1], Michael Bunce[2] and Mark Gibberd[1]*

[1] Centre for Crop and Disease Management, School of Molecular and Life Sciences, Curtin University, Perth, WA, Australia,
[2] TRENDLab, School of Molecular and Life Sciences, Curtin University, Perth, WA, Australia

In the absence of a primary crop host, secondary plant hosts may act as a reservoir for fungal plant pathogens of agricultural crops. Secondary hosts may potentially harbor heteroecious biotrophs (e.g., the stripe rust fungus *Puccinia striiformis*) or other pathogens with broad host ranges. Agricultural grain production tends toward monoculture or a limited number of crop hosts over large regions, and local weeds are a major source of potential secondary hosts. In this study, the fungal phyllospheres of 12 weed species common in the agricultural regions of Western Australia (WA) were compared through high-throughput DNA sequencing. Amplicons of D2 and ITS were sequenced on an Illumina MiSeq system using previously published primers and BLAST outputs analyzed using MEGAN. A heatmap of cumulative presence–absence for fungal taxa was generated, and variance patterns were investigated using principal components analysis (PCA) and canonical correspondence analysis (CCA). We observed the presence of several major international crop pathogens, including basidiomycete rusts of the *Puccinia* spp., and ascomycete phytopathogens of the *Leptosphaeria* and *Pyrenophora* genera. Unrelated to crop production, several endemic pathogen species including those infecting Eucalyptus trees were also observed, which was consistent with local native flora. We also observed that differences in latitude or climate zones appeared to influence the geographic distributions of plant pathogenic species more than the presence of compatible host species, with the exception of Brassicaceae host family. There was an increased proportion of necrotrophic Ascomycete species in warmer and drier regions of central WA, compared to an increased proportion of biotrophic Basidiomycete species in cooler and wetter regions in southern WA.

Keywords: mycobiome, phyllosphere, fungi, plant pathogen, weeds

OPEN ACCESS

Edited by:
Dhanushka Nadeeshan Wanasinghe, Kunming Institute of Botany, China

Reviewed by:
Sajeewa S. N. Maharachchikumbura, University of Electronic Science and Technology of China, China
Mathabatha Evodia Setati, Stellenbosch University, South Africa

****Correspondence:***
James K. Hane
james.hane@curtin.edu.au

Specialty section:
This article was submitted to Fungi and Their Interactions, a section of the journal Frontiers in Microbiology

Received: *09 July 2020*
Accepted: *28 October 2020*
Published: *24 November 2020*

Citation:
Michael PJ, Jones D, White N, Hane JK, Bunce M and Gibberd M (2020) Crop-Zone Weed Mycobiomes of the South-Western Australian Grain Belt. Front. Microbiol. 11:581592. doi: 10.3389/fmicb.2020.581592

INTRODUCTION

In the absence of a primary crop host, secondary plant hosts may act as a reservoir for fungal plant pathogens that cause economically significant crop diseases. In a crop monoculture environment, readily available secondary hosts are commonly weed species (Narayanasamy, 2011), which may serve as important sources of inoculum for agricultural crops that are sown or planted in the following season. Outbreaks of crop diseases caused by fungi are typically monitored based on observations of disease symptoms or fungal spore counts, making early preventative control

measures (i.e., fungicide application) difficult or impossible to implement (Lindahl and Kuske, 2013). Epidemiological studies have highlighted the need for eradication of potential sources of inoculum (i.e., weeds) between growing seasons so as to restrict the incidence and spread of disease to newly planted crops (Narayanasamy, 2011). These studies have focused on a small number of pathogen species infecting major crops; however, little is known of the larger pool of fungal species that co-inhabit the plant phyllosphere (leaf surface and interior). Pathogens constitute a relatively minor proportion of a fungal community—or mycobiome (Sapkota et al., 2015; Donovan et al., 2018). Limited information on the composition of plant mycobiomes is available, with the majority of mycobiome surveys focusing on soil and human body environments (Donovan et al., 2018). Plant pathogens are not commonly featured, despite occasional reports in non-plant mycobiomes (Weyrich et al., 2017; Donovan et al., 2018). A study by Sapkota et al. (2015) found most fungal species to be common to all host cereal crops, with the exception of a few crop-specific pathogens, indicating that the majority of fungi in leaves are not host-specific and that host-specific pathogens live in a "sea" of non-specific fungi. How plant-associated mycobiomes react to inter-species interactions, environmental conditions, host genotype, and agronomic practices is largely unexplored.

Over 99,000 fungal species are documented; however, total diversity is estimated to be between 2.2 and 3.8 million species (Hawksworth and Lücking, 2017). Traditional culture-dependent studies have revealed an immense diversity of fungal communities colonizing the plant phyllosphere, although more recent metagenomic studies indicate that this may be underestimated (Peršoh, 2015). Phyllosphere mycobiomes are highly variable among leaves within an individual plant, with factors such as leaf position, canopy height, and leaf age shown to influence leaf-to-leaf variability (Kinkel, 1997). An overlap between fungi found in the phyllosphere and the air spora has also been reported (Levetin and Dorsey, 2006). A study on beech tree (*Fagus sylvatica*) looking at variability at four different spatial scales (tree, branch, group of leaves, and individual leaf) found the majority of variation occurred at the smallest spatial scale (i.e., between individual leaves), with intra-host variability of phyllosphere fungal populations distinctly greater than inter-host variability (Cordier et al., 2012). However, within a single tree canopy, mycobiome profiles become more similar with decreasing distance, suggesting that these differences may be minimized for a single sapling when it is small and has only a few leaves. When analyzing differences between individual trees, dissimilarity between mycobiomes was linked with genetic rather than geographic distance between trees (Cordier et al., 2012). Several studies have also found that genetic makeup of the plant host at both the species and cultivar/ecotype level was the major factor influencing fungal diversity on plant leaves (Joshee et al., 2009; Bálint et al., 2013; Hunter et al., 2015; Sapkota et al., 2015), with spatial and seasonal factors also having significant but lesser impacts. While Sapkota et al. (2015) found crop genotype was the principal factor in explaining mycobiome diversity in cereal phyllospheres, within each of the individual crops studied (wheat, winter, and spring barley), location also played an important role. Blixt et al. (2010) found that variation of community composition was greater between fields of diseased wheat (*Triticum aestivum*) than within fields as did Zimmerman and Vitousek (2012) who found that among-site diversity of a single tree species (*Metrosideros polymorpha*) contributed more than within-site diversity to the overall fungal community richness. In addition to spatial, temporal, and genetic factors, phyllosphere mycobiomes are also influenced by management factors, with fungicide use shown to have an impact on the composition of cereal leaf mycobiomes (Karlsson et al., 2014; Sapkota et al., 2015).

In this study, we identified the plant pathogens present in the phyllosphere mycobiomes of commonly found weed species adjacent to cropping fields. We also report differences in weed mycobiome composition in response to host species and spatial/climatic factors. This information on weed host-specific or region-specific association with pathogen species may become an increasingly important factor in developing new methods for crop disease management in the future.

MATERIALS AND METHODS

Field Sampling

Leaves from 12 common agricultural weed species (**Table 1**) were sampled from 15 locations across the Western Australian grain belt (**Figure 1**) over a 5-day period during autumn, 2016. All sites were located on road verges adjacent to cropping paddocks and selected for maximum number of weed species present. Environmental and spatial data were recorded for each site (**Table 2**), with climatic data from 1950 to 2020. The optimal number of individual plants required per site to be representative of the fungal biota present needed to be determined. Therefore, 10 leaves of two weed species (*Lolium rigidum* and *Raphanus raphanistrum*) were sampled from one site (Yoting) and the preliminary metagenome sequencing results were used to benchmark the appropriate number of leaves for sequencing (**Supplementary Figure 1**). Thus, for each weed species sampled at each site, one uppermost leaf from 10 individual plants were stored at 4°C during transport to the lab (24–48 h), then at −20°C. Each set of 10 leaves was randomly sub-sampled using a 5-mm disc punch and combined in a 5-ml sample vial until full prior to DNA sequencing.

DNA Extraction and Quantification

DNA was extracted using a QIAmp Plant Mini Kit (Qiagen, Venlo) with a modified protocol including (1) 600 μl of API digestion buffer; (2) overnight sample digestion at 65°C; and (3) the following day, addition of 195 μl of P3 to digests prior to putting on ice. A total volume of 400 μl was placed in a QIAcube, after which the remaining QIAmp Plant Mini Kit protocol was followed. Genomic DNA (gDNA) extracts were eluted in 100 μl of AE buffer and stored at −20°C, and then quantified and assessed for quality via real-time quantitative polymerase chain reaction (qPCR) at three dilutions (1:1, 1:10, and 1:100). Primer pairs were designed to amplify a highly conserved region of the 26S gene between the D1/D2 domains (Q1_F 5′ GTTGTTTGGGAATGCAGCTC 3′ and QB3_R 5′

TABLE 1 | Sampling incidence of 12 common weed species from 15 locations within the Western Australian grain belt.

Species	Family	Group	Common name	No. locations sampled
Arctotheca calendula L.	Asteraceae	Dicotyledon	Capeweed	14
Avena fatua L.	Poaceae	Monocotyledon	Wild oats	13
Bromus spp.	Poaceae	Monocotyledon	Brome grass	4
Echium plantagineum L.	Boraginaceae	Dicotyledon	Patterson's curse	2
Emex australis Steinh.	Polygonaceae	Dicotyledon	Doublegee	3
Erodium spp.	Geraniaceae	Dicotyledon	Erodium	4
Hordeum spp.	Poaceae	Monocotyledon	Barley grass	1
Hypochaeris radicata L.	Asteraceae	Dicotyledon	Flatweed	8
Lolium rigidum Gaud	Poaceae	Monocotyledon	Annual ryegrass	14
Oxalis pes-caprae L.	Oxalidaceae	Dicotyledon	Soursob	1
Raphanus raphanistrum L.	Brassicaceae	Dicotyledon	Wild radish	12
Vulpia spp.	Poaceae	Monocotyledon	Silvergrass	2

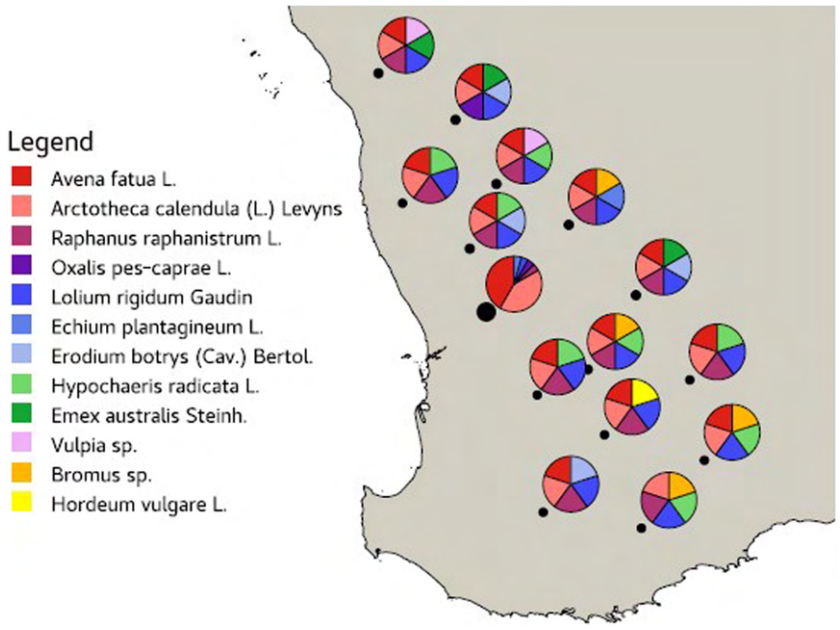


FIGURE 1 | Map of the Western Australian grain belt showing the 15 sampling locations (black dot) and the weed species sampled at each site.

AGTGCTTTTCATCTTTCCCTCAC 3′). qPCR was performed in 25-μl reactions containing 1 × PCR Gold Buffer, 2.5 mM $MgCl_2$, 0.4 mg/ml BSA, 0.25 mM of each dNTP, 0.4 μM of forward and reverse primer, 0.25 μl of AmpliTaq Gold, 0.6 μl of SYBR Green, and 2 μl of gDNA. The qPCR cycling conditions included an initial heat denaturation at 95°C for 5 min, 40 cycles of 95°C for 30 s, 52°C for 30 s, and 72°C for 45 s, and then a final extension at 72°C for 10 min. From the qPCR results, an optimal DNA concentration was selected for DNA sequencing, which was free of inhibition and yielded DNA of sufficient quality, as reported to facilitate reproducible quantitative data by (Murray et al., 2011).

High-Throughput DNA Sequencing

The D2 and ITS2 amplicons were sequenced on an Illumina MiSeq system utilizing previously published primers that were modified with a unique 8-bp Multiplex Identifier tag (MID-tag) and MiSeq adaptors for paired-end sequencing. For the D2 domain primers, U1_F (Putignani et al., 2008) and NL4_R (Kurtzman and Robnett, 1998) were utilized, and for the ITS2 region, fITS7_F (Ihrmark et al., 2012) and ITS4_R (White et al., 1990) were used. Independent MID-tagged qPCR setup for samples and controls were prepared in a physically separate ultra-clean laboratory and were carried out using each primer set in 25-μl reactions containing 1 × PCR Gold Buffer, 2.5 mM $MgCl_2$, 0.4 mg/ml BSA, 0.25 mM of each dNTP, 0.4 μM of forward and reverse MID-tag primer, 0.25 μl of AmpliTaq Gold, 0.6 μl of SYBR Green, and 2 μl of gDNA. The cycling conditions for qPCR using the U1_F/NL4_R (52°C annealing) and fITS7_F/ITS4_R (54°C annealing) primer sets were as follows: initial heat denaturation at 95°C for 5 min, followed by 40

TABLE 2 | Spatial and climatic details (averaged over 1950–2020) for each location.

Location	Latitude (°S)	Longitude (°E)	Annual rainfall (mm)	Annual mean temp (°C)	Annual min temp (°C)	Annual max temp (°C)	No. weeds sampled
Badgingarra	−30.2111	115.4599	521	19.2	12.3	25.6	5
Borden	−34.0256	118.2634	377	15.8	9.8	21.7	5
Buntine	−29.9852	116.5618	319	19.4	12.2	26.5	6
Harrismith	−32.9292	117.8300	356	16.5	9.6	22.8	5
Hyden	−32.2848	118.8303	342	17.1	9.9	24.5	5
Kalannie	−30.4651	117.4111	307	19.2	12.0	26.0	6
Kojonup	−33.8383	117.1103	505	15.5	9.1	21.6	5
Morawa	−29.0817	115.9865	308	20.0	12.8	27.6	6
Mullewa	−28.6862	115.1780	352	19.6	13.0	26.6	6
Newdegate	−33.2284	119.0020	357	16.0	9.5	22.9	5
Nungarin	−31.2907	118.1990	296	18.2	11.4	25.3	6
Pingelly	−32.4618	117.0406	427	16.7	9.9	23.7	5
Toodyay	−31.4870	116.4431	485	18.1	11.2	24.9	3
Walebing	−30.7444	116.2491	432	18.3	11.6	25.1	6
Yoting	−32.1621	117.6333	325	17.4	10.2	24.2	4

Data sourced from https://www.longpaddock.qld.gov.au/silo/.

cycles of 95°C for 30 s; 52°C or 54°C for 30 s (annealing step); and 72°C for 45 s followed by final extension at 72°C for 10 min. Multiplex Identifier-tagged PCR amplicons were generated in duplicate for each sample and pooled together to minimize the effects of PCR stochasticity. The pooled amplicons were quantified on a LabChip (with high-sensitivity chip) and then combined to produce a final library of equimolar ratio per sample. The final library was then purified on a PippinPrep with a size selection gate of 300–600 base pair capture following the manufacturer's protocol (PerkinElmer). The purified library was diluted with purified water and re-run on the LabChip to determine the volume required (2 nM) for Illumina MiSeq paired-end sequencing. For each MID-tagged qPCR assay, extraction and PCR controls were included, and if found to contain amplifiable DNA, these reactions were incorporated into the pooled MID-tagged DNA sequencing library. Illumina MiSeq sequencing was performed using a MiSeq Reagent Kit v2 (500 cycles) 250-bp paired-end protocol as per the manufacturer's instruction. Paired-end reads were stitched using Illumina's MiSeq Reporter software.

DNA Sequence Quality Filtering and Analyses

Stitched sequences were separated from unstitched, which had low overall phred scores and were discarded. MID-tags and gene-specific primers were trimmed from the sequences allowing for no mismatch in length or base composition using QIIME1 (extract_barcodes.py) and de-multiplexed using QIIME2 (Caporaso et al., 2010). De-multiplexed reads were then trimmed and filtered using v1.9 (≤6 undetermined bases, quality cutoffs: 5′ = 25 and 3′ = 22) (Martin, 2011). Reads were then error-corrected using deblur (Amir et al., 2017) within QIIME2 (Caporaso et al., 2010), using a custom database of ITS [combining UNITE (Kõljalg et al., 2013), ITS2 (Ankenbrand et al., 2015), and NCBI bioproject: PRJNA177353] and LSU [combining RDP (Cole et al., 2013) and NCBI bioproject: PRJNA51803] sequences, but discarding those below 258 bp for ITS and 240 bp for D2 sequences. Similar sequences to these clustered consensus sequences in the NCBI nt and GSS databases were also found using BLAST + v2.2.6 (blastn). The BLAST databases were filtered before searching to exclude sequences with NCBI taxonomic IDs in the subtree below "12908" (unknown and environmental samples) or containing the keyword "Uncultured" in the sequence name.

Data Analysis

BLAST outputs were analyzed using MEGAN v6.11.2 (Huson et al., 2016) (Weighted LCA, Min Score: 50, Max Expected: 1e-5, Min Percent Identity: 70, Top Percent: 10.0, Min Support Percent: 0.05, Min Support: 1, Percent to cover: 50). Taxonomic lineages for OTUs were found and manipulated using custom scripts[1] based on the NCBI taxonomy database. OTUs with taxonomic assignments were associated with the samples and summarized to generate presence–absence profiles. Presence–absence profiles of fungal taxa were generated for each sample using a cumulative approach, whereby the presence of a low-level taxon also automatically assigned presence to its higher-level taxa. Patterns of variance were investigated using principal components analysis (PCA) using the scikit-learn python package (Pedregosa et al., 2011) as well as by performing canonical correspondence analysis (CCA) using the R package "vegan" (Oksanen et al., 2019). Species richness was also measured using the Shannon–Weaver and Simpson diversity indices in the R package "vegan." To infer the specific identities of OTUs, OTU sequences were searched with BLAST against the custom database developed for this study (see above). BLAST matches were filtered to contain

[1]https://github.com/darcyabjones/acc-to-tax

alignments with greater than 95% identical matches, and covering at least 90% of the OTU sequence.

RESULTS AND DISCUSSION

Crop-Zone Weeds Are Host to a Wide Range of Plant-Associated Fungi

Overall, 2391 and 1819 OTUs were detected from the sequenced ITS and D2 loci, respectively (**Supplementary Table 1**). As an indicator of the specificity of identification, 1231 ITS and 1086 D2 OTUs were directly assigned to NCBI taxonomy identifiers at genus level or lower via BLASTN to Genbank (nt and GSS), while 7 and 15 could not be assigned to any taxonomic identifiers (**Supplementary Table 2**). Most OTUs were assigned to taxa at or below the fungal classes Dothideomycetes, Tremellomycetes, and Sordariomycetes (**Supplementary Table 3**). Species richness was highest in weeds within the Brassicaceae and Polygonacae, followed by other host families that were not dissimilar. Among the fungal species taxa detected in weed samples, the heteroecious rusts of the *Puccinia* spp. (Zhao et al., 2016) (*Puccinia graminis*, *Puccinia bassiae*, *Puccinia coronata*, and *Puccinia malvacearum*) were prominently represented across multiple West Australian weed hosts (**Supplementary Table 3**). These reports of common Australian weeds as alternate hosts for the *Puccinia* rusts are analogous to a growing collection of reports across Europe, the United States, and Asia (Roelfs, 1982; Jin et al., 2010; Wang and Chen, 2013; Zhao et al., 2016). In this study, *P. graminis* was observed on *Bromus*; *P. bassiae* on *Echium* and *Avena*; *P. coronata* on *Bromus* and *Lolium*; and *P. malvacearum* on *Bromus*, *Lolium*, *Avena*, *Echium*, and *Raphanus*. Additional species of agricultural relevance that were identified tended to infect multiple hosts, including the following: *Leptosphaeria maculans*, *Leptosphaeria biglobosa*, *Drechslera campanulata* (syn. *Pyrenophora semeniperda*; ring spot of oat), and *Drechslera nobleae* (syn. *Pyrenophora phaeocomes*, pathogen of rye and pastures). Species endemic to Australia, including several species associated with native eucalypts (*Cladoriella eucalypti, Fusculina eucalypti, Plectosphaera eucalypti, Phacidiella eucalypti, Phaeococcomyces eucalypti*, and *Saccharata eucalyptorum*) (**Supplementary Table 2**), were detected, which were consistent with the local environment but unrelated to crop production. Several genera that contain well-known crop pathogen species were also detected, however, their species-level identities were inconclusive. These included *Cladosporium*, *Zymoseptoria*, *Phaeosphaeria*, *Bipolaris*, *Pyrenophora*, *Venturia*, *Colletotrichum*, *Fusarium*, *Taphrina*, and *Ustilago*. We looked for (using covariance analysis of presence–absence variation) consistent association of two or more species across samples, as well as the inverse mutual exclusion of species (**Supplementary Table 4**), but unlike similar previous studies (Sapkota et al., 2017), we did not observe strong evidence of species complexes or potential bio-control agents across these samples. While culture-independent methods, which are based on the extraction and analysis of nucleic acids, can provide a comprehensive overview of microbial communities, they have drawbacks such as the detection of DNA from dead cells (Müller and Ruppel,

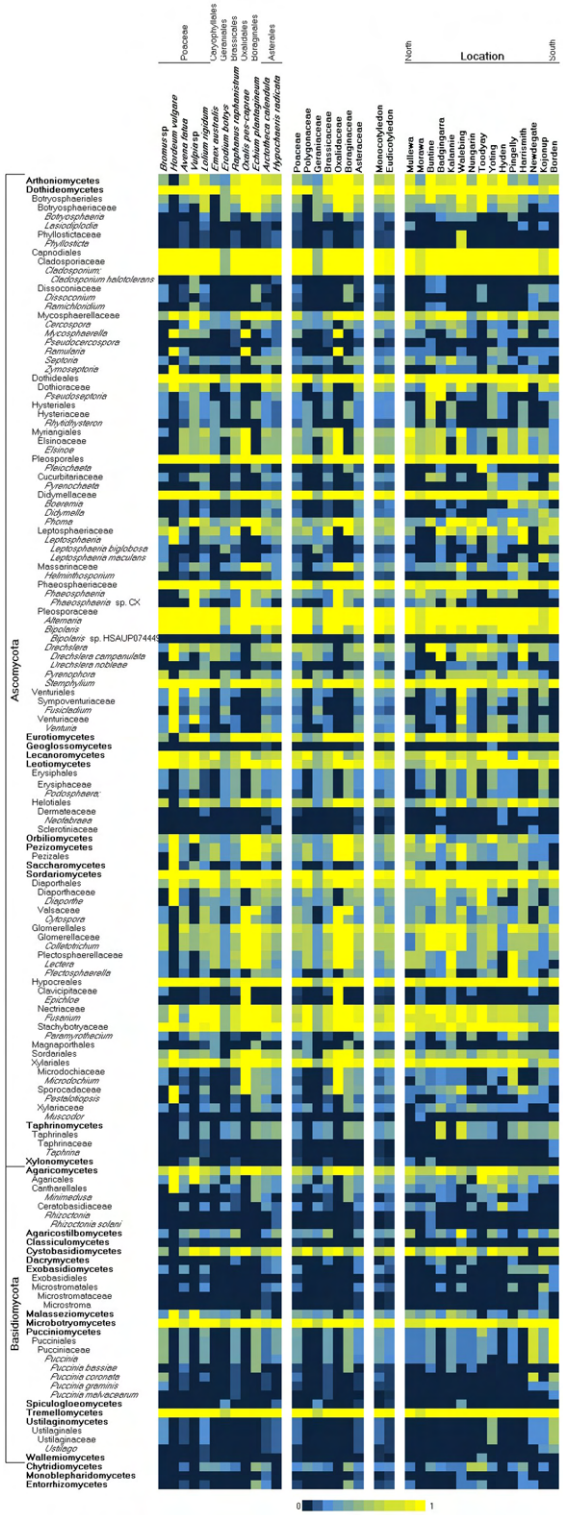


FIGURE 2 | Heatmap of the observed presence of common fungal pathogen genera/species by weed host species and collection location (left/North to right/South). A presence score from 0 to 1 (shown in blue–yellow) indicates the proportion of observed presence of a taxon across all samples of a common species or location, where 0 indicates total absence and 1 indicates ubiquity.

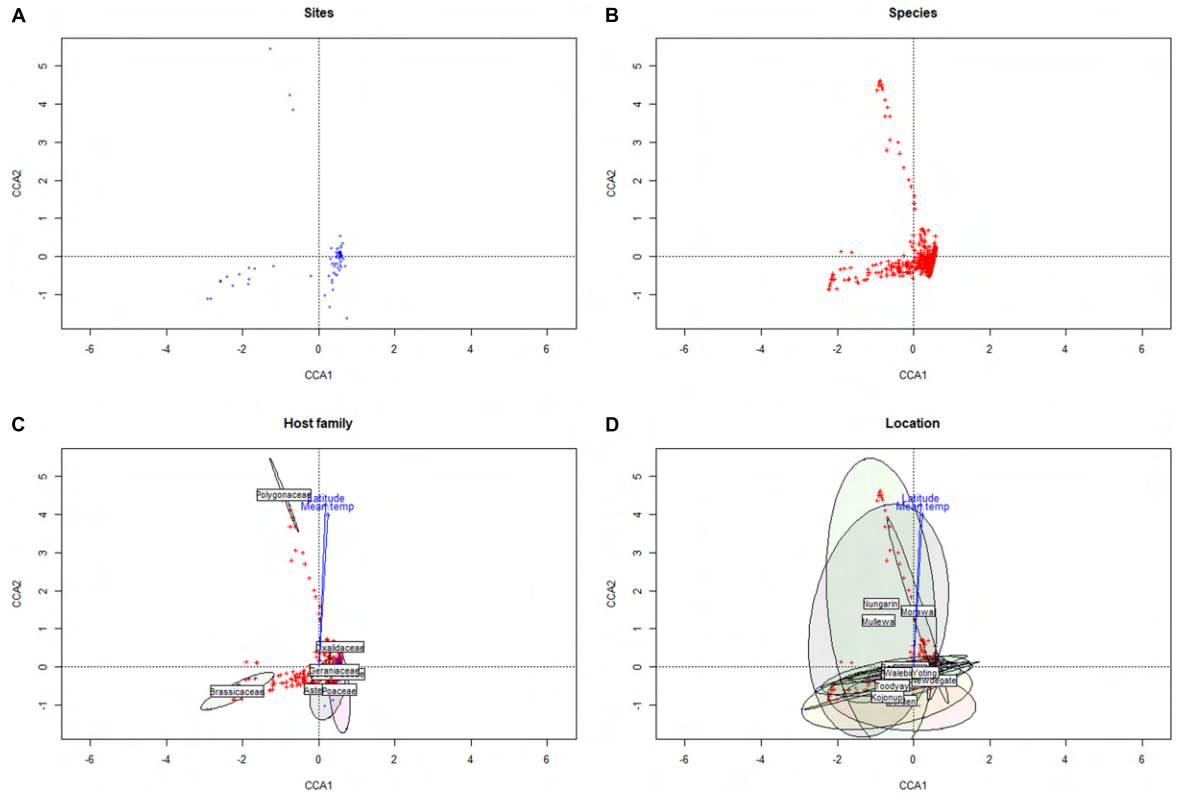


FIGURE 3 | Canonical correspondence analysis of sampled sites **(A)** and microbial taxa **(B)** versus climate conditions and geographic locations, with groupings presented by taxonomic family **(C)** and by location **(D)**.

2014) and dependence on well-annotated databases (Blackwell, 2011; Aylward et al., 2017). It has been reported that up to 20% of fungal sequences in major databases such as GenBank may be misidentified (Bridge et al., 2003; Nilsson et al., 2008). However, the internal transcribed spacer regions (ITS1 and ITS2) have been demonstrated to be taxonomic markers for fungi due to their length and discriminative sequence variation (Lindahl and Kuske, 2013). Nevertheless, an issue we encountered with the taxonomic mapping of OTUs was that mappings to the species level were not consistently achieved. Species level (or lower taxa) would be required in order to reliably assign a crop disease to an OTU and infer its potential host range. Species-level mapping was dependent on multiple factors including representation bias across different taxa in sequence databases, variable map-ability of reads across different taxa, and potential for some species to be present but undetectable. In this study, OTU alignments more often supported the genus-level over species-level mappings; hence, this limited our ability to infer fully the disease risks posed by crop-zone weeds. We have presented a taxonomic summary of genera likely to contain plant-pathogenic species in **Figure 2**, although in most cases, the presence of pathogenic species was inconclusive, as endophytic species may have also been detected.

Regional Biases

Within the Western Australian regions sampled, there appeared to be an overall trend for a relatively increased presence of Ascomycota in northern sample sites and a corresponding relative increase in Basidiomycota (particularly biotrophic rusts of the *Puccinia* spp.) in southern sites (**Figure 2**). Multivariate clustering of samples tended to form groups based on geographic locations rather than by the weed host (**Supplementary File 1**), suggesting that location and/or climate zone was a major factor influencing the weed phyllosphere. This was also supported by CCA (**Figure 3** and **Supplementary File 2**), which indicated that the phyllosphere composition was influenced by the continuous variables, latitude and mean temperature. Species distributions in the CCA corroborated the initial observation of a necrotrophic bias in hotter and dryer northern regions and a corresponding biotrophic bias in cooler and wetter southern regions (**Figure 3** and **Supplementary File 2**). CCA also indicated clustering of hosts of the Polygonaceae family with a strong association with northern latitudes and increased temperatures, which was due to the three *Emex* (syn. doublegee) samples having been exclusively obtained from sites in the northern wheatbelt (Mullewa, Morawa, and Nungarin). However, a host cluster by Brassicaceae (composed of 13 *R. raphanistrum* samples) along CCA1 could not be explained by latitude or temperature vectors and suggests that unique host-specific pathogens may be present on this weed species. The other host families were represented at all sites and did not exhibit any clustering by CCA.

This approximate north-to-south division may reflect that the northern regions in this study are warmer and drier, which may

suppress saprophytic and/or pathogen-suppressive microbial activity that would break down the stubble containing pathogen inoculum—a common problem with cereal necrotrophs (typically of the Ascomycota). Conversely, the southernmost regions of WA are cooler and have increased rainfall, which likely increases the spread and survivability of heteromecic/obligate biotrophic pathogens (typically of the Basidiomycota). Although crop pathogen species were in the relative minority within the phyllosphere, this appears to be consistent to other studies linking climate zones to the distribution of fungal pathogens (Bebber et al., 2014; Fisher et al., 2018).

CONCLUSION

This survey of the weed phyllospheres local to cereal crop-growing regions of Western Australia demonstrates the utility of this approach for the monitoring of plant pathogenic species and could be adapted for the purpose of monitoring for pathogen reservoirs and emerging crop disease risks. We report the presence of several important crop pathogen species within the phyllospheres of local weed hosts and observe that host range and climate zone are both important factors in determining their geographic distributions.

DATA AVAILABILITY STATEMENT

Sequence data was deposited in NCBI BioProject: https://www.ncbi.nlm.nih.gov/bioproject/672330.

AUTHOR CONTRIBUTIONS

PM, JH, and MG contributed to the conception and design of the study. PM conducted the field sampling. NW contributed to DNA extraction and sequencing. DJ and JH contributed to DNA sequence quality filtering and data analysis. PM, JH, DJ, and NW contributed to the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This study was conducted within the Centre for Crop and Disease Management, a cooperative research center funded by the Grains and Research Development Corporation (GRDC) and Curtin University.

ACKNOWLEDGMENTS

This work was supported by resources provided by the Pawsey Supercomputing Centre with funding from the Australian and Western Australian Governments.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fmicb.2020.581592/full#supplementary-material

Supplementary Figure 1 | Fungal species accumulation curves across 10 samples of *Lolium rigidum* and *Raphanus raphanistrum* leaves sampled from Yoting.

Supplementary Table 1 | OTU frequencies for ITS and D2 regions versus host taxonomy and sample locations.

Supplementary Table 2 | Assignment of taxa to OTUs via BLAST.

Supplementary Table 3 | Combined ITS and D2 relative frequencies of microbial taxa versus host taxonomy and sample locations. Each microbial taxon detected by either an ITS or D2 OTU sequence is expressed in terms of a number ranging from 0 to 1, where 1 indicates universal occurrence across host taxa or sample locations, and 0 indicates no occurrences.

Supplementary Table 4 | Co-variance tests for association of taxon pairs across samples. Taxon pairs with high or low co-variance may indicate biologically relevant associations or mutually exclusions across samples in this study.

Supplementary File 1 | Multivariate analyses of microbial taxa testing for relationships versus host species or sample locations.

Supplementary File 2 | Canonical correspondence analyses of microbial taxa testing for relationships between host taxa, pathogen taxa, sample location and climate.

REFERENCES

Amir, A., Mcdonald, D., Navas-Molina, J. A., Kopylova, E., Morton, J. T., Xu, Z. Z., et al. (2017). Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* 2:e0191-16. doi: 10.1128/mSystems.00191-16

Ankenbrand, M. J., Keller, A., Wolf, M., Schultz, J., and Förster, F. (2015). ITS2 database V: twice as much. *Mol. Biol. Evol.* 32, 3030–3032. doi: 10.1093/molbev/msv174

Aylward, J., Steenkamp, E. T., Dreyer, L. L., Roets, F., Wingfield, B. D., and Wingfield, M. J. (2017). A plant pathology perspective of fungal genome sequencing. *IMA Fungus* 8, 1–15. doi: 10.5598/imafungus.2017.08.01.01

Bálint, M., Tiffin, P., Hallström, B., O'hara, R. B., Olson, M. S., Fankhauser, J. D., et al. (2013). Host genotype shapes the foliar fungal microbiome of balsam poplar (*Populus balsamifera*). *PLoS One* 8:e53987. doi: 10.1371/journal.pone.0053987

Bebber, D. P., Holmes, T., and Gurr, S. J. (2014). The global spread of crop pests and pathogens. *Glob. Ecol. Biogeogr.* 23, 1398–1407. doi: 10.1111/geb.12214

Blackwell, M. (2011). The fungi: 1, 2, 3. 5.1 million species? *Am. J. Bot.* 98, 426–438. doi: 10.3732/ajb.1000298

Blixt, E., Olson, Å, Lindahl, B., Djurle, A., and Yuen, J. (2010). Spatiotemporal variation in the fungal community associated with wheat leaves showing symptoms similar to stagonospora nodorum blotch. *Eur. J. Plant Pathol.* 126, 373–386. doi: 10.1007/s10658-009-9542-z

Bridge, P. D., Roberts, P. J., Spooner, B. M., and Panchal, G. (2003). On the unreliability of published DNA sequences. *New Phytol.* 160, 43–48. doi: 10.1046/j.1469-8137.2003.00861.x

Caporaso, J. G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F. D., Costello, E. K., et al. (2010). QIIME allows analysis of high-throughput community sequencing data. *Nat. Methods* 7, 335–336. doi: 10.1038/nmeth.f.303

Cole, J. R., Wang, Q., Fish, J. A., Chai, B., Mcgarrell, D. M., Sun, Y., et al. (2013). Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 42, D633–D642. doi: 10.1093/nar/gkt1244

Cordier, T., Robin, C., Capdevielle, X., Desprez-Loustau, M.-L., and Vacher, C. (2012). Spatial variability of phyllosphere fungal assemblages: genetic distance

predominates over geographic distance in a European beech stand (*Fagus sylvatica*). *Fungal Ecol.* 5, 509–520. doi: 10.1016/j.funeco.2011.12.004

Donovan, P. D., Gonzalez, G., Higgins, D. G., Butler, G., and Ito, K. (2018). Identification of fungi in shotgun metagenomics datasets. *PLoS One* 13:e0192898. doi: 10.1371/journal.pone.0192898

Fisher, M. C., Hawkins, N. J., Sanglard, D., and Gurr, S. J. (2018). Worldwide emergence of resistance to antifungal drugs challenges human health and food security. *Science* 360, 739–742. doi: 10.1126/science.aap7999

Hawksworth, D. L., and Lücking, R. (2017). Fungal diversity revisited: 2.2 to 3.8 million species. *Microbiol. Spectr.* 5, 1–2. doi: 10.1128/microbiolspec.FUNK-0052-2016

Hunter, P. J., Pink, D. A., and Bending, G. D. (2015). Cultivar-level genotype differences influence diversity and composition of lettuce (*Lactuca* sp.) phyllosphere fungal communities. *Fungal Ecol.* 17, 183–186. doi: 10.1016/j.funeco.2015.05.007

Huson, D. H., Beier, S., Flade, I., Górska, A., El-Hadidi, M., Mitra, S., et al. (2016). MEGAN community edition - interactive exploration and analysis of large-scale microbiome sequencing data. *PLoS Comput. Biol.* 12:e1004957. doi: 10.1371/journal.pcbi.1004957

Ihrmark, K., Bödeker, I., Cruz-Martinez, K., Friberg, H., Kubartova, A., Schenck, J., et al. (2012). New primers to amplify the fungal ITS2 region-evaluation by 454-sequencing of artificial and natural communities. *FEMS Microbiol. Ecol.* 82, 666–677. doi: 10.1111/j.1574-6941.2012.01437.x

Jin, Y., Szabo, L. J., and Carson, M. (2010). Century-old mystery of *Puccinia striiformis* life history solved with the identification of *Berberis* as an alternate host. *Phytopathology* 100, 432–435. doi: 10.1094/Phyto-100-5-0432

Joshee, S., Paulus, B. C., Park, D., and Johnston, P. R. (2009). Diversity and distribution of fungal foliar endophytes in New Zealand podocarpaceae. *Mycol. Res.* 113, 1003–1015. doi: 10.1016/j.mycres.2009.06.004

Karlsson, I., Friberg, H., Steinberg, C., and Persson, P. (2014). Fungicide effects on fungal community composition in the wheat phyllosphere. *PLoS One* 9:e111786. doi: 10.1371/journal.pone.0111786

Kinkel, L. L. (1997). Microbial population dynamics on leaves. *Annu. Rev. Phytopathol.* 35, 327–347. doi: 10.1146/annurev.phyto.35.1.327

Kõljalg, U., Nilsson, R. H., Abarenkov, K., Tedersoo, L., Taylor, A. F., Bahram, M., et al. (2013). Towards a unified paradigm for sequence-based identification of fungi. *Mol. Ecol.* 22, 5271–5277. doi: 10.1111/mec.12481

Kurtzman, C. P., and Robnett, C. J. (1998). Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. *Antonie Van Leeuwenh.* 73, 331–371. doi: 10.1023/A:1001761008817

Levetin, E., and Dorsey, K. (2006). Contribution of leaf surface fungi to the air spora. *Aerobiologia* 22, 3–12. doi: 10.1007/s10453-005-9012-9

Lindahl, B. D., and Kuske, C. R. (2013). "Metagenomics for study of fungal ecology," in *The Ecological Genomics of Fungi*, ed. F. Martin (Hoboken, NJ: Wiley-Blackwell), 279–303. doi: 10.1002/9781118735893

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet J.* 17, 10–12. doi: 10.14806/ej.17.1.200

Müller, T., and Ruppel, S. (2014). Progress in cultivation-independent phyllosphere microbiology. *FEMS Microbiol. Ecol.* 87, 2–17. doi: 10.1111/1574-6941.12198

Murray, D. C., Bunce, M., Cannell, B. L., Oliver, R., Houston, J., White, N. E., et al. (2011). DNA-based faecal dietary analysis: a comparison of qPCR and high throughput sequencing approaches. *PLoS One* 6:e25776. doi: 10.1371/journal.pone.0025776

Narayanasamy, P. (2011). "Detection of fungal pathogens in the environment," in *Microbial Plant Pathogens Detection and Disease Diagnosis*, ed. P. Narayanasamy (Cham: Springer), 201–244. doi: 10.1007/978-90-481-9735-4

Nilsson, R. H., Kristiansson, E., Ryberg, M., Hallenberg, N., and Larsson, K.-H. (2008). Intraspecific ITS variability in the kingdom Fungi as expressed in the international sequence databases and its implications for molecular species identification. *Evol. Bioinform.* 4, 193–201. doi: 10.4137/EBO.S653

Oksanen, J., Blanchet, F. G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., et al. (2019). *vegan: Community Ecology Package. R Package Version 2.5-6.* Available online at: https://CRAN.R-project.org/package=vegan (accessed September 11, 2020).

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in python. *J. Mach. Learn. Res.* 12, 2825–2830.

Peršoh, D. (2015). Plant-associated fungal communities in the light of meta'omics. *Fungal Divers.* 75, 1–25. doi: 10.1007/s13225-015-0334-9

Putignani, L., Paglia, M. G., Bordi, E., Nebuloso, E., Pucillo, L. P., and Visca, P. (2008). Identification of clinically relevant yeast species by DNA sequence analysis of the D2 variable region of the 25-28S rRNA gene. *Mycoses* 51, 209–227. doi: 10.1111/j.1439-0507.2007.01472.x

Roelfs, A. P. (1982). Effects of barberry eradication on stem rust in the United States. *Plant Dis.* 66, 177–181. doi: 10.1094/Pd-66-177

Sapkota, R., Jørgensen, L. N., and Nicolaisen, M. (2017). Spatiotemporal variation and networks in the mycobiome of the wheat canopy. *Front. Plant Sci.* 8:1357. doi: 10.3389/fpls.2017.01357

Sapkota, R., Knorr, K., Jørgensen, L. N., O'hanlon, K. A., and Nicolaisen, M. (2015). Host genotype is an important determinant of the cereal phyllosphere mycobiome. *New Phytol.* 207, 1134–1144. doi: 10.1111/nph.13418

Wang, M. N., and Chen, X. (2013). First report of Oregon grape (*Mahonia aquifolium*) as an alternate host for the wheat stripe rust pathogen (*Puccinia striiformis* f. sp. *tritici*) under artificial inoculation. *Plant Dis.* 97, 839–839. doi: 10.1094/Pdis-09-12-0864-Pdn

Weyrich, L. S., Duchene, S., Soubrier, J., Arriola, L., Llamas, B., Breen, J., et al. (2017). Neanderthal behaviour, diet, and disease inferred from ancient DNA in dental calculus. *Nature* 544, 357–361. doi: 10.1038/nature21674

White, T. J., Bruns, T., Lee, S., and Taylor, J. (1990). "Amplification and direct sequencing of fungal ribosomal RNA genes for phylogenetics," in *PCR Protocols: a Guide to Methods and Applications*, eds M. Innis, D. Gelfand, J. Sninsky, and T. J. White (Michigan: Academic Press), 315–322. doi: 10.1016/b978-0-12-372180-8.50042-1

Zhao, J., Wang, M., Chen, X., and Kang, Z. (2016). Role of alternate hosts in epidemiology and pathogen variation of cereal rusts. *Annu. Rev. Phytopathol.* 54, 207–228. doi: 10.1146/annurev-phyto-080615-095851

Zimmerman, N. B., and Vitousek, P. M. (2012). Fungal endophyte communities reflect environmental structuring across a Hawaiian landscape. *Proc. Natl. Acad. Sci. U.S.A.* 109, 13022–13027. doi: 10.1073/pnas.1209872109

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

CHAPTER 8 — THEME 2

Low Amplitude Boom-and-Bust Cycles Define the Septoria Nodorum Blotch Interaction

This chapter is also published in:
Frontiers in Plant Science, 2020, vol. 10, p. 1785
https://doi.org/10.3389/fpls.2019.01785

8.1 Declaration

Title	Low Amplitude Boom-and-Bust Cycles Define the Septoria Nodorum Blotch Interaction.
Authors	Huyen T. T. Phan, **Darcy A. B. Jones**, Kasia Rybak, Kejal N. Dodhia, Francisco J. Lopez-Ruiz, Romain Valade, Lilian Gout, Marc-Henri Lebrun, Patrick C. Brunner, Richard P. Oliver and Kar-Chun Tan
Reference	2020. *Frontiers in Plant Science*, 10, 1785.
DOI	`https://doi.org/10.3389/fpls.2019.01785`

This thesis chapter is submitted in the form of a collaboratively-written and peer-reviewed manuscript. As such, not all work contained within this chapter can be attributed to the Ph. D. candidate.

The following contributions were made by the Ph.D candidate (DABJ) and co-authors:

- K-CT and HTTP conceived and designed the study.
- KD and FL-R collected and isolated fungal samples.
- HTTP, KR, and KD conducted all experiments.
- **DABJ** and HTTP performed genetic and statistical analyses.
- **DABJ**, K-CT, HTTP, PCB, and RPO interpreted and analysed results.
- K-CT and HTTP wrote the manuscript.
- **DABJ**, FL-R, M-HL, RV, PCB, and RO contributed feedback and edited the manuscript.
- All authors read and approved the manuscript.

I, Darcy Jones, hereby certify that this attribution statement is an accurate record of my contribution to the research presented in this chapter.

Darcy Jones

I certify that this attribution statement is an accurate record of Darcy Jones' contribution to the research presented in this chapter.

Huyen T. T. Phan	Kasia Rybak	Kejal N. Dodhia
Francisco J. Lopez-Ruiz	Romain Valade	Lilian Gout
Marc-Henri Lebrun	Richard P. Oliver	Kar-Chun Tan

Patrick Brunner has unfortunately passed.

frontiers in Plant Science

ORIGINAL RESEARCH
published: 31 January 2020
doi: 10.3389/fpls.2019.01785

Low Amplitude Boom-and-Bust Cycles Define the Septoria Nodorum Blotch Interaction

Huyen T. T. Phan[1†], Darcy A. B. Jones[1†], Kasia Rybak[1], Kejal N. Dodhia[1], Francisco J. Lopez-Ruiz[1], Romain Valade[2], Lilian Gout[3], Marc-Henri Lebrun[3], Patrick C. Brunner[4], Richard P. Oliver[1] and Kar-Chun Tan[1]*

[1] Centre for Crop and Disease Management, School of Molecular and Life Sciences, Curtin University, Perth, WA, Australia, [2] ARVALIS Institut du Végétal Avenue Lucien Brétignières, Bâtiment INRA Bioger, Thiverval-Grignon, France, [3] UMR INRA Bioger Agro-ParisTech, Thiverval-Grignon, France, [4] Plant Pathology, Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland

OPEN ACCESS

Edited by:
Andres Mäe, Estonian Crop Research Institute, Estonia

Reviewed by:
Angela Feechan, University College Dublin, Ireland
Graham Robert David McGrann, Science and Advice for Scottish Agriculture (SASA), United Kingdom

***Correspondence:**
Kar-Chun Tan
Kar-Chun.Tan@curtin.edu.au

†These authors have contributed equally to this work

Specialty section:
This article was submitted to Plant Microbe Interactions, a section of the journal Frontiers in Plant Science

Received: *09 August 2019*
Accepted: *20 December 2019*
Published: *31 January 2020*

Citation:
Phan HTT, Jones DAB, Rybak K, Dodhia KN, Lopez-Ruiz FJ, Valade R, Gout L, Lebrun M-H, Brunner PC, Oliver RP and Tan K-C (2020) Low Amplitude Boom-and-Bust Cycles Define the Septoria Nodorum Blotch Interaction. Front. Plant Sci. 10:1785. doi: 10.3389/fpls.2019.01785

Introduction: Septoria nodorum blotch (SNB) is a complex fungal disease of wheat caused by the Dothideomycete fungal pathogen *Parastagonospora nodorum*. The fungus infects through the use of necrotrophic effectors (NEs) that cause necrosis on hosts carrying matching dominant susceptibility genes. The Western Australia (WA) wheatbelt is a SNB "hot spot" and experiences significant under favorable conditions. Consequently, SNB has been a major target for breeders in WA for many years.

Materials and Methods: In this study, we assembled a panel of 155 WA *P. nodorum* isolates collected over a 44-year period and compared them to 23 isolates from France and the USA using 28 SSR loci.

Results: The WA *P. nodorum* population was clustered into five groups with contrasting properties. 80% of the studied isolates were assigned to two core groups found throughout the collection location and time. The other three non-core groups that encompassed transient and emergent populations were found in restricted locations and time. Changes in group genotypes occurred during periods that coincided with the mass adoption of a single or a small group of widely planted wheat cultivars. When introduced, these cultivars had high scores for SNB resistance. However, the field resistance of these new cultivars often declined over subsequent seasons prompting their replacement with new, more resistant varieties. Pathogenicity assays showed that newly emerged isolates non-core are more pathogenic than old isolates. It is likely that the non-core groups were repeatedly selected for increased virulence on the contemporary popular cultivars.

Discussion: The low level of genetic diversity within the non-core groups, difference in virulence, low abundance, and restriction to limited locations suggest that these populations more vulnerable to a population crash when the cultivar was replaced by one that was genetically different and more resistant. We characterize the observed

pattern as a low-amplitude boom-and-bust cycle in contrast with the classical high amplitude boom-and-bust cycles seen for biotrophic pathogens where the contrast between resistance and susceptibility is typically much greater. Implications of the results are discussed relating to breeding strategies for more sustainable SNB resistance and more generally for pathogens with NEs.

Keywords: septoria nodorum blotch, SSR, effector, population, wheat

INTRODUCTION

Parastagonospora nodorum causes septoria nodorum blotch (SNB) of wheat (*Triticum* spp.) (Solomon et al., 2006). The fungus causes significant yield losses in many wheat growing regions world-wide (Eyal et al., 1987; Loughman, 1989; Oliver et al., 2009). *P. nodorum* reproduces sexually before the start of the growing season forming air-borne ascospores that can disperse over long distances and asexually throughout the growing season for short distance rain-splash dispersal (Eyal et al., 1987). Mixed reproduction allows the rapid production of adapted recombinant genotypes *via* sexual reproduction and their fixation and multiplication in the asexual phases (Brasier et al., 1999; McDonald and Linde, 2002).

The development of SNB is dictated by interactions between several proteinaceous necrotrophic effectors (NEs) secreted by *P. nodorum* and dominant susceptibility genes in the host (Friesen and Faris, 2010; Oliver et al., 2012). A compatible interaction results in host tissue death and disease. Three NE genes have been identified and fully characterised for their role in SNB. *ToxA* encode a protein that causes necrosis on wheat varieties that carry the dominant susceptibility gene *Tsn1* (Friesen et al., 2006). Two other wheat pathogens also contain *ToxA*; *Pyrenophora tritici-repentis* (*Ptr*) (Ciuffetti et al., 1997) and *Bipolaris sorokiniana* (*Bs*) (McDonald et al., 2018). A global survey of *P. nodorum* isolates identified 17 *SnToxA* haplotypes encoding nine unique protein isoforms (McDonald et al., 2013). *Tox1* encodes a cysteine-rich protein with a chitin binding-like motif at the carboxyl terminus. Tox1–induced chlorosis requires the expression of the host sensitivity gene *Snn1* located on wheat chromosome 1BS (Liu et al., 2004; Liu et al., 2012). *Tox3* encodes a small cysteine-rich protein and sensitivity is controlled by *Snn3* which is located on wheat chromosome 5BS (Shi et al., 2016). These effectors possess multiple protein isoforms (McDonald et al., 2013).

SNB remains a problematic disease worldwide despite a greater understanding on the role of NEs in the establishment of disease in the *P. nodorum* — wheat pathosystem. In Australia, host resistance in current cultivars is partial at best (Zaicou-Kunesch et al., 2018). The degree of resistance is not a simple reflection of the number of matching NEs in the fungus and susceptibility genes in host cultivars. Molecular genetic analyses have shown that SNB is impacted by the complex interactions of multiple wheat QTLs (Shankar et al., 2008; Friesen and Faris, 2010; Tan and Oliver, 2017). SNB is further complicated by differences in effector isoform activity. For *ToxA* at least, protein isoforms exhibit large differences in necrosis-inducing activities (Tan et al., 2012). Isoforms that induced the most rapid necrosis on *Tsn1* wheat caused the most fungal sporulation.

The population structure of the Western Australia (WA) *P. nodorum* population has been examined in several studies using various genetic markers. Murphy et al. (2000) analyzed isolates collected over 13 years using four restriction-fragment-length-polymorphism probes and observed a high level of genetic diversity but no distinct subpopulations. Stukenbrock et al. (2006) examined the global migration pattern of the pathogen using EST-SSR markers (Stukenbrock et al., 2005). A high global migration rate of *P. nodorum* was observed, with the Australian population acting as a sink for foreign immigrants. Like Murphy et al. (2000); Stukenbrock et al. (2006) also observed a high level of genotypic diversity within the Australian *P. nodorum* population even though it was sampled from a single field in 2001.

The aim of this study was to undertake a more complete analysis of the genetic structure of the population using a set of isolates collected over a 44-year period. Such a collection collected over a long-time span is extremely rare and mainly restricted to well-studied model fungi like *Puccinia triticina* (Ordonez and Kolmer, 2009) and *Neurospora crassa* (Turner et al., 2001). Another objective was to investigate differences or trends in virulence of the structured populations and seek for their explanations. By combining pathogen population genetic data with data on the cultivation of wheat, we hoped to shed light on the microevolutionary process operating in time and space.

MATERIALS AND METHODS

Fungal Reisolation

A set of 155 Australian *P. nodorum* isolates collected in WA between 1972 and 2016 was used in this study (**Table S1**, **Figure 1**). Fifty-five were obtained from the Department of Primary Industries and Regional Development (DPIRD). A further 100 were isolated from leaf samples with SNB symptoms using two following methods: the traditional method obtained 73 single pycnidial isolations as previously described (McDonald et al., 1994); and the second method used 10 μg/ml boscalid to suppress *Ptr*. To do this, the infected leaf was embedded onto tap water agar supplemented with 10 mg/L boscalid plus 100 mg/L ampicillin, 30 mg/L streptomycin, 50 mg/L neomycin sulphate to suppress bacterial growth. Growing hyphae were then transferred onto V8PDA agar medium supplemented with the same antibiotics and grown until pycnidiation to generate pure

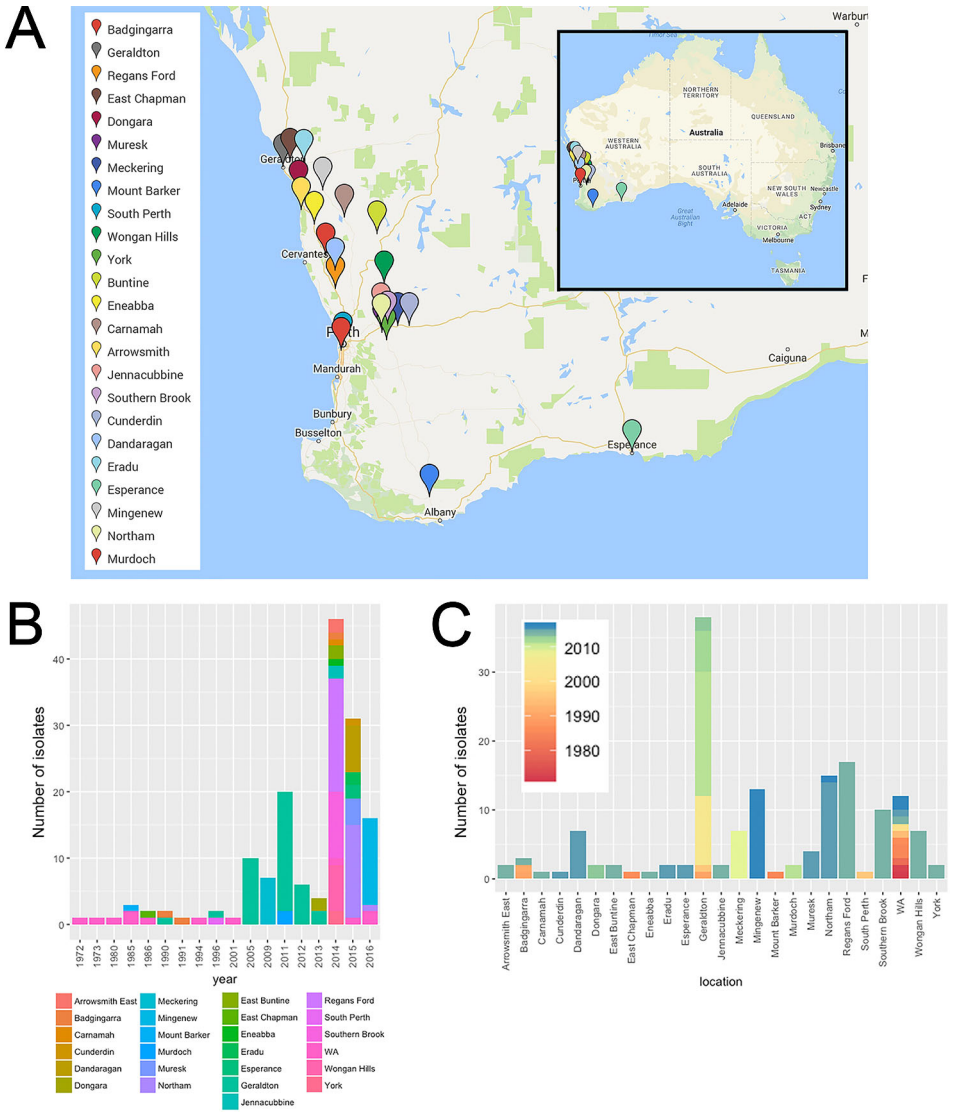


FIGURE 1 | Spatial-longitudinal distribution of Australian *P. nodorum* isolates collected between 1972-2016: **(A)** Sampling sites (Google Maps, CA, USA) isolates with unknown collection location in WA were designated as "WA" and were not included in the map; **(B)** number of isolates per sampling year; **(C)** per sampling location.

cultures. To test whether this isolation method biased towards fungicide tolerant isolates, the EC_{50} concentration of boscalid was determined using a microtitre plate assay (Mair et al., 2016).

Eighteen *P. nodorum* isolates from France isolated as previously described (McDonald et al., 1994) one from Denmark and four from the USA were also included in the study. One *Ptr*, one *Parastagonospora avenae* f.sp. *triticae* (*Pat*), one *Phoma* sp. (15FG039), two *Pyrenophora teres* f. sp. *maculata* (*Ptm*), and one *Pyrenophora teres* f. sp. *teres* (*Ptt*) isolates were included in the study as phylogenetically distinct outgroup controls.

SSR Marker Design

Twelve SSR markers designed from *P. nodorum* ESTs were tested (Stukenbrock et al., 2005). Only SNOD1, SNOD23, SNOD3, SNOD5, and SNOD8 demonstrated consistent amplification and were subsequently used in this study. We designed 25 SSR primer pairs (**Table S2**) from the 20 largest genomic scaffolds of the *P. nodorum* SN15 assembly to increase genome-wide marker coverage (Syme et al., 2016).

Nei's diversity index (H_{exp}) was used to test for the effect of each marker used in this study (Nei, 1978). The genetic diversity of the *P. nodorum* collection was calculated omitting each of the

markers in turn to assess the effect of each SSR marker. H_{exp} indicated that none of these markers gave biased genetic diversity values as they were in a similar range and compatible to that derived from the four EST-SSR previously published markers (SNOD1, SNOD3, SNOD5, and SNOD8).

SSR Genotyping

SSR genotypes were obtained using a multiplex-ready PCR technique (Hayden et al., 2008). PCR products were subsequently electrophoresed using an ABI-PRISM 3730xl automated sequencer (Applied Biosystems, CA, USA). PCR amplicon lengths were determined by GeneMarker V1.91.

Population Analyses

Amplicon lengths representing SSR alleles were rounded to the nearest multiple of the repeat length and analysed using the R statistical programming language (Development Core Team R, 2008). To assess the number of loci required to discriminate individuals in the population, genotype accumulation curves were calculated using poppr v2.6.0 (Kamvar et al., 2014). Bootstrapped (n = 100) phylogenetic trees were estimated using Bruvo's distance (Bruvo et al., 2004) and unweighted-pair-group-method-with-arithmetic-mean (UPGMA) hierarchical clustering algorithm in poppr. Trees were rooted using four *Pyrenophora* sp., *P. avenae* and a *Phoma* sp. as outgroups.

Summary statistics including: allele frequency, Simpson's index (λ), H_{exp}, and evenness for SSR loci were obtained using poppr (Grunwald et al., 2003). Uninformative markers (i.e., those with < 2 variant isolates or where 99% of isolates possess one major allele) were identified using poppr and removed from further analysis.

SSR variation between isolates was summarised using principal component analysis (PCA) from the ade4 and adegenet R packages (Chessel et al., 2004). Isolates were clustered based on SSR data using snapclust (Jombart et al., 2010).

Relationships between the clustered groups of isolates identified by snapclust and the alleles contributing to their separation were dissected using discriminant analysis of principal components (DAPC) (Jombart et al., 2010) in adegenet for the Australian population. Given a set of isolates with assigned groups, DAPC finds linear functions of PCs of the SSR genotype data that maximise differences between groups and minimise differences within groups. The number of PCs to include in the model was determined by performing cross-validation DAPC using a range of PC retention numbers each cross-validated 50 times (90%:10% training: validation stratified splits) so that it minimised the mean squared error of reclassification. Alleles with loading values above 0.03 were considered to have a major influence in discriminating isolate groups. Posterior probabilities of group membership resultant from DAPC models were used to assess groupings, detect possible admixture between groups and identify anomalous isolates.

For each identified population, the number of multilocus genotypes (MLG) observed, number of expected MLG, Shannon-Wiener Index of MLG diversity (Shannon, 1997), Stoddart and Taylor's index of MLG diversity (Stoddart and Taylor, 1988), λ index (Simpson, 1949) and H_{exp} (Nei, 1978) were calculated using poppr (Kamvar et al., 2014). These parameters were calculated taking into account differences in sample sizes between these groups. A clone-corrected dataset was generated by retaining a representation of duplicated isolates identical in SSR genotypes. The Prevosti's distance model-free method was used to calculate the genetic variation between groups observed from the Australian *P. nodorum* panel (Prevosti et al., 1975).

Determination of the Mating Genotype and Index-of-Association (I_A) in *P. nodorum* Isolates

Mating types in *P. nodorum* isolates were determined by PCR amplification of genomic DNA with specific primers MAT1-1/2 and MAT2-1/2 (Bennett et al., 2003). A Chi square test (χ^2) was used to determine if the observed *MAT1*:*MAT2* ratio departed from the standard 1:1. I_A, rbarD ($\hat{r}_d$) and I_A test which uses 999 permutation of the data were deployed to determine if populations are in linkage disequilibrium (Smith et al., 1993) using poppr (Kamvar et al., 2014).

Commercial Wheat Varieties and Disease Rating

SNB resistance ratings of wheat cultivars rated from 1990 until 2018 were obtained from the DPIRD historical archive as 'Crop Variety Guides' (https://www.agric.wa.gov.au/library) and was converted to a numerical system for calculations: 10, very susceptible; 9, susceptible-very susceptible; 8, susceptible; 7, moderately susceptible-susceptible; 6, moderately susceptible; 5, moderately resistant-moderately susceptible/intermediate.

