School of Science and Engineering
Department of Environment and Agriculture

Effector gene prediction from fungal pathogen genome assemblies

Alison Clare Testa

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

March 2016

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 that has been accepted for the award of any other degree or diploma in any university.

Signature: _________________________ Date: ________________

Alison Clare Testa

Abstract

Fungal phytopathogens interact with their plant hosts largely via the secretion of effectors, which manipulate and elicit plant defences. Knowledge of fungal effectors has provided tangible benefits to the agricultural industry, and as such the search for novel effectors continues to be an active area of research. With the advent of whole genome sequencing, it has become common practice to sequence and assemble the genome of a pathogen, predict the gene content, and arrive at a set of putative effectors for further analysis. In doing so there is a heavy reliance on the accuracy of each of these processes. In Chapter 1 I review the relevant literature relating to fungal effectors and their prediction, and in Chapter 2 I provide a more focused review of necrotrophic effectors. The presented research makes a novel contribution to the field of effector discovery through the development of improved fungal gene prediction techniques, a detailed investigation into AT-rich regions, relevant to fungal evolution and the genomic context of effectors, and the provision of improved genomic resources and effector candidate lists for two important barley pathogens, *Pyrenophora teres* f. *teres* and *Pyrenophora teres* f. *maculata.*

To address the need for high levels of accuracy in gene prediction, part of this research, described in Chapter 3, focused on improving fungal gene prediction through the incorporation of RNA-seq data. The resulting software tool, CodingQuarry, uses a novel method to incorporate RNA-seq data into hidden Markov model gene prediction. CodingQuarry was demonstrated to provide more accurate gene predictions than competing software tools when tested on the model organisms *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. In Chapter 4 a more specific investigation into the success of effector gene prediction is described. *Ab initio* prediction methods were found to miss many effectors. These failures were due to factors such as atypical codon usage, high cysteine content, and the signal peptide sequence accounting for a large proportion of the coding sequence. Despite this, where effector gene loci were supported by RNA-seq data, RNA-seq driven predictions made by CodingQuarry and BRAKER1 were able to capture most effector genes. To reduce this reliance on RNA-seq data, the functionality of CodingQuarry was extended to target the prediction of secreted, cysteine-rich genes with atypical codon usage, improving the success rate of the *ab initio* prediction of effectors. CodingQuarry will therefore also be applicable where effectors or effector-like genes are not expressed under the conditions used to generate the RNA-seq data, or where RNA-seq data is not available for the species or isolate of interest.

Many effector genes are found within or close to AT-rich genomic regions. I therefore focussed on AT-rich regions within fungal genomes, aiming to document their prevalence and properties. A novel method, performed by the software tool OcculterCut, was developed to facilitate the rapid assessment and characterisation of AT-rich regions. In a pilot study described in Chapter 5, OcculterCut was used to assess the AT-rich genome content of the wheat pathogen *Zymoseptoria*

tritici. In Chapter 6, a comprehensive survey of the AT-rich region content of 538 published fungal genome assemblies is described. AT-rich bimodal genomes were found to be common within the Pezizomycotina subphylum and plant-associated fungi therein. Coding sequences in AT-rich genome regions were shown to have differing amino acid and codon usage to coding sequences in GC-equilibrated regions, indicating that this genomic context influences gene evolution and has implications for gene prediction.

Chapter 7 describes a practical application of the techniques explored in this Ph. D. and other recent progress in the field of effector discovery. PacBio long-read genome assemblies, high quality RNA-seq driven annotations and effector candidate lists were generated for *P. teres* f. *teres* and *P. teres* f. *maculata.* Both forms were found to have a high repetitive content and exemplify the bimodal AT-rich genome type. As such, the genomic context of genes in relation to these regions was included as a line of evidence in assembling lists of effector candidates. The improved genomic resources and assessment of putative effector genes provide a powerful resource for further investigation into these pathogens.

Acknowledgments

Firstly, I would like to thank Richard Oliver, my primary supervisor, for helping me to move to a career in bioinformatics and for his generosity in sharing ideas, advice, and knowledge. I would also like to thank my co-supervisor, James Hane, whose bioinformatics expertise has been invaluable. Thanks also go to Simon Ellwood, whose patience and time while I was learning a new discipline was extremely helpful. I have thoroughly enjoyed my time at the CCDM (and ACNFP), and I would like to thank all staff and students for their encouragement and support.

Many thanks to Curtin University for supporting me through the first 6 months of my Ph. D. with a Curtin Interim Scholarship, and to the Australian Federal Government and Grains Research and Development Corporation for awarding me a Grains Research Scholarship (GRS10564) and Australian Postgraduate Award.

Parts of this research were undertaken with the assistance of resources provided at the NCI Specialised Facility in Bioinformatics at The University of Queensland through the National Computational Merit Allocation Scheme supported by the Australian Government, and using resources provided by the Pawsey Supercomputing Centre with funding from the Australian Government and the Government of Western Australia. Special thanks go to the Pawsey Supercomputing Centre for providing training, events, and user support.

Chapter 1 Literature review

Crop losses due to pathogens are economically damaging to the agricultural industry (1,2) and threaten food security (3). Certain species of oomycetes, nematodes, bacteria, and viruses are known to have phytopathogenic lifestyles, but the majority of crop losses are due to plant diseases caused by fungi. In Australia, diseases caused by fungal pathogens account for the majority of crop losses in wheat and barley (1,2) – the country's two most important crops. Furthermore, in Australia fungal diseases on pulse crops such as field pea, narrowleaf lupin, chickpea, and faba bean have the potential to cause 100 per cent yield losses (4), making these crops high-risk for farmers. As such, understanding how fungal pathogens interact with their plant host and cause disease is vital to controlling and minimising crop losses in Australia and the rest of the world.

1.1 Fungal pathogens of interest

Certain fungal pathogen species are of particular relevance to this Ph. D. thesis either because they are included in the bioinformatic analysis described in Chapters 4-7, or because they have characteristics particularly relevant to the discussion of fungal effector biology, pathogen evolution, and effector gene prediction later in this review. In the next sections, I give a brief introduction to these species, the diseases they cause, and their available genomic resources. An introduction to *Passalora fulva* is given as this pathogen has become a model species for understanding avirulence effectors, and is therefore frequently mentioned in the remainder of this literature review. *Pyrenophora tritici-repentis* and *Parastagonospora nodorum* are model systems for understanding proteinaceous necrotrophic effectors (NEs) and their interactions, which are discussed in more detail in Chapter 2. The repeat and gene content of *P. tritici-repentis* is also compared to closely related barley-infecting *Pyrenophora teres* species in Chapter 7. Research that has been conducted on the genome sequence and effector complement of *Leptosphaeria maculans* with regards to AT-rich regions and repeat-induced point mutation (RIP) is highly relevant to Chapter 6, where AT-rich regions within fungal genomes are explored in detail. *Zymoseptoria tritici* is discussed in Chapter 5 as part of a review article, along with an analysis of the AT-rich region content of the assembly. *Pyrenophora teres* f. *teres* and *Pyrenophora teres* f. *maculata* are the focus of Chapter 7 and thus briefly introduced here. These species are all within the Pezizomycotina subphylum (class Dothideomycetes), which contains numerous plant pathogens (5) many of which now have whole genome sequence resources publically available (Figure 1).

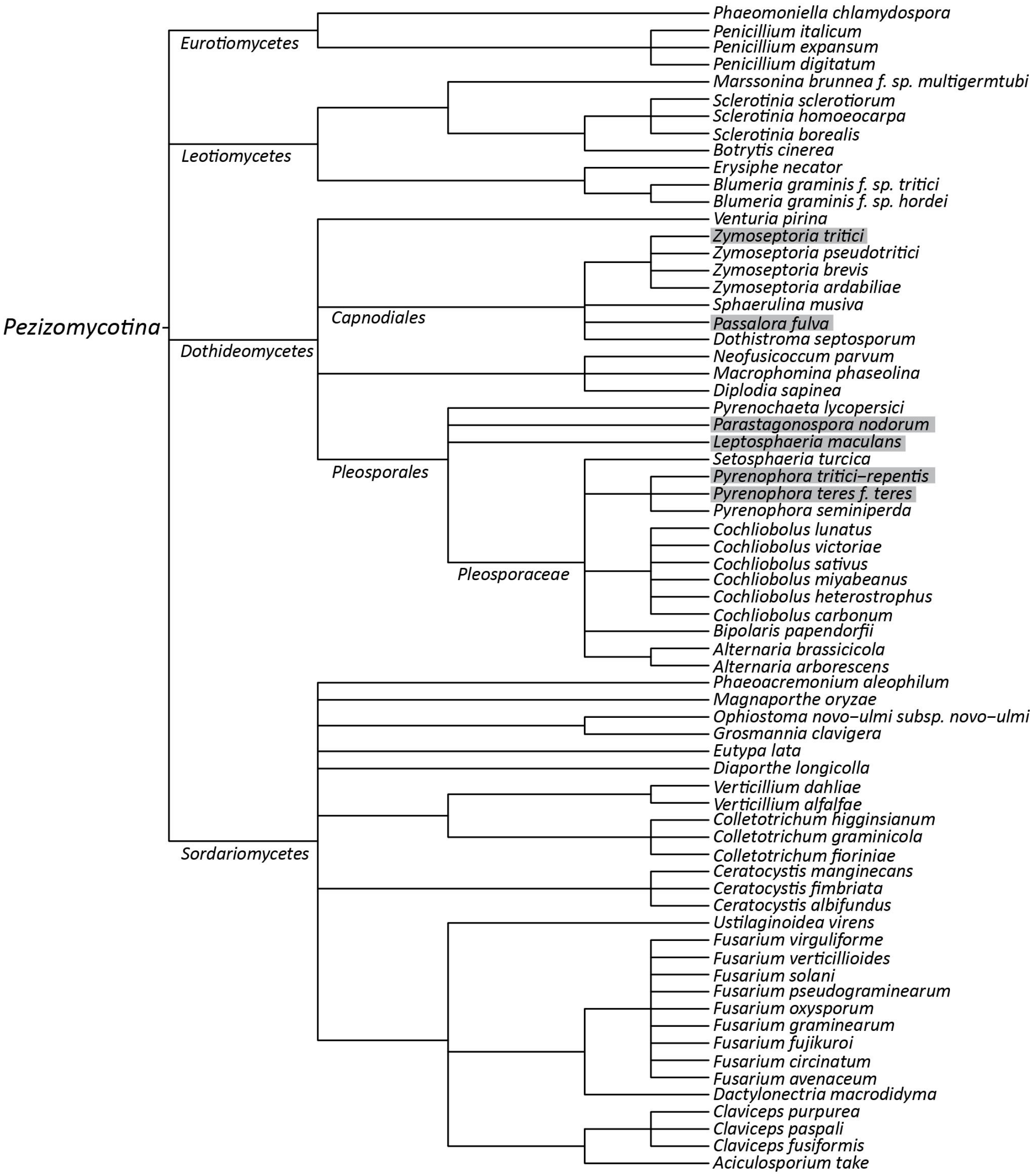


Figure 1 A cladogram showing Dothideomycete plant pathogens species that have publically available genomes sequences. The species highlighted in grey are of particular relevance to this body of research.

1.1.1 *Parastagonospora nodorum*

P. nodorum (formerly *Stagonospora nodorum*) is a necrotrophic fungal pathogen of wheat and the causal agent of *Stagonospora nodorum* blotch (SNB) on wheat. Symptoms appear as oval or round necrotic lesions on the leaves, necrotic bands on stems, and lesions on glumes (6). Disease symptoms are easily confused with those of tan spot, caused by *P. tritici-repentis,* which often co-infects wheat (7). This disease has been of responsible for particularly severe crop losses in the wheat growing regions of Western Australia, amounting to an estimated 9% of total yield losses (1). During infection, spores land on the leaf and germinate, forming hyphae that enter the leaf via a penetration structure known as a hyphopodium or through stomata (6). The genome sequence of *P. nodorum* (reference strain SN15) was the earliest publication of a Dothideomycete pathogen genome, and the study also included the analysis of a library of 7,750 expressed sequence tags (ESTs) (post quality filtering) (8). The initial genome assembly was 37 Mb contained within 107 scaffolds and had a relatively low repetitive content of 4.5% (8). Additional strains SN4 and SN79 were later sequenced and published for comparative genomic purposes (9). Recently, the genome sequence has been updated, and RNA-seq and proteomic data used to conduct a comprehensive re-annotation (10). Genomic resources, along with established molecular techniques, a worldwide collection of isolates, and genetically characterised host-mapping populations have contributed to *P. nodorum* becoming a model organism for the study of similar necrotrophic pathogens (7). Furthermore, the characterisation and subsequent study of *P. nodorum* effectors have contributed to this species becoming a model for understanding the inverse gene-for-gene interaction described in the review of necrotrophic effectors (NEs) in Chapter 2.

1.1.2 *Pyrenophora tritici-repentis*

P. tritici-repentis is a necrotrophic fungal pathogen that causes tan spot disease on wheat, with symptoms appearing as round or oval shaped necrotic lesions on the leaves (11). *P. tritici-repentis* is part of the *Pyrenophora* genus, which contains barley pathogens *P. teres* f. *teres*, *P. teres* f. *maculata*, and *Pyrenophora graminicola*, as well as a seed pathogen *Pyrenophora semeniperda*. The *P. tritici-repentis* reference (isolate Pt-1C-BFP) genome was sequenced using Sanger technology with scaffold placement assisted by an optical map (12). The resulting assembly was 37.8 Mbp contained within just 47 scaffolds. A non-pathogenic isolate and an additional pathogenic isolate were sequenced using Illumina short-read technology (12). Comparisons between the three assemblies indicated that transposon activity had played a key role in the evolution of pathogenicity in this species (12).

1.1.3 *Pyrenophora teres* f. *teres* and *Pyrenophora teres* f. *maculata*

P. teres f. *teres* and *P. teres* f. *maculata* are fungal pathogens that cause net form of net blotch and spot form of net blotch (respectively) on barley. The disease symptoms of both forms appear on the leaves, but have a different appearance. Net form of net blotch appears as necrotic lines that run along the leaf veins and are connected by occasional transverse streaks of necrosis, forming a crisscross or net pattern (13). Spot form of net blotch presents as oval or round necrotic lesions surrounded by a yellow chlorotic halo (13). Phylogenetic analysis suggests that *P. teres* f. *teres* and *P. teres* f. *maculata* diverged from one another around 519 kya ago, and diverged from *P. tritici-repentis* around 8.04 Mya ago (14). The genome sequence of *P. teres* f. *teres* isolate 0-1 was published in 2010 (15), assembled from paired-end short-read data from the Solexa sequencing platform. The resulting assembly was 42 Mbp contained within 146,737 scaffolds, although most of these were short fragments and 80% of the genome was contained within 6,684 scaffolds. Whole genome sequence resources are not publically available for *P. teres* f. *maculata.*

1.1.4 *Leptosphaeria maculans*

L. maculans is a fungal pathogen that infects brassica crops, in particular oilseed rape (also known as canola). It causes stem canker or blackleg, with disease symptoms appearing as greyish-green collapsed cotyledon and leaf tissue (16). *L. maculans* is considered to be a hemibiotroph (16). The genome sequence of *L. maculans* 'brassicae' (isolate v23.1.3) was published in 2011 (17). The assembly was 45 Mbp contained within 76 scaffolds and was found to have a relatively high component of repeats, accounting for around one third of the assembly (17). Interestingly, the genome was reported to be composed of a mosaic of GC-equilibrated regions and gene sparse AT-rich regions, formed through the activity of repeat-induced point mutation (RIP) (17). This bimodal genome type is explored in detail in Chapter 6.

1.1.5 *Zymoseptoria tritici*

Z. tritici (syn. *Mycosphaerella graminicola*) is a wheat pathogen and the cause of septoria leaf blotch, with disease symptoms appearing as necrotic lesions with chlorotic borders on leaves. Unlike the other pathogens described so far, which are within the order Pleosporales, *Z. tritici* is from the order Capnodiales (Figure 1). *Z. tritici* is phylogenetically distant from *S. nodorum* and *P. tritici-repentis*, however differentiating the diseases each cause can be difficult due to their frequent co-infection of wheat and similar symptoms. During infection, hyphae extend along the leaf and enter the plant leaf through stomata (18,19). *Z. tritici* has a period of latent infection before lesions appear, at which point the pathogen switches to a necrotrophic phase. This two-phase mode of infection means that *Z. tritici* is generally described as a hemibiotroph. The reference genome sequence (IPO323) was published in 2011 (18). Whole-genome shotgun sequencing using three libraries with differing insert sizes (2-3, 6-8, and 35-40 kb) resulted in a near complete genome assembly, consisting of 21 assembled chromosomes amounting to 37.9 Mb (18). The assembly was found to be around 17% repetitive with genomic evidence of repeat-induced point mutation (RIP) (20). Bioinformatic resources and research on *Z. tritici* are reviewed in Chapter 5, along with an analysis of the AT-rich region content of the genome assembly.

1.1.6 *Passalora fulva*

P. fulva (syn. *Cladosporium fulvum*) is a fungal pathogen that causes leaf mould on susceptible *Solanum lycopersicon* (tomato) plants. The taxonomic classification of *P. fulva* places it within the Mycosphaerellaceae family, along with *Z. tritici* and the pine pathogen *Dothistroma septosporum*. During infection by *P. fulva*, conidia germinate on the leaf surface producing hyphae that enter the plant leaf through stomata (21). Disease symptoms appear as white mould on the leaf that turns brown when the fungus sporulates, as well as leaf wilting and curling due to stomata blockages (21). *P. fulva* is considered to be a non-obligate biotroph (22). The genome sequence of *P. fulva* was published in 2012, along with that of the closely related *D. septosporum* (22). The genome was sequenced using paired end shotgun 454 reads, assembling into 61 Mb within 2,664 scaffolds (22). *P. fulva* was found to have a high repetitive content of around 47% (22). An assessment of the AT-rich region content of the *P. fulva* genome is made in Chapter 6.

1.2 Fungal Effectors

Plants are considered to have two main lines of defence against pathogen colonisation: basal and effector triggered (23–26). Basal immunity relies on recognising molecular patterns that typify pathogens, termed pathogen-associated molecular patterns (PAMPs), triggering a defence response from the plant resulting in PAMP-triggered immunity (PTI). In fungal pathogen-plant interactions, a well-known example of a PAMP is chitin, which forms a key structural component of fungal cell walls. Recognition of chitin by the plant has been shown to trigger the production of plant chitinases that degrade chitin thus suppressing fungal growth. Plant defence responses can also be initiated by the recognition of their own degraded polymer molecules, termed danger- or damage-associated molecules (DAMPs)(24,25,27). For example, one way fungi derive nutrition is from plant carbohydrates through the activity of carbohydrate degrading enzymes (Carbohydrate-Active enZymes - CAZys). The activity of these enzymes releases molecules from the plant cell wall, which are then recognised by the plant as DAMPs. The recognition of these ubiquitous PAMPS and DAMPs differs from the recognition of specific molecules, termed effectors, that forms the second line of plant defence.