The *P. nodorum* collection period into three eras corresponding to major shifts in wheat cultivar adoption. Crop planting data was obtained from DPIRD as "Crop Variety Guides" (https://www.agric.wa.gov.au) based on CBH Group reports (https://www.cbh.com.au). Coefficient correlation was determined independently for each period based on the P values of the test between frequency of each groups and frequency for all top-grown wheat lines for Era I (1984–2001), II (2001–2013) and III (2013–present).

Whole Plant Seedling Infection Assay

Three isolates from each group (Group 1: FG63, MuS3, FG49; Group 2: Meck1, WAC13077, WAC13955; Group 3: WAC8635, WAC13418, SN15; Group 4: WAC13404, WAC13632, WAC13525; Group 5: 16FG165, 16FG167, 16FG168) were chosen for seedling infection assay on seven top-grown wheat lines from the three eras (Era I: Halberd, Eradu, Cadoux; Era II: Carnamah, Calingiri, Wyalkatchem; Era III: Mace). The disease assay for seedling plants was carried out following the method described in Solomon et al. (2003). Briefly pycnidiospore suspension was prepared to a concentration $1x10^6$ spores/ml in 0.05% (w/v) gelatine. All wheat lines were planted in a randomised design in three replicates. Two-week-old seedlings were sprayed with the spore-suspensions until runoff and kept

under 100% relative humidity at 21°C under a 12-h photoperiod for 48 h. The plants were kept moist using a fine misting system in the growth chamber for five more days before disease symptom was scored in a scale from 1 to 9. A score of one indicates no disease symptoms whereas a score of nine indicates a fully necrotised plant. Significant differences in disease scores for each group in each era and their interactions were determined based on analysis of variance (ANOVA) and Tukey's Post Hoc test using statistical functions in core R and agricolae v1.3 R package (https://cran.r-project.org/web/packages/agricolae). Each isolate was inoculated on the individual cultivars with three replicates and then the data put together for variance analysis between eras.

RESULTS

Isolation and Assembly of the *P. nodorum* Isolate Panel

A collection of 155 WA (this study; Solomon et al., 2004b; Syme et al., 2018) and 23 non-Australian [this study; (Friesen et al., 2006; Syme et al., 2018; Richards et al., 2019)] *P. nodorum* isolates, plus single isolates of five closely-related species were used in this study (**Figure 1**). WA isolates were collected from 24 known locations across the WA wheat belt. Most isolates were collected from 2005 onwards comprising of 93.6% of the entire WA collection. Isolates from Geraldton, collected between 1990 and 2013, is the largest regional group which comprised of 24.5% of isolates in the WA collection. The 12 WA isolates collected between 1972 and 2016 do not have sampling site information and were assigned as "unknown WA" (**Figure 1**; **Table S1**). Through the sampling process, we developed a new method to isolate *P. nodorum* from fresh infected leaf tissues using 10 μg/ml boscalid (**Figure S1**). This method permitted the separation of co-infecting *Ptr* (**Figure S1**, **Table S3**).

SSR Marker Development

We developed 25 new microsatellite loci from the genome sequence of the *P. nodorum* reference isolate SN15 and added 5 EST_SSR markers from Stukenbrock et al. (2005). A genotype-accumulation-curve indicated that 27 loci developed in this study and from Stukenbrock et al. (2005) was sufficient to capture all observed genetic diversity existing in the Australian *P. nodorum* collection (**Figure S2A**). Two SSR markers (SSR3 and SNOD23) were found to be uninformative within *P. nodorum* isolates and were omitted from further analysis. The genetic diversity of the *P. nodorum* collection was calculated omitting each of the markers in turn to assess the effect of each SSR marker. H_{exp} indicated that none of these markers gave biased genetic diversity values (**Figure S2B**) as they were in a similar range and compatible to those derived from the four previously published EST-SSR markers (SNOD1, SNOD3, SNOD5, and SNOD8).

A clonality test of the Australian *P. nodorum* isolates revealed two pairs with identical SSR profiles (1. 16FG161 and 16FG162; 2. 16FG169 and 16FG170). Isolates 16FG162 and 16FG170 were removed, thus retaining 153 non-clonal isolates for further analyses.

TABLE 1 | Number of alleles detected, λ index, H_{exp}, and evenness for each SSR marker.

Marker	Allele number	λ	H_{exp}	Evenness
SNOD1	13	0.28	0.28	0.33
SNOD23	4	0.04	0.04	0.32
SNOD3	5	0.47	0.47	0.80
SNOD5	10	0.64	0.64	0.67
SNOD8	10	0.33	0.34	0.45
SSR1	11	0.82	0.83	0.82
SSR3	1	0.00	0.00	NA
SSR4	22	0.85	0.85	0.57
SSR5	9	0.77	0.77	0.74
SSR6	7	0.33	0.33	0.53
SSR7	20	0.87	0.88	0.66
SSR8	8	0.73	0.74	0.72
SSR9	19	0.78	0.78	0.60
SSR10	16	0.86	0.86	0.74
SSR12	12	0.77	0.78	0.68
SSR14	33	0.91	0.92	0.61
SSR15	14	0.87	0.88	0.78
SSR16	26	0.93	0.94	0.82
SSR17	18	0.77	0.77	0.52
SSR18	11	0.72	0.72	0.64
SSR19	16	0.51	0.51	0.44
SSR20	6	0.63	0.64	0.82
SSR21	27	0.92	0.93	0.72
SSR22	13	0.78	0.79	0.59
SSR23	14	0.87	0.88	0.79
SSR24	14	0.66	0.66	0.52
SSR25	11	0.68	0.68	0.54
SSR26	10	0.29	0.29	0.35
SSR27	22	0.90	0.91	0.70
SSR28	15	0.77	0.78	0.61
Mean	13.90	0.66	0.66	0.62

The properties of the 28 informative SSR markers were calculated from the clone-corrected panel (**Table 1**). An average of 14.7 alleles/locus were observed. Four loci (SSR14, SSR16, SSR21 and SSR27) had a high λ index (> 0.90). Overall genetic diversity with an average H_{exp} of 0.71 was reported (**Table 2**).

An UPGMA tree was generated to examine the phylogenetic relationship between all fungal isolates (**Figure 2**). The *P. nodorum* isolates clustered away from the other species confirming the validity of the species. Amongst the *P. nodorum* isolates, most Australian *P. nodorum* isolates were connected by relatively long branches, indicating a high level of genetic diversification. Apart from two isolates (US isolate

TABLE 2 | Gene, genotypic diversity and linkage disequilibrium of Australian *P. nodorum* isolates and its associated discriminant analysis of principal components (DAPC) groups.

Group	*n*	$\hat{G}$	λ	H_{exp}
All	153	98.71	0.99	0.71
Group 1	66	100.00	0.98	0.67
Group 2	56	100.00	0.98	0.71
Group 3	7	100.00	0.86	0.31
Group 4	14	87.50	0.92	0.20
Group 5	10	100.00	0.90	0.09

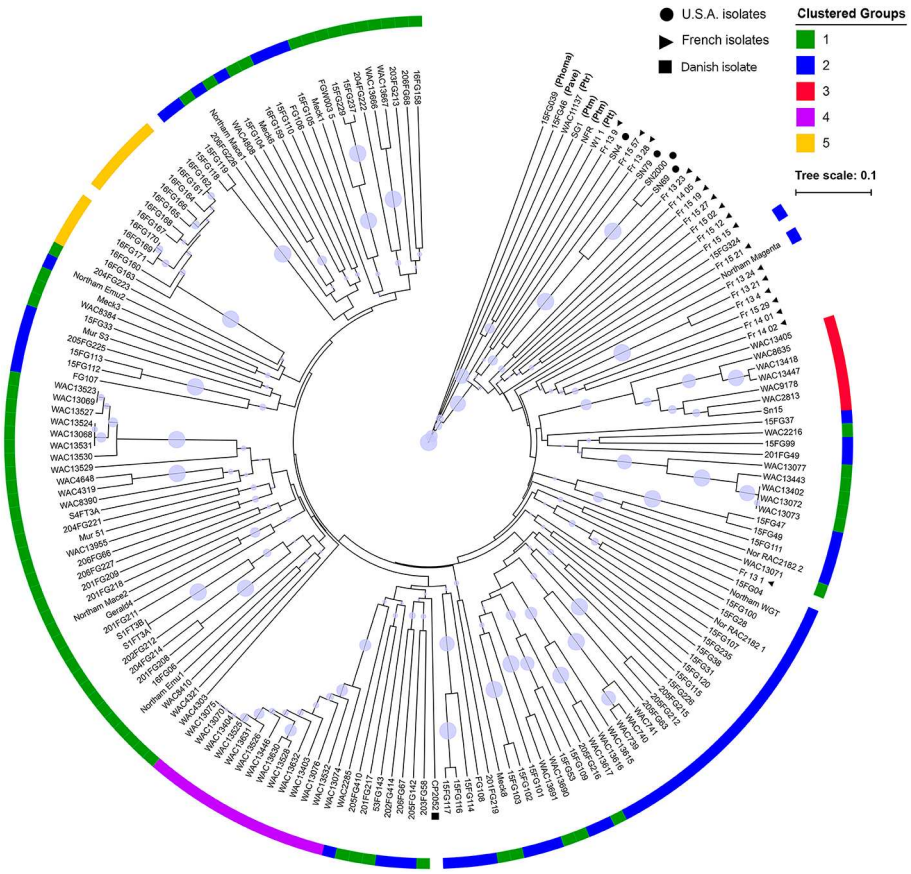


FIGURE 2 | A UPGMA tree of 184 P. *nodorum* and other species constructed using Bruvo's distance with non-parametric bootstrapping. Non-shading blocks indicate the absence of effector genes (**Table S1**). Clades with bootstrap support < 30% are collapsed into polytomies. Branch bubbles represent bootstrap support for the remaining clades (30%–100% support). Clustered groups were constructed from DAPCs. Non-*P. nodorum* isolates have their abbreviated species names in brackets. Isolates from France, USA and Denmark are indicated by black circles, triangles and a square, respectively. No group assignment was done for the isolates 16FG162 and 16FG170 as they were removed from the cluster analysis due to clonality.

CP2052 and French isolate Fr_13_1) that clustered with the Australian, all the non-Australian isolates were located on distinct clades (**Figure 2**).

Evidence of Core and Transient Populations in the Australian *P. nodorum* Panel

PCA revealed five groups were observed in the scatterplot built from PC1 (7.4% of the variance) and PC2 (6.6% of the variance) (**Figure 3A** and **Table S4**). The low levels of variance explained by PC1 and PC2 were possibly due to the large and complex diversity that exists in the collection, it is therefore necessary to use snapclust and supervised DAPC methods with multiple PCAs for further analysis. Snapclust analysis divided the population into five groups as presented in **Figure 3B**. The groups were further interrogated using DAPC. The DAPC models were evaluated by excluding isolates from training and predicting their group assignment (cross-validation). The level of correct reassignment was high for the selected five *P. nodorum* group model, with the maximum mean of successful prediction (0.983) obtained using the first 10 PCs only (**Figure 3C**). All subsequent DAPC analyses used the 10 PCs (explaining 39.3% of the total variance).

Comparative analysis between groups using DAPC indicated that Groups 1 and 2 were closely related, whereas Groups 3, 4, and 5 were distant from each other and from Groups 1 and 2 (**Figure 3D**). DAPC retained the five-group structure with only two isolates (15FG33 and WAC2285) reassigned between Group 1 (n = 56) and 2 (n = 66). Isolates assigned to Group 3 (n = 7), 4 (n = 14) and 5 (n = 10) using snapclust remained unchanged with DAPC.

80% of isolates were members of groups 1 and 2 and were embedded throughout the phylogenetic tree, while groups 3, 4, and 5 formed distinct clades (**Figure 2**). Overall genetic diversity within Groups 1 and 2 was high (mean H_{exp} = 0.69), but lower within Groups 3, 4, 5 (H_{exp} = 0.31, 0.20, and 0.09) (**Table 2**). Assessment of the genetic distance between *P. nodorum* groups using the Prevosti's model-free method indicates a higher level of genetic similarity between Groups 1 and 2 (0.39) compared with

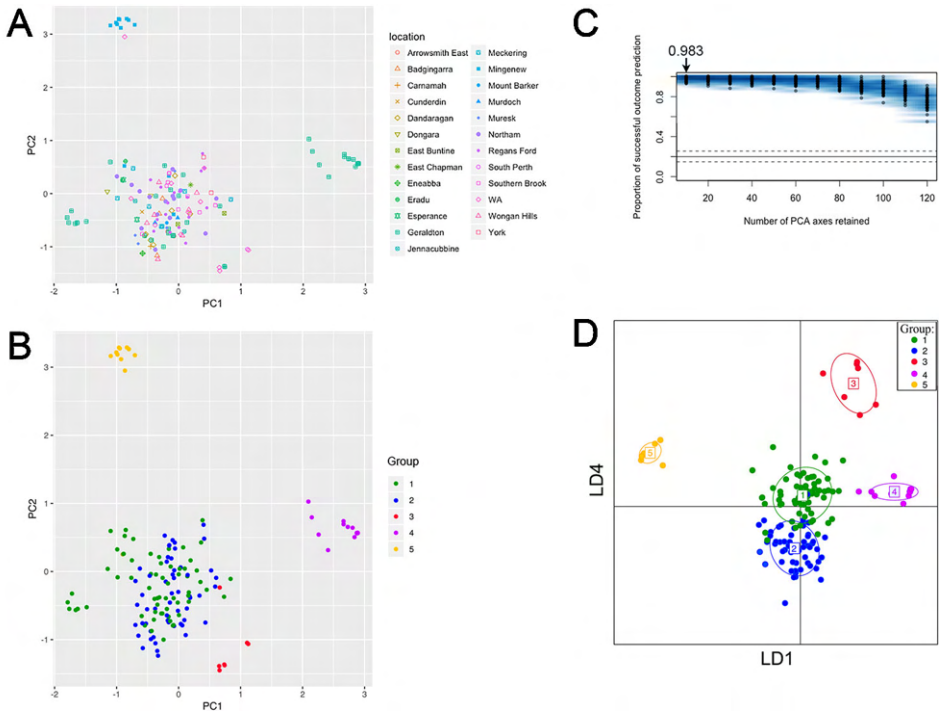


FIGURE 3 | PC and DAPC analyses of population structure among 153 clone-corrected Australian *P. nodorum* isolates. First two PCs are shown where each point represents an isolate. **(A)** Sampling locations are indicated. **(B)** Isolates presented by genetic groups based on unbiased maximum-likelihood genetic clustering. **(C)** DAPC cross-validation test was performed for 10 to 120 PCs. Optimal number PCs to achieve the highest proportion of correct prediction outcome for the five-group model is shown. **(D)** PC1 to PC10 were DAPC-transformed to generate a simulated scatterplot displayed as LD1 and LD4 functions.

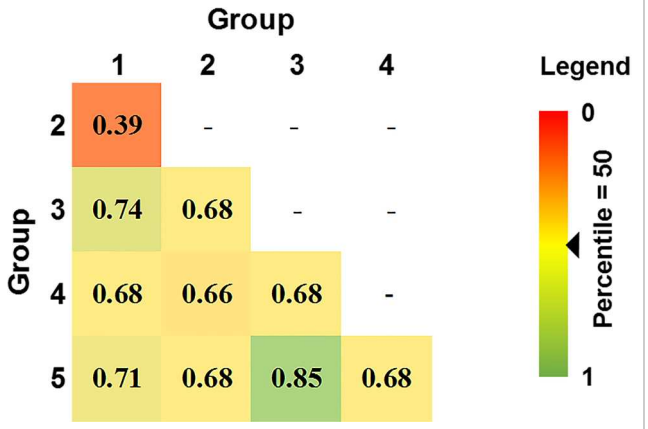


FIGURE 4 | Pair-wise genetic distances between five DAPC Australian *P. nodorum* groups using Prevosti's distance. Genetic distances are ranked by percentile score. 0 and 1 denotes 0 and 100th percentile, respectively.

Groups 3, 4, 5 (0.66 to 0.85) (Prevosti et al., 1975) (**Figure 4**). The high level of genetic distances indicated limited gene flow between Groups 1/2 and Groups 3/4/5.

When the distributions of each group across sampling locations (**Figure 5A**) was considered, Groups 1 and 2 comprised isolates collected from most sampling locations. Group 3 consisted isolates sampled from two known locations (South Perth and Geraldton) and three isolates from unknown locations in WA. All Group 4 members derived from Geraldton whereas most Group 5 isolates were sampled from Mingenew. We then determined if the population structure was subjected to shifts over time (**Figure 5B**). Group 1 consisted of isolates sampled between 1972 and 2016. Group 2 consisted of isolates collected between 2011 and 2015. Members of Group 3 were found between 1994 and 2011. Isolates from Group 4 were sampled between 2005 and 2011. All isolates belonging to Group 5 were sampled in 2016. We concluded that Group 1 is ubiquitous in time and space and form the core Australian *P. nodorum* population whereas Groups 3 and 4 are transient and limited to specific time and locations. Groups 2 and 5 consisted of emerging isolates that became prominent since 2014 however, isolates belonging to the former group are much more prevalent from samples collected. Therefore, it can be concluded that Groups 1 and 2 form the core populations whereas Groups 3, 4, and 5 are considered smaller non-core populations. It can further be deduced that Groups 1 and 5 are emergent populations whereas Groups 3 and 4 are considered transient.

A DAPC analysis of the Australian isolate collection was also performed with non-Australian *P. nodorum* isolates using geographically defined groupings. It was observed that

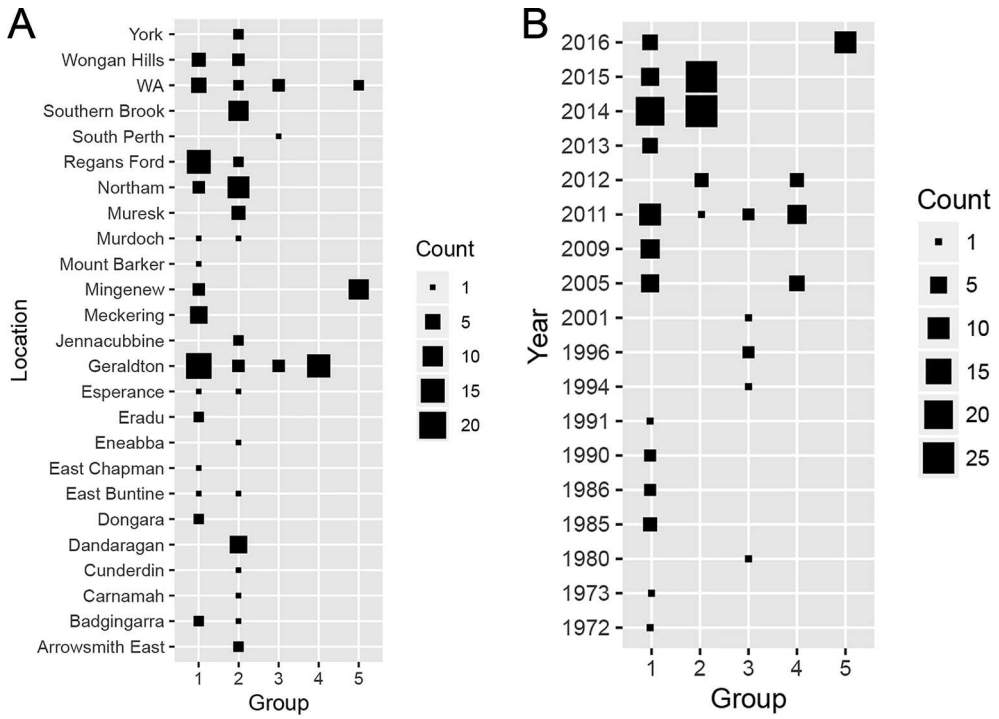


FIGURE 5 | Distribution of discriminant analysis of principal components (DAPC)-grouped Australian *P. nodorum* isolates over sampling **(A)** locations and **(B)** times.

TABLE 3 | I_A, rbarD and mating type assignments of the Australian *P. nodorum* groups.

Population	N	I_A	rbarD	*P* (rbarD)	*MAT1-1*	*MAT1-2*	*P* (χ2)
Group 1	66	0.83	0.03	0.001	32	33	0.90
Group 2	56	0.61	0.02	0.001	25	31	0.42
Group 3	7	1.72	0.10	0.004	7	0	0.01
Group 4	14	1.80	0.11	0.001	0	14	0.00
Group 5	10	0.07	0.01	0.371	10	0	0.00
Total	153	0.99	0.04	0.001	74	78	0.75

*rbarD is the standardised I_A denoted as $\bar{r}_d$ (**Figure S4**).*

non-Australian isolates are distinct from Australian isolates (**Figure S3**).

Determination of Mating Type Distribution and I_A

We used PCR to assign mating types to our panel. A χ^2 test demonstrated that the distribution of *MAT1-1* and *MAT1-2* across Groups 1 and 2 fitted the 1:1 ratio (**Table 3**). In contrast, isolates from Groups 3, 4, 5 had only a single mating type; Groups 3 and 5 carried *MAT1-1* while Group 4 carried *MAT1-2* (**Table 3**).

I_A was determined for all groups of *P. nodorum* and rbarD significance test were used to determine if a population is in linkage disequilibrium. For this test, $P < 0.05$ would reject the null hypothesis of no linkage among markers (**Figure S4**). Difference to the mating type ratio and the rbarD significant test revealed that all Australian *P. nodorum* groups except Group 5 were in linkage disequilibrium signifying that the assortative recombination level among these groups is not high (Goss et al., 2014). Finding from this analysis indicated that sexual reproduction possibly only happened in Group 5 whereas asexual sporulation likely to be the main mode of reproduction in the other four groups.

Shifts in *P. nodorum* Population Structure Correspond to Wheat Cultivar Adoption

In an attempt to rationalise the *P. nodorum* population structure, we compared it to the prevalence of wheat cultivars where data was available (**Figures 6A, B**). Only wheat cultivars with ≥10% of the area sown in any one year were selected for analysis. We divided the period 1984 to date into three eras based on changes in cultivar adoption. Cultivars Eradu, Halberd, Calingiri dominated Era I. Carnamah, Spear, and Wyalkatchem defined Era II whereas Era III has a remarkable predominance of a single cultivar Mace. Groups 3 and 4 isolates were found in Eras I and II whereas Group 5 only emerged in Era III. A correlation coefficient test indicated significant associations between the emergence of Groups 3, 4, 5 and the frequency of wheat cultivar adoption in Era I, II and III ($r = 0.89$, df = 29, p-value = $3.01e^{-11}$ for wheats in Era I and Group 3; $r = 0.83$, df = 29, $P = 9.22e^{-9}$ for wheats in Era II and Group 4; $r = 0.83$, df = 29, $P = 9.15e^{-9}$ for wheats in Era III and Group 5).

We then examined the SNB resistance rating of popular wheat cultivars between 1990 and 2018 (**Figure 6C**) as determined by official DPIRD trials. Popular cultivars in Era I

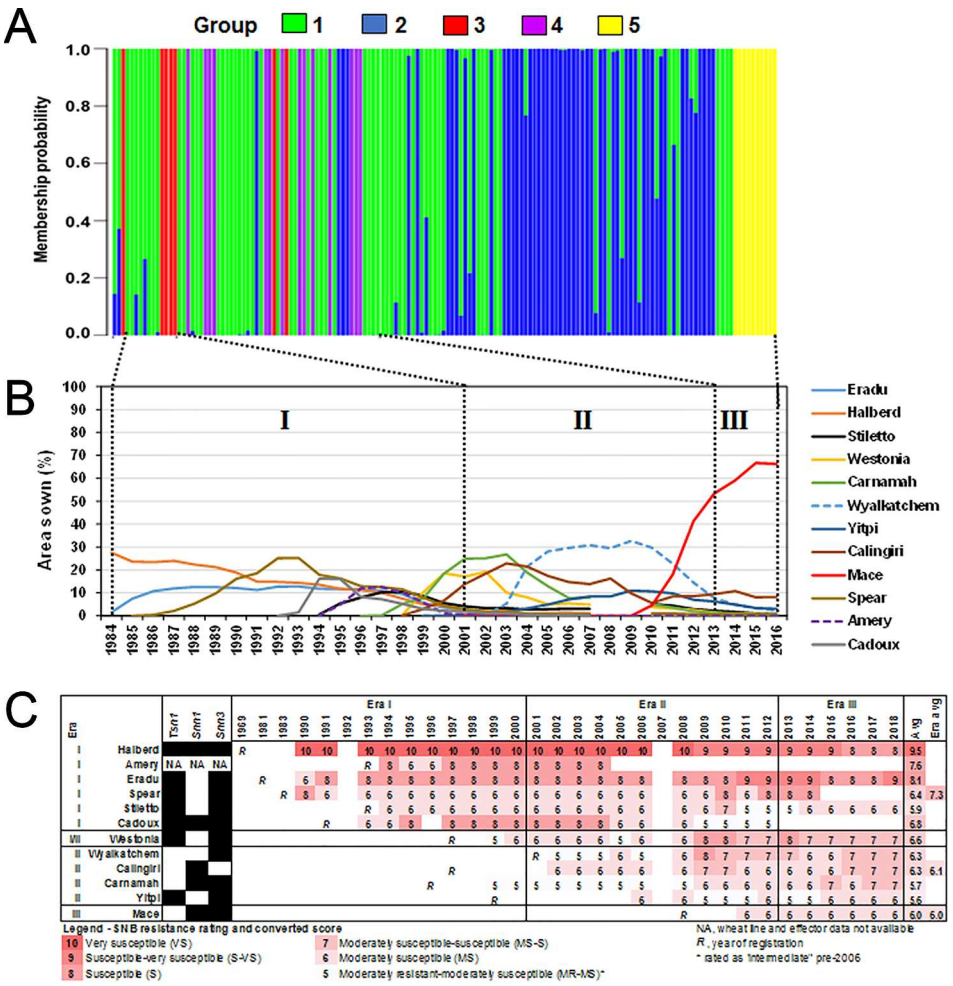


FIGURE 6 | Shifts in the *P. nodorum* population structure over time using **(A)** discriminant analysis of principal components (DAPC) membership probability. Isolates were arranged in chronological order from 1976 (left) to 2016 (right). **(B)** Major wheats sown in WA between 1984–2016 (CBH Group, Australia). Wheats grown on at least 10% of the total area sown in any given year were shown. Roman numerals indicate designated eras of major changes in wheat cultivar adoption coinciding with shifts in the *P. nodorum* population. **(C)** SNB resistance rating of major wheats from 1990 to 2018 (DPIRD, Australia).

were on average at least one unit more susceptible to SNB than popular cultivars in Era II and Era III, indicating a slow but significant overall improvement in wheat breeding for SNB. In seven cases (Amery, Eradu, Cadoux, Westonia, Wyalkatchem, Calingiri, and Carnamah), the cultivar resistance rating declined in the years after release. The most dramatic cases are Eradu and Wyalkatchem which declined 3 units during the period of adoption. In three other cases (Halberd, Yitpi, and Mace) there was no sustained change during the period of use. There is some evidence of an increase in resistance rating after a cultivar had been dropped (Halberd and Cadoux).

Whole Plant Infection Assay

To identify factors that promote the formation of the three transient groups which coincided with the three Eras of mass adoption of a few or a single top-grown wheat lines from 1984 to 2016 and the shift of Group 1 to Group 2, we choose three isolates from each group to test their performance on the top-grown wheat lines during those periods. Data from the whole plant infection assay was split into two sets: disease scores of the three transient (Groups 3/4/5) and the core (Groups 1/2) on the same set of seven top-grown wheat lines (**Figure 7**). Of the seven wheat lines; Halberd, Eradu and Cadoux were the representatives for Era I, Carnamah, Calingiri and Wyalkatchem for Era II; and Mace for Era III.

Significant differences in isolates from different groups and wheat lines from the three Eras were observed when the disease data of Groups 1 and 2 was subjected to ANOVA, but there was no significant variation for their interactions ($P = 1.13e^{-04}$, $2.12e^{-04}$, and 0.34, respectively). Since their interactions are not significant, two main factors: wheat lines from three Eras and isolates belonging to Groups 1 and 2 were analysed separately. Isolates from Group 1 were significantly less pathogenic than those from Group 2 ($P = 2.12e^{-04}$, **Figure 7**) and wheat lines from Era III

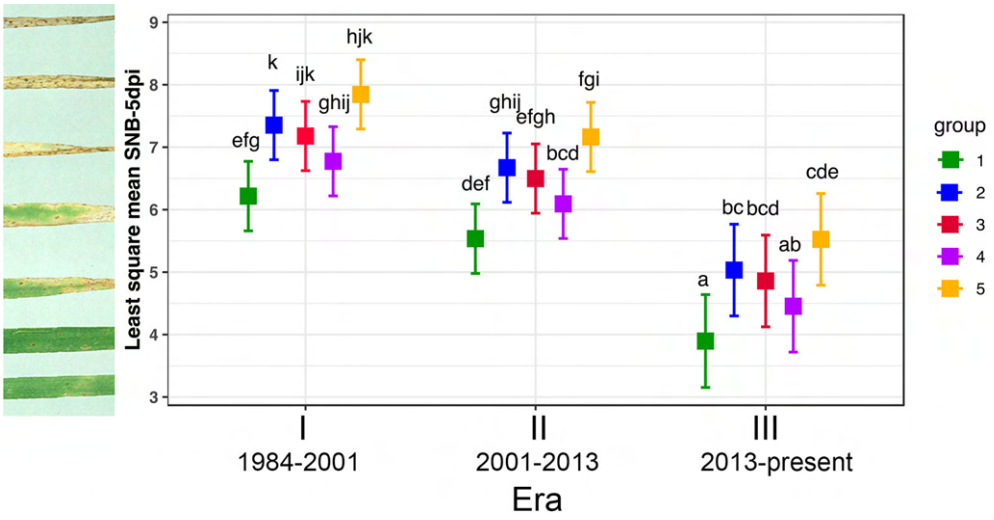


FIGURE 7 | Tukey's Post Hoc tests on average virulence score of three *P. nodorum* isolates from each genotype group infecting seven representative wheat lines from three Eras. Levels not connected by the same letter are significantly different ($P < 0.05$). Individual disease scores are described in **Table S5**.

were significantly more resistant than those in Eras I and II ($P = 1.13e^{-04}$, **Figure 7**).

For the three transient groups, ANOVA revealed that there are significant differences among isolates, wheat lines from different eras and their interactions ($P = <$ 2e-16, 1.09e-07, and 8.93e-03, respectively). Since their interaction was significant, main effects (isolates and wheat lines from different Eras) were not considered independently and Tukey's Post Hoc test was deployed. **Figure 7** showed that all isolates from three eras were pathogenic and performed equally well on Era I wheat lines; for wheat lines in Era II, isolates from Groups 3 and 4 were similar ($P = 0.99$) and significantly less pathogenic than isolates from Group 5 ($P = 0.04$ and 0.006, respectively); in Era III, isolates from Group 5 were significantly more aggressive than those from Group 4 ($P = 1.25e^{-05}$) and those from Group 4 were somewhere in the middle (**Figure 7**). Based on all wheat lines tested, Group 5 was the most aggressive with disease score significantly higher than all other groups, followed by the modern core Group 2 and the least pathogenic group was the old core Group 1 ($P = 0.005$).

DISCUSSION

SNB and tan spot are currently the two major fungal diseases of wheat in WA (Murray and Brennan, 2009). Both pathogens frequently co-exist on wheat and disease symptoms are often hard to distinguish (Loughman et al., 1993; Shankar et al., 2013). Difficulties in *P. nodorum* isolation prompted us to develop an effective method using boscalid to suppress *Ptr* to a sufficient level that allow *P. nodorum* growth from infected plant material to attain greater isolation efficiency.

In this study, we assembled a long time-span panel of isolates albeit the numbers of isolate from before 2000 is low. Such longitudinal collections are rare and we sought to determine how it could help us understand how the pathogen had evolved in the past few decades and assist in creating sound disease control strategies. Previous studies found that *P. nodorum* populations exist without any discernible structure even at the continental scales (Keller et al., 1997a; Keller et al., 1997b; Stukenbrock et al., 2006; Blixt et al., 2008). Keller et al. (1997a) examined the structure of field populations of *P. nodorum* from Texas, Oregon, and Switzerland and found 96% of the total genetic diversity within the geographically distant populations. Only 3% genetic dissimilarity was detected among nine wheat fields representing three geographical regions in Switzerland and different wheat cultivars (Keller et al., 1997b). We used robust statistical clustering methods to demonstrate that the Australian *P. nodorum* population distributed into five groups. To our knowledge, this is the first study to demonstrate evidence of a distinct genetic structure in a *P. nodorum* population.

Two isolates collected from Denmark and France were found to be closely related to Australian isolates (**Figure 2**). In a global *P. nodorum* population diversity study by Stukenbrock et al. (2006), it was reported that Australia is a sink population where traces of the pathogen from Europe were identified. As these two isolates clustered within Group 2 (**Figure 2**), it is possible that the ancestral lineage of these isolates were foreign incursions that contributed to the appearance and expansion of Group 2 isolates in the core population. More studies on European *P. nodorum* isolates are needed to verify this hypothesis.

Both mating types *MAT1-1* and *MAT1-2* were previously detected in the WA *P. nodorum* (Murphy et al., 2000; Solomon et al., 2004a) with a frequency not significantly different from 1:1, suggestive of regular sexual reproduction. Ascospores have been detected confirming that sexual reproduction does occur (Murphy et al., 2000; Bathgate and Loughman, 2001). This study also found an equal mixture of the two mating types in the entire collection but with a distinct pattern in the five groups.

An equal ratio was found in both Groups 1 and 2. However, the I_A analysis rejected the null hypothesis of no linkage among markers tested, indicating that sexual reproduction is rare. This phenomenon could be due to Group 1 and 2 possess a high genetic diversity base with sexual reproduction commonly occurring in the past (Murphy et al., 2000, Bathgate and Loughman, 2001) and the diversity maintained over time with recently dominant asexual reproduction.

A strict pairing between isolates of similar genetic backgrounds is also another possible explanation for the observation. Numerous attempts to cross the isolates in our laboratory in different conditions reported to be suitable for the mating conditions of *P. nodorum* have been made without any success and differences in chromosome numbers are a common feature of plant fungal pathogens especially in these Dothideomycetes fungal species (Cooley and Caten, 1991; Galazka and Freitag, 2014; Bertazzoni et al., 2018; Moller et al., 2018). The possible restricted mating between similar isolates may have led to retention of equal mating type ratio and the low observed recombination frequency among different genetic backgrounds. If this is the case, Groups 1 and 2 pose the highest potential for rapid adaptation as we note that these groups possess greater genetic diversity and variants pools of known and potential effectors and virulence profiles.

Group 2 isolates were consistently more virulent than Group 1 and this effect has become more pronounced in Era III when cv. Mace has predominated. We suggest that the shift from Group 1 genotypes to the more contemporary Group 2 may be *via* selection for virulence on Mace (**Figure 6B**).

Groups 3, 4, and 5 each possessed a single mating type and a single ToxA, 1 and 3 isoform (**Table S1**). The small number of detected isolates compared to the core isolates, differences in virulence, low genetic variance within groups and the large distance between these groups may be explained by: changes in cultivar usage, population genetic bottlenecks, or new populations founded by a small number of well-adapted or foreign introduced isolates.

The origins of the non-core groups are currently being addressed. There are possibilities including they could be local diversifications since large genetic diversity also exists within the core groups which were adapted to the newly grown wheat lines; or they were derived from migration events from distant populations and selected due to the ability to infect a particular sets of wheat lines; or they were hybrids between local core and non-Australian isolates. Examining unique alleles for each of the five groups we found that the transient groups possess 14 unique and shared 135 common alleles with the two core Group 1 and 2 (**Table S6**). Although the numbers of isolates for Groups 3/4/5 are small, this study does indicate that among the available local and non-Australian candidates whoever is/are the most suitable for/adaptive to the current hosts at a particular time will prevail and expand and become a group of its own which is detectable.

The fact that Group 5 was collected only in 2016 and in one location, it is not absolutely certain that it is a transient based on time frame as data for post 2016 would be required. However, based on Group 5's properties of small number of isolates, limited spatial abundance, single mating-type and single ToxA, 1 and 3 isoform profile (**Table S1**), Group 5 seems to fit well with an emergent set. Likewise, Group 3 spanned over 30-year period but was considered transient due to its other characters such as small number of isolates, limited spatial abundance, single mating-type and single effector isoform profile. Additional data collected for the following years would be needed to warrant a more comprehensive understanding and more precise conclusions on Australian *P. nodorum* evolutionary processes and population genetic structure.

Shifts in fungal populations and disease expression on crops can be attributed to many factors including host shift/range expansion, foreign introductions and climatic changes (Elad and Pertot, 2014; Gladieux et al., 2015). The occurrence of SNB 'boom-and-bust' cycles in the last 44 years were evident when we combined evidence gathered from *P. nodorum* genotypes, wheat cultivar popularity and SNB resistance ratings. Released in 1969, Halberd remained a popular wheat cultivar until the late 90s and was highly susceptible to SNB and later replaced by Wyalkatchem. Reduction in the commercial adoption of Halberd resulted in a significant increase in the SNB resistance. We suspect the replacement of Halberd caused a shift in the *P. nodorum* population available at that time as it adapted. This pattern of the temporary appearance of the transient populations suggests that these groups were selected by virtue of their greater virulence on contemporary cultivars. The selection might explain the frequently observed decline in the resistance rating of the cultivars in the years following their peak adoption. The transient *P. nodorum* groups can reproduce asexually many times within a growing season which can result in an epidemic caused by a small subset of the population may be enough to skew the mating type ratio to the point where one mating type allele was dominant. It has been shown that the mating type locus is linked to virulence in other fungal pathogens (Lee et al., 1999; Hsueh and Heitman, 2008). The low genetic diversity and maybe absence of the other mating type would then render the transient populations vulnerable to the introduction of a new cultivar as it would be unable to generate the appropriate recombinants quickly enough. Instead, the diverse genetic pool of Groups 1 and 2 and/or new introductions from overseas may be able to compensate for their relative lack of virulence by their ability to recombine modestly virulent genotypes which leads to new and distinct transient groups capable to adapt to newly introduced cultivars.

SNB ratings of commercial wheat cultivars were scored on a ten-point scale (10 being "very susceptible"). Throughout the period in question, the highest resistance scores were 5, described as MR-MS. No cultivars have been assigned any better scores. Within that 5-point range, some cultivars declined in resistance by 3 units but 1 or 2 units was more common. We therefore describe this pattern of ubiquitous disease and modest declines in resistance over 5–10 years as a low amplitude boom and bust cycle in contrast to the classical full range boom and bust cycles seen for biotrophic diseases such as the cereal rusts (Agrios, 2005). As Brown (2015) points out, the classic boom and bust cycle is only found when single major R-genes are used to control

a pathogen and where the loss of avirulence to the R-gene does not impact overall pathogenicity of the strain. Durable resistance has been achieved to biotrophic diseases like wheat powdery mildew and to SNB in the United Kingdom by the use of diverse germplasm sources and pragmatic selection for minor resistance genes based on field phenotyping. A similar breeding strategy was used for SNB in WA and was broadly successful in increasing the resistance levels and durability of adopted cultivars. Individual cultivars retained their resistance rating for 2–5 years and the reduction in the overall resistance rating was quite modest. The advent of effector-assisted breeding has accelerated the removal of susceptibility alleles and allowed more rapid breeding of cultivars with improved resistance to SNB and tan spot caused by *Ptr* through the removal of *Tsn1* (Vleeshouwers and Oliver, 2014).

This pattern of both non-core populations of specific and limited genetic diversity alongside with core populations of ubiquitous and rich in genetic variation implies that the selection of isolates used in the testing of new breeding lines and cultivars can be improved. The current system uses an uncontrolled mixture of stored and newer isolates (Shankar *pers. comm.*). Our studies indicate that the *P. nodorum* population from WA can be divided in to five groups. The transient groups from the past (Groups 3 and 4) can be safely ignored as they have been apparently driven to extinction. The first priority would be to screen for resistance to the current dominant emergent Group 5. We can predict that the use of cultivar/s with good resistance to Group 5 will lead to its rapid elimination and its replacement from within the core populations and/or foreign and/or local/foreign hybrids. Hence to achieve long term and durable resistance new cultivars should also be selected for resistance to Groups 1 and 2.

In more general terms, these studies indicate that the annual collection of isolates should be a priority for the control of all crop diseases. Examination of neutral genetic markers can be used to estimate population differentiation. Detection of skewed mating type ratios within sub-groups is a predictor of rapid adaptation to a current cultivar or other agronomic factor, such as a fungicide regime. Overall though these studies emphasise the value of cultivar diversity even in the absence of high amplitude boom and bust cycles (Wolfe, 1985; van den Bosch et al., 2014).

DATA AVAILABILITY STATEMENT

The datasets generated for this study can be found as supplemental data.

AUTHOR CONTRIBUTIONS

K-CT and HP conceived the experiment. K-CT, HP, and RO wrote the manuscript. HP, DJ, KR, and KD conducted all experiments. Results were analysed by K-CT, HP, PB, RO, and DJ. FL-R, M-HL, RV, PB, LG, and RO provided intellectual feedback and edited the manuscript.

FUNDING

This study was supported by CCDM, a joint initiative of Curtin University and the Grains Research and Development Corporation (CUR00023).

ACKNOWLEDGMENTS

We thank DPIRD for providing SNB resistance data, Dr Tim Friesen (USDA) for US and Denmark *P. nodorum* isolates and the CBH Group for supplying crop statistic data, and Dr Bruce McDonald (ETH Zurich) for scientific discussions. The article is dedicated to the memory of Dr Patrick Brunner for his tireless work on fungal evolutionary genetics and septoria biology.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fpls.2019.01785/full#supplementary-material

REFERENCES

Agrios, G. N. (2005). *Plant pathology. 5th edn.* (Amsterdam, Boston: Elsevier Academic Press).

Bathgate, J. A., and Loughman, R. (2001) Ascospores are a source of inoculum of *Phaeosphaeria nodorum, P. avenaria* f. sp *avenaria* and *Mycosphaerella graminicola* in Western Australia. *Aust. Plant Pathol.* 30, 317–322. doi: 10.1071/AP01043

Bennett, R. S., Yun, S. H., Lee, T. Y., Turgeon, B. G., Arseniuk, E., Cunfer, B. M., et al. (2003). Identity and conservation of mating type genes in geographically diverse isolates of *Phaeosphaeria nodorum*. *Fungal. Genet. Biol.* 40, 25–37. doi: 10.1016/S1087-1845(03)00062-8

Bertazzoni, S., Williams, A. H., Jones, D. A., Syme, R. A., Tan, K. C., and Hane, J. K. (2018). Accessories make the outfit: accessory chromosomes and other dispensable DNA Regions in plant-pathogenic fungi. *Mol. Plant Microb. Interact.* 31, 779–788. doi: 10.1094/MPMI-06-17-0135-FI

Blixt, E., Olson, A., Hogberg, N., Djurle, A., and Yuen, J. (2008). Mating type distribution and genetic structure are consistent with sexual recombination in the Swedish population of *Phaeosphaeria nodorum*. *Plant Pathol.* 57, 634–641. doi: 10.1111/j.1365-3059.2008.01826.x

Brasier, C. M., Cooke, D. E., and Duncan, J. M. (1999). Origin of a new Phytophthora pathogen through interspecific hybridization. *Proc. Natl. Acad. Sci. U. S. A.* 96, 5878–5883. doi: 10.1073/pnas.96.10.5878

Brown, J. K. (2015). Durable resistance of crops to disease: a Darwinian perspective. *Annu. Rev. Phytopathol.* 53, 513–539. doi: 10.1146/annurev-phyto-102313-045914

Bruvo, R., Michiels, N. K., D'Souza, T. G., and Schulenburg, H. (2004). A simple method for the calculation of microsatellite genotype distances irrespective of ploidy level. *Mol. Ecol.* 13, 2101–2106. doi: 10.1111/j.1365-294X.2004.02209.x

Chessel, D., Dufour, A. B., and Thioulouse, J. (2004). The ade4 package - I: one-table methods. *R. News* 4, 5–10.

Ciuffetti, L. M., Tuori, R. P., and Gaventa, J. M. (1997). A single gene encodes a selective toxin causal to the development of tan spot of wheat. *Plant Cell* 9, 135–144. doi: 10.1105/tpc.9.2.135

Cooley, R. N., and Caten, C. E. (1991). Variation in electrophoretic karyotype between strains of *Septoria nodorum*. *Mol. Gen. Genet.* 228, 17–23. doi: 10.1007/BF00282442

Development Core Team R. (2008). R Core Team. R A Language and Environment for Statistical Computing 2014.

Elad, Y., and Pertot, I. (2014). Climate change impacts on plant pathogens and plant diseases. *J. Crop Improvement* 28, 99–139. doi: 10.1080/15427528.2014.865412

Eyal, Z., Scharen, A. L., Prescott, J. M., and van Ginkel, M. (1987). *The Septoria Diseases of Wheat: Concepts and Methods of Disease Management* (Texcoco, Mexico: CIMMYT).

Friesen, T. L., and Faris, J. D. (2010). Characterization of the wheat–*Stagonospora nodorum* disease system: what is the molecular basis of this quantitative necrotrophic disease interaction. *Canadian J. Plant Pathol.* 32, 20–28. doi: 10.1080/07060661003620896

Friesen, T. L., Stukenbrock, E. H., Liu, Z. H., Meinhardt, S., Ling, H., Faris, J. D., et al. (2006). Emergence of a new disease as a result of interspecific virulence gene transfer. *Nature Genet.* 38, 953–956. doi: 10.1038/ng1839

Galazka, J. M., and Freitag, M. (2014). Variability of chromosome structure in pathogenic fungi-of 'ends and odds'. *Curr. Opin. Microbiol.* 20, 19–26. doi: 10.1016/j.mib.2014.04.002

Gladieux, P., Feurtey, A., Hood, M. E., Snirc, A., Clavel, J., Dutech, C., et al. (2015). The population biology of fungal invasions. *Mol. Ecol.* 24, 1969–1986. doi: 10.1111/mec.13028

Goss, E. M., Tabima, J. F., Cooke, D. E. L., Restrepo, S., Fry, W. E., Forbes, G. A., et al. (2014). The Irish potato famine pathogen Phytophthora infestans originated in central Mexico rather than the Andes. *Proc. Natl. Acad. Sci. U. S. A.* 111, 8791–8796. doi: 10.1073/pnas.1401884111

Grunwald, N. J., Goodwin, S. B., Milgroom, M. G., and Fry, W. E. (2003). Analysis of genotypic diversity data for populations of microorganisms. *Phytopathology* 93, 738–746. doi: 10.1094/PHYTO.2003.93.6.738

Hayden, M. J., Nguyen, T. M., Waterman, A., and Chalmers, K. J. (2008). Multiplex-ready PCR: a new method for multiplexed SSR and SNP genotyping. *BMC Genom.* 9, 80. doi: 10.1186/1471-2164-9-80

Hsueh, Y. P., and Heitman, J. (2008). Orchestration of sexual reproduction and virulence by the fungal mating-type locus. *Curr. Opin. Microbiol.* 11, 517–524. doi: 10.1016/j.mib.2008.09.014

Jombart, T., Devillard, S., and Balloux, F. (2010). Discriminant analysis of principal components: a new method for the analysis of genetically structured populations. *BMC Genet.* 11, 94. doi: 10.1186/1471-2156-1111-1194

Kamvar, Z. N., Tabima, J. F., and Grunwald, N. J. (2014). Poppr: an R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *Peerj* 2:e281. doi: 10.7717/peerj.7281

Keller, S. M., McDermott, J. M., Pettway, R. E., Wolfe, M. S., and McDonald, B. A. (1997a). Gene flow and sexual reproduction in the wheat glume blotch pathogen *Phaeosphaeria nodorum* (anamorph *Stagonospora nodorum*). *Phytopathol* 87, 353–358. doi: 10.1094/PHYTO.1997.87.3.353

Keller, S. M., Wolfe, M. S., McDermott, J. M., and McDonald, B. A. (1997b). High genetic similarity among populations of phaeosphaeria nodorum across wheat cultivars and regions in Switzerland. *Phytopathology* 87, 1134–1139. doi: 10.1094/PHYTO.1997.87.11.1134

Lee, N., Bakkeren, G., Wong, K., Sherwood, J. E., and Kronstad, J. W. (1999). The mating-type and pathogenicity locus of the fungus *Ustilago hordei* spans a 500-kb region. *Proc. Natl. Acad. Sci. U. S. A.* 96, 15026–15031. doi: 10.1073/pnas.96.26.15026

Liu, Z. H., Faris, J. D., Meinhardt, S. W., Ali, S., Rasmussen, J. B., and Friesen, T. L. (2004). Genetic and physical mapping of a gene conditioning sensitivity in wheat to a partially purified host-selective toxin produced by *Stagonospora nodorum*. *Phytopathology* 94, 1056–1060. doi: 10.1094/PHYTO.2004.94.10.1056

Liu, Z., Zhang, Z., Faris, J. D., Oliver, R. P., Syme, R., McDonald, M. C., et al. (2012). The cysteine rich necrotrophic effector SnTox1 produced by *Stagonospora nodorum* triggers susceptibility of wheat lines harboring *Snn1*. *PLoS Pathog.* 8, e1002467. doi: 10.1371/journal.ppat.1002467

Loughman, R., Wilson, R. E., and Thomas, G. J. (1993). The influence of disease complexes involving *Leptosphaeria* (*Septoria*) *nodorum on* detection of resistance to 3 leaf-spot diseases in wheat. *Euphytica* 72, 31–42. doi: 10.1007/BF00023770

Loughman, R. (1989). Identifying wheat leaf diseases. Western Australian Department of Agriculture Farmnote 9/89.

Mair, W. J., Deng, W. W., Mullins, J. G. L., West, S., Wang, P. H., Besharat, N., et al. (2016). Demethylase inhibitor fungicide resistance in *Pyrenophora teres* f. sp *teres* associated with target site modification and inducible overexpression of *Cyp51*. *Front. Microbiol.* 7. doi: 10.3389/fmicb.2016.01279

McDonald, B. A., and Linde, C. (2002). The population genetics of plant pathogens and breeding strategies for durable resistance. *Euphytica* 124, 163–180. doi: 10.1023/A:1015678432355

McDonald, B. A., Miles, T., Nelson, L. R., and Pettway, R. E. (1994). Genetic variability in nuclear DNA in field populations of *Stagonospora nodorum*. *Phytopathology* 84, 250–255. doi: 10.1094/Phyto-84-250

McDonald, M. C., Oliver, R. P., Friesen, T. L., Brunner, P. C., and McDonald, B. A. (2013). Global diversity and distribution of three necrotrophic effectors in *Phaeosphaeria nodorum* and related species. *New Phytol.* 199, 241–251. doi: 10.1111/nph.12257

McDonald, M. C., Ahren, D., Simpfendorfer, S., Milgate, A., and Solomon, P. S. (2018). The discovery of the virulence gene ToxA in the wheat and barley pathogen *Bipolaris sorokiniana*. *Mol. Plant Pathol.* 19, 432–439. doi: 10.1111/mpp.12535

Moller, M., Habig, M., Freitag, M., and Stukenbrock, E. H. (2018). Extraordinary genome instability and widespread chromosome rearrangements during vegetative growth. *Genetics* 210, 517–529. doi: 10.1534/genetics.118.301050

Murphy, N. E., Loughman, R., Appels, R., Lagudah, E. S., and Jones, M. G. K. (2000). Genetic variability in a collection of *Stagonospora nodorum* isolates from Western Australia. *Australian J. Agri. Res.* 51, 679–684. doi: 10.1071/AR99107

Murray, G. M., and Brennan, J. P. (2009). Estimating disease losses to the Australian wheat industry. *Austr. Plant Pathol.* 38, 558–570. doi: 10.1071/AP09053

Nei, M. (1978). Estimation of average heterozygosity and genetic distance from a small number of individuals. *Genetics* 89, 583–590.

Oliver, R. P., Rybak, K., Solomon, P. S., and Ferguson-Hunt, M. (2009). Prevalence of ToxA-sensitive alleles of the wheat gene *Tsn1 in* Australian and Chinese wheat cultivars. *Crop Pasture Sc.* 60, 348–352. doi: 10.1071/CP08259

Oliver, R. P., Friesen, T. L., Faris, J. D., and Solomon, P. S. (2012). *Stagonospora nodorum*: from pathology to genomics and host resistance. *Annu. Rev. Phytopathol.* 50, 23–43. doi: 10.1146/annurev-phyto-081211-173019

Ordonez, M. E., and Kolmer, J. A. (2009). Differentiation of molecular genotypes and virulence phenotypes of *Puccinia triticina* from common wheat in North America. *Phytopathology* 99, 750–758. doi: 10.1094/PHYTO-99-6-0750

Prevosti, A., Ocana, J., and Alonso, G. (1975). Distances between populations of *Drosophila subobscura*, based on chromosome arrangement frequencies. *Theor. Appl. Genet.* 45, 231–241. doi: 10.1007/BF00831894

Richards, J. K., Stukenbrock, E. H., Carpenter, J., Liu, Z., Cowger, C., Faris, J. D., et al. (2019). Local adaptation drives the diversification of effectors in the fungal wheat pathogen Parastagonospora nodorum in the United States. *PLoS Genet.* 15, e1008223. doi: 10.1371/journal.pgen.1008223

Shankar, M., Walker, E., Golzar, H., Loughman, R., Wilson, R. E., and Francki, M. G. (2008). Quantitative trait loci for seedling and adult plant resistance to Stagonospora nodorum in wheat. *Phytopathology* 98, 886–893. doi: 10.1094/PHYTO-98-8-0886

Shankar, M., Mather, D., Francki, M., Jorgensen, D., Golzar, H., Chalmers, K., et al. (2013). Strategies for genetic enhancement of resistance to yellow spot and stagonospora nodorum blotch in wheat. *W. A. Crop Updates Perth.* 1–3.

Shannon, C. E. (1997). The mathematical theory of communication. 1963. MD Comput 14, 306–317.

Shi, G., Zhang, Z., Friesen, T. L., Bansal, U., Cloutier, S., Wicker, T., et al. (2016). Marker development, saturation mapping, and high-resolution mapping of the Septoria nodorum blotch susceptibility gene Snn3-B1 in wheat. *Mol. Genet. Genom.* 291, 107–119. doi: 10.1007/s00438-015-1091-x

Simpson, E. H. (1949). Measurement of diversity. *Nature* 163, 688. doi: 10.1038/163688a0

Smith, J. M., Smith, N. H., Orourke, M., and Spratt, B. G. (1993). How clonal are bacteria. *Proc. Natl. Acad. Sci. U. S. A.* 90, 4384–4388. doi: 10.1073/pnas.90.10.4384

Solomon, P. S., Thomas, S. W., Spanu, P., and Oliver, R. P. (2003). The utilisation of di/tripeptides by *Stagonospora nodorum* is dispensable for wheat infection. *Physiol. Mol. Plant P.* 63, 191–199. doi: 10.1016/j.pmpp.2003.12.003

Solomon, P. S., Parker, K., Loughman, R., and Oliver, R. P. (2004a). Both mating types of *Phaeosphaeria* (anamorph *Stagonospora*) *nodorum are* present in Western Australia. *Eur. J. Plant Pathol.* 110, 763–766. doi: 10.1023/B:EJPP.0000041565.42836.c1

Solomon, P. S., Tan, K. C., Sanchez, P., Cooper, R. M., and Oliver, R. P. (2004b). The disruption of a Gα subunit sheds new light on the pathogenicity of *Stagonospora nodorum* on wheat. *Mol. Plant-Microbe Interact.* 17, 456–466. doi: 10.1094/MPMI.2004.17.5.456

Solomon, P. S., Lowe, R. G. T., Tan, K.-C., Waters, O. D. C., and Oliver, R. P. (2006). *Stagonospora nodorum*: cause of stagonospora nodorum blotch of wheat. *Mol. Plant Pathol.* 7, 147–156. doi: 10.1111/j.1364-3703.2006.00326.x

Stoddart, J. A., and Taylor, J. F. (1988). Genotypic diversity: estimation and prediction in samples. *Genetics* 118, 705–711.

Stukenbrock, E. H., Banke, S., Zala, M., McDonald, B. A., and Oliver, R. P. (2005). Isolation and characterization of EST-derived microsatellite loci from the fungal wheat pathogen *Phaeosphaeria nodorum*. *Mol. Ecol. Notes* 5, 931–933. doi: 10.1111/j.1471-8286.2005.01120.x

Stukenbrock, E. H., Banke, S., and McDonald, B. A. (2006). Global migration patterns in the fungal wheat pathogen *Phaeosphaeria nodorum*. *Mol. Ecol.* 15, 2895–2904. doi: 10.1111/j.1365-294X.2006.02986.x

Syme, R. A., Tan, K. C., Hane, J. K., Dodhia, K., Stoll, T., Hastie, M., et al. (2016). Comprehensive annotation of the *Parastagonospora nodorum* reference genome using next-generation genomics, transcriptomics and proteogenomics. *PLoS One* 11, e0147221. doi: 10.1371/journal.pone.0147221

Syme, R. A., Tan, K. C., Rybak, K., Friesen, T. L., McDonald, B. A., Oliver, R. P., et al. (2018). Pan-Parastagonospora comparative genome analysis-effector prediction and genome evolution. *Genome Biol. Evol.* 10, 2443–2457. doi: 10.1093/gbe/evy192

Tan, K. C., and Oliver, R. P. (2017). Regulation of proteinaceous effector expression in phytopathogenic fungi. *PLoS Pathog.* 13, e1006241. doi: 10.1371/journal.ppat.1006241

Tan, K. C., Ferguson-Hunt, M., Rybak, K., Waters, O. D., Stanley, W. A., Bond, C. S., et al. (2012). Quantitative variation in effector activity of ToxA isoforms from *Stagonospora nodorum* and *Pyrenophora tritici-repentis*. *Mol. Plant-Microbe Interact.* 25, 515–522. doi: 10.1094/MPMI-10-11-0273

Turner, B. C., Perkins, D. D., and Fairfield, A. (2001). Neurospora from natural populations: a global study. *Fungal. Genet. Biol.* 32, 67–92. doi: 10.1006/fgbi.2001.1247

van den Bosch, F., Oliver, R., van den Berg, F., and Paveley, N. (2014). Governing principles can guide fungicide-resistance management tactics. *Annu. Rev. Phytopathol.* 52, 175–195. doi: 10.1146/annurev-phyto-102313-050158

Vleeshouwers, V. G., and Oliver, R. P. (2014). Effectors as tools in disease resistance breeding against biotrophic, hemibiotrophic, and necrotrophic plant pathogens. *Mol. Plant-Microbe Interact.* 27, 196–206. doi: 10.1094/MPMI-10-13-0313-IA

Wolfe, M. S. (1985). The current status and prospects of multiline cultivars and variety mixtures for disease resistance. *Annu. Rev. Phytopathol.* 23, 251–273. doi: 10.1146/annurev.py.23.090185.001343

Zaicou-Kunesch, C., Trainor, G., Shackley, B., Curry, J., Nicol, D., Shankar, M., et al. (2018). 2018 Wheat variety sowing guide for Western Australia. *Department Food Agri. Perth.* 1–44. ISSN 1833 7236

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

CHAPTER 9 — THEME 3

A specific fungal transcription factor controls effector gene expression and orchestrates the establishment of the necrotrophic pathogen lifestyle on wheat

This chapter is also published in:
Scientific Reports, 2019, vol. 9, article 15884
https://doi.org/10.1038/s41598-019-52444-7

This chapter is submitted as supplementary material and should not contribute to assessment of this thesis. It is included here as an example of related research contributions made during the candidacy.

9.1 Declaration

Title	A specific fungal transcription factor controls effector gene expression and orchestrates the establishment of the necrotrophic pathogen lifestyle on wheat.
Authors	**Darcy A. B. Jones**, Evan John, Kasia Rybak, Huyen T. T. Phan, Karam B. Singh, Shao-Yu Lin, Peter S. Solomon, Richard P. Oliver and Kar-Chun Tan.
Reference	2019. *Scientific Reports*, 9(1), 15884.
DOI	`https://doi.org/10.1038/s41598-019-52444-7`

This supplementary chapter has been submitted as part of another student's thesis (Evan John). As such, it is included here only as an example of contributions to related work conducted during the candidate's Ph. D. and should not contribute to assessment. This thesis supplementary chapter is submitted in the form of a collaboratively-written peer-reviewed manuscript.

The following contributions were made by the Ph.D candidate (DABJ) and co-authors:

- K-CT conceived the experiment.
- KR and HTTP performed the experiment and extracted RNA.
- **DABJ** performed RNAseq processing, differential gene expression, and GO term enrichment analyses.
- EJ performed enrichment analyses of DNA motifs in promoters of differentially expressed genes.
- S-YL performed yeast-1-hybrid assays.
- KCT, **DABJ**, and EJ wrote the paper.
- All authors read, edited, and approved the manuscript.

I, Darcy Jones, hereby certify that this attribution statement is an accurate record of my contribution to the research presented in this chapter.

Darcy Jones

I certify that this attribution statement is an accurate record of Darcy Jones' contribution to the research presented in this chapter[1].

Evan John	Kasia Rybak	Huyen T. T. Phan
Karam B. Singh	Peter S. Solomon	Richard P. Oliver
Kar-Chun Tan		

[1] We were unable to contact Shao-Yu Lin.

SCIENTIFIC REPORTS
nature research

OPEN

A specific fungal transcription factor controls effector gene expression and orchestrates the establishment of the necrotrophic pathogen lifestyle on wheat

Darcy A. B. Jones[1,4], Evan John[1,4], Kasia Rybak[1], Huyen T. T. Phan[1], Karam B. Singh[1,2], Shao-Yu Lin[3], Peter S. Solomon[3], Richard P. Oliver[1] & Kar-Chun Tan[1*]

The fungus *Parastagonospora nodorum* infects wheat through the use of necrotrophic effector (NE) proteins that cause host-specific tissue necrosis. The Zn_2Cys_6 transcription factor PnPf2 positively regulates NE gene expression and is required for virulence on wheat. Little is known about other downstream targets of PnPf2. We compared the transcriptomes of the *P. nodorum* wildtype and a strain deleted in *PnPf2* (*pf2-69*) during *in vitro* growth and host infection to further elucidate targets of PnPf2 signalling. Gene ontology enrichment analysis of the differentially expressed (DE) genes revealed that genes associated with plant cell wall degradation and proteolysis were enriched in down-regulated DE gene sets in *pf2-69* compared to SN15. In contrast, genes associated with redox control, nutrient and ion transport were up-regulated in the mutant. Further analysis of the DE gene set revealed that PnPf2 positively regulates twelve genes that encode effector-like proteins. Two of these genes encode proteins with homology to previously characterised effectors in other fungal phytopathogens. In addition to modulating effector gene expression, PnPf2 may play a broader role in the establishment of a necrotrophic lifestyle by orchestrating the expression of genes associated with plant cell wall degradation and nutrient assimilation.