Effectors are molecules produced by the pathogen that interact with the host to allow the pathogen to evade host recognition, trigger defence responses, and/or cause or increase virulence.
Most known effectors are secreted proteins, although there are some examples of small RNAs (28) and secondary metabolites (29) also acting as effectors. Secondary metabolite effectors in the form of host-selective toxins (HSTs) are reviewed in more detail in Chapter 2. Effectors are either secreted into the extracellular space and interact in the apoplast or xylem of the plant, or are taken up by the plant cell and interact within the cytoplasm. In the case of obligate biotrophs species such as *Blumeria* spp., *Puccinia* spp., and *Melampsora lini*, effectors are delivered via specialised feeding structures called haustoria that penetrate the host cell (30). The species introduced in this review —

and indeed most known fungal pathogens — are not obligate biotrophs and the mode by which effectors enter the host cell is unknown (31). Some effectors function to inhibit PTI and overcome basal plant immunity. For example, certain effectors have been shown to function to prevent PTI that can occur through recognition of chitin. The *P. fulva* effector Ecp6 has been shown to bind to chitin molecules, effectively hiding them from the plants recognition mechanisms and preventing PTI (32–34). There are also subsets of effectors that interact with the host in a different or additional way to those whose only known function is to inhibit PTI. These effectors can be can be grouped into the set of effectors that are avirulence determinants, termed avirulence effectors (Avrs), and necrotrophic effectors (NEs).

1.2.1 Avirulence effectors

Encompassed within the broader category of effectors are avirulence (Avr) genes. An Avr gene product interacts with the product of a matching host resistance (R) gene, triggering plant defences (23,26). This usually culminates in a hypersensitive response (HR) by the plant leading to immunity. The HR is a defence mechanism enacted by the plant whereby, though localised cell death, the plant cuts off the pathogens food source thus preventing infection. As such, the presence of the fungal effector and plant resistance genes leads to effector-triggered immunity. This gene–for–gene interaction was initially hypothesised in the 1940s with reference to the interaction between *Melampsora lini,* the causal agent of flax rust, and flax (35,36). The first fungal avirulence effector to be cloned was *Avr9* from the tomato pathogen *Passalora fulva* (37), around 50 years after Flors landmark paper. Since then, a number of fungal avirulence effectors have been identified and characterised, in some cases with their corresponding host R gene.

In the presence of a non-resistant host, many Avrs are often either known to or presumed to act as virulence factors, contributing to pathogenicity. For some avirulence effectors, a primary role as virulence factors has been determined. For example, the *P. fulva* effector Avr4 protects fungal hyphae against hydrolysis by plant chitinases by binding to chitin (38,39). Avr2 inhibits cysteine proteases, which play a role in plant defences (39,40). In the presence of their respective R genes (*Cf-4* and *Cf-2*), both proteins interact with the pathogen to induce a HR. In these cases, the primary role of the effector is overridden by host recognition, resulting in the effector also being an avirulence determinant. It is presumed that Avr type effectors have an effector function in the absence of host recognition but this has not only been demonstrated for a subset of the known Avrs.

The evolution and change of Avr genes and their matching plant R genes is an arms race between host and pathogen. The pathogen population evolves effectors to overcome plant defences, and the plant population responds by evolving mechanisms by which to specifically recognise effectors. In the absence of host recognition, loss of an Avr gene confers a fitness penalty, the degree of which depends on the intrinsic function of the effector. When challenged with a resistant host, loss or modification of the *Avr* gene allows the pathogen to evade host recognition. This can occur through deletions, point mutations, and/or transposon insertion within the Avr gene (23,26). Importantly, the mode by which host resistance is overcome differs between different *Avr* genes and pathogens. For example, the only observed route to overcome *Rlm4* recognition of the *L. maculans* effector protein AvrLm4-7 is a single mutation altering a glycine residue to an arginine (41). In contrast, complete deletion of the *L. maculans* Avr gene *AvrLm1* was found to be the most common adaptation to overcome *Rlm1* mediated resistance (42). In *P. fulva*, single point mutations to *Avr4* result in unstable but functional proteins that are not recognised by the R gene *Cf-4* (43,44). This method of recognition avoidance differs from those seen affecting *Avr4E*, which is more commonly deleted or rendered non-functional (44). These examples suggest different fitness penalties associated with the loss of different effector genes and that selection favours different kinds of adaptations (44). Additionally, pathogens and individual avirulence genes may differ in the evolutionary mechanisms available to bypass resistance and in their evolutionary potential — a point discussed in more detail later in this chapter.

1.2.2 Necrotrophic effectors

Another sub-set of effectors are “necrotrophic” effectors (NEs), including what has been traditionally termed host-selective toxins (45). Due to the difficultly in distinctly classifying pathogens according to their lifestyle (discussed further in Chapter 2), for the purposes of this review NEs are not solely considered to be effectors produced by necrotrophic pathogens, but rather effectors that interact with the host in a way that promotes or has the potential to promote infection by inducing necrosis or chlorosis. This definition thus encompasses the effectors produced by necrotrophic pathogens, but also extends to similarly functioning effectors produced by pathogens that do not exclusively employ a necrotrophic lifestyle. NEs have been shown to function in many plant pathogen interactions, including those involving the wheat pathogens *P. nodorum* and *P. tritici-repentis* introduced earlier in this chapter. Strong evidence has been presented that NEs also function in the barley–*P. teres* interaction (46), making this class of effectors of particular relevance to this body of research. Briefly, one of the important similarities between certain NEs and Avrs is that they are specifically recognised by the host. A crucial difference is that host recognition of a NE can lead to susceptibility, as in some cases the pathogen exploits host resistance mechanisms to its advantage. A detailed review of NEs is included separately in Chapter 2.

1.2.3 Pathogen accelerated evolutionary mechanisms and effector evolution

Pathogens are under strong selective pressure to overcome host resistance. Plant pathogenic fungi are concentrated in the Dothideomycete, Leotiomycete, and Sordariomycete families within the Pezizomycotina subphylum. There are common themes to the evolutionary mechanisms seen among these pathogenic fungi that may contribute to their success as pathogens. In many cases these evolutionary mechanisms have been shown to contribute to the pathogens ability to modify, expand, or reduce its effector repertoire to overcome plant resistance. Here, genome evolution through transposon activity, repeat-induced point mutation, dispensable chromosomes, and horizontal gene transfer (HGT) are reviewed. While not the only mechanisms of fungal genome evolution, these mechanisms operate in addition to normal evolution and have been identified as being particularly relevant to the evolution of effectors.

Effectors have been observed to frequently be closely associated with repetitive elements and in some cases repeat dense regions (12,17,47,48). Transposon activity contributes to genetic diversity. There is enormous variation in the repetitive content and overall size of fungal pathogen genomes (49), as well as examples of pathogen genomes being larger than closely related non-pathogenic species. Obligate biotrophs in particular have been observed to have extremely large repeat rich genomes (50). For example, the barley powdery mildew pathogen *Blumeria graminis* f. sp. *hordei* assembled to around 120 Mbp, with a transposable elements accounting for most (64%) of the genome (51). To put this genome size into context, the genomes of necrotrophic plant pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea* — also from the Leotiomycete family — have genomes of around 38–39 Mb and lower repetitive components of around 7% and less than 1% respectively (52). It has been proposed that the rampant propagation of transposons seen in the genomes of obligate biotrophs has been a key component to their evolution and the evolution of their effectors (50). The “bigger can be better” concept (53) has been used to describe such genomes — not solely limited to the obligate biotrophs — where the evolutionary advantages conferred by transposon activity and genome expansion outweigh the fitness penalty that arises from having to repair and replicate a large genome. In other cases a direct involvement of transposon activity in adaptation to host resistance has been observed (54,55). In some cases repeats have accumulated to form repeat dense regions, resulting in a state of genome compartmentalisation.

Some pathogens have non-essential chromosomes, termed dispensable chromosomes, which are specific to certain pathogen isolates or lineages (56–59). For example, *formae speciales* of *Fusarium oxysporum* can be distinguished by the presence or absence of certain lineage specific chromosomes (57,60). These dispensable chromosomes tend to be repeat rich, devoid of housekeeping genes, and in some cases have been shown to harbour effectors or other genes with a role in pathogenicity or host specificity. The wheat pathogen *Z. tritici* has 8 dispensable or accessory chromosomes that have been described as a “cradle for adaptive evolution” (61). This terminology refers to the potential for these chromosomes to diversify more rapidly than the core genome, similarly to repeat-rich regions, due to the fitness cost of mutations being lower on non-essential chromosomes compared to the core chromosomes.

Certain fungi — particularly within the Pezizomycotina subphylum — have a genome defence mechanism called repeat-induced point mutation (RIP). RIP was first observed in the in the genome of the model organism *Neurospora crassa* (62), peppering repeats (over 400bp and 80% similarity in

N. crassa (63,64)) with cytosine to thymine (C to T) mutations (and guanine to adenine (G to A) mutations on the reverse strand). These mutations introduce stop codons into the open-reading frames of transposons, rendering them inactive and protecting the organism from the detrimental effects of transposon activity. The resulting genomic signature of RIP activity is a low GC-content in RIP affected regions, and in most cases a TpA bias resulting from the preferential mutation of CpA dinucleotides (65). The presence and activity of RIP could be expected to prevent the discussed benefits of transposon activity to pathogen evolution, and the absence of the genes necessary for RIP from the genomes of obligate biotrophs has been noted (51). Despite this, many pathogen genomes have evidence of RIP activity and, curiously, evidence of RIP capability is seen in some pathogens that also have a high repeat content (49). A well-known example of this is seen in the genome assembly of *L. maculans*, in which around one third of the genome assembly is repetitive (17). In *L. maculans*, repeats have accumulated into repeat dense regions degraded by RIP, giving the genome a bimodal GC-content that alternates between AT-rich and GC-equilibrated genome regions. Adding interest to this, *L. maculans* effector genes (*AvrLm1*, *AvrLm2*, *AvrLm4-7*, *AvrLm6*, and *AvrLmJ1*) were found to be located within these otherwise gene-sparse regions (66–69). Furthermore, RIP has been demonstrated to be instrumental in the pseudogenisation of effectors, allowing *L. maculans* to rapidly evolve to evade host recognition and overcome plant resistance (17,69–71). As such, there is evidence that RIP is not only a genome defence mechanism, but also a driver of evolution.

The terminology “two-speed” genome evolution was first coined in reference to the transposon-rich regions in the genome of the Oomycete pathogen *Phytophthora infestans* (72). The two-speed genome evolution terminology has since been frequently used in fungal genomics to describe where transposon activity, dispensable chromosomes, and in some cases RIP contribute to a situation where parts of a fungal genome diversify and evolve more rapidly than others (61,73,74).

In addition to transposon-mediated mechanisms for genome evolution and adaptation, fungal pathogens are also able to evolve via the acquisition of genetic material via HGT. It is possible that HGT allows fungal pathogens to acquire novel effectors from other species. This was the case for *P. tritici-repentis*, which famously acquired the *ToxA* effector from *P. nodorum* via HGT (75). Strong evidence has also been presented that the pea pathogenicity (PEP) gene cluster in *Nectria haematococca* was acquired through HGT (76). While HGT is not exclusive to fungal pathogens (77,78) or indeed fungi, it has been shown to be more prevalent with the pathogen-rich Pezizomycotina subphylum (77). Furthermore it has been suggested that HGT may play an important role in the evolution of phytopathogens (79,80).

1.2.4 Crop protection

Breeding crop varieties for resistance to pathogens was conducted long before its genetic basis was understood. For example, in Australia the pioneering work of William Farrer in breeding for wheat rust resistance was carried out in the late 19th century, prior to the rediscovery of Mendel's insights into inheritance (81). These days, knowledge of effectors assists in the development of molecular markers, which allow new lines to be screened for resistance. More recently, the provision of effector protein solutions to breeders has allowed assays to be carried out on new lines, offering a rapid and comparatively high-throughput method by which to assess resistance. This strategy has been highly successful in reducing sensitivity of Australian wheat cultivars to NEs produced by *S. nodorum* and *P. tritici-repentis,* with savings due to disease reduction estimated at $50 million (82). NEs and crop protection are reviewed in more detail in Chapter 2. An advantage of breeding for resistance is that it reduces the reliance on other disease control measures such as fungicides. Although these breeding strategies are highly effective, the arms race between host and pathogen continues. Once identified or bred, resistant crop varieties are widely used, exerting enormous selective pressure on virulent pathogen species or virulent alleles within a population. As a result, the continued monitoring and knowledge of a pathogen's effector complement are necessary for the continued provision and selection of crop varieties.

1.3 Identifying novel effectors

Given the importance of effectors to virulence and tangible benefits of effector discovery to industry, the identification of effectors is a principal area of research. Pioneering effector / avirulence gene discovery generally used map-based cloning techniques, but with the advent of whole genome sequencing a number of additional methods have emerged. Genome sequencing of fungi began with model organisms; firstly the yeast *Saccharomyces cerevisiae*, and some years later filamentous fungi *N. crassa* (83) and *Aspergillus nidulans* (84). In 2005 *Magnaporthe grisea* marked the first filamentous fungal pathogen to have its genome sequence published (85). Since then, hundreds of fungal genome sequences have been published, including many plant pathogens. Whole genome sequence data has become a powerful resource in selecting putative effectors for laboratory confirmation, reviewed in the following paragraphs and summarised in Figure 2.

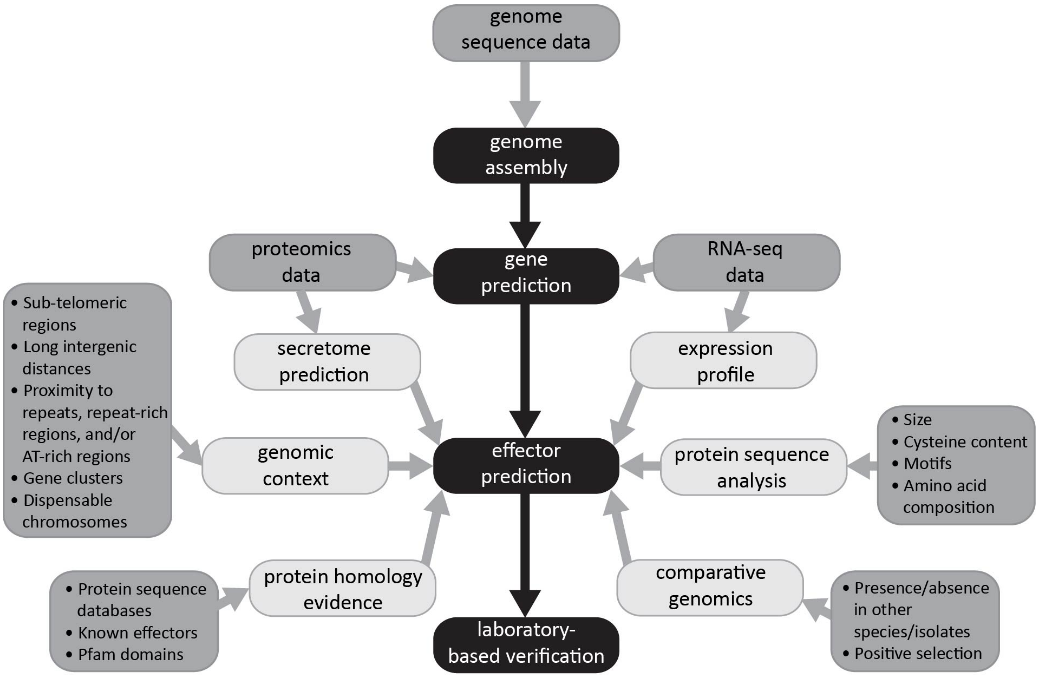


Figure 2 A flow chart showing the main steps (black boxes) in effector gene prediction. Sources of evidence feeding into effector prediction are shown in light grey, and contributing data/information is listed in the dark grey boxes.

1.3.1 Selecting effector candidates

The identification of candidate effectors from larger sets of genes is recognised as a continuing bioinformatic challenge (86). Experimental testing of putative effectors is time consuming and labour intensive, ideally suited to small sets of candidate effectors. Unfortunately, the genomes of fungal pathogens contain hundreds of genes that could potentially encode effectors (24). Reducing the size of this set by bioinformatic methods can be a trade off between strict criteria that may exclude some effectors versus lax criteria that include a large number of other secreted proteins. Bioinformatic methods used to predict effectors generally relate to properties of their protein sequence, homology, comparative genomics, genomic context, as well as the incorporation of expression data and proteomic data where available (86,87).

1.3.1.1 Protein sequence analysis

Early discoveries of small, cysteine-rich, species-specific effectors set in place the idea that these protein characteristics are typical of effectors. These characteristics have been used to conduct genome wide searches for effector candidates. While using these criteria has in some cases been successful, we now have numerous examples of effectors that would have — if relying on these criteria alone — gone undetected. For example, the *L. maculans* effector protein AvrLm1 only has one cysteine (66). Furthermore, there is variation in the published literature in what qualifies a protein as cysteine-rich, with examples defining cysteine-rich as 3% cysteine (88), 4% cysteine (60), and a percentage one standard deviation above the mean expected value (87). While known effectors are generally small, some are larger than what would typically be defined as small. For example, the *M. lini* effector protein AvrM is 314 amino acids long and 36 kDa and the *Glomus intraradices* effector protein SP7 is 36 kDa and 311 amino acids. Both of these are longer than some of the commonly employed definitions of "small" such as < 30kDa (87), 150 amino acids (88), and 200 amino acids (60). In identifying effector candidates in *P. nodorum*, predicted genes were given a cumulative score using criteria relating to secretion prediction, size, cysteine content, proximity to repeats, presence in other isolates, proteomic evidence, oomycete effector RLXR or *ToxA*-like RGD motifs, and expression profile. An advantage of this technique was that a small size and high cysteine content were not required characteristics, but rather contributed to the overall likelihood a protein may be an effector. This technique resulted in the identification of *Tox1*. Sperschneider et al. (89) recently applied a machine learning method to effector prediction from sets of secreted proteins. This method was validated on known effectors, and does not use pre-determined criteria or cutoffs based on perceived effector properties. It is possible that using this method will unveil novel effectors that are not small and cysteine-rich.