The fungus *Parastagonospora nodorum* causes septoria nodorum blotch (SNB) of wheat. *P. nodorum* uses necrotrophic effectors (NEs) to cause tissue necrosis and facilitate infection of hosts possessing dominant susceptibility genes. The genes encoding three of these NEs are known: *SnToxA*, *SnTox1*, and *SnTox3*. *SnToxA* encodes a 13.2 kDa mature protein that causes necrosis on wheat cultivars that possess the dominant susceptibility gene *Tsn1*[1,2]. Near-identical copies of *ToxA* have been found in two other wheat fungal pathogens, *Pyrenophora tritici-repentis* (*Ptr*)[3] and *Bipolaris sorokiniana*[4]. These may have been horizontally acquired, presumably from *P. nodorum*[1]. SnTox1 encodes a 10.3 kDa cysteine-rich mature protein that causes necrosis and confers virulence on wheat cultivars possessing *Snn1*[5]. SnTox3 is also a cysteine-rich NE. Sensitivity to the effector is conferred by either *Snn3-B1* or *Snn3-D1* located on wheat chromosomes 5BS and 5DS, respectively[6,7]. Genetic studies and protein purification assays indicate that *P. nodorum* possesses many more unidentified effectors associated with SNB[8].

SnToxA, *SnTox1* and *SnTox3* are highly expressed during early infection but their expression is greatly decreases during saprophytic growth on the necrotised host tissue[9]. However, else was known about factors affecting their regulation until recently. Studies of TFs in *P. nodorum* have also provided some insights into effector gene regulation. Deletion of the APSES-class TF gene *SnStuA* in *P. nodorum* resulted in mutants with abnormal vegetative growth, loss of sporulation and a complete loss of virulence on wheat[10]. The expression of *SnTox3* was significantly down-regulated in the mutant, though the loss in virulence is likely attributable to pleotropic effects incurred by the mutation. A C_2H_2 zinc finger TF PnCon7 that binds to the promoter region of *SnTox3* was

[1]School of Molecular and Life Sciences, Centre for Crop and Disease Management, Curtin University, Bentley, 6102, Perth, Western Australia, Australia. [2]CSIRO Agriculture and Food, Wembley, Western Australia, Australia. [3]Division of Plant Sciences, Research School of Biology, The Australian National University, Canberra, ACT, Australia. [4]These authors contributed equally: Darcy A. B. Jones and Evan John. *email: Kar-Chun.Tan@curtin.edu.au

identified using a combination of yeast-1-hybrid (Y1H) and DNase footprinting, suggesting that PnCon7 may directly regulate *SnTox3* expression[11]. Silencing of *PnCon7* drastically reduced *SnTox3* expression, suggesting that PnCon7 may be a direct regulator[11].

Cho *et al*.[12] identified and characterised a Pleosporales-specific zinc-finger TF gene *Abpf2* from *Alternaria brassicicola* using gene knockout methods. Mutants lacking *AbPf2* were non-pathogenic on various brassica hosts. Gene expression analysis using RNAseq identified eight putative candidate effector genes that were positively regulated by AbPf2. A BLAST search of AbPf2 against the *P. nodorum* predicted protein set identified a conserved homolog, PnPf2[9]. Functional analysis revealed that PnPf2 is a positive regulator of *SnToxA* and *SnTox3* expression and mutants lacking *PnPf2* were only infective on *Snn1* wheat lines[9]. Based on all evidence observed, we hypothesise that PnPf2 regulates the expression of novel effectors in *P. nodorum*. Firstly, *P. nodorum* SN15 carrying *SnToxA*, *SnTox1* and *SnTox3* deletions (*toxa13*) retained the ability to produce culture filtrate that cause host-specific chlorosis[13] and remained highly pathogenic on many modern bread wheat lines[14]. Secondly, genetic analysis revealed new quantitative trait loci for SNB were detected on wheat mapping populations[14,15]. It is possible that these QTL may be associated with novel dominant susceptibility genes[8]. Lastly, SN15 carrying deletions in both *PnPf2* and *SnTox1* lost the ability to infect all wheat lines tested including those that demonstrated susceptibility to *P. nodorum toxa13*[9]. This strongly suggests that PnPf2 positively regulates the expression of novel effector genes. To investigate this hypothesis and dissect other biochemical aspects of PnPf2 regulation, we used RNAseq to compare the gene expression profiles of a *P. nodorum pnpf2* mutant with the wildtype strain under conditions that are conducive for effector gene expression.

Results

***PnPf2* is required for full hyphal proliferation during host infection.** The transcriptome of the *P. nodorum* reference wildtype strain SN15 was compared to the *PnPf2*-deleted strain *pf2-69* grown under two conditions. Firstly, we sampled RNA during early infection at three days *in planta* (*ip*) where *PnPf2*, *SnToxA*, *SnTox1* and *SnTox3* are maximally expressed. Wheat cv. Halberd (*Tsn1*, *Snn1*, *Snn3*) was used as a host as it is susceptible to SN15 and *pf2-69*[9]. Secondly, SN15 and *pf2-69* were grown for three days *in vitro* (*iv*) in Fries 3 broth which is conducive for SnTox1 and SnTox3 production[9]. Vegetative growth of *pf2-69* was comparable to SN15[9]. Paired-end Illumina HiSeq technology was used as an RNAseq sequencing platform. The latest SN15 genome revision produced 13,563 predicted genes[16]. Deep sequencing produced more than 90% fungal transcripts that aligned to predicted genes for all samples (Supplementary Data S1 and Table 1). *In vitro* and *ip* samples returned an average of 24 million and 290 million read pair fragments (including plant reads), respectively. Between 18 and 22 million read pairs, representing an average of 6.94% of the total reads, aligned to the SN15 genome for the SN15 *ip* treatment (Supplementary Table S1). Between 3.4 and 5.9 million reads (average 1.57% of total) from *pf2-69 ip* growth aligned to the SN15 genome. The low proportion of fungal reads from *pf2-69* suggests reduced biomass during infection. Quantitative PCR of genomic DNA extracted from three days post infected wheat cv. Halberd confirmed that the biomass of *pf2-69* was significantly lower than strains carrying a functional copy of *PnPf2* (Fig. 1a,b).

Analysis of differentially expressed (DE) genes. Genes were considered DE in a contrast of isolate or treatment if tests of absolute $\log_2$ fold change >1 were consistently significant ($P_{adj} < 0.05$) for three test methods described below (Supplementary Data S3). Because *pf2-69 ip* samples had considerably fewer reads than other samples, an additional filter requiring *pf2-69 ip* samples to have $>= 10$ counts per million (CPM) for a gene to be called down-regulated was used for high-confidence DE prediction sets. For SN15 *ip* and *iv* treatments, 1,889 genes were up-regulated and 1393 were down-regulated *ip* (Supplementary Table S2). A total of 1,736 genes were up-regulated and 706 genes were down-regulated between the *pf2-69 ip* and *iv* treatments. For *ip* comparisons, 303 genes were significantly reduced whereas 449 were up-regulated in *pf2-69* over SN15. Additional DE genes were observed using relaxed criteria, allowing genes where any of the three tests are significant (<3 tests) (Supplementary Table S2). The main difference between the three tests results were in how they handle contrasts involving samples with few or no reads aligned to the gene. Additional genes involving *pf2-69 ip* samples with <10 CPM were identified using the same relaxed criteria. In total, 269 gene were down-regulated in *pf2-69 ip* compared to *iv* growth and had fewer than 10 CPM in *pf2-69* during *ip* growth. Similarly, 163 genes were down-regulated in *pf2-69* during infection compared to SN15 and had fewer than 10 CPM in *pf2-69* during *ip* growth (Supplementary Table S2).

A principal component analysis (PCA) plot for PC1 and PC2 was constructed based on normalised fragment counts per gene to describe the variation between and within each treatment (Fig. 1c). The biological replicates tightly clustered together, with each treatment strongly differentiated from the others. This indicates that sample treatment and sequencing did not contribute to systematic biases that could not be removed by normalisation. PC1 captured 71% of the total variance and discriminated *iv* from *ip* samples. PC2 captured 12% of the variance and discriminated SN15 from *pf2-69*.

We then examined *SnToxA*, *SnTox1* and *SnTox3* expression profiles (Fig. 1d). As expected, the expression of SnToxA and SnTox3 was almost abolished in *pf2-69 ip*. SnTox3 expression was also highly reduced in *pf2-69 iv*. *SnToxA* is poorly expressed in SN15 and *pf2-69* during *iv* growth. *SnTox1* expression was significantly higher in SN15 compared to *pf2-69*. *SnTox1* is still strongly expressed during *ip* growth and had the lowest fold change difference between SN15 and *pf2-69 ip* compared to *SnToxA* and *SnTox3*.

PnPf2 regulates genes that encode effector-like proteins. To identify candidate effector genes positively regulated by PnPf2, we analysed genes that were down-regulated in *pf2-69* that possessed a secretory signal peptide (but no transmembrane domains outside of the signal peptide) and were predicted to be effector-like by EffectorP[17]. Twelve genes that showed a similar expression profile to *SnToxA* (ie. down-regulated in *pf2-69 ip* compared to SN15 *ip* and up-regulated *ip* in both strains) were identified (Fig. 2). In contrast, *SnTox1* and

SN15 gene	PhiBase	Functional prediction	Size (kDa)	Length (aa)	SN79-1087 gene	Mutations (aa)	Notes
SNOG_01146	Homolog of *MoCDIP4* effector.	Cleavage of cellulose chains. CAZy family AA9 (formerly GH61)	23.5	229	03796-RA	D28E	—
SNOG_02755	—	Family with unknown function. Members in pathogens and non-pathogens. Incl biotrophs and necrotrophs.	41.6	409	02992-RA	GGQNNGQGQNNGQ31G, QNN82Q, G313GN	Repeat motif copy number variation
SNOG_02980	—	SGNH hydrolase-type esterase. Possible lipase or pectinase.	25.9	247	02810-RA	—	—
SNOG_08150	—	—	14.1	131	01518-RA	F3S, S131W	
SNOG_10736	—	—	48.8	522	10887-RA	N187NANAGNNANANAG, GANAGNNANAGAAAGNAAGNNANAGN244G, NANAG280N, GNN300G, G342GN	Repeat motif copy number variation
SNOG_12350	—	—	10.8	109	12820-RA	—	
SNOG_13939	—	Family with unknown function. Members in other pathogens. Incl biotrophs and necrotrophs	17.5	171	06645-RA	T24P, V67I, A104AAQVSISPSLTVTMMMWRNSSADAC	Intron splice site SNP in SN79-1087 creates large insertion
SNOG_14243	—	SGNH hydrolase-type esterase. Possible acetyl xylan esterase.	25.8	246	09528-RA	—	—
SNOG_15270	Homolog of *Xyn11A*	Xylanase. CAZy family GH11.	25.3	231	04223-RA	—	Numerous paralogs.
SNOG_30077	—	—	7.1	66	03763-RA	F39A, P49L, S60A, RACC63VSSRESRMRVDTILMLLYSALAAHLVPVPKVGV	SNP interrupts stop codon in SN79-1087, extended protein.
SNOG_30352	—	—	8.4	79	07626-RA	A12T	
SNOG_30359	—	—	8.3	76	07575-RA	—	—

Table 1. A functional summary of PnPf2-regulated candidate effector genes and their status in *P. nodorum* SN19-1087.

SnTox3 were the only effector genes categorised in their respective expression profile categories (Fig. 2). The expression profiles of these candidate effector genes in SN15 and *pf2-69* three days post-infection were validated using qRT-PCR (Supplementary Fig. S1). Apart from SNOG_10736, SNOG_13939 and SNOG_02980, the qRT-PCR-based expression profile of all other candidate effector genes between SN15 and *pf2-69* was consistent with findings from the RNAseq data. The expression profiles of the 12 candidate effector genes in SN15 were examined between three and 10 days post-infection using available microarray gene expression data[18] and qRT-PCR analyses performed in this study (Supplementary Fig. S2). SNOG_08150, SNOG_13939, SNOG_30077, SNOG_30352 and SNOG_30359 demonstrated similar expression profiles to *SnToxA*, *SnTox1* and *SnTox3* where gene expression peaked at three dpi and decreased to almost non-detectable levels at seven and 10 dpi, coinciding with host tissue necrosis.

Four of the 12 candidate effectors possess Pfam domains (Supplementary Data S4 and Table 1). SNOG_01146 and SNOG_15270 possess a glycosyl hydrolase family domain. SNOG_02980 and SNOG_14243 both possess a hydrolase-type esterase family domain. A BLAST search of PHIbase[19] indicated that SNOG_01146 displays significant amino acid sequence similarities to MoCDIP4 (*Magnaporthe oryzae* cell death–inducing protein P4) of the rice blast fungus *M. oryzae*[20] whereas SNOG_15270 is similar to the *Botrytis cinerea* partial virulence determinant gene *Xyn11A* which encodes a xylanase[21]. Pfam domains were not observed for the other six candidates (Supplementary Data S4 and Table 1) but SNOG_08150, 12350, 30352, 30359 and 30077 encode small cysteine-rich (<20 kDa) proteins and BlastP analyses of SNOG_02755. 08150, 10736, 12350 and 13939 revealed significant hits to other fungal hypothetical proteins, whereas SNOG_30352, 30359 and 30077 appear to be unique to *P. nodorum* based on tBlastN searches.

P. nodorum SN79-1087 is non-pathogenic on wheat and lacks *SnToxA*, *1* and *3*[5]. We decided to investigate if these 12 candidate effectors are present or altered in SN79-1087[22]. BlastP and tBlastN analysis revealed five genes were identical between SN15 and SN79-1087. SNOG_02755 and 10736 are also present in SN79-1087, but both have in-frame deletions in low-complexity amino acid repeat regions. Changes in amino acid sequence were observed for seven gene homologs in SN79-1087 (Table 1). Frame shifts or premature stop codons were not observed for these genes.

PnPf2 regulates depolymerase and nutrient assimilation gene expression *in planta*. To investigate changes in overall biochemical processes between SN15 and *pf2-69* during *iv* and *ip* growth, we assessed DE genes for enrichment of GO terms[23] (Fig. 3). GO terms were assigned to all genes where possible using InterProScan[24] and dbCAN[25].

During *iv* growth, genes categorised under oxidoreductase activities, flavin adenine dinucleotide binding and catalytic activity were significantly up-regulated in *pf2-69* (Fig. 3a). The majority of these genes encode cytochrome P450s, FAD binding proteins and oxidases (Supplementary Data S3 and S5). GO network analysis

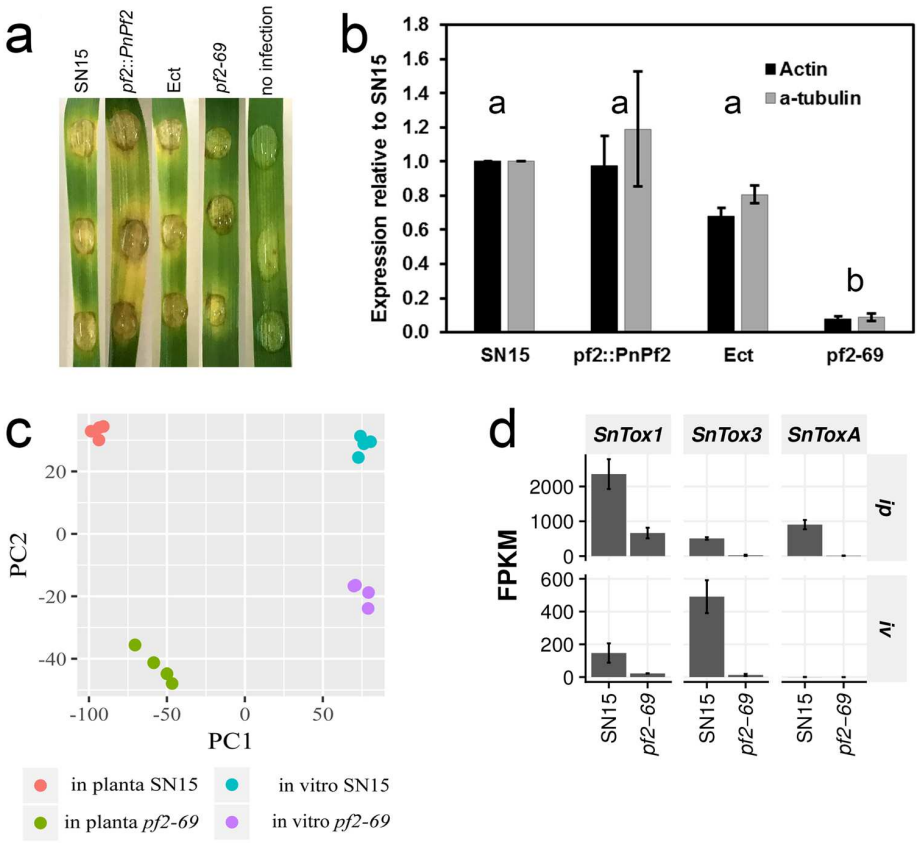


Figure 1. Infection, biomass and RNAseq analysis. (**a**) The onset of chlorotic symptoms was observed for SN15, *pf2-69*, Ect and *pf2*::*PnPf2*-infected wheat at three dpi. (**b**) Q-PCR quantification of biomass via fungal gDNA. Average biomass levels determined from PCR amplification of actin and α-tubulin not connected by the same letter are significantly different ($P < 0.05$) based on ANOVA ($n = 3$). (**c**) Comparing transcriptomes of SN15 and *pf2-69* sampled *in vitro* and *in planta* using PCA from DESeq. 2 normalised reads (Supplementary Data S2). PnPf2 plays a bigger regulatory role during infection compared to *in vitro* growth. PC1 and PC2 explains 71% and 12% of the total variance, respectively. (**d**) Comparative RNAseq expression profiling of *SnToxA*, *SnTox1* and *SnTox3* in SN15 and *pf2-69* under *in vitro* (*iv*) and *in planta* (*ip*) conditions. Bars show mean FPKM estimated by Cufflinks for each sample ($n = 4$), and error bars indicate standard deviation. Expression of all three effectors was reduced in the *in planta pf2-69* samples compared with SN15.

revealed that differentially expressed genes associated with oxidoreductase activities are central to biological processes related to respiratory electron transport chain, steroid metabolism, redox, carbohydrate metabolism and generation of precursor metabolites and energy (Fig. 3b).

During *ip* growth, molecular functions (MFs) associated with hydrolase, glucosidase and peptidase activities tended to be down-regulated in *pf2-69* (Fig. 3c). The MF hydrolase activity associated with hydrolysing N-glycosyl compounds consisted of 155 genes. Of these, the expression of 39 genes were significantly lower in *pf2-69*. Similarly, the MF associated with another hydrolase activity associated with hydrolysing N-glycosyl compounds consisted of 32 genes, of which eight genes were significantly down-regulated in *pf2-69* compared to SN15. The MF hydrolase activity consisted of 1,168 genes. Of these, 64 were expressed at lower levels in *pf2-69*. The majority of genes annotated encode plant cell wall degrading enzymes (CWDEs) and other carbohydrate depolymerases such as β-xylosidases, acetyl xylan esterases, glucanases and glucosidases (Supplementary Data S3 & S5). Arabinose is a major constituent of the plant cell wall. GO enrichment indicates that PnPf2 regulates arabinose metabolism in *P. nodorum*. Of the six genes associated with α-L-arabinofuranosidase activity, five were expressed at lower levels in *pf2-69* (Fig. 3c,d).

For protein degradation, 240 genes encode proteins with predicted peptidase activity were differentially expressed (GO:0008233) (Fig. 3c). Of these, 29 were down-regulated in *pf2-69* compared to SN15. Additionally, 63 genes encoding proteins with putative metallopeptidase activity were identified from the genome. Of these, 14 were down-regulated in *pf2-69 ip* compared to SN15. The MF associated with metallocarboxypeptidase activity (GO:0004181) consisted of nine genes where the expression of seven was reduced in *pf2-69*. For the MF associated with serine-type peptidase activity, 22 of 131 genes were expressed at lower levels in *pf2-69*. CAZyme and Interpro analyses of genes classified under GO:0008233, 0004181, 0008237 and 0008236 indicate that most encode peptidases and esterases (Supplementary Data S3 and S5).

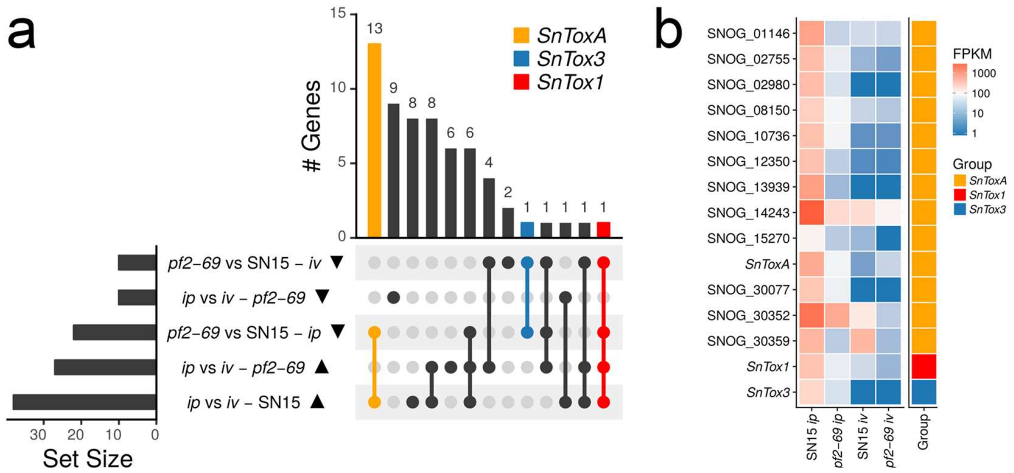


Figure 2. Identification of SN15 candidate effector genes positively regulated by PnPf2. (**a**) An UpSet plot demonstrating the number of candidate effector genes that displayed similar expression profiles. Rows in the matrix represents sets of differentially expressed effector candidates in a contrast, with the solid triangles indicating the direction of expression change. Lines connecting rows of the matrix indicate an intersection between the sets, and the vertical bar chart indicates the number of genes that are common to those sets. Set intersections containing known effectors are indicated with colour. (**b**) A heatmap showing mean FPKM ($n = 4$) profiles of the candidate effector genes that share a common expression profile with *SnToxA*.

GO analysis revealed that cellular redox potential in *pf2-69* was perturbed during *ip* growth in addition to a similar defect observed during *iv* growth. A MF associated with oxidoreductase (GO:0016491) activity was enriched in up-regulated genes in *pf2-69 ip* compared to SN15 (Fig. 3c). The majority of DE genes encoding oxidases, cytochrome P450s, reductases and dehydrogenases (Supplementary Data S3 and S5) are associated with a biological role in carbohydrate metabolism (Fig. 3d). In addition, MFs linked to transport activities were enriched with genes that were similarly up-regulated in *pf2-69 ip* compared to SN15 (Fig. 3c). Genes associated with the transport function encode sugar and amino acid transporters (Fig. 3d; Supplementary Data S3 and S5).

Identification of DNA motifs enriched in the promoters of PnPf2-regulated genes. We hypothesised that a shared *pf2-69* DE patterns implied a common transcriptional regulator. Therefore, promoters of these gene sets may harbor over-represented motif(s) functioning as potential *PnPf2* transcription factor binding site(s) (TFBS). Analysis of the promoters from the respective *pf2-69* DE gene groupings revealed three such motifs (Fig. 4). The motif WMGGVCCGAA, enriched in *pf2-69 iv* and *ip* down-regulated gene promoters, is similar to an enriched motif associated with AbPf2 down-regulated genes in *A. brassicicola*[12] and is characteristic of a Zn_2Cys_6 TFBS[26,27]. A second motif resembling a C_2H_2 TFBS (RTSYGGGGWA) was significantly enriched in *pf2-69 ip* down-regulated gene promoters. The third motif (CTGYGCCGCA) also resembled a C_2H_2 TFBS and was enriched in *pf2-69 iv* up-regulated gene promoters. The identification of unique enriched motifs in the separate datasets suggests that PnPf2 may act as an indirect regulator or its binding site specificity can be influenced by other regulators of target genes.

Absence of interaction between PnPf2 and the putative consensus motif on *SnToxA* and *SnTox3* promoters. Inspection of the *SnToxA* and *SnTox3* promoter region revealed at least one occurrence of the WMGGVCCGAA motif consensus sequences that was absent from *SnTox1*. For *SnToxA*, the consensus sequence was identified at 218, 364 and 416 bp upstream of the transcriptional start site. The consensus sequence was also observed at two sites in the *PtrToxA* promoter of *Ptr*. For *SnTox3*, the consensus sequence was identified at 679 bp upstream of the transcriptional start site. This consensus sequence was not observed in the promoter region of *SnTox1*. Therefore, it was hypothesised that WMGGVCCGAA functions as a PnPf2 binding site (Pf2BS). A yeast 1-hybrid (Y1H) assay was performed in order to determine whether PnPf2 can directly interact with the putative binding site represented in the *SnToxA* promoter. No significant interaction was observed between PnPf2 and four tandem repeats of the Pf2BS (Fig. 5a). Western blot analysis confirmed the presence of the PnPf2 protein indicating that the absence of Y1H interaction was not the result of the lack of protein (Fig. 5b).

Identification of DE TF genes. We then screened for putative TF genes that were DE between SN15 and *pf2-69* from the high confidence DE gene set to explore the possibility that PnPf2 operates indirectly. We limited our search to genes that encode proteins with TF domains found in fungi[28]. A total of 20 DE putative TFs were identified covering both *iv* and *ip* treatments. Based on distinct InterPro classifications[24], this set consisted of five basic leucine zippers, one zinc knuckle, one myc-type, one CCHC-type, one p53-like, one C_2H_2, one homeodomain-like, six fungal specific Zn_2Cys_6 and three unspecified fungal TFs (Table 2). A BLAST search of PHIbase[19] revealed that seven of these DE TF genes have strong matches to other fungal TFs associated with virulence (Table 2). Three of these belong to the fungal-specific Zn_2Cys_6 class (SNOG_03490, 07307 and 08440), one homeodomain-like (SNOG_08237) and three basic-leucine zippers (SNOG_04486, 13689 and 16487).

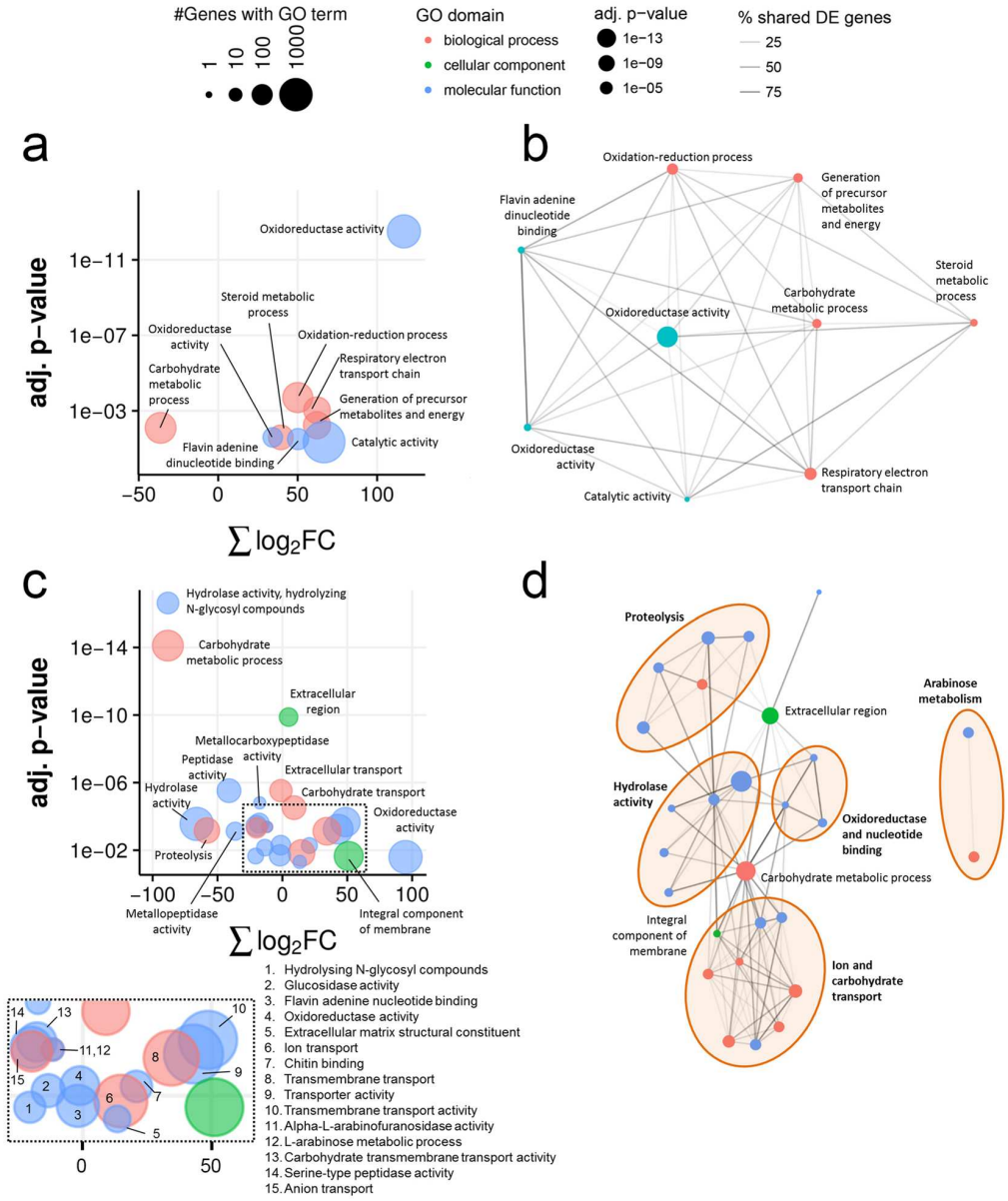


Figure 3. An illustrated summary of gene ontology (GO) term enrichment analysis between *pf2-69* and SN15 *in vitro* (**a**,**b**) and *in planta* (c and d). Bubble plots illustrate GO terms that were over-represented in differentially expressed (DE) genes as the sum of all $\log_2$ fold changes of *pf2-69* relative to SN15 and the statistical significance of enrichment tests for GO terms (**a**,**c**). Bubble size indicates the number of genes annotated with that GO term. The relationships between significantly over-represented GO terms were highlighted using network analysis (**b**,**d**). Nodes represent a single GO term and are connected if a DE gene is annotated with both terms, with the shade of the edges indicating the proportion of genes with both GO terms that are differentially expressed. Node sizes indicate the statistical significance of GO term enrichment tests. Detailed GO term analysis is deposited as Supplementary Data S5. Interactive GO enrichment and network plots are deposited as Supplementary Data S6.

Discussion

Regulation of downstream target genes including those that encode effector-like proteins by members of the Pf2 Zn_2Cys_6 family was first reported in Cho *et al.*[12] in *A. brassisicola*. The comparative RNAseq approach employed in that study derived from *A. brassicicola*-infected *A. thaliana* tissue which yielded a total of 8.5 to 9.3 × 10^5 reads from the WT and *abpf2* mutant sample (approximately 0.5% of total reads) that mapped to the *A. brassicola* genome, respectively. Much higher fungal read counts were obtained in this study through the use of deep sequencing across four biological replicates resulting in more read information to exhaustively identify DE genes between SN15 and *pf2-69* during *ip* growth (Supplementary Table S1).

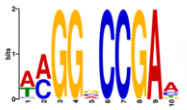
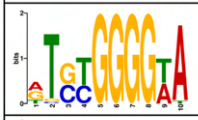
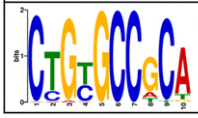
Motif positional weight matrix	Predicted from promoter set	Fungal matches JASPAR NR 2018 Database		Significantly enriched in promoter sets ($P < 0.05$)
	*pf2-69*_down	MA0429.1 (YLL054C)	Zn_2Cys_6	*pf2-69_ip*_down *pf2-69_iv*_down
	*pf2-69_ip*_down	MA0339.1 (MIG3) MA0441.1 (ZMS1) MA0338.1 (MIG2) MA0431.1 (TDA9)	C_2H_2 C_2H_2 C_2H_2 C_2H_2	*pf2-69_ip*_down
	*pf2-69_iv*_up	MA0394.1 (STP 1) MA0395.1 (STP2)	C_2H_2 C_2H_2	*pf2-69_iv*_up

Figure 4. Identification of motifs displaying enrichment in promoters of DE genes. Promoter sets used to model the motifs are listed in the second column. The third column contains motif matches to known fungal TFBSs in the JASPAR 2018 non-redundant database and their associated TF family. The fourth column lists the treatment groups displaying enrichment of the respective motif in the promoter set. Motif frequency and statistical analysis are described in Supplementary Data S7.

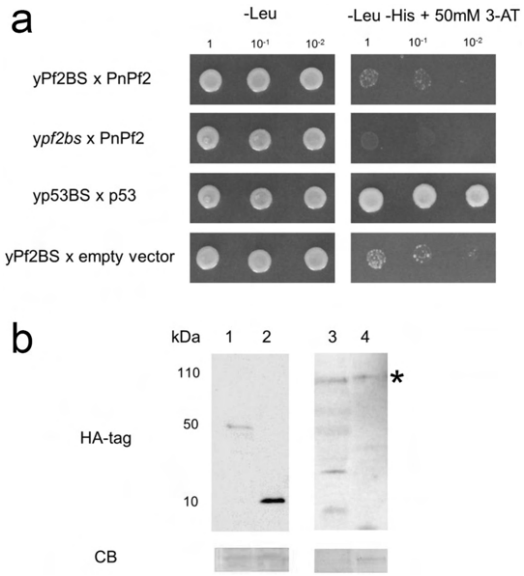


Figure 5. Y1H analysis of PnPf2 and putative promoter motif interaction. (**a**) yPf2BS expressing PnPf2 grew on the -Leu medium; however, was not able to grow on the -Leu -His medium. 50 mM 3-AT was added to the -Leu -His plate for preventing possible histidine leakage. Mutated PnPf2 binding site (pf2bs) and empty vector were manipulated as negative controls. *p53* interaction was used as positive control. Dilutions of yeast cells are indicated. (**b**) Western blots using HA antibody and coomassie blue (CB) staining of yeast cell extracts. 1: yp53BS x p53. 2: yPf2BS x empty vector. 3: yPf2BS x PnPf2. 4: ypf2bs x PnPf2. The PnPf2-GAL4AD-HA tag protein band is indicated (*). These are cropped images from different gel and blot photos. Original photos are supplied as Supplementary Fig. S3 where possible.

RNAseq confirmed *SnToxA* and *SnTox3* down-regulation in *pf2-69* but the expression of *SnTox1* was significantly higher in SN15 than *pf2-69* than our previous observation[9]. The possibility that PnPf2 plays a minor regulatory role in *SnTox1* regulation requires further investigation. Culture filtrates derived from *pnpf2* mutants caused chlorosis on *Snn1* wheat lines although the symptom was slightly weaker than with SN15[9]. Nevertheless, *SnTox1* is still strongly expressed in *pf2-69* during infection and is sufficient to produce detectable SnTox1 activity in the culture filtrate and confer virulence on *Snn1* wheat lines[9].

It is not known if *A. brassicicola* uses effectors to modulate host infection. However, Cho *et al.*[12] identified eight genes that encode small-secreted proteins with effector-like hallmarks positively regulated by AbPf2. Candidate effector genes were identified in this study that showed the same differential expression patterns as *SnToxA* and have effector-like properties. One of the effector candidates displayed significant sequence similarities to a known effector and pathogenicity factor. SNOG_01146 possesses a glycosyl hydrolase 61 domain and showed amino acid

Gene	*pf2-69* vs SN15 (*ip*)	*pf2-69* vs SN15 (*iv*)	Interpro description	Top PHI-BLAST gene hit (Pathogen*)	Mutant phenotype	E value (% aa identity)	Reference
SNOG_00166	same	down	Basic-leucine zipper domain	GzbZIP020 (*Fg*)	Unaffected pathogenicity	3.34E-52 (50)	Son, *et al.*[40]
SNOG_00439	down	same	Transcription factor domain, fungi	GzZC252 (*Fg*)	Unaffected pathogenicity	1E-150 (42)	Son, *et al.*[40]
SNOG_03490	same	down	Zn_2C_6 fungal-type DNA-binding domain	GzZC232 (*Fg*)	Reduced virulence	0 (52)	Son, *et al.*[40]
—	—	—	—	*MoPRO1* (*Mo*)	Unaffected pathogenicity	0 (52)	Lu, *et al.*[79]
—	—	—	—	*ProA* (*Ef*)	Hypervirulence	0 (45)	Tanaka, *et al.*[80]
SNOG_04486	same	down	Basic leucine zipper domain	GzbZIP001 (*Fg*)	Reduced virulence	8.3E-118 (40)	Son, *et al.*[40]
SNOG_05500	up	same	Zinc knuckle CX2CX4HX4C	GzCCHC008 (*Fg*)	Unaffected pathogenicity	1.62E-98 (40)	Son, *et al.*[40]
SNOG_06105	same	up	Transcription factor domain	GzZC238 (*Fg*)	Unaffected pathogenicity	7.1E-139 (45)	Son, *et al.*[40]
SNOG_07070	same	up	Zn_2C_6 fungal-type DNA-binding domain	GzZC211 (*Fg*)	Unaffected pathogenicity	3.11E-63 (28)	Son, *et al.*[40]
SNOG_07307	same	up	Zn_2C_6 fungal-type DNA-binding domain	*Cca1* (*Mo*)	Loss of pathogenicity	1.2E-23 (53)	Lu, *et al.*[79]
SNOG_07556	up	same	Myc-type, basic helix-loop-helix (bHLH) domain	GzbHLH014 (*Fg*)	Unaffected pathogenicity	2.06E-08 (36)	Son, *et al.*[40]
SNOG_08237	down	same	Homeodomain-like	*MoHox5* (*Mo*)	Reduced virulence	1.23E-79 (57)	Kim, *et al.*[81]
—	—	—	—	GzHOME004 (*Fg*)	Unaffected pathogenicity	2.21E-64 (51)	Son, *et al.*[40]
SNOG_08440	same	down	Zn_2C_6 fungal-type DNA-binding domain	*AtrR* (*Af*)	Reduced virulence	4.9E-34 (25)	Hagiwara, *et al.*[41]
SNOG_08565	up	same	Zn_2C_6 fungal-type DNA-binding domain	GzZC243 (*Fg*)	Unaffected pathogenicity	1.1E-117 (39)	Son, *et al.*[40]
SNOG_11322	up	same	Zinc finger, CCHC-type	GzCCHC008 (*Fg*)	Unaffected pathogenicity	4.71E-33 (39)	Son, *et al.*[40]
SNOG_12086	up	same	Zn_2C_6 fungal-type DNA-binding domain	FZC87 (*Mo*)	Unaffected pathogenicity	1.09E-25 (45)	Son, *et al.*[40]
SNOG_12740	same	up	p53-like transcription factor	GzP53L005 (*Fg*)	Unaffected pathogenicity	1.88E-29 (43)	Son, *et al.*[40]
SNOG_13359	up	up	Basic-leucine zipper domain	*atfD* (*Af*)	Unaffected pathogenicity	1.47E-10 (35)	Pereira Silva *et al.*[82]
SNOG_13689	up	up	Basic-leucine zipper domain	*CgAP1* (*Cg*)	Loss of pathogenicity	2.97E-10 (47)	Li *et al.*[83]
SNOG_15627	same	up	Zinc finger, C_2H_2	GzC2H091 (*Fg*)	Unaffected pathogenicity	1.15E-17 (27)	Son, *et al.*[40]
SNOG_16487	up	same	Basic-leucine zipper domain	GzbZIP007 (*Fg*)	Reduced virulence	2.05E-14 (56)	Son, *et al.*[40]
SNOG_30247	up	same	Transcription factor domain, fungi	GzZC239 (*Fg*)	Unaffected pathogenicity	2.51E-17 (23)	Son, *et al.*[40]

Table 2. A description of DE putative *P. nodorum* TF genes, domains and amino acid (aa) identity to characterised orthologs in other fungal pathogens. **Fg*, *Fusarium graminearum*; *Mo*, *Magnaporthe oryzae*; *Af*, *Aspergillus fumigatus*; *Ef*, *Epichloe festucae*; *Cg*, *Colletotrichum gloeosporioides*.

similarity to MoCDIP4. MoCDIP4 was identified as an apoplastic effector secreted by *M. oryzae* that causes cell death in rice[20]. Moreover, MoCDIP4 also induces cell death in non-host eudicots. In addition, these effectors are small, cysteine rich and expressed highly during early infection. SNOG_15270 is an homolog of *Xyn11A* which encodes an endo-ß-1,4-xylanase in *B. cinerea*. Deletion of *Xyn11A* in *B. cinerea* caused a significant reduction in virulence and growth on xylan[21]. All 12 candidate effector genes are also present in SN79-1087. Seven of these candidate proteins encode altered protein sequences in SN79-1087, which may explain some difference in pathogenicity. Five proteins possess changes in amino acid residues. It was previously observed that ToxA isoforms differ greatly in necrosis-inducing activities on *Tsn1* wheats and affect the speed of asexual sporulation[29]. It is interesting to note that SNOG_02755 and 10736 polypeptides contain short amino acid sequence repeats that are partially deleted in SN79-1087. Several well-studied fungal and oomycete effectors contain repeats that possess functional roles in cellular localisation, host recognition and plant cell wall binding[30]. Additionally, recent studies have indicated that differential expression of effector genes between *P. nodorum* isolates affect their contributions to SNB of wheat[15,31,32]. The expression of these candidate genes in SN79-1087 requires further study.

GO enrichment revealed that PnPf2 functions as a positive regulator of a large subset of plant CWDEs and proteases during infection. Additionally, the removal of *PnPf2* resulted in a general up-regulation in expression of nutrient transporter genes during infection. It is still not known whether this change is caused directly by the absence of PnPf2, or indirectly via another mechanism regulated by PnPf2. Comparative transcriptomic analysis of *A. brassicicola* identified only 13 genes that encode hydrolytic enzymes including two pectate lyases, were regulated by AbPf2[12]. Deep sequencing used in this study provided a higher resolution insight into CAZyme regulation exerted by the Pf2 Zn_2Cys_6 class. Quantifying the contributions of plant CWDEs to phytopathogenicity is difficult because many fungal phytopathogens possess expanded gene families that result in functional redundancies[33]. For example, early studies on the causal agent of northern leaf spot of maize *Cochliobolus carbonum* (eg.[34–37]) did not find a clear role for CWDEs in fungal virulence. This is not to imply that CWDEs are dispensable for fungal virulence. It was reported that feruloyl esterases from *Valsa mali*[38], a AbPf2-regulated pectate lyase from *A. brassicicola*[39] and an endo-β-1,4-xylanase from *B. cinerea*[21] function as virulence factors. Since plant CWDEs deconstruct the plant cell wall and liberate simple carbohydrates for assimilation and growth, it remains to be determined if SN15 can outcompete *pf2-69* during co-infection on *Snn1* wheats as the former can express a much larger repertoire of extracellular hydrolytic enzymes. RNAseq read counts suggested that *pf2-69* accumulated much less biomass than SN15 at three dpi. This is surprising as *pf2-69* retained the ability to causes lesions on *Snn1* wheat lines comparable to SN15 as previously observed[9]. It is probable that SnTox1 secreted by *pf2-69* during infection is the main cause of necrosis rather than the accumulation of fungal biomass at the lesion.

Analysis of the *pf2-69* DE gene sets identified three distinct over-represented motifs (Fig. 4). The most notable of these is the WMGGVCCGAA motif associated with genes under PnPf2 positive regulation, as this motif was

observed at multiple sites along the *Sn/PtrToxA* and *SnTox3* promoters and also enriched in AbPf2-regulated gene promoters[12]. We hypothesised that it functions as a PnPf2 binding site as it resembles a Zn_2Cys_6 TFBS[26,27]. However, Y1H assay indicated that PnPf2 did not bind to the motif. This suggests either PnPf2 does not function as a direct regulator of *SnToxA*, *SnTox3* and DE genes through interaction with the WMGGVCCGAA motif, or that necessary PnPf2 post-translational modifications/interactions are not compatible with the Y1H system. It was noted however that six other Zn_2Cys_6-type TF genes were differentially expressed between *pf2-69* and SN15 (Table 2). Of these, only two were down-regulated but may serve as alternate candidates for direct regulation targeting the WMGGVCCGAA motif. A BLAST search of these against PHIbase revealed pathogenicity-associated functions in fungal homologues. SNOG_03490 is 52% identical to GzZC232 of *Fusarium graminearum*, the causal agent of fusarium head blight of wheat and is required for full virulence[40]. SNOG_08440 is homologous to a Zn_2Cys_6-type TF gene *AtrR* of *Aspergillus fumigatus*, an opportunistic fungal pathogen of mammals[41]. AtrR is a regulator of ergosterol biosynthesis pathway genes most notably *Cyp51*, a target for fungicide control. Deletion of *AtrR* resulted in impaired fungal growth and attenuated virulence on mice[41]. The other enriched motifs were characteristic of C_2H_2 binding sites[27] however, only one DE TF of this class was identified - SNOG_15627 (Table 2). SNOG_15627 expression was up-regulated in *pf2-69* under *iv* condition but remained unchanged during *ip* growth. SNOG_15627 demonstrated weak similarity to a characterised TF in *F. graminearum* shown to be dispensable for pathogenicity on wheat[40]. As the CTGYGCCGCA motif was enriched in the *pf2-69 iv* up-regulated gene promoters, it is possible that SNOG_15627 functions as a direct regulator. PnCon7 is the only characterised C_2H_2 TF in *P. nodorum* involved in SnTox3-mediated disease and direct regulation[11]. However, the cis-regulatory element of PnCon7 differs to both predicted C_2H_2 binding sites observed in this study.

We propose a model to explain the role of *PnPf2* during early host infection based on evidence observed in this study (Fig. 6). The removal of PnPf2 drastically diminishes effector expression and so restricts the number of hosts on which *P. nodorum* is virulent[9] (Fig. 6a). Both mutant and wild type strains are able to infect but the reduced ability to produce effectors and cell wall degrading enzymes means that *pf2-69* is delayed in accessing bulk nutrients that come from the early stages of cell necrosis (Fig. 6b). The mutant has reduced access to nutrients stored as complex carbohydrates or compartmentalised in plant cells leading to a reduction in growth during host infection. Increased expression of transporter proteins may be an attempt to scavenge freely available nutrients possibly from the apoplastic space[42] (Fig. 6c). In addition, we have identified candidate effector genes that are homologous to virulence factors and effectors in other phytopathogens. It is evident that PnPf2 functions to coordinate the expression of a subset of DE genes identified in this study through other TFs. Studies are currently under way to functionally characterise effector candidates and DE TF genes for their role in effector regulation and pathogenicity on wheat.

Methods

Infection assays. Whole plant infection assay on two week-old wheat seedlings was performed as previously described[43]. Disease severity was visually determined and scored. A score of zero indicates no disease symptoms. A score of nine indicates a fully necrotised plant. Detached leaf infection assays on two-week old wheat cv. Halberd leaves was performed as previously described[43,44].

Biomass analysis using quantitative (Q)-PCR. Q-PCR was to determine fungal biomass from infected wheat. Wheat cv. Halberd was infected with *P. nodorum* pycnidiospores as described above. Infection was allowed to develop for three days prior to sampling. The inoculated leaf section was excised and collected. Following this, gDNA was extracted using a Biosprint genomic DNA extraction kit (Qiagen, Venlo, Netherlands). Q-PCR was essentially carried out as described in Brouwer *et al.*[45] using the primer pair alTubulinqPCRf/r and ActinqPCRhp2F/R (Supplementary Table S3).

RNA extraction and handling. RNA isolation and *in planta* gene expression analyses were performed as described in Rybak *et al.*[9] using three day post-infected lesions excised from detached wheat. Library construction and sequencing was performed by the Ramaciotti Centre for Genomics (The University of NSW, Australia). Briefly, the TruSeq Stranded mRNA-seq method was used to prepare all libraries. Following this, sequencing was performed on an Illumina HiSeq. 2500 platform (San Diego, CA, USA) to generate 125 bp paired-end reads. Deep sequencing of all *in planta* samples were carried on individual lanes in the flowcell to ensure maximum sequence data was obtained from low fungal biomass. Samples derived from *in vitro* growth conditions were multiplexed into a single lane. The experiment was performed with four biological replicates.

RNAseq QC and read trimming. The quality of reads in the FastQ files were assessed using FastQC v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc) before trimming adapter sequences using cutadapt v1.12[46]. Adapter trimmed reads were then filtered into sets belonging to SN15 and wheat using BBSplit v36.67 (https://sourceforge.net/projects/bbmap) using the *P. nodorum* genome[16]. Fungal reads were aligned to the SN15 genome using STAR v2.5.0a[47]. Novel splice sites were identified in a first pass alignment of the adapter-trimmed reads of all samples combined. Sample reads were then aligned individually using the novel splice sites identified in the first pass.

Determining differential gene expression in RNAseq. Fragments overlapping annotated features in the genome were counted using the SubRead featureCount v1.5.1 program using the union mode[48]. Differentially expressed (DE) genes were determined using the R packages EdgeR v3.16.4[49], DESeq. 2 v1.14.1[50] and Limma v3.30.6[51]. DE genes were determined from tests of log_2 fold changes (LFC) against the null hypothesis $-1 \leq LFC \leq 1$ (i.e. $H_a = |LFC| > 1$) using a BH-adjusted *P*-value significance threshold of 0.05. Tests were also

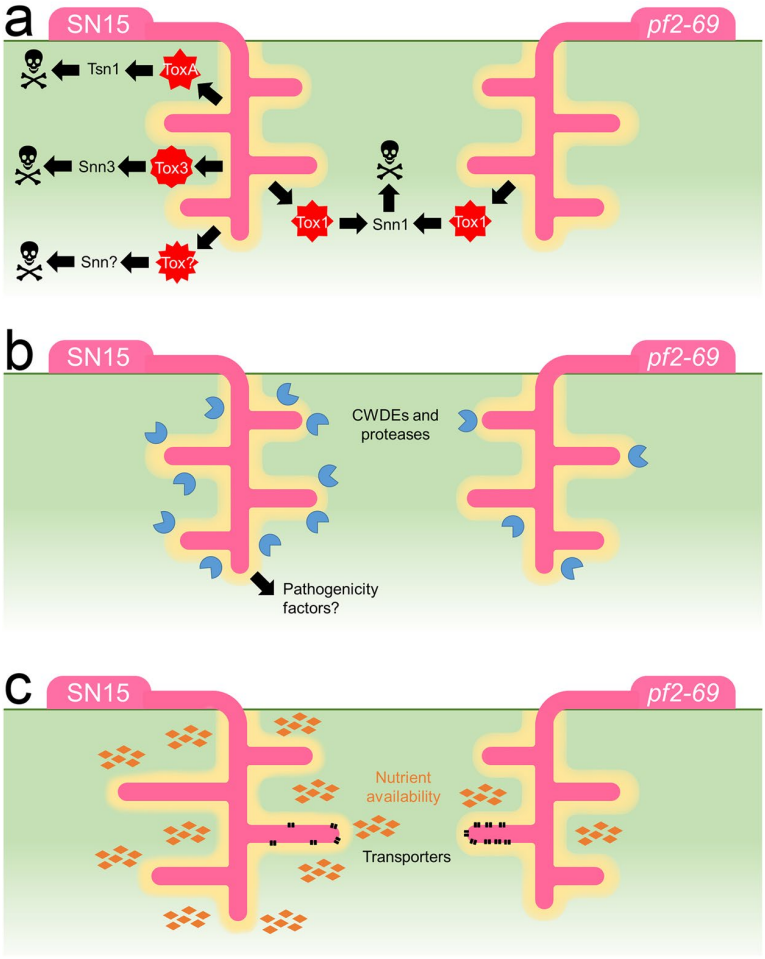


Figure 6. The proposed model for the role of PnPf2 during infection. (**a**) The deletion of *PnPf2* diminishes effector expression and effector-triggered susceptibility in wheat carrying *Tsn1* and *Snn3*. (**b**) In addition, PnPf2 functions as a positive regulator of CWDE expression *in planta*. (**c**) Without the full complement of CWDEs being produced, *pf2-69* has a reduced ability to breakdown plant polymers and complex carbohydrates for assimilation needed during *in planta* growth.

performed against the null hypothesis $LFC \neq 0$, to be used where greater sensitivity (but lower confidence) was required. Unless otherwise specified, all results refer to thresholded tests ($|LFC| > 1$). Genes that were determined to be DE from tests by all three programs were taken as high-confidence DE sets. For contrasts involving samples with fewer than 10 million fragments (*pf2-69 ip*), genes from these samples were required to have a minimum of 10 CPM to be considered as differentially expressed in the high confidence sets. Normalised FPKM statistics were obtained using Cufflinks v2.2.1[52]. Genes with mean FPKM > 100 were considered to be highly expressed.

Functional annotation. Functional annotations for existing genes were determined using InterProScan v5.19-58.0[24]. Additionally, carbohydrate active enzymes were predicted using HMMER v3.1b2 (hmmer.org) and dbCAN v5[25]. GO terms[23] for each gene were found from combined dbCAN and InterProScan results, including matches from: Pfam[53], TIGRFAM[54], SMART[55], PIRSF[56], PANTHER[57], HAMAP[58], Prosite[59], ProDom[60], PRINTS[61], and CATH-Gene3D[62]. Likely protein locations were determined using SignalP v4.1[63], TargetP v1.1[64], and TMHMM v2.0c[65]. Predicted proteins with a signal peptide and no transmembrane domains outside of the first 27 amino acids were considered to be secreted. Proteins with effector-like properties were determined using EffectorP v1.0[66] and were considered to be effector-like if they were also predicted to be secreted using the criteria above. Candidate genes were searched for in SN79-1087 (NCBI, GCA_002267025.1) using Spaln v2.3.3[67]. Overlapping SN79-1087 genes were extracted and protein sequences were aligned using the needle command using EMBOSS[68].

Functional enrichment of differentially expressed genes. Over-representation of GO terms in high-confidence differentially expressed gene sets were performed using the R package Goseq v1.26.0[69]. Due to differences in the ability of DESeq. 2, EdgeR, and Limma to handle features with few aligned fragments;

enrichment of effector-like or secreted transcripts were determined using the union of differentially expressed genes from all three prediction packages.

QRT-PCR determination of gene expression. Total RNA extraction from infected wheat cv. Halberd and *P. nodorum* mycelia from *in vitro* growth was extracted as described earlier. QRT-PCR was performed using a Quantitect SYBR Green RT-PCR kit (Qiagen, Valencia, CA, USA) and a Bio-Rad (Hercules, CA, USA) CFX96 system. *P. nodorum* SN15 gDNA was used as a quantitative standard. The expression value of each gene was normalised against the housekeeping gene actin (*Act1*) using the primer pair ActinqPCRf and ActinqPCRr[70].

Analysis of promoters for enriched motifs. Common DNA motifs were discovered from the promoter regions 1.5 kbp upstream (or to the next annotated gene) of predicted transcription start sites of DE genes. Weeder 2.0[71] was used to search for enriched motifs in these promoters. A full set of SN15 predicted gene promoters was used for background frequencies with the redundancy filter set at 0.5. Utilising the consensus option in MEME v5.0.1[72], position weight matrices (PWMs) for top non-redundant motifs from each subset were derived for downstream analysis with MEMEsuite tools[73]. Each PWM motif was assessed for overrepresentation in *pf2-69* DE subsets similar to Cho *et al.*[12]. Motif occurrences were first counted using FIMO[74] and promoters with at least one occurrence were regarded as positive. Significance of over-representation in DE gene promoter sets was determined using Fisher's exact test with Bonferroni corrected *P*-values ($P_{adj} < 0.05$)[75] as compared with the full promoter set of SN15. TOMTOM[76] was used to search the JASPAR NR 2018 databases for matches ($E < 1$) to published fungal TFBSs in order to characterise the over-represented motifs.

Y1H assay. The construction of yeast reporter strain and Y1H screening was carried out based on the method of Ouwerkerk and Meijer[77] with modifications. Y1H bait constructs were prepared by cloning three repeats of the p53 binding site (p53BS) (5′-AGACATGCCT-3′) using the primer pair p53BS-F1/R1[78], four repeats of the putative *SnToxA* PnPf2 binding site (Pf2BS) (5′-AAGGACCGA-3′) using the primer pair Pf2BS-F1/R1 and four repeats of pf2bs (5′-AAGGAAATA-3′) using the primer pair pf2bs-F1/R1 into pINT1-HIS3NB (provided by Dr. P.B.F Ouwerkerk, Leiden University) (Supplementary Table S3). Repeats of binding sites were cloned into pINT1-HIS3NB. Each construct was linearised, transformed into the yeast strain Y187 (Clontech, CA, USA) and selected on YPAD supplemented with G418. Bait strains were grown on selective media (-His) containing 3-amino-1,2,4-triazole (Sigma-Aldrich, MO, USA). Mating of the yeast bait strains with the prey strains was conducted by mixing the two strains together and grown on YPAD medium. Confirmation of the specific interaction between the bait sequence and the target protein was performed by reintroduce the prey construct into the bait strain. The prey construct pGADT7-p53 was built by cloning partial *p53* from pGBKT7-53 (Clontech, CA, USA) into pGADT7. Similarly, *PnPf2* was amplified from cDNA using the primer pair Pf2-F2/R3 and ligated into pGADT7.

Data availability

All data generated or analysed during this study are included in this published article (and its Supplementary Information files)

Received: 3 July 2019; Accepted: 17 October 2019;
Published online: 04 November 2019

References

1. Friesen, T. L. *et al.* Emergence of a new disease as a result of interspecific virulence gene transfer. *Nature Genetics* **38**, 953–956 (2006).
2. Faris, J. D. *et al.* A unique wheat disease resistance-like gene governs effector-triggered susceptibility to necrotrophic pathogens. *Proceedings of the National Academy of Sciences USA* **107**, 13544–13549 (2010).
3. Ciuffetti, L. M., Tuori, R. P. & Gaventa, J. M. A single gene encodes a selective toxin causal to the development of tan spot of wheat. *Plant Cell* **9**, 135–144 (1997).
4. McDonald, M. C., Ahren, D., Simpfendorfer, S., Milgate, A. & Solomon, P. S. The discovery of the virulence gene ToxA in the wheat and barley pathogen Bipolaris sorokiniana. *Mol. Plant Pathol.* **19**, 432–439 (2018).
5. Liu, Z. *et al.* The cysteine rich necrotrophic effector SnTox1 produced by *Stagonospora nodorum* triggers susceptibility of wheat lines harboring *Snn1*. *PLoS Pathogens* **8**, e1002467 (2012).
6. Liu, Z. *et al.* SnTox3 acts in effector triggered susceptibility to induce disease on wheat carrying the *Snn3* gene. *PLoS Pathogens* **5**, e1000581 (2009).
7. Zhang, Z. *et al.* Two putatively homoeologous wheat genes mediate recognition of SnTox3 to confer effector-triggered susceptibility to *Stagonospora nodorum*. *The Plant Journal* **65**, 27–38 (2011).
8. Friesen, T. L., Faris, J. D., Solomon, P. S. & Oliver, R. P. Host-specific toxins: effectors of necrotrophic pathogenicity. *Cellular Microbiology* **10**, 1421–1428 (2008).
9. Rybak, K. *et al.* A functionally conserved Zn2 Cys6 binuclear cluster transcription factor class regulates necrotrophic effector gene expression and host-specific virulence of two major Pleosporales fungal pathogens of wheat. *Mol. Plant Pathol.* **18**, 420–434 (2017).
10. IpCho, S. V. *et al.* The transcription factor StuA regulates central carbon metabolism, mycotoxin production, and effector gene expression in the wheat pathogen *Stagonospora nodorum*. *Eukaryot. Cell* **9**, 1100–1108 (2010).
11. Lin, S. Y., Chooi, Y. H. & Solomon, P. S. The global regulator of pathogenesis PnCon7 positively regulates Tox3 effector gene expression through direct interaction in the wheat pathogen Parastagonospora nodorum. *Mol. Microbiol* (2018).
12. Cho, Y., Ohm, R. A., Grigoriev, I. V. & Srivastava, A. Fungal-specific transcription factor AbPf2 activates pathogenicity in *Alternaria brassicicola*. *The Plant Journal* **75**, 498–514 (2013).
13. Tan, K. C. *et al.* Functional redundancy of necrotrophic effectors - consequences for exploitation for breeding. *Frontiers in Plant Science* **6**, 501 (2015).
14. Phan, H. T. T. *et al.* Novel sources of resistance to Septoria nodorum blotch in the Vavilov wheat collection identified by genome-wide association studies. *Theor Appl Genet* (2018).
15. Phan, H. T. *et al.* Differential effector gene expression underpins epistasis in a plant fungal disease. *The Plant Journal* **87**, 343–354 (2016).