The identification of common motifs has been a successful strategy in identifying effectors in oomycete pathogens. For example, the RLXR motif is present in the N-terminal sequence of many known oomycete effectors and has been shown to be necessary for translocation into the host cell. While it has been suggested that there may be common motifs among fungal effectors (90,91), conclusive evidence of this has not yet been found. The RGD sequence in *P. tritici-repentis* and *P. nodorum* effector *ToxA* has been shown to been necessary for localisation into the host cell (92), and it has been suggested that this may be a motif common to other effectors. Interestingly, the RGD motif is not conserved in the *ToxA*-homologue effector in *C. heterostrophus* (93).

1.3.1.2 Protein homology evidence

Effectors are often species-specific and an absence of homologues in related species has been used to identify candidate effectors. While in many cases this still holds true, weak homologues to effectors have increasingly being identified with growing genomic resources, meaning homology to known effectors can also be used to identify novel effectors. One example of this is the necrotrophic effector *ToxA*, which was thought to be present only in *P. tritici-repentis* and *P. nodorum* (via a recent horizontal gene transfer event (75)). More recent studies, using increased genomic resources for related species, have found that *ToxA* is part of a larger gene family with homologues present in many species (93). The *Cochliobolus heterostrophus* homologue of *ToxA*, *ChToxA*, was found to function as a necrotrophic effector in its interaction with the host plant, corn (93). Other examples of effectors with homologues in other species further demonstrate that effectors are not necessarily species specific (94,95). In addition to the general protein sequence databases such as Uniprot and NCI's protein databank, there are some databases of particular use to pathogenomics such as the Pathogen Host Interaction Database, which contains proteins and the results of analysis on their interaction with their hosts. Homology to protein family (Pfam (96)) domains can be used as evidence that a protein is not an effector (24), or in some cases (such as the LysM domain) used as evidence that a protein may be an effector (97).

1.3.1.3 Comparative genomics

Comparative genomics can help to identify novel effectors. In one of the early applications of whole genome sequence data to effector discovery, the genome sequences of multiple isolates of *M. oryzae* were compared and differences associated with differing avirulence phenotypes (98). This resulted in the characterization of 3 effectors: *AVR-Pia*, *AVR-Pii*, and *AVR-Pik/km/kp*. Other studies have also compared isolates of the same species to identify candidate effectors by correlating presence and absence variation with different phenotypes or looking for genes with evidence of positive selection (87). In another method, publically available pathogenic and non-pathogenic species were compared to find genes that were enriched in pathogen genome but absent from non-pathogen genomes (90). The developed pipeline presented candidate effectors for *Fusarium oxysporum*, *Fusarium graminearum*, and *P. nodorum*. Encouragingly, selected sets of likely effector candidates included the known *P. nodorum* and *F. oxysporum* effectors.

1.3.1.4 Genomic context

Whole genome sequences of fungi have helped to illuminate the genomic context of effectors, in many cases showing effectors to be close to repeats or within repeat-rich or unstable genome regions. This information has in turn in some cases been used as a line of evidence that a gene may be an effector. These genome regions are gene-sparse, meaning the set of genes within such regions is generally small. In the case of *L. maculans*, repeat regions are RIP degraded, AT-rich, and harbour numerous effectors and effector candidates (17,66,67). It is possible that knowledge of AT-rich regions may be relevant to the prediction of effectors in other pathogens.

1.3.1.5 Expression profile

Known effectors are up-regulated during infection, and *in planta* expression or an expression profile similar to known effectors can be used as evidence that a gene may encode an effector. For example, in identifying *P. nodorum* effector candidates microarray data was used to identify genes with an expression profile similar to characterised effectors *SnToxA* and *SnTox3*, resulting in the identification of *SnTox1* (87). The advent of RNA-seq has offered another high-throughput method for expression profiling and does not rely on the genome annotation as with microarray experiments, meaning it can be used to identify novel transcripts (99). *In planta* libraries from different infection time point can be used to find genes that are up-regulated during infection. This can then be used as a line of evidence that these genes may encode effectors (100–102).

1.3.1.6 Proteomic data

Protein samples can be analysed using mass spectrometry, returning peptide weights that can then matched to a reference set of protein sequences, evidencing the presence of that protein in the sample. This proteomics data can be used as another line of evidence when finding candidate effectors. For example, proteomics data from proteins secreted by the pathogen can provide additional evidence — for example, alongside secretion prediction — that a protein is secreted (103).

In some cases a particular size fraction of proteins can be separated and shown to trigger a hypersensitive response or disease symptoms when infiltrated into the host plant and proteomic analysis of these active fractions can then be used to identify candidate effectors. This approach was used to analyse active fractions from *P. nodorum*, resulting in the identification of *SnTox3* (104).

1.3.1.7 Structural analysis

There have been some studies that support the notion that effectors that do not share sequence homology may be structurally similar. For example, a similar β-sandwich fold has been observed in the experimentally determined structure of ToxA from *P. tritici-repentis* and AvrL567 from *Melampsora lini* (105). Structural similarity between AvrPiz-t, AvrPia, and Avr1-CO39 avirulence effectors from *Magnaporthe grisea* and ToxB from *P. tritici-repentis* has also been observed in experimentally determined structures (106). This finding was applied to predict novel effector candidates in *M. grisea* by searching for proteins with a predicted secondary structure similar to that seen in the experimentally determined *M. grisea* effectors (106).

1.3.2 Pathogenomics for effector prediction

The selection of effector candidates builds on the foundations of the pathogen genome assembly and the predicted gene set. As such, in addition to the described challenges of selecting putative effectors from larger gene lists, there are challenges presented by generating accurate genome assemblies and gene predictions. Some of these challenges are not limited to pathogenomics and are common to fungal genomics and indeed the genomics of any organism, although some are particularly relevant to pathogenomics and the study of effectors.

1.3.2.1 Genome assembly

Genome assembly is the first step in the studying the genome of an organism. The choice of sequencing platform has a great bearing on both cost and the quality of the resulting assembly. Sanger sequencing was used in early fungal genome sequencing projects (83–85), and generally results in a high quality assembly. However, Sanger sequencing is expensive, time-consuming, and labour intensive (107) and therefore unsuitable for many fungal genome sequencing projects. Next generation sequencing has allowed many fungal species to be rapidly sequenced at low cost, although the resulting assemblies can be fragmented, owing to the difficulty in assembling repetitive regions from short read data. The common result is that the bulk of coding regions are assembled, as is usually judged by the proportion of conserved eukaryotic orthologues present in the genome (108–110), although repetitive regions and genes encoded within repeat rich may not be present within the assembly. Further compounding this, some fungal species have a mechanism called repeat induced point mutation (RIP), which peppers repeat regions with C to T and G to A transitions resulting in AT-rich repeat regions (111). Some sequencing platforms are biased toward higher coverage in higher GC regions (112), meaning these AT-rich repeat regions receive lower sequencing coverage than other regions. In terms of effector prediction, if a gene has not been assembled it will not be included in gene lists from which effectors are selected. When considering that many effector genes have been found within or close repetitive regions and in some cases AT-rich regions (17,66,104), incomplete or fragmented assemblies are particularly at risk of missing effectors (113) or obscuring their genomic context. Recently, PacBio long-read sequencing has emerged as a cost effective sequencing platform. Reads typically have a high error rate, but can either be corrected using low error rate short-reads (114) or, with sufficient coverage, can self-corrected (115). Using self-correction, shorter PacBio reads are aligned to the longer reads consensus corrected reads are formed and subsequently assembled (115). PacBio sequencing and assembly has been applied to fungal pathogens (116), and it has been suggested that this technology may be beneficial to effector prediction by way of better assembly of repetitive regions (113).

1.3.2.2 Gene prediction

Gene prediction forms the basis for downstream analysis of organisms. Gene prediction methods have traditionally been divided into extrinsic and intrinsic or *ab initio* methods. *Ab initio* or intrinsic methods use information contained within the genome alone to make predictions. *Ab initio* methods usually make predictions by distinguishing coding sequences, introns, and intergenic sequences based differing patterns of nucleotide frequencies associated with each feature. Extrinsic methods use information from outside the genome sequence to assist in annotation. This evidence can be in the form of protein homology, proteomic peptide alignments, or transcript sequence alignments. Often methods employ a combinatory approach, either within a single tool or by running multiple different gene prediction tools in a pipeline.

1.3.2.2.1 Extrinsic methods

Protein homology can be used to assist in gene prediction by aligning known proteins from another species to the genome. Regions of protein homology can indicate the presence of a gene, and detailed alignments can delineate intron and exon boundaries. A popular program employing protein homology is GeneWise (117), which has now been replaced by Exonerate. Accuracy depends on the level of homology between the known protein sequence and the homologous version within the genome of interest, with accuracy decreasing with decreasing homology (117). Homology tends to drop off at the start and end of the protein, meaning predictions that are incomplete at the 3' or 5' end of the coding sequence are common (118). Protein sequences in databases are often the result of prediction themselves, and there is a growing risk of propagating errors made in other gene prediction efforts. Furthermore, in fungal genomics and in particular pathogenomics, a high proportion of the genes within an organism are either unique or not present in a well-annotated model organism (85). This means a heavy reliance *ab initio* gene prediction and experimental evidence from the species or isolate of interest is inevitable.

Expressed sequence tags (ESTs) are short (~100bp) sequences of cDNA. These can be aligned to the genome and assist in gene prediction by indicating an expressed region and in some cases intron boundaries. Full-length cDNAs, sequenced using Sanger sequencing, can be informative for gene prediction but are prohibitively costly and rarely used (118). Deep sequencing of cDNA using short read next generation sequencing technologies, termed RNA-seq, has emerged as a cost effective way to improve gene prediction. Reads can be aligned to the genome and assembled into transcripts, or assembled and the resulting transcript sequences subsequently aligned to the genome. Spliced read / transcript alignments clearly delineate intron boundaries, and in many case read coverage spans the full length of the transcribed region. Studies reporting improvements to gene annotation via the use of RNA-seq data are now numerous (119–122). One of the drawbacks of this type of evidence is that it relies on the genes having been expressed under the conditions used to generate the data.

Data from proteomics is another type of data that can be used in gene prediction both to validate gene models predicted by other methods or provide evidence to support gene model corrections or the annotation of novel genes (123). Short peptide hits are typically aligned to the predicted proteome or to a 6-frame translation of the genome or transcriptome. Proteomic data has the distinct advantage over transcript data in that it confirms translation of a protein, rather than just expression, and gives information about the translation frame. Proteomic data does not elucidate the full structure of the gene, and needs to be used in conjunction with other methods. Proteomics can be particularly useful in confirming the translation of short open reading frames within transcribed regions (118), and can be a powerful resource to use in conjunction with transcriptomic data (10,48). This is relevant to the prediction of effector genes, as they are often small and can therefore be difficult to distinguish from randomly occurring ORFs which are not translated.

1.3.2.2.2 Ab initio prediction

Hidden Markov models (HMMs) were first applied to eukaryotic gene prediction in the 90s (124), and since have dominated the domain of *ab initio* gene prediction. HMM gene predictions utilise a set of states based on the known structural features of genes. These states may differ from one predictor to the next but typically include states to describe exons, introns, and intergenic sequences. The most likely series of states to describe a nucleotide sequence is determined computationally. Each state has a content model, which is used to calculate the probability that a particular stretch of nucleotide sequence belongs to that state. For example, for states describing coding sequence, the content model is typically a periodic Markov chain, usually of order 4 or 5 (124,125). As such, parameters are based on the frequency of k-mers 5 or 6 nucleotides long and are calculated separately for each reading frame. The prediction accuracy of HMMs is far from perfect, and annotations based solely on *ab initio* prediction will contain a significant proportion of errors. Training sets of genes are required to derive the parameters necessary to make predictions and it has been shown that predictions are most accurate when derived from the species of interest (126). Gathering a set of genes for training purposes can be a challenge. The HMM gene predictor SNAP addressed this by using a bootstrapping method whereby parameters from a related species were used to make predictions which were then used to train species specific parameters, thereby improving gene prediction accuracy (126). GeneMark-ES was the first eukaryotic HMM gene predictor to employ a self-training method, requiring only the genome of interest as input (127). Extrinsic methods can also be used to generate a training set of genes. Aligning a well-conserved set of proteins to the genome of a novel species (for example, using GeneWise or Exonerate) may deliver a sufficiently large set of genes for training. Transcript data can also be used to obtain a training set of genes structures.

1.3.2.2.3 Pipelines and combinatory approaches

Different prediction methods and implementations tend to have differing strengths and weaknesses. As such, pipelines combining gene prediction software and/or results have been developed such as EVidenceModeler (128), JAMg , SnowyOwl (129), MAKER (130,131), and BRAKER1 (132). AUGUSTUS pioneered the incorporation of EST evidence into HMM gene prediction (133), by awarding gene models in agreement with this data higher probability scores. Protocols for AUGUSTUS have since extended its use to RNA-seq data. The incorporation of RNA-seq data and other forms of evidence into gene prediction continues to be an active area of study.

1.3.2.2.4 Application to effector prediction

Downstream analysis of predicted protein sets to find effector candidates relies on accurate gene prediction to minimise incorrect coding sequences being chosen as effector candidates and ensure genuine effectors are part of the predicted proteome. Gene prediction of effectors has been suggested to be poor (82,113). For example, *Avr5* (*P. fulva*) (48) and *SnTox1* (*P. nodorum*) (R. Syme, personal communication) were both missed by gene prediction software. Furthermore, several properties of effectors suggest gene prediction accuracy could be particularly problematic for effectors. Many effectors have only weak or no homology to proteins in others species, meaning that homology based gene prediction does not help to annotate these genes. Effectors are generally small, and the prediction of small genes and small exons is known to be a challenge in gene prediction. The possible role of HGT, as was the case for *SnToxA* (75), means that codon usage for an effector may be atypical for the species it has been transferred into. It is also possible that effector genes located close to RIP affected or hyper-mutated regions may have differing codon and/or amino acid biases, contributing to prediction inaccuracies. Effectors with an unusually high cysteine content may be a particular challenge for gene prediction. Cysteines are under-represented in coding sequences when considering their expected frequency based on nucleotide frequencies. In a comprehensive gene mining effort carried out on plant species *Arabidopsis thaliana* and *Oryza sativa* it was found that small cysteine-rich proteins had been poorly predicted in previous predicted gene sets, and many novel examples of these resembling antimicrobial peptides were identified (134). This finding has been echoed in fungal gene annotation update papers where the identification of novel small cysteine-rich peptides has been reported (10,135).

1.4 Thesis introduction

The field of effector discovery has progressed substantially with the study of fungal pathogen genome sequences. Despite this progress, the number of pathogen species with genomic resources exceeds the number of species with characterised and cloned effectors. Furthermore, in species where effectors have been identified additional effectors remain undiscovered. There is abundant room for improvement in the prediction of effectors, both in the way putative effectors are short-listed from larger gene lists and in the pathogenomics that precedes this. This body of research makes a novel contribution to fungal effector gene prediction through improvements to gene prediction, an improved understanding of AT-rich regions in fungal genomes, and the provision of improved genomic resources and effector candidate lists for barley pathogens *P. teres* f. *teres* and *P. teres* f. *maculata*.

Chapter 2 supplements this main literature review with a detailed review of research into NEs. For completeness, both secondary metabolite and proteinaceous NEs are reviewed, however the research chapters of this thesis address the prediction of proteinaceous effectors. Many pathogenic species are known to employ NEs, including *P. teres* f. *teres* (Chapter 7). This review highlights similarities between the properties of NEs and Avrs, supporting the use of similar effector prediction methods for proteinaceous effectors from the two groups. Past work by the Centre for Crop and Disease Management (formally the Australian Centre for Necrotrophic Fungal Pathogens) on NEs has been instrumental in breeding resistant wheat varieties in Australia, and these past successes motivate this research.

As discussed, gene prediction is an important step in effector prediction pipelines. Chapter 3, presented as a published paper, explores automated gene prediction in fungal genomes through the incorporation of RNA-seq data. A novel method was used to incorporate RNA-seq derived data into hidden Markov model gene prediction. Benchmarking the resulting tool, CodingQuarry, against high-quality gene sets from model fungal species *S. cerevisiae* and *Schizosaccharomyces pombe* demonstrated 91.3% and 90.4% of genes were predicted correctly (respectively) — 4–5% better than the next best performing competing software. Further benefits of the CodingQuarry methodology were demonstrated in CodingQuarry's ability to handle gene loci with overlapping UTRs, frequently seen in fungal genomes due to their high gene density.

In Chapter 4 the specific application of gene prediction techniques to the prediction of fungal effectors is explored. Known effectors are generally poorly predicted by *ab initio* methods. Effector gene prediction improvements are found using RNA-seq driven methods employed by CodingQuarry and BRAKER1, but only where effector loci are well supported by RNA-seq data. CodingQuarry is extended to target the prediction of secreted effector-like genes. This improves prediction without relying on RNA-seq data at effector loci. This approach can therefore be used to identify effector-like genes that are not expressed under the conditions used for RNA-seq or where RNA-seq data is not available for the species or isolate of interest.

Motivated by reports of effector encoding genes within or close to AT-rich regions, Chapter 5 and Chapter 6 report on an investigation into AT-rich regions in fungal genomes. In this study I aim to develop a consistent method for AT-rich region identification and then apply this to determine the prevalence and properties of AT-rich regions. Chapter 5 describes a pilot study investigating AT-rich regions in *Z. tritici*, which is presented as part of a published review article. Chapter 6 gives a detailed description of a novel method to characterise AT-rich regions. This automates genome segmentation to allow the identification of regions with differing GC-content, and automatically determines whether the genome has distinct AT-rich regions. Importantly, the choice of GC-content cut-off to define AT-rich regions is done on a per genome basis. Chapter 6 also describes the application of OcculterCut in a detailed investigation into the prevalence and properties of AT-rich regions in over 500 fungal genomes. This study demonstrates that AT-rich regions are common within the genomes of fungal species within the Pezizomycotina subphylum. Analysis of the gene content of AT-rich regions highlights possible implications for gene prediction accuracy in AT-rich regions due to differing patterns of amino acid use and atypical codon usage (when compared to coding sequences in GC-equilibrated genome regions).

Chapter 7 describes a practical application of pathogenomics and effector gene prediction methods to two barley pathogens *P. teres* f. *teres* and *P. teres* f. *maculata*. PacBio single-molecule real-time (SMRT) long-read and Illumina short-read genome sequence data is used to generate high quality genome assemblies. The new *P. teres* f. *teres* (isolate Won1-1) assembly was 14.7 Mb larger than the previously published assembly (isolate 0-1). Analysis of the repeat content found the *P. teres* f. *teres* assembly to be 44.2% repeats and the *P. teres* f. *maculata* assembly to be 25.7%. The predictions of CodingQuarry, AUGUSTUS, and GeneMark-ET were combined using EVidenceModeler to generate a high-quality RNA-seq driven annotation of the coding sequences of each species. Additional steps were taken to ensure that putative secreted, RNA-seq supported coding sequences were included in the annotation. Finally, the properties of coding sequences that are typically associated with effectors were assessed and recorded. This provides a resource for future research, as the bioinformatic assessment of putative effectors can be used to complement expression profiles, proteomic data, and comparisons with other isolates.

Chapter 2 Necrotrophic effectors

One of the traditionally held views of necrotrophic pathogens has been that they have simplistic modes of infection when compared to biotrophic pathogens. Biotrophs were thought to have complex host-specific mechanisms by which they subvert and evade host defences, whereas necrotrophs were thought to blindly unleash a tirade of destructive non-host-specific mechanisms to kill host tissues. The complex mechanisms of infection hypothesised for biotrophs have certainly been shown to be true, as is evident when considering the intricate host-specific nature of the avirulence effectors secreted by biotrophic pathogens discussed elsewhere in this chapter. Yet in a revolutionary time for necrotrophic pathogen research, complex species-specific host manipulation has also been shown to function in necrotrophic fungal pathogen-plant interactions.

Amidst this, the idea that pathogens fit neatly into two or even three classes describing their lifestyle has fallen out of favour (50,136). It has been commonplace to describe pathogens as biotrophs or necrotrophs — the former deriving nutrients from live cells and the latter from dead cells. An additional class of hemibiotrophs has been used describe pathogens that infect in a two-phase manner, beginning with a symptomless latent phase before switching to necrotrophy. For example, the hemibiotrophic wheat pathogen *Zymoseptoria tritici* has a latent (biotrophic) phase of more than 10 days before switching to necrotrophy (19), effectively living as both a biotroph and a necrotroph. While these clear-cut lifestyle categories are attractively simple, the defining features of each group have eroded away with anomalies and exceptions — and not only due to the aforementioned insights regarding modes of infection. For example, with no clear definition of what length latent phase qualifies a pathogen as hemibiotrophic, we can easily find examples within the published literature of pathogens receiving conflicting classifications. Furthermore, arguably every necrotroph has at least a short latent phase and could therefore be considered a hemibiotroph (50).

By these arguments it is clear that there would be little value in considering the effectors produced by necrotrophic pathogens as different from the effectors produced by biotrophic pathogens. There is, however, value in considering the sub-set of “necrotrophic” effectors (NEs) as effectors that have a positive or potentially positive effect on pathogenicity (from the pathogen’s point of view) by inducing necrosis or chlorosis in specific genotypes of the host species. By this definition, and in agreement previous publications (45,137), the set of NEs encompasses what have traditionally been termed host-selective toxins (HSTs). These effectors are capable of independently re-producing all or some of the disease symptoms induced by the pathogen and are a diverse group (Table 1), including proteins and structurally diverse secondary metabolites.

Table 1 A list of published fungal necrotrophic effectors.*

Pathogen	Host	Effector	Structure	R-gene	Target	Reference
Cochliobolus heterostrophus	corn	T-toxin	linear polyketide	?	URF13	(138)
	corn	ChToxA	protein	?	?	(93)
Pyronellaea zeae-maydis	oats	PM-toxin	linear polyketide	?	URF13	(139)
Cochliobolus carbonum	corn	HC-toxin	cyclic tetrapeptide	*Hm1*	Histone-deacetylases	(140)
Cochliobolus victoriae	oats	victorin	cyclic chlorinated pentapeptide	*Vb / LOV1*	NB-LRR protein	(141)
Alternaria alternata	Japanese pear	AK-toxin	epoxy-decatrienoic ester	?	?	(142)
	strawberry	AF-toxin	epoxy-decatrienoic ester	?	?	(143)
	tangerine	ACT-toxin	epoxy-decatrienoic ester	?	?	(59)
	apple	AM-toxin	cyclic tetrapeptide	?	?	(144)
	tomato	AAL-toxin	aminopentol ester	*Asc-1*	?	(145)
	rough lemon	ACR(L)-toxin	terpenoid	?	?	(146)
Parastagonospora nodorum	wheat	ToxA	protein	*Tsn1*	ToxABP1, PR1-5	(75,147)
	wheat	Tox1	protein	*Snn1*	?	(87)
	wheat	Tox3	protein	*Snn3*	?	(104)
Pyrenophora tritici-repentis	wheat	ToxA	protein	*Tsn1*	ToxABP1	(148)
	wheat	ToxB	protein	*Tsc2*	?	(149)
Botrytis cinerea	multiple	NEP1-like	proteins	?	?	(150)
Rhynchosporium secalis	barley	NIP1	protein	*Rrs1*	PM ATPase	(151)

** Included are proteinaceous NEs (produced by fungal pathogens targeting plant hosts) that have been cloned and secondary metabolite NEs that have had their structure characterized.*

Some NEs are involved in an inverse gene–for–gene interaction with host susceptibility genes. In an inverse gene-for-gene interaction, the effector gene product interacts with the product of a matching host susceptibility gene, resulting in host susceptibility (45,152). Multiple studies now support that certain pathogens use the host defences to their advantage (153,154) tricking the plant into killing its own cells rather than the pathogen directly killing the plant cells. This can occur in a similar manner to avirulence and resistance (Avr and R) gene product interactions, in that the necrotrophic effector and recognition gene product interact inducing a hypersensitive response (HR). Local cell death, while preventing colonisation by pathogens reliant on live tissue, provides a food source for a pathogen capable of deriving nutrients from dead tissue. As such, presence of both the NE and matching recognition gene lead to a susceptible interaction, and absence of either the NE or the recognition gene confers resistance. This has been frequently termed effector triggered susceptibility (ETS).

2.1 Secondary metabolite host-selective toxins as necrotrophic effectors

Some fungal pathogens produce secondary metabolite toxins that interact with certain hosts to induce infection. These toxins have traditionally been termed host-selective toxins (HSTs), and form a structurally and chemically diverse group (155). The activity of HSTs and their role in infection is sufficient to consider them effectors, and supported by parallels between the function of these toxins and other effectors. Here, some of the well-known secondary metabolite NEs are described.

The first HST to be discovered was AK-toxin in 1933, a necrosis inducing epoxy-decatrienoic acid ester produced by the Japanese pear infecting pathotype of *Alternaria alternata* (156). Deletion of one or more of the genes involved in the biosynethesis of AK-Toxin has been shown to result in loss of virulence (157). Since the discovery of AK-toxin, other pathotypes of *A. alternata* have also been found to produce secondary metabolite NEs (Table 1). In each case the NE is the single determinant of pathogenicity and host range (158). The chemical structures of six of the seven *Alternaria* spp. HSTs have been determined, the exception being AT-toxin from the tobacco infecting pathotype. AK-Toxin and the toxins of the strawberry and pear infecting pathotypes, AT-Toxin and ACT-toxin, are structurally analogous esters of a 9,10-epoxy-8-hydroxy-9-methyl-decatrienoic acid (158).

Victorin is a cyclic pentapeptide produced by *Cochliobolus victoriae*, the fungal pathogen responsible for Victoria blight on oat (*Avena sativa*). Victorin induces cell death in certain varieties of oat leading to susceptibility. The susceptibility gene, named *Vb*, has not been cloned but strong evidence supports it being the same gene that confers resistance to crown rust, caused by the obligate biotroph *Puccinia coronata* (159). The more experimentally tractable *Arabidopsis thaliana* was also found to be sensitive to victorin (160), and the sensitivity gene LOV1, encoding a nucleotide-binding leucine-rich repeat (NB-LRR) protein resembling an R-gene, has been cloned (161).

A related Cochliobolus species, *Cochliobolus carbonum*, causes Northern corn leaf spot and ear rot on corn. *C. carbonum* produces a cyclic tetrapeptide called HC-toxin, which responsible for increased virulence on susceptible corn plants (140). HC-toxin does not elicit defence responses but rather acts to suppress plant defences. Susceptibility is conferred by lack of the *Hm1* gene, which encodes an enzyme that detoxifies HC-toxin (162). *Hm1* was the first plant resistance gene to be cloned. While HC-toxin is produced by a necrotrophic pathogen and is a HST, it is different to the other effectors described here in that it supresses rather than evokes host defences, and in that susceptibility is conferred by the lack rather than presence of a host gene.

Cochliobolus heterostrophus race T produces a polyketide toxin called T-toxin. Race T isolates are highly virulent on corn varieties carrying the Texas male sterile cytoplasm (*T-cms*), due to an interaction between T-toxin and a protein named URF13 present in these cultivars (163,164). This interaction interferes with the functioning of the mitochondria resulting in cell death and increased virulence. *Pyronellaea zeae-maydis* (formally *Mycosphaerella zeae-maydis*) produces secondary metabolite analogous to T-toxin named PM-toxin, which is required for virulence on *T-cms* corn varieties (139).

The wheat pathogen *Pyrenophora tritici-repentis* secretes two well-characterised proteinaceous effectors (discussed in the next section) as well as a less well-characterised secondary metabolite toxin named ToxC (165). ToxC induces chlorosis on susceptible wheat genotype. The structure of ToxC and genes required for its synthesis are as yet unknown.

2.2 Proteinaceous necrotrophic effectors

The necrotrophic effector *ToxA* was identified in wheat pathogen *P. tritici-repentis* in the late 90s and has since been arguably one of the most studied proteinaceous NEs. Sensitivity to *ToxA* results in necrosis and relies on the presence of the recognition gene *Tsn1*. *Tsn1* is one of few NE recognition genes that have been cloned, and encodes a NB-LRR protein. During a sensitive interaction, ToxA localises to the mesophyll cells (166) and binds to a chloroplast localised protein named ToxABP1 and plastocyanin (167,168). Reactive oxygen species then accumulate followed by cell death. Determination of the ToxA crystal structure revealed a solvent exposed loop harbouring an RGD motif (169), which has been shown to be required for translocation into the host cell (92). Until recently*, P. tritici-repentis* could not be transformed, hampering efforts to identify novel effectors due to their symptoms being masked by ToxA (170). The development of a *ToxA* knockout (170) marks a major step forward in the study of this pathogen. A first report of NE epistasis was recently made in the *P. tritici-repentis* – wheat interaction, where ToxA was shown to, in interaction with certain wheat genotypes, reduce the spread of chlorosis typically induced by other NEs (171).

A second *P. tritici-repentis* effector, *ToxB*, has also been cloned (149). *ToxB* encodes a small 6.6 kDa, 64 amino acid long protein that induces chlorosis on susceptible wheat lines. While *ToxA* is ubiquitous in Australian isolates of *P. tritici-repentis*, *ToxB* is generally absent (172). *ToxB* is present in multiple copies in some isolates, with as many as 9 copies estimated to be present in race 5 isolates (173). A *ToxB* homologue, *toxb*, sharing 86% amino acid similarity with *ToxB* was also found in non-pathogenic *P. tritici-repentis* isolates (173).

The *Parastagonospora nodorum*-wheat interaction has developed into a model system for understanding this inverse gene-for-gene interaction. Three *S. nodorum* effectors have, to date, been cloned. The first of these, *SnToxA*, was identified due to its near perfect homology (99.7% amino acid identity) to the characterised *P. tritici-repentis* effector *ToxA*. Strong evidence supports this being an intriguing instance of recent (within the last century) horizontal gene transfer in which *P. tritici-repentis* was the recipient species (75). It has been demonstrated that *SnToxA* interacts with the same wheat susceptibility gene, *Tsn1*, as the *P. tritici-repentis ToxA* (75). ToxA has been shown to interact with a dimeric pathogenesis-related protein named PR1-5 in wheat, with results suggesting a site-specific interaction between the proteins may cause ToxA-induced necrosis (147). In a proteomic and metabolomic study of wheat in response to ToxA, ToxA was found to disrupt photosynthesis and cause an oxidative burst (153). The results of this study suggested that the *P. nodorum* ToxA has a comparable mode of action to *P. tritici-repentis* ToxA (153). Metabolomic profiling of ToxA-infiltrated wheat found plant secondary metabolites were induced in response to the effector, and demonstrated that serotonin acts as a phytoalexin against *P. nodorum* (174).

Following the identification of *SnToxA* in *P. nodorum*, a mapping and proteomics approach led to the identification of *SnTox3* — a gene encoding a 25.8 kDa protein with six cysteine residues that interacts with the wheat resistance gene *Snn3* (104). Subsequent studies showed SnTox3 to induce cell death by a differing mechanism to that employed by ToxA (175).

The identification of *SnTox1* utilised the published genome sequence (8), using a bioinformatic scoring system based on the protein properties, genomic context, and inter-isolate comparisons to arrive at a list of putative effectors (87). This remains an unusual study in the sense that, while publications frequently report putative effectors, the bioinformatic determination of effector candidates and resulting cloned effector were reported as a single publication. *SnTox1* encodes an exceptionally cysteine-rich protein — even by effector standards — containing 16 cysteine residues (13.7% cysteine). Recently, a triple knockout strain was developed that lacked these three known effectors and the virulence of the knockout strain assessed on a range of wheat lines (176). Phytotoxic activity showed that other effectors exist and are secreted into culture filtrate. The continued bioinformatic analysis and recent improvements to genomic resources is likely to illuminate additional novel effectors in this species (10).

It is suspected that proteinaceous NEs similar to those described above also play a role in other plant pathogen interactions. Strong evidence supports the activity of NEs in the barley–*Pyrenophora teres* f. *teres* interaction (46,177–179), but a NE is yet to be cloned from this species. *Z. tritici* is also likely to secrete proteinaceous NEs (180,181) and given the importance of this pathogen (182) and available bioinformatic resources (183), it likely that there will soon be developments in this regard.

The NIP1 effector protein produced by the barley pathogen *Rhynchosporium secalis* is unusual as it is both a NE and an Avr (184). In the presence of the *Rrs1* gene, NIP1 is recognised, triggering defence responses and conferring resistance (151,184). In the absence of *Rrs1*, NIP1 is a virulence factor, inducing necrosis (151,184). This is interesting, as both interactions induce necrosis but with different disease outcomes. Furthermore, different alleles of *NIP1* were found to induce differing levels of toxicity, however the frequency of the different alleles in populations indicated that higher toxicity was not being selected for (184). This example hints to the complex nature of NEs, in that the timing and amount of necrosis may be important. Other NEs of *R. secalis* NIP2 and NIP3 both induce necrosis but are not recognised in avirulence interactions.

2.3 Innate function and NE interactions

Whether NEs have an innate function in the absence of host-recognition, similar to that shown for some Avrs, has been a topic of speculation. Evidence was found to support that in the absence of recognition by *LOV1*, victorin inhibition of TRX-h5 undermined plant defences (154). This was also supported by increased susceptibility to *Pseudomonas syringae* pv. *maculicola*, a biotrophic pathogen, in the presence of victorin (154). This could mean victorin acts as a virulence factor in the absence of host recognition. Slightly more mycelial growth was observed in a *ToxA* expressing *P. tritici-repentis* isolate compared to a non-*ToxA* expressing *P. tritici-repentis* isolate when grown on ToxA-insensitive wheat (171). This result implies that ToxA may have an innate function in the absence of host recognition. Appressoria production has also been correlated with *P. tritici-repentis ToxB* expression, implying that ToxB may additionally function in ways relating to pathogen fitness or host colonisation (185,186).

In pathogens with multiple NEs, variation in the frequency of each of the effectors has been observed. In a study screening *P. nodorum* isolates for presence/absence of effectors significant differences in the NE effector frequencies were observed (187). These differences were hypothesised to reflect differences in the host sensitivity gene repertoire of wheat grown in different regions (187). Whether there is a fitness penalty associated with expressing unrecognised NEs is unknown, however known examples of NEs are highly expressed and in the absence of host recognition it is reasonable to expect that selection would favour gene loss over superfluous expression. As discussed it is possible that in the absence of recognition certain NEs may still confer weak fitness advantages. It could also be possible that some NEs are recognised and trigger a non-beneficial (from the pathogen's point of view) defence response or interaction with other pathogens / organisms they are in contact with.

Origin and evolution

The observation that effectors and avirulence genes are often located close to repeats or repeat dense regions also holds true for many NEs. Repeat activity has been linked to the evolution of effectors. In the case of Avrs this has been associated with the ability to rapidly evolve resistance through mutation, transposon interruption, or deletion of an Avr when pressured by a resistant host

(23,25,26). In reference to NEs, it has been suggested that proximity to repeats may increase their mobility within the genome or between species (188). This could facilitate effector duplications, as are seen in the *ToxB* duplications in races of *P. tritici-repentis* associated with increased virulence (173) or the HGT of *ToxA* discussed above. If there is in fact is some fitness penalty associated with expressing unrecognised effectors, the evolutionary advantage of these being in unstable genomic regions can be explained analogously to Avrs. AT-rich regions, arising from repetitive regions being subjected to repeat-induced point mutation (RIP), have been associated with Avrs being able to rapidly accumulate mutations due to RIP (17,66,71), which can also affect non-repetitive sequence (including coding sequence) close to repeats. A close proximity to AT-rich sequence has also been noted for some NEs. For example, *P. nodorum's SnTox3* that was noted to be close to a region of AT-rich RIP-affected sequence around 10 kb long (104). Similarly, the genes required for the production *C. heterostrophus* secondary metabolite NE T-toxin were noted to exist within 1.2 Mb of AT-rich sequence (189).

As with Avrs, NEs have been observed to be under positive selection (87,150,190). In the *P. nodorum* wheat interaction, variations in *SnToxA* isoforms were found relate to quantitate variations in effector activity (191). These findings supported the hypothesis that variation in *SnToxA* was due to selective pressure favouring increased virulence (191). In a similar manner to Avrs and R genes, it is also likely that positive selection at NE and host recognition gene loci is indicative of the arms race between host and pathogen.

The horizontal transition of the *ToxA* effector gene *from P. nodorum* into *P. tritici* raises the possibility that HGT may have played a key role in the evolution of other pathogens and the recent emergence of certain crop diseases (192). In the case of *ToxA* both pathogens employ *ToxA* as a NE, but it is possible that pathogens employing NEs could have acquired them via HGT from an organism in which they have an avirulence function. HGT has been shown to be more prevalent within the pathogen-rich Pezizomycotina subphylum (77), and it has been hypothesised that HGT may be responsible for the recent emergence of other crop diseases (80).