16. Syme, R. A. *et al.* Comprehensive annotation of the *Parastagonospora nodorum* reference genome using next-generation genomics, transcriptomics and proteogenomics. *PLoS One* **11**, e0147221 (2016).
17. Sperschneider, J. *et al.* EffectorP: predicting fungal effector proteins from secretomes using machine learning. *New Phytologist* **210**, 743–761 (2016).
18. Ipcho, S. V. *et al.* Transcriptome analysis of *Stagonospora nodorum*: gene models, effectors, metabolism and pantothenate dispensability. *Mol. Plant Pathol.* **13**, 531–545 (2012).
19. Urban, M. *et al.* PHI-base: a new interface and further additions for the multi-species pathogen-host interactions database. *Nucleic Acids Res.* **45**, D604–D610 (2017).
20. Chen, S. *et al.* Identification and characterization of in planta-expressed secreted effector proteins from Magnaporthe oryzae that induce cell death in rice. *Mol Plant Microbe Interact* **26**, 191–202 (2013).
21. Brito, N., Espino, J. J. & Gonzalez, C. The endo-beta-1,4-xylanase Xyn11A is required for virulence in *Botrytis cinerea*. *Mol. Plant-Microbe Interact.* **19**, 25–32 (2006).
22. Richards, J. K., Wyatt, N. A., Liu, Z., Faris, J. D. & Friesen, T. L. Reference Quality Genome Assemblies of Three Parastagonospora nodorum Isolates Differing in Virulence on Wheat. *G3 (Bethesda)* **8**, 393–399 (2018).
23. Ashburner, M. *et al.* Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25**, 25–29 (2000).
24. Jones, P. *et al.* InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236–1240 (2014).
25. Yin, Y. *et al.* dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res.* **40**, W445–451 (2012).
26. MacPherson, S., Larochelle, M. & Turcotte, B. A fungal family of transcriptional regulators: the zinc cluster proteins. *Microbiology and molecular biology reviews: MMBR* **70**, 583–604 (2006).
27. Khan, A. *et al.* JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework. *Nucleic Acids Res.* **46**, D1284 (2018).
28. Shelest, E. Transcription Factors in Fungi: TFome Dynamics, Three Major Families, and Dual-Specificity TFs. *Front Genet* **8**, 53 (2017).
29. Tan, K. C. *et al.* Quantitative variation in effector activity of ToxA isoforms from *Stagonospora nodorum* and *Pyrenophora tritici-repentis*. *Mol. Plant-Microbe Interact.* **25**, 515–522 (2012).
30. Mesarich, C. H., Bowen, J. K., Hamiaux, C. & Templeton, M. D. Repeat-containing protein effectors of plant-associated organisms. *Front Plant Sci* **6**, 872 (2015).
31. Faris, J. D., Zhang, Z., Rasmussen, J. B. & Friesen, T. L. Variable expression of the Stagonospora nodorum effector SnToxA among isolates is correlated with levels of disease in wheat. *Molecular plant-microbe interactions: MPMI* **24**, 1419–1426 (2011).
32. Peters, A. R., Zhang, Z., Richards, J. K., Friesen, T. L. & Faris, J. D. Genetics of variable disease expression conferred by inverse gene-for-gene interactions in the wheat-Parastagonospora nodorum pathosystem. *Plant Physiol* (2019).
33. Kubicek, C. P., Starr, T. L. & Glass, N. L. Plant Cell Wall-Degrading Enzymes and Their Secretion in Plant-Pathogenic Fungi. *Annual Review of Phytopathology, Vol 52* **52**, 427–451 (2014).
34. Apel-Birkhold, P. C. & Walton, J. D. Cloning, disruption, and expression of two endo-beta 1, 4-xylanase genes, XYL2 and XYL3, from Cochliobolus carbonum. *Appl. Environ. Microbiol.* **62**, 4129–4135 (1996).
35. Kim, H. *et al.* Mutational analysis of beta-glucanase genes from the plant-pathogenic fungus Cochliobolus carbonum. *Mol Plant Microbe Interact* **14**, 1436–1443 (2001).
36. Ahn, J. H., Sposato, P., Kim, S. I. & Walton, J. D. Molecular cloning and characterization of cel2 from the fungus Cochlibolus carbonum. *Biosci. Biotech. Bioch.* **65**, 1406–1411 (2001).
37. Gorlach, J. M., Van Der Knaap, E. & Walton, J. D. Cloning and targeted disruption of MLG1, a gene encoding two of three extracellular mixed-linked glucanases of Cochliobolus carbonum. *Appl. Environ. Microbiol.* **64**, 385–391 (1998).
38. Xu, M. *et al.* The feruloyl esterase genes are required for full pathogenicity of the apple tree canker pathogen Valsa mali. *Mol. Plant Pathol.* **19**, 1353–1363 (2018).
39. Cho, Y. *et al.* A Pectate Lyase-Coding Gene Abundantly Expressed during Early Stages of Infection Is Required for Full Virulence in Alternaria brassicicola. *PLoS One* **10**, e0127140 (2015).
40. Son, H. *et al.* A phenome-based functional analysis of transcription factors in the cereal head blight fungus, Fusarium graminearum. *PLoS Pathogens* **7**, e1002310 (2011).
41. Hagiwara, D. *et al.* A Novel Zn2-Cys6 Transcription Factor AtrR Plays a Key Role in an Azole Resistance Mechanism of Aspergillus fumigatus by Co-regulating cyp51A and cdr1B Expressions. *PLoS Pathog* **13**, e1006096 (2017).
42. Solomon, P. S. & Oliver, R. P. The nitrogen content of the tomato leaf apoplast increases during infection by *Cladosporium fulvum*. *Planta* **213**, 241–249 (2001).
43. Solomon, P. S., Tan, K. C. & Oliver, R. P. Mannitol 1-phosphate metabolism is required for sporulation *in planta* of the wheat pathogen *Stagonospora nodorum*. *Mol. Plant-Microbe Interact.* **18**, 110–115 (2005).
44. Benedikz, P. W., Mappledoram, C. J. & Scott, P. R. A laboratory technique for screening cereals for resistance to *Septoria nodorum* using detached seedling leaves. *Trans. Br. Mycol. Soc.* **77**, 667–668 (1981).
45. Brouwer, M. *et al.* Quantification of disease progression of several microbial pathogens on Arabidopsis thaliana using real-time fluorescence PCR. *FEMS Microbiol. Lett.* **228**, 241–248 (2003).
46. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, http://journal.embnet.org/index.php/embnetjournal/article/view/200/458 (2011).
47. Dobin, A. *et al.* STAR: ultrafast universal RNA-seq aligner. *Bioinformatics* **29**, 15–21 (2013).
48. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).
49. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).
50. Love, M. I., Huber, W. & Anders, S. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq. 2. *Genome Biol* **15**, 550 (2014).
51. Ritchie, M. E. *et al.* limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res.* **43**, e47 (2015).
52. Trapnell, C. *et al.* Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* **7**, 562–578 (2012).
53. Punta, M. *et al.* The Pfam protein families database. *Nucleic Acids Res.* **40**, D290–301 (2012).
54. Haft, D. H. *et al.* TIGRFAMs: a protein family resource for the functional identification of proteins. *Nucleic Acids Research* **29**, 41–43 (2001).
55. Letunic, I., Doerks, T. & Bork, P. SMART: recent updates, new developments and status in 2015. *Nucleic Acids Res.* **43**, D257–260 (2015).
56. Wu, C. H. *et al.* PIRSF: family classification system at the Protein Information Resource. *Nucleic Acids Res.* **32**, D112–114 (2004).
57. Mi, H., Poudel, S., Muruganujan, A., Casagrande, J. T. & Thomas, P. D. PANTHER version 10: expanded protein families and functions, and analysis tools. *Nucleic Acids Res.* **44**, D336–342 (2016).
58. Pedruzzi, I. *et al.* HAMAP in 2015: updates to the protein family classification and annotation system. *Nucleic Acids Res.* **43**, D1064–1070 (2015).
59. Sigrist, C. J. A. *et al.* New and continuing developments at PROSITE. *Nucleic Acids Research* **41**, E344–E347 (2013).
60. Bru, C. *et al.* The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.* **33**, D212–215 (2005).
61. Attwood, T. K. *et al.* The PRINTS database: a fine-grained protein sequence annotation and analysis resource–its status in 2012. *Database (Oxford)* **2012**, bas019 (2012).

62. Dawson, N. L. *et al*. CATH: an expanded resource to predict protein function through structure and sequence. *Nucleic Acids Research* **45**, D289–D295 (2017).
63. Petersen, T. N., Brunak, S., von Heijne, G. & Nielsen, H. SignalP 4.0: discriminating signal peptides from transmembrane regions. *Nat Methods* **8**, 785–786 (2011).
64. Emanuelsson, O., Nielsen, H., Brunak, S. & von Heijne, G. Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. *Journal of Molecular Biology* **300**, 1005–1016 (2000).
65. Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology* **305**, 567–580 (2001).
66. Sperschneider, J. *et al*. EffectorP: predicting fungal effector proteins from secretomes using machine learning. *The New phytologist* (2015).
67. Gotoh, O. A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Res.* **36**, 2630–2638 (2008).
68. Rice, P., Longden, I. & Bleasby, A. EMBOSS: the European Molecular Biology Open Software Suite. *Trends Genet* **16**, 276–277 (2000).
69. Young, M. D., Wakefield, M. J., Smyth, G. K. & Oshlack, A. Gene ontology analysis for RNA-seq: accounting for selection bias. *Genome Biol* **11**, R14 (2010).
70. Tan, K. C. *et al*. A signaling-regulated, short-chain dehydrogenase of *Stagonospora nodorum* regulates asexual development. *Eukaryot. Cell* **7**, 1916–1929 (2008).
71. Pavesi, G., Mereghetti, P., Mauri, G. & Pesole, G. Weeder Web: discovery of transcription factor binding sites in a set of sequences from co-regulated genes. *Nucleic Acids Res.* **32**, W199–203 (2004).
72. Bailey, T. L. & Elkan, C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc. Int. Conf. Intell. Syst. Mol. Biol.* **2**, 28–36 (1994).
73. Bailey, T. L. *et al*. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–208 (2009).
74. Grant, C. E., Bailey, T. L. & Noble, W. S. FIMO: scanning for occurrences of a given motif. *Bioinformatics* **27**, 1017–1018 (2011).
75. Armstrong, R. A. When to use the Bonferroni correction. *Ophthal Physl Opt* **34**, 502–508 (2014).
76. Gupta, S., Stamatoyannopoulos, J. A., Bailey, T. L. & Noble, W. S. Quantifying similarity between motifs. *Genome Biology* **8** (2007).
77. Ouwerkerk, P. B. & Meijer, A. H. Yeast one-hybrid screens for detection of transcription factor DNA interactions. *Methods Mol Biol* **678**, 211–227 (2011).
78. Wang, Y., Schwedes, J. F., Parks, D., Mann, K. & Tegtmeyer, P. Interaction of p53 with its consensus DNA-binding site. *Mol. Cell. Biol.* **15**, 2157–2165 (1995).
79. Lu, J., Cao, H., Zhang, L., Huang, P. & Lin, F. Systematic analysis of Zn2Cys6 transcription factors required for development and pathogenicity by high-throughput gene knockout in the rice blast fungus. *PLoS Pathogens* **10**, e1004432 (2014).
80. Tanaka, A. *et al*. ProA, a transcriptional regulator of fungal fruiting body development, regulates leaf hyphal network development in the Epichloe festucae-Lolium perenne symbiosis. *Mol. Microbiol.* **90**, 551–568 (2013).
81. Kim, S. *et al*. Homeobox transcription factors are required for conidiation and appressorium development in the rice blast fungus Magnaporthe oryzae. *PLoS Genet* **5**, e1000757 (2009).
82. Pereira Silva, L. *et al*. Genome-wide transcriptome analysis of Aspergillus fumigatus exposed to osmotic stress reveals regulators of osmotic and cell wall stresses that are SakA(HOG1) and MpkC dependent. *Cell Microbiol* **19** (2017).
83. Li, X., Wu, Y., Liu, Z. & Zhang, C. The function and transcriptome analysis of a bZIP transcription factor CgAP1 in Colletotrichum gloeosporioides. *Microbiol Res* **197**, 39–48 (2017).

Acknowledgements

This study was supported by the Centre for Crop and Disease Management, a joint initiative of Curtin University and the Grains Research and Development Corporation [research grant CUR00023 (Programme 3)]. EJ was supported by the Australian Government Research Training Program Scholarship. We thank Dr. James Hane for bioinformatic discussions.

Author contributions

K.C.T. conceived the experiment. D.A.B.K., E.J., K.R., H.T.T.P. and S.Y.L. performed the experiment. K.B.S., P.S.S. and R.P.O. provided additional experimental design ideas and critical feedbacks. K.C.T., D.A.B.J. and E.J. wrote the paper. D.A.B.J., E.J., K.B.S., P.S.S., R.P.O. and K.C.T. edited the paper.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41598-019-52444-7.

Correspondence and requests for materials should be addressed to K.-C.T.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

CHAPTER 10 — THEME 3

Chromosome-level genome assembly and manually-curated proteome of model necrotroph *Parastagonospora nodorum* Sn15 reveals a genome-wide trove of effector-like homologs, and redundancy of virulence-related functions within an accessory chromosome

This chapter is also published in:
BMC Genomics, 2021, vol. 22, article 382
https://doi.org/10.1186/s12864-021-07699-8

This chapter is submitted as supplementary material and should not contribute to assessment of this thesis. It is included here as an example of related research contributions made during the candidacy and as necessary context for chapter 11.

10.1 Declaration

Title	Chromosome-level genome assembly and manually-curated proteome of model necrotroph Parastagonospora nodorum Sn15 reveals a genome-wide trove of effector-like homologs, and redundancy of virulence-related functions within an accessory chromosome.
Authors	Stefania Bertazzoni, **Darcy A. B. Jones**, Huyen T. Phan, Kar-Chun Tan, and James K. Hane

This supplementary chapter may contribute toward another student's thesis (Stefania Bertazzoni) in the future. As such, it is included here to provide context and as an example of contributions to related work conducted during the candidate's Ph. D. and should not contribute to assessment. This thesis supplementary chapter is submitted in the form of a collaboratively-written manuscript ready for journal submission.

The following contributions were made by the Ph.D candidate (DABJ) and co-authors:

- K-CT and JKH conceived the experiment.
- SB, **DABJ**, and JKH performed bioinformatics analyses.
- SB and JKH wrote the manuscript.
- JKH, HTTP, and KCT edited the manuscript.
- All authors read, edited, and approved the manuscript.

I, Darcy Jones, hereby certify that this attribution statement is an accurate record of my contribution to the research presented in this chapter.

Darcy Jones

I certify that this attribution statement is an accurate record of Darcy Jones' contribution to the research presented in this chapter.

Stefania Bertazzoni | Huyen T. T. Phan | Kar-Chun Tan

James K. Hane

Bertazzoni *et al. BMC Genomics* (2021) 22:382
https://doi.org/10.1186/s12864-021-07699-8

BMC Genomics

RESEARCH ARTICLE Open Access

Chromosome-level genome assembly and manually-curated proteome of model necrotroph *Parastagonospora nodorum* Sn15 reveals a genome-wide trove of candidate effector homologs, and redundancy of virulence-related functions within an accessory chromosome

Stefania Bertazzoni[1], Darcy A. B. Jones[1], Huyen T. Phan[1*], Kar-Chun Tan[1*] and James K. Hane[1,2*]

Abstract

Background: The fungus *Parastagonospora nodorum* causes septoria nodorum blotch (SNB) of wheat (*Triticum aestivum*) and is a model species for necrotrophic plant pathogens. The genome assembly of reference isolate Sn15 was first reported in 2007. *P. nodorum* infection is promoted by its production of proteinaceous necrotrophic effectors, three of which are characterised – ToxA, Tox1 and Tox3.

Results: A chromosome-scale genome assembly of *P. nodorum* Australian reference isolate Sn15, which combined long read sequencing, optical mapping and manual curation, produced 23 chromosomes with 21 chromosomes possessing both telomeres. New transcriptome data were combined with fungal-specific gene prediction techniques and manual curation to produce a high-quality predicted gene annotation dataset, which comprises 13,869 high confidence genes, and an additional 2534 lower confidence genes retained to assist pathogenicity effector discovery. Comparison to a panel of 31 internationally-sourced isolates identified multiple hotspots within the Sn15 genome for mutation or presence-absence variation, which was used to enhance subsequent effector prediction. Effector prediction resulted in 257 candidates, of which 98 higher-ranked candidates were selected for in-depth analysis and revealed a wealth of functions related to pathogenicity. Additionally, 11 out of the 98 candidates also exhibited orthology conservation patterns that suggested lateral gene transfer with other cereal-pathogenic fungal species. Analysis of the pan-genome indicated the smallest chromosome of 0.4 Mbp length to be an accessory chromosome (AC23). AC23 was notably absent from an avirulent isolate and is predominated by mutation hotspots with an increase in non-synonymous mutations relative to other chromosomes. Surprisingly, AC23 was deficient in effector candidates, but contained several predicted genes with redundant pathogenicity-related functions.

* Correspondence: Huyenphan.Phan@curtin.edu.au; Kar-Chun.Tan@curtin.edu.au; James.Hane@curtin.edu.au
[1]Centre for Crop & Disease Management, Curtin University, Perth, Australia
Full list of author information is available at the end of the article

Conclusions: We present an updated series of genomic resources for *P. nodorum* Sn15 – an important reference isolate and model necrotroph – with a comprehensive survey of its predicted pathogenicity content.

Background

The fungus *Parastagonospora nodorum* causes septoria nodorum blotch (SNB) of wheat (*Triticum aestivum*) and is a model species for necrotrophic plant pathogens. In order to provide insight on the evolutionary history and gene repertoire of this pathogen, a genome assembly of *Parastagonospora nodorum* model isolate Sn15 was first reported in 2007 [1]. This used Sanger shotgun sequencing of a genomic BAC library which produced a 37.5 Mbp draft genome reference with 108 scaffolds and 10,762 genes. It was the first species among the class Dothideomycetes for which a whole-genome reference was available [1] and has been used as a model species for cereal necrotrophs. This draft genome resource contributed to the discovery of three proteinaceous necrotrophic effectors (NEs) corresponding to known gene loci - ToxA [2], Tox1 [3] and Tox3 [4] - which are major host-specific virulence determinants in *P. nodorum*. The presence of additional NEs have been detected via their interaction with quantitative trait loci (QTL) corresponding to host sensitivity loci, but the genes encoding these effectors have not yet been identified [5–11] and others may have not yet been uncovered. In order to discover novel effectors in *P. nodorum* and in other fungal plant pathogens, it is important to ensure that the genome assembly and gene annotations are as accurate and reliable as possible.

Recent advances in long-read genome sequencing technologies, and established genetic and physical mapping techniques, have made whole-chromosome assembly of microbial genomes readily achievable [12–18]. Three decades ago, chromosome size and number estimates via pulsed-field gel electrophoresis (PFGE) of 11 *P. nodorum* isolates had estimated a range from 14 to 19, totalling 28 to 32 Mbp, ranging from 0.4 to 3.5 Mbp in length, with the smallest observed only in wheat and barley-infecting isolates [19]. The *P. nodorum* Sn15 genome assembly was progressively improved over subsequent years. It was updated in 2013 reducing the number of scaffolds from 108 to 91 [20], and again in 2016 with revised gene annotations that were supported by protein and transcriptome alignments and manual curation [21–23]. Leveraging these resources, comprehensive analyses of its genomic landscape and genome-based processes contributing to pathogenic adaptations have extended to transposable elements (TE) and gene repeats [1, 24, 25], repeat-induced point mutations (RIP) [24–26], mesosynteny [27], and multiple comparative genomics studies [14, 15, 23, 28, 29]. Initially, the Sn15 reference isolate was compared to a hyper-virulent isolate (Sn4) and a non-aggressive isolate (Sn79–1087) lacking known effector genes *ToxA*, *Tox3* and *Tox1* [20, 23]. These newly gained information and resources have played a vital role in studying important pathogenicity gene candidates. Subsequent comparison to an international panel (across 10 countries) of 22 *P. nodorum* and 10 *Parastagonospora avenae* isolates indicated presence-absence variation (PAV) - with notable absences in the 'avirulent' Sn79–1087 isolate assembly - of known effector loci and of large regions (i.e. scaffolds 44, 45, 46 and 51) [23], which was supplemented by a predictive analysis of accessory chromosome (AC) or region (AR) sequence properties (scaffolds 50 and 69) [29]. These large PAV regions were indicative of ACs/ARs that are associated with host-specific virulence in numerous fungal species [30], but this could not be confirmed with an unfinished genome assembly. In 2018, long-read-based genome assemblies were generated for 3 *P. nodorum* isolates (Sn4, Sn79–1087 and Sn2000) with 22 to 24 contigs [15]. Analysis of the Sn4 genome revealed that 'contig23' (~ 0.48 Mbp) was absent in Sn79–1087 and therefore considered an AC. This study also used transcriptome data from the Sn15 reference isolate to 'auto-annotate' genes in Sn4, and subsequently trained gene prediction software on the Sn4 annotations, which was used to perform in silico prediction of genes in the remaining isolates [15]. A follow up study in 2019 compared these four assemblies to NGS-based assemblies for a panel of 197 isolates from the United States, highlighted widespread diversifying selection within predicted effector loci and across the AC Sn4 contig23, reinforced the impact of the known ToxA, Tox1 and Tox3 effectors, and predicted 17 candidate effector loci with high levels of diversifying selection [31].

The recent updates to *P. nodorum* genomic resources enable consideration of the genomic landscape and sequence features which are relevant to pathogenicity or adaptation at the chromosome-scale, such as repeat-rich regions and mutation hotspots [26, 30, 32]. Long-read-based methods have significantly improved genome assembly of these previously challenging regions [12, 16, 18, 33, 34] and scaffold lengthscan be further improved with genome-finishing techniques including optical restriction [14, 35] or chromosome interaction mapping [36]. In fungal pathogens with "two-speed" genomes, repeat-rich regions typically accumulate mutations more rapidly than conserved gene-rich regions [26], leading to compartmentalisation of pathogen genomes into stable

GC-equilibrated regions and AT-rich 'mutation hotspots', which can include pathogenicity-associated ACs or ARs [30]. For pathogenicity loci not residing within ACs, growing evidence supports their frequent location in sub-telomeric mutation hotspots [37–39] which may also be ARs. The segregation bias of certain gene functions to the sub-telomere may be associated with the role of heterochromatin found at sub-telomere region in regulating gene expression during infection [39] and protection of the core genome from interspersion of sub-telomeric heterochromatin [40].

The presence or absence of effector genes, or ACs/ARs that contain them, can determine host/cultivar-specific virulence for several pathogen species [30]. Bioinformatic methods for effector prediction are usually of a reductive nature, which filter the complete gene set down to an candidate effector subset based upon multiple criteria [41]. These methods typically require effector gene annotations not to have been missed in the complete gene set (at either assembly or gene prediction steps), and directly benefit from the proper application of transcriptome data to gene annotation, which for gene-dense genomes like those of fungi can pose a technical challenge [42]. In this study, we present an updated chromosome-scale genome assembly for *P. nodorum* reference isolate Sn15, combining long-read data and optical mapping to arrive at a near complete telomere-to-telomere assembly of 23 chromosomes. Sn15 gene annotations have also been updated integrating new transcriptome data and extensive manual curation, which will ensure its reliability and ongoing utility as a model necrotroph. Insights from comparative genomics analysis is presented for comparisons of the Sn15 reference isolate versus the Sn4, Sn2000 and Sn79–1087 long-read assemblies, and an international panel of NGS-based assemblies for 28 other *Parastagonospora* isolates. This has highlighted mutation hotspots and locational biases across the 23 chromosomes of Sn15, including a 0.4 Mbp accessory chromosome and several telomeric ARs. New effector gene predictions for Sn15 are also provided, integrating the wealth of past data for Sn15 with new data including PAV and diversifying selection across the international pan-genome. These aggregated resources for *P. nodorum* Sn15 will offer novel research opportunities and serve as a useful tool to enhance ongoing efforts to breed for crop disease resistance.

Results

A chromosome-level reference genome assembly for *P. nodorum* Sn15

In order to complete the Sn15 assembly, a combination of long read sequencing using PacBio technology and optical mapping were used. PacBio DNA sequencing generated 368,822 raw reads of 50 bp to 41 Kbp in length at ~71X coverage. Self-correction resulted in 118,028 corrected reads, totalling 1.31 Gbp with an average length of 10 Kbp. Corrected reads were assembled into a draft assembly of 36 gapless contigs ranging from 3.5 Mbp to 37 Kbp, with a total length of 37.4 Mbp at 33.6X coverage. Only 844 corrected reads (0.7%) were not assembled. One of the 36 contigs corresponded to the previously published mitochondrial DNA sequence [GenBank: EU053989] and was discarded. An optical map produced 23 maps with an estimated total length of 39.26 Mbp (Supplementary Text 1). Thirty out of the 35 assembled contigs (36.88 Mbp) aligned to the 23 optical maps. The 5 contigs that did not align were short (38 to 120 Kbp, or 0.84% of the contig assembly) and highly repetitive, with no predicted genes. The curated scaffolds of contigs aligned to the 23 optical maps - subsequently referred to as 'chromosomes'- were numbered in descending size order based on the physical lengths predicted by the optical map (Supplementary Table 5). Fourteen chromosomes contained no gaps, and 8 gaps were added to join non-overlapping contigs within chromosomes 2, 3, 4, 6, 7, 10 and 20. Terminal 'TTAGG' tandem repeats indicating telomeres were observed at both ends of 21 chromosomes, with 2 having a single telomere. New repetitive regions comprised ~ 0.4% of the assembly. The new assembly had 286 fewer gaps than the previous version [20, 23] and there was a ~ 4 Kbp increase in the average length of AT-rich regions, a reduction of incompletely assembled AT-rich regions (− 46) and an increase in fully assembled AT-rich regions (+ 33) (Supplementary Table 3).

A revised set of gene annotations aggregated from multiple sources of evidence, including new *in planta* RNA-seq, fungal-specific gene finding software and manual curation

An estimate of the representation of the core gene content in the updated Sn15 assembly via BUSCO (v5.1.2) indicated 99.1% completeness versus the "fungi" dataset (fungi_odb10, 2020-09-10). The combination of various gene prediction methods (see methods), incorporating recently published in vitro and *in planta* RNA-seq data [23, 43], fungal-specific gene prediction software, and manual curation, resulted in 16,431 predicted genes. This gene set was split into two subsets: a higher confidence set (Set A), and a lower confidence set to allow more sensitivity for subsequent pathogenicity gene predictions (Set B) (Fig. 1, Table 1A). Set A included 13,893 high confidence genes models with higher levels of support, whereas set B contained 2538 putative genes with either shorter coding sequence length or less RNA-seq support (Table 1B). Compared to the previously published annotation [20, 23], average gene length decreased by 70 bp and gene density increased by 2.8 genes per

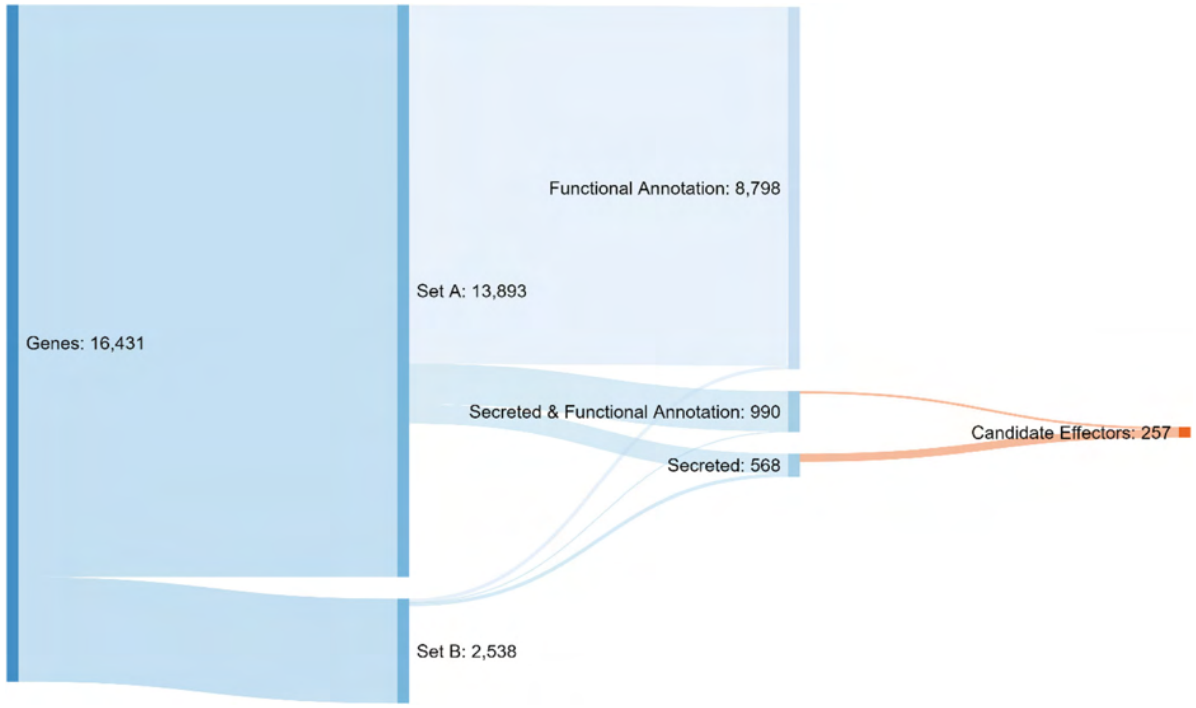


Fig. 1 Summary of predicted genes of *Parastagonospora nodorum* Sn15. Predicted genes in Set A and B were separated based on predicted secretion signal peptides or informative functional annotations, from which effector candidates were predicted

Mbp (Table 1A). Set B annotations were on average length 4 times shorter than those of Set A and in 86% of cases were a single exon (Table 1A). Of the 16,341 genes, 9788 were informatively functionally annotated (i.e. a conserved domain), and 990 of these also had a predicted secretion signal peptide (Fig. 1). The predicted secretome comprised 1568 genes of which 257 (1.5% of total genes and 25.3% of the secretome) were effector candidates (Fig. 1, Table 1). Across the Sn15 genome, gene density was inversely correlated with density of repetitive DNA (Fig. 2), with genes distributed at a relatively even density (~ 450 Mbp) except for accessory chromosome 23 (AC23 which was gene sparse (~ 380 Mbp) (Fig. 2). A lower proportion (36.7%) of loci were assigned functional annotations within AC23, which was 13% less than average. The known necrotrophic effector genes *ToxA*, *Tox1* and *Tox3* were all located within subtelomeric regions of chromosome 4, 10 and 11 respectively, with *ToxA* also notably residing in the middle of a large (~ 570 Kbp) repeat-rich region (Fig. 2).

Comparative genomics

In comparisons of the Sn15 genome to alternate isolates, the Sn15 genome exhibited multiple large PAV regions (Fig. 2, Supplementary Table 6, Supplementary Table 7). Prior pan-genome and in silico studies using the previously published Sn15 assembly as its reference genome had indicated scaffolds 44, 45, 50 and 51 as regions of the genome with PAVs [20, 23, 29] (Supplementary Table 8).

Scaffold 50 corresponded to a sub-telomeric region of chromosome 8 (Supplementary Table 8). New reports of additional variable regions derived from this study include regions of chromosomes 7, 8 and 10. Chromosome 7 contained a ~ 455 Kbp region that is potentially duplicated in some isolates, but is represented in single copy in the current Sn15 assembly. Chromosomes 8 and 10 contained ~ 88 Kbp and ~ 10 Kbp PAV regions respectively. The PAV on chromosome 10 contained no genes, and the PAV on chromosome 8 contained 27 genes (Supplementary Table 9) but did not contain any predicted effector candidates (Supplementary Table 4).

Former scaffolds 44 and 45 corresponded to the ~ 444 Kbp AC23 of this study (Supplementary Table 8). Pan-genome alignment of Sn15 chromosomes with other *P. nodorum* and *P. avenae* isolates indicated that chromosome 23 was absent in *P. avenae* and the non-aggresive *P. nodorum* isolate Sn79–1087 (Fig. 2), which suggested that it lacked genes required for viability and was an accessory chromosome. In contrast, the majority of other "core" chromosomes were well conserved across *Parastagonospora* spp. AC23 also exhibited higher overall levels of non-synonymous mutations indicating diversification across this population relative to the Sn15

Table 1 Summary of new gene annotations of *P. nodorum* reference isolate Sn15. A) Comparison of high-confidence Set A and low confidence Set B to previous annotation versions and B) summary of data supporting gene annotations

A) Summary	Previous studies [20, 23]	This study (Set A)			This study (Set B)		
Number of genes	**13,569**	**13,869**			**2534**		
Number of mRNAs	**13,944**	**14,160**			**2557**		
Average gene length, bp	**1558**	**1488**			**373**		
Number of exons	**36,557**	**36,447**			**3070**		
Average exons/gene	**2.6**	**2.6**			**1.2**		
Average exon length (bp)	**556**	**539**			**299**		
Gene density (genes / Mbp)	**369**	**372**			**440 (Set A + Set B)**		
B) Supporting Data	**Loci**	**FPKM > 50**	**FPKM > 5**	**FPKM < 5**	**SignalP**	**SignalP+ EffectorP**	**Region not in previous assembly**	**Fuctional annotation**
Set A	13,869	4033	9968	4030	1488	340	19	9505
In previous studies [20, 23]	13,663	3742	9633	4030	1463	323		9456
Not in previous studies [20, 23]	206	291	335	0	25	17		49
Set B	2534	220	2508	26	97	61	23	75
In previous study [1]	117	5	117	0	19	6		0
New or modified genes	2417	215	2391	26	77	55		75

reference isolate (Supplementary Table 10). However the mutation profile of AC23 contained two regions separated by a large repeat island – each side corresponding to scaffolds 44 and 45 of the previously Sn15 assembly [20, 23] - which exhibited distinctly different mutation rates (Fig. 2). Comparison of AC23 to other Sn15 chromosomes did not indicate that it had originated from duplication of core chromosomes (Supplementary Figure 1), however homologous (non-repetitive) regions in *Pyrenophora tritici repentis* [14, 44] and *Bipolaris* spp. [28, 29] genomes tended to be located in sub-telomeric regions (Supplementary Figure 2).

The previous scaffold 51 corresponded to a ~ 74 kbp region within a repeat-rich sub-telomeric region of chromosome 4 (Supplementary Table 8), which also contained the effector gene *ToxA*. This sub-telomeric region had below average GC content, correspondingly high repeat content (~ 18.3% higher than the genome average), increased mutation density and less than half of the average gene density (Supplementary Table 6). The 9 predicted loci within this region had an average DN/DS of 1.9, more than double the genome average (Supplementary Table 6). Alignment of this region between the Sn15 assembly presented in this study, and the long read assemblies of Sn4, Sn2000 and Sn79–1087, showed structural variations that may indicate that breakage-fusion bridge (BFB)-mediated rearrangements (distal translocations between chromosomes lacking telomere caps) may have occurred in one or more of these isolates (Supplementary Figure 3) [45]. Comparisons of this region to corresponding regions containing ToxA homologs in related species *Pyrenophora tritici-repentis* [14, 44], *Bipolaris maydis* [29] and *B. sorokiniana* [28], indicated further chromosome structure diversity. The sub-telomeric *ToxA* region of chromosome 4 in *P. nodorum* appeared to be consistent with *Bipolaris* spp. where it was also found in sub-telomeric locations. In contrast, the *P. tritici-repentis ToxA* -containing chromosome appeared to be a product of the breakage of the *P. nodorum ToxA* region, followed by chromosome fusions resulting this region being flanked by sequences corresponding to *P. nodorum* chromosomes 14 and 19 (Fig. 3).

Discussion

The chromosome-level assembly for *P. nodorum* reference isolate Sn15 improved detection of pathogenicity gene-rich regions

The new chromosome-level genome assembly of *P. nodorum* Sn15 created by this study has established the correct number of chromosomes for this pathogen, which was previously underestimated by PFGE to range from 14 to 19 [19], and is consistent with 22–23 observed in assemblies of other isolates [15, 31]. PFGE fragment resolution accuracy requires at least ~ 1% difference in chromosome size [46], meaning 6 out of the 23 assembled sequences were within a potentially unresolvable size range (Supplementary Table 5). The

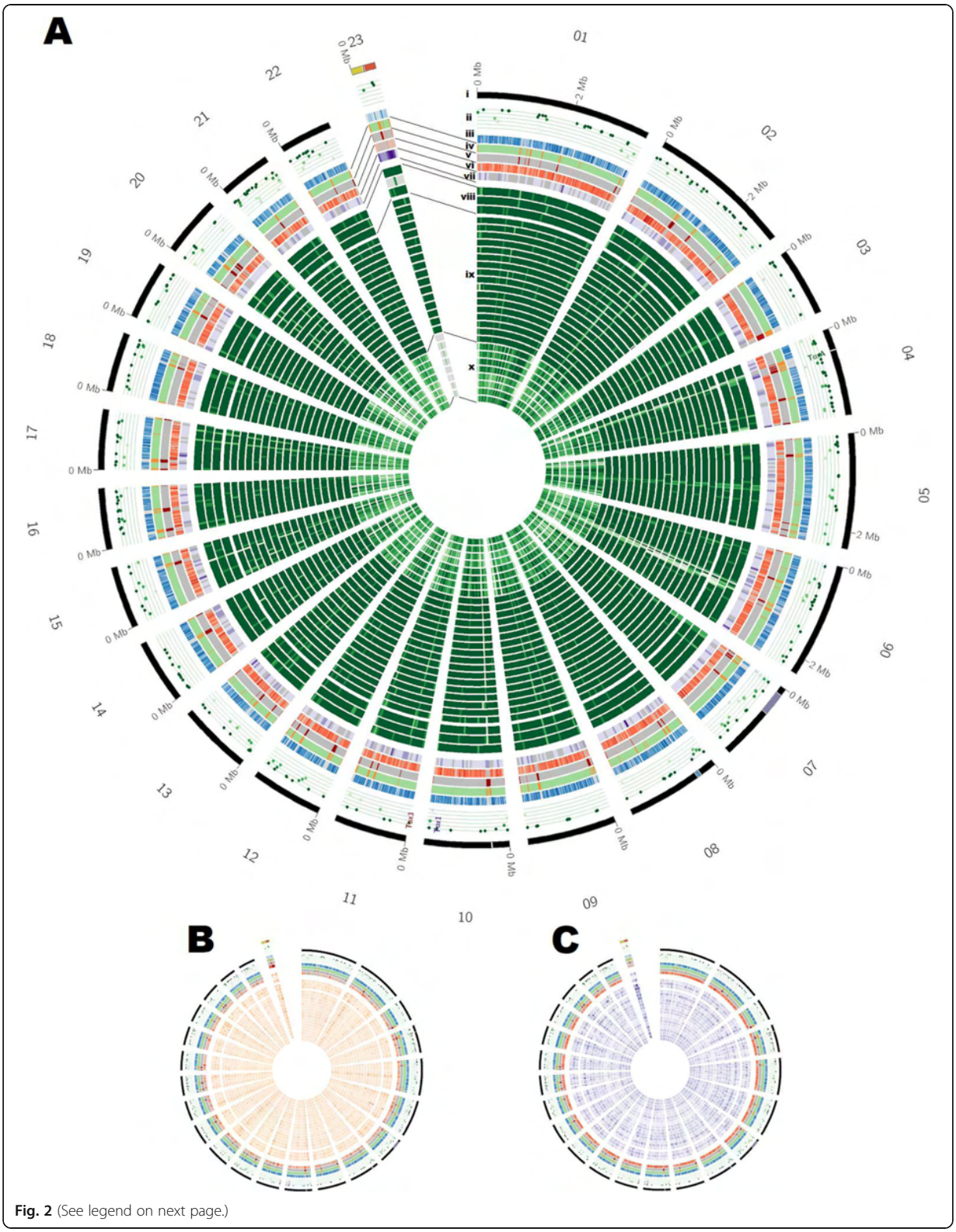


Fig. 2 (See legend on next page.)

(See figure on previous page.)
Fig. 2 Sequence comparisons of the new genome assembly of the *Parastagonospora nodorum* Sn15 reference isolate with alternate *P. nodorum* isolates and *P. avenae* isolates, within 50 Kbp windows, for: **a** Presence-absence variation (PAV) indicated by percent coverage of MUMmer matches (green), **b** SNP density (red), and **c** the ratio of non-synonymous to synonymous SNP mutations (DN/DS) relative to Sn15 (purple). Rings indicate (in inwards order): i) Sn15 chromosome (black); ii) loci predicted by EffectorP and score from 0 to 1 (dark green); iii) gene presence (blue); iv) AT-rich regions (orange); v) repeat regions (red); vi) average SNP mutation density from (**b**) (orange); vii) average DN/DS from (**c**) (purple); viii) PAV versus alternate isolates Sn4, Sn79–1087 and Sn2000; ix) *P. nodorum* isolate draft assemblies; x) *P. avenae* isolate draft assemblies. *P. nodorum* Sn15 accessory chromosome 23 (AC23) has been highlighted with regions corresponding to scaffolds 44 (yellow) and 45 (red), previously reported in Syme et al. 2018 to be conditionally-dispensable and under positive selection

difference in chromosome number between the two studies therefore is justified. This study also presents 21 out of 23 chromosomes with both telomeric ends and 14 gapless chromosomes. In addition, the new chromosome-level genome assembly for Sn15 was also supported by transcriptome data and manually curated gene annotations, and related bioinformatic resources for the *P. nodorum* Sn15 reference isolate were updated, enhancing these important resource for studying molecular host-pathogen interactions and for effector discovery [41, 47, 48].

Chromosome-level analysis of the genomic landscape can enable detection of compartmentalised mutation 'hot spots' that may contain pathogenicity loci - a commonly reported feature for "two-speed" genomes which have been broadly affected by transposon activity and

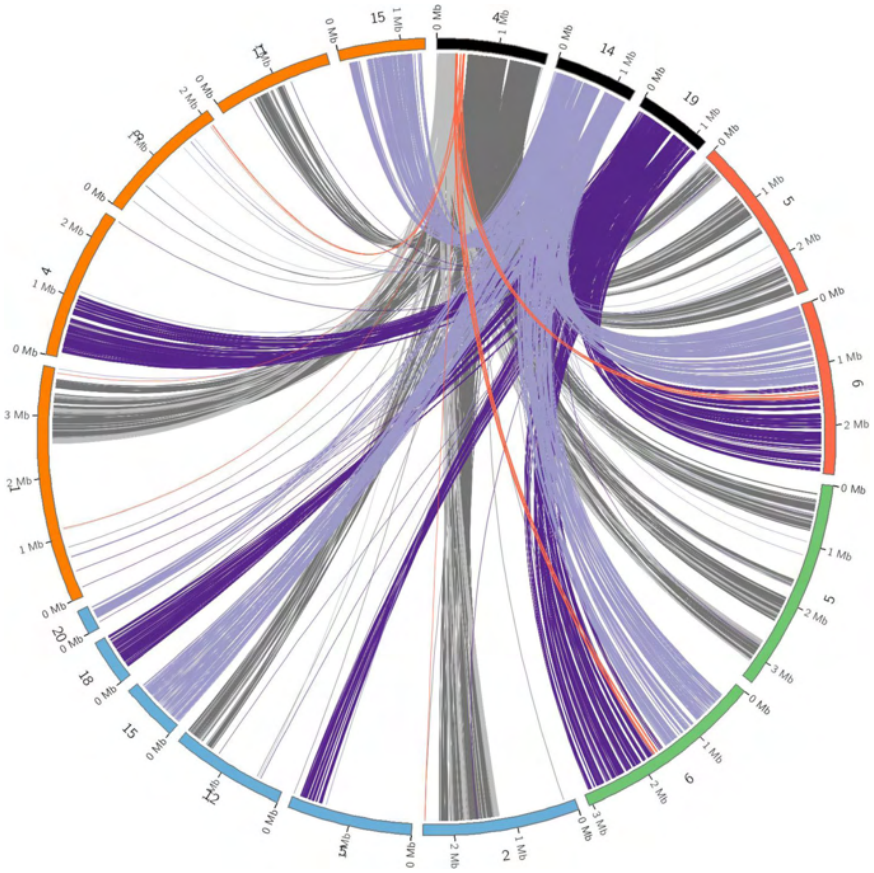


Fig. 3 Sequence similarity comparisons between *ToxA*-containing and related sequences of (A, black) *P. nodorum* Sn15 (chromosomes 4, 14 and 19); (B, red) *Pyrenophora tritici-repentis* BFP (chromosomes 5 and 6); (C, green) *Pyrenophora tritici-repentis* M4 (chromosomes 5 and 6); *Bipolaris maydis* (blue) (scaffolds 2, 5, 12, 15, 18 and 20); and *Bipolaris sorokiniana* CS10 (orange) (chromosomes 1, 4, 8, 1 and 15). Matches with *P. nodorum* Sn15 chromosome 4 are coloured grey, with the *ToxA*-containing region highlighted in red, and matches with *P. nodorum* Sn15 chromosomes 14 and 19 and coloured light and dark purple respectively

repeat-induced point mutation (RIP) [24, 26]. As genome assemblies have been improved towards chromosome-scale representation, there have also been several reports of pathogenicity genes within sub-telomeric locations in other pathosystems [15, 17, 37, 38, 40]. Thus, the new Sn15 assembly presented new opportunities to predict novel pathogenicity-related genes within the 'two-speed' regions that are repeat-rich or conditionally-dispensable [26, 30, 32]. Presumably, the apparent bias of pathogenicity gene locations within mini-chromosome or sub-telomeric regions could be associated with BFB formation [45] and mesosyntenic rearrangements [27, 32] between chromosome termini. Indeed, pan-genome comparisons indicated that the 0.4 Mbp AC23 was an accessory chromosome, with gene content relevant to pathogenicity (see below). The updated Sn15 assembly also highlighted additional regions not present in previous assembly versions, which comprised ~ 152 Kbp of repetitive DNA (0.4% of the genome) and 86,468 bp of non-repetitive DNA. While these represented a very small proportion of the genome, they may have special significance for plant pathology as they are more likely to contain effector or other pathogenicity genes. The relative placement of these regions in the genomic landscape was also important in assessing their likely roles in pathogenicity adaptation [26, 30, 32] as a parameter for effector prediction [41].

Candidate effector genes were derived from extensive gene annotation data for *P. nodorum* Sn15

Considerable efforts have been made across previous studies to ensure the reliability and ongoing applicability to plant pathology research of the annotated gene set for *P. nodorum* Sn15 [1, 20–23, 42], particularly for the purpose of effector and pathogenicity gene discovery. The revised Sn15 gene set includes a primary set of 13,893 genes (Set A) and a lower confidence set of 2538 (Set B) which was retained to enhance the sensitivity and capacity of effector gene predictions. The total number of predicted genes has increased since the previous annotation version [23], and is also higher than the currently reported average across the Ascomycota [49]. However we note general trends across all species, that while reducing assembly fragmentation can reduce the total number of predicted genes [50], the addition of significantly improved transcriptome data [51] or gene prediction methods [42] can increase this number. Functional annotations were assigned to 59.5% of predicted genes (Set A + B, excluding non-specific features e.g. coiled-coils, intrinsic disorder).

Across the whole genome, 257 effector candidate genes were predicted (Fig. 1, Supplementary Table 4), a number comparable to similar fungal pathogen genome surveys [41]. Effector candidate genes exhibited the typical features expected of effectors, including: secretion, low molecular weight, cysteine richness, diversifying selection, association with mutation hotspots, and where functional annotations were assigned these had a common pathogenicity-related theme (Supplementary Table 4). Secretion was predicted for 1558 genes (9.5% of Set A + B), of which 257 (16.5% of predicted secretome) were effector candidates and 12 were predicted to localise to the chloroplast (including the confirmed effector ToxA) (Supplementary Table 4). Effector candidate loci were typically found within either 5–10 or 20–25 Kbp of AT-rich regions, which was not the case across the whole gene set (Supplementary Table 4, Supplementary Table 11). This is consistent with reports of RIP and effector location bias within AT-rich mutation hotspots [26]. The *ToxA* locus was 4039 bp and *Tox3* was 1860 bp from their nearest respective AT-rich regions. The *Tox1* locus was located > 200 Kbp from its nearest AT-rich region, however all 3 effector loci were also located within sub-telomeric regions (Fig. 2). This association between telomeres and effector-rich mutation hotspots is also reported in other pathogen species [30, 32]. Comparison of orthologs between the Australian reference isolate Sn15 and the US isolate Sn4 [15], indicated 14 out of the 17 previously published Sn4 candidate effectors were also predicted among the Sn15 candidates (Supplementary Table 12). These Sn4 candidates – which included *Tox1* - were previously reported to exhibit diversifying selection that was specific to one of the 2 major US sub-populations [31].

Functionally-redundant genes may be associated with potential pathogenic properties of accessory chromosome 23

Surprisingly, AC23 which exhibited typical characteristics of ACs [15, 30, 31, 52] - and may correspond to anecdotal reports of a ~ 0.4 Mbp AC specific to *P. nodorum* wheat and barley-infecting isolates [53] - had a relatively low density of effector candidate loci (Supplementary Table 4, Supplementary Table 10). Six candidate effector loci were predicted on AC23 in a previous study [23], with two of these (SNOG_16226 and SNOG_16236) re-predicted (with ranked scores of 10 and 9 respectively) in the more stringent predictions of this study. AC23 also encoded multiple genes with other pathogenicity-related and/or redundant functions, which may indicate tandem duplications or multiple BFB events. These functions included: Ulp1 protease (SNOG_16274, SNOG_16214), RING/FYVE/PHD-type zinc finger proteins (SNOG_16310, SNOG_16333), valyl-trna synthase (SNOG_16268, SNOG_16213, SNOG_16211), and UstYa-like protein (mycotoxin biosynthesis) (SNOG_16357) (Supplementary Table 13). Ulp1 protease is involved in the modification of SMT3, a

ubiquitin-like protein of the SUMO family which suppresses MIF2 mutations. MIF2 is a centromere protein that regulates stability of di-centromeric mini-chromosomes in baker's yeast [54]. Its presence on AC23 is notable given that AC23 is a mini-chromosome and is therefore more likely to be unstable. UstYa-like proteins are involved in the secondary metabolite synthesis of cyclic peptide mycotoxins including ustiloxin and chyclochlorotine [55], however the products of many remain unknown. FYVE domain zinc finger proteins reportedly may bind to phospholipid PI3P [56], which could potentially facilitate host cell uptake.. The genes and functions listed above represent candidate pathogenicity loci residing on AC23 which are of high importance for further investigation.

A trove of effector and pathogenicity gene homologs were predicted among candidate effector-loci

We previously observed that the deletion of *ToxA*, *1* and *3* in *P. nodorum* SN15 resulted in a mutant that retained near-WT level of virulence on most commercially adopted wheat varieties [57, 58]. This suggested that SN15 that lacked *ToxA*, *Tox1* and *Tox3* may have produced undiscovered effectors or other virulence factors to functionally compensated for the loss of these major NE genes [59, 60]. In addition, biochemical and genetic characterisation of US *P. nodorum* isolates identified evidence of other NEs [5–11]. This prompted us to apply a bioinformatic approach to predict for NE candidates in the near-complete SN15 genome that are relevant to the Australian cereal industry. From the prediction analysis, *ToxA*, *Tox1* and *Tox3* ranked highly among the top 98 Sn15 candidates with ranked scores of 5 and above (Supplementary Table 10). While remaining candidates are unconfirmed, among these we observed a wealth of assigned functions or matches strongly suggesting roles in pathogenicity. *SNOG_13622* and *SNOG 08876* encode for CFEM domain proteins, which have roles in iron acquisition and several of which have been reported with roles in virulence [61]. *SNOG_42372* and *SNOG_07772* encode for chitin-binding LysM domain proteins which offer protection from PTI in the host [62]. *SNOG_07596* encodes a thaumatin-like protein, which when produced by host plants are pathogenesis-related (PR) proteins involved in defence, however fungal homologs have also been reported with roles in virulence [63]. *SNOG_03746* encodes a knottin-like protein. Knottins are cytotoxins that are best represented by snake and arachnid venoms, with the first fungal report of a knottin in the poplar rust *Melampsora larici-populina* [64]. *SNOG_30910* encodes a homolog to phospholipase A2 - which cleaves sn-2 acyl bond between 2 phospholipids that releases arachidonic acid and lysophosphatidic acid - and is also a common domain in spider, insect and snake venoms that disrupt cell membranes [65]. *SNOG_00200* encodes a product similar to *Alternaria alternata* allergen 1 (AA1-like). The AA1-like family [66] contains the *V. dahliae* effector PevD1, which binds the host thaumatin PR5 [67]. *SNOG_00182*, *SNOG_02182*, and *SNOG_16063* encode ribotoxins, which have a conserved sarcin/ricin loop (SRL) structure that cleaves specific sequences in the host rRNA, leading to ribosome inactivation and cell death by apoptosis [68]. *SNOG_13722* encodes a cerato-platanin, which induces phytoalexin synthesis and causes necrosis [69]. *SNOG_06012* encodes a protein similar to gamma crystallin/yeast killer toxin, which is a pore-forming cytotoxin [70]. *SNOG_01218* encodes a subtilisin, a serine protease family that is frequently reported in fungi to promote virulence [71]. *SNOG_03959* encodes a protein similar to a cyclophillin-like/peptidyl-prolyl cis-trans isolmerase (PPIase) [72], which in humans is well known for interfering with the immunosuppressive drug cyclosporin A, but is widespread across eukaryotes and has been reported as virulence determinants in several fungi including: *Leptosphaeria* spp., *Botrytis cinerea*, *Cryphonectria parasitica*, *Puccinia triticina*, *M. oryzae*, and *Lhellinus sulphurascens*, as well as various oomycete species of the *Phytophthora* genus [73]. *SNOG_08289* encodes a pectin/pectate lyase, which are reported in many fungi to promote virulence [74]. *SNOG_11034* encodes a protein similar to Egh16, an appressorially-located virulence factor of *Blumeria graminis* f. sp. *hordei* with broadly conserved homologs across several pathogenic fungal species [75]. *SNOG_15608* encodes a cutinase, which may be involved in host surface penetration [76]. *SNOG_02399, SNOG_03334, SNOG_40970, SNOG_08150, SNOG_04779* all encode for proteins with lipid interacting domains. *SNOG_11842* encodes a Hce2 effector homolog - which is named after Homologs of *C. fulvum* ECP2, a necrosis inducing effector. There are 3 defined classes of proteins with Hce2 domains, of which SNOG_11842 belongs to class I, the smallest and most common class [77]. Many of the above candidates with pathogenicity-related functional annotations are also expressed higher *in planta* (IP) relative to in vitro (IV) by a factor of 5, however the *Tox3* IP:IV is only 2 indicating this lower values may also be relevant in host-pathogen interactions. A lower-ranked candidate (*SNOG_06459*) with a ranked score below 5 is also mentioned here as it encoded a cerato-ulmin homolog. Cerato-ulmin is a hydrophobin, which is not a functional class normally reported to be directly involved in pathogenicity, but has been reported as a potential virulence factor in dutch elm disease [78]. Its mode of action is not like a typical effector however, as its role appears to be to protect spores from desiccation, which leads to increased spore survivability and transmission.

Multiple effector candidate loci were predicted to be laterally-transferred with other cereal-pathogenic fungal species

Of the 98 highly-ranked effector candidates, 11 showed a conservation pattern indicating potential lateral transfer when compared to a panel of whole gene sets of > 150 fungal species (Supplementary Table 4) [79]. This included *SNOG_16571* (*ToxA*), *SNOG_20078* (*Tox1*), *SNOG_13622* (CFEM domain), *SNOG_15952* (ribotoxin-like), *SNOG_00152, SNOG_01658, SNOG_20100, SNOG_08426, SNOG_07039, SNOG_00726,* and *SNOG_14618*. These had rare orthology relationships indicating potential lateral gene transfer (LGT) with *Pyrenophora* spp., *Setosphaeria turcica*, *Alternaria brassicicola*, *Verticillium dahliae*, *Leptospaheria maculans* and *Coccidioides immitus*. Aside from *SNOG_13622*, *SNOG_15952*, and known effectors *ToxA* and *Tox1*, this group of effector candidates had no predicted functional annotations. As expected, *SNOG_08981* (*Tox3*) was not included in this set and has so far been reported to have no known homologs.

Conclusions

The *P. nodorum* isolate Sn15 was the first representative of the class Dothideomycetes with a genomic survey report [1], and has since become an important reference and model necrotroph with a significant set of accumulated genomic, transcriptomic, proteomic and bioinformatic resources supporting its genome and gene data [15, 20–25, 31, 43]. This study updates these resources in the context of a chromosome-scale assembly, identifying genome features relevant to pathogenicity i.e. sub-telomeric regions, accessory chromosomes and mutations hotspots. This has provided genomic context to subsequent predictions of candidate genes encoding effectors and other pathogenicity factors. In contrast to the earliest Sn15 genome study, effector candidates were supported by a wealth of functional annotation and comparative genomics data indicating strong homology to known effectors and other pathogenicity genes. This study is an important step forward for the further characterisation of *P. nodorum* chromosome structure and its role in pathogenicity, particularly in highly mutable and potentially effector-rich regions of the genome including AC23. Additionally, the increased representation of repeat-rich regions and provision of curated gene annotations within them, is of high value to ongoing efforts to characterise and understand fungal effectors. We anticipate future studies will utilise the effector predictions provided in this study to confirm new novel *P. nodorum* effectors, and potentially discover a role for AC23 in promoting virulence.

Methods

Genome sequencing and assembly

Genomic DNA of *P.nodorum (syn. Phaeosphaeria nodorum, Stagonospora nodorum, Leptosphaeria.*

nodorum,. Septoria nodorum) strain Sn15 [20] – originally isolated in Western Australia by the Dept. Primary Industries and Regional Development (DPIRD: Agriculture & Food) - was sequenced via Pacific Biosciences P5-C3 chemistry with 4 SMRT cells, at the Génome Québec Innovation Centre (McGill University, Montreal, QC, Canada). The longest 25% of reads were self-corrected and assembled using Canu v1.0 (–pacbio-raw, expected genome size 39 Mbp) [13]. Assembly base-calls were corrected with Pilon v1.16 [80] using Illumina reads [20] which were mapped to the assembly with Bowtie2 v2.3.3.1 [81]. Mitochondrial contigs assembled by the above methods were identical to a previously published Sn15 mtDNA [GenBank: EU053989] [1] therefore the old mtDNA record was not updated by this new assembly..

Optical maps were used to order and orient the Canu-assembled Sn15 contigs into a complete genome. Sn15 protoplasts were extracted from hyphae as per Solomon et al [82], which was adjusted to 1x10e8 with GMB (0.125 M EDTA pH 8, 0.9 M sorbitol) at 42 °C. Protoplasts were added 1:1 to 1% low melt agarose (SeaPlaque GTG in 2% sorbitol and 50 mM EDTA) and poured into Plug Mold (Bio-Rad Laboratories, Munich, Germany) and set at 4 °C for 30 min. The plug was added to 5 ml Proteinase K solution (1 mg/ml Proteinase K, 100 mM EDTA pH 8.0, 0.2% Na deoxycholate, 10 mM Tris pH 8.0 and 1% N-lauroyl sarcosine) and incubated at 50 °C overnight, then added to sterile wash buffer (20 mM Tris pH 8, 50 mM EDTA pH 8) for 4 h changing the solution every hour. Clean plugs were transferred into 0.5 M EDTA at 9.5 pH and stored at 4 °C until shipment at room temperature. High molecular weight DNA was extracted from protoplasts as per Syme et al [23] and digested with *Spe*I, resulting in 63,440 fragments with an average size of ~ 315 Kbp. Optical maps were generated and manually curated with MapSolver™ (OpGen, MD, USA). Contig joins were made by inserting a 100 bp unknown (N) gap. Where the optical map indicated contig mis-assemblies, potential breakpoints were inspected for a localised drop in aligned read coverage. Chromosome scaffolds were numbered in descending size order based on the estimated physical lengths derived from the optical map. Chromosome scaffolds were assessed by Quast v5.2 [83], BUSCO 5.1.2 [84] and by coverage depth of alignments of raw and corrected SMRT reads by bwa-mem (0.7.17-r1188, –x pacbio) [85] via SAMtools [86] and BEDtools [87]. The assembled Sn15 chromosome scaffolds of this study were compared to previously published assembly versions (Supplementary

Table 1) [15, 23] with MUMmer 3.0 (nucmer -maxmatch, show-cords) [88] and alignments were visualised with Dot [89]. The SN15 reference genome data is available under BioProject: PRJNA686477. The updated SN15 genome assembly is deposited under [Genome: GCA_016801405.1/ASM1680140v1] and [NUC: CP069023.1 - CP069045.1].

Annotation of genome features

Repetitive DNA regions within the Sn15 genome assembly were analysed by three methods. The presence and overall proportion of AT-rich regions were calculated with OcculterCut v1.1 [26]. Annotation of repeat regions was performed using RepeatMasker 4.0.6 (sensitive mode, rmblastn version 2.2.27+) [90] in four separate analyses, using: a) a published set of de novo repeats derived from a previous Sn15 genome assembly [24], b) RepBase (taxon "Fungi") [91], c) LTRharvest of the GenomeTools suite [92], and d) a newly predicted set of de novo repeats generated by RepeatModeler v1.0.8 [93] (–engine ncbi -pa 15). Subsequent repeat analyses requiring a repeat-masked input used the output derived from the new de novo repeat dataset (Supplementary Table 2, Supplementary Table 3). Tandem repeats were predicted using Tandem Repeat Finder (Parameters: Match = 2, Mismatch = 7, Delta = 7, PM = 80, PI = 10, Minscore = 180, MaxPeriod = 2000) [94]. Telomere regions were identified by terminal "TTAGGG" tandem repeats [95].

Protein-coding gene loci were annotated incorporating multiple transcriptome datasets from previous studies [23, 43]. RNA-seq reads from a prior study [43] were trimmed with cutadapt v.9.1 (paired end mode, –-quality-cutoff = 30, –-minimum-length 25, –n 3) [96] and de-duplicated with khmer v2.0 and screed v0.9 (normalize-by-median.py, –C 30 -M 100e9) [97, 98]. Fungal RNA-seq reads were derived from a mixed *in planta* library during early infection (3 dpi) when known effectors are maximally expressed [60]. Fungal reads were separated from wheat sequences with BBsplit v36.11 (BBmap, Seal v36.11, –Xmx200g) [99] to screen against the *Ttriticum aestivum* assembly TGAC v1.30 (GCA_900067645.1) [100]. Filtered reads were mapped to the new Sn15 assembly with STAR v2.5.2b (alignReads, –-outSAMstrandField intronMotif --outFilterIntronMotifs) [101]. Transcript assembly was performed with Trinity v2.2.0 (–-seqType fa --trinity_complete --full_cleanup --jaccard_clip) [102]. RNAseq reads were aligned to the SN15 assembly with TopHat v2.2.6.0 (defaults) [103]. Relative expression was quantified with Cufflinks [104] and Stringtie v1.3.3b (params -m 50 -B -e -p 8) [105].

A final set of gene annotations was generated by combining annotations from multiple sources. Initial transcriptome-based predictions were made with PASA v2.0.2 [106], incorporating: Trinity transcripts; open-reading frames generated with Transdecoder v2.0.1 [102]; CodingQuarry (CQ) v2.0 predictions based on TopHat outputs [42]; a second round of predictions generated using Coding-Quarry "Pathogen Mode" (CQPM) (*A. testa*, 2016) within regions between the initial CQ predictions. De novo prediction was performed with GeneMark-ES v4.32 (–-ES, –-fungus) [107]. Previously published gene annotations for Sn15 were aligned to the new assembly with AAT r03052011 [108] using CDS features (–-dds '-f 100 -i 20 -o 75 -p 70 -a 2000' --filter '-c 10' --gap2 '-× 1') and protein sequences (–-dps '-f 100 -i 30 -a 200' --filter '-c 10' --nap '-× 10'). All predicted gene annotations described above were assigned relative weight scores (AAT protein mapping 1, EST 5, AAT CDS mapping, GeneMark-ES 1, CQ/CQPM 10, transdecoder 10, PASA 9) and were then integrated into a single annotation set via EvidenceModeler (EVM) v1.1.1 [106]. Every locus of the EVM annotation set was manually curated using Webapollo [109] alongside supporting evidence from the various prediction methods described above, InterProScan domains aligned to the genome [110], aligned RNAseq reads [23, 43] and annotated repeat features (see above). New loci were manually annotated within intergenic regions if supported by RNAseq alignments. The resulting set of manually curated annotations were filtered (Table 2) for either: 1) orthologous best hit to the 13,690 gene models from the previously published annotation [20, 23] or; 2) coding regions (CDS) of > 300 bp in length and with RNAseq read coverage of > 50 fragments per kilobase of transcripts per million mapped reads (FPKM). This filtered set of manually curated genes is subsequently referred to as the primary gene set (Set A). The remaining predictions that failed this filter were retained as a secondary gene set (Set B) if CDS length was > 90 bp and RNAseq depth was > 5 FPKM or if homologous to a previously annotated gene [1, 20, 23]. To be consistent with previous publications on *P. nodorum* genomics, in this study gene annotations corresponding to loci that had been numbered in previous studies [1, 20, 23] retained their previous locus number despite non-sequential order along the new assembly. New annotations not corresponding to previously annotated loci were numbered from SNOG_40000 onwards.