P. tritici-repentis effectors ToxA and ToxB are part of a small set of proteinaceous effectors that have been structurally characterized (169,193). Interestingly, a β-sandwich fold showing some structural similarity to ToxA has since observed in the structure of avirulence effectors AvrL567 from *Melampsora lini* (105). ToxB was found to have a differing yet analogous β-sandwich structure (193), recently shown to have structural similarity to AvrPiz-t, AvrPia, and Avr1-CO39 avirulence effectors from *Magnaporthe oryzae* (106). These examples of structural similarity between Avrs and NEs are intriguing, and suggest these different effector types may have a common evolutionary origin (106). As more becomes known about the structure of effectors it will be interesting to see which structural features are shared between Avrs and NEs and what role these features play in host recognition.

Necrotrophic effectors and crop protection

Necrotrophic effectors have played a central role in several agricultural disasters. In the early 1940s, certain oat varieties were widely planted in North America due to their resistance to crown rust, caused by the obligate biotroph pathogen *Puccinia coronata*. These same cultivars were highly susceptible to Victoria blight, caused by *C. victoriae,* due to susceptibility triggered by the HST victorin. In the 1970s, an epidemic of southern corn leaf blight caused by *C. heterostrophus* was due to widespread susceptibility triggered by the HST T-toxin.

Knowledge of necrotrophic effectors and the host “resistance” genes they interact with can be used to breed out sensitivity genes, establishing resistant varieties. In Australia the provision of effector ToxA, Tox1, and Tox3 protein solutions to breeders allowed rapid screening of new wheat cultivars for resistance to *P. tritici-repentis* (ToxA) and *P. nodorum* (ToxA, Tox1, and Tox3). This strategy saw the percentage of ToxA-sensitive wheat grown in the area reduce from around 30% to 17% from the 2009–2010 to 2012–2013, equating to a saving of $50 million (82). Furthermore, it has been shown that there is no yield penalty associated with insensitivity to these effectors at least under the conditions prevailing in WA (194). An assessment of the sensitivity of Australian wheat cultivars to the *P. nodorum* effectors ToxA, Tox1, and Tox3 found sensitivity to one or more of the known effectors to be common, and resistance to SNB can be further improved by continuing these effector breeding strategies (195). Further studies of *P. nodorum* have shown that additionally using a culture filtrate containing the remaining as yet uncharacterised effectors is a valid strategy for assessing sensitivity to *P. nodorum* (176). Despite this, functional redundancy exists between effectors, supporting the need to clone the remaining effectors and use these separately (176).

Breeding out a resistance gene conferring sensitivity through interaction with a necrotrophic effector could — in theory — also have the effect of breeding out a functional resistance to a pathogen with the corresponding avirulence gene. The only example of this is *C. victoriae*’s NE victorin and an Avr produced by *P. coronata* which are suspected to target the same resistance gene in oat. In all other cases the targets of necrotrophic effectors are presumed to be resistance genes owing to their leucine rich repeat structure (for example, *Tsn1* and *LOV1*) or role in activating defence responses, but no role in avirulence interactions has been found. These R genes may have evolved to recognise Avr effector employing pathogens (and not necessarily fungi) that are no longer around. As such, in practice breeding out certain plant resistance genes that interact with necrotrophic effectors has been highly effective at reducing losses pathogens infecting through ETS, with no identifiable yield penalties.

2.4 Future directions

Genomics of fungal pathogens promised to find novel NEs, and in some cases whole genome sequencing has delivered on this promise with the characterisation of novel NEs (87). Furthermore, most of the pathogens discussed within this chapter now have whole genome resources publically available, and some have benefited from the incorporation of transcriptomic and proteomic datasets (10). There are clear similarities between Avrs and NEs - both are recognised by matching host genes, effectors within both sets have been observed to be under positive selection, and each set contains examples where effector-encoding genes are located within plastic genome regions. Additionally, proteinaceous effectors within both sets tend to encode small, secreted, cysteine-rich proteins. Recently reported structural similarities between some NEs and Avrs hint at a common evolutionary origin. In light of this, bioinformatic pipelines and techniques for the prediction of NEs are indistinct from those used for the prediction of other effectors. As such, the ability to predict NEs accurately — as with the prediction of all effectors — relies on accurate genome assembly, gene predictions, and the selection of effectors from larger lists. These needs support the focus on gene prediction (Chapter 3 and Chapter 4) and the genomic context of effectors (Chapter 5 and Chapter 6).

Chapter 3 CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts

3.1 Attribution statement

Title:	CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts
Authors:	Alison C Testa, James K Hane, Simon R Ellwood, and Richard P Oliver
Citation:	Testa et al. BMC Genomics (2015) 16:170

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 Ph. D. candidate (ACT) made the following contributions to this chapter:

- ACT wrote the software and performed all bioinformatic analysis
- ACT wrote manuscript with assistance from JKH

The following contributions were made by co-authors:

- RPO and ACT conceived the study
- JKH provided feedback and suggestions about gene prediction methods
- JKH assisted in editing and writing the manuscript
- RPO, JKH, and SRE read and revised the manuscript

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

________________________________ ________________

Alison C Testa Date

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

________________________________ ________________

Richard P Oliver Date
(Principal supervisor and co-author)

________________________________ ________________

James K Hane (co-author) Date

_______________________________ ________________

Simon R Ellwood (co-author) Date

BMC Genomics

SOFTWARE Open Access

CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts

Alison C Testa[1,2*], James K Hane[1], Simon R Ellwood[1] and Richard P Oliver[1]

Abstract

Background: The impact of gene annotation quality on functional and comparative genomics makes gene prediction an important process, particularly in non-model species, including many fungi. Sets of homologous protein sequences are rarely complete with respect to the fungal species of interest and are often small or unreliable, especially when closely related species have not been sequenced or annotated in detail. In these cases, protein homology-based evidence fails to correctly annotate many genes, or significantly improve *ab initio* predictions. Generalised hidden Markov models (GHMM) have proven to be invaluable tools in gene annotation and, recently, RNA-seq has emerged as a cost-effective means to significantly improve the quality of automated gene annotation. As these methods do not require sets of homologous proteins, improving gene prediction from these resources is of benefit to fungal researchers. While many pipelines now incorporate RNA-seq data in training GHMMs, there has been relatively little investigation into additionally combining RNA-seq data at the point of prediction, and room for improvement in this area motivates this study.

Results: CodingQuarry is a highly accurate, self-training GHMM fungal gene predictor designed to work with assembled, aligned RNA-seq transcripts. RNA-seq data informs annotations both during gene-model training and in prediction. Our approach capitalises on the high quality of fungal transcript assemblies by incorporating predictions made directly from transcript sequences. Correct predictions are made despite transcript assembly problems, including those caused by overlap between the transcripts of adjacent gene loci.
Stringent benchmarking against high-confidence annotation subsets showed CodingQuarry predicted 91.3% of *Schizosaccharomyces pombe* genes and 90.4% of *Saccharomyces cerevisiae* genes perfectly. These results are 4-5% better than those of AUGUSTUS, the next best performing RNA-seq driven gene predictor tested. Comparisons against whole genome *Sc. pombe* and *S. cerevisiae* annotations further substantiate a 4-5% improvement in the number of correctly predicted genes.

Conclusions: We demonstrate the success of a novel method of incorporating RNA-seq data into GHMM fungal gene prediction. This shows that a high quality annotation can be achieved without relying on protein homology or a training set of genes. CodingQuarry is freely available (https://sourceforge.net/projects/codingquarry/), and suitable for incorporation into genome annotation pipelines.

Keywords: Generalised hidden Markov model, Gene annotation, Fungi, Gene prediction

* Correspondence: 13392554@student.curtin.edu.au
[1]Centre for Crop and Disease Management, Department of Environment and Agriculture, School of Science, Curtin University, Bentley, WA 6102, Australia
[2]Postal address: Department of Environment and Agriculture Centre for Crop and Disease Management, GPO Box U1987, Perth 6845, Western Australia

Background

Whole-genome sequencing has enabled investigations into the gene content of living many organisms and forms the foundation for further study of gene expression, proteomics and epigenetics. After assembly of a novel genome, gene annotation is often the first step in analysing the gene content of an organism. Accurate annotation of the exonic structure of genes is crucial to the success of all subsequent functional and comparative analyses.

Problems that can potentially be caused by incorrect gene annotation are numerous and can lead to incorrect assessments of the lifestyle and ecology of an organism. In comparative genomics where orthologous genes or conserved functional domains are compared between species/isolates, the estimated numbers of such genes/domains can be distorted by less than perfect annotations (as described by Hane et al. [1], S Text 1). Prediction of extracellular secretion, which can be determined by a short signal peptide at the N-terminus, can miss secreted proteins if the start codon of a gene has been incorrectly annotated. Mis-annotating the start of protein translation could either cut off the signal peptide or bury it within the annotation. While a seemingly benign annotation error, the consequences for downstream research could be detrimental, particularly as the biotic interactions or industrial applications of microbes are largely determined by their secretomes. Additionally, translated protein sequences of novel species are often submitted to databases such as NCBI [2] and Uniprot [3]. It is commonplace to use these database entries to support the annotation of related species or isolates, meaning errors present in the pioneer annotation may be repeated. When these new annotations based on false assumptions are added to databases, there is not only a propagation of errors, but also a perceived strengthening of homology evidence for incorrect protein sequences.

In recent years, correction of *in silico* predicted gene annotations with RNA-seq derived transcripts and read alignments has enabled vastly improved genome annotations and corrections of annotated gene structures [4-6]. Short read and/or assembled transcript alignments are typically used to correct the coordinates of intron-exon boundaries in existing gene annotations or predictions [7], to train gene predictors [8], and can also be incorporated directly into gene prediction by hybrid gene predictors [9,10]. Since their initial application to gene prediction [11], generalised hidden Markov models (GHMMs) have played an important role in genome annotation. Various GHMM gene predictors [12-15] continue to be incorporated into annotation pipelines [16-18], some of which are capable of making use of RNA-seq data. For example, AUGUSTUS [9,10,14] allows the user to generate hint files from RNA-seq read/transcript alignments that are then used to improve prediction accuracy. More recently, a new version of GeneMark-ES [15], named GeneMark-ET [8] allows the incorporation of RNA-seq data into its automated gene model training. These gene finders are both applicable to a broad range of eukaryotic genomes. A number of pipelines have also been developed that utilise available gene prediction software and RNA-seq data to generate annotations. Some examples of such pipelines are Maker [16,19], EVidenceModeler [7], JAMg [20], SnowyOwl [18] and the insect genome annotation pipeline OMIGA [21]. The continued development of pipelines such as these relies on the availability and development of component software such as GHMM gene predictors.

Fungal genomics has applications in areas such as agriculture [22-24], medicine [25,26], biomass conversion [27,28] and food/beverage production [29,30]. This broad industry relevance and the continued growth in the number of new fungal species with sequenced genomes emphasises the importance of fungal gene annotation. Fungal genomes differ from those higher eukaryotes in that they are gene dense with short introns [31,32]. They also exhibit less alternate splicing when compared to other eukaryotes, with a higher proportion of mRNA isoforms arising from retained introns [33]. Manual annotation is considered to be the most reliable method of producing a high quality genome annotation, but this is time consuming and can be a bottleneck in genome studies [34]. Consequently fungal genome annotations are typically derived from *ab initio* predictions, spliced EST/transcript alignments and protein homology [34]. For many fungi, closely related species have either not been sequenced or their genomes have not been annotated in detail. This can mean that sets of homologous proteins for use in protein homology annotation are either small or unreliable. In such cases, gene prediction relies more on EST/transcript alignments and *ab initio* predictions.

Currently available gene prediction software and pipelines are typically intended for application across a broad range of eukaryotes, with comparatively few being specific to fungi. GipsyGene [35] is a GHMM gene predictor that was developed for fungi, with particular attention given to modelling fungal introns correctly. A version of GeneMark-ES [15], a self-training GHMM, also uses an intron model designed for fungi. However, neither of these incorporates RNA-seq data. SnowyOwl [18] is a recently developed pipeline designed specifically to annotate fungal genomes using RNA-seq data and homology information. Although designed for fungi, SnowyOwl selects from GHMM predictions made by AUGUSTUS [9,10,14], a gene predictor that was optimised for application across a broad range of eukaryotes.

In this study we present the gene prediction tool CodingQuarry. It is designed to make protein-coding gene sequence predictions through the use of assembled or aligned RNA-seq transcripts in both GHMM training and prediction. CodingQuarry is differentiated from other gene predictors by the combined use of gene predictions made directly from both transcript and genome sequences.

The choice to tailor CodingQuarry to the prediction of fungal genes and to use assembled, aligned transcripts rather than raw read alignments relates to some key differences between fungal genomes and those of higher eukaryotes. Firstly, fungi exhibit significantly less alternative splicing than higher eukaryotes. Consequently, the task of transcript assembly is simpler, resulting in a higher proportion of correctly assembled full-length transcripts [36]. Secondly, fungi have smaller introns than higher eukaryotes [32]. Recent studies indicate short introns are reconstructed in transcript assembly with a higher success rate than long introns [37]. These transcript assembly advantages make it feasible to generate coding sequence annotations directly from assembled transcript sequences, a process that is more likely to be error prone in higher eukaryotes.

A major consequence of the high gene density observed in fungi is a high proportion of instances whereby the untranslated regions (UTRs) of adjacent transcripts overlap in terms of their positions on genomic DNA. Overlap can be between 3′ and 5′ UTRs of adjacent genes on the same strand, or between 5′ and 5′ or 3′ and 3′ UTRs of adjacent genes on opposite strands. Overlaps from the latter example, particularly in the case of 3′ to 3′, are referred to as sense-antisense (S-AS) overlaps. S-AS overlaps have been observed to occur rarely in many species, but are widespread in fungi [38,39]. Essentially this means that in gene-dense fungal genomes, mapped RNA-seq reads belonging to adjacent genes may support regions of coverage that span two or more loci. This is a more severe problem when ‘unstranded’ RNA-seq chemistries are used, as S-AS overlaps can be distinguished through the use of stranded RNA-seq data. CodingQuarry is designed to work with assembled, aligned transcripts derived from either stranded or unstranded RNA-seq data and to specifically address the problem of merged transcripts, such that these transcript assembly errors do not translate to coding sequence annotation errors or omitted gene loci.

For the purpose of demonstrating CodingQuarry’s performance we have selected two exemplar fungal species, which possess highly reliable sequence and annotation resources: *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe. S. cerevisiae*, commonly known as Baker’s yeast, has long been a model organism and is important to the wine making, baking and brewing industries. *Sc. pombe*, commonly known as fission yeast, is also a model organism. These two species are estimated to have diverged from a common ancestor up to 1000 million years ago [40,41] and are representative of distantly related fungal sub-phyla. In this study we have used the high-quality annotations of these fungi to benchmark the sensitivity and specificity of CodingQuarry, and compare it to other gene predictors.

Implementation

Data sets for benchmarking

To test the accuracy of predictions made by CodingQuarry and other gene predictors, we utilised assembled genome sequences, RNA-seq reads and up-to-date gene annotations of two model fungi: *S. cerevisiae* and *Sc. pombe*.

The *Sc. pombe* (isolate 972h-) genome, annotation and protein sequences were downloaded from PomBase [42] and RNA-seq reads [SRA: SRX040571] were downloaded from NCBI [43]. The reads were trimmed using Cutadapt [44], aligned to the genome using TopHat [45,46] (version 2.0.19, −-mate-inner-dist 280, −-mate-std-dev 70, −-min-intron-length 10, −-max-intron-length 5000, −-min-segment-intron 10, −-max-segment-intron 5000) and assembled using Cufflinks [47] (version 2.1.1, −-min-intron-length 10, −max-intron-length 5000, −-overlap-radius 10, −-min-isoform-fraction 0.4, −-library-type fr-firststrand). The RNA-seq data used for *Sc. Pombe* was stranded (i.e. the strand of genomic DNA that produced the mRNA fragment is known). To simulate a transcript assembly from unstranded RNA-seq data, TopHat and Cufflinks were also re-run as above with the modified parameter ‘−library-type fr-unstranded’.

The *S. cerevisiae* (isolate S288c) genome, annotation and protein sequences were downloaded from the Saccharomyces Genome Database [48] and RNA-seq reads [SRA: SRR1198662-8] were downloaded from NCBI. Reads were trimmed, aligned and assembled using the same method as described above for *Sc. pombe* (stranded only, −-mate-inner-dist 200, −-mate-std-dev 40).

Although both *Sc. pombe* and *S. cerevisiae* are annotated to a high standard, it was desirable to identify a stringent subset of their genes that are of very high-confidence. This is because not all genes are verified to the same degree, and some are therefore more likely to be accurate than others. It is still possible that the full annotations contain errors that are artefacts of the prediction tools, data and methods used to generate them. Comparing predictions against a high-confidence set excludes some annotations that are lower confidence, and is likely to give a better assessment of the accuracy of gene predictors. Annotations within these high-confidence subsets were required to exactly match sequences in Uniprot’s [3] reviewed database and to be listed and as possessing protein level evidence. There were 1,898 of these for *Sc. pombe* and 5,224 for *S. cerevisiae*. Nevertheless, as CodingQuarry’s intended purpose is to predict genes across entire fungal genomes, we also report its performance benchmarked to the less stringent full datasets of 5,124 *Sc. pombe* genes and 6,575 *S. cerevisiae* genes.

CodingQuarry prediction method outline

CodingQuarry predicts genes in 2 stages. The first stage involves prediction of genes directly from transcript sequences derived from regions of the genome supported by RNA-seq in GFF (General Feature Format) [49], such as derived from Cufflinks [47]. The second stage complements the first and involves additional predictions based on genomic sequences. In both stages GHMMs are used to predict genes, however, these differ in their structure and in how they incorporate RNA-seq data into their predictions. The GHMMs used in both stages are also trained automatically using the RNA-seq data. The final predicted annotation produced by CodingQuarry is a combination of predictions made in stages 1 and 2.

Stage 1: Training and prediction from transcript sequences

The coordinates of transcribed regions (in GFF format) relative to the assembled genome sequence are used to extract the sequences of a set of virtually spliced transcripts (i.e. intron sequences are removed). A generalised hidden Markov model (GHMM) is used to make gene predictions directly from this set of transcript sequences. Predicted coding-sequences are then converted back to their relative genomic coordinates, with transcript splicing being accounted for in this process.