Various software and databases were used to assign functional annotations for both gene sets (A and B). OcculterCut v1.1 [26] predicted AT-rich regions and distances to the nearest AT-rich region for each locus. InterProScan (5.27–66.0) [110] was used to generate a broad range of functional annotations (InterPro, Pfam, Gene3D, Superfamily, MobiDB). PHIbase v4.2 [48] was searched to assign homology to known effectors. SignalP

Table 2 Criteria used to predict the primary gene prediction set (Set A) for P. nodorum Sn15, and the secondary set (Set B) which contains low-confidence gene predictions for the purpose of extracting a small proportion of strong effector candidates

Criteria	Set A (high confidence)	Set B (putative)
CDS length	> 300 bp / > 100 aa	(> 90 bp / > 30 aa
Gene expression	*AND*	*AND*
Orthology/Homology	> 50 FPKM	> 5 FPKM)
	OR	*OR*
	Orthology (reciprocal best hit) to previous annotation [20, 23]	Homology to previous annotation set [20, 23]

v4 [111]. was used to predict extracellular secretion, EffectorP v2.0 [47] was used to predict effector functions, and Localizer v1.0 (–e mode) [112] was used to predict potential host-cell sub-cellular localisation. The dbCAN r07/20/2017 [113] and AntiFam r3.0 databases were both searched using hmmsearch (–-cut_ga) [114] to predict carbohydrate-active enzymes (CAZymes) and pseudogenes respectively.

Comparative pan-genomics

The new Sn15 genome reference was compared with draft assemblies of 18 *P. nodorum* isolates [20, 23] and 10 *P. avenae* isolates [BioProject: PRJNA476481], as well as long-read assemblies of 3 isolates of *P. nodorum*: Sn79–1087 [BioProject: PRJNA398070; Genome: GCA_002267025.1], Sn4 [BioProject: PRJNA398070; Genome: GCA_002267045.1] and Sn2000 [BioProject: PRJNA398070; Genome: GCA_002267005.1] [15, 31] (Supplementary Table 1). Whole-genome alignments and variant calling was performed via MUMmer v3 (nucmer --maxmatch, show-coords -T -H -r) [88] and the percent coverage of matches of each isolate relative to Sn15 was calculated within 50 Kbp windows via BEDtools v2.26.0 (makewindows, coverage) [87]. PAVs were calculated from nucmer (delta-filter – 1) alignments via BEDtools v2.26.0 (genomecov -bga) [87]. SNPs were calculated from nucmer alignments (show-snps -rlHTC) [88] and SNP density was calculated in 10 Kbp windows via BEDtools v2.26.0 (genomecov -bga) [87]. SNPs were analysed with SnpEff [115] relative to the new Sn15 gene annotations and the non-synonymous/synonymous mutation (Dn/Ds) ratio was calculated for every Sn15 gene both: 1) versus individual isolates and 2) averaged over all isolate comparisons. For visualisation by CIRCOS v 0.69–3 [116], Dn/Ds ratios were averaged across all genes within 50 Kbp windows via BEDtools v2.26.0 (map -c 4 -o mean) [87].

Prediction of necrotrophic effector (NE) candidate gene loci

Putative effector genes were predicted based on the ranking of cumulative scores assigned from effector-associated gene or protein properties, as has been reported in previous studies [20, 23]. In this study the features (Supplementary Table 4) used to assign scores were: predicted secretion by SignalP 4.1 [111] (1 point); molecular weight < 30 kDa (1 point); %cysteines > 4% (1 point); DN:DS > 1.5 (1 point); distance of 0–5 from an AT-rich region as predicted by OcculterCut [26] (1 point); EffectorP [47] score > 0.9 (1 point); an *in planta* to in vitro differential expression ratio (IP:IV) > 5 (2 points) or < 1 (– 2 points); sub-telomeric location within genome (within 500 Kbp of sequence end, 2 points); presence (– 3 points) or absence (3 points) of ortholog in low-virulence isolate Sn79–1087; presence of ortholog predicted as an effector candidate in US isolate Sn4 [31] (3 points); assigned an effector/toxin-like functional annotation (3 points) and; predicted lateral-gene transfer event with a fungal pathogen species [79] (3 points).

Abbreviations

AC: Accessory chromosome; AR: Accessory region; BFB: Breakage-fusion bridge; EDTA: Ethylenediaminetetraacetic acid; IP: *In planta*; IV: In vitro; P5-C3: 5th generation polymerase, 3rd generation chemistry; PAV: Presence-absence variation; PFGE: Pulsed-field gel eletrophoresis; RIP: Repeat-induced point mutation; SMRT: Single-molecule real-time; SNB: Septoria nodorum blotch; SNP: Single nucleotide polymorphism

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12864-021-07699-8.

Additional file 1: Supplementary Figure 1. Comparison of non-repetitive regions of accessory chromosome 23 (AC23, red) to other Sn15 chromosomes (black), indicating that it is not the product of duplication of a core, sister chromosome. The GC content of AC23 is indicated by the linear plot, and local repeat density is indicated in the heat map below (red). Nucleotide matches > 200 bp are indicated by grey arcs

Additional file 2: Supplementary Figure 2. Comparison of non-repetitive regions of accessory chromosome 23 (AC23, black) of > 100 bp in length (grey arcs) to *P. tritici-repentis* BFP chromosomes 1,3 4, and 11 (red), *P. tritici-repentis* M4 chromosomes 1, 3, 4, 6 and 10 (green), *Bipolaris maydis* scaffold 16 (blue) and *B. sorokiniana* chromosomes 2, 4, and 9. This comparison indicated a trend of telomeric proximity in the relative matching regions of related species.

Additional file 3: Supplementary Figure 3. Alignment of locally collinear blocks (LCBs) via Mauve, indicating large sections of similarity with structural rearrangements between Chromosome 4 of *P. nodorum* isolate Sn15 (top) and corresponding chromosomal sequences of isolates Sn4, Sn2000, and Sn79–1087, presented at the whole chromosome level (A) and within ~ 700–800 Kb of the telomere (B).

Additional file 4: Supplementary Table 1. Summary of draft (A) and high-quality (B) genome assemblies of *Parastagonospora* spp. alternate isolates used in this study for comparative genomics versus the Australian reference isolate Sn15.

Additional file 5: Supplementary Table 2. Comparison of repetitive sequence masking in the new *P. nodorum* Sn15 genome assembly using 3 different repeat libraries applied sequentially in 3 iterations.

Additional file 6: Supplementary Table 3. De novo repeat sequences predicted within the *Parastagonospora nodorum* Sn15 genome assembly.

Additional file 7: Supplementary Table 4. Summary of *Parastagonospora nodorum* gene properties and predicted effector candidate genes.

Additional file 8: Supplementary Table 5. Summary of assembled sequence lengths in the new *P. nodorum* Sn15 genome assembly, and estimates of their potential to be unresolved by PFGE comparing a 1% size error range to the size difference with the next longest sequence.

Additional file 9: Supplementary Table 6. Properties of selected large regions of the Sn15 assembly exhibiting presence absence variation (PAV) across the *Parastagonospora* population.

Additional file 10: Supplementary Table 7. Presence-absence variation (PAV) matrix for comparison of *Parastagonospora nodorum* Sn15 genes versus all other *Parastagonospora* spp. isolates included in this study.

Additional file 11: Supplementary Table 8. Summary of scaffold sequences from Syme et al. 2013 corresponding to chromosomes of new optical map-assisted long-read genome assembly for *Parastagonospora nodorum* Sn15

Additional file 12: Supplementary Table 9. Summary of genes and their functional annotations within the Chromosome 8 PAV region.

Additional file 13: Supplementary Table 10. Summary of average SNP density and DN/DS selection metrics across the *Parastagonospora* spp. population, relative to the *P. nodorum* Sn15 reference genome assembly.)

Additional file 14: Supplementary Table 11. Summary of gene and effector candidate gene distances from nearest AT-rich regions in the *P. nodorum* Sn15 assembly.

Additional file 15: Supplementary Table 12. Orthology comparison between gene predictions of *P. nodorum* Sn15 and Sn4, indicating Sn15 orthologs to Sn4 effector candidates under diversifying selection identified in Richards et al. 2019.

Additional file 16: Supplementary Table 13. Summary of gene content and functional annotation for *P. nodorum* Sn15 accessory chromosome 23 (AC23).

Additional file 17: Supplementary Text 1. Notes on the *P. nodorum* Sn15 genome assembly and the integration of optical mapping data.

Acknowledgements
We thank Prof. Richard Oliver for his expertise and guidance, and Ms. Julie Lawrence for her assistance with laboratory work relating to optical mapping.

Authors' contributions
KCT and JKH conceived the experiment and supervised the research. SB performed the majority of bioinformatics analysis, with assistance from DABJ and JKH. SB and JKH drafted the manuscript and JKH, HTP and KCT substantially revised the manuscript. All authors read and approved the manuscript.

Funding
This study was supported by the Centre for Crop and Disease Management, a joint initiative of Curtin University and the Grains Research and Development Corporation (Research Grant CUR00023). This research was undertaken with the assistance of resources and services from the Pawsey Supercomputing Centre and the National Computational Infrastructure (NCI), which is supported by the Australian Government (Research Grant Y95).

Availability of data and materials
The SN15 reference genome and protein data is available under BioProject: PRJNA686477. The updated SN15 genome assembly sequence consisting of 23 chromosomes is deposited under [Genome: GCA_016801405.1/ASM1680140v1; NUC: CP069023.1 - CP069045.1]. P. nodorum SN15 transcriptome data was previously deposited under BioProject:PRJNA632579. Alternate isolate data for *P. nodorum* and *P. avenae* was previously deposited under BioProject:PRJNA47648. Alternate reference isolate data was obtained for Sn79–1087 from BioProject:PRJNA398070 / GCA_002267025.1, for Sn4 from BioProject:PRJNA398070 / GCA_002267045.1, and for Sn2000 from BioProject:PRJNA398070 / GCA_002267005.1.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Author details
[1]Centre for Crop & Disease Management, Curtin University, Perth, Australia. [2]Curtin Institute for Computation, Curtin University, Perth, Australia.

Received: 22 October 2020 Accepted: 11 May 2021
Published online: 25 May 2021

References
1. Hane JK, Lowe RG, Solomon PS, Tan K-C, Schoch CL, Spatafora JW, et al. Dothideomycete–plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen Stagonospora nodorum. Plant Cell. 2007;19(11):3347–68. https://doi.org/10.1105/tpc.107.052829.
2. Liu Z, Friesen TL, Ling H, Meinhardt SW, Oliver RP, Rasmussen JB, et al. The Tsn1–ToxA interaction in the wheat–Stagonospora nodorum pathosystem parallels that of the wheat–tan spot system. Genome. 2006;49(10):1265–73. https://doi.org/10.1139/g06-088.
3. Liu Z, Zhang Z, Faris JD, Oliver RP, Syme R, McDonald MC, et al. The cysteine rich Necrotrophic effector SnTox1 produced by Stagonospora nodorum triggers susceptibility of wheat lines harboring Snn1. PLoS Pathog. 2012;8(1):e1002467. https://doi.org/10.1371/journal.ppat.1002467.
4. Liu Z, Faris JD, Oliver RP, Tan K-C, Solomon PS, McDonald MC, et al. SnTox3 acts in effector triggered susceptibility to induce disease on wheat carrying the Snn3 gene. PLoS Pathog. 2009;5(9):e1000581. https://doi.org/10.1371/journal.ppat.1000581.
5. Abeysekara NS, Friesen TL, Keller B, Faris JD. Identification and characterization of a novel host–toxin interaction in the wheat–Stagonospora nodorum pathosystem. Theor Appl Genet. 2009;120(1):117–26. https://doi.org/10.1007/s00122-009-1163-6.
6. Friesen TL, Chu C, Xu SS, Faris JD. SnTox5–Snn5: a novel S tagonospora nodorum effector–wheat gene interaction and its relationship with the SnToxA–Tsn1 and SnTox3–Snn3–B1 interactions. Mol Plant Pathol. 2012; 13(9):1101–9. https://doi.org/10.1111/j.1364-3703.2012.00819.x.
7. Friesen TL, Meinhardt SW, Faris JD. The Stagonospora nodorum-wheat pathosystem involves multiple proteinaceous host-selective toxins and corresponding host sensitivity genes that interact in an inverse gene-for-gene manner. Plant J. 2007;51(4):681–92. https://doi.org/10.1111/j.1365-313X.2007.03166.x.
8. Gao Y, Faris J, Liu Z, Kim Y, Syme R, Oliver R, et al. Identification and characterization of the SnTox6-Snn6 interaction in the Parastagonospora nodorum–wheat pathosystem. Mol Plant-Microbe Interact. 2015;28(5):615–25. https://doi.org/10.1094/MPMI-12-14-0396-R.
9. Lo Presti L, Lanver D, Schweizer G, Tanaka S, Liang L, Tollot M, et al. Fungal effectors and plant susceptibility. Annu Rev Plant Biol. 2015;66(1):513–45. https://doi.org/10.1146/annurev-arplant-043014-114623.
10. McDonald MC, Solomon PS. Just the surface: advances in the discovery and characterization of necrotrophic wheat effectors. Curr Opin Microbiol. 2018; 46:14–8. https://doi.org/10.1016/j.mib.2018.01.019.
11. Shi G, Friesen TL, Saini J, Xu SS, Rasmussen JB, Faris JD. The wheat Snn7 gene confers susceptibility on recognition of the Parastagonospora nodorum necrotrophic effector SnTox7. Plant Genome. 2015;8(2):1–10.
12. van Dam P, Fokkens L, Ayukawa Y, van der Gragt M, ter Horst A, Brankovics B, et al. A mobile pathogenicity chromosome in Fusarium oxysporum for

infection of multiple cucurbit species. Sci Rep. 2017;7(1):9042. https://doi.org/10.1038/s41598-017-07995-y.
13. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27(5):722–36. https://doi.org/10.1101/gr.215087.116.
14. Moolhuijzen P, See PT, Hane JK, Shi G, Liu Z, Oliver RP, et al. Comparative genomics of the wheat fungal pathogen Pyrenophora tritici-repentis reveals chromosomal variations and genome plasticity. BMC Genomics. 2018;19(1):279. https://doi.org/10.1186/s12864-018-4680-3.
15. Richards JK, Wyatt NA, Liu Z, Faris JD, Friesen TL. Reference quality genome assemblies of three *Parastagonospora nodorum* isolates differing in virulence on wheat. G3. 2018;8(2):393.
16. Van Kan JA, Stassen JH, Mosbach A, Van Der Lee TA, Faino L, Farmer AD, et al. A gapless genome sequence of the fungus Botrytis cinerea. Mol Plant Pathol. 2017;18(1):75–89. https://doi.org/10.1111/mpp.12384.
17. Wyatt NA, Richards JK, Brueggeman RS, Friesen TL. A comparative genomic analysis of the barley pathogen Pyrenophora teres f. teres identifies Subtelomeric regions as drivers of virulence. Mol Plant-Microbe Interact. 2020;33(2):173–88. https://doi.org/10.1094/MPMI-05-19-0128-R.
18. Derbyshire M, Denton-Giles M, Hegedus D, Seifbarghy S, Rollins J, van Kan J, et al. The complete genome sequence of the phytopathogenic fungus Sclerotinia sclerotiorum reveals insights into the genome architecture of broad host range pathogens. Genome Biol Evol. 2017;9(3):593–618. https://doi.org/10.1093/gbe/evx030.
19. Cooley RN, Caten CE. Variation in electrophoretic karyotype between strains of Septoria nodorum. Mol Gen Genet. 1991;228(1–2):17–23. https://doi.org/10.1007/BF00282442.
20. Syme RA, Hane JK, Friesen TL, Oliver RP. Resequencing and comparative genomics of Stagonospora nodorum: sectional gene absence and effector discovery. G3. 2013;3(6):959–69.
21. Bringans S, Hane JK, Casey T, Tan K-C, Lipscombe R, Solomon PS, et al. Deep proteogenomics; high throughput gene validation by multidimensional liquid chromatography and mass spectrometry of proteins from the fungal wheat pathogen Stagonospora nodorum. BMC Bioinformatics. 2009;10(1):301. https://doi.org/10.1186/1471-2105-10-301.
22. Ipcho SV, Hane JK, Antoni EA, Ahren D, Henrissat B, Friesen TL, et al. Transcriptome analysis of Stagonospora nodorum: gene models, effectors, metabolism and pantothenate dispensability. Mol Plant Pathol. 2012;13(6):531–45. https://doi.org/10.1111/j.1364-3703.2011.00770.x.
23. Syme RA, Tan K-C, Rybak K, Friesen TL, McDonald BA, Oliver RP, et al. Pan-Parastagonospora comparative genome analysis—effector prediction and genome evolution. Genome Biol Evol. 2018;10(9):2443–57. https://doi.org/10.1093/gbe/evy192.
24. Hane JK, Oliver RP. RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. BMC Bioinformatics. 2008;9(1):478. https://doi.org/10.1186/1471-2105-9-478.
25. Hane JK, Oliver RP. In silico reversal of repeat-induced point mutation (RIP) identifies the origins of repeat families and uncovers obscured duplicated genes. BMC Genomics. 2010;11(1):655. https://doi.org/10.1186/1471-2164-11-655.
26. Testa AC, Oliver RP, Hane JK. OcculterCut: a comprehensive survey of AT-rich regions in fungal genomes. Genome Biol Evol. 2016;8(6):2044–64. https://doi.org/10.1093/gbe/evw121.
27. Hane JK, Rouxel T, Howlett BJ, Kema GHJ, Goodwin SB, Oliver RP. A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. Genome Biol. 2011;12(5):R45. https://doi.org/10.1186/gb-2011-12-5-r45.
28. McDonald MC, Taranto AP, Hill E, Schwessinger B, Liu Z, Simpfendorfer S, et al. Transposon-mediated horizontal transfer of the host-specific virulence protein ToxA between three fungal wheat pathogens. mBio. 2019;10(5):e01515.
29. Ohm RA, Feau N, Henrissat B, Schoch CL, Horwitz BA, Barry KW, et al. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes Fungi. PLoS Pathog. 2012;8(12):e1003037. https://doi.org/10.1371/journal.ppat.1003037.
30. Bertazzoni S, Williams AH, Jones DA, Syme RA, Tan K-C, Hane JK. Accessories make the outfit: accessory chromosomes and other dispensable DNA regions in plant-pathogenic Fungi. Mol Plant-Microbe Interact. 2018;31(8):779–88. https://doi.org/10.1094/MPMI-06-17-0135-FI.
31. Richards JK, Stukenbrock EH, Carpenter J, Liu Z, Cowger C, Faris JD, et al. Local adaptation drives the diversification of effectors in the fungal wheat pathogen Parastagonospora nodorum in the United States. PLoS Genet. 2019;15(10):e1008223. https://doi.org/10.1371/journal.pgen.1008223.
32. Croll D, McDonald BA. The accessory genome as a cradle for adaptive evolution in pathogens. PLoS Pathog. 2012;8(4):e1002608. https://doi.org/10.1371/journal.ppat.1002608.
33. Gardiner DM, Benfield AH, Stiller J, Stephen S, Aitken K, Liu C, et al. A high-resolution genetic map of the cereal crown rot pathogen Fusarium pseudograminearum provides a near-complete genome assembly. Mol Plant Pathol. 2018;19(1):217–26. https://doi.org/10.1111/mpp.12519.
34. Plissonneau C, Hartmann FE, Croll D. Pangenome analyses of the wheat pathogen Zymoseptoria tritici reveal the structural basis of a highly plastic eukaryotic genome. BMC Biol. 2018;16(1):1–16.
35. Mendelowitz L, Pop M. Computational methods for optical mapping. Gigascience. 2014;3(1):33. https://doi.org/10.1186/2047-217X-3-33.
36. Burkhardt AK, Childs KL, Wang J, Ramon ML, Martin FN. Assembly, annotation, and comparison of Macrophomina phaseolina isolates from strawberry and other hosts. BMC Genomics. 2019;20(1):802. https://doi.org/10.1186/s12864-019-6168-1.
37. Farman ML. Telomeres in the rice blast fungus Magnaporthe oryzae : the world of the end as we know it. FEMS Microbiol Lett. 2007;273(2):125–32. https://doi.org/10.1111/j.1574-6968.2007.00812.x.
38. Rehmeyer C, Li W, Kusaba M, Kim Y-S, Brown D, Staben C, et al. Organization of chromosome ends in the rice blast fungus, Magnaporthe oryzae. Nucleic Acids Res. 2006;34(17):4685–701. https://doi.org/10.1093/nar/gkl588.
39. Soyer JL, El Ghalid M, Glaser N, Ollivier B, Linglin J, Grandaubert J, et al. Epigenetic control of effector gene expression in the plant pathogenic fungus Leptosphaeria maculans. PLoS Genet. 2014;10(3):e1004227. https://doi.org/10.1371/journal.pgen.1004227.
40. Tashiro S, Nishihara Y, Kugou K, Ohta K, Kanoh J. Subtelomeres constitute a safeguard for gene expression and chromosome homeostasis. Nucleic Acids Res. 2017;45(18):10333–49. https://doi.org/10.1093/nar/gkx780.
41. Jones DA, Bertazzoni S, Turo CJ, Syme RA, Hane JK. Bioinformatic prediction of plant-pathogenicity effector proteins of fungi. Curr Opin Microbiol. 2018;46:43–9. https://doi.org/10.1016/j.mib.2018.01.017.
42. Testa AC, Hane JK, Ellwood SR, Oliver RP. CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics. 2015;16(1):170. https://doi.org/10.1186/s12864-015-1344-4.
43. Jones DAB, John E, Rybak K, Phan HTT, Singh KB, Lin S-Y, et al. A specific fungal transcription factor controls effector gene expression and orchestrates the establishment of the necrotrophic pathogen lifestyle on wheat. Sci Rep. 2019;9(1):15884. https://doi.org/10.1038/s41598-019-52444-7.
44. Manning VA, Pandelova I, Dhillon B, Wilhelm LJ, Goodwin SB, Berlin AM, et al. Comparative genomics of a plant-pathogenic fungus, Pyrenophora tritici-repentis, reveals transduplication and the impact of repeat elements on pathogenicity and population divergence. G3. 2013;3(1):41–63.
45. McClintock B. The fusion of broken ends of chromosomes following nuclear fusion. Proc Natl Acad Sci. 1942;28(11):458–63. https://doi.org/10.1073/pnas.28.11.458.
46. Ferris MM, Yan X, Habbersett RC, Shou Y, Lemanski CL, Jett JH, et al. Performance assessment of DNA fragment sizing by high-sensitivity flow cytometry and pulsed-field gel electrophoresis. J Clin Microbiol. 2004;42(5):1965–76. https://doi.org/10.1128/JCM.42.5.1965-1976.2004.
47. Sperschneider J, Dodds PN, Gardiner DM, Singh KB, Taylor JM. Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. Mol Plant Pathol. 2018;19(9):2094–110. https://doi.org/10.1111/mpp.12682.
48. Urban M, Cuzick A, Seager J, Wood V, Rutherford K, Venkatesh SY, et al. PHI-base: the pathogen–host interactions database. Nucleic Acids Res. 2020;48(D1):D613–20. https://doi.org/10.1093/nar/gkz904.
49. Mohanta TK, Bae H. The diversity of fungal genome. Biol Proced Online. 2015;17(1):8–8. https://doi.org/10.1186/s12575-015-0020-z.
50. Denton JF, Lugo-Martinez J, Tucker AE, Schrider DR, Warren WC, Hahn MW. Extensive error in the number of genes inferred from draft genome assemblies. PLoS Comput Biol. 2014;10(12):e1003998. https://doi.org/10.1371/journal.pcbi.1003998.
51. Magrini V, Gao X, Rosa BA, McGrath S, Zhang X, Hallsworth-Pepin K, et al. Improving eukaryotic genome annotation using single molecule mRNA sequencing. BMC Genomics. 2018;19(1):172. https://doi.org/10.1186/s12864-018-4555-7.

52. Syme RA, Martin A, Wyatt NA, Lawrence JA, Muria-Gonzalez MJ, Friesen TL, et al. Transposable element genomic fissuring in Pyrenophora teres is associated with genome expansion and dynamics of host-pathogen genetic interactions. Front Genet. 2018;9:130. https://doi.org/10.3389/fgene.2018.00130.
53. Zolan ME. Chromosome-length polymorphism in fungi. Microbiol Rev. 1995;59(4):686–98. https://doi.org/10.1128/MR.59.4.686-698.1995.
54. Mossessova E, Lima CD. Ulp1-SUMO crystal structure and genetic analysis reveal conserved interactions and a regulatory element essential for cell growth in yeast. Mol Cell. 2000;5(5):865–76. https://doi.org/10.1016/S1097-2765(00)80326-3.
55. Nagano N, Umemura M, Izumikawa M, Kawano J, Ishii T, Kikuchi M, et al. Class of cyclic ribosomal peptide synthetic genes in filamentous fungi. Fungal Genet Biol. 2016;86:58–70. https://doi.org/10.1016/j.fgb.2015.12.010.
56. Stahelin RV, Long F, Diraviyam K, Bruzik KS, Murray D, Cho W. Phosphatidylinositol 3-phosphate induces the membrane penetration of the FYVE domains of Vps27p and Hrs. J Biol Chem. 2002;277(29):26379–88. https://doi.org/10.1074/jbc.M201106200.
57. Phan HTT, Rybak K, Bertazzoni S, Furuki E, Dinglasan E, Hickey LT, et al. Novel sources of resistance to Septoria nodorum blotch in the Vavilov wheat collection identified by genome-wide association studies. Theor Appl Genet. 2018;131(6):1223–38. https://doi.org/10.1007/s00122-018-3073-y.
58. Tan K-C, Phan HTT, Rybak K, John E, Chooi YH, Solomon PS, et al. Functional redundancy of necrotrophic effectors – consequences for exploitation for breeding. Front Plant Sci. 2015;6:501.
59. Phan HTT, Rybak K, Furuki E, Breen S, Solomon PS, Oliver RP, et al. Differential effector gene expression underpins epistasis in a plant fungal disease. Plant J. 2016;87(4):343–54. https://doi.org/10.1111/tpj.13203.
60. Rybak K, See PT, Phan HT, Syme RA, Moffat CS, Oliver RP, et al. A functionally conserved Zn (2) Cys (6) binuclear cluster transcription factor class regulates necrotrophic effector gene expression and host-specific virulence of two major Pleosporales fungal pathogens of wheat. Mol Plant Pathol. 2017;18(3):420–34. https://doi.org/10.1111/mpp.12511.
61. Zhu W, Wei W, Wu Y, Zhou Y, Peng F, Zhang S, et al. BcCFEM1, a CFEM domain-containing protein with putative GPI-anchored site, is involved in pathogenicity, conidial production, and stress tolerance in Botrytis cinerea. Front Microbiol. 2017;8:1807. https://doi.org/10.3389/fmicb.2017.01807.
62. Kombrink A, Thomma BPHJ. LysM effectors: secreted proteins supporting fungal life. PLoS Pathog. 2013;9(12):e1003769. https://doi.org/10.1371/journal.ppat.1003769.
63. Zhang J, Wang F, Liang F, Zhang Y, Ma L, Wang H, et al. Functional analysis of a pathogenesis-related thaumatin-like protein gene TaLr35PR5 from wheat induced by leaf rust fungus. BMC Plant Biol. 2018;18(1):76. https://doi.org/10.1186/s12870-018-1297-2.
64. de Guillen K, Lorrain C, Tsan P, Barthe P, Petre B, Saveleva N, et al. Structural genomics applied to the rust fungus Melampsora larici-Populina reveals two candidate effector proteins adopting cystine knot and NTF2-like protein folds. Sci Rep. 2019;9(1):18084. https://doi.org/10.1038/s41598-019-53816-9.
65. Burke JE, Dennis EA. Phospholipase A2 structure/function, mechanism, and signaling. J Lipid Res. 2009;50(Suppl (Suppl)):S237–42.
66. Chruszcz M, Chapman MD, Osinski T, Solberg R, Demas M, Porebski PJ, et al. Alternaria alternata allergen Alt a 1: a unique β-barrel protein dimer found exclusively in fungi. J Allergy Clin Immunol. 2012;130(1):241–247.e249.
67. Zhang Y, Gao Y, Liang Y, Dong Y, Yang X, Qiu D. Verticillium dahliae PevD1, an Alt a 1-like protein, targets cotton PR5-like protein and promotes fungal infection. J Exp Bot. 2019;70(2):613–26. https://doi.org/10.1093/jxb/ery351.
68. Olombrada M, Peña C, Rodríguez-Galán O, Klingauf-Nerurkar P, Portugal-Calisto D, Oborská-Oplová M, et al. The ribotoxin α-sarcin can cleave the sarcin/ricin loop on late 60S pre-ribosomes. Nucleic Acids Res. 2020;48(11):6210–22. https://doi.org/10.1093/nar/gkaa315.
69. Baccelli I. Cerato-platanin family proteins: one function for multiple biological roles? Front Plant Sci. 2015;5:769.
70. Antuch W, Güntert P, Wüthrich K. Ancestral βγ-crystallin precursor structure in a yeast killer toxin. Nat Struct Biol. 1996;3(8):662–5. https://doi.org/10.1038/nsb0896-662.
71. Figueiredo J, Sousa Silva M, Figueiredo A. Subtilisin-like proteases in plant defence: the past, the present and beyond. Mol Plant Pathol. 2018;19(4):1017–28. https://doi.org/10.1111/mpp.12567.
72. Singh K, Winter M, Zouhar M, Ryšánek P. Cyclophilins: less studied proteins with critical roles in pathogenesis. Phytopathology. 2017;108(1):6–14. https://doi.org/10.1094/PHYTO-05-17-0167-RVW.
73. Viaud MC, Balhadère PV, Talbot NJ. A Magnaporthe grisea cyclophilin acts as a virulence determinant during plant infection. Plant Cell. 2002;14(4):917–30. https://doi.org/10.1105/tpc.010389.
74. Zhao Z, Liu H, Wang C, Xu J-R. Comparative analysis of fungal genomes reveals different plant cell wall degrading capacity in fungi. BMC Genomics. 2013;14(1):274. https://doi.org/10.1186/1471-2164-14-274.
75. Grell MN, Mouritzen P, Giese H. A Blumeria graminis gene family encoding proteins with a C-terminal variable region with homologues in pathogenic fungi. Gene. 2003;311:181–92. https://doi.org/10.1016/S0378-1119(03)00610-3.
76. Schäfer W. The role of cutinase in fungal pathogenicity. Trends Microbiol. 1993;1(2):69–71. https://doi.org/10.1016/0966-842X(93)90037-R.
77. Zhang M, Xie S, Zhao Y, Meng X, Song L, Feng H, et al. Hce2 domain-containing effectors contribute to the full virulence of Valsa Mali in a redundant manner. Mol Plant Pathol. 2019;20(6):843–56. https://doi.org/10.1111/mpp.12796.
78. Temple B, Horgen PA, Bernier L, Hintz WE. Cerato-ulmin, a hydrophobin secreted by the causal agents of Dutch elm disease, is a parasitic fitness factor. Fungal Genet Biol. 1997;22(1):39–53. https://doi.org/10.1006/fgbi.1997.0991.
79. Hane JK. Effector-like ROG Predictions. https://effectordb.com/lgt-effector-predictions-summary/? Accessed 20 Oct 2020.
80. Walker BJ, Abeel T, Shea T, Priest M, Abouelliel A, Sakthikumar S, et al. Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. PLoS One. 2014;9(11):e112963. https://doi.org/10.1371/journal.pone.0112963.
81. Langmead B, Salzberg SL. Fast gapped-read alignment with bowtie 2. Nat Methods. 2012;9(4):357–9. https://doi.org/10.1038/nmeth.1923.
82. Solomon PS, Thomas SW, Spanu P, Oliver RP. The utilisation of di/tripeptides by Stagonospora nodorum is dispensable for wheat infection. Physiol Mol Plant Pathol. 2003;63(4):191–9. https://doi.org/10.1016/j.pmpp.2003.12.003.
83. Gurevich A, Saveliev V, Vyahhi N, Tesler G. QUAST: quality assessment tool for genome assemblies. Bioinformatics. 2013;29(8):1072–5. https://doi.org/10.1093/bioinformatics/btt086.
84. Seppey M, Manni M, Zdobnov EM. BUSCO: assessing genome assembly and annotation completeness. Methods Mol Biol. 1962;2019:227–45.
85. Li H. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM. arXiv preprint arXiv:13033997; 2013.
86. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The sequence alignment/map format and SAMtools. Bioinformatics. 2009;25(16):2078–9. https://doi.org/10.1093/bioinformatics/btp352.
87. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2. https://doi.org/10.1093/bioinformatics/btq033.
88. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, et al. Versatile and open software for comparing large genomes. Genome Biol. 2004;5(2):R12. https://doi.org/10.1186/gb-2004-5-2-r12.
89. Dot. https://github.com/dnanexus/dot. Accessed 20 Oct 2020.
90. RepeatMasker Open-4.0. http://www.repeatmasker.org. Accessed 20 Oct 2020.
91. Bao W, Kojima KK, Kohany O. Repbase update, a database of repetitive elements in eukaryotic genomes. Mob DNA. 2015;6(1):11. https://doi.org/10.1186/s13100-015-0041-9.
92. Gremme G, Steinbiss S, Kurtz S. GenomeTools: a comprehensive software library for efficient processing of structured genome annotations. IEEE/ACM Trans Comput Biol Bioinform. 2013;10(3):645–56. https://doi.org/10.1109/TCBB.2013.68.
93. Flynn JM, Hubley R, Goubert C, Rosen J, Clark AG, Feschotte C, et al. RepeatModeler2 for automated genomic discovery of transposable element families. Proc Natl Acad Sci. 2020;117(17):9451–7. https://doi.org/10.1073/pnas.1921046117.
94. Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999;27(2):573–80. https://doi.org/10.1093/nar/27.2.573.
95. Qi X, Li Y, Honda S, Hoffmann S, Marz M, Mosig A, et al. The common ancestral core of vertebrate and fungal telomerase RNAs. Nucleic Acids Res. 2013;41(1):450–62. https://doi.org/10.1093/nar/gks980.
96. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. 2011;17(1):3.

97. Brown CT, Howe A, Zhang Q, Pyrkosz AB, Brom TH. A reference-free algorithm for computational normalization of shotgun sequencing data. arXiv preprint arXiv:12034802; 2012.
98. Crusoe MR, Alameldin HF, Awad S, Boucher E, Caldwell A, Cartwright R, et al. The khmer software package: enabling efficient nucleotide sequence analysis. F1000Res. 2015;4:900.
99. Bushnell B. BBMap: a fast, accurate, splice-aware aligner. Berkeley: Lawrence Berkeley National Lab.(LBNL); 2014.
100. Clavijo BJ, Venturini L, Schudoma C, Accinelli GG, Kaithakottil G, Wright J, et al. An improved assembly and annotation of the allohexaploid wheat genome identifies complete families of agronomic genes and provides genomic evidence for chromosomal translocations. Genome Res. 2017;27(5): 885–96. https://doi.org/10.1101/gr.217117.116.
101. Dobin A, Davis CA, Schlesinger F, Drenkow J, Zaleski C, Jha S, et al. STAR: ultrafast universal RNA-seq aligner. Bioinformatics. 2013;29(1):15–21. https://doi.org/10.1093/bioinformatics/bts635.
102. Haas BJ, Papanicolaou A, Yassour M, Grabherr M, Blood PD, Bowden J, et al. De novo transcript sequence reconstruction from RNA-seq using the trinity platform for reference generation and analysis. Nat Protoc. 2013;8(8):1494–512. https://doi.org/10.1038/nprot.2013.084.
103. Kim D, Pertea G, Trapnell C, Pimentel H, Kelley R, Salzberg SL. TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions. Genome Biol. 2013;14(4):R36. https://doi.org/10.1186/gb-2013-14-4-r36.
104. Trapnell C, Roberts A, Goff L, Pertea G, Kim D, Kelley DR, et al. Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and cufflinks. Nat Protoc. 2012;7(3):562–78. https://doi.org/10.1038/nprot.2012.016.
105. Pertea M, Kim D, Pertea GM, Leek JT, Salzberg SL. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. Nat Protoc. 2016;11(9):1650–67. https://doi.org/10.1038/nprot.2016.095.
106. Haas BJ, Salzberg SL, Zhu W, Pertea M, Allen JE, Orvis J, et al. Automated eukaryotic gene structure annotation using EVidenceModeler and the program to assemble spliced alignments. Genome Biol. 2008;9(1):R7. https://doi.org/10.1186/gb-2008-9-1-r7.
107. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. 2008;18(12):1979–90. https://doi.org/10.1101/gr.081612.108.
108. Huang X, Adams MD, Zhou H, Kerlavage AR. A tool for analyzing and annotating genomic sequences. Genomics. 1997;46(1):37–45. https://doi.org/10.1006/geno.1997.4984.
109. Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, et al. Web Apollo: a web-based genomic annotation editing platform. Genome Biol. 2013;14(8):R93. https://doi.org/10.1186/gb-2013-14-8-r93.
110. Jones P, Binns D, Chang H-Y, Fraser M, Li W, McAnulla C, et al. InterProScan 5: genome-scale protein function classification. Bioinformatics. 2014;30(9): 1236–40. https://doi.org/10.1093/bioinformatics/btu031.
111. Nielsen H. Predicting secretory proteins with SignalP. In: Protein Function Prediction: Springer; 2017. p. 59–73.
112. Sperschneider J, Catanzariti A-M, DeBoer K, Petre B, Gardiner DM, Singh KB, et al. LOCALIZER: subcellular localization prediction of both plant and effector proteins in the plant cell. Sci Rep. 2017;7(1):1–14.
113. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2012; 40(W1):W445–51. https://doi.org/10.1093/nar/gks479.
114. Eberhardt RY, Haft DH, Punta M, Martin M, O'Donovan C, Bateman A. AntiFam: a tool to help identify spurious ORFs in protein annotation Database 2012; 2012.
115. Cingolani P, Platts A, Wang LL, Coon M, Nguyen T, Wang L, et al. A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff: SNPs in the genome of Drosophila melanogaster strain w1118; iso-2; iso-3. Fly. 2012;6(2):80–92. https://doi.org/10.4161/fly.19695.
116. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009;19(9):1639–45. https://doi.org/10.1101/gr.092759.109.

Publisher's Note

CHAPTER 11 — THEMES 2 & 3

Novel effector candidates and large accessory genome revealed by population genomic analysis of *Parastagonospora nodorum*

A revised preprint of this chapter is available at:
bioRxiv https://doi.org/10.1101/2021.09.01.458590

11.1 Declaration

Title	Novel effector candidates and large accessory genome revealed by population genomic analysis of Parastagonospora nodorum.
Authors	**Darcy A. B. Jones**, Kasia Rybak, Stefania Bertazzoni, Kar-Chun Tan, Huyen T. T. Phan, James K. Hane

This thesis chapter is submitted in the form of a collaboratively-written manuscript ready for journal submission. As such, not all work contained within this chapter can be attributed to the Ph. D. candidate.

The following contributions were made by the Ph.D candidate (DABJ) and co-authors:

- **DABJ**, K-CT, and HTTP conceived and designed the study.
- KR and HTTP performed DNA extraction.
- **DABJ**, SB, and JKH performed bioinformatics and data analysis.
- **DABJ** and JKH wrote the manuscript.
- **DABJ**, K-CT, HTTP, and JKH edited the manuscript.
- All authors read and approved the manuscript.

I, Darcy Jones, hereby certify that this attribution statement is an accurate record of my contribution to the research presented in this chapter.

Darcy Jones

I certify that this attribution statement is an accurate record of Darcy Jones' contribution to the research presented in this chapter.

Kasia Rybak | Stefania Bertazzoni | Kar-Chun Tan

Huyen T. T. Phan | James K. Hane

Abstract

The necrotrophic pathogen *Parastagonospora nodorum* (septoria nodorum blotch (SNB)) causes significant wheat yield losses globally. Since the mid 2000s, *P. nodorum* has emerged as a model necrotroph species and has accumulated significant genomic and bioinformatic resources, which are currently moving towards population-level pan-genome studies. These can be leveraged to provide novel insights into fungal pathogen evolution and pathogenicity effector interactions relevant to local crop disease outbreaks. In this study, we examine 156 isolates representing a regional population from the Western Australian (WA) wheat-belt region, which were also compared against 17 internationally-sourced isolates. We observed a highly diverse local population, within which were numerous small and highly-similar clusters of isolates from hotter and drier regions. Pan-genome-level assembly and orthologous gene datasets were generated — from which 3579 effector candidates were predicted, 2291 of which exhibited presence-absence variation (PAV) across the population, and 1362 were specific to WA isolates — enhancing previous effector discovery efforts. Across the pan-genome, we observed an abundance of mutations (including repeat-induced point mutation (RIP)), which were distributed in 'hot-spots' within the genomic landscape and also exhibited spatial distribution bias in terms of isolate locations. However three characterised effector loci (*ToxA*, *Tox1* and *Tox3*) were located within regions of lower overall diversity, yet were still flanked within large hotspot regions when observed at a larger (i.e. chromosome-level) scale. RIP was observed to be widespread across the genome, but non-synonymous RIP-like mutations appeared to be strongly selected against in this population — which is consistent with expectations for the genome of a RIP-active necrotroph. Overall this study has progressively improved bioinformatic resources for the necrotroph model species *P. nodorum*, as well as advancing approaches for the study of fungal pan-genomes with a view towards developing a region-specific understanding of host-pathogen interactions.

11.2 Introduction

Parastagonospora nodorum is a necrotrophic pathogen causing septoria nodorum blotch (SNB) of wheat (*Triticum spp.*) (Solomon et al., 2006) which leads to significant yield losses in Australia (Murray & Brennan, 2009). *P. nodorum* is primarily spread by infected seed, infested debris or by wind dispersed sexual ascospores (Solomon et al., 2006). Secondary infections can occur when water splash spreads asexual pycnidiospores to higher leaves and glumes, causing further necrotic patches and crop loss. *P. nodorum* is observed to be highly diverse in the field (McDonald et al., 2012; Stukenbrock et al., 2006), and appears to regularly reproduce sexually (Bathgate & Loughman, 2001; Murphy et al., 2000; Sommerhalder et al., 2006). This suggests a high adaptive capacity of *P. nodorum* populations, where application of strong selective pressures may be quickly overcome by extant diversity.

P. nodorum infection heavily relies on the activity of necrotrophic effector (NE) proteins, which are secreted into the host and cause cell death upon recognition by host susceptibility (S)-proteins (Tan et al., 2010). Three NEs have been characterised in *P. nodorum* to date: ToxA (Liu et al., 2006), Tox1 (Liu et al., 2012) and Tox3 (Liu et al., 2009). At least five additional

host specific necrosis phenotypic interactions have been described in the *P. nodorum*-wheat pathosystem (Abeysekara et al., 2009; Friesen et al., 2012; Friesen et al., 2007; Friesen et al., 2008; Gao et al., 2015; Phan et al., 2018; Shi et al., 2015; Z. Zhang et al., 2011), indicating the presence of additional undiscovered NEs. Identification of the *ToxA* NE has led to deployment of a ToxA resistant wheat cultivar (Tan et al., 2014), and the presence of additional major disease resistance quantitative trait loci (QTL) encourages further development of disease resistant cultivars. However, the known epistatic interactions of Tox1 and Tox2 over Tox3 (Friesen et al., 2007; Phan et al., 2016) indicates that the contributions of different effector-receptor interactions to virulence are complex, and that reliable markers of S-genes and knowledge of the epistasis interactions are important for future disease breeding efforts. The discovery of novel NEs in *P. nodorum* (and other fungal pathogens) remains an important element of crop protection research (Vleeshouwers & Oliver, 2014). Fungal effector discovery relies heavily on genomic and bioinformatic resources (D. A. Jones et al., 2018), and the increased accessibility of sequencing has resulted in a considerable increase in the rate of effector discovery (Kanja & Hammond-Kosack, 2020).

P. nodorum was among the first fungal species for which a reference genome sequence was generated (for the Western Australian (WA)) isolate SN15) (Hane et al., 2007), and was the first species within the class Dothideomycetes, which comprises several prominent cereal necrotroph and hemibiotroph species (Aylward et al., 2017; Ohm et al., 2012). The SN15 reference isolate has predominated molecular plant pathology studies of *P. nodorum* since, becoming an important model for the cereal necrotrophs (Solomon et al., 2006). Significant resources have accumulated over time for the SN15 isolate and *P. nodorum* in general, including transcriptomic (Hane et al., 2007; Ipcho et al., 2012; D. A. B. Jones et al., 2019; Richards et al., 2018; Syme et al., 2016), proteomic (Bringans et al., 2009; Syme et al., 2016), and metabolomic (Chooi et al., 2014; Gummer et al., 2013; R. G. T. Lowe et al., 2008; Muria-Gonzalez et al., 2020) datasets. Notable recent additions to the growing pool of *P. nodorum* data include the chromosome-scale genome assemblies of four reference isolates (including SN15) (Bertazzoni et al., 2021; Richards et al., 2018, chapter 10).

Parastagonospora nodorum pan-genomics

Improving cost and availability of genome and transcriptome sequencing has significantly advanced plant pathology, enabling conceptual shifts from the focussed study of a single reference isolate, to large-scale comparative genomics between numerous isolates over a few short years. Three pan-genomic comparative studies of *P. nodorum* have been conducted to date, comparing global collections of isolates (Pereira et al., 2020; Syme et al., 2018), and populations within the USA (Richards et al., 2019). Syme et al. (2018) compared the genomes of *P. nodorum* isolates from Iran, Finland, Sweden, Switzerland, South Africa, the USA, and Australia. They observed frequent presence-absence variation (PAV) in the effectors *ToxA*, *Tox1*, and *Tox3*, and possible accessory genomic regions with potential roles in virulence. Additionally, numerous genes were observed to be under positive selection pression, including

several effector candidates. They observed distinct regions with large numbers of mutations and some regions with consistently high dN/dS ratios, indicating positive selection and enrichment of adaptively relevant genes in specific genomic regions. Pereira et al. (2020) sequenced the genomes of *P. nodorum* isolates from Australia, Iran, South Africa, Switzerland, and the USA to investigate *P. nodorum* adaptation to fungicide use. The highest genetic diversity was found in the Iranian population, consistent with the hypothesis that *P. nodorum* co-evolved with wheat during early domestication (Ghaderi et al., 2020), while low diversity was observed within the Australian isolates. A genome-wide association study identified several loci correlated with azole resistance, with higher incidence of fungicide resistance and the identified resistance associated alleles in the Swiss population. Richards et al. (2019) sequenced 197 *P. nodorum* isolates collected from Spring, Winter and Durum wheat cultivars across the growing region of the USA. Two major US sub-populations were identified corresponding to geographical features and wheat lines grown, with one sub-population almost completely lacking *ToxA*. Both populations had diversified at different loci, indicating distinct selective pressures and resulting in different sets of effector candidates predicted by genome-wide association. The USA population study also highlighted several patterns of gene PAV between isolates. The known effector genes *ToxA*, *Tox1* and *Tox3* were absent in 37%, 5% and 41% of US isolates respectively. Gene PAV was commonly associated with transposons in USA isolates, yet there was no significant association with frequency of secreted or effector-like proteins.

Collectively, these genomic studies have highlighted the high diversity of *P. nodorum* genomes, and that studies focussing only on reference isolates are missing considerable amounts of information. The 400 kb accessory chromosome 23 (AC23), which is missing in the avirulent isolate Sn79-1087, may have a role in host-specific virulence (Bertazzoni et al., 2018) and has a high background rate of mutation (Bertazzoni et al., 2021; Richards et al., 2018, chapter 10). Similarly, regional biases for high AT-base content and repeat-induced point mutation (RIP)-like activity suggests the presence of rapidly mutating accessory regions, primarily within repeat-rich stretches of AC23 and sub-telomeric regions (Bertazzoni et al., 2021; Richards et al., 2018; Syme et al., 2018, chapter 10). Numerous effector candidates have been derived from these genomic studies, utilising features such as signatures of positive selection across isolates, similarity to known effectors, signal peptides, EffectorP results (Sperschneider, Dodds, Gardiner, et al., 2018; Sperschneider et al., 2016), genomic context (e.g. AT-richness or distance to transposable elements (TEs)), presence or absence in avirulent isolates, and genome wide association (Richards et al., 2019; Richards et al., 2018; Syme et al., 2018).

In the Australian reference isolate SN15, many of these genomic features have been combined with a broad range of experimental and bioinformatic indicators to refine effector candidate lists, including: *in planta* gene expression (D. A. B. Jones et al., 2019), putative lateral gene transfer with other cereal-pathogenic fungi (https://effectordb.com/lgt-effector-predictions-summary), and predicted effector-like gene and protein properties (Bertazzoni et al., 2021; Syme et al., 2018, chapter 10).

The Western Australian *P. nodorum* pan-genome

Despite the long history of study of the Australian *P. nodorum* reference isolate SN15, relatively little was known about the genomic diversity of the *P. nodorum* population in WA until recently. A study of 28 short sequence repeat (SSR) loci compared a WA population of 155 isolates collected over 44 years, and contrasted this population with 23 international isolates sourced from France and the USA (Phan et al., 2020). This SSR study identified two core admixed clusters in WA, with three low-diversity satellite clusters that were geographically and temporally restricted. The population shift broadly correlated with historical shifts in wheat cultivar preference, particularly after the mass adoption of the ToxA insensitive cultivar "Mace" in 2013 which covers nearly 70% of the area sown (Trainor et al., 2018). Although wheat cultivar disease resistance has increased over time, more recently sampled isolates from emergent clusters were more pathogenic than older isolates.

In this study, we further dissect the evolutionary history of a WA *P. nodorum* population previously analysed by Phan et al. (2020) using whole genome sequencing. Further, we compare these genomes with isolates previously sequenced by Syme et al. (2018), and identify novel effector candidates in *P. nodorum*. Guided by phylogeographic and population structure analyses, we observe a highly diverse population in WA, with numerous small highly similar clusters collected from hotter and drier regions of Western Australia. We present the first report of novel sequences and genes within the growing pool of *P. nodorum* pangenome data, specific to locally-adapted isolates. We also combine new WA pangenome data with existing resources to define new effector candidate orthologous groups, extending effector candidate discovery beyond reference isolates. Overall, this growing wealth of new pathogen population data has enhanced our understanding of the pathogenicity gene content, genome architecture and population dynamics for the model cereal necrotroph, *P. nodorum*.

11.3 Methods

11.3.1 DNA extraction and sequencing of Western Australian *P. nodorum* isolates

A total of 156 Western Australian (WA) *P. nodorum* isolates described previously were sequenced using short-read Illumina sequencing [NCBI BioProject: PRJNA612761]. Genomic DNA of 141 isolates was extracted using the method described by Xin and Chen (2012), and was sequenced by the Australian Genome Research Facility (Melbourne, Australia) using 125 bp paired end (PE) reads on an Illumina HiSeq2500, using an Illumina TruSeq PCR-free library with a target fragment size of 600 bp. Genomic DNA of 17 additional isolates (including two isolates sequenced with the original 141 isolates, 14FG141 and Mur_S3) was extracted using a Qiagen DNeasy Plant Mini kit (Venlo, Netherlands. Catalogue ID: 69104) and sequenced by Novogene (Beijing, China), using 150bp PE reads on an Illumina HiSeq2500, also using an Illumina TruSeq PCR-free library and a target fragment size of 350bp. Genome data for 15 additional

international *P. nodorum* isolates [NCBI BioProject: PRJNA476481] were sourced from an earlier pan-genome study (Syme et al., 2018) (Supplementary table S1). The chromosome-scale genome sequence for the *P. nodorum* SN15 reference isolate has been previously described (Bertazzoni et al., 2021, chapter 10). Chromosome-scale genome assemblies for the SN4, SN2000 and SN79-1087 isolates from the US [NCBI BioProject: PRJNA398070], were sourced from Richards et al. (2018).

Reads were trimmed for adapter content and very low-quality sequence regions using CutAdapt (version 1.18) in two passes allowing three trims per pass, trimming ends with less than Phred score 2, screened against TruSeq universal adapters (Martin, 2011). Reads with an average Phred score less than 5 or with a length after trimming less than 50 bp were discarded. Potential laboratory and technical contaminants such as PhiX were filtered out using BBduk version 38.38 (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbduk-guide/) requiring read kmer coverage of 0.7, using the UniVec database (https://www.ncbi.nlm.nih.gov/tools/vecscreen/univec/) and the PhiX genome (NCBI RefSeq: NC_001422.1) (Sanger et al., 1978) as bait templates. Potential sample contaminants were searched for using Kraken (version 2.0.7) (Wood et al., 2019) searching against a database constructed from all NCBI Refseq bacterial, archaeal, protozoan, viral, and fungal genomes (downloaded: 2019-03-16), as well as the human GRCh38 genome (Wood & Salzberg, 2014) (Supplementary table S2). The four published reference *P. nodorum* genomes (Bertazzoni et al., 2021; Richards et al., 2018, chapter 10) were also included as a positive set. Reads were aligned to the four reference *P. nodorum* genomes using BBmap version 38.38 (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/bbmap-guide/), to evaluate insert size and completeness. Quality control statistics for each step were collected using FastQC version 0.11.8 (https://www.bioinformatics.babraham.ac.uk/projects/fastqc/), BBmap and Samtools (Li et al., 2009), and were collated using MultiQC (Ewels et al., 2016) (Supplementary data S1). Code for performing quality control (QC) steps is available online at https://github.com/darcyabjones/qcflow (commit: e77715d).

11.3.2 Read alignment and variant calling

Short reads from all samples were aligned to the *P. nodorum* SN15 reference genome (Bertazzoni et al., 2021, chapter 10) using Stampy version 1.0.32 allowing multi-mapping reads (`--sensitive --substitutionrate=0.0001 --xa-max=3 --xa-max-discordant=10`) (Lunter & Goodson, 2011). PCR duplicate reads in the aligned binary alignment map (BAM) files were marked using Picard version 2.18.29 (http://broadinstitute.github.io/picard/). Short variants were then predicted from the BAM files using GATK version 4.1.0.0 (McKenna et al., 2010; Poplin et al., 2018) and the following "bootstrapped" pipeline: 1) Call individual variants using `gatk HaplotypeCaller`. 2) Combine variants from all samples and run joint genotype prediction using `gatk GenotypeGVCFs`. 3) filter first pass variants with extreme statistics for each mutation type (snp, indel, or mixed) using `gatk VariantFiltration`. 4) Recalibrate the base quality scores in the BAM files using the predicted variants, using `gatk BaseRecalibrator` and `gatk ApplyBQSR`. 5) Steps 1-4 were repeated using the recalibrated BAM files until there

was no difference in base quality score recalibration (BQSR) statistics between successive iterations. Filters applied at each bootstrap iteration are detailed in supplementary table S3. An initial set of variants was found from the 140 isolates initially sequenced (excluding the resequenced SN15 isolate), using 5 bootstrap iterations to converge the BQSR scores. Nineteen additional sequenced isolates and 18 previously sequenced isolates (Syme et al., 2018), were then included in the analysis using the previously identified variant loci as starting points for two further bootstrap iterations including alignments from all isolates. Unfiltered variants were taken from the final bootstrap step and filtered more stringently. Individual genotype depth and genotype quality scores were visualised using the R (R Core Team, 2019) packages vcfR (Knaus & Grünwald, 2017) and ggplot2 (Wickham, 2016) to determine appropriate statistic cutoff thresholds, and were soft filtered using `gatk VariantFiltration` to have a minimum genotype "DP" statistic of 8 and minimum "GQ" score of 30. Variant locus "DP" and "MQ" score ranges to include were determined based on scores in non-repetitive regions visualised using a vcfR "chromoplot". Other variant loci statistics were visualised using ggplot2 to determine cutoff thresholds, and loci were filtered using `gatk VariantFiltration` separately for each variant type using the selected thresholds. All filtering parameters used during variant prediction are presented in supplementary table S3. Filtered variants with effects on the translation of annotated genes (Bertazzoni et al., 2021, chapter 10) (i.e. non-synonymous or nonsense mutations) were identified using SnpEff (Cingolani et al., 2012). Variants called relative to the SN15 reference genome are available at https://doi.org/10.6084/m9.figshare.13340975.

11.3.3 Phylogeny estimation and population structure analysis

Single nucleotide polymorphisms (SNPs) for phylogenetic and population analyses were selected by excluding SNPs missing in more than 30% of isolates or with a SnpEff impact prediction of "HIGH" or "MODERATE". To reduce the potential impact of small scale linkage disequilibrium, SNPs were selected using PLINK version 1.9 (Purcell et al., 2007) by taking SNPs with the highest minor allele frequency where two SNPs within 10 kb have an R^2 corellation greater than 0.6 (`--indep-pairwise 10 kb 1 0.60`). Resulting SNPs were then filtered to have a minimum non-major allele frequency of 0.05 using BCFtools (Li, 2011) (`--min-af "0.05:nonmajor"`). SNPs were converted to a sequence alignment and the substitution model best fitting the data was predicted using ModelFinder in IQTree version 2.0.3 (`-st DNA -m "MF+ASC" -mset "GTR,JC,F81,K80,HKY,K81" -cmax 15 -rcluster 25 -safe`) (Kalyaanamoorthy et al., 2017). The best performing model was selected and used to construct a maximum likelihood tree using IQTree with 10000 UFBoot (Hoang et al., 2018) and 1000 SH-aLRT iterations (`-bb 10000 -bnni -alrt 1000 -st DNA`) (Minh et al., 2020). Phylogenetic trees were plotted using the R version 4.0.2 (R Core Team, 2019) packages phytools v0.7-47 and ggtree v2.2.4 (Yu et al., 2017). Phylogenetic trees from this study and the tree published by Phan et al. (2020) were visually compared using tanglegrams, implemented in the dendextend (version 1.14.0) package (Galili, 2015).

Population structure was inferred from SNP data using STRUCTURE version 2.3.4 (Pritchard

et al., 2000). To select an appropriate number of subpopulations (K) to model, STRUCTURE was run with 10000 replicate burn-in period and 20000 MCMC replicates for a range of values of K between 1 and 12, rerunning 8 times for each K to account for random starting points. The optimal value of K was selected using STRUCTURE HARVESTER (Earl & vonHoldt, 2012) using the method described by Evanno et al. (2005). STRUCTURE was run using the selected value of K, using a 20000 replicate burn-in period and 100000 MCMC replications, running with 8 random seeds and selecting the run with the highest log probability of data given the model.

Population statistics were calculated and visualised using R version 4.0.2 (R Core Team, 2019) packages: ade4 v1.7-15 (Thioulouse et al., 1997), adegenet v2.1.3 (Jombart, 2008), poppr v2.8.6 (Kamvar et al., 2014), and vcfR v1.12.0 (Knaus & Grünwald, 2017). The filtered SNP variants were imported into R and locus heterozygosity and G'_{st} (Hedrick, 2005) scores were computed using vcfR (Knaus & Grünwald, 2017). To account for effects near identical multilocus genotypes in population statistics calculations, the maximum likelihood (ML) distances from IQTree were used to identify multi-locus genotypes (MLGs) that are highly similar and collapse them using poppr. For each cluster identified by STRUCTURE, the Shannon, Simpson, and inverse Simpson indices (Hurlbert, 1971) were calculated, for both MLG collapsed and uncollapsed data using poppr. To test for linkage disequilibrium in population clusters, isolates from each population cluster were selected and a permutation test of $\bar{r}_d$ (a modification of I_A) (Agapow & Burt, 2001) with 999 permutations was run for polymorphic loci within those isolates using poppr. To test for any correlation of geographic distance with phylogenetic distance in the WA isolate population, a Mantel test with 999 replications was performed using ade4 comparing ML distances from IQTree with euclidean distances of GPS coordinates from Australian clone corrected isolates with known sampling locations and distinct collapsed multilocus genotypes. Patterns of variance were assessed using principal component analysis (PCA) of the filtered SNPs from all isolates with distinct collapsed MLGs using the adegenet package.

11.3.4 Genome assembly

Overlapping 150 bp paired end reads from the WA isolates were stitched using BBmerge version 38.38 (Bushnell et al., 2017) using the strict mode, kmer size of 62 bp, allowing assembly extension up to 50 bp from the ends of reads (rem mode), and using error correction to assist merging (ecct option). Genomes were assembled using Spades version 3.13.0 (Bankevich et al., 2012) (`--careful --cov-cutoff auto`). Different kmers were used depending on input read length (Supplementary table S1) (125 bp PE samples: 21,31,51,71,81,101; 150 bp PE samples 31,51,71,81,101,127; 201FG217 [125 bp PE]: 21,31,51,71). Mitochondrial genomes (mtDNA) were assembled using Novoplasty version 2.7.2 (Dierckxsens et al., 2017), using the Sn15 mitochondrial genome assembly [NCBI RefSeq: EU053989.1] (Hane et al., 2007) as a seed sequence, and with kmers between 31 and 81 (Supplementary table S5). The k-mer resulting in assemblies with the fewest number of contigs within an expected total assembly size of 47-52 Kb was manually selected, and designated the mtDNA sequence of that isolate. Code

for generating the mtDNA assemblies is available online at https://github.com/darcyabjones/mitoflow. Nuclear genome assemblies were then filtered for mtDNA sequences by aligning reads to assembled scaffolds using BBmap, and aligning mitochondrial scaffolds to assemblies using minimap2 git commit 371bc95 (Li, 2018). Genomic scaffolds were considered to be mitochondrial if the alignment coverage with mitochondrial contigs was greater than 95%, and the median read depth was in the top 0.8% overall read depth. Some manual assignment of very short contigs near the threshold cutoffs was undertaken.

Genome assembly quality control statistics were collected using Quast version 5.0.2 (Gurevich et al., 2013), bbtools stats version 38.38 (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/statistics-guide/), and KAT version 2.4.2 (Mapleson et al., 2017). Code for running post-assembly quality control and selection of mitochondrial scaffolds is available at https://github.com/darcyabjones/postasm (commit: c94c3b9).

11.3.5 Determining presence-absence variation relative to the SN15 reference isolate

Genome assemblies were aligned to the reference isolate SN15 [NCBI Assembly: GCA_016801405.1](Bertazzoni et al., 2021, chapter 10), and to long-read assemblies for isolates SN4 [NCBI Assembly: GCA_002267005.1], SN2000 [NCBI Assembly: GCA_002267045.1], Sn79 [NCBI Assembly: GCA_002267025.1] using nucmer version 4.0.0beta2 (`--maxmatch`) (Marçais et al., 2018). Alignments were converted to BED format, from which alignment coverage was computed (`bedtools genomecov -bga`) and combined (`bedtools unionbedg`) using BEDTools version 2.28.0 into a bedgraph file (Quinlan & Hall, 2010). Mean coverage in non-overlapping 50Kb windows of the genome visualised using the R package circlize (Gu et al., 2014).

Regions of presence-absence variation (PAV) were extracted from the bedgraph using a custom python script (available online at https://github.com/darcyabjones/mumflow/blob/master/bin/find_pavs.py), and coverage blocks were converted to simple presence-absence values (0 or 1) by collapsing coverage count >= 1 to 1. Adjacent blocks with identical PAV values in all isolates were merged, and non-overlapping PAV blocks were also merged where for a set of 3 adjacent blocks, the outer 2 blocks had identical PAV values in all isolates and the centre block was <= 50 bp. Code for running these steps is available at https://github.com/darcyabjones/mumflow.

11.3.6 Annotation of DNA repeats and non-protein coding gene features

Transposable elements (TEs) were predicted using a combination of tools: EAHelitron git commit c4c3dca (https://github.com/dontkme/EAHelitron), LTRharvest (Ellinghaus et al., 2008) and LTRdigest from genometools version 1.5.10 (Steinbiss et al., 2009), MiteFinder git commit 833754b (Hu et al., 2018), RepeatModeler version 1.0.11 (http://www.repeatmasker.org

/RepeatModeler), and RepeatMasker version 4.0.9p2 (http://www.repeatmasker.org) using the species "Parastagonospora nodorum". Putative TE protein coding regions in the genomes were identified using MMSeqs2 version 9-d36de (Steinegger & Söding, 2017), searching protein profiles from selected Pfam families (Supplementary table S6), GyDB families (Llorens et al., 2011), and a custom multiple sequence alignment (MSA) database based on protein collections in TransposonPSI (http://transposonpsi.sourceforge.net/) and LTR_retriever (Ou & Jiang, 2018) (available online at https://github.com/darcyabjones/pante/tree/master/data/proteins).

Predicted TE sequences from EAHelitron, MiteFinder, RepeatModeler, and MMSeqs protein finding were combined and clustered using VSEARCH version 2.14.1 (Rognes et al., 2016) requiring cluster members to have >=70% identity to the cluster seed sequence (`--cluster_fast combined.fasta --id 0.90 --weak_id 0.7 --iddef 0 --qmask dust`). Clusters were filtered based on frequency and conservation across the population, requiring presence in >= 4 distinct genomic locations to be considered present in a genome, and requiring the cluster to be present in >= 20% of the total population. Filtered clusters were aligned using DECIPHER version 2.10.0 (Wright, 2015), and classified into subtypes using RepeatClassifier (part of RepeatModeler), which was then used as a final customised library to map repeat locations for each assembly with a final round of RepeatMasker. Genes encoding rRNA and tRNA were predicted with RNAmmer version 1.2 (Lagesen et al., 2007) and tRNAscan-SE version 2.0.3 (T. M. Lowe & Chan, 2016), respectively. These TE and non-coding RNA predictions were used to "soft-mask" genomes using BEDTools (Quinlan & Hall, 2010). Code to run these steps is available at https://github.com/darcyabjones/pante/ (commit: 2de5d08).

11.3.7 Annotation of protein-coding genes

Proteins predicted from previous *P. nodorum* SN15 (Bertazzoni et al., 2021, chapter 10), SN4, and SN79 annotations (Richards et al., 2018) were aligned to each genome using Spaln version 2.3.3 (`spaln -KP -LS -M3 -O0 -Q7 -ya1 -yX -yL20 -XG20000`) (Iwata & Gotoh, 2012). Additionally all fungal proteins from the UniRef 50 database release 2019_08 (https://www.uniprot.org/uniref/, downloaded: 2019-10-29, query: 'taxonomy:"Fungi [4751]" AND identity:0.5') were aligned to genomes using Exonerate version 2.4.0 (`--querytype protein --targettype dna --model protein2genome --refine region --percent 70 --score 100 --geneseed 250 --bestn 2 --minintron 5 --maxintron 15000 --showtargetgff yes --showalignment no --showvulgar no`) (Slater & Birney, 2005) with pre-filtering by MMSeqs2 (`-e 0.00001 --min-length 10 --comp-bias-corr 1 --split-mode 1 --max-seqs 50 --mask 0 --orf-start-mode 1`) (Steinegger & Söding, 2017). Published RNAseq data from *P. nodorum* SN15 *in vitro* and 3 days post infection of wheat leaves (D. A. B. Jones et al., 2019) (available in NCBI GEO project: GSE150493; NCBI SRA accessions: SRX8337785, SRX8337784, SRX8337783, SRX8337782, SRX8337777, SRX8337776, SRX8337775, and SRX8337774) were assembled using Trinity v2.8.4 (`--jaccard_clip --SS_lib_type FR`) (Grabherr et al., 2011) and aligned to genomes using Spaln version 2.3.3 (`-LS -O0 -Q7 -S3 -yX -ya1 -Tphaenodo -yS -XG 20000 -yL20`) (Iwata & Gotoh, 2012), and GMAP version 2019-05-

12 (Wu & Watanabe, 2005). RNAseq reads were also aligned to all genomes using STAR version 2.7.0e (Dobin et al., 2013) and assembled into transcript annotations using StringTie version 1.3.6 (`--fr -m 150`) (Pertea et al., 2015).

Genes were initially predicted for each genome using multiple tools: PASA2 version 2.3.3 (`-T --MAX_INTRON_LENGTH 15000 --ALIGNERS blat --transcibed_is_aligned_oriented --TRANSDECODER --stringent_alignment_overlap 30.0`) (Haas et al., 2003), GeneMark-ET (`--soft_mask 100 --fungus`) (Lomsadze et al., 2014), CodingQuarry version 2.0 (including the "Pathogen Mode" with signal peptides predicted using SignalP version 5.0b (Armenteros, Tsirigos, et al., 2019)) (Testa et al., 2015), Augustus git commit 8b1b14a (independently for both forward and backward strands; `--hintsFile=hints.gff3 --strand=${STRAND} --allow_hinted_splicesites='gtag,gcag,atac,ctac' --softmasking=on --alternatives-from-evidence=true --min_intron_len=5`) (Stanke et al., 2008), and GeMoMa version 1.6.1 (transferring SN15 annotations only) (Keilwagen et al., 2018). Gene predictions using PASA2 used hints from assembled RNASeq transcripts aligned to the genomes with GMAP and BLAT. Augustus gene predictions used transcript alignments by GMAP, intron locations from STAR read alignments, and Spaln and protein alignments by Spaln as hints.

Pan-genomic gene sets may be prone to annotation errors in which orthologous loci are incorrectly annotated in some isolates, leading to false absences. To improve annotation consistency between isolates, protein predictions from PASA, Augustus, and CodingQuarry from all isolates were clustered using MMSeqs2 (90% identity and 98% reciprocal coverage), and annotations corresponding to proteins from representative members of the clusters were transferred to all isolates using GeMoMa as described earlier. Annotations and alignments from Genemark-ET, CodingQuarry, Augustus, PASA, both GeMoMa configurations, Exonerate, Spaln protein and transcript alignments, and GMAP alignments were combined using EVidenceModeler (git commit 73350ce) (`--min_intron_len 5`) (Haas et al., 2008). Because EVidenceModeler does not support prediction of non-standard splice sites or overlapping genes in different strands, Augustus (with all hints and the same parameters described earlier) was used to predict additional genes in regions of the genomes with hints that didn't overlap the EvidenceModeler predicted genes on the same strand. Protein predictions were searched against AntiFam (Eberhardt et al., 2012) using HMMER version 3.2.1 (`--cut_ga`) and matches to pseudogenes were removed. Genes within the merged Augustus and EVidenceModeler gene sets were marked as "low confidence" if supported only by Spaln or GMAP transcript alignments, Exonerate protein alignments, or transfers of annotations between isolates performed via GeMoMa (excluding the initial set of genes transferred from curated SN15 annotations). In SN15, genes only supported by the above tools or Augustus were also marked as low-confidence. "Low-confidence" genes that overlapped other genes on either strand by more than 30% of their length were removed. We corrected errors in the CDS coordinates where phases of gene annotations may lead to incorrect translations in some downstream pipelines by searching against all proteins from other isolates without stop codons, and all Pezizomycota proteins from UniRef-90 (filter: 'taxonomy: "Pezizomycotina [147538]" AND identity:0.9'; downloaded:

2020-05-13) using blastx version 2.10.0 (`-strand plus -max_intron_length 300 -evalue 1e-5`) (Camacho et al., 2009). Genes with an in-phase BLAST match lacking internal stop codons were fixed and retained, genes with an out-of-phase BLAST match with internal stop codons were marked as pseudogenes, and those with no BLAST match and internal stops were discarded. Genes overlapping predicted rRNA genes by more than 50% of their length were also discarded, and genes with exons overlapping assembly gaps were split into multiple fragmented genes, where each fragmented annotation was >= 60 bp.

Gene prediction completeness was evaluated for each isolate using BUSCO version 3 (git commit 1554283) versus the "pezizomycotina_odb9" dataset (Waterhouse et al., 2018), and additional statistics were collected by genometools version 1.5.10 (Gremme et al., 2013). The updated SN15 annotations were compared to previously published gene annotation versions (Bertazzoni et al., 2021, chapter 10) using ParsEval/AEGeAn version 0.15.0 (Standage & Brendel, 2012). The new SN15 annotations were identified using BEDTools, by excluding all new mRNA predictions overlapping original SN15 "A" and "B" mRNA annotations from Bertazzoni et al. (2021, chapter 10) on the same strand >= 20% by length (`bedtools subtract -a new -b old -s -A -F 0.2`), and were designated as the "C" gene annotations.

11.3.8 Orthology & positive selection

Orthology relationships for predicted proteins were predicted using Proteinortho version 6.0.30 (`-singles -selfblast`) (Lechner et al., 2011) using Diamond version 2.0.8 (Buchfink et al., 2021) as the search tool. Alternative protein isoforms present in the SN15 annotations were included in the orthology finding. Orthologous clusters were assigned identifiers prefixed by "SNOO" (summarised in supplementary table S10. Complete data available at https://doi.org/10.6084/m9.figshare.12966971.v3). Orthologous clusters with members in isolate SN15 were assigned the names matching the corresponding "SNOG" transcript names (Hane et al., 2007) for all loci with members in the orthogroup. For example "SNOO_434350AB" contains "SNOG_434350" isoforms A and B, and "SNOO_033200A149040A" contains both 'SNOG_033200A" and "SNOG_149040A". Clusters without members in SN15 were assigned sequential numbers starting from 50,000 and without isoform identifiers.

For each orthogroup a representative isoform was selected for differentially spliced genes by selecting the member with the closest length to the mean sequence length of the orthogroup, resulting in a single sequence per locus. These representative CDS sequences of predicted orthogroups were codon-aligned using DECIPHER version 2.16.1 (Wright, 2015), and gene trees were estimated using FastTree version 2.1.11 (Price et al., 2010). Orthogroup codon multiple sequence alignments and the gene trees were used to test for positive selection in the orthogroups using the BUSTED method in the HYPHY package version 2.5.15 (Murrell et al., 2015; Pond et al., 2005). A p-value threshold of 0.01 was used to determine positively selected orthogroups. Position specific positive selection tests were performed for the known effectors ToxA, Tox1, and Tox3 using the FUBAR method in the HYPHY package (Murrell et al., 2013; Pond et al., 2005).

For the purposes of PAV and pangenomic comparisons, to account for alternate isoforms present across multiple orthogroups, temporary "locus groups" were constructed by combining orthogroups that share common loci. Copy numbers of locus groups were calculated for each isolate as the number of distinct loci. Large regions of presence-absence variation were identified by hierarchically clustering the locus groups and samples by the locus group copy numbers using UPGMA clustering and the manhattan distance metric for orthogroups and 1 - Pearson's correlation coefficient as a distance metric for the isolates. Manually selected clusters of locus groups showing presence-absence patterns were aligned to the SN15 genome where possible using MashMap version 2.0 (`-s 500 --filter_mode none`) (Jain et al., 2018).

Orthogroups were designated as "core" if all isolates contained at least one member in the parent locus group, as "accessory" if more than one isolate but not all isolates had at least one member in the locus group, and as "singleton" if the locus group was only detected in a single isolate. Additionally, locus groups were designated as "multicopy" if any isolate had more than one member.

11.3.9 Functional analysis & effector candidate prediction

Predicted whole protein functions were found by searching the Swiss-Prot database version 2020_02 (Bairoch & Apweiler, 2000) using MMSeqs2 version 11-e1a1c (`--start-sens 3 -s 7.5 --sens-steps 3 -a`) (Steinegger & Söding, 2017). Matches were considered reliable for functional annotation if they covered >= 70% of both sequences, with >=60% sequence identity, and an e-value < 1e-10. Functional domains were predicted using InterProScan (P. Jones et al., 2014; Mitchell et al., 2019). Additionally, gene ontology (GO)-terms and predicted product names were predicted using the web-servers of PANNZER (Koskinen et al., 2015) and eggNOG-Mapper (Huerta-Cepas et al., 2017). GO-term predictions from InterProScan, PANNZER, and eggNOG-Mapper were combined and filtered to exclude terms in the GO "do_not_annotate anti-slim" set (available at: http://geneontology.org/docs/download-ontology, downloaded: 2020-05-15) to remove uninformative terms, forming the final GO-term set for the predicted proteomes.

Effector-like sequences were predicted using the Predector pipeline (https://github.com/ccdmb/predector, version: 0.1.0-alpha; chapter 3), which incorporates several software analyses including SignalP versions 3.0, 4.1g, 5.0b (Armenteros, Tsirigos, et al., 2019; Bendtsen et al., 2004; Petersen et al., 2011), DeepSig (Savojardo et al., 2018), TargetP version 2.0 (Armenteros, Salvatore, et al., 2019), DeepLoc version 1.0 (Armenteros et al., 2017), TMHMM version 2.0c (Krogh et al., 2001), Phobius version 1.01 (Käll et al., 2004), EffectorP versions 1 and 2 (Sperschneider, Dodds, Gardiner, et al., 2018; Sperschneider et al., 2016), ApoplastP version 1 (Sperschneider, Dodds, Singh, et al., 2018), LOCALIZER version 1 (Sperschneider et al., 2017), homology searches against dbCAN version 8 using HMMER version 3.3 (Mistry et al., 2013; H. Zhang et al., 2018), and sequence matches against PHI-base version 4.9 (Urban et al., 2020) using MMSeqs2 version 11.e1a1c (Steinegger & Söding, 2017).

Information from Predector, InterProScan, Pannzer, eggNOG-mapper, positive selection

and orthogroup analyses were combined into a single table (summarised in supplementary table S10. Complete data available at https://doi.org/10.6084/m9.figshare.12966971.v3).

Enriched and depleted GO terms were detected for core, accessory, and singleton orthogroups (and their multicopy subsets), accessory orthogroups contained in between 20% and 80% of isolates, orthogroups predicted to be under positive selection anywhere in the gene tree and with more than 20% of members, for orthogroups not found in SN15, manually selected clusters of orthogroups identified from hierarchical clustering (Supplementary table S9), *P. nodorum* Sn15 genes on accessory chromosome 23, and *P. nodorum* Sn15 genes with a ratio of RIP-like mutations over transitions within 1000 bp of the gene over the 95th percentile, using two-tailed hypergeometric tests implemented in the GOATOOLS version 1.0.6 (Klopfenstein et al., 2018). For each of these sets, two-tailed hypergeometric/Fisher's exact tests were also used to test for enrichment of genes lacking any GO term assignments, genes annotated as secreted by the Predector pipeline, and genes annotated as secreted and with a positive EffectorP 2 prediction using the SciPy Python package (Virtanen et al., 2020).

11.4 Results

11.4.1 Quality control of input sequence data

The majority of sequencing read pairs were assigned by Kraken2 to either *Parastagonospora nodorum* (average 74.1%) or were unclassified (average 25.6%) (Supplementary table S2). A small number of read pairs matched other organisms across kingdoms, but no fastq files had more than 1% of reads assigned to any non-fungal taxon nor were any consistent trends observed. Sequencing read quality was generally high, with only a single lane of sequencing (of three) for the isolate WAC13068 failing quality control (QC) because of low base quality (Supplementary data S1). Mean insert sizes for isolates sequenced with 125 bp reads ranged between 441 and 709 bp (mean 632 bp), with standard deviations between 263bp and 760bp. Insert sizes for isolates sequenced with 150bp reads varied between 275bp and 321bp (mean 305 bp) with standard deviations between 208 bp and 295 bp.

11.4.2 Prediction of mutations across the *P. nodorum* pan-genome relative to the SN15 reference isolate

Short variants (single nucleotide polymorphisms (SNPs), insertions, and deletions) were predicted from aligned short-read sequenced isolates, yielding 895,000 variant loci after filtering (Supplementary table S3, data 2-8) corresponding to 1 mutation for every 41 bases in the genome on average. The majority of variants were SNPs (830,761), compared to 33,943 and 30,296 insertions and deletions. The majority of SNPs were C↔T or G↔A mutations (296,688 and 296,927, respectively), with an overall transition to transversion ratio of 2.5. Relative to the genome of the SN15 reference isolate, 7.4% of variants were within exon features, 3236 mutations resulted in truncation (gain of stop), 660 mutations resulted in loss of a stop codon,

466 mutations resulting in loss of start codon, 249 and 310 mutations were in splice site acceptor or donor sites respectively, 9896 were frameshift variants, 148,914 were missense mutations, 205,602 were synonymous variants, and there were 1648 and 1591 disruptive in-frame deletions and insertions, respectively.

11.4.3 Phylogeny and structure of the local Western Australian *P. nodorum* population

A subset of SNPs for phylogenetic and population genetics analyses were selected (maximum missing genotypes of 30%, minimum non-major allele frequency of 0.05, and filtering correlated loci within 10 kb) resulting in 45,194 SNP loci from all illumina sequenced isolates.

A phylogenetic tree was estimated from these SNPs using IQTree (Minh et al., 2020) (Figure 11.1). The resulting tree grouped isolates from Western Australian (WA) and non-Australian isolates into distinct clades. Leaf lengths were generally long within the WA clade, however, six groups of isolates with very short branch lengths were observed, with sizes ranging from 3 to 14 isolates (Supplementary data S9, Supplementary figure S1). Some internal tree nodes had low UFBoot support values (< 95), indicating that some of the high level relationships were poorly resolved. A major split with high SH-aLRT but low UFBoot support was observed, which tended to separate isolates with long branch lengths from a diverse range of locations from a second clade comprised mostly of several highly similar clades which were generally collected from northern regions of the WA wheat growing area (Geraldton and Dandaragan). No obvious correspondence between effector haplotype profiles and high level phylogenetic clades was observed; but at lower level clades effector haplotype profiles appear to be conserved, particularly where member isolates have short branch lengths (Figure 11.1).

Analysis of population structure from SNP data with STRUCTURE (Figure 11.1; Supplementary table S4), predicted nine clusters, with the non-Australian isolates forming a single cluster (Structure cluster 8). Within the WA isolates, a single main population was observed (cluster 4) with 7 small satellite clusters. Few isolates were unambiguously assigned to the main cluster (21 of 102 had a posterior probability of > 0.85), with small but appreciable posterior probabilities contributed by satellite clusters and the international cluster. Noting this, we refer to isolates with the highest posterior probability of assignment to cluster 4 as members. These members generally had long branch lengths in the phylogenetic tree, and were found in multiple clades. This main cluster corresponds to the two main clusters (1 and 2) presented in Phan et al. (2020, chapter 8) (referred to hereafter as clusters P1-*5*) (Supplementary figure S2 and S3). The seven remaining clusters identified by STRUCTURE corresponded to clades in the phylogenetic tree where all leaves had very short lengths, indicating that all members were highly similar to each other. Cluster 1 consists of 3 isolates collected from Geraldton in 2012 (Supplementary figure S4 and S*5*). Cluster 2 consists of 8 isolates collected from Geraldton in 200*5* and 2011. Cluster 3 consists of 14 isolates collected from Geraldton in 200*5*, 2011, and 2012. Cluster *5* consists of *5* isolates collected from Geraldton in 200*5* and 2011. Cluster 6 consists of 7 isolates collected from Geraldton, South Perth, or WA (unknown specific location)

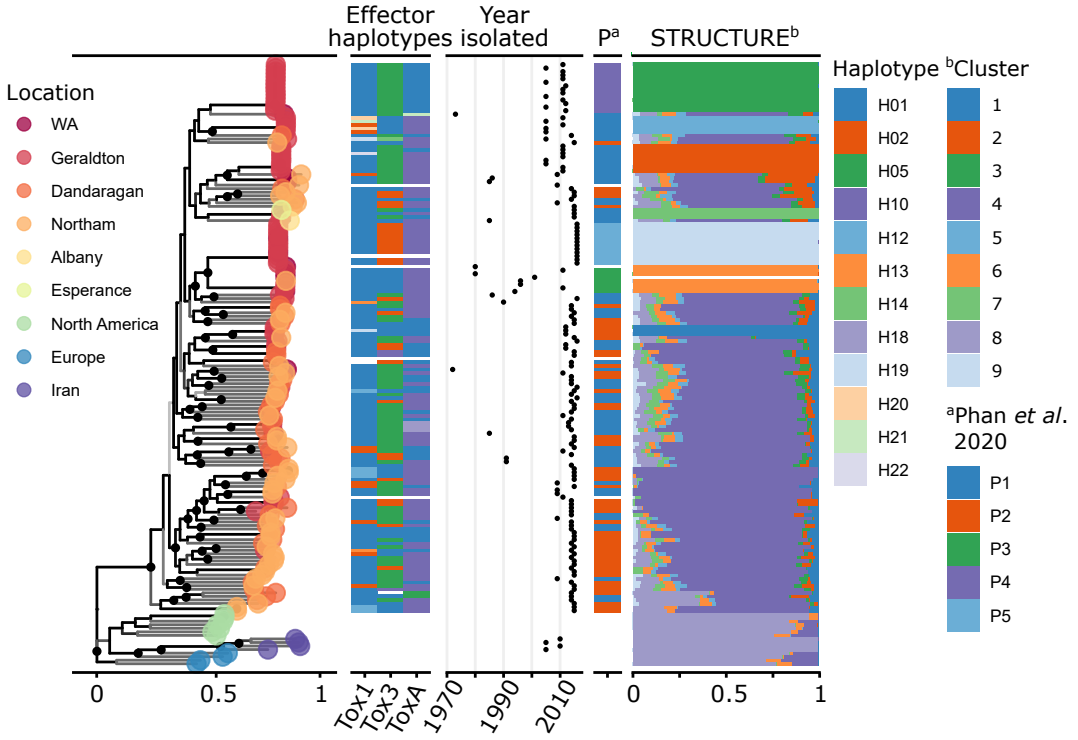


Figure 11.1: The structure and features of the Western Australian (WA) *Parastagonospora nodorum* population. The tree on the left shows the predicted phylogeny of WA and internationally-sampled *P. nodorum* isolates, with colours corresponding to sampling regions, summarised from supplementary figure S1. Dots on the tree indicate where clades have >= 95% UFBoot confidence, and clade shade indicates SH-aLRT scores, with black indicating high support. Effector haplotype profiles for three confirmed effector loci are shown based on data from Phan et al. (2020, chapter 8). A haplotype with white indicates that the isolate has not been haplotyped, or was unable to be amplified by PCR. On the right shows the results of population structure analysis, with colours indicating discrete population clusters, and the relative size of the bars indicating posterior probability of an isolate belonging to that cluster. The section indicated as "P" shows the clusters assigned to that isolate by the simple sequence repeats (SSR) study from Phan et al. (2020, chapter 8), where white indicates isolates that were not present in that study.

between the years 1980 and 2011. Cluster 9 consists of 12 isolates collected from Mingenew (near Geraldton) or WA in 2016. Clusters 3, 6, and 9 correspond to the satellite clusters P3, P4, and P5 presented by Phan et al. (2020, chapter 8), respectively. The remaining clusters further separate the main populations identified by Phan et al. (2020, chapter 9), with cluster 1 separating from P2, and clusters 7, 5, and 2 separating from P1. The Tox1 effector haplotypes of members in cluster 5 were highly variable and contained rare variants in the WA population.

Population diversity statistics indicated that individuals within clusters except the core WA cluster (4) and the international cluster (8) are highly similar (Table 11.1; Supplementary table S4). F_{ST} and average locus expected heterozygosity suggests that the alleles are nearly fixed in these clusters. The mean locus G'_{ST} (a normalised variant of F_{ST} that accounts for

multi-allelic markers) (Hedrick, 2005) was 0.44, indicating that there is some differentiation between subpopulations. To reduce the effects of nearly identical isolates on local diversity estimates, reduced multi-locus genotypes (MLGs) were defined using the IQTree maximum likelihood (ML) distance estimates using complete linkage clustering and a cutoff threshold of 0.1 (Table 11.1). All clusters other than 4 and 8 (the main WA population and the international isolates) were composed of a single MLG using this strategy. Cluster 4 comprised 89 MLGs (of 102 individuals) and cluster 8 comprised 14 MLGs (of 15 individuals). For each sampling location and year, a single individual of each MLG was selected for further comparison, referred to as the "clone-corrected" subset. A mantel test comparing the ML distance matrix with the distance matrix derived from the isolate sampling GPS coordinates indicated no significant correlation between genetic distance and geographic distance in the WA "clone-corrected" population (Mantel test, 999 replications, p-value = 0.381). Principal component analysis (PCA) of the clone corrected samples showed a clear separation of international isolates from those of WA in PC1 which explained 4.7% of the total variance, but no other principal components showed structure in the data correlated with sampling location or year (Supplementary figure S6 and S7).

Permutation tests of the $\bar{r}_d$ index of association indicated that all clusters except 1 and 7 were in linkage disequilibrium (999 replications, p-value < 0.05) (Table 11.1). Clusters 1 and 7 both contain only three members, so the test may be underpowered in those cases. Repeated tests for clusters 4 and 8 using the "clone corrected" subset were also significant, indicating that linkage disequilibrium was not associated with isolate clonality.

11.4.4 Comparative genomics across the local Western Australian *P. nodorum* population indicated telomeric or transposon-rich mutation 'hotspots'

Short variant mutation frequencies were observed to occur in "hotspots" throughout the SN15 genome, which were often, but not exclusively, telomeric (Figure 11.2, Supplementary table S8). The accessory chromosome 23 (AC23) was observed to have a higher overall SNP density compared to other chromosomes. Isolates WAC2813, WAC9178, WAC2810, WAC8635, WAC13405, WAC13418, WAC13447 (all from population cluster 6) all had very few SNPs relative to SN15, forming the inner blue circle present in Figure 11.2. Care was taken that certain biological and technical factors did not unduly influence our interpretation of SNP density at the whole-genome level. A region on chromosome 07 of between 140,979 bp and 623,833 bp, which was identified in a previous study (Bertazzoni et al., 2021, chapter 10) as a potential sequencing artifact exhibited an absence of SNPs in our analysis. Low SNP counts around the ribosomal DNA (rDNA) tandem repeat array located on the end of chromosome 3 (Bertazzoni et al., 2021, chapter 10) and other areas that overlap repeat regions may have also been caused by read-alignment depth filtering rather than the absence of variants. Repeat-induced point mutations (RIP) is an important feature indicating 'hypermutation' compartments throughout

Table 11.1: Population diversity statistics based on 45,194 single nucleotide polymorphism (SNP) variant loci (44,532 biallelic) from the *P. nodorum* population. Isolates are assigned to populations based on the STRUCTURE cluster with the highest posterior probability. No isolates had the same SNP profile, so multi-locus genotypes (MLGs) were defined as having a genetic distance less than 0.1. The column "# loci" indicates the number of loci within a subpopulation that were variable and had no missing genotypes. The mean locus Simpson's index (Mean λ. AKA expected heterozygosity) is calculated for every locus under analysis, for a biallelic SNP in a haploid organism the maximum possible value is 0.5. The Simpson's index (λ), Stoddart-Taylor's index (G), and Shannon's diversity (H) are calculated based on MLGs. The index of association (I_A) and it's normalised form $\bar{r}_d$ indicate linkage disequilibrium in each population cluster, and was calculated only on variant loci with no missing data (# loci) within each cluster. $\bar{r}_d$ p-values indicate the result of a permutation test (999 permutations) for higher $\bar{r}_d$ than would be expected by random allele distribution across isolates.

cluster	# isolates	# MLG	# loci	F_{ST}	Mean λ	λ	G	H	I_A	$\bar{r}_d$	$\bar{r}_d$ p-value
1	3	1	155	0.99	0.002	0	1	0	0.11	7e-3	0.325
2	8	1	149	0.99	0.002	0	1	0	3.77	0.03	0.001
3	14	1	156	0.99	0.003	0	1	0	0.58	0.004	0.026
4	102	89	10353	0.05	0.311	0.99	81.28	4.45	1.53	0.006	0.001
5	5	1	114	0.99	0.002	0	1	0	2.38	0.021	0.001
6	7	1	33	1.0	0.001	0	1	0	1.27	0.040	0.021
7	3	1	114	0.99	0.001	0	1	0	-0.03	-2e-3	0.363
8	15	14	6695	0.0	0.314	0.92	13.24	2.62	11.85	0.013	0.001
9	12	1	87	1.0	0.001	0	1	0	0.54	0.006	0.048
Total	169	110	45194		0.320						

fungal genomes (Hane et al., 2015). The ratio of RIP-like dinucleotide changes over all transition SNPs across the pan-genome relative to the SN15 reference genome indicated that RIP-like mutations are over-represented and also tend to be localised in hotspots (Figure 11.3). Hotspots of RIP-like mutation tended to co-locate with regions rich in transposable elements in the SN15 genome, though there are several regions that are enriched for RIP-like mutations for a small group of isolates.

11.4.5 Comparative genomics across the WA *P. nodorum* pan-genome

The average assembly size for the 156 WA *Parastagonospora nodorum* isolates was 37.8 Mb with a median L50 (N50 count) of 17, meaning that the 17 largest contigs contain at least 50% of the total genome size. The median N50 (N50 length) and NG50 lengths were 793431 and 783556 (Supplementary table S5), meaning that 50% of the assembly size was containined in contigs that are at least 783 kb long. Mitochondrial assemblies produced between 1 and 3 contigs in all cases, with a median size of 49591 bp. A common ~1000bp repeated region was observed in the mitochondrial assemblies, which appeared to be the cause of the fragmented assemblies (data not shown). The WA isolates were predicted to be highly complete with a median number of genes predicted by Genemark of 13037, with average completeness estimated via BUSCO at 98.94%, only one isolate (15FG111) had completeness estimated below 98%.

In comparison to the SN15 reference assembly, the majority of the SN15 genome sequence

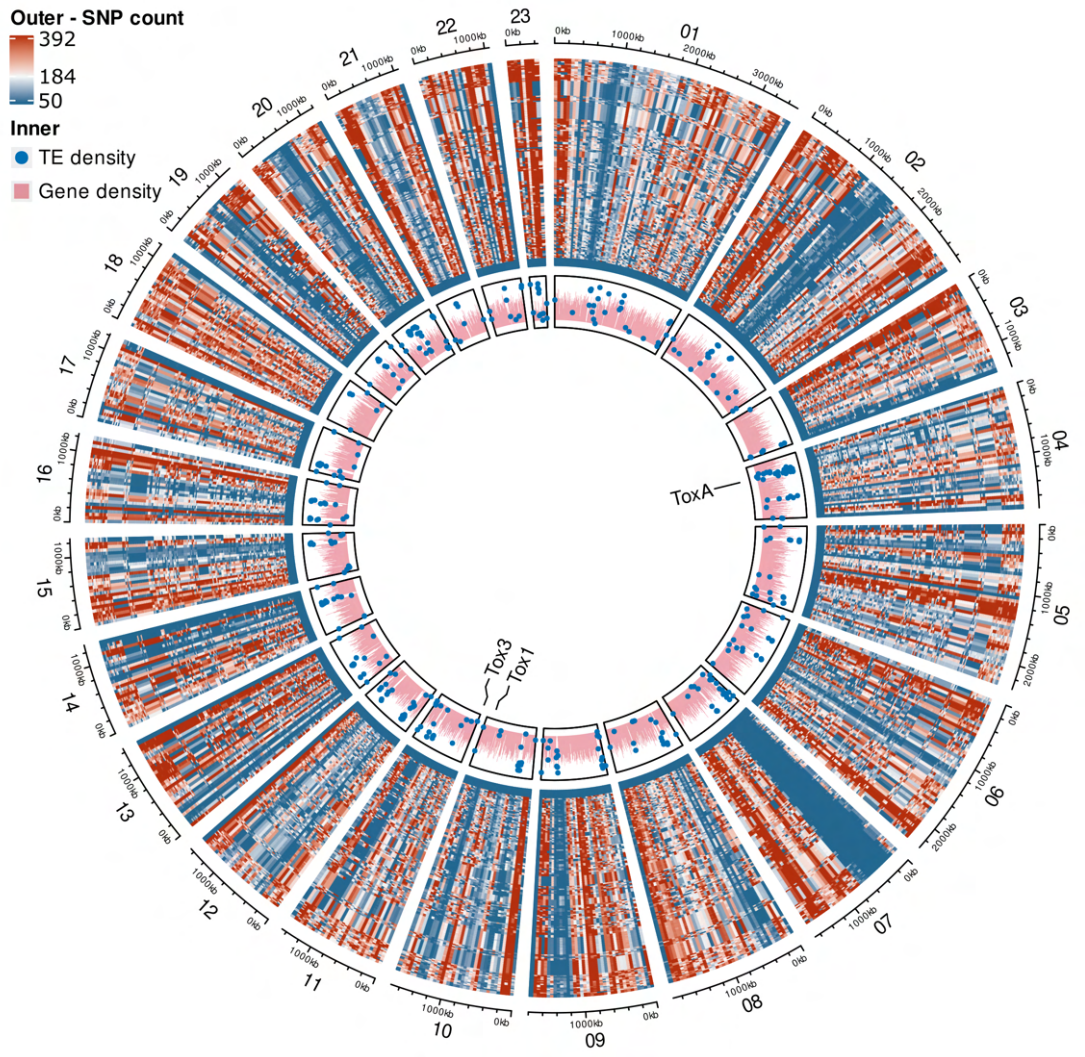


Figure 11.2: A circos plot showing SNP density over each of the 23 chromosomes in the SN15 genome assembly. The innermost track shows the proportion of bases covered by genes (CDS features, red) and transposable elements (TE, blue dots) in non-overlapping 10kb windows. For TEs, windows with TE base coverage less than 10% are not plotted. The heatmap shows SNP counts in 50 kb non-overlapping windows for each of the Western Australian isolates in the outer track (Supplementary table S8), with the colour scale boundaries set by the 10th, 50th, and 90th percentiles (50, 184, and 392, respectively).

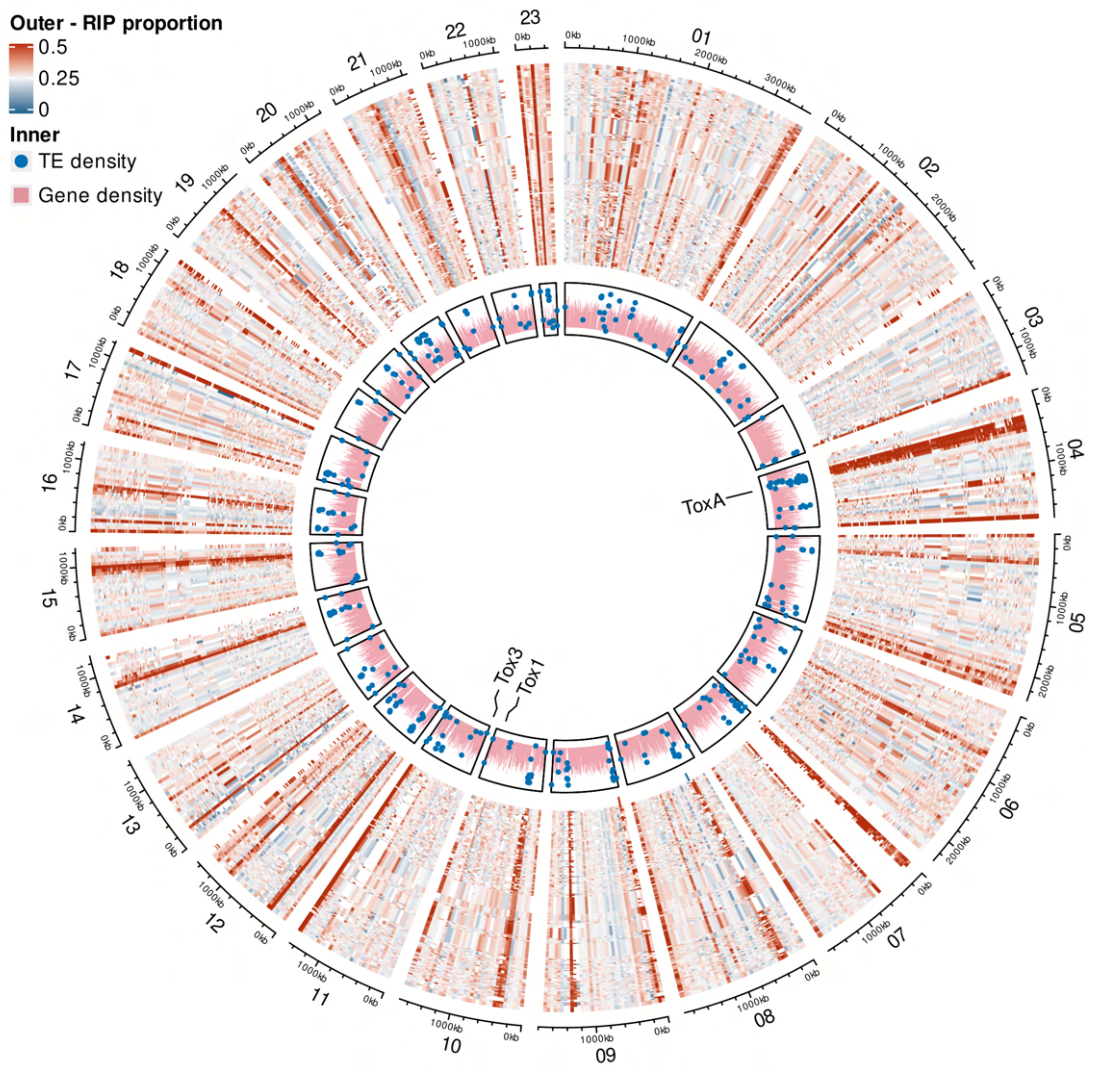


Figure 11.3: A circos plot showing the proportion of RIP-like (CA↔TA or TG↔TA) mutations over transition (C↔T or G↔A) mutations for each of the 23 chromosomes in the SN15 genome assembly. The innermost track shows the proportion of bases covered by genes (CDS features) and transposable elements (TE) in 10kb non-overlapping windows. Windows with TE base coverage less than 10% are not plotted. The heatmap in the outermost track shows the proportion of RIP-like mutations over the number of transition mutations in 50 kb non-overlapping windows for each isolate (Supplementary table S8). Windows with fewer than 20 SNPs are plotted in white to avoid high ratios caused by a small number of RIP-like mutations. By chance, 25% of transition mutations would be expected to be part of a RIP-like dinucleotide pair change.

was conserved with resequenced isolate assemblies, with small regions of presence-absence variation (PAV) tending to be observed at telomere ends or large internal repeat regions (Figure 11.4; Supplementary table S8). In addition to being missing in isolate SN79 (Richards et al., 2018), the accessory chromosome 23 was wholly absent in isolate 'Northam_Magenta' and partially absent in isolates 16FG160-162, 16FG163_2, 16FG164-171, and WAC13403 (Figure 11.4). We observed large duplications of genomic regions in some isolates (Figure 11.4). On SN15 chromosome 19 there was a large duplication of a region between 850 kb and 1 Mb in length in isolates WAC13404, WAC13075, and WAC13525. The first 200 kb was duplicated in isolate Meck8 relative to SN15 chromosome 22. Isolate WAC13631 had a large duplication relative to SN15 chromosome 12, of between 750 kb and 1.1 Mb.

Gene prediction resulted in a median gene count of 18294 across WA isolates, with a minimum of 17633 in isolate SN79 and a maximum of 19125 in isolate RSID03 (Supplementary table S6 and S7). The median length of coding domain sequences (CDSs) was 894 bp and was 53 bp for introns. The median BUSCO gene completeness was 3135 or 99.3%, with a median of 16 fragmented and 3 missing loci. Orthology clustering of predicted protein products from all *P. nodorum* isolates (including non-WA isolates) produced 34381 'orthogroups', 14098 of which were core to the population (13628 single copy and 470 multi-copy), and 11460 were dispensable (10043 single copy, 1417 multi-copy) (Supplementary table S9). An additional 8823 singleton groups were identified (8490 single copy, 333 multi-copy) which were only observed in a single isolate. To detect orthogroups with any members potentially under positive selection at any site, dN/dS branch site tests were run for all non-singleton orthogroups using the BUSTED algorithm in HyPhy (Pond et al., 2005). This identified 5306 orthogroups that were undergoing diversifying selection at any point with p-values < 0.01. Of these, 732 orthogroups had more than 20% of sequences within the orthogroup predicted to be subject to positive selection.

Multiple codon alignment of the coding sequences of the three known effector loci *ToxA*, *Tox1*, and *Tox3* (Supplementary data S10) indicated the occurrence of non-synonymous and RIP-like mutations in these loci. The *Tox1* orthogroup (SNOO_200780A) was absent in RSID36 RSID37, and RSID39 (Supplementary table S9) and was present as a single copy in all other genomes, though it had a truncated C-terminus in WAC13443 and SN79. Some branches of the orthogroup were predicted to be under positive selection (HyPhy BUSTED test, p-value = 0.0008), and three codons showed significant position specific positive selection (HyPhy FUBAR test, posterior probability > 0.90) at alignment codons 108 (GAC↔TCC; p=0.9194), 113 (ACC↔CCC; 0.9296), and 117 (CGA↔GTA↔CAA; 0.9892). The alignment variants at T113P and R117{V,Q} are restricted to WA isolates, while the variant at D108S is restricted to a subset of the international isolates, including SN4. Overall, there were 13 distinct AA and nucleotide CDS sequences. The *Tox3* orthogroup (SNOO_089810AB) was present in 160 isolates, but absent in SN79, SN2000, and 8 of the 16 remaining non-Australian isolates. The *Tox3* orthogroup was not predicted to be under positive selection (HyPhy BUSTED, p-value>0.05), and no specific positions were detected to be under positive selection using HyPhy FUBAR. Three distinct Tox3 codon sequences were observed resulting in two distinct AA sequences, with 28 WA

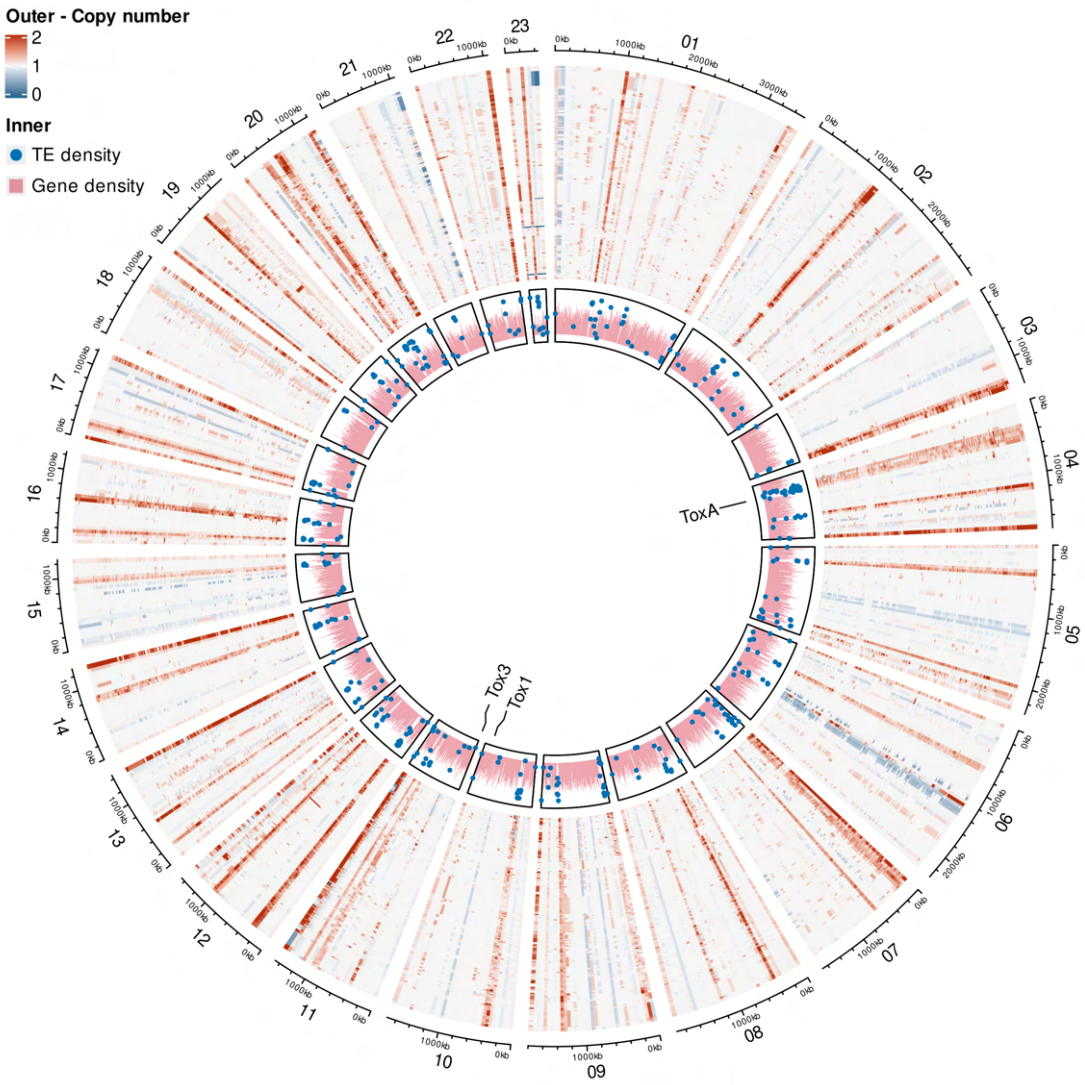


Figure 11.4: A circos plot showing each *Parastagonospora nodorum* genome assembly alignment coverage for each of the 23 chromosomes in *P. nodorum* SN15. The innermost track shows the proportion of bases covered by genes (CDS features) and transposable elements (TE) in 10kb non-overlapping windows. Windows with TE base coverage less than 10% are not plotted. The heatmap on the outside track shows average alignment coverage of each isolate genome assembly to SN15 in 50 kb non-overlapping windows (Supplementary table S8).

isolates (including all isolates from population cluster 9) and one Swedish isolate (RSID28) possessing 3 mutations resulting in the codon changes N78D (AAT↔GAT), R102L (CTA↔CGA), and D104E (GAA↔GAT). The A↔G mutation in codon 78 is next to a T nucleotide, which follows the canonical RIP CA↔TA dinucleotide change on the complementary strand. The *ToxA* orthogroup (SNOO_165710A) was present in most isolates but absent in one WA isolate (201FG209), and 11 international isolates, including SN79. The *ToxA* orthogroup was not predicted to be under positive selection by HyPhy BUSTED, and no individual sites were predicted to be under positive selection using HyPhy FUBAR. There were two mutations causing amino-acid change, I130V (ATT↔GTT) and E125D (GAA↔GAT). The I130V variant was restricted to WA isolates and the A↔G mutation is next to a T nucleotide, which follows a RIP-like dinucleotide change pattern. The E125D variant was present in seven isolates, of which 2 were from WA. Five distinct ToxA CDSs were observed, each with distinct AA sequences, however all of the predicted genes lacked a C-terminal region (which included the RGD motif) annotated in SN15, suggesting that gene predictions had missed an additional exon.

Effector prediction using Predector (https://github.com/ccdmb/predector; chapter 3) identified 779 effector candidates in the SN15 reference isolate which were predicted to be secreted and with positive EffectorP 2 scores, and 1348 effector candidates with Predector scores greater than zero, of which 132 were homologous to known fungal effectors or had virulence associated Pfam domains. Across the entire pangenome, 2055 orthogroups were predicted to be secreted and have a positive EffectorP 2 prediction, 3398 orthogroups had members with a Predector effector score greater than zero. Of the predector candidates 997 were predicted to be secreted and have a positive EffectorP 2 prediction, 145 contained effector homologues or virulence related Pfam domains, 411 had members significantly under positive selection, 55 were under positive selection in more than 20% of orthogroup sequences, 750 were accessory orthogroups, and 1405 were singleton orthogroups.

Orthogroups showed some PAVs spanning large regions of the genome (Figure 11.5; Supplementary table S9). The largest of these, indicated in Figure 11.5, is only present in 31 isolates and contains 385 orthogroups, which is comparable in gene number to AC23 which contains 218 predicted genes in SN15. Assembled scaffolds containing orthogroups in PAV group 1 were generally shorter than 200 kb and most did not align to any chromosomes in the SN15 genome (Supplementary data S11). These scaffolds did align to scaffolds in other isolates possessing the PAV group, with a high level of collinearity, indicating that these are contiguous regions in the genomes.

11.4.6 Accessory regions and candidate effector loci are enriched in RIP-like mutations and unknown functions across the WA pan-genome

Statistically significant enrichment (two-tailed hypergeometric test, BH FDR corrected p-value < 0.05) of predicted protein functions (by gene ontology (GO) terms) were observed within

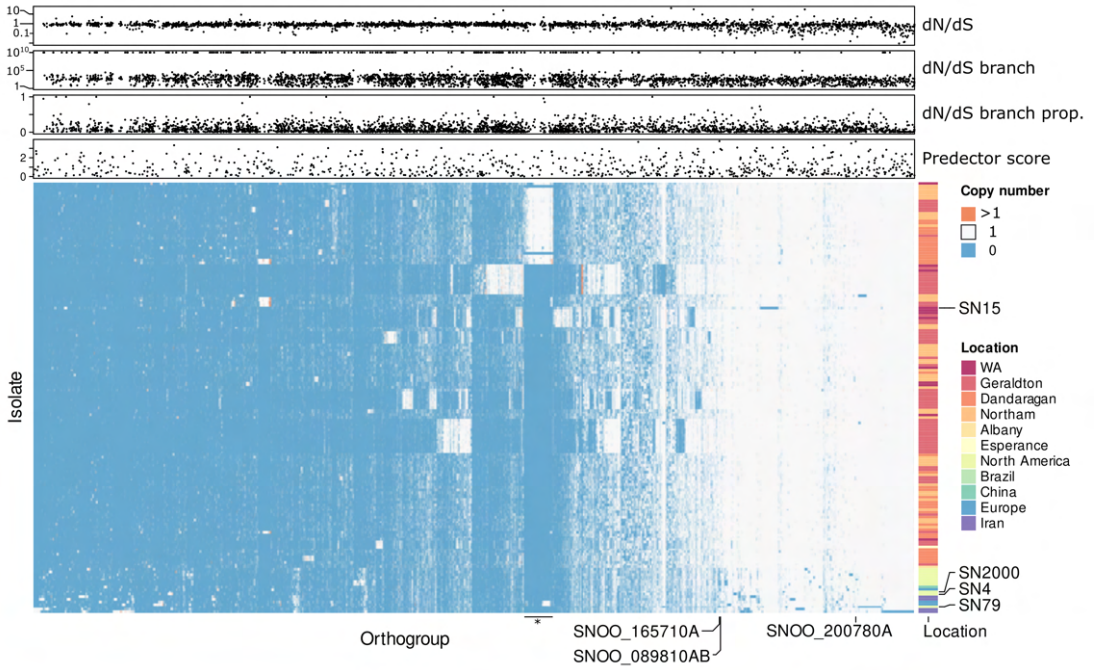


Figure 11.5: Dispensable and multi-copy orthogroups for each isolate in the *P. nodorum* pan-genome. "Orthogroups" is used as a general term to cover both orthologues and paralogues, which have not been separated here. Heatmap rows represent each *P. nodorum* isolate, and columns indicate each of the dispensable or multicopy orthogroups. Heatmap colour indicates the number of copies of an orthogroup each isolate has. Orthogroups absent (blue), present with single copy genes (white) and present with multicopy genes (orange) are shown. The columns of orthogroups containing ToxA (SNOO_165710A), Tox3 (SNOO_089810AB), and Tox1 (SNOO_200780A) are indicated. Locations that isolates were collected from are indicated on the right hand side colour bar. The rows corresponding to reference *P. nodorum* isolates are indicated. The top 3 scatter plots indicate orthogroups with any members with significant positive selection tests ($p < 0.01$). "dN/dS" indicates the overall dN/dS for the whole orthogroup. "dN/dS branch" indicates the dN/dS at the branch predicted to be under the highest selection. "dN/dS branch prop" indicates the proportion of sequences in the orthogroup predicted to be under positive selection. "Predector score" indicates where the highest scoring member of an orthogroup was greater than 0.

various subsets of pan-genome orthogroups. The core pan-genome was enriched for 3509 GO terms generally associated with core functions including biosynthesis, cell cycle control, and transport (Supplementary table S11). Conversely, the accessory pan-genome showed no enrichment for any GO terms, and was depleted in 1698 terms which were generally associated with core functions. Similarly, singleton orthogroups (present in only one isolate) were depleted in 1708 GO terms associated with core functions, as were orthogroups not present in the SN15 reference isolate (2946 depleted). Multicopy subsets of the core, accessory, and singleton pan-genomes, and the subset of the accessory pan-genome containing between 20% and 80% of isolates showed similar patterns of enrichment and depletion as their respective supersets. Positively selected orthogroups were not significantly enriched or depleted in any GO terms, but when restricted to orthogroups where more than 20% of sequences were predicted to be subject to positive selection were depleted in 139 GO terms with core functions (e.g. transport, metabolic processes, response to stimulus).

Genes of the SN15 reference isolate that were RIP-affected (above the RIP ratio 95th percentile, 0.432) were depleted in 54 core GO terms. AC23 contained two of the four genes in SN15 predicted to be involved in dehydroaustinol biosynthesis, but otherwise GO terms were all depleted. The PAV group was not enriched for any GO terms, and was depleted in 32 terms with general core functions. Overall, the enrichment tests above did not reveal any clear associations between GO terms and features within the genomic landscape. The majority of orthogroups (28767 of 34381) had no GO terms assigned. A complementary series of enrichment tests for a lack of GO terms ('no-GO') were performed using fishers exact tests with an uncorrected p-value threshold of 0.05. The core pan-genome was depleted for no-GO orthogroups, whereas the accessory and singleton pan-genome, positively selected orthogroups with more than 20% of sequences under selection, and the PAV group were all significantly enriched in no-GO orthogroups. RIP-affected orthogroups and SN15 loci on AC23 were also enriched for no-GO.

Enrichments tests were also performed for predicted secreted proteins and effector-like proteins (comprising secretion and EffectorP 2). The accessory pan-genome, singleton pan-genome, orthogroups absent in SN15, and positively selected orthogroups were enriched for secreted proteins. However the core pangenome and PAV group were depleted in secreted proteins. Similarly, the accessory pan-genome and positively selected orthogroups were enriched in effector-like orthogroups, while the core and singleton pan-genomes were depleted.

To find effector candidates a representative member of each orthogroup for each distinct locus was selected (Supplementary table S10). For orthogroups with members in the reference isolate SN15, all distinct loci were included selecting the protein isoform with the closest sequence length to the average orthogroup length. Representative members of other orthogroups were selected by taking the member with the closest sequence length to the average orthogroup length, then by the highest predector score. Orthogroups with a Predector score greater than zero, or with a signal peptide predicted by any method and an EffectorP 2 score greater than 0.5 were selected to be effector candidates. This identified 3579 candi-

date orthogroups, of which 788 and 1504 were in the accessory and singleton pangenomes, respectively. The WA isolates contained 1362 candidates (181 accessory) that were not predicted in the international isolates, of which 411 were restricted to the non-core populations (not population cluster 4; 64 accessory). The core WA population (cluster 4) possessed 842 distinct candidates not present in other clusters (64 accessory), while clusters 6 (which includes SN15) and 8 (international) possessed 96 (1 accessory) and 375 (52 accessory) unique candidates. The reference isolates SN15, SN4, SN2000, and SN79 were missing 1732 candidate orthogroups (317 accessory). There were 66 candidate orthogroups predicted to be under positive selection in at least 20% of orthogroup members, including a Tox3 homologue (SNOO_010970A). The large cluster of orthogroups with PAV identified in Figure 11.5 contained 18 candidate orthogroups.

From these 3579 candidates, 1809 orthogroups with known functions or similarity to known effectors were selected (Table 11.2; Supplementary table S10). Five Tox1 (SNOO_200780A, SNOO_304660A, SNOO_423420AB, SNOO_436740A, SNOO_531030), four Tox3 (SNOO_089810AB, SNOO_010970A, SNOO_438650A, SNOO_503200), and one ToxA (SNOO_165710A) homologous orthogroups were identified, including each of the original sequences from the SN15 isolate. Numerous other effector homologues were found, including 19 MoCDIP4 (Chen et al., 2012), 7 XYLA (Pollet et al., 2009; Sperschneider et al., 2015), 7 CfTom1 (Ökmen et al., 2013; Pareja-Jaime et al., 2008), 2 FGL1 (Voigt et al., 2005), 2 AVR-Pita/AVR-Pita2 (Chuma et al., 2011; Dai et al., 2010), 2 Zt6 (Kettles et al., 2018), 2 HCE2/Ecp2 (Stergiopoulos et al., 2012), 2 NEP/NLP (Oome et al., 2014), 2 BEC2 (Schmidt et al., 2014) and 5 other CFEM domain containing proteins, 2 Cgfl (Fungalysin peptidase) (Sanz-Martín et al., 2016), and one each of BEC1019 (Y. Zhang et al., 2019), NIS1 (Irieda et al., 2019; Yoshino et al., 2012), MoBas2 (Mosquera et al., 2009), MoCDIP1 (Chen et al., 2012), MoMSP1 (Wang et al., 2016), MoSPD5/MoBas4 (Mosquera et al., 2009; Sharpee et al., 2017), PevD1 (Bu et al., 2014), and ZtNIP2 (M'Barek et al., 2015). Other notable functions and families identified among these effector candidates include Peptidases, Nucleases, Cupredoxins, CAP-superfamily proteins, Egh16-like virulence factors, Osmotin/Thaumatin-like proteins, Killer toxin KP4, tuberculosis necrotizing toxin, SnoaL-like/NTF2-like domain superfamily, RmlC-like cupin domain superfamily, TolB-like/major royal jelly protein, Ubiquitin and biotin related functions, and WD40/Ankyrin/Kelch repeat-containing proteins. A single UstYa-like mycotoxin biosynthesis protein, and several proteins related to metabolite biosynthesis or detoxification were also found.