The GHMM used in stage 1 uses fixed length states to describe the gene start and Kozak sequence [50] and gene stop codon, and variable length states to describe gene coding sequences, UTRs, and non-coding transcripts. To address the issue of merged transcripts, this model allows a single transcript sequence to contain multiple genes, via the creation of a "middle UTR" state. Where UTRs of adjacent transcripts overlap in terms of their relative corresponding positions on the genomic DNA, a single transcript sequence as derived from RNA-seq can contain multiple gene loci. A pictorial example of this is shown in Figure 1, section Bi, in which the middle UTR state is used to allow the correct prediction of two genes on the same strand within a merged transcript sequence. In the case of unstranded RNA-seq, prediction errors arising from transcript sequences merged due to S-AS UTR overlap are corrected in stage 2.

The coding regions are modelled using a fifth-order, three-periodic Markov chain. The 5′, 3′ and 'middle' UTRs, as well as non-coding transcripts are modelled using a fifth-order (non-periodic) Markov chain. A second-order weighted array matrix over a region of 11 nucleotides up to and including the ATG start codon models the Kozak sequence and gene start. Length distributions of the coding region state, UTR states and non-coding transcript state are modelled using smoothed length frequencies.

A self-training method is used, where parameters are initially estimated from the longest open reading frame (beginning with a methionine) in each transcript. The GHMM is then successively run and retrained twice to refine the parameters. There are some restrictions placed on the sequences that are used for retraining, based on the general principle of preferential exclusion of some correct sequences rather than risking including false-positives. Training of the "gene" state is therefore restricted to coding sequence lengths greater than or equal to 600 nucleotides to guard against the inclusion of false-positive predictions in the training set. Similarly, open reading frames in UTR regions greater than or equal to 300 nucleotides are removed from the UTR training set to guard against the inclusion of coding sequences. Where there are overlapping genes in the prediction, the longer gene is retained in the training set and the shorter overlapping gene(s) are discarded.

Importantly, this method is distinct from methods where the transcript/EST alignment is used to inform a GHMM prediction from genome sequence. The main advantage of the initial prediction from transcript sequences is that the predicted annotation will have intron boundaries that agree exactly with the intron boundaries in the transcripts to genome alignment. Another advantage is that where the transcript assembly indicates that there is an alternative splicing, prediction from transcripts allows the coding sequences splicing alternatives to be predicted.

Stage 2: Prediction from the genome sequence

After the prediction from transcript sequences is carried out in stage 1, there may still be a number of errors and omissions in the predicted annotation (see Figure 1, section D). These predictions are therefore added to, and in some cases replaced by predictions made from genome sequence.

The stage 1 predicted gene set is used to train a second, different GHMM which is designed to make predictions from assembled genome sequence. This genome based GHMM includes additional states to model introns, a feature not previously required in the spliced transcript-based GHMM used in stage 1. Another difference is that the GHMM used for transcript sequences models the 5′ and 3′ UTR regions, whereas the GHMM used for prediction from genome sequence models these regions as part of larger "intergenic" regions. The stage 2 GHMM intron model used has fixed length states for the donor, acceptor and branch point sequences, modelled by first-order weighted array matrices. Variable

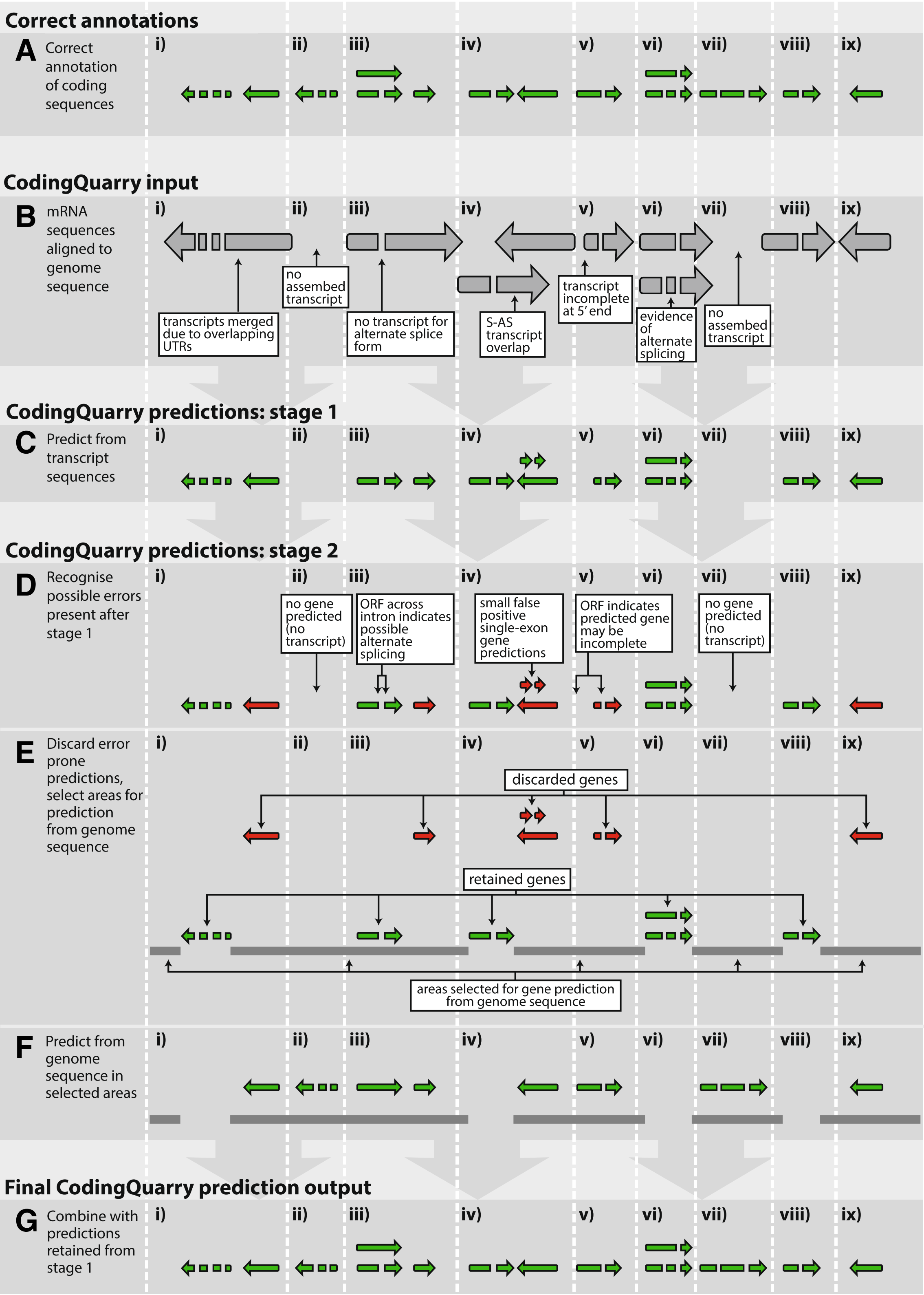

Correct annotations
A Correct annotation of coding sequences
i) ii) iii) iv) v) vi) vii) viii) ix)
CodingQuarry input
B mRNA sequences aligned to genome sequence
transcripts merged due to overlapping UTRs
no assembed transcript
no transcript for alternate splice form
S-AS transcript overlap
transcript incomplete at 5′ end
evidence of alternate splicing
no assembed transcript
CodingQuarry predictions: stage 1
C Predict from transcript sequences
CodingQuarry predictions: stage 2
D Recognise possible errors present after stage 1
no gene predicted (no transcript)
ORF across intron indicates possible alternate splicing
small false positive single-exon gene predictions
ORF indicates predicted gene may be incomplete
no gene predicted (no transcript)
E Discard error prone predictions, select areas for prediction from genome sequence
discarded genes
retained genes
areas selected for gene prediction from genome sequence
F Predict from genome sequence in selected areas
Final CodingQuarry prediction output
G Combine with predictions retained from stage 1

Figure 1 (See legend on next page.)

(See figure on previous page.)

Figure 1 CodingQuarry flow diagram. Examples are shown of correct annotations of coding sequences, **(A)** and a typical CodingQuarry input; assembled transcripts aligned to the genome **(B)**. The stages used within CodingQuarry to predict coding sequences are shown **(C-G)**. Firstly, coding sequences are predicted from transcript sequences (introns are removed) using a GHMM **(C)**. Possible prediction errors after this step are coloured red, and notes show how these are identified **(D)**. These error prone predicted genes are discarded **(E)**, and regions are selected for prediction from genome sequence **(F)**. The resulting prediction is output by CodingQuarry **(G)**, which merges the retained predictions from transcript sequences **(E)** with the predictions from selected areas of the genome sequence **(F)**. Sections of the example genome sequence and annotations have been labelled **i-x** in each part of the diagram **(A-G)**, and marked with vertical dotted lines. These sections are labelled to facilitate in-text references to the diagram in the Implementation section of this manuscript. Labels **i-x** correspond to the same genome sections through **A-G**.

length states are used to model the regions between these fixed length states. The intron model is based on research showing that fungal introns have high information content in the 5′ splice site, 3′ splice site and branch point regions [32], and is similar to the intron model used by GeneMark-ES [15]. During training, the acceptor/donor lengths are automatically adjusted by CodingQuarry to suit the fungi being predicted on. The acceptor/donor is extended to the furthest out nucleotide position (up to a maximum length) with a statistically significant difference in nucleotide composition when compared to the adjacent intron region. A Chi-square test (p-value 0.01) is used to test for statistical significance. The acceptor and donor are taken to extend 2 nucleotides into the adjacent exon, and can be up to a maximum length of 22 nucleotides. During prediction, the lengths of these states is fixed. The maximum intron size is set as 10% longer than the longest intron evidenced by the transcript alignment, unless this value is greater than 5,000, in which case the maximum is limited to 5,000. The user can choose to disable the intron model length restrictions of CodingQuarry in order to allow it to be used for species with longer intron lengths.

In prediction from transcript sequences (stage 1), the location of introns is inferred from the transcript to genome alignment, and the assembled transcript sequences are used to model the UTRs. When predicting genes from genome sequence in stage 2, RNA-seq data is also incorporated in GHMM prediction, but in a different way. Where there is supporting evidence from RNA-seq data, the prediction of introns is restricted by the transcript alignment. Intron boundaries (donor and acceptor sites) are disallowed in areas where there is an aligned transcript sequence on the same strand. This restriction is relaxed within 50 nucleotides of the transcript end, where introns may be predicted *ab initio*, in the same manner as in regions without evidence of transcription. Introns are only allowed to occur where the first 2 nucleotides of the intron donor and last two nucleotides of the intron acceptor sites are GT and AG respectively.

After the stage 1 genes are used for training, certain predicted genes that are likely to be inaccurate are discarded and areas of the genome are selected for prediction from genome sequence. Discarded stage 1 predicted genes include single-exon genes and genes suspected to be incomplete (described in detail below). The areas selected for prediction from genome sequence are the areas flanking the retained stage 1 genes, as well as loci where alternative splice forms may exist. These steps, and the motivation for them, are discussed in more detail in the following paragraphs, and Figure 1, sections D, E and F give examples and summarise this process.

Where an assembled RNA-seq transcript, aligned relative to the genome sequence, overlaps another assembled transcript on the opposite strand, the transcript's predicted UTR can contain all or part of the coding sequence from the adjacent transcript on the opposite strand. In stage 1, genes are predicted in a single direction in a single transcript, that is, although multiple genes are permitted to be predicted in a single transcript, they must all be in the same direction. As a result, where prediction from transcript sequences is carried out on UTR regions containing coding sequence on the opposite strand, we have observed a tendency to predict small false-positive single-exon genes (see Figure 1, section Div). This is because the reverse-complement of a coding sequence has a slightly higher G:C content and contains fewer stop codons than typically occur in UTRs, therefore these regions are often a closer match to the coding sequence model. This problem occurs even more frequently when unstranded RNA-seq is used and adjacent transcripts on opposite strands are assembled into a single locus. To correct this, all single-exon genes from stage 1 are discarded and predictions from genome sequence are carried out in those regions. Although single-exon genes are used for training, this is restricted to coding sequences over 600 nucleotides so that these small false-positives do not contaminate the training set. When genes are predicted from genome sequence in step 2, the prediction is allowed to be on either strand and these false-positive predictions therefore do not occur, leading to better results.

Transcript assemblies are likely to contain some low coverage, incomplete transcripts (Figure 1, section Bv). Attempting to predict a complete gene from an incomplete transcript sequence can lead to errors due to absent start or stop codons (Figure 1, section Dv). If the transcript is incomplete at the 3′ end and the gene's stop codon is outside the transcript sequence then it is likely that no gene will be predicted from the transcript. If the 5′ end of the transcript is incomplete then the predicted gene will have an incorrect start codon, or be completely missed (Figure 1, section Dv). In these circumstances, a prediction from genome sequence is likely to be more accurate. Where the open reading frame of a coding sequence predicted in stage 1 can be extended beyond the bounds of the supporting assembled transcript, there is the possibility that the assembled transcript sequence and resulting predicted coding sequence are incomplete at the 5′ end. Such genes are therefore identified as genes that are suspected to be incomplete. Therefore, the stage 1 prediction is removed and stage 2 genome-based predictions are then carried out (Figure 1, sections D-Fv). Any intron sites supported by the partial transcripts will restrict the location of introns predicted in step 2 and the gene prediction is thus operating as an RNA-seq informed predictor, rather than completely *ab initio*.

In an effort to identify alternate splicing during stage 2, if the removal of an intron can extend an ORF across it without terminating at a stop codon, additional predictions from genome sequence in stage 2 are allowed in these regions (Figure 1, sections D-Fiii). This process allows correct predictions to be made where a transcript has been assembled with a false-positive intron, or where an alternatively spliced transcript retaining the intron sequence was not included in the transcript assembly, possibly due to low RNA-seq abundance.

In addition to correcting some of the inaccuracies in gene prediction from transcript sequences, prediction from genome sequence allows *ab initio* prediction of any genes that were not expressed under the experimental conditions used (Figure 1, sections ii and vii). Gene predictions of this kind are *ab initio* and therefore subject to greater uncertainties. In light of this, the final outputs of CodingQuarry make note of whether a final gene prediction was derived from transcript (stage 1) or genome-based (stage 2) prediction processes.

Post prediction filtering

The final stage of annotation that CodingQuarry carries out is the removal of genes likely to be false-positive predictions. Any gene with a coding sequence that translates to less than 30 amino acids is removed from the annotation. Where alternative splice variants are predicted, only variants with at least one unique intron, or 10 or more unique amino acids are retained. Finally, any gene predicted overlapping a larger gene on the opposite strand is removed where less than 20% of its coding sequence lies outside the bounds of the larger gene. As discussed earlier, false-positive predictions of this kind a common where transcripts overlap one another. While nested genes of this kind are known to occur, they are considered to be rare [51].

Gene discovery

Often one of the key interests of RNA-seq studies for annotation purposes is to discover previously unannotated genes in areas with evidence of transcription. For example, laterally transferred genes, which are of high relevance in fungal genomics [52,53], may be missed in homology or GHMM-based predictions due to a lack of homologs in closely related species or atypical codon usage patterns. To assist in this process, CodingQuarry forces a gene prediction in transcripts that have no overlapping gene prediction after the complete annotation run. This uses the same hidden Markov model as in stage 1, however the probability of a state transition to a non-coding transcript state is set to zero. These genes are not intended to be included in the main set of predicted annotations and are output separately as a set of "dubious" genes. Further efforts to verify which of these genes are genuine could include searches for pfam/antifam domains [54,55], blast searches to databases or experimental verification. However, this set is certain to contain a high proportion of false-positive genes, in part due to open reading frames occurring by chance within non-coding transcripts.

Merged transcripts

One of the final outputs of CodingQuarry reports the IDs of assembled transcripts suspected to be instances of transcripts merged in assembly due to overlapping UTRs. This output is based on the genes predicted by CodingQuarry, and reports the number and DNA strand orientations of the theoretical constituent transcripts. Reporting the orientation is important for unstranded RNA-seq data, where instances of sense-antisense (S-AS) overlap between UTRs can lead to transcripts on opposite strands assembling into single loci.

Training and running other gene predictors for benchmarking

Comparisons were made with AUGUSTUS [9,10,14], and TransDecoder [56]. AUGUSTUS (using hints) and TransDecoder both leverage RNA-seq data and as such have comparable features with CodingQuarry. Though GeneMark-ET also uses RNA-seq data to assist annotation, comparisons were not possible at the time of submission due to its application to fungi being under development. It is important to note that GeneMark-ET uses RNA-seq data to assist in automated training, rather than to also subsequently inform and influence predictions.

AUGUSTUS was trained using the online training server [57] taking a FASTA file of assembled transcripts (in this case from TopHat-aligned RNA-seq read coverage generated by Cufflinks) and the genome sequence as input. This pipeline uses PASA [17] to generate a training set of genes from the transcript data, aligns the transcripts to the genome and uses hints generated from the alignment to assist in gene prediction. This pipeline does not train an untranslated region (UTR) model from assembled RNA-seq transcripts. Intron hints were also generated directly from the read to genome alignment generated by Tophat, however predicting with the hints produced by the training server produced predictions with better sensitivity and specificity when compared to the accepted annotations, and these results were therefore used for comparisons with CodingQuarry.

TransDecoder predicts genes from transcript sequences and uses the transcript-to-genome alignment to place predictions on the genome. Pfam domain searches are also used by TransDecoder to support gene predictions. TransDecoder was run using the TopHat/Cufflinks transcript assembly as per the instructions on the cited web page [56].

Quantifying gene prediction accuracy

Measures of nucleotide, exon, intron and gene sensitivity and specificity, as described by Burset and Guigo [58], were used to compare the high-confidence sets with the various predictions. Sensitivity is the proportion of a given feature (nucleotides/exons/introns/genes) in the high-confidence set that are correctly predicted. Specificity is the proportion of features in the predicted set that are correct (i.e. exactly match the high-confidence set). A correct nucleotide prediction was defined to be a nucleotide within a predicted coding region that is also within a coding region of the high-confidence set. An incorrect nucleotide prediction was defined to be a nucleotide within a predicted coding region that is within an intron or intergenic region in the high-confidence set. A correct exon/intron was defined to be where the exon/intron boundaries in the predicted set were an exact match to the exon/intron boundaries in the high-confidence set. An incorrect exon/intron was defined to be where the exon/intron boundaries in the predicted set did not exactly match one of the exons/introns in the high-confidence set. A gene was defined to be correctly predicted if the gene was exactly the same as in the high-confidence set, and incorrect if the high-confidence set did not contain gene that matched exactly.

Where comparisons were made with the full set of genes in the annotation, all genes in the prediction and in the annotation were used to calculate the values of sensitivity and specificity. Where comparisons were made with the high-confidence annotation subsets, the region over which each of these values were calculated was bounded by the high-confidence set gene boundaries, and any overlapping gene in the predicted set.

Results and discussion

Sensitivity and specificity values were calculated at the nucleotide, exon, intron and gene-level for CodingQuarry predictions and predictions made by TransDecoder and AUGUSTUS. Comparisons were made between predictions and high-confidence subsets (Table 1), and the full sets (Table 2) of *Sc. pombe* and *S. cerevisiae* gene annotations. CodingQuarry can be seen to outperform the other gene predictors in many of the measures. Impressively, CodingQuarry achieved a ~90% gene-level sensitivity when comparing predictions with the high-confidence subsets. This means that CodingQuarry predicts around 90% of the high-confidence set genes perfectly, which is around 4-5% more than the next best gene-level sensitivity result, belonging to AUGUSTUS (with hints), and around 10% better than TransDecoder, which also makes predictions from transcript sequences.

An important consideration is that although both CodingQuarry and AUGUSTUS both use GHMMs, CodingQuarry operates very differently to AUGUSTUS. The main difference is that CodingQuarry combines predictions made initially from transcript sequences together with predictions from genome sequences. We assert that this is an important point in favour of CodingQuarry being considered for wider incorporation into automated annotation pipelines. Consensus between the predictions of different programs/tools can strengthen the confidence in the gene structure, particularly where genes are predicted by different methods. For example, CodingQuarry and AUGUSTUS predict 4,294 genes *Sc. pombe* genes identically, 95.0% of which exactly match the *Sc. pombe* annotation. In the case of *S. cerevisiae*, CodingQuarry and AUGUSTUS predict 4,813 genes identically, 95.4% of which are correct. This demonstrates that these

Table 1 Comparisons between predictions and high-confidence gene sets for *Sc. pombe* and *S. cerevisiae*

	Nucleotide		Exon		Intron		Gene	
	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp
Sc. pombe (1898/5124 genes in high-confidence set)								
CodingQuarry	**99.3**	**99.7**	**93.4**	**93.6**	94.5	96.7	**91.3**	**89.0**
AUGUSTUS	99.2	99.1	92.0	91.4	**95.7**	92.6	86.9	88.9
TransDecoder	95.4	99.3	84.5	86.3	88.5	**97.0**	80.2	73.5
S. cerevisiae (5206/6575 genes in high-confidence set)								
CodingQuarry	**99.2**	**99.8**	**90.0**	90.0	**79.0**	67.6	**90.4**	91.1
AUGUSTUS	97.5	99.7	84.7	**90.9**	74.4	**77.0**	85.0	**91.5**
TransDecoder	92.2	99.5	79.9	74.8	73.9	67.4	80.1	68.0

Sensitivity (Sn) is the proportion of a given feature (nucleotides/exons/introns/genes) in the high-confidence set that are correctly predicted. Specificity (Sp) is the proportion of features in the predicted set that are correct. Sensitivity and specificity calculations for nucleotides are made on nucleotides within coding regions. Further descriptions of these measures are given in the Implementation subsection titled "Quantifying prediction results". The highest scores in each column for *Sc. pombe* and *S. cerevisiae* are shown in boldface.

Table 2 Whole-genome comparisons between predictions and current *Sc. pombe* and *S. cerevisiae annotations*

	Nucleotide		Exon		Intron		Gene	
	Sn	Sp	Sn	Sp	Sn	Sp	Sn	Sp
Sc. pombe (all 5124 genes)								
CodingQuarry	**98.6**	98.9	**90.3**	89.4	92.6	95.2	**87.5**	83.0
AUGUSTUS	98.0	**99.3**	89.0	**90.6**	**94.2**	92.7	83.1	**87.7**
TransDecoder	93.4	99.2	80.8	85.4	85.3	**96.6**	76.3	72.5
S. cerevisiae (all 6575 genes)								
CodingQuarry	**97.2**	99.5	**76.1**	87.2	**64.4**	65.8	**76.6**	88.3
AUGUSTUS	95.4	99.6	71.1	**88.9**	60.5	**69.3**	71.5	**89.8**
TransDecoder	87.8	**99.7**	67.1	75.0	60.5	70.1	67.8	68.0

Sensitivity (Sn) is the proportion of a given feature (nucleotides/exons/introns/genes) in the annotation that are correctly predicted. Specificity (Sp) is the proportion of features in the predicted set that are correct. Sensitivity and specificity calculations for nucleotides are made on nucleotides within coding regions. Further descriptions of these measures are given in the Implementation subsection titled "Quantifying prediction results" The highest scores in each column for *Sc. pombe* and *S. cerevisiae* are shown in boldface.

subsets of genes have a higher specificity than either of the programs do individually, and can be considered higher confidence. If gene predictors operate in very similar ways, the fact that predictions agree is less significant.

The improved accuracy of CodingQuarry over alternative gene predictors is not achieved through protein homology-based prediction or refinement. Accurate gene predictions are therefore achievable when reliable sets of homologous proteins are not available. Such situations can arise when considering newly sequenced fungi, where closely related fungal species have not been sequenced or well annotated. However, if reliable homology evidence is available, CodingQuarry's results have the potential to be further refined and improved by post-prediction annotation tools that merge predicted annotations with multiple sources of supporting evidence, such as EVidenceModeller [7] or Maker2 [19].

The closest competitor to CodingQuarry is AUGUSTUS, which derives all its gene predictions from genome sequences. However, when predicting genes from gene-dense genomes, the short intergenic distances make it possible for an intergenic region between two adjacent genes to be falsely annotated as an intron thus predicting a single merged gene where there should be two or more separate genes. We observed 32 and 25 instances of this in the AUGUSTUS predicted gene sets for *Sc. pombe* and *S. cerevisiae* respectively. When predicting directly from transcript sequences with CodingQuarry this is unlikely to occur, as introns are not predicted during stage 1 and adjacent genes would therefore need to be separated by an ORF to be falsely predicted as a single gene. As such, we see just one case of this error occurring in CodingQuarry predictions for *S. cerevisiae*, and two for *Sc. pombe*. This demonstrates an advantage to using CodingQuarry when annotating gene-dense fungal genomes. Notably, this advantage is also observed for TransDecoder, which also predicts from transcript sequences, with no cases of this error in the *S. cerevisiae* prediction and just one in *Sc. pombe*. However, TransDecoder achieved a much lower overall quality of prediction, with a ~10% lower sensitivity and ~10-20% lower specificity than CodingQuarry when compared to the high-confidence subsets and full sets of annotations (Tables 1 and 2). TransDecoder is intended to be used as part of a prediction pipeline and generates a set of genes to be used for training gene predictors. It is important to note that for its intended purpose, TransDecoder performs extremely well. However, based on the results shown in Tables 1 and 2, CodingQuarry was able to generate a larger and more accurate training set of genes.

As explained in the methods section, the predictions made by CodingQuarry are a combination of predictions from transcript sequences (stage 1), and predictions made from genome sequence (stage 2). A filtering step then removes genes likely to be false-positive predictions. The gene-level sensitivity and specificity of CodingQuarry, when compared to full *Sc. pombe* datasets, after each of these stages is displayed in Figure 2A. Figure 2A shows that the initial step of creating a training set using the longest ORF in each transcript has low values of sensitivity and specificity. An ~8% gene-level sensitivity and ~6% specificity improvement to predictions is made in stage 1, where these annotations are replaced by GHMM predicted genes. Part of the reason for this is that during stage 1, multiple genes predictions are allowed to be made within a single transcript, allowing a large number of genes residing in incorrectly "merged" transcripts to still be predicted. The second prediction stage again results in a jump in prediction accuracy, this time improving the gene-level sensitivity by ~8% and specificity by ~2%. This is due to the addition of genes predicted *ab initio* in regions without RNA-seq transcript coverage and the prediction of genes in regions where the transcript assembly is incomplete. Single-exon genes are also re-predicted stage 2. The final filtering step gives the final output CodingQuarry prediction. This step serves to improve specificity via the removal of false-positive genes, and therefore had little effect of the gene-level sensitivity (Figure 2A).

We observed variation in the accuracy of gene predictions made by all assessed gene predictors when comparing the results for *Sc. pombe* with those for *S. cerevisiae*. Fungal species have many complex differences relating to characteristics such as the number and size of introns [32], prevalence of alternative splicing [59], and gene density [60]. It is therefore reasonable to expect that gene prediction accuracy may vary across differing fungal species, and this can be seen in occurring in other published studies [15]. For predictions generated by

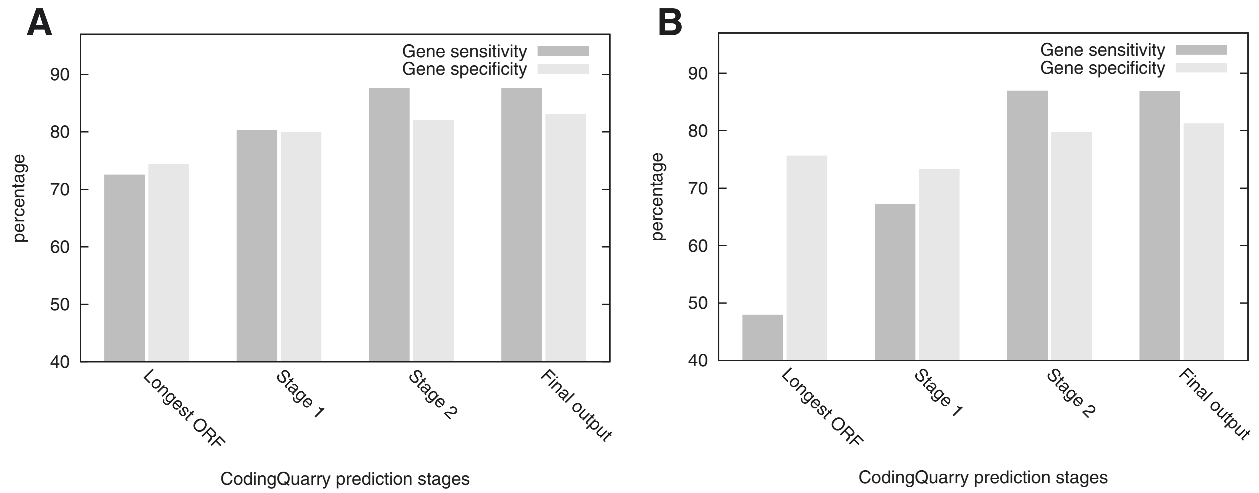


Figure 2 Changes in CodingQuarry prediction accuracy at various stages of prediction of *Sc. pombe* genes. The gene-level sensitivity and specificity is shown at various stages (See Figure 1 and Methods) within a CodingQuarry run. Results show comparisons with *Sc. pombe* where **A)** (left-hand panel) RNA-seq data strand information was used and **B)** (right-hand panel) strand information was ignored**.** Longest ORF is the initial training set, found by taking the longest open reading frame in each transcript to be a gene, stage 1 predictions are made from transcript sequences, stage 2 adds to and replaces some of stage 1 predictions by predicting from genome sequence. Filtering of likely false-positive genes (see Implementation section) takes place before a set of predicted genes is output as the "final output". This output is the annotation generated by CodingQuarry.

CodingQuarry, a possible explanation is the contribution of RNA-seq evidence and how this could influence prediction accuracy. In the case of *Sc. pombe*, around 84% of the predicted genes result from a stage 1 transcript based predictions. However, the stage 1 component of the predicted genes is around 5% lower in S. cerevisiae. As these predictions RNA-seq driven, they are expected to be higher confidence, and it is therefore reasonable to expect the results to be better for *Sc. pombe* than *S. cerevisiae*.

Although stranded RNA-seq data is now readily available, a large quantity of non-stranded RNA-seq data is publically available. It is therefore important that CodingQuarry can deal with transcript assemblies resulting from either stranded or unstranded RNA-seq. Figure 2 shows gene-level sensitivity and specificity of *S. pombe* gene predictions made at stages within CodingQuarry with RNA-seq data where stranded information was ignored (Figure 2B), and where stranded information was included (Figure 2A). Gene level sensitivity and specificity for CodingQuarry's final output predictions on *Sc. pombe* were less than 1% and 2% different between unstranded and stranded runs (respectively) (Figure 2). This result supports of the efficacy of CodingQuarry in overcoming issues in unstranded RNA-seq datasets. Comparisons between Figure 2A and B show that CodingQuarry predictions using the unstranded transcript assembly showed a ~25% improvement in gene level sensitivity going from stage 1 to stage 2 – further supporting the validity of the various processes employed in stage 2 to correct for annotation errors. We surmise that this is a direct result of sense-antisense (S-AS) transcript overlap resulting in merged transcripts composed of transcripts on opposite strands. This confounds prediction from transcript sequence, where genes are expected to be in the same direction as the transcript. As explained in the methods section, and evident from Figure 2A, this is corrected in stage 2 leading to comparable final outputs.

CodingQuarry reports on assembled transcripts which, according to the coding sequence predictions, may be multiple transcripts merged together in the assembly process. Where stranded RNA-seq is used, this is only a problem for overlapping transcripts on the same strand. For the *Sc. pombe* stranded RNA-seq experiment, there were 507 instances reported by CodingQuarry of likely transcript fusions. Of these, 64 were suspected to be the result of fusion of more than 2 transcripts. For *S. cerevisiae* there were more fusions detected: 1,060, with 452 of those suspected to result from the fusion of more than 2 transcripts. Given that different organisms of the same phyla can have very different gene densities and spacing, the higher number of fusions present in the *S. cerevisiae* transcript assembly is not surprising. Where transcripts are assembled from unstranded RNA-seq, there is the possibility of merged transcripts arising from S-AS transcript overlap. Although the splice sites in the transcript-to-genome alignment can help to separate these transcripts, it remains a problem where one or more of the transcripts align without introns. For *Sc. pombe*, the version of the transcript assembly generated without using strand information contained 1,219 instances where one transcript was suspected to be the fusion of multiple transcripts. 630 of these were suspected to be instances of transcripts fusions involving transcripts on opposite strands.

CodingQuarry has been designed for and tested on fungal genomes. It achieves a higher level of accuracy than competing methods by mixing predictions made from assembled transcript sequences with predictions made from assembled genome sequences. In theory, changes to the intron model used for prediction to allow the prediction of longer introns when predicting genes from assembled genome sequence would allow CodingQuarry to be applied to higher eukaryotes. However, in practice, the transcript assembly quality for RNA-seq datasets from higher eukaryotes does not result in enough correctly assembled full-length transcripts for this method to be advantageous. The limitations of transcript assembly quality to gene prediction have been previously noted [8]. Examples of factors contributing to this are the RNA-seq alignment/assembly being complicated by larger introns, and a higher prevalence of alternative splicing, as discussed in the Background section of this manuscript. It is therefore the opinion of the authors that it is unlikely that CodingQuarry would deliver similar improvement in genomes of higher eukaryotes as in fungal genomes, however this is something that may be explored in future studies.

CodingQuarry also outputs an additional set of "dubious" genes, as candidates for gene discovery. As described in the methods section, these genes are forced predictions in transcripts that, after running CodingQuarry steps 1 and 2, do not have an overlapping coding sequence prediction. 632 "dubious" genes are predicted for *Sc. pombe*, and 444 for *S. cerevisiae*. Of these, 25 and 16 overlap a gene in the annotation of *Sc. pombe* and *S. cerevisiae* respectively in the same coding frame. BLAST [61] was used to search for alignments between the protein sequences of dubious genes predictions with no coding sequence shared with genes in the annotation, and NCBI's non-redundant database. Seven of these *Sc. pombe* genes aligned to an entry in nr with a protein level identity of 40% or better and e-value less than 10^{-5}. Of these, six lay completely within a gene annotation on the opposite strand. For *S. cerevisiae,* 21 novel genes aligned to an entry in nr with a protein level identity of 40% or better and e-value better than 10^{-5}, 10 of which lay completely within a annotated gene on the opposite strand. This result can either be viewed as the possibility of unannotated proteins in the test genome annotations, or, possible contamination of the nr database with translated sequences from non-coding RNA. We hope that this feature will assist researchers in gene discovery, however these predictions should be treated cautiously and we do not recommend their inclusion in a formal annotation dataset or submitted to databases without further validation.

Conclusions

We have demonstrated the success of our method of using RNA-seq derived data in GHMMs for fungal gene prediction. For researchers studying the genomes of newly sequenced fungi, for which protein homology resources are absent or unreliable, CodingQuarry can be used as a single step in predicting protein-coding gene sequences with high accuracy. For more detailed annotation efforts, CodingQuarry offers an appropriate starting point for further refinement of annotations with additional supporting evidence.

Availability and requirements

Project name: CodingQuarry.
Project home page: https://sourceforge.net/projects/codingquarry/.
Operating system(s): Platform independent.
Programming language: C++.
Other requirements: OpenMP.
License: GNU.
Any restrictions to use by non-academics: No.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

ACT and RPO designed the study. ACT developed the software, performed bioinformatics analysis. ACT and JKH wrote the manuscript. RPO, SRE and JKH revised the manuscript. All authors read and approved the final manuscript.

Acknowledgements

Many thanks to Rob Syme for trialling CodingQuarry during its development and providing valuable feedback about usability and outputs. This research was undertaken with the assistance of resources provided at the NCI Specialised Facility in Bioinformatics at The University of Queensland through the National Computational Merit Allocation Scheme supported by the Australian Government. The studentship of ACT was supported by an Australian Postgraduate Award (APA), funded by the Australian Federal Government, and the Grains Research & Development Corporation (GRS10564).

Received: 15 September 2014 Accepted: 13 February 2015
Published online: 11 March 2015

References

1. Hane JK, Anderson JP, Williams AH, Sperschneider J, Singh KB. Genome sequencing and comparative genomics of the broad host-range pathogen Rhizoctonia solani AG8. PLoS genetics. 2014;10:e1004281.
2. Coordinators NR. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2013;41:D8–20.
3. Consortium TU. Update on activities at the Universal Protein resource (UniProt) in 2013. Nucleic Acids Res. 2013;41:D43–7.
4. Zhao C, Waalwijk C, De Wit PJGM, Tang D, Van Der Lee T. RNA-Seq analysis reveals new gene models and alternative splicing in the fungal pathogen *Fusarium graminearum*. BMC Genomics. 2013;14:21.
5. Bruno VM, Wang Z, Marjani SL, Euskirchen GM, Martin J, Sherlock G, et al. Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq. Genome Res. 2010;20:1451–8.
6. Cerqueira GC, Arnaud MB, Inglis DO, Skrzypek MS, Binkley G, Simison M, et al. The *Aspergillus* Genome Database: multispecies curation and

incorporation of RNA-Seq data to improve structural gene annotations. Nucleic Acids Res. 2014;42:D705–10.