Table 11.2: Selected effector candidate orthogroups with functional annotations in the *P. nodorum* pangenome. Of the 1908 effector candidates with functional predictions, this table includes all candidates with effector homologues and a Predector score > 1, the 5 top ranking (by Predector) orthogroups that were under positive selection, the top 5 candidates specific to any populations identified by structure, and the top 5 candidates restricted to Western Australian or non-core (satellite) WA population clusters. Population cluster 8 includes the international isolates, and cluster 4 represents the core WA cluster. For each candidate orthogroup SN15 members were selected as representatives, or in the case of orthogroups absent in SN15 the member with the highest Predector score. Membership of PAV clusters shown in Figure 11.5 is indicated by the corresponding number, where applicable. Products in bold show where a candidate matched a known necrotrophic or avirulence effector. All orthogroup candidates presented here have a signal peptide predicted by at least one method, and none any predicted transmembrane (TM) domains.

orthogroup	Pangenome[1]	Refence isolates[2]				Population cluster[3]									Product	dN/dS[4]	Predector	EffectorP1	EffectorP2	ApoplastP	LOCALIZER[5]	Length	# Cysteine
		SN15	SN79	SN4	SN2000	1	2	3	4	5	6	7	8	9									
SNOO_047630A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	chitin binding domain-containing protein	-	3.94	1.00	0.98	0.88	-	80	10
SNOO_616900	S	-	-	-	-	0	0	0	0	0	0	0	1	0	aldose epimerase	-	3.79	0.96	0.93	0.80	-	202	0
SNOO_165710A	A	Y	Y	Y	-	3	8	14	101	5	8	3	7	12	**ToxA**	-	3.76	1.00	0.95	0.48	C	178	2
SNOO_200780A	A	Y	Y	Y	Y	3	8	14	102	5	8	3	15	12	**ToxI**	-	3.73	1.00	0.97	0.64	-	117	16

Continued on next page

[1](C)ore, (A)ccessory, or (S)ingleton pangenome based on whether an orthogroup was present in all, some, or one isolate

[2]Presence or absence of the orthogroup in the four reference isolates

[3]Numbers of isolates with an orthogroup copy in each of the population clusters. The reference isolates were not included in the STRUCTURE analysis, so SN15 was assigned to cluster 6 where it was assigned by Phan et al. (2020), and the remaining reference isolates were assigned to cluster 8 with the international isolates

[4]tests for positive selection in orthogroups were significant ($p<0.01$, number members in orthogroup > 0.1)

[5]N=nuclear localized, C=chloroplast localized

Continued from previous page

orthogroup	Pangenome[1]	Refence isolates[2]				Population cluster[3]									Product	dN/dS[4]	Predector	EffectorP1	EffectorP2	ApoplastP	LOCALIZER[5]	Length	# Cysteine
		SN15	SN79	SN4	SN2000	1	2	3	4	5	6	7	8	9									
SNOO_576590	S	-	-	-	-	0	0	0	0	0	0	0	2	0	Gamma-crystallin-like family protein	-	3.32	1.00	0.94	0.84	-	118	5
SNOO_423420AB	A	Y	-	-	-	0	1	14	54	5	8	3	4	0	Tox1-like protein	-	3.08	1.00	0.95	0.72	-	124	16
SNOO_109810A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoCDIP4-like glucanase	-	3.04	0.83	0.82	0.73	-	233	2
SNOO_638140	S	-	-	-	-	0	0	0	0	0	0	0	1	0	glycoside hydrolase	-	2.93	0.99	0.85	0.76	-	110	1
SNOO_107590A	A	Y	Y	Y	Y	3	8	14	102	5	8	3	17	12	FgXYLA-like xylanase	-	2.93	0.44	0.82	0.80	-	241	2
SNOO_160630A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	Zt6-like endoribonuclease	-	2.92	0.99	0.88	0.56	-	174	4
SNOO_522100	A	-	-	-	-	3	0	1	26	0	0	0	0	0	RWD-like superfamily protein	-	2.88	1.00	0.78	0.26	-	108	1
SNOO_616750	S	-	-	-	-	0	0	0	0	0	0	0	1	0	glycoside hydrolase	-	2.82	0.93	0.83	0.86	-	159	2
SNOO_304660A	C	Y	Y	Y	Y	3	8	15	102	5	8	3	18	12	Tox1-like protein	-	2.81	1.00	0.93	0.50	-	84	8
SNOO_089810AB	A	Y	Y	-	-	3	8	14	99	5	8	3	8	12	**Tox3**	-	2.78	0.87	0.89	0.27	-	230	6
SNOO_568060	A	-	-	-	-	0	0	0	0	0	0	0	2	0	arabinofuranosidase	-	2.71	0.99	0.88	0.70	-	168	4
SNOO_531830	A	-	-	-	-	0	0	0	15	0	0	0	0	0	RWD-like superfamily protein	-	2.71	1.00	0.74	0.20	-	108	1
SNOO_128450A	A	Y	Y	Y	Y	3	8	14	101	5	8	3	17	12	FgXYLA-like xylanase	-	2.63	0.04	0.79	0.88	-	230	0
SNOO_137220A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoMSP1-like cerato-platanin	-	2.58	0.65	0.69	0.91	-	136	4
SNOO_121270A	C	Y	Y	Y	Y	3	8	15	102	5	8	3	18	12	MoCDIP4-like glucanase	-	2.56	0.20	0.68	0.78	-	232	2
SNOO_096500A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	FgXYLA-like xylanase	-	2.54	0.06	0.64	0.87	-	221	2

Continued on next page

Continued from previous page

orthogroup	Pangenome[1]	Refence isolates[2]				Population cluster[3]									Product	dN/dS[4]	Predector	EffectorP1	EffectorP2	ApoplastP	LOCALIZER[5]	Length	# Cysteine
		SN15	SN79	SN4	SN2000	1	2	3	4	5	6	7	8	9									
SNOO_427940A436170A	A	Y	-	-	-	3	0	13	51	5	17	0	0	12	RWD-like superfamily protein	Y	2.54	1.00	0.69	0.18	-	107	1
SNOO_120370A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoCDIP4-like glucanase	-	2.50	0.29	0.73	0.82	-	246	4
SNOO_436740A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	Tox1-like protein	-	2.48	1.00	0.94	0.76	-	79	8
SNOO_138880A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoBas5/MoSPD5-like protein	-	2.44	1.00	0.91	0.36	-	106	6
SNOO_531030	A	-	-	-	-	0	1	2	8	0	0	0	3	0	Tox1-like protein	-	2.43	1.00	0.89	0.33	N	128	7
SNOO_527620	A	-	-	-	-	0	0	0	12	5	0	0	0	0	Major royal jelly family protein	-	2.38	0.90	0.78	0.65	-	213	0
SNOO_573060	A	-	-	-	-	0	0	0	2	0	0	0	0	0	outer membrane enzyme PagP beta-barrel family protein	-	2.33	0.87	0.76	0.78	-	236	0
SNOO_001310A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoCDIP4-like glucanase	-	2.33	0.30	0.60	0.89	-	229	4
SNOO_149380A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoCDIP4-like glucanase	-	2.33	0.40	0.67	0.89	C	220	4
SNOO_108620A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	CfTom1-like xylanase	-	2.29	0.01	0.62	0.63	-	332	4
SNOO_152700A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	FgXYLA-like xylanase	-	2.26	0.76	0.78	0.75	-	231	0
SNOO_010970A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	Tox3-like protein	Y	2.23	0.52	0.64	0.25	C	224	6
SNOO_086040A	A	Y	Y	Y	Y	3	8	14	102	5	8	3	17	12	FgXYLA-like xylanase	-	2.15	0.25	0.71	0.77	-	256	1
SNOO_019220A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoCDIP4-like glucanase	-	2.10	0.21	0.56	0.80	-	255	2

Continued on next page

Continued from previous page

orthogroup	Pangenome[1]	Refence isolates[2]				Population cluster[3]									Product	dN/dS[4]	Predector	EffectorP1	EffectorP2	ApoplastP	LOCALIZER[5]	Length	# Cysteine
		SN15	SN79	SN4	SN2000	1	2	3	4	5	6	7	8	9									
SNOO_002000A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	VdPevD1-like Alternaria alternata allergen 1 family protein	-	1.85	0.71	0.56	0.87	-	167	6
SNOO_575600	S	-	-	-	-	0	2	0	0	0	0	0	0	0	acetylxylan esterase	-	1.83	0.02	0.62	0.74	-	302	4
SNOO_160810A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoCDIP4-like glucanase	-	1.83	0.06	0.58	0.78	-	238	5
SNOO_098820A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoCDIP4-like pectin lyase	-	1.82	0.01	0.57	0.91	C	354	9
SNOO_008150A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoCDIP4-like glucanase	-	1.77	0.00	0.24	0.83	-	230	2
SNOO_026870A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoCDIP4-like glucanase	-	1.77	0.06	0.33	0.80	-	222	2
SNOO_423720A	A	Y	Y	Y	Y	3	8	14	102	5	8	3	16	12	LysM domain superfamily protein	Y	1.75	0.95	0.67	0.58	-	113	5
SNOO_108650A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	FgXYLA-like xylanase	-	1.68	0.04	0.56	0.76	-	230	2
SNOO_065020A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	FGL1-like lipase	-	1.63	0.00	0.36	0.80	-	304	6
SNOO_503200	A	-	-	Y	-	3	8	14	102	5	0	3	16	12	Tox3-like protein	-	1.63	1.00	0.71	0.21	-	161	4
SNOO_510220	A	-	-	-	-	0	8	14	33	5	0	3	7	0	acetylcholinesterase/ carboxylesterase	Y	1.63	0.81	0.67	0.23	-	196	5
SNOO_159340A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	CfTom1-like xylanase	-	1.56	0.09	0.55	0.76	-	320	2
SNOO_140400A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	FgXYLA-like xylanase	-	1.54	0.00	0.24	0.86	-	227	2
SNOO_100710A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoCDIP4-like glucanase	-	1.50	0.07	0.45	0.87	-	249	4
SNOO_123500A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoBas2-like protein	-	1.49	0.88	0.24	0.85	-	109	6
SNOO_035930A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	CfTom1-like xylanase	-	1.31	0.00	0.50	0.75	-	356	4

Continued on next page

Continued from previous page

orthogroup	Pangenome[1]	Refence isolates[2]				Population cluster[3]									Product	dN/dS[4]	Predector	EffectorP1	EffectorP2	ApoplastP	LOCALIZER[5]	Length	# Cysteine
		SN15	SN79	SN4	SN2000	1	2	3	4	5	6	7	8	9									
SNOO_102410A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	Zt6-like endoribonuclease	-	1.27	0.08	0.45	0.72	-	138	4
SNOO_098950A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	NEP-like protein	-	1.25	0.02	0.37	0.76	-	238	2
SNOO_069800A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	CfTom1-like xylanase	-	1.10	0.00	0.10	0.89	-	393	10
SNOO_550640	A	-	-	-	-	0	0	0	5	0	0	0	0	0	RWD-like superfamily protein	Y	1.09	1.00	0.57	0.44	-	71	0
SNOO_059190A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoCDIP4-like glucanase	-	1.05	0.20	0.53	0.48	-	254	2
SNOO_057170AB	S	Y	-	-	-	0	0	0	0	0	1	0	0	0	PLAC8 motif-containing protein	-	1.04	1.00	0.91	0.38	-	91	9
SNOO_105220A	A	Y	Y	Y	Y	3	8	14	102	5	8	3	17	12	AVR-Pita2-like peptidase	-	1.03	0.00	0.16	0.85	-	350	8
SNOO_106950A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	Cgfl-like metalloprotease	-	0.94	0.00	0.12	0.81	-	630	4
SNOO_054410A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoCDIP4-like glucanase	-	0.93	0.00	0.23	0.77	-	333	5
SNOO_008890A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoCDIP4-like glucanase	-	0.90	0.00	0.17	0.82	-	230	2
SNOO_006170A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	FGL1-like lipase	-	0.88	0.02	0.52	0.71	-	318	6
SNOO_125540A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	ZtNIP2-like protein	-	0.88	0.87	0.47	0.37	-	167	4
SNOO_438650A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	Tox3-like protein	-	0.87	0.83	0.45	0.48	-	180	5
SNOO_127270A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	MoCDIP4-like glucanase	-	0.86	0.00	0.06	0.86	-	324	10
SNOO_117410A	C	Y	Y	Y	Y	3	8	14	102	5	8	3	18	12	CfTom1-like xylanase	-	0.86	0.00	0.31	0.16	-	354	5
SNOO_544730	A	-	-	-	-	0	0	0	0	0	0	0	0	6	RWD-like superfamily protein	-	0.84	1.00	0.68	0.21	-	52	0

11.5 Discussion

The transition from the sequencing of a single or few reference genomes to larger populations has broadened the scope of comparative genomics of plant pathogens. This includes the identification of: additional accessory genome content missing from the reference isolate (Badet & Croll, 2020); spatial distribution of virulence loci, and; region-specific selection pressures (Richards et al., 2019). To this end, we investigated the interplay between the population structure and genomic features relevant to plant pathogenicity, in a Western Australian (WA) population of *Parastagonospora nodorum*.

11.5.1 Population structure

Previously we have analysed the WA *P. nodorum* population using short sequence repeat (SSR) markers (Phan et al., 2020, chapter 8), and identified five distinct clusters of isolates. Two main clusters were proposed to represent a gradual change over time in response to wheat cultivar use, while the three remaining clusters were highly similar and were proposed to be clonally expanded populations. Interestingly, although the two core populations had a 1:1 ratio of mating type loci, the core population was observed to be in linkage disequilibrium, suggesting a predominantly asexually reproducing population. In contrast, this study indicated that the WA *P. nodorum* population is dominated by a single diverse main population, with seven satellite clusters that are highly similar. The fact that the core population was not split into two as previously observed may be explained by the marker type and number, the number of clusters selected to find, and the clustering method employed, where STRUCTURE (used in this study) explicitly models gene flow. The relative uncertainty of individual assignment to cluster 4 (the main population), suggests that there is some latent structure in the data, but this appears not to correspond to distinct reproductively isolated populations. In addition to the three satellite clusters identified by Phan et al. (2020, chapter 8), the greater resolution gained by the use of many single nucleotide polymorphisms (SNPs) identified an additional 4 minor clusters, which were subdivided from the two previously identified core clusters (Supplementary figure S3). The relatively low diversity observed in these seven clusters suggests that these are very recent expansions or clonal subpopulations. Analysis of the index of association indicated that all but two small sub-populations are in linkage disequilibrium, suggesting that limited sexual reproduction is occurring in populations. It should be noted that the two exceptional population clusters (1 and 7) contained few isolates and their lack of linkage disequilibrium may be inaccurate. Linkage disequilibrium across WA *P. nodorum* sub-populations was also reported by Phan et al. (2020, chapter 8), but they observed that clusters P1 and P2 (corresponding to cluster 4 in this study) had a 1:1 mating type ratio which indicates the potential for sexual reproduction, and that cluster P5 (corresponding to cluster 9 in this study) was in linkage equilibrium. The finding that the international cluster is in linkage disequilibrium is also unexpected as it would imply that the global *P. nodorum* population is largely asexual, in conflict with numerous previous reports (Caten & Newton, 2000; Keller et al., 1997; Murphy

et al., 2000; Sommerhalder et al., 2006; Stukenbrock et al., 2006). Although the $\bar{r}_d$ values are significantly different from a random background, they are still relatively low so the population may be exhibiting a mixture of clonal and non-clonal reproduction. Alternatively, the permutation method of $\bar{r}_d$ may not be appropriate for samples with large numbers of SNPs in relatively high density (compared to low-throughput marker studies). Phan et al. (2020, chapter 8) suggested that the non-core populations may be hybrids of local and internationally introduced *P. nodorum* isolates. This is not supported by the phylogenetic tree or population structure analysis reported in this study; however, STRUCTURE analysis indicated a cluster of six isolates from population 4 collected from the Northam and Dandaragan regions which had a high posterior probability of assignment to the international subpopulation, suggesting possible historic exchange of genetic material.

In the overall WA population there was no correlation observed between sampling location and genetic distance when nearly clonal isolates were excluded, suggesting that there are no geographic barriers to migration. Similarly, principal component analysis (PCA) only indicated an axis of variance separating international isolates from those from WA, with no other principal components showing association with sampling location or year (Supplementary figure S6 and S7). Long range wind dispersal of sexual ascospores has long been known to occur in the WA *P. nodorum* population (Bathgate & Loughman, 2001), though dispersal by infected seed is often reported internationally (Bennett et al., 2005; Cunfer, 1978, 1998). Both of these mechanisms may explain the lack of geography-dependent population structure observed in this study. Although seed borne dispersal is more likely if the population is mostly asexual, we were unable to find publications describing seed borne epidemics of *P. nodorum* in Australia.

The majority of isolates from the satellite clusters were collected from northern regions of the WA sampling area, predominantly Geraldton and Mingenew. Within the sampling zones presented in this study, the average rainfall in the south west regions is typically higher than in the dryer northern regions (Bureau of Meteorology, http://www.bom.gov.au/climate/current/annual/wa/summary.shtml, retrieved 2020-09-19). In addition to splash dispersal of secondary inoculum and a general positive correlation of rainfall with *P. nodorum* disease load (Shaw et al., 2008; Solomon et al., 2006), rain impacts can indirectly enable long-distance air travel (Kim et al., 2019), and may have contributed to the increased diversity of south-western regions. Similarly, high temperature during harvest time has been observed to be negatively correlated with *P. nodorum* disease load (Shaw et al., 2008), which may favour stronger populations in the southern regions. Numerous environmental factors can influence the lifecycle of *P. nodorum* which may explain these northern clonal sub-populations. However, the samples used in this study were not collected with population genetics analyses in mind, and an intentionally designed experiment may yet reveal the existence of similar structure in the other regions of the WA population. It appears that there is variance in the WA population, but this remains cryptic and is not explained by barriers to gene flow.

The presence of necrotrophic effector loci, or in some cases specific allele variants, is a direct determinant of crop disease outcomes in combination with the corresponding host

sensitivity loci (Vleeshouwers & Oliver, 2014). A US-based pan-genome study previously indicated alternate sets of candidate effector loci between two major *P. nodorum* sub-populations (Richards et al., 2019), highlighting the importance of region-specific genomic analysis and refinement of effector predictions in local isolates. The widespread surveillance of effector profiles within pathogen populations has great potential for crop disease management tailored to specific regions. In this WA-based study, low diversity sub-populations tended to have a conserved haplotype profile for the 3 known effector loci *ToxA*, *Tox1* and *Tox3* (Figure 11.1). In contrast sub-population cluster 5 exhibited notable diversity in its *Tox1* effector haplotypes. Overall this study indicates the potential for resistant cultivars to be broadly recommended for growing across regions identified to have low diversity, but less reliably for regions with higher diversity. Given the extreme potential for genome plasticity in fungal genomes (Croll et al., 2013; Hane et al., 2011; McClintock, 1941; Testa et al., 2016), it is encouraging that conserved effector haplotype profiles were observed in several cases. Identification of geographic regions exhibiting high variability of effector haplotype profiles within a narrow timeframe may also become an important element of crop disease monitoring in the future.

11.5.2 Genomic structure

A previous comparison of *P. nodorum* isolates from the USA and Australia (Bertazzoni et al., 2021; Richards et al., 2018, chapter 10), indicated that the smallest chromosome (AC23) is an accessory chromosome with high levels of mutation, diversifying selection, and numerous gene duplications for redundant pathogenicity-related functions. Both AC23 and the region of chromosome 4 encoding the known effector gene *ToxA*, appeared to exhibit structural mutations that may be influenced by breakage-fusion-bridge (BFB) formation. BFBs are hypothesised to be a driver of accessory chromosome formation and evolution (Bertazzoni et al., 2018; Croll et al., 2013) and of intrachromosomal recombination events that cumulatively lead to an inter-species conservation pattern termed "mesosynteny" (Hane et al., 2011). The largest duplications or absences in WA isolates relative to SN15 were predominantly located near telomeric regions or on AC23 (Figure 11.4). This is consistent with the known enrichment of structural rearrangement in subtelomeres (Hocher & Taddei, 2020) and proposed prevalence of BFB-mediated rearrangement across the Dothideomycetes (Croll et al., 2013). Complete and partial absences of SN15 AC23 were observed in some WA isolates, suggesting that large structural mutations are occurring in the field (Figure 11.4).

SNP density across the WA pan-genome was also consistently highest on AC23 (Figure 11.2) and near telomeres; however, several intrachromosomal mutation hotspots were also observed, where overall gene and repeat densities appeared relatively normal. The three known effector loci *ToxA*, *Tox1* and *Tox3* are all located at or near telomeres in SN15, which also corresponded with SNP hotspots. Additionally, a large orthogroup presence-absence variation (PAV) cluster was observed in a subset of WA isolates, which may represent a previously undescribed accessory chromosome or genomic regions. Future long-read sequencing of these isolates may resolve the structure and history of these fragmented regions.

Overall, the variable genes and genome regions across the pan-genome (PAV and SNP) did not highlight any clear association with known gene functions. Indeed, the gene ontology resource used in this study define function very broadly, and are biased towards conserved functions. Variable regions, including accessory, diversifying, repeat-rich and repeat-induced point mutation (RIP)-mutated, appeared to be bereft of known gene functions, but were conversely enriched in effector-like candidate loci. We observed 181 effector candidates that were absent in the international isolates and present in more than one WA isolate, of which 68 were restricted to a single WA subpopulation. This suggests that WA may have a distinct pathogenicity gene profile; however, the number of non-Australian isolates used in study was relatively small, so further comparison with international isolates may find these candidates elsewhere. The large PAV cluster of orthogroups which may represent an accessory chromosome or large chromosomal PAV, contained 18 effector candidates, including two putative cupredoxins. Five Tox1 and four Tox3 homologous orthogroups in the pan-genome, which were found in different frequencies across the different populations. This suggests that these necrotrophic effectors (NEs) have duplicated and diversified sufficiently that they are predicted as distinct orthogroups. Expansion and diversification of effectors within pathogen genomes appears to be a common phenomenon (de Guillen et al., 2015; Praz et al., 2017) in plant pathogens, and these homologues are strong effector candidates which may confer distinct phenotypes from their characterised homologues. Numerous homologues of effectors from other species were also identified, of which the necrotrophic effectors Zt6 and ZtNIP2 homologues are of particular interest to *P. nodorum*. Zt6 is a ribotoxin which cleaves non-self ribosomal sarcin-ricin loops in both wheat and microbial competitors (Kettles et al., 2018). ZtNIP1 induces light-dependent necrosis in wheat with differential responses between cultivars (M'Barek et al., 2015). Interestingly, numerous other homologues of Avirulence elicitors and biotrophic effectors were also identified. Many of these have functions related to nutrient and sugar scavenging, but may still promote virulence in *P. nodorum* more directly. For example, numerous MoCDIP4 homologues were identified which induces cell death in non-host plants of *Magnaporthe oryzae* (Chen et al., 2012). Apart from the possibility that *P. nodorum* benefits from this resistance elicitation, it was recently reported that MoCDIP4 interferes with mitochondrial homeostasis, which in turn inhibits mitochondrially mediated resistance responses (G. Xu et al., 2020). Interestingly, we previously identified a MoCDIP4-like effector candidate in *P. nodorum*, SNOG_01146 (SNOO_011460A), that was significantly upregulated *in planta* compared to *in vitro* and in a mutant isolate lacking the *PnPf2* transcription factor (D. A. B. Jones et al., 2019). Other effector-like orthogroups without fungal effector homologues but with potentially virulence related functions were identified. The putative aldose epimerase SNOO_616900 was one of the highest ranked candidates and was only observed in *P. nodorum* isolate RSID03. This appears to be a truncated copy of SNOO_063350A, which is only absent in isolate RSID03 and has a Predector score of 1.62. In *Phytophthora sojae* the apoplastic aldose 1-epimerase AEP1 functions as a virulence factor by scavenging apoplastic aldose, and can trigger cell death and pattern triggered immunity in *Nicotiana benthamiana* (Y. Xu et al., 2021).

Comparative genomics across the WA pan-genome also indicated the prevalence of RIP-like mutations in variable regions, and was also associated with candidate effector loci. The presence of distinct genome wide patterns of RIP-like mutation between isolates would indicate that RIP has been actively occurring within a recent time frame. However RIP-like mutations did not correlate with diversifying selection observed across the whole genome, or even within highly ranked effector loci. RIP occurs during pre-meiosis, however, the observation of linkage disequilibrium suggests that the WA population may not be regularly undergoing sexual recombination. Although previous studies indicate *P. nodorum* has the potential for meiosis in WA (Bathgate & Loughman, 2001; Murphy et al., 2000), it is yet to be determined if widespread RIP has only occurred in the past or if recent selection pressures have eliminated background isolate diversity. Investigations of the role of RIP in biotrophic and hemibiotrophic pathogens (Gervais et al., 2017; Testa et al., 2016) have indicated that RIP-mediated loss of a recognized avirulence effector may confer a selective advantage. We speculate that the nature of necrotrophic effectors with inverse gene-for-gene interactions with host sensitivity receptors (Fenton et al., 2009; Thrall et al., 2016) may conceal the full influence of RIP in necrotrophic effector diversification, as loss of function mutations are unlikely to be advantageous and would be selected against in the population. The remaining detectable RIP would be observable only for genome regions which do not significantly contribute a selective advantage. Nevertheless the *ToxA* locus resides in a large RIP hotspot on chromosome 4, and the confirmed effector loci *ToxA*, *Tox1* and *Tox3* retain a small number of non-synonymous RIP-like SNPs. The infrequent but constant potential for RIP to introduce potentially virulence enhancing mutations remains an important consideration for genome-guided disease risk assessment in necrotrophs.

11.6 Conclusion

Population-level pan-genome approaches are the next frontier of plant pathogen bioinformatics, which may eventually lead to affordable genome-based crop disease diagnostics and surveillance at a local level. Trends in pathogen genomics have begun to abandon intensive study of a single reference isolate, and are steadily progressing towards regionally-customised and data-driven assessments of pathogen gene-content, particularly with regards to effector genes (Badet & Croll, 2020; Richards et al., 2019). In this study, we analyse a local Western Australian population of the wheat pathogen *Parastagonospora nodorum* and identify multiple genome features of relevance to this pathosystem. We observed an apparently high potential for genome adaptability, suggested by the presence of active RIP and other mutations, but this was not readily observed to drive diversification of its three known highly conserved necrotrophic effectors. Mutation hotspots were identified which were rich in effector candidates and genes of unknown function, and often also classified as dispensable, sub-telomeric or large repeat-rich regions. In a spatial context, we observed regional 'hot' and 'cold-spots' of population diversity, that may be linked to climatic factors affecting spore dispersal. Across the local pan-genome, we observed the diversity of haplotype profiles of 3 known effector

genes to be conserved in regions with lower overall diversity. A total of 3579 novel effector candidates were predicted across all isolates, with 2291 of these exhibiting PAV across the genomes and 1362 restricted to WA isolates. Overall this study has progressively improved bioinformatic resources for the *P. nodorum* pathogen, as well as advancing approaches for the study of fungal pan-genomes with a view towards developing a region-specific understanding of host-pathogen interactions.

11.7 Acknowledgements

This study was supported by the Centre for Crop and Disease Management, a joint initiative of Curtin University and the Grains Research and Development Corporation (Research Grant CUR00023). This research was undertaken with the assistance of resources and services from the Pawsey Supercomputing Centre and the National Computational Infrastructure (NCI), which is supported by the Australian Government. This research is supported by an Australian Government Research Training Program (RTP) Scholarship.

11.8 Data availability

All sequencing data, genomes and annotations generated are available under NCBI Bioproject: PRJNA612761. Gene annotations for isolates SN2000, SN4, and SN79, as well as short variant predictions relative to SN15 are deposited online at https://doi.org/10.6084/m9.figshare.13340975. Complete functional annotations, orthogroup assignments, CDS alignments and trees, and positive selection tests are available online at https://doi.org/10.6084/m9.figshare.12966971.v3.

11.9 Supplementary material

All supplementary material is available online at https://doi.org/10.6084/m9.figshare.13325915.v2.

Supplementary table S1. Additional published genomes used in this study.

Supplementary table S2. Summary of Illumina sequencing read contamination detection. Counts of read pairs assigned to taxonomic groups are summarised for each barcoded sample corresponding to FASTQ file pairs. The column "provider" indicates the sequencing centre that performed the sequencing, either the Australian Genome Research Facility (AGRF. Melbourne, Australia), or Novogene (Beijing, China). The columns "percentage_fragments_under_taxon" and "number_fragments_under_taxon" indicates for higher taxonomic levels (e.g. Kingdom, Phylum or Family) the number or proportion of read pairs assigned to children of this taxonomic id. The column "number_fragments_at_taxon" indicates the number of read pairs that are specifically assigned to this taxonomic id, not including any child taxon. The column

"rank_code" indicates the taxonomic level of the taxon, i.e. (U)nclassified, (D)omain, (K)ingdom, (P)hylum, (C)lass, (O)rder, (F)amily, (G)enus, or (S)pecies.

Supplementary table S3. Parameters used to filter short variants by quality, and statistics of variants in the filtered set. The sheet "Bootstrap filtering parameters" shows the filtering parameters used during GATK (McKenna et al., 2010) base quality recalibration bootstrapping. The sheet "Filtering parameters" shows the filtering parameters selected for the final variant filtering set, based on data presented in supplementary data S5-S8. Remaining sheets summarise snpEff (Cingolani et al., 2012) results from the final filtered variant set. The sheet "variants per gene" shows predicted variant effects on genes present in SN15 annotated by Bertazzoni et al. (2021, chapter 10).

Supplementary table S4. Population diversity statistics and results of STRUCTURE analysis. The sheet "STRUCTURE" shows the posterior probabilities of each isolates' assignment to the cluster, alongside sampling metadata and results of a previous clustering analysis by (Phan et al., 2020, chapter 8).

Supplementary table S5. Genome assembly for all isolates sequenced in this study. Statistics were collected using BBtools stats (https://jgi.doe.gov/data-and-tools/bbtools/bb-tools-user-guide/statistics-guide/) and QUAST (Gurevich et al., 2013).

Supplementary table S6. Summaries statistics of transposable elements, rRNA and tRNA genes, and repeat annotations for each assembled genome. PFAM domains associated with transposable element genes used to soft mask repetitive gene regions are in the sheet "Pfam TE associated".

Supplementary table S7. Summary statistics of gene predictions for each isolate. Numbers are provided for each prediction method. EVM refers to EvidenceModeler predictions. Augustus "gapfiller" refers to the local gene finding using Augustus in regions where EVM has not annotated a gene but where supporting evidence exists. Final predictions are given by the columns labelled "filtered", while filtered excluded refers to the genes that were discarded due to insufficient supporting evidence.

Supplementary table S8. SNP, counts RIP-like SNP ratios, and genome assembly alignment coverage data used to plot circular heatmaps in figures 2, 3, and 4. Tables are in the same order as plotted on the figures, where the innermost row in the circular heatmap correspond to the left-most column in the table.

Supplementary table S9. Orthogroup counts for each isolate used to plot figure 5. The table is in the same order as the heatmap, but transposed so that the top-left column in the table corresponds to the top-right column in the figure. The column "selected" indicates the PAV clusters selected for further investigation.

Supplementary table S10. Functional annotation, selection, presence absence data for each orthogroup. A single representative sequence for each orthogroup was selected as the member

with the highest Predector score. The full orthogroup annotation can be found at https://doi.org/10.6084/m9.figshare.12966971.v3.

Supplementary table S11. GO term and effector enrichment tests for predicted functions and groups of orthogroups.

Supplementary data S1. MultiQC reports (https://multiqc.info/) of read trimming and quality control for Illumina sequencing reads.

Supplementary data S2. Boxplots showing short variant (SNP, insertion/deletion, Mixed) genotype quality (GQ) statistics for each isolate. Each chromosome in SN15 is shown on a separate page in the PDF.

Supplementary data S3. Violin plots showing short variant (SNP, insertion/deletion, Mixed) genotype read depth (DP) statistics for each isolate. Each chromosome in SN15 is shown on a separate page in the PDF.

Supplementary data S4. Bar plots showing amounts of missing short variant genotype information for each isolate. Each chromosome in SN15 is shown on a separate page in the PDF.

Supplementary data S5. SNP locus quality statistics visualised for each chromosome in SN15 on separate pages in the PDF. Plots were created using the R package vcfR (Knaus & Grünwald, 2017).

Supplementary data S6. Insertion and Deletion (INDEL) locus quality statistics visualised for each chromosome in SN15 on separate pages in the PDF. Plots were created using the R package vcfR (Knaus & Grünwald, 2017).

Supplementary data S7. Mixed variant (multi-nucleotide variations, or insertions/deletions with SNPs at the same locus) locus quality statistics visualised for each chromosome in SN15 on separate pages in the PDF. Plots were created using the R package vcfR (Knaus & Grünwald, 2017).

Supplementary data S8. Kernel density estimate plots showing the distributions of short variant locus quality statistics. Statistics presented include the BaseQRankSum (approximate Z-scores from a test comparing base qualities of reference and alternate alleles), FS (the Phred-scaled probability that there is read strand bias at the site), MQ (the root mean square mapping quality over all the reads at the site), MQRankSum (Approximate Z-scores from a test comparing MQ of reference and alternate alleles), QD (Variant quality QUAL divided by the read alignment depth DP), ReadPosRankSum (approximate Z-scores from a test for read strand bias), and SOR (another test for read strand bias where high values indicate bias).

Supplementary data S9. Maximum likelihood phylogenetic tree estimated from 45,194 SNPs using IQTree (Minh et al., 2020). The file is in Newick format. Clade confidence values show SH-aLRT and UFBoot support separated by '/'.

Supplementary data S10. MSA and trees of ToxA, 1, 3 CDS/codon-aligned regions from pan-genome, to support prevalence of RIP-like SNPs across pan-genome in confirmed effector loci.

Supplementary data S11. Example dot plot alignments between scaffolds and chromosomes containing orthogroups in PAV clusters selected from Figure 11.5. Alignments and plots were generated using MashMap (Jain et al., 2018).

Supplementary figure S1. Phylogeographic representation of the WA *P. nodorum* populations, with phylogeny generated from whole-genome SNP data relative to alignment to the SN15 reference genome, and yellow lines indicating the approximate location of sampling. Isolates that were sampled in WA but without specific location information do not have points plotted on the map. For these isolates the sampling years are presented in the pie chart and each wedge represents either one or two isolates.

Supplementary figure S2. Tanglegram comparison of predicted SNP phylogeny with the SSR predicted tree from Phan et al. (2020, chapter 8). Clades highlighted with colour indicate where the clade structure is preserved between both trees.

Supplementary figure S3. Comparison of population cluster assignment between this study and as identified by Phan et al. (2020, chapter 8). Numbers in the heatmap indicate the number of isolates present in the clusters of each analysis. NA indicates assignment of isolates included in one analysis but not the other (e.g. group 8 from this study consists of international isolates which were not present in the previous study).

Supplementary figure S4. Numbers of isolates in clusters from each sampling location.

Supplementary figure S5. Numbers of isolates in clusters from each sampling year.

Supplementary figure S6. The first six principal components computed from bi-allelic SNP data plotted for each sampling location.

Supplementary figure S7. The first six principal components computed from bi-allelic SNP data plotted against each sampling year.

11.10 References

Abeysekara, N. S., Friesen, T. L., Keller, B., & Faris, J. D. (2009). Identification and characterization of a novel host–toxin interaction in the wheat–*Stagonospora nodorum* pathosystem. *Theoretical and Applied Genetics*, *120*(1), 117–126. https://doi.org/10.1007/s00122-009-1163-6

Agapow, P.-M., & Burt, A. (2001). Indices of multilocus linkage disequilibrium. *Molecular Ecology Notes*, *1*(1-2), 101–102. https://doi.org/10.1046/j.1471-8278.2000.00014.x

Armenteros, J. J. A., Salvatore, M., Emanuelsson, O., Winther, O., Heijne, G. v., Elofsson, A., & Nielsen, H. (2019). Detecting sequence signals in targeting peptides using deep learning. *Life Science Alliance*, *2*(5). https://doi.org/10.26508/lsa.201900429

Armenteros, J. J. A., Sønderby, C. K., Sønderby, S. K., Nielsen, H., & Winther, O. (2017). DeepLoc: Prediction of protein subcellular localization using deep learning. *Bioinformatics*, *33*(21), 3387–3395. https://doi.org/10.1093/bioinformatics/btx431

Armenteros, J. J. A., Tsirigos, K. D., Sønderby, C. K., Petersen, T. N., Winther, O., Brunak, S., Heijne, G. v., & Nielsen, H. (2019). SignalP 5.0 improves signal peptide predictions using deep neural networks. *Nature Biotechnology*, *37*(4), 420–423. https://doi.org/10.1038/s41587-019-0036-z

Aylward, J., Steenkamp, E. T., Dreyer, L. L., Roets, F., Wingfield, B. D., & Wingfield, M. J. (2017). A plant pathology perspective of fungal genome sequencing. *IMA Fungus*, *8*(1), 1–15. https://doi.org/10.5598/imafungus.2017.08.01.01

Badet, T., & Croll, D. (2020). The rise and fall of genes: Origins and functions of plant pathogen pangenomes. *Current Opinion in Plant Biology*, *56*, 65–73. https://doi.org/10.1016/j.pbi.2020.04.009

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, *28*(1), 45–48. https://doi.org/10.1093/nar/28.1.45

Bankevich, A., Nurk, S., Antipov, D., Gurevich, A. A., Dvorkin, M., Kulikov, A. S., Lesin, V. M., Nikolenko, S. I., Pham, S., Prjibelski, A. D., Pyshkin, A. V., Sirotkin, A. V., Vyahhi, N., Tesler, G., Alekseyev, M. A., & Pevzner, P. A. (2012). SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology*, *19*(5), 455–477. https://doi.org/10.1089/cmb.2012.0021

Bathgate, J. A., & Loughman, R. (2001). Ascospores are a source of inoculum of *Phaeosphaeria nodorum*, *P. avenaria* f. sp. *avenaria* and *Mycosphaerella graminicola* in Western Australia. *Australasian Plant Pathology*, *30*(4), 317. https://doi.org/10.1071/AP01043

Bendtsen, J. D., Nielsen, H., von Heijne, G., & Brunak, S. (2004). Improved Prediction of Signal Peptides: SignalP 3.0. *Journal of Molecular Biology*, *340*(4), 783–795. https://doi.org/10.1016/j.jmb.2004.05.028

Bennett, R. S., Milgroom, M. G., & Bergstrom, G. C. (2005). Population Structure of Seedborne *Phaeosphaeria nodorum* on New York Wheat. *Phytopathology*, *95*(3), 300–305. https://doi.org/10.1094/PHYTO-95-0300

Bertazzoni, S., Jones, D. A. B., Phan, H. T., Tan, K.-C., & Hane, J. K. (2021). Chromosome-level genome assembly and manually-curated proteome of model necrotroph *Parastagonospora nodorum* sn15 reveals a genome-wide trove of candidate effector homologs, and redundancy of virulence-related functions within an accessory chromosome. *BMC Genomics*, *22*(382). https://doi.org/10.1186/s12864-021-07699-8

Bertazzoni, S., Williams, A. H., Jones, D. A., Syme, R. A., Tan, K.-C., & Hane, J. K. (2018). Accessories Make the Outfit: Accessory Chromosomes and Other Dispensable DNA Regions in Plant-Pathogenic Fungi. *Molecular Plant-Microbe Interactions*, *31*(8), 779–788. https://doi.org/10.1094/MPMI-06-17-0135-FI

Bringans, S., Hane, J. K., Casey, T., Tan, K.-C., Lipscombe, R., Solomon, P. S., & Oliver, R. P. (2009). Deep proteogenomics; high throughput gene validation by multidimensional liquid chromatography and mass spectrometry of proteins from the fungal wheat pathogen *Stagonospora nodorum*. *BMC Bioinformatics*, *10*(1), 301. https://doi.org/10.1186/1471-2105-10-301

Bu, B., Qiu, D., Zeng, H., Guo, L., Yuan, J., & Yang, X. (2014). A fungal protein elicitor PevD1 induces Verticillium wilt resistance in cotton. *Plant Cell Reports*, *33*(3), 461–470. https://doi.org/10.1007/s00299-013-1546-7

Buchfink, B., Reuter, K., & Drost, H.-G. (2021). Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nature Methods*, *18*(4), 366–368. https://doi.org/10.1038/s41592-021-01101-x

Bushnell, B., Rood, J., & Singer, E. (2017). BBMerge – Accurate paired shotgun read merging via overlap. *PLOS ONE*, *12*(10), e0185056. https://doi.org/10.1371/journal.pone.0185056

Camacho, C., Coulouris, G., Avagyan, V., Ma, N., Papadopoulos, J., Bealer, K., & Madden, T. L. (2009). BLAST+: Architecture and applications. *BMC Bioinformatics*, *10*(1), 421. https://doi.org/10.1186/1471-2105-10-421

Caten, C. E., & Newton, A. C. (2000). Variation in cultural characteristics, pathogenicity, vegetative compatibility and electrophoretic karyotype within field populations of *Stagonospora nodorum*. *Plant Pathology*, *49*(2), 219–226. https://doi.org/10.1046/j.1365-3059.2000.00441.x

Chen, S., Songkumarn, P., Venu, R. C., Gowda, M., Bellizzi, M., Hu, J., Liu, W., Ebbole, D., Meyers, B., Mitchell, T., & Wang, G.-L. (2012). Identification and Characterization of In planta–Expressed Secreted Effector Proteins from *Magnaporthe oryzae* That Induce Cell Death in Rice. *Molecular Plant-Microbe Interactions*, *26*(2), 191–202. https://doi.org/10.1094/MPMI-05-12-0117-R

Chooi, Y.-H., Muria-Gonzalez, M. J., & Solomon, P. S. (2014). A genome-wide survey of the secondary metabolite biosynthesis genes in the wheat pathogen *Parastagonospora nodorum*. *Mycology*, *5*(3), 192–206. https://doi.org/10.1080/21501203.2014.928386

Chuma, I., Isobe, C., Hotta, Y., Ibaragi, K., Futamata, N., Kusaba, M., Yoshida, K., Terauchi, R., Fujita, Y., Nakayashiki, H., Valent, B., & Tosa, Y. (2011). Multiple Translocation of the *AVR-Pita* Effector Gene among Chromosomes of the Rice Blast Fungus *Magnaporthe oryzae* and Related Species. *PLOS Pathogens*, *7*(7), e1002147. https://doi.org/10.1371/journal.ppat.1002147

Cingolani, P., Platts, A., Wang, L. L., Coon, M., Nguyen, T., Wang, L., Land, S. J., Lu, X., & Ruden, D. M. (2012). A program for annotating and predicting the effects of single nucleotide polymorphisms, SnpEff. *Fly*, *6*(2), 80–92. https://doi.org/10.4161/fly.19695

Croll, D., Zala, M., & McDonald, B. A. (2013). Breakage-fusion-bridge Cycles and Large Insertions Contribute to the Rapid Evolution of Accessory Chromosomes in a Fungal Pathogen. *PLOS Genetics*, *9*(6), e1003567. https://doi.org/10.1371/journal.pgen.1003567

Cunfer, B. M. (1978). The Incidence of *Septoria nodorum* in Wheat Seed. *Phytopathology*, *68*(6), 832. https://doi.org/10.1094/Phyto-68-832

Cunfer, B. M. (1998). Seasonal availability of inoculum of *Stagonospora nodorum* in the field in the southeastern US. *Cereal Research Communications*, *26*(3), 259–263.

Dai, Y., Jia, Y., Correll, J., Wang, X., & Wang, Y. (2010). Diversification and evolution of the avirulence gene *AVR-Pita1* in field isolates of *Magnaporthe oryzae*. *Fungal Genetics and Biology*, *47*(12), 973–980. https://doi.org/10.1016/j.fgb.2010.08.003

de Guillen, K. d., Ortiz-Vallejo, D., Gracy, J., Fournier, E., Kroj, T., & Padilla, A. (2015). Structure Analysis Uncovers a Highly Diverse but Structurally Conserved Effector Family in Phytopathogenic Fungi. *PLOS Pathogens*, *11*(10), e1005228. https://doi.org/10.1371/journal.ppat.1005228

Dierckxsens, N., Mardulyn, P., & Smits, G. (2017). NOVOPlasty: De novo assembly of organelle genomes from whole genome data. *Nucleic Acids Research*, *45*(4), e18–e18. https://doi.org/10.1093/nar/gkw955

Dobin, A., Davis, C. A., Schlesinger, F., Drenkow, J., Zaleski, C., Jha, S., Batut, P., Chaisson, M., & Gingeras, T. R. (2013). STAR: Ultrafast universal RNA-seq aligner. *Bioinformatics*, *29*(1), 15–21. https://doi.org/10.1093/bioinformatics/bts635

Earl, D. A., & vonHoldt, B. M. (2012). STRUCTURE HARVESTER: A website and program for visualizing STRUCTURE output and implementing the Evanno method. *Conservation Genetics Resources*, *4*(2), 359–361. https://doi.org/10.1007/s12686-011-9548-7

Eberhardt, R. Y., Haft, D. H., Punta, M., Martin, M., O'Donovan, C., & Bateman, A. (2012). AntiFam: A tool to help identify spurious ORFs in protein annotation. *Database*, *2012*. https://doi.org/10.1093/database/bas003

Ellinghaus, D., Kurtz, S., & Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *BMC Bioinformatics*, *9*(1), 18. https://doi.org/10.1186/1471-2105-9-18

Evanno, G., Regnaut, S., & Goudet, J. (2005). Detecting the number of clusters of individuals using the software structure: A simulation study. *Molecular Ecology*, *14*(8), 2611–2620. https://doi.org/10.1111/j.1365-294X.2005.02553.x

Ewels, P., Magnusson, M., Lundin, S., & Käller, M. (2016). MultiQC: Summarize analysis results for multiple tools and samples in a single report. *Bioinformatics*, *32*(19), 3047–3048. https://doi.org/10.1093/bioinformatics/btw354

Fenton, A., Antonovics, J., & Brockhurst, M. A. (2009). Inverse-Gene-for-Gene Infection Genetics and Coevolutionary Dynamics. *The American Naturalist*, *174*(6), E230–E242. https://doi.org/10.1086/645087

Friesen, T. L., Chu, C., Xu, S. S., & Faris, J. D. (2012). *SnTox5-Snn5*: A novel *Stagonospora nodorum* effector-wheat gene interaction and its relationship with the SnToxA- *Tsn1* and SnTox3- <*i*>*Snn3* </i> - *B1* interactions: Characterization of the SnTox5-*Snn5*interaction. *Molecular Plant Pathology*, *13*(9), 1101–1109. https://doi.org/10.1111/j.1364-3703.2012.00819.x

Friesen, T. L., Meinhardt, S. W., & Faris, J. D. (2007). The *Stagonospora nodorum*-wheat pathosystem involves multiple proteinaceous host-selective toxins and corresponding host sensitivity genes that interact in an inverse gene-for-gene manner. *The Plant Journal*, *51*(4), 681–692. https://doi.org/10.1111/j.1365-313X.2007.03166.x

Friesen, T. L., Zhang, Z., Solomon, P. S., Oliver, R. P., & Faris, J. D. (2008). Characterization of the Interaction of a Novel *Stagonospora nodorum* Host-Selective Toxin with a Wheat Susceptibility Gene. *Plant Physiology*, *146*(2), 682–693. https://doi.org/10.1104/pp.107.108761

Galili, T. (2015). Dendextend: An R package for visualizing, adjusting and comparing trees of hierarchical clustering. *Bioinformatics*, *31*(22), 3718–3720. https://doi.org/10.1093/bioinformatics/btv428

Gao, Y., Faris, J. D., Liu, Z., Kim, Y. M., Syme, R. A., Oliver, R. P., Xu, S. S., & Friesen, T. L. (2015). Identification and Characterization of the SnTox6-Snn6 Interaction in the *Parastagonospora nodorum*–Wheat Pathosystem. *Molecular Plant-Microbe Interactions*, *28*(5), 615–625. https://doi.org/10.1094/MPMI-12-14-0396-R

Gervais, J., Plissonneau, C., Linglin, J., Meyer, M., Labadie, K., Cruaud, C., Fudal, I., Rouxel, T., & Balesdent, M.-H. (2017). Different waves of effector genes with contrasted genomic location are expressed by *Leptosphaeria maculans* during cotyledon and stem colonization of oilseed rape. *Molecular Plant Pathology*, *18*(8), 1113–1126. https://doi.org/10.1111/mpp.12464

Ghaderi, F., Sharifnabi, B., Javan-Nikkhah, M., Brunner, P. C., & McDonald, B. A. (2020). *SnToxA* , *SnTox1* , and *SnTox3* originated in *Parastagonospora nodorum* in the Fertile Crescent. *Plant Pathology*, ppa.13233. https://doi.org/10.1111/ppa.13233

Grabherr, M. G., Haas, B. J., Yassour, M., Levin, J. Z., Thompson, D. A., Amit, I., Adiconis, X., Fan, L., Raychowdhury, R., Zeng, Q., Chen, Z., Mauceli, E., Hacohen, N., Gnirke, A., Rhind, N., Palma, F. d., Birren, B. W., Nusbaum, C., Lindblad-Toh, K., … Regev, A. (2011). Full-length

transcriptome assembly from RNA-Seq data without a reference genome. *Nature Biotechnology*, *29*(7), 644–652. https://doi.org/10.1038/nbt.1883

Gremme, G., Steinbiss, S., & Kurtz, S. (2013). GenomeTools: A Comprehensive Software Library for Efficient Processing of Structured Genome Annotations. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, *10*(3), 645–656. https://doi.org/10.1109/TCBB.2013.68

Gu, Z., Gu, L., Eils, R., Schlesner, M., & Brors, B. (2014). Circlize implements and enhances circular visualization in R. *Bioinformatics*, *30*(19), 2811–2812. https://doi.org/10.1093/bioinformatics/btu393

Gummer, J. P. A., Trengove, R. D., Oliver, R. P., & Solomon, P. S. (2013). Dissecting the role of G-protein signalling in primary metabolism in the wheat pathogen *Stagonospora nodorum*. *Microbiology*, *159*(Pt_9), 1972–1985. https://doi.org/10.1099/mic.0.067009-0

Gurevich, A., Saveliev, V., Vyahhi, N., & Tesler, G. (2013). QUAST: Quality assessment tool for genome assemblies. *Bioinformatics*, *29*(8), 1072–1075. https://doi.org/10.1093/bioinformatics/btt086

Haas, B. J., Delcher, A. L., Mount, S. M., Wortman, J. R., Smith Jr, R. K., Hannick, L. I., Maiti, R., Ronning, C. M., Rusch, D. B., Town, C. D., Salzberg, S. L., & White, O. (2003). Improving the Arabidopsis genome annotation using maximal transcript alignment assemblies. *Nucleic Acids Research*, *31*(19), 5654–5666. https://doi.org/10.1093/nar/gkg770

Haas, B. J., Salzberg, S. L., Zhu, W., Pertea, M., Allen, J. E., Orvis, J., White, O., Buell, C. R., & Wortman, J. R. (2008). Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome Biology*, *9*(1), R7. https://doi.org/10.1186/gb-2008-9-1-r7

Hane, J. K., Lowe, R. G. T., Solomon, P. S., Tan, K.-C., Schoch, C. L., Spatafora, J. W., Crous, P. W., Kodira, C., Birren, B. W., Galagan, J. E., Torriani, S. F. F., McDonald, B. A., & Oliver, R. P. (2007). Dothideomycete–Plant Interactions Illuminated by Genome Sequencing and EST Analysis of the Wheat Pathogen Stagonospora nodorum. *The Plant Cell*, *19*(11), 3347–3368. https://doi.org/10.1105/tpc.107.052829

Hane, J. K., Rouxel, T., Howlett, B. J., Kema, G. H., Goodwin, S. B., & Oliver, R. P. (2011). A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome Biology*, *12*(5), R45. https://doi.org/10.1186/gb-2011-12-5-r45

Hane, J. K., Williams, A. H., Taranto, A. P., Solomon, P. S., & Oliver, R. P. (2015). Repeat-Induced Point Mutation: A Fungal-Specific, Endogenous Mutagenesis Process. In M. A. van den Berg & K. Maruthachalam (Eds.), *Genetic Transformation Systems in Fungi* (pp. 55–68). Cham, Springer International Publishing. https://doi.org/10.1007/978-3-319-10503-1_4

Hedrick, P. W. (2005). A standardized genetic differentiation measure. *Evolution*, *59*(8), 1633–1638.

Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q., & Vinh, L. S. (2018). UFBoot2: Improving the Ultrafast Bootstrap Approximation. *Molecular Biology and Evolution*, *35*(2), 518–522. https://doi.org/10.1093/molbev/msx281

Hocher, A., & Taddei, A. (2020). Subtelomeres as Specialized Chromatin Domains. *BioEssays*, *42*(5), 1900205. https://doi.org/10.1002/bies.201900205

Hu, J., Zheng, Y., & Shang, X. (2018). MiteFinderII: A novel tool to identify miniature inverted-repeat transposable elements hidden in eukaryotic genomes. *BMC Medical Genomics*, *11*(5), 101. https://doi.org/10.1186/s12920-018-0418-y

Huerta-Cepas, J., Forslund, K., Coelho, L. P., Szklarczyk, D., Jensen, L. J., von Mering, C., & Bork, P. (2017). Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-

Mapper. *Molecular Biology and Evolution*, *34*(8), 2115–2122. https://doi.org/10.1093/molbev/msx148

Hurlbert, S. H. (1971). The Nonconcept of Species Diversity: A Critique and Alternative Parameters. *Ecology*, *52*(4), 577–586. https://doi.org/10.2307/1934145

Ipcho, S. V. S., Hane, J. K., Antoni, E. A., Ahren, D., Henrissat, B., Friesen, T. L., Solomon, P. S., & Oliver, R. P. (2012). Transcriptome analysis of *Stagonospora nodorum*: Gene models, effectors, metabolism and pantothenate dispensability. *Molecular Plant Pathology*, *13*(6), 531–545. https://doi.org/10.1111/j.1364-3703.2011.00770.x

Irieda, H., Inoue, Y., Mori, M., Yamada, K., Oshikawa, Y., Saitoh, H., Uemura, A., Terauchi, R., Kitakura, S., Kosaka, A., Singkaravanit-Ogawa, S., & Takano, Y. (2019). Conserved fungal effector suppresses PAMP-triggered immunity by targeting plant immune kinases. *Proceedings of the National Academy of Sciences*, *116*(2), 496–505. https://doi.org/10.1073/pnas.1807297116

Iwata, H., & Gotoh, O. (2012). Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Research*, *40*(20), e161–e161. https://doi.org/10.1093/nar/gks708

Jain, C., Koren, S., Dilthey, A., Phillippy, A. M., & Aluru, S. (2018). A fast adaptive algorithm for computing whole-genome homology maps. *Bioinformatics*, *34*(17), i748–i756. https://doi.org/10.1093/bioinformatics/bty597

Jombart, T. (2008). Adegenet: A R package for the multivariate analysis of genetic markers. *Bioinformatics*, *24*(11), 1403–1405. https://doi.org/10.1093/bioinformatics/btn129

Jones, D. A. B., John, E., Rybak, K., Phan, H. T. T., Singh, K. B., Lin, S.-Y., Solomon, P. S., Oliver, R. P., & Tan, K.-C. (2019). A specific fungal transcription factor controls effector gene expression and orchestrates the establishment of the necrotrophic pathogen lifestyle on wheat. *Scientific Reports*, *9*(1), 1–13. https://doi.org/10.1038/s41598-019-52444-7

Jones, D. A., Bertazzoni, S., Turo, C. J., Syme, R. A., & Hane, J. K. (2018). Bioinformatic prediction of plant–pathogenicity effector proteins of fungi. *Current Opinion in Microbiology*, *46*, 43–49. https://doi.org/10.1016/j.mib.2018.01.017

Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., Pesseat, S., Quinn, A. F., Sangrador-Vegas, A., Scheremetjew, M., Yong, S.-Y., Lopez, R., & Hunter, S. (2014). InterProScan 5: Genome-scale protein function classification. *Bioinformatics*, *30*(9), 1236–1240. https://doi.org/10.1093/bioinformatics/btu031

Käll, L., Krogh, A., & Sonnhammer, E. L. L. (2004). A Combined Transmembrane Topology and Signal Peptide Prediction Method. *Journal of Molecular Biology*, *338*(5), 1027–1036. https://doi.org/10.1016/j.jmb.2004.03.016

Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A., & Jermiin, L. S. (2017). ModelFinder: Fast model selection for accurate phylogenetic estimates. *Nature Methods*, *14*(6), 587–589. https://doi.org/10.1038/nmeth.4285

Kamvar, Z. N., Tabima, J. F., & Grünwald, N. J. (2014). *Poppr*: An R package for genetic analysis of populations with clonal, partially clonal, and/or sexual reproduction. *PeerJ*, *2*, e281. https://doi.org/10.7717/peerj.281

Kanja, C., & Hammond-Kosack, K. E. (2020). Proteinaceous effector discovery and characterization in filamentous plant pathogens. *Molecular Plant Pathology*. https://doi.org/10.1111/mpp.12980

Keilwagen, J., Hartung, F., Paulini, M., Twardziok, S. O., & Grau, J. (2018). Combining RNA-seq data and homology-based gene prediction for plants, animals and fungi. *BMC Bioinformatics*, *19*(1), 189. https://doi.org/10.1186/s12859-018-2203-5

Keller, S. M., McDermott, J. M., Pettway, R. E., Wolfe, M. S., & McDonald, B. A. (1997). Gene Flow and Sexual Reproduction in the Wheat Glume Blotch Pathogen *Phaeosphaeria nodorum* (Anamorph *Stagonospora nodorum*). *Phytopathology*, *87*(3), 353–358. https://doi.org/10.1094/PHYTO.1997.87.3.353

Kettles, G. J., Bayon, C., Sparks, C. A., Canning, G., Kanyuka, K., & Rudd, J. J. (2018). Characterization of an antimicrobial and phytotoxic ribonuclease secreted by the fungal wheat pathogen *Zymoseptoria tritici*. *The New Phytologist*, *217*(1), 320–331. https://doi.org/10.1111/nph.14786

Kim, S., Park, H., Gruszewski, H. A., Schmale, D. G., & Jung, S. (2019). Vortex-induced dispersal of a plant pathogen by raindrop impact. *Proceedings of the National Academy of Sciences*, *116*(11), 4917–4922. https://doi.org/10.1073/pnas.1820318116

Klopfenstein, D. V., Zhang, L., Pedersen, B. S., Ramírez, F., Warwick Vesztrocy, A., Naldi, A., Mungall, C. J., Yunes, J. M., Botvinnik, O., Weigel, M., Dampier, W., Dessimoz, C., Flick, P., & Tang, H. (2018). GOATOOLS: A Python library for Gene Ontology analyses. *Scientific Reports*, *8*(1), 10872. https://doi.org/10.1038/s41598-018-28948-z

Knaus, B. J., & Grünwald, N. J. (2017). Vcfr: A package to manipulate and visualize variant call format data in R. *Molecular Ecology Resources*, *17*(1), 44–53. https://doi.org/10.1111/1755-0998.12549

Koskinen, P., Törönen, P., Nokso-Koivisto, J., & Holm, L. (2015). PANNZER: High-throughput functional annotation of uncharacterized proteins in an error-prone environment. *Bioinformatics*, *31*(10), 1544–1552. https://doi.org/10.1093/bioinformatics/btu851

Krogh, A., Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. *Journal of Molecular Biology*, *305*(3), 567–580. https://doi.org/10.1006/jmbi.2000.4315

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H.-H., Rognes, T., & Ussery, D. W. (2007). RNAmmer: Consistent and rapid annotation of ribosomal RNA genes. *Nucleic Acids Research*, *35*(9), 3100–3108. https://doi.org/10.1093/nar/gkm160

Lechner, M., Findeiß, S., Steiner, L., Marz, M., Stadler, P. F., & Prohaska, S. J. (2011). Proteinortho: Detection of (co-)orthologs in large-scale analysis. *BMC Bioinformatics*, *12*(1). https://doi.org/10.1186/1471-2105-12-124

Li, H. (2011). A statistical framework for SNP calling, mutation discovery, association mapping and population genetical parameter estimation from sequencing data. *Bioinformatics*, *27*(21), 2987–2993. https://doi.org/10.1093/bioinformatics/btr509

Li, H. (2018). Minimap2: Pairwise alignment for nucleotide sequences. *Bioinformatics*, *34*(18), 3094–3100. https://doi.org/10.1093/bioinformatics/bty191

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., & Durbin, R. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, *25*(16), 2078–2079. https://doi.org/10.1093/bioinformatics/btp352

Liu, Z., Faris, J. D., Oliver, R. P., Tan, K.-C., Solomon, P. S., McDonald, M. C., McDonald, B. A., Nunez, A., Lu, S., Rasmussen, J. B., & Friesen, T. L. (2009). SnTox3 Acts in Effector Triggered Susceptibility to Induce Disease on Wheat Carrying the *Snn3* Gene. *PLOS Pathogens*, *5*(9), e1000581. https://doi.org/10.1371/journal.ppat.1000581

Liu, Z., Friesen, T. L., Ling, H., Meinhardt, S. W., Oliver, R. P., Rasmussen, J. B., & Faris, J. D. (2006). The Tsn1–ToxA interaction in the wheat–*Stagonospora nodorum* pathosystem parallels that of the wheat–tan spot system. *Genome*, *49*(10), 1265–1273. https://doi.org/10.1139/g06-088

Liu, Z., Zhang, Z., Faris, J. D., Oliver, R. P., Syme, R., McDonald, M. C., McDonald, B. A., Solomon, P. S., Lu, S., Shelver, W. L., Xu, S., & Friesen, T. L. (2012). The Cysteine Rich Necrotrophic Effector

SnTox1 Produced by *Stagonospora nodorum* Triggers Susceptibility of Wheat Lines Harboring Snn1. *PLOS Pathogens*, *8*(1), e1002467. https://doi.org/10.1371/journal.ppat.1002467

Llorens, C., Futami, R., Covelli, L., Domínguez-Escribá, L., Viu, J. M., Tamarit, D., Aguilar-Rodríguez, J., Vicente-Ripolles, M., Fuster, G., Bernet, G. P., Maumus, F., Munoz-Pomer, A., Sempere, J. M., Latorre, A., & Moya, A. (2011). The Gypsy Database (GyDB) of mobile genetic elements: Release 2.0. *Nucleic Acids Research*, *39*(suppl_1), D70–D74. https://doi.org/10.1093/nar/gkq1061

Lomsadze, A., Burns, P. D., & Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research*, *42*(15), e119–e119. https://doi.org/10.1093/nar/gku557

Lowe, R. G. T., Lord, M., Rybak, K., Trengove, R. D., Oliver, R. P., & Solomon, P. S. (2008). A metabolomic approach to dissecting osmotic stress in the wheat pathogen *Stagonospora nodorum*. *Fungal Genetics and Biology*, *45*(11), 1479–1486. https://doi.org/10.1016/j.fgb.2008.08.006

Lowe, T. M., & Chan, P. P. (2016). tRNAscan-SE On-line: Integrating search and context for analysis of transfer RNA genes. *Nucleic Acids Research*, *44*(W1), W54–W57. https://doi.org/10.1093/nar/gkw413

Lunter, G., & Goodson, M. (2011). Stampy: A statistical algorithm for sensitive and fast mapping of Illumina sequence reads. *Genome Research*, *21*(6), 936–939. https://doi.org/10.1101/gr.111120.110

Mapleson, D., Garcia Accinelli, G., Kettleborough, G., Wright, J., & Clavijo, B. J. (2017). KAT: A K-mer analysis toolkit to quality control NGS datasets and genome assemblies. *Bioinformatics*, *33*(4), 574–576. https://doi.org/10.1093/bioinformatics/btw663

Marçais, G., Delcher, A. L., Phillippy, A. M., Coston, R., Salzberg, S. L., & Zimin, A. (2018). MUMmer4: A fast and versatile genome alignment system. *PLOS Computational Biology*, *14*(1), e1005944. https://doi.org/10.1371/journal.pcbi.1005944

Martin, M. (2011). Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal*, *17*(1), 10–12. https://doi.org/10.14806/ej.17.1.200

M'Barek, S. B., Cordewener, J. H. G., Ghaffary, S. M. T., van der Lee, T. A. J., Liu, Z., Mirzadi Gohari, A., Mehrabi, R., America, A. H. P., Robert, O., Friesen, T. L., Hamza, S., Stergiopoulos, I., de Wit, P. J. G. M., & Kema, G. H. J. (2015). FPLC and liquid-chromatography mass spectrometry identify candidate necrosis-inducing proteins from culture filtrates of the fungal wheat pathogen *Zymoseptoria tritici*. *Fungal Genetics and Biology*, *79*, 54–62. https://doi.org/10.1016/j.fgb.2015.03.015

McClintock, B. (1941). The Stability of Broken Ends of Chromosomes in *Zea mays*. *Genetics*, *26*(2), 234–282.

McDonald, M. C., Razavi, M., Friesen, T. L., Brunner, P. C., & McDonald, B. A. (2012). Phylogenetic and population genetic analyses of *Phaeosphaeria nodorum* and its close relatives indicate cryptic species and an origin in the Fertile Crescent. *Fungal Genetics and Biology*, *49*(11), 882–895. https://doi.org/10.1016/j.fgb.2012.08.001

McKenna, A., Hanna, M., Banks, E., Sivachenko, A., Cibulskis, K., Kernytsky, A., Garimella, K., Altshuler, D., Gabriel, S., Daly, M., & DePristo, M. A. (2010). The Genome Analysis Toolkit: A MapReduce framework for analyzing next-generation DNA sequencing data. *Genome Research*, *20*(9), 1297–1303. https://doi.org/10.1101/gr.107524.110

Minh, B. Q., Schmidt, H. A., Chernomor, O., Schrempf, D., Woodhams, M. D., von Haeseler, A., & Lanfear, R. (2020). IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution*, *37*(5), 1530–1534. https://doi.org/10.1093/molbev/msaa015

Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A., & Punta, M. (2013). Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Research*, *41*(12), e121–e121. https://doi.org/10.1093/nar/gkt263

Mitchell, A. L., Attwood, T. K., Babbitt, P. C., Blum, M., Bork, P., Bridge, A., Brown, S. D., Chang, H.-Y., El-Gebali, S., Fraser, M. I., Gough, J., Haft, D. R., Huang, H., Letunic, I., Lopez, R., Luciani, A., Madeira, F., Marchler-Bauer, A., Mi, H., … Finn, R. D. (2019). InterPro in 2019: Improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Research*, *47*(D1), D351–D360. https://doi.org/10.1093/nar/gky1100

Mosquera, G., Giraldo, M. C., Khang, C. H., Coughlan, S., & Valent, B. (2009). Interaction Transcriptome Analysis Identifies *Magnaporthe oryzae* BAS1-4 as Biotrophy-Associated Secreted Proteins in Rice Blast Disease. *The Plant Cell*, *21*(4), 1273–1290. https://doi.org/10.1105/tpc.107.055228

Muria-Gonzalez, M. J., Yeng, Y., Breen, S., Mead, O., Wang, C., Chooi, Y.-H., Barrow, R. A., & Solomon, P. S. (2020). Volatile Molecules Secreted by the Wheat Pathogen *Parastagonospora nodorum* Are Involved in Development and Phytotoxicity. *Frontiers in Microbiology*, *11*. https://doi.org/10.3389/fmicb.2020.00466

Murphy, N. E., Loughman, R., Appels, R., Lagudah, E. S., & Jones, M. G. K. (2000). Genetic variability in a collection of *Stagonospora nodorum* isolates from Western Australia. *Australian Journal of Agricultural Research*, *51*(6), 679–684. https://doi.org/10.1071/ar99107

Murray, G. M., & Brennan, J. P. (2009). Estimating disease losses to the Australian wheat industry. *Australasian Plant Pathology*, *38*(6), 558–570. https://doi.org/10.1071/AP09053

Murrell, B., Moola, S., Mabona, A., Weighill, T., Sheward, D., Kosakovsky Pond, S. L., & Scheffler, K. (2013). FUBAR: A Fast, Unconstrained Bayesian AppRoximation for Inferring Selection. *Molecular Biology and Evolution*, *30*(5), 1196–1205. https://doi.org/10.1093/molbev/mst030

Murrell, B., Weaver, S., Smith, M. D., Wertheim, J. O., Murrell, S., Aylward, A., Eren, K., Pollner, T., Martin, D. P., Smith, D. M., Scheffler, K., & Kosakovsky Pond, S. L. (2015). Gene-Wide Identification of Episodic Selection. *Molecular Biology and Evolution*, *32*(5), 1365–1371. https://doi.org/10.1093/molbev/msv035

Ohm, R. A., Feau, N., Henrissat, B., Schoch, C. L., Horwitz, B. A., Barry, K. W., Condon, B. J., Copeland, A. C., Dhillon, B., Glaser, F., Hesse, C. N., Kosti, I., LaButti, K., Lindquist, E. A., Lucas, S., Salamov, A. A., Bradshaw, R. E., Ciuffetti, L., Hamelin, R. C., … Grigoriev, I. V. (2012). Diverse Lifestyles and Strategies of Plant Pathogenesis Encoded in the Genomes of Eighteen Dothideomycetes Fungi. *PLOS Pathogens*, *8*(12), e1003037. https://doi.org/10.1371/journal.ppat.1003037

Ökmen, B., Etalo, D. W., Joosten, M. H. A. J., Bouwmeester, H. J., de Vos, R. C. H., Collemare, J., & de Wit, P. J. G. M. (2013). Detoxification of α-tomatine by *Cladosporium fulvum* is required for full virulence on tomato. *New Phytologist*, *198*(4), 1203–1214. https://doi.org/10.1111/nph.12208

Oome, S., Raaymakers, T. M., Cabral, A., Samwel, S., Böhm, H., Albert, I., Nürnberger, T., & Van den Ackerveken, G. (2014). Nep1-like proteins from three kingdoms of life act as a microbe-associated molecular pattern in Arabidopsis. *Proceedings of the National Academy of Sciences*, *111*(47), 16955–16960. https://doi.org/10.1073/pnas.1410031111

Ou, S., & Jiang, N. (2018). LTR_retriever: A Highly Accurate and Sensitive Program for Identification of Long Terminal Repeat Retrotransposons. *Plant Physiology*, *176*(2), 1410–1422. https://doi.org/10.1104/pp.17.01310

Pareja-Jaime, Y., Roncero, M. I. G., & Ruiz-Roldán, M. C. (2008). Tomatinase from *Fusarium oxysporum* f. sp. *lycopersici* is required for full virulence on tomato plants. *Molecular plant-microbe interactions*, *21*(6), 728–736. https://doi.org/10.1094/MPMI-21-6-0728

Pereira, D., McDonald, B. A., & Croll, D. (2020). The genetic architecture of emerging fungicide resistance in populations of a global wheat pathogen. *bioRxiv*, 2020.03.26.010199. https://doi.org/10.1101/2020.03.26.010199

Pertea, M., Pertea, G. M., Antonescu, C. M., Chang, T.-c., Mendell, J. T., & Salzberg, S. L. (2015). StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. *Nature Biotechnology*, *33*(3), 290–295. https://doi.org/http://dx.doi.org.dbgw.lis.curtin.edu.au/10.1038/nbt.3122

Petersen, T. N., Brunak, S., Heijne, G. v., & Nielsen, H. (2011). SignalP 4.0: Discriminating signal peptides from transmembrane regions. *Nature Methods*, *8*(10), 785–786. https://doi.org/10.1038/nmeth.1701

Phan, H. T. T., Jones, D. A. B., Rybak, K., Dodhia, K. N., Lopez-Ruiz, F. J., Valade, R., Gout, L., Lebrun, M.-H., Brunner, P. C., Oliver, R. P., & Tan, K.-C. (2020). Low Amplitude Boom-and-Bust Cycles Define the Septoria Nodorum Blotch Interaction. *Frontiers in Plant Science*, *10*, 1785. https://doi.org/10.3389/fpls.2019.01785

Phan, H. T. T., Rybak, K., Bertazzoni, S., Furuki, E., Dinglasan, E., Hickey, L. T., Oliver, R. P., & Tan, K.-C. (2018). Novel sources of resistance to Septoria nodorum blotch in the Vavilov wheat collection identified by genome-wide association studies. *Theoretical and Applied Genetics*, *131*(6), 1223–1238. https://doi.org/10.1007/s00122-018-3073-y

Phan, H. T. T., Rybak, K., Furuki, E., Breen, S., Solomon, P. S., Oliver, R. P., & Tan, K.-C. (2016). Differential effector gene expression underpins epistasis in a plant fungal disease. *The Plant Journal*, *87*(4), 343–354. https://doi.org/10.1111/tpj.13203

Pollet, A., Beliën, T., Fierens, K., Delcour, J. A., & Courtin, C. M. (2009). *Fusarium graminearum* xylanases show different functional stabilities, substrate specificities and inhibition sensitivities. *Enzyme and Microbial Technology*, *44*(4), 189–195. https://doi.org/10.1016/j.enzmictec.2008.12.005

Pond, S. L. K., Frost, S. D. W., & Muse, S. V. (2005). HyPhy: Hypothesis testing using phylogenies. *Bioinformatics*, *21*(5), 676–679. https://doi.org/10.1093/bioinformatics/bti079

Poplin, R., Ruano-Rubio, V., DePristo, M. A., Fennell, T. J., Carneiro, M. O., Van der Auwera, G. A., Kling, D. E., Gauthier, L. D., Levy-Moonshine, A., Roazen, D., Shakir, K., Thibault, J., Chandran, S., Whelan, C., Lek, M., Gabriel, S., Daly, M. J., Neale, B., MacArthur, D. G., & Banks, E. (2018). Scaling accurate genetic variant discovery to tens of thousands of samples. *bioRxiv*, 201178. https://doi.org/10.1101/201178

Praz, C. R., Bourras, S., Zeng, F., Sánchez-Martín, J., Menardo, F., Xue, M., Yang, L., Roffler, S., Böni, R., Herren, G., McNally, K. E., Ben-David, R., Parlange, F., Oberhaensli, S., Flückiger, S., Schäfer, L. K., Wicker, T., Yu, D., & Keller, B. (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*(3), 1301–1314. https://doi.org/10.1111/nph.14372

Price, M. N., Dehal, P. S., & Arkin, A. P. (2010). FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLOS ONE*, *5*(3), e9490. https://doi.org/10.1371/journal.pone.0009490

Pritchard, J. K., Stephens, M., & Donnelly, P. (2000). Inference of Population Structure Using Multilocus Genotype Data. *Genetics*, *155*(2), 945–959. Retrieved August 20, 2020, from https://www.genetics.org/content/155/2/945

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M. A. R., Bender, D., Maller, J., Sklar, P., de Bakker, P. I. W., Daly, M. J., & Sham, P. C. (2007). PLINK: A Tool Set for Whole-Genome Association and Population-Based Linkage Analyses. *The American Journal of Human Genetics*, *81*(3), 559–575. https://doi.org/10.1086/519795

Quinlan, A. R., & Hall, I. M. (2010). BEDTools: A flexible suite of utilities for comparing genomic features. *Bioinformatics*, *26*(6), 841–842. https://doi.org/10.1093/bioinformatics/btq033

R Core Team. (2019). *R: A Language and Environment for Statistical Computing*. Vienna, Austria, R Foundation for Statistical Computing. https://www.R-project.org/

Richards, J. K., Stukenbrock, E. H., Carpenter, J., Liu, Z., Cowger, C., Faris, J. D., & Friesen, T. L. (2019). Local adaptation drives the diversification of effectors in the fungal wheat pathogen *Parastagonospora nodorum* in the United States. *PLOS Genetics*, *15*(10), e1008223. https://doi.org/10.1371/journal.pgen.1008223

Richards, J. K., Wyatt, N. A., Liu, Z., Faris, J. D., & Friesen, T. L. (2018). Reference Quality Genome Assemblies of Three *Parastagonospora nodorum* Isolates Differing in Virulence on Wheat. *G3: Genes, Genomes, Genetics*, *8*(2), 393–399. https://doi.org/10.1534/g3.117.300462

Rognes, T., Flouri, T., Nichols, B., Quince, C., & Mahé, F. (2016). VSEARCH: A versatile open source tool for metagenomics. *PeerJ*, *4*, e2584. https://doi.org/10.7717/peerj.2584

Sanger, F., Coulson, A. R., Friedmann, T., Air, G. M., Barrell, B. G., Brown, N. L., Fiddes, J. C., Hutchison, C. A., Slocombe, P. M., & Smith, M. (1978). The nucleotide sequence of bacteriophage φX174. *Journal of Molecular Biology*, *125*(2), 225–246. https://doi.org/10.1016/0022-2836(78)90346-7

Sanz-Martín, J. M., Pacheco-Arjona, J. R., Bello-Rico, V., Vargas, W. A., Monod, M., Díaz-Mínguez, J. M., Thon, M. R., & Sukno, S. A. (2016). A highly conserved metalloprotease effector enhances virulence in the maize anthracnose fungus *Colletotrichum graminicola*. *Molecular Plant Pathology*, *17*(7), 1048–1062. https://doi.org/10.1111/mpp.12347

Savojardo, C., Martelli, P. L., Fariselli, P., & Casadio, R. (2018). DeepSig: Deep learning improves signal peptide detection in proteins. *Bioinformatics*, *34*(10), 1690–1696. https://doi.org/10.1093/bioinformatics/btx818

Schmidt, S. M., Kuhn, H., Micali, C., Liller, C., Kwaaitaal, M., & Panstruga, R. (2014). Interaction of a *Blumeria graminis* f. sp. *hordei* effector candidate with a barley ARF-GAP suggests that host vesicle trafficking is a fungal pathogenicity target: *Blumeria graminis* effector candidates. *Molecular Plant Pathology*, *15*(6), 535–549. https://doi.org/10.1111/mpp.12110

Sharpee, W., Oh, Y., Yi, M., Franck, W., Eyre, A., Okagaki, L. H., Valent, B., & Dean, R. A. (2017). Identification and characterization of suppressors of plant cell death (SPD) effectors from *Magnaporthe oryzae*. *Molecular Plant Pathology*, *18*(6), 850–863. https://doi.org/10.1111/mpp.12449

Shaw, M. W., Bearchell, S. J., Fitt, B. D. L., & Fraaije, B. A. (2008). Long-term relationships between environment and abundance in wheat of *Phaeosphaeria nodorum* and *Mycosphaerella graminicola*. *New Phytologist*, *177*(1), 229–238. https://doi.org/10.1111/j.1469-8137.2007.02236.x

Shi, G., Friesen, T. L., Saini, J., Xu, S. S., Rasmussen, J. B., & Faris, J. D. (2015). The Wheat *Snn7* Gene Confers Susceptibility on Recognition of the *Parastagonospora nodorum* Necrotrophic Effector SnTox7. *The Plant Genome*, *8*(2), plantgenome2015.02.0007. https://doi.org/10.3835/plantgenome2015.02.0007

Slater, G. S. C., & Birney, E. (2005). Automated generation of heuristics for biological sequence comparison. *BMC Bioinformatics*, *6*, 31–31. https://doi.org/10.1186/1471-2105-6-31

Solomon, P. S., Lowe, R. G. T., Tan, K.-C., Waters, O. D. C., & Oliver, R. P. (2006). *Stagonospora nodorum*: Cause of stagonospora nodorum blotch of wheat. *Molecular Plant Pathology*, *7*(3), 147–156. https://doi.org/10.1111/j.1364-3703.2006.00326.x

Sommerhalder, R. J., McDonald, B. A., & Zhan, J. (2006). The Frequencies and Spatial Distribution of Mating Types in *Stagonospora nodorum* Are Consistent with Recurring Sexual Reproduction. *Phytopathology*, *96*(3), 234–239. https://doi.org/10.1094/PHYTO-96-0234

Sperschneider, J., Catanzariti, A.-M., DeBoer, K., Petre, B., Gardiner, D. M., Singh, K. B., Dodds, P. N., & Taylor, J. M. (2017). LOCALIZER: Subcellular localization prediction of both plant and effector proteins in the plant cell. *Scientific Reports*, *7*(1), 1–14. https://doi.org/10.1038/srep44598

Sperschneider, J., Dodds, P. N., Gardiner, D. M., Singh, K. B., & Taylor, J. M. (2018). Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. *Molecular Plant Pathology*, *19*(9), 2094–2110. https://doi.org/10.1111/mpp.12682

Sperschneider, J., Dodds, P. N., Singh, K. B., & Taylor, J. M. (2018). ApoplastP: Prediction of effectors and plant proteins in the apoplast using machine learning. *New Phytologist*, *217*(4), 1764–1778. https://doi.org/10.1111/nph.14946

Sperschneider, J., Gardiner, D. M., Dodds, P. N., Tini, F., Covarelli, L., Singh, K. B., Manners, J. M., & Taylor, J. M. (2016). EffectorP: Predicting fungal effector proteins from secretomes using machine learning. *New Phytologist*, *210*(2), 743–761. https://doi.org/10.1111/nph.13794

Sperschneider, J., Gardiner, D. M., Thatcher, L. F., Lyons, R., Singh, K. B., Manners, J. M., & Taylor, J. M. (2015). Genome-Wide Analysis in Three Fusarium Pathogens Identifies Rapidly Evolving Chromosomes and Genes Associated with Pathogenicity. *Genome Biology and Evolution*, *7*(6), 1613–1627. https://doi.org/10.1093/gbe/evv092

Standage, D. S., & Brendel, V. P. (2012). ParsEval: Parallel comparison and analysis of gene structure annotations. *BMC Bioinformatics*, *13*(1), 187. https://doi.org/10.1186/1471-2105-13-187

Stanke, M., Diekhans, M., Baertsch, R., & Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics*, *24*(5), 637–644. https://doi.org/10.1093/bioinformatics/btn013

Steinbiss, S., Willhoeft, U., Gremme, G., & Kurtz, S. (2009). Fine-grained annotation and classification of de novo predicted LTR retrotransposons. *Nucleic Acids Research*, *37*(21), 7002–7013. https://doi.org/10.1093/nar/gkp759

Steinegger, M., & Söding, J. (2017). MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nature Biotechnology*, *35*, 1026–1028. https://doi.org/10.1038/nbt.3988

Stergiopoulos, I., Kourmpetis, Y. A. I., Slot, J. C., Bakker, F. T., De Wit, P. J. G. M., & Rokas, A. (2012). In Silico Characterization and Molecular Evolutionary Analysis of a Novel Superfamily of Fungal Effector Proteins. *Molecular Biology and Evolution*, *29*(11), 3371–3384. https://doi.org/10.1093/molbev/mss143

Stukenbrock, E. H., Banke, S., & McDonald, B. A. (2006). Global migration patterns in the fungal wheat pathogen *Phaeosphaeria nodorum*. *Molecular Ecology*, *15*(10), 2895–2904. https://doi.org/10.1111/j.1365-294X.2006.02986.x

Syme, R. A., Tan, K.-C., Hane, J. K., Dodhia, K., Stoll, T., Hastie, M., Furuki, E., Ellwood, S. R., Williams, A. H., Tan, Y.-F., Testa, A. C., Gorman, J. J., & Oliver, R. P. (2016). Comprehensive Annotation of the *Parastagonospora nodorum* Reference Genome Using Next-Generation Genomics, Transcriptomics and Proteogenomics. *PLOS ONE*, *11*(2), e0147221. https://doi.org/10.1371/journal.pone.0147221

Syme, R. A., Tan, K.-C., Rybak, K., Friesen, T. L., McDonald, B. A., Oliver, R. P., & Hane, J. K. (2018). Pan-Parastagonospora Comparative Genome Analysis—Effector Prediction and Genome Evolution. *Genome Biology and Evolution*, *10*(9), 2443–2457. https://doi.org/10.1093/gbe/evy192

Tan, K.-C., Oliver, R. P., Solomon, P. S., & Moffat, C. S. (2010). Proteinaceous necrotrophic effectors in fungal virulence. *Functional Plant Biology*, *37*(10), 907–912. https://doi.org/10.1071/FP10067

Tan, K.-C., Waters, O. D. C., Rybak, K., Antoni, E., Furuki, E., & Oliver, R. P. (2014). Sensitivity to three *Parastagonospora nodorum* necrotrophic effectors in current Australian wheat cultivars

and the presence of further fungal effectors. *Crop and Pasture Science*, *65*(2), 150–158. https://doi.org/10.1071/CP13443

Testa, A. C., Hane, J. K., Ellwood, S. R., & Oliver, R. P. (2015). CodingQuarry: Highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics*, *16*(1), 170. https://doi.org/10.1186/s12864-015-1344-4

Testa, A. C., Oliver, R. P., & Hane, J. K. (2016). OcculterCut: A Comprehensive Survey of AT-Rich Regions in Fungal Genomes. *Genome Biology and Evolution*, *8*(6), 2044–2064. https://doi.org/10.1093/gbe/evw121

Thioulouse, J., Chessel, D., Dole´dec, S., & Olivier, J.-M. (1997). ADE-4: A multivariate analysis and graphical display software. *Statistics and Computing*, *7*(1), 75–83. https://doi.org/10.1023/A:1018513530268

Thrall, P. H., Barrett, L. G., Dodds, P. N., & Burdon, J. J. (2016). Epidemiological and Evolutionary Outcomes in Gene-for-Gene and Matching Allele Models. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.01084

Trainor, G., Zaicou-Kunesch, C., Curry, J., Shackley, B., & Nicol, D. (2018). 2019 Wheat variety sowing guide for Western Australia. Department of Primary Industries; Regional Development. Retrieved September 29, 2020, from https://www.agric.wa.gov.au/sites/gateway/files/2019%20Wheat%20Variety%20Guide-web.pdf

Urban, M., Cuzick, A., Seager, J., Wood, V., Rutherford, K., Venkatesh, S. Y., De Silva, N., Martinez, M. C., Pedro, H., Yates, A. D., Hassani-Pak, K., & Hammond-Kosack, K. E. (2020). PHI-base: The pathogen–host interactions database. *Nucleic Acids Research*, *48*(D1), D613–D620. https://doi.org/10.1093/nar/gkz904

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Jarrod Millman, K., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … Contributors, S. 1. 0. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/https://doi.org/10.1038/s41592-019-0686-2

Vleeshouwers, V. G. A. A., & Oliver, R. P. (2014). Effectors as Tools in Disease Resistance Breeding Against Biotrophic, Hemibiotrophic, and Necrotrophic Plant Pathogens. *Molecular Plant-Microbe Interactions*, *27*(3), 196–206. https://doi.org/10.1094/MPMI-10-13-0313-IA

Voigt, C. A., Schäfer, W., & Salomon, S. (2005). A secreted lipase of *Fusarium graminearum* is a virulence factor required for infection of cereals. *The Plant Journal*, *42*(3), 364–375. https://doi.org/10.1111/j.1365-313X.2005.02377.x

Wang, Y., Wu, J., Kim, S. G., Tsuda, K., Gupta, R., Park, S.-Y., Kim, S. T., & Kang, K. Y. (2016). *Magnaporthe oryzae*-Secreted Protein MSP1 Induces Cell Death and Elicits Defense Responses in Rice. *Molecular Plant-Microbe Interactions*, *29*(4), 299–312. https://doi.org/10.1094/MPMI-12-15-0266-R

Waterhouse, R. M., Seppey, M., Simão, F. A., Manni, M., Ioannidis, P., Klioutchnikov, G., Kriventseva, E. V., & Zdobnov, E. M. (2018). BUSCO Applications from Quality Assessments to Gene Prediction and Phylogenomics. *Molecular Biology and Evolution*, *35*(3), 543–548. https://doi.org/10.1093/molbev/msx319

Wickham, H. (2016). *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org

Wood, D. E., Lu, J., & Langmead, B. (2019). Improved metagenomic analysis with Kraken 2. *Genome Biology*, *20*(1), 257. https://doi.org/10.1186/s13059-019-1891-0

Wood, D. E., & Salzberg, S. L. (2014). Kraken: Ultrafast metagenomic sequence classification using exact alignments. *Genome Biology*, *15*(3), R46. https://doi.org/10.1186/gb-2014-15-3-r46

Wright, E. S. (2015). DECIPHER: Harnessing local sequence context to improve protein multiple sequence alignment. *BMC Bioinformatics*, *16*(1), 322. https://doi.org/10.1186/s12859-015-0749-z

Wu, T. D., & Watanabe, C. K. (2005). GMAP: A genomic mapping and alignment program for mRNA and EST sequences. *Bioinformatics*, *21*(9), 1859–1875. https://doi.org/10.1093/bioinformatics/bti310

Xin, Z., & Chen, J. (2012). A high throughput DNA extraction method with high yield and quality. *Plant Methods*, *8*(1), 26. https://doi.org/10.1186/1746-4811-8-26

Xu, G., Zhong, X., Shi, Y., Liu, Z., Jiang, N., Liu, J., Ding, B., Li, Z., Kang, H., Ning, Y., Liu, W., Guo, Z., Wang, G.-L., & Wang, X. (2020). A fungal effector targets a heat shock–dynamin protein complex to modulate mitochondrial dynamics and reduce plant immunity. *Science Advances*, *6*(48), eabb7719. https://doi.org/10.1126/sciadv.abb7719

Xu, Y., Zhang, Y., Zhu, J., Sun, Y., Guo, B., Liu, F., Huang, J., Wang, H., Dong, S., Wang, Y., & Wang, Y. (2021). *Phytophthora sojae* apoplastic effector AEP1 mediates sugar uptake by mutarotation of extracellular aldose and is recognized as a MAMP. *Plant Physiology*. https://doi.org/10.1093/plphys/kiab239

Yoshino, K., Irieda, H., Sugimoto, F., Yoshioka, H., Okuno, T., & Takano, Y. (2012). Cell Death of *Nicotiana benthamiana* Is Induced by Secreted Protein NIS1 of *Colletotrichum orbiculare* and Is Suppressed by a Homologue of CgDN3. *Molecular Plant-Microbe Interactions*, *25*(5), 625–636. https://doi.org/10.1094/MPMI-12-11-0316

Yu, G., Smith, D. K., Zhu, H., Guan, Y., & Lam, T. T.-Y. (2017). Ggtree: An R package for visualization and annotation of phylogenetic trees with their covariates and other associated data (G. McInerny, Ed.). *Methods in Ecology and Evolution*, *8*(1), 28–36. https://doi.org/10.1111/2041-210X.12628

Zhang, H., Yohe, T., Huang, L., Entwistle, S., Wu, P., Yang, Z., Busk, P. K., Xu, Y., & Yin, Y. (2018). dbCAN2: A meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, *46*(W1), W95–W101. https://doi.org/10.1093/nar/gky418

Zhang, Y., Xu, K., Yu, D., Liu, Z., Peng, C., Li, X., Zhang, J., Dong, Y., Zhang, Y., Tian, P., Guo, T., & Li, C. (2019). The Highly Conserved Barley Powdery Mildew Effector BEC1019 Confers Susceptibility to Biotrophic and Necrotrophic Pathogens in Wheat. *International Journal of Molecular Sciences*, *20*(18), 4376. https://doi.org/10.3390/ijms20184376

Zhang, Z., Friesen, T. L., Xu, S. S., Shi, G., Liu, Z., Rasmussen, J. B., & Faris, J. D. (2011). Two putatively homoeologous wheat genes mediate recognition of SnTox3 to confer effector-triggered susceptibility to *Stagonospora nodorum*. *The Plant Journal*, *65*(1), 27–38. https://doi.org/10.1111/j.1365-313X.2010.04407.x

CHAPTER 12

Conclusion

This thesis aimed to develop and apply technical solutions to assist large scale fungal genomics sequencing projects, with a particular focus on determining virulence factors and the structure of natural populations of *Parastagonospora nodorum*. The project was organised under three broad research themes: 1. Developing tools to predict fungal habits and virulence related proteins, 2. Investigating spatial and temporal changes a population of *P. nodorum*, and 3. Analysing the genomes of a *P. nodorum* population.

12.1 Theme 1. Developing tools to predict fungal habits and virulence related proteins

A major theme of this thesis was the development of computational tools to predict effector genes and efficiently process and annotate a large number of genomes. In chapter 3 we developed an effector prediction pipeline, Predector, which performs numerous secretion and protein property prediction analyses and combines the results using a novel learn-to-rank machine learning method. Protein secretion and property based effector prediction is an important early step in pathogen genomics research, and is often one of the first analyses performed after gene prediction (Sperschneider et al., 2015). However, despite its ubiquity in the research community there is little agreement on the best combination of tools, and no published pipelines that combine secretion and effector prediction as a unified analysis. We showed that by explicitly combining secretion and effector prediction, Predector outperforms the commonly used hard filter of a single secretion prediction tool and EffectorP (Sperschneider et al., 2018; Sperschneider et al., 2016) in terms of both classification and ranking. We also demonstrate that ranking approaches should be preferred over the application of successive hard filters, where high confidence candidates will be nearer the top of the list, but less confident candidates (which might have been excluded by hard-filters) are still separated from the bulk of proteins in the sorted list. Furthermore, Predector automates and optimises the installation and running of numerous software, which saves considerable time. Overall, we think that Predector will be a considerable asset to the plant-pathology research community, saving time and providing consistency, and hope to continue its development as new requirements and analyses emerge.

In chapter 4 we investigated the application of sensitive sequence search techniques and clustering to identify remotely homologous families of fungal effector proteins, and the potential relationships between them. With the increasing rate of fungal genome sequencing and effector discovery, several effector families have been described which generally share little sequence identity but are structurally similar (Wirthmueller et al., 2013). These families include the MAX (de Guillen et al., 2015), ToxA-like (Lu et al., 2015; Schmidt et al., 2016), and RALPH (Praz et al., 2017; Spanu, 2017) effectors. The method developed here, RemEff, uses iteratively increasing sensitivity sequence search and graph clustering techniques, to identify clusters of proteins with low sequence similarity but possibly sharing structural features or an ancient evolutionary history. In addition to identifying numerous homologues of known effectors,

we identified 26 clusters containing two or more known effectors, suggesting a shared history. Most significantly we identified that the previously described AvrLm6-like family (Shiller et al., 2015) may form part of a larger family including the effectors FolSIX5 (Lievens et al., 2009), MoBas4 (Mosquera et al., 2009), MoSPD5 (Sharpee et al., 2017), and CbNip1 (Ebert, 2018), which may have a common function associated with host membrane binding. We also identified that the *Blumeria graminis* effector AvrPm3$^{a2/f2}$ is also a possible member of the RALPH family (Praz et al., 2017; Spanu, 2017), along with a protein that suppresses its activity SvrPm3$^{a1/f1}$. Other previously described relationships between ToxA and AvrFom2 (Schmidt et al., 2016), or the MAX family of effectors (de Guillen et al., 2015) were not identified in this analysis. Because of the large memory and processing time requirements involved in clustering a dataset of this size, and the potential for false positive alignments, we filtered hidden markov model (HMM)-HMM alignments to have relatively strict e-value and coverage requirements. In contrast, ToxA and MAX homology relationships were identified previously by manual inspection of PSI-BLAST and HMMER sequence matches, as well as tertiary structural determination. Future studies may be able to identify these relationships by relaxing sequence alignment criteria and using memory optimised clustering methods. Alternatively, for the purposes of maintaining a database of effector HMMs an effector centric cluster seeding approach similar to what is done by Pfam could be used (Finn et al., 2014). The possibility of identifying effectors by sensitive search techniques is significant in our understanding. Effector discovery is a difficult process and the lack of usefulness of homology based prediction methods is often cited as a leading hindrance in the field (D. A. Jones et al., 2018). As more genomes become sequenced, with increased resolution of closely related species, new effector families are likely to be identified. We anticipate that tailor-made databases and sensitive sequence search techniques, such as those used and provided by RemEff, will become an important step in future effector discovery efforts.

In chapter *5*, we re-evaluated the current "trophic" classification system, and developed a pipeline to annotate fungal genomes using this updated system based on carbohydrate active enzymes. The more granular understanding of pathogen infection strategies has highlighted the arbitrary nature of the classical symbiont, biotroph, hemibiotroph, and necrotroph classification system. Most necrotrophs have some latent phase where the pathogen attempts to evade recognition, which might be characterised as biotrophic behaviour. The distinction between symbiosis and biotrophy is similarly fraught as they are both arguably degrees of pathogens, and the distinction may simply come from the "decision" of their plant hosts. By focussing on the enzymes releasing the primary carbon source used by fungal pathogens (i.e. carbohydrate-active enzymes (CAZymes)), the CATAStrophy method offers a useful insight into the ways that plant associated fungi have co-evolved with their hosts and how they obtain food. This may be particularly useful when descriptive categories conflict. For example if a pathogen typically described as a biotroph was found to have a suite of CAZymes more similar to a hemibiotroph (mesotroph) or necrotroph (polymertroph) we might question our current understanding of how the pathogen operates or the evolutionary history of the pathogen.

Previously, a relationship between pathogen lifestyle and repeat-induced point mutation (RIP) has also been observed, where plant pathogens with a non-obligate biotrophic phase were more often found to have a higher proportion of AT-rich genomic regions (indicative of historic RIP activity) (Testa et al., 2016). By assessing pathogen lifestyle through the lens of genome adaptation, using indicators such as CAZyme content or AT-richness, we can ask questions of the adaptive and mutational processes that shape pathogens with common habit. We see this updated classification system as a first step towards reshaping how pathologists categorise pathogens in the genomics age.

In addition to these effector and trophic prediction tools, several pipelines were developed in chapter 11 to support the assembly and annotation of a large number of *P. nodorum* genomes. Most notably two pipelines for annotating genes (PanAnn) and transposable elements (TEs) (PanTE) were developed and focussed on highly sensitive prediction and transferring annotations across genomes. Incorporating information from gene annotations of other isolates is often overlooked in fungal genomics, or performed in an ad-hoc manner after orthologue comparative studies, as missing orthologues can often be found in genomes using translated search tools such as tBLASTn (Deng et al., 2017). Failure to identify orthologues in specific genomes may result in incorrect conclusions about candidate genes for a phenotype of interest, for example differential virulence on a host. In the context of effector discovery, sensitive gene prediction methods are extremely important as an unpredicted gene can never be considered for effector candidacy, and accurately annotating effector genes has historically been difficult (Cook et al., 2019; Testa et al., 2015). Similarly, identifying candidate effectors from presence-absence profiles generated from an incomplete orthology analysis might miss genuine candidates or include bad candidates. This has consequences for numerous downstream analyses and experiments, such as candidate association with a phenotype of interest. Therefore, it is important to establish an accurate set of gene predictions and infer accurate orthology relationships from the outset. We observed that including comparative steps in genome annotation pipelines enhanced consistency and sensitivity, and should become a standard practice in the pathogen genomics community, particularly as more pangenomics studies are undertaken.

12.2 Theme 2. Spatial survey and pattern detection

A second theme of this thesis is investigating the trends change in pathogen populations over time and geographic space, focussing on a Western Australian (WA) *P. nodorum* population. Local variability of WA pathogen populations had been largely ignored, with population level analyses of *P. nodorum* investigating global variability or European or U. S. A. populations, including few Australian isolates or sampling locations (Ghaderi et al., 2020; Lin et al., 2020; Richards et al., 2019; Stukenbrock et al., 2006; Syme et al., 2018).

In chapter 8 we presented the first investigation of a large WA *P. nodorum* population using short sequence repeat (SSR) markers. We found evidence of a diverse population of *P. nodorum*

in WA. Two clusters of isolates comprised the majority of the population, which appeared to correspond to a gradual shift in allele frequencies over time. We speculated that this shift between the two core clusters over time, occurring most starkly after 2013, may be in response to cultivar usage preference particularly the widespread adoption of a ToxA insensitive wheat variety. However, this shift also coincided with increased *P. nodorum* sampling efforts in 2014, which may have created artificial structure in the data. In addition to the two core clusters, we identified 3 satellite clusters of isolates which were geographically restricted. Most of these populations were collected from northern regions of the WA wheat growing areas, around Mingenew and Geraldton, and isolates are highly similar within clusters. We proposed that these clusters represent short term clonal or near clonal expansions, and suggest that there might be specific environmental characteristics of hotter and drier regions that promote clonal population expansion. Notably, one of these clusters (group 3) spanned a sampling period of 31 years, which suggests that clonal or near clonal lineages may persist for a long time.

In chapter 11, we built on the initial study from 8 by performing whole genome sequencing of isolates in the WA *P. nodorum* population. Population clusters predicted using single nucleotide polymorphism (SNP) data were similar to those presented in chapter 8, with notable differences. We observed that the core population was present as a single cluster rather than two as previously found, and that there were 4 additional satellite clusters of closely related isolates from WA. A phylogeny estimated from SNPs showed that the satellite clusters corresponded to distinct clades in the tree and members were genetically similar, with the core cluster consisting of a heterogeneous group of distantly related individuals which could not be grouped further. We suggested that rather than representing distinct reproductively isolated populations, these satellite clusters are probably related individuals which are similar because of clonality, or common sampling time or location. As in chapter 10, the cluster of highly similar isolates sampled between 1980 and 2011 (group 3) was also observed in chapter 8 (cluster 6), despite the larger number of markers. The isolates within the core WA and international clusters are highly dissimilar, which suggested that there was enough information to differentiate them from the other clusters, but not enough to divide them further. If we exclude the nearly clonal clusters from consideration, as they may not represent true sub-populations, these data suggested that the WA population has no distinct structure and that there are no major barriers to movement of isolates or gene flow. Analyses of linkage disequilibrium suggested that the core and international clusters had higher than expected levels of linkage disequilibrium, which is indicative of clonality. We note however that the standardised indices $\bar{r}_d$ were very low, and hypothesise that a mixture of sexual and asexual reproduction is occurring in these populations. All other clusters were highly similar, and so even without significantly high $\bar{r}_d$ indices we expect that these clusters have emerged as a result of clonal reproduction or a very recently emerged lineage of isolates.

In both studies presented in chapters 8 and 11, the conclusions we could draw about the population structure and dynamics were limited by the sampling scheme and available metadata. There was no consistency of sampling locations between years, and the sampling locations

were heavily biased to only two growing regions (Northam and Geraldton). Additionally, precise locational information, crop history, paddock blocking factors, or sampling dates were generally incomplete, necessitating the restriction of covariate analyses to only consider year and sampling region. These issues are a consequence of the opportunistic sampling scheme employed for numerous purposes rather than being an experimental oversight. These isolates were collected from a variety of sources, but the majority were collected as part of a fungicide resistance monitoring programme. Future studies would benefit from a more rigorous sampling scheme and experimental design, dedicated to addressing questions about disease monitoring, epidemiology, and population structure. Nevertheless, these two studies offer a valuable early insight into the WA population, and suggest that further investment in this area is warranted. The observation of small expanded clusters of isolates, particularly in hotter and drier regions, is a significant step understanding the short-term epidemiology of *P. nodorum*, which appears to involve both sexual and asexual reproduction. Some of these expanded near-clonal clusters are highly virulent toward modern wheat cultivars, and an understanding of their emergence may be critical to managing future epidemics. The lack of geographic barriers to reproduction observed suggests long-range spore dispersal or human mediated movement, of which the latter could potentially be managed alongside monitoring programs. Similarly, regional knowledge of a local population's effector content, general virulence, or fungicide resistance might inform growers in selecting effective resistant crop cultivars or fungicide modes of action, which would be of considerable value to the wheat industry. Together, these analyses provide an important first step toward understanding the WA *P. nodorum* population, and investigating the practical consequences of local population characteristics.

12.3 Theme 3. Analysing the genomes of a *P. nodorum* population.

The final aim of this project was to perform an in-depth analysis of the Western Australian (WA) *P. nodorum* pangenome. In chapter 11 we assembled, comprehensively annotated, and compared the genomes of the WA *P. nodorum* population and a number of previously published *P. nodorum* genomes from Europe, the USA, and Iran (Richards et al., 2018; Syme et al., 2018). Alignment of the assembled genomes to the SN15 reference isolate identified a single WA isolate that was lacking accessory chromosome (AC) 23, and a cluster of WA isolates missing an arm of AC 23. Single nucleotide polymorphism (SNP) density observed to be uneven across the genomes, indicating the presence of "mutation hotspots" which have long been hypothesised to drive gene diversification and pathogen adaptability. Repeat-induced point mutation (RIP)-like dinucleotide changes contributed the largest proportion of mutations present in the population, and the genomic regions exhibiting high frequency of RIP-like mutation were different between groups of isolates, indicating that RIP has been active relatively recently in the WA population. Ortho-group analysis (encompassing both orthologues and paralogues) indicated the presence of a large pangenome, with 14098 core orthogroups and 11460 accessory

groups. This is in contrast with most previous *P. nodorum* gene annotations, which had only 13569 genes (Syme et al., 2016). These updated gene numbers indicate that the accessory genomes of individual isolates may be significant, and supports the SN15 annotations presented by Bertazzoni et al. (2021, chapter 10) which contained 16403 predicted genes. Although this study transferred all of those gene predictions across to new isolates, the consistently large number of genes predicted as a result of multiple annotation methods and annotation transfer suggests that the true number of genes present in *P. nodorum* may be larger than previously thought. These increased gene numbers may also be attributable to the use of deep RNA sequencing data from *P. nodorum* to support gene annotation software and curation, which included transcripts extracted from both *in vitro* culture and plant infection treatments (D. A. B. Jones et al., 2019, chapter 9). Although there are likely to be numerous false positive gene predictions, we prioritised sensitivity over specificity for fungal pathogen annotation as genes of interest are often difficult to predict (Testa et al., 2015) and because effector candidates can only be selected from the pool of protein coding predictions. We identified numerous predicted proteins with effector-like properties using the Predector pipeline developed in chapter 3. Notably, this included several homologues of the known necrotrophic effectors Tox1 and Tox3, as well as effectors from other pathogens such as *Magnaporthe oryzae* AVR-Pita (Chuma et al., 2011; Dai et al., 2010) and *Zymoseptoria tritici* Zt6 (Kettles et al., 2018). Effector candidates were enriched in RIP affected regions and the accessory pangenome, and among orthogroups subject to positive selection pressure. This suggests that this population of *P. nodorum* isolates is actively mutating and potentially adapting to hosts via diversification of virulence factors. Further, the enrichment of effector candidates with a high ratio of RIP-like mutations relative to transitions in isolate SN15, suggests that RIP may have a role in diversification of effectors. However, we did not observe any relationship between orthogroups with a high ratio of RIP-like mutations and orthogroups under positive selection. This may be because RIP mutations causing amino-acid changes will rarely have beneficial effects on gene function, and so any potentially beneficial RIP mutations are vastly outnumbered by those with neutral or negative effects on gene function. This may also hint at the differences in selection pressure between necrotrophic and avirulence effectors observed by Gervais et al. (2017), where necrotrophic effectors were present in the same genomic regions as the "core" genes and not in the RIP-affected AT-rich regions alongside the avirulence (Avr) effectors. Necrotrophs benefit from effector and virulence factor diversification by their potential increased protein activity or interaction with host susceptibility factors, rather than by avoiding recognition as is the case for avirulence effectors in biotrophs. This higher requirement for mutation benefit and selection against pseudo-genised necrotrophic effectors may explain some differences in the contribution of RIP to the genome evolution between biotrophs and necrotrophs.

Overall, the first pangenomic analysis of the WA *P. nodorum* population has enhanced our understanding of the genome dynamics of *P. nodorum*, and identified numerous strong effector candidates for future investigation. The observation of a large number of short variants, many of which carry a RIP-like dinucleotide pattern, suggests that the population is highly mutable

which means that any deployed control measures exerting a strong selective pressure may be quickly overcome by extant or novel variants in the population. We also identified several homologues of the known necrotrophic effectors Tox1 and Tox3, which may have duplicated and mutated to form distinct pathogenicity factors as observed in other pathogens (de Guillen et al., 2015; Praz et al., 2017; Spanu, 2017). The number of publicly available fungal protein sequences and sensitivity of sequence search techniques has increased considerably over the past decade, and the observation of numerous effector homologues in this pangenome, including some effectors from other pathogen species, is suggestive of a shift towards homology being a reliable means of effector identification. Effector databases like RemEff (chapter 4) and sensitive sequence search techniques are now an important tool in annotation fungal pathogen genomes. We expect that these analyses and the resources developed for them will be an asset to the *P. nodorum* research community and the WA wheat industry.

12.4 Overall Significance

During this thesis, we have developed high-throughput computational solutions to annotate fungal pathogen genomes and investigate the nature of pathogen populations, pathogenicity, and effector proteins. The Predector pipeline described in chapter 3 combines several existing common analyses and adds several new analyses (e.g. a curated effector sequence database) in bioinformatic effector prediction. This unified pipeline will greatly simplify future genomics studies, and enhance the sensitivity and reproducibility of effector candidate generation and analyses across multiple pathogen species. The concepts of remote homology and sensitive sequence comparison techniques presented in chapter 4 continues an important conceptual shift in the effector research community, from the traditional view of narrow lineage specific fungal effectors towards larger more conserved families that have diversified considerably. The resulting database of "remote-homologue" effector families will be a useful tool for effector discovery, and may give rise to new ways of effector categorisation. The investigations of the Western Australian (WA) *Parastagonospora nodorum* population presented in chapter 8 and chapter 11 offer opportunities and challenges to the Australian wheat growing community in managing disease. *P. nodorum*'s high population diversity and demonstrated potential to rapidly expand in favourable niches poses a high risk of epidemic outbreak, particularly in hotter and drier regions. The methods used in pangenomic comparisons have highlighted that genomic regions undergoing rapid mutation and/or containing genes with functions important to pathogenicity may have genuine roles in pathogen adaptation to environments and hosts. The further dissection of pathogen population and genome dynamics will likely be an ongoing research topic. Finally, the large fully annotated pan-genome established in chapter 11, one of the first for fungal pathogens, is a considerable asset to the *P. nodorum* research community and the pathogen genomics community at large. Together, this thesis has meaningfully contributed to the collective understanding of how fungal pathogens operate, particularly *P. nodorum*, from the molecular level to the population level.

12.5 References

Bertazzoni, S., Jones, D. A. B., Phan, H. T., Tan, K.-C., & Hane, J. K. (2021). Chromosome-level genome assembly and manually-curated proteome of model necrotroph *Parastagonospora nodorum* sn15 reveals a genome-wide trove of candidate effector homologs, and redundancy of virulence-related functions within an accessory chromosome. *BMC Genomics*, *22*(382). https://doi.org/10.1186/s12864-021-07699-8

Chuma, I., Isobe, C., Hotta, Y., Ibaragi, K., Futamata, N., Kusaba, M., Yoshida, K., Terauchi, R., Fujita, Y., Nakayashiki, H., Valent, B., & Tosa, Y. (2011). Multiple Translocation of the *AVR-Pita* Effector Gene among Chromosomes of the Rice Blast Fungus *Magnaporthe oryzae* and Related Species. *PLOS Pathogens*, *7*(7), e1002147. https://doi.org/10.1371/journal.ppat.1002147

Cook, D. E., Valle-Inclan, J. E., Pajoro, A., Rovenich, H., Thomma, B. P., & Faino, L. (2019). Long-read annotation: Automated eukaryotic genome annotation based on long-read cdna sequencing. *Plant Physiology*, *179*(1), 38–54. https://doi.org/10.1104/pp.18.00848

Dai, Y., Jia, Y., Correll, J., Wang, X., & Wang, Y. (2010). Diversification and evolution of the avirulence gene *AVR-Pita1* in field isolates of *Magnaporthe oryzae*. *Fungal Genetics and Biology*, *47*(12), 973–980. https://doi.org/10.1016/j.fgb.2010.08.003

de Guillen, K. d., Ortiz-Vallejo, D., Gracy, J., Fournier, E., Kroj, T., & Padilla, A. (2015). Structure Analysis Uncovers a Highly Diverse but Structurally Conserved Effector Family in Phytopathogenic Fungi. *PLOS Pathogens*, *11*(10), e1005228. https://doi.org/10.1371/journal.ppat.1005228

Deng, C. H., Plummer, K. M., Jones, D. A. B., Mesarich, C. H., Shiller, J., Taranto, A. P., Robinson, A. J., Kastner, P., Hall, N. E., Templeton, M. D., & Bowen, J. K. (2017). Comparative analysis of the predicted secretomes of Rosaceae scab pathogens *Venturia inaequalis* and *V. pirina* reveals expanded effector families and putative determinants of host range. *BMC Genomics*, *18*(1), 339. https://doi.org/10.1186/s12864-017-3699-1

Ebert, M. K. (2018). *Effector biology of the sugar beet pathogen Cercospora beticola* (Doctoral dissertation). Wageningen University. https://doi.org/10.18174/453825

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J., & Punta, M. (2014). Pfam: The protein families database. *Nucleic Acids Research*, *42*(Database issue), D222–D230. https://doi.org/10.1093/nar/gkt1223

Gervais, J., Plissonneau, C., Linglin, J., Meyer, M., Labadie, K., Cruaud, C., Fudal, I., Rouxel, T., & Balesdent, M.-H. (2017). Different waves of effector genes with contrasted genomic location are expressed by *Leptosphaeria maculans* during cotyledon and stem colonization of oilseed rape. *Molecular Plant Pathology*, *18*(8), 1113–1126. https://doi.org/10.1111/mpp.12464

Ghaderi, F., Sharifnabi, B., Javan-Nikkhah, M., Brunner, P. C., & McDonald, B. A. (2020). *SnToxA* , *SnTox1* , and *SnTox3* originated in *Parastagonospora nodorum* in the Fertile Crescent. *Plant Pathology*, ppa.13233. https://doi.org/10.1111/ppa.13233

Jones, D. A. B., John, E., Rybak, K., Phan, H. T. T., Singh, K. B., Lin, S.-Y., Solomon, P. S., Oliver, R. P., & Tan, K.-C. (2019). A specific fungal transcription factor controls effector gene expression and orchestrates the establishment of the necrotrophic pathogen lifestyle on wheat. *Scientific Reports*, *9*(1), 1–13. https://doi.org/10.1038/s41598-019-52444-7

Jones, D. A., Bertazzoni, S., Turo, C. J., Syme, R. A., & Hane, J. K. (2018). Bioinformatic prediction of plant–pathogenicity effector proteins of fungi. *Current Opinion in Microbiology*, *46*, 43–49. https://doi.org/10.1016/j.mib.2018.01.017

Kettles, G. J., Bayon, C., Sparks, C. A., Canning, G., Kanyuka, K., & Rudd, J. J. (2018). Characterization of an antimicrobial and phytotoxic ribonuclease secreted by the fungal wheat pathogen *Zymoseptoria tritici*. *The New Phytologist*, *217*(1), 320–331. https://doi.org/10.1111/nph.14786

Lievens, B., Houterman, P. M., & Rep, M. (2009). Effector gene screening allows unambiguous identification of *Fusarium oxysporum* f. sp. *lycopersici* races and discrimination from other formae speciales. *FEMS Microbiology Letters*, *300*(2), 201–215. https://doi.org/10.1111/j.1574-6968.2009.01783.x

Lin, M., Ficke, A., Cockram, J., & Lillemo, M. (2020). Genetic Structure of the Norwegian *Parastagonospora nodorum* Population. *Frontiers in Microbiology*, *11*. https://doi.org/10.3389/fmicb.2020.01280

Lu, S., Gillian Turgeon, B., & Edwards, M. C. (2015). A ToxA-like protein from *Cochliobolus heterostrophus* induces light-dependent leaf necrosis and acts as a virulence factor with host selectivity on maize. *Fungal Genetics and Biology*, *81*, 12–24. https://doi.org/10.1016/j.fgb.2015.05.013

Mosquera, G., Giraldo, M. C., Khang, C. H., Coughlan, S., & Valent, B. (2009). Interaction Transcriptome Analysis Identifies *Magnaporthe oryzae* BAS1-4 as Biotrophy-Associated Secreted Proteins in Rice Blast Disease. *The Plant Cell*, *21*(4), 1273–1290. https://doi.org/10.1105/tpc.107.055228

Praz, C. R., Bourras, S., Zeng, F., Sánchez-Martín, J., Menardo, F., Xue, M., Yang, L., Roffler, S., Böni, R., Herren, G., McNally, K. E., Ben-David, R., Parlange, F., Oberhaensli, S., Flückiger, S., Schäfer, L. K., Wicker, T., Yu, D., & Keller, B. (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*(3), 1301–1314. https://doi.org/10.1111/nph.14372

Richards, J. K., Stukenbrock, E. H., Carpenter, J., Liu, Z., Cowger, C., Faris, J. D., & Friesen, T. L. (2019). Local adaptation drives the diversification of effectors in the fungal wheat pathogen *Parastagonospora nodorum* in the United States. *PLOS Genetics*, *15*(10), e1008223. https://doi.org/10.1371/journal.pgen.1008223

Richards, J. K., Wyatt, N. A., Liu, Z., Faris, J. D., & Friesen, T. L. (2018). Reference Quality Genome Assemblies of Three *Parastagonospora nodorum* Isolates Differing in Virulence on Wheat. *G3: Genes, Genomes, Genetics*, *8*(2), 393–399. https://doi.org/10.1534/g3.117.300462

Schmidt, S. M., Lukasiewicz, J., Farrer, R., Dam, P. v., Bertoldo, C., & 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*(1), 307–318. https://doi.org/10.1111/nph.13584

Sharpee, W., Oh, Y., Yi, M., Franck, W., Eyre, A., Okagaki, L. H., Valent, B., & Dean, R. A. (2017). Identification and characterization of suppressors of plant cell death (SPD) effectors from *Magnaporthe oryzae*. *Molecular Plant Pathology*, *18*(6), 850–863. https://doi.org/10.1111/mpp.12449

Shiller, J., Van de Wouw, A. P., Taranto, A. P., Bowen, J. K., Dubois, D., Robinson, A., Deng, C. H., & Plummer, K. M. (2015). A Large Family of *AvrLm6*-like Genes in the Apple and Pear Scab Pathogens, *Venturia inaequalis* and *Venturia pirina*. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.00980

Spanu, P. D. (2017). Cereal immunity against powdery mildews targets RNase-Like Proteins associated with Haustoria (RALPH) effectors evolved from a common ancestral gene. *New Phytologist*, *213*(3), 969–971. https://doi.org/10.1111/nph.14386

Sperschneider, J., Dodds, P. N., Gardiner, D. M., Singh, K. B., & Taylor, J. M. (2018). Improved prediction of fungal effector proteins from secretomes with EffectorP 2.0. *Molecular Plant Pathology*, *19*(9), 2094–2110. https://doi.org/10.1111/mpp.12682

Sperschneider, J., Gardiner, D. M., Dodds, P. N., Tini, F., Covarelli, L., Singh, K. B., Manners, J. M., & Taylor, J. M. (2016). EffectorP: Predicting fungal effector proteins from secretomes using machine learning. *New Phytologist*, *210*(2), 743–761. https://doi.org/10.1111/nph.13794

Sperschneider, J., Williams, A. H., Hane, J. K., Singh, K. B., & Taylor, J. M. (2015). Evaluation of Secretion Prediction Highlights Differing Approaches Needed for Oomycete and Fungal Effectors. *Frontiers in Plant Science*, *6*. https://doi.org/10.3389/fpls.2015.01168

Stukenbrock, E. H., Banke, S., & McDonald, B. A. (2006). Global migration patterns in the fungal wheat pathogen *Phaeosphaeria nodorum*. *Molecular Ecology*, *15*(10), 2895–2904. https://doi.org/10.1111/j.1365-294X.2006.02986.x

Syme, R. A., Tan, K.-C., Hane, J. K., Dodhia, K., Stoll, T., Hastie, M., Furuki, E., Ellwood, S. R., Williams, A. H., Tan, Y.-F., Testa, A. C., Gorman, J. J., & Oliver, R. P. (2016). Comprehensive Annotation of the *Parastagonospora nodorum* Reference Genome Using Next-Generation Genomics, Transcriptomics and Proteogenomics. *PLOS ONE*, *11*(2), e0147221. https://doi.org/10.1371/journal.pone.0147221

Syme, R. A., Tan, K.-C., Rybak, K., Friesen, T. L., McDonald, B. A., Oliver, R. P., & Hane, J. K. (2018). Pan-Parastagonospora Comparative Genome Analysis—Effector Prediction and Genome Evolution. *Genome Biology and Evolution*, *10*(9), 2443–2457. https://doi.org/10.1093/gbe/evy192

Testa, A. C., Hane, J. K., Ellwood, S. R., & Oliver, R. P. (2015). CodingQuarry: Highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. *BMC Genomics*, *16*(1), 170. https://doi.org/10.1186/s12864-015-1344-4

Testa, A. C., Oliver, R. P., & Hane, J. K. (2016). OcculterCut: A Comprehensive Survey of AT-Rich Regions in Fungal Genomes. *Genome Biology and Evolution*, *8*(6), 2044–2064. https://doi.org/10.1093/gbe/evw121

Wirthmueller, L., Maqbool, A., & Banfield, M. J. (2013). On the front line: Structural insights into plant–pathogen interactions. *Nature Reviews Microbiology*, *11*(11), 761–776. https://doi.org/10.1038/nrmicro3118

Appendix A - Permission to use copyright material

JOHN WILEY AND SONS LICENSE
TERMS AND CONDITIONS

Sep 02, 2021

This Agreement between Curtin University -- Darcy Jones ("You") and John Wiley and Sons ("John Wiley and Sons") consists of your license details and the terms and conditions provided by John Wiley and Sons and Copyright Clearance Center.

License Number	5075730047011
License date	May 25, 2021
Licensed Content Publisher	John Wiley and Sons
Licensed Content Publication	New Phytologist
Licensed Content Title	ApoplastP: prediction of effectors and plant proteins in the apoplast using machine learning
Licensed Content Author	Jana Sperschneider, Jennifer M. Taylor, Karam B. Singh, et al
Licensed Content Date	Dec 15, 2017
Licensed Content Volume	217
Licensed Content Issue	4
Licensed Content Pages	15
Type of use	Dissertation/Thesis

Requestor type	University/Academic
Format	Print and electronic
Portion	Figure/table
Number of figures/tables	1
Will you be translating?	No
Title	Fighting fungal pathogens with big data: new computational approaches for effector discovery and crop disease management
Institution name	Curtin University
Expected presentation date	Jun 2021
Portions	Figure 2 section A.
Requestor Location	Curtin University 8/91 Carrington Street Fremantle, Western Australia 6160 Australia Attn: Curtin University
Publisher Tax ID	EU826007151
Total	0.00 AUD

Terms and Conditions

TERMS AND CONDITIONS

This copyrighted material is owned by or exclusively licensed to John Wiley & Sons, Inc. or one of its group companies (each a"Wiley Company") or handled on behalf of a society with which a Wiley Company has exclusive publishing rights in relation to a particular work (collectively "WILEY"). By clicking "accept" in connection with completing this licensing transaction, you agree that the following terms and conditions apply to this transaction (along with the billing and payment terms and conditions established by the Copyright Clearance Center Inc., ("CCC's Billing and Payment terms and conditions"), at the time that

you opened your RightsLink account (these are available at any time at http://myaccount.copyright.com).

Terms and Conditions

- The materials you have requested permission to reproduce or reuse (the "Wiley Materials") are protected by copyright.

- You are hereby granted a personal, non-exclusive, non-sub licensable (on a stand-alone basis), non-transferable, worldwide, limited license to reproduce the Wiley Materials for the purpose specified in the licensing process. This license, **and any CONTENT (PDF or image file) purchased as part of your order,** is for a one-time use only and limited to any maximum distribution number specified in the license. The first instance of republication or reuse granted by this license must be completed within two years of the date of the grant of this license (although copies prepared before the end date may be distributed thereafter). The Wiley Materials shall not be used in any other manner or for any other purpose, beyond what is granted in the license. Permission is granted subject to an appropriate acknowledgement given to the author, title of the material/book/journal and the publisher. You shall also duplicate the copyright notice that appears in the Wiley publication in your use of the Wiley Material. Permission is also granted on the understanding that nowhere in the text is a previously published source acknowledged for all or part of this Wiley Material. Any third party content is expressly excluded from this permission.

- With respect to the Wiley Materials, all rights are reserved. Except as expressly granted by the terms of the license, no part of the Wiley Materials may be copied, modified, adapted (except for minor reformatting required by the new Publication), translated, reproduced, transferred or distributed, in any form or by any means, and no derivative works may be made based on the Wiley Materials without the prior permission of the respective copyright owner.**For STM Signatory Publishers clearing permission under the terms of the STM Permissions Guidelines only, the terms of the license are extended to include subsequent editions and for editions in other languages, provided such editions are for the work as a whole in situ and does not involve the separate exploitation of the permitted figures or extracts,** You may not alter, remove or suppress in any manner any copyright, trademark or other notices displayed by the Wiley Materials. You may not license, rent, sell, loan, lease, pledge, offer as security, transfer or assign the Wiley Materials on a stand-alone basis, or any of the rights granted to you hereunder to any other person.

- The Wiley Materials and all of the intellectual property rights therein shall at all times remain the exclusive property of John Wiley & Sons Inc, the Wiley Companies, or their respective licensors, and your interest therein is only that of having possession of and the right to reproduce the Wiley Materials pursuant to Section 2 herein during the continuance of this Agreement. You agree that you own no right, title or interest in or to the Wiley Materials or any of the intellectual property rights therein. You shall have no rights hereunder other than the license as provided for above in Section 2. No right, license or interest to any trademark, trade name, service mark or other branding ("Marks") of WILEY or its licensors is granted hereunder, and you agree that you shall not assert any such right, license or interest with respect thereto

- NEITHER WILEY NOR ITS LICENSORS MAKES ANY WARRANTY OR REPRESENTATION OF ANY KIND TO YOU OR ANY THIRD PARTY, EXPRESS, IMPLIED OR STATUTORY, WITH RESPECT TO THE MATERIALS OR THE ACCURACY OF ANY INFORMATION CONTAINED IN THE MATERIALS, INCLUDING, WITHOUT LIMITATION, ANY IMPLIED WARRANTY OF MERCHANTABILITY, ACCURACY, SATISFACTORY QUALITY, FITNESS FOR A PARTICULAR PURPOSE, USABILITY,

INTEGRATION OR NON-INFRINGEMENT AND ALL SUCH WARRANTIES ARE HEREBY EXCLUDED BY WILEY AND ITS LICENSORS AND WAIVED BY YOU.

- WILEY shall have the right to terminate this Agreement immediately upon breach of this Agreement by you.
- You shall indemnify, defend and hold harmless WILEY, its Licensors and their respective directors, officers, agents and employees, from and against any actual or threatened claims, demands, causes of action or proceedings arising from any breach of this Agreement by you.
- IN NO EVENT SHALL WILEY OR ITS LICENSORS BE LIABLE TO YOU OR ANY OTHER PARTY OR ANY OTHER PERSON OR ENTITY FOR ANY SPECIAL, CONSEQUENTIAL, INCIDENTAL, INDIRECT, EXEMPLARY OR PUNITIVE DAMAGES, HOWEVER CAUSED, ARISING OUT OF OR IN CONNECTION WITH THE DOWNLOADING, PROVISIONING, VIEWING OR USE OF THE MATERIALS REGARDLESS OF THE FORM OF ACTION, WHETHER FOR BREACH OF CONTRACT, BREACH OF WARRANTY, TORT, NEGLIGENCE, INFRINGEMENT OR OTHERWISE (INCLUDING, WITHOUT LIMITATION, DAMAGES BASED ON LOSS OF PROFITS, DATA, FILES, USE, BUSINESS OPPORTUNITY OR CLAIMS OF THIRD PARTIES), AND WHETHER OR NOT THE PARTY HAS BEEN ADVISED OF THE POSSIBILITY OF SUCH DAMAGES. THIS LIMITATION SHALL APPLY NOTWITHSTANDING ANY FAILURE OF ESSENTIAL PURPOSE OF ANY LIMITED REMEDY PROVIDED HEREIN.
- Should any provision of this Agreement be held by a court of competent jurisdiction to be illegal, invalid, or unenforceable, that provision shall be deemed amended to achieve as nearly as possible the same economic effect as the original provision, and the legality, validity and enforceability of the remaining provisions of this Agreement shall not be affected or impaired thereby.
- The failure of either party to enforce any term or condition of this Agreement shall not constitute a waiver of either party's right to enforce each and every term and condition of this Agreement. No breach under this agreement shall be deemed waived or excused by either party unless such waiver or consent is in writing signed by the party granting such waiver or consent. The waiver by or consent of a party to a breach of any provision of this Agreement shall not operate or be construed as a waiver of or consent to any other or subsequent breach by such other party.
- This Agreement may not be assigned (including by operation of law or otherwise) by you without WILEY's prior written consent.
- Any fee required for this permission shall be non-refundable after thirty (30) days from receipt by the CCC.
- These terms and conditions together with CCC's Billing and Payment terms and conditions (which are incorporated herein) form the entire agreement between you and WILEY concerning this licensing transaction and (in the absence of fraud) supersedes all prior agreements and representations of the parties, oral or written. This Agreement may not be amended except in writing signed by both parties. This Agreement shall be binding upon and inure to the benefit of the parties' successors, legal representatives, and authorized assigns.
- In the event of any conflict between your obligations established by these terms and conditions and those established by CCC's Billing and Payment terms and conditions,

these terms and conditions shall prevail.

- WILEY expressly reserves all rights not specifically granted in the combination of (i) the license details provided by you and accepted in the course of this licensing transaction, (ii) these terms and conditions and (iii) CCC's Billing and Payment terms and conditions.

- This Agreement will be void if the Type of Use, Format, Circulation, or Requestor Type was misrepresented during the licensing process.

- This Agreement shall be governed by and construed in accordance with the laws of the State of New York, USA, without regards to such state's conflict of law rules. Any legal action, suit or proceeding arising out of or relating to these Terms and Conditions or the breach thereof shall be instituted in a court of competent jurisdiction in New York County in the State of New York in the United States of America and each party hereby consents and submits to the personal jurisdiction of such court, waives any objection to venue in such court and consents to service of process by registered or certified mail, return receipt requested, at the last known address of such party.

WILEY OPEN ACCESS TERMS AND CONDITIONS

Wiley Publishes Open Access Articles in fully Open Access Journals and in Subscription journals offering Online Open. Although most of the fully Open Access journals publish open access articles under the terms of the Creative Commons Attribution (CC BY) License only, the subscription journals and a few of the Open Access Journals offer a choice of Creative Commons Licenses. The license type is clearly identified on the article.

The Creative Commons Attribution License

The Creative Commons Attribution License (CC-BY) allows users to copy, distribute and transmit an article, adapt the article and make commercial use of the article. The CC-BY license permits commercial and non-

Creative Commons Attribution Non-Commercial License

The Creative Commons Attribution Non-Commercial (CC-BY-NC)License permits use, distribution and reproduction in any medium, provided the original work is properly cited and is not used for commercial purposes.(see below)

Creative Commons Attribution-Non-Commercial-NoDerivs License

The Creative Commons Attribution Non-Commercial-NoDerivs License (CC-BY-NC-ND) permits use, distribution and reproduction in any medium, provided the original work is properly cited, is not used for commercial purposes and no modifications or adaptations are made. (see below)

Use by commercial "for-profit" organizations

Use of Wiley Open Access articles for commercial, promotional, or marketing purposes requires further explicit permission from Wiley and will be subject to a fee.

Further details can be found on Wiley Online Library
http://olabout.wiley.com/WileyCDA/Section/id-410895.html

Other Terms and Conditions:

v1.10 Last updated September 2015

Questions? customercare@copyright.com or +1-855-239-3415 (toll free in the US) or +1-978-646-2777.