7. 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:R7.
8. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acid Res. 2014;42:1–8.
9. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics (Oxford, England). 2008;24:637–44.
10. Stanke M, Schöffmann O, Morgenstern B, Waack S. Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC bioinformatics. 2006;7:62.
11. Kulp DHD, Reese MG, Eeckman FH. A generalized hidden Markov model for the recognition of human genes in DNA. In: Proc Int Conf Intell Syst Mol Biol:. 1996;1996:134–42.
12. Lukashin AV, Borodovsky M. GeneMark. hmm: new solutions for gene finding. Nucleic acids research. 1998;26:1107–15.
13. Korf I. Gene finding in novel genomes. BMC bioinformatics. 2004;5:59.
14. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003;19:ii215–25.
15. Ter-Hovhannisyan V. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome. 2008;18:1979–90.
16. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 2008;18:188–96.
17. Haas BJ. Improving the *Arabidopsis* genome annotation using maximal transcript alignment assemblies. Nucleic Acids Res. 2003;31:5654–66.
18. Reid I, O'Toole N, Zabaneh O, Nourzadeh R, Dahdouli M, Abdellateef M, et al. SnowyOwl: accurate prediction of fungal genes by using RNA-Seq and homology information to select among ab initio models. BMC Bioinformatics. 2014;15:229.
19. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. 2011;12:491.
20. Papanicolaou A: Just Annotate My genome (JAMg). Volume 1. CSIRO; 2014. doi: 10.4225/08/54AB80F5105DE.
21. Liu J, Xiao H, Huang S, Li F. OMIGA: optimized maker-based insect genome annotation. Mol Genet Genomics. 2014;289:567–73.
22. Dean R. The Top 10 fungal pathogens in molecular plant pathology. Plant Pathol. 2012;13:414–30.
23. Oliver RP, Solomon PS. New developments in pathogenicity and virulence of necrotrophs. Curr Opin Plant Biol. 2010;13:415–9.
24. Ellwood SR, Syme R, Moffat CS, Oliver RP. Evolution of three *Pyrenophora* cereal pathogens: recent divergence, speciation and evolution of non-coding DNA. Fungal Genet Biol. 2012;49:825–9.
25. Van Den Berg M, Albang R, Albermann K, Badger JH, Daran J-M, Driessen AJM, et al. Genome sequencing and analysis of the filamentous fungus *Penicillium chrysogenum*. Nat Biotechnol. 2008;26:1161–8.
26. Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, et al. The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. Science (New York, NY). 2005;307:1321–4.
27. Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, Baker SE, et al. Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). Nat Biotechnol. 2008;26:553–60.
28. Dashtban M, Schraft H, Qin W. Fungal bioconversion of lignocellulosic residues; opportunities & perspectives. Int J Biol Sci. 2009;5:578–95.
29. de Vos WM. Advances in genomics for microbial food fermentations and safety. Curr Opin Biotechnol. 2001;12:493–8.
30. Cullen D. The genome of an industrial workhorse. Nat Biotechnol. 2007;25:189–90.
31. Galagan JE, Henn MR, Ma L-J, Cuomo C, Birren B. Genomics of the fungal kingdom: insights into eukaryotic biology. Genome Res. 2005;15:1620–31.
32. Kupfer DM, Drabenstot SD, Buchanan KL, Lai H, Zhu H, Dyer DW, et al. Introns and splicing elements of five diverse fungi introns and splicing elements of five diverse fungi †. 2004.
33. McGuire AM, Pearson MD, Neafsey DE, Galagan JE. Cross-kingdom patterns of alternative splicing and splice recognition. Genome Biol. 2008;9:R50.
34. Haas BJ, Zeng Q, Pearson MD, Cuomo CA, Jennifer R. Approaches to fungal genome annotation. Mycology. 2012;2:118–41.
35. Neverov AD, Gelfand MS, Mironov AA. GipsyGene : a statistics-based gene recognizer for fungal genomes. Biophysics. 2003;48:S71–5.
36. Grabherr MG, Haas BJ, Yassour M, Levin JZ, Thompson D, Amit I, et al. Full-length transcriptome assembly from RNA-Seq data without a reference genome. Nat Biotechnol. 2011;29:644–52.
37. Steijger T, Abril JF, Engström PG, Kokocinski F, Abril JF, Akerman M, et al. Assessment of transcript reconstruction methods for RNA-seq. Nat Methods. 2013;10:1177–84.
38. Wang L, Jiang N, Wang L, Fang O, Leach LJ, Hu X, et al. 3' untranslated regions mediate transcriptional interference between convergent genes both locally and ectopically in *Saccharomyces cerevisiae*. PLoS Genet. 2014;10:e1004021.
39. Guida A, Lindstädt C, Maguire SL, Ding C, Higgins DG, Corton NJ, et al. Using RNA-seq to determine the transcriptional landscape and the hypoxic response of the pathogenic yeast Candida parapsilosis. BMC Genomics. 2011;12:628.
40. Hedges SB. The origin and evolution of model organisms. Nat Rev Genet. 2002;3:838–49.
41. Forsburg SL. The yeasts Saccharomyces cerevisiae and Schizosaccharomyces pombe: models for cell biology research. Gravit Space Biol Bull. 2005;18:3–10.
42. Rhind N, Chen Z, Yassour M, Thompson D, Haas BJ, Habib N, et al. Comparative functional genomics of the fission yeasts. Science (New York, NY). 2011;332:930–6.
43. Leinonen R, Sugawara H, Shumway M. The sequence read archive. Nucleic Acids Res. 2011;39:D19–21.
44. Martin M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet journal. 2011;17(1):10.
45. 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:R36.
46. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics (Oxford, England). 2009;25:1105–11.
47. Trapnell C, Williams B, Pertea G, Mortazavi A, Kwan G, Van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28:511–5.
48. Christie KR, Weng S, Balakrishnan R, Costanzo MC, Dolinski K, Dwight SS, et al. Saccharomyces Genome Database (SGD) provides tools to identify and analyze sequences from Saccharomyces cerevisiae and related sequences from other organisms. Nucleic Acids Res. 2004;32:D311–4.
49. GFF (General Feature Format) specifications document https://www.sanger.ac.uk/resources/software/gff/spec.html.
50. Kozak M. Point mutations define a sequence flanking the AUG initiator codon that modulates translation by eukaryotic ribosomes. Cell. 1986;44:283–92.
51. Kumar A. An overview of nested genes in eukaryotic genomes. Eukaryot Cell. 2009;8:1321–9.
52. Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H, Faris JD, et al. Emergence of a new disease as a result of interspecific virulence gene transfer. Nat Genet. 2006;38:953–6.
53. Marcet-Houben M, Gabaldón T. Acquisition of prokaryotic genes by fungal genomes. Trends Genet. 2010;26:5–8.
54. 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:bas003.
55. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42:D222–30.
56. Haas BJ, Papanicolaou A. TransDecoder (Find Coding Regions Within Transcripts) http://transdecoder.github.io.
57. Hoff KJ, Stanke M. WebAUGUSTUS–a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Res. 2013;41:W123–8.
58. Burset M, Guigó R. Evaluation of gene structure prediction programs. Genomics. 1996;34:353–67.
59. Gru KO. Fungal alternative splicing is associated with multicellular complexity and virulence: a genome-wide multi-species study. DNA Res. 2013:1-13.
60. Ralph A, Dean AL-P, Kole C, editors. Genomics of Plant-Associated Fungi: Monocot Pathogens. New York: Springer; 2014.
61. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.

Chapter 4 Effector gene prediction

4.1 Attribution statement

Title: Effector gene prediction

Authors: Alison C Testa, Richard P Oliver, and James K Hane

This thesis chapter is submitted in the form of a collaboratively-written manuscript.

The Ph. D. candidate (ACT) made the following contributions to this chapter:

- ACT wrote the software and conducted the bioinformatics analysis
- ACT summarised the data and generated the figures and tables
- ACT wrote the manuscript

The following contributions were made by co-authors:

- RPO, JKH, and ACT conceived the study
- RPO and JKH read and reviewed the manuscript

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

________________________________ ________________

Alison C Testa Date

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

________________________________ ________________

Richard P Oliver Date
(Principal supervisor and co-author)

________________________________ ________________

James K Hane (co-author) Date

4.2 Abstract

Fungal phytopathogens interact with their host via the secretion of a battery of effector proteins. While accounting for a tiny portion of the full gene content these organisms, our interest in these effector genes is disproportionately high due to their crucial role in infection. In attempting to locate novel effectors, downstream analysis often relies on automated gene prediction methods to form gene lists from which candidate effectors are chosen. In this study, a selection of gene prediction tools were run on four plant pathogen species possessing known effectors (*Passalora fulva*, *Leptosphaeria maculans*, *Magnaporthe oryzae*, and *Fusarium oxysporum* f. sp. *lycopersici*) using publically available genome sequences and RNA-seq data. The success of different gene prediction software/methods at predicting the 34 known effectors contained within these genomes was compared. Results showed the prediction of effectors to be poor and that correct prediction of effectors was reliant on RNA-seq support at effector gene loci. Even so, some effectors, such as the *P. fulva* effector *Avr4E*, were still missed by RNA-seq driven methods. To address this issue, a novel "pathogen" mode of gene prediction via the software tool CodingQuarry, named CodingQuarry-PM, was developed (https://sourceforge.net/projects/codingquarry/). CodingQuarry-PM specifically targets the prediction of secreted, cysteine-rich genes with atypical codon usage. Using this method, most effectors that had been previously missed by gene predictors were correctly predicted. Further application to 9 fungal phytopathogen species revealed previously unidentified genes that may encode effectors.

4.3 Background

Fungal plant pathogens cause billions of dollars of crop losses worldwide each year and jeopardize food security (3). Plants are generally considered to have two lines of defence against such pathogens. The first involves recognition of ubiquitous pathogen associated molecular patterns (PAMPs) and danger- or damage-associated molecular patterns (DAMPs), whereas the second involves recognition of specific molecules produced by the pathogen called effectors (25,26). Fungal effectors interact with the plant host to assist the pathogen in avoiding host recognition, promote or cause increased virulence, and/or trigger host defence responses. Knowledge of fungal effectors has been a powerful tool in breeding resistant crop varieties (82), motivating continued efforts to discover novel effectors.

Before whole genome resources became available, effectors were typically located using map-based cloning techniques (66) or after direct protein purification. With the advent of whole genome sequencing and ever reducing sequencing costs, the formulation of lists of candidate effectors from predicted gene sets has been increasingly used in searching for proteinaceous effectors (24,87,88,196). Known effectors have been observed to be small, secreted, and often cysteine-rich, motivating the frequent use of criteria based on these characteristics in candidate searches (86,87). Recently, a machine learning method has been successfully employed as an alternative to imposing *ad hoc* cut-offs on protein size and cysteine content, finding tryptophan levels and charge to also be informative in distinguishing effectors from other secreted proteins (89). Other criteria such as a close proximity to repeats or unstable genome regions, high dN/dS ratios, long adjacent intergenic distances, a lack of orthologues, homology to known effectors, up-regulation *in planta*, and presence and absence variation across isolates have also been used (86,87). Despite progress in this area, narrowing down a predicted gene set to a smaller set of effector candidates remains a bioinformatic challenge.

Gene prediction forms the basis for many lines of downstream investigation, including effector candidate prediction. In the past, gene prediction strategies were classed as *intrinsic* or *extrinsic,* the former using information gleaned from the genome of interest alone and the latter relying on outside information such as protein homology and expressed sequence tags (118,197). In recent years this distinction has blurred, with many programs and pipelines mixing methods and evidence types to maximize gene prediction accuracy (128,130). The increase in genome sequencing has led to far more comprehensive and taxonomically broad protein databases available to provide protein homology evidence in gene prediction. Additionally, RNA-seq (deep sequencing of cDNA) has revolutionized gene prediction, with read/transcript to genome alignments delineating gene loci and intron boundaries. On the heels of deep transcriptome data, numerous gene prediction programs and pipelines have emerged or been updated to incorporate it with conventional methods, offering unprecedented gene prediction accuracy (129,132,198).

While gene prediction accuracy has greatly improved in recent years, it has been suggested that the prediction of effector genes is unreliable (82,113). Although there has been no formal assessment of the accuracy of gene prediction on known effectors, some of the gene/protein characteristic that are typically associated with effectors have been linked to gene prediction accuracy. Effectors are often species-specific or lack close homologues, meaning the incorporation of homology evidence does not improve their prediction. Gene prediction in the absence of supporting homology relies heavily on hidden Markov model or similar methods, which have inherent inaccuracies. Gene prediction on small genes and small exons is known to be less accurate (197), which is relevant to effectors and indeed much of the pathogen secretome. Effectors are often cysteine-rich, which in certain plant species has been observed to be associated with genes being mistaken for intergenic sequence (134). Furthermore, there are now numerous publications of genome updates where annotation improvement has resulted in the addition of numerous small, cysteine-rich proteins to their gene sets (10). Gene prediction trains on patterns of short stretches of nucleotides (typically 5 or 6 nucleotides long) and as such incorporates patterns of codon usage present within the set of genes used for training. Genes with atypical codon usage can be problematic for gene prediction. Effectors have frequently been observed to reside in rapidly mutable genome regions and found to be under diversifying selection (17,42,87,190). These factors can affect codon usage and may contribute to parameters from training sets being ill fitted to the coding sequences of effectors. In some cases this is due to an accumulation of C to T mutations by a process called repeat-induced point mutation (RIP), prevalent within the pathogen rich Pezizomycotina subphylum (17,65). Recent lateral gene transfers, such as the *Parastagonospora nodorum* to *Pyrenophora tritici-repentis ToxA* effector transfer (75), retain the codon usage patterns of their original genome, which may differ from their recipient genome and thus impede gene prediction. As such, it is possible that inaccuracies in prediction of effectors go beyond the typical inaccuracies associated with gene prediction.

Gene prediction accuracy confounds our ability to bioinformatically select effector candidates. Where some of the coding sequence is predicted but there are errors in the prediction, assessment of the coding sequence with reference to effector criteria can be impeded. For example, if the start site of the gene is incorrectly predicted, the subsequent prediction of a signal peptide may produce an incorrect result due to the signal peptide being cut short or buried within the predicted protein sequence. Other measures based on the protein coding sequence, such as cysteine levels, can also be distorted when the coding sequence is only partially correct. The potential downstream effects of an incorrect gene prediction depend on how much of the coding sequence in correctly predicted and what downstream analysis is conducted. Genes that are “missed”, that is, mistakenly annotated as intergenic sequence, do not appear in predicted gene sets and are as such excluded from effector candidate lists.

In previous work, contributing to the improvements in the automated incorporation of RNA-seq data into gene prediction, we developed CodingQuarry, a self-training gene predictor that used a novel method of incorporating RNA-seq data into hidden Markov model prediction (198). Our goal was to provide a tool capable of accurately predicting genes in fungal species, without requiring training gene sets or protein homology evidence that is often unavailable when researching novel species. This is particularly necessary in the study of pathogens, which typically have a large species-specific gene complement. We then made use of genomes and annotations of model fungal organisms *Saccharomyces cerevisae* and *Schizosaccharomyces pombe*, demonstrating the success of CodingQuarry at fungal gene prediction. In this work we extend our application of CodingQuarry to fungal phytopathogens, emphasising the prediction of known effector genes. We include a “pathogen” mode of operation (CodingQuarry-PM), which models the portion of coding sequence encoding the signal peptide separately to the mature peptide coding sequence to particularly suit genes encoding secreted proteins. Parameters are trained to suit to genes with a high-cysteine content and atypical codon usage, further suiting this model to effector-like genes likely to be challenging for normal prediction strategies. The new prediction method currently focuses on predicting effector-like genes that have been missed in previous rounds by conventional gene predictors (usually mistaken for intergenic sequences) by attempting to predict novel genes from the intergenic regions between previously annotated loci.

4.4 Methods

4.4.1 Generating predicted gene sets

RNA-seq reads for each species were downloaded from NCBI's SRA (199) (Table 2). Reads were quality trimmed and filtered for adapter sequences using Trimmomatic (200) (ILLUMINACLIP:adapters.fa:2:30:10, LEADING:30, TRAILING:30, SLIDINGWINDOW:5:30 MINLEN:35), aligned to their respective genomes using TopHat (v. 2.1.0) (122,201) (--min-intron-length 10, --max-intron-length 5000, --min-segment-intron 10, --max-segment-intron 5000) and assembled using Cufflinks (v. 2.2.1) (202) (--min-intron-length 10, --max-intron-length 5000, --overlap-radius 10). Where there were multiple RNA-seq runs for a single species, the resulting sequence alignment files were merged together into a single file to supply as input to BRAKER1 (v. 1.6) (132), but assembled separately (using Cufflinks) and the assemblies subsequently merged (using the Cuffmerge tool from Cufflinks).

Table 2 RNA-seq data accessions

Species	Isolate	NCBI accession	Bases	Conditions
Leptosphaeria maculans	IBCN18	SRR1130246	132.6M	*in planta* 7 days post inoculation
		SRR1141251	227M	*in planta* 7 days post inoculation
		SRR1151403	2G	*in planta* 14 days post inoculation
		SRR1151404	2.2G	*in planta* 14 days post inoculation
		SRR1151407	6.2G	*in vitro* 7 day cultures grown in oilseed rape medium
Passalora fulva	CBS 131901	SRR1171044	1.2G	*in vitro* 3 day potato-dextrose broth, 24 hours Gamborg B5 liquid medium
		SRR1171045	1.2G	*in vitro* 3 day potato-dextrose broth, 24 hours Gamborg B5 liquid medium minus nitrogen
		SRR1171046	1.2G	*in vitro* 4 day potato-dextrose broth
Fusarium oxysporum f. sp. lycopersici	4287	SRR190806	3G	Unspecified
Magnaporthe oryzae	70-15	SRR081552	3.9G	Unspecified
		SRR081553	3.9G	Unspecified
		SRR081554	3.4G	Unspecified
		SRR081555	3.9G	Unspecified
		SRR081556	3.6G	Unspecified

De novo repeat identification was carried out on each genome separately using RepeatModeler (203) and the resulting repeat libraries used to soft-mask each of the genomes using RepeatMasker (204). CodingQuarry (198) (v. 2.0) (default parameters, with –d specifying unstranded RNA-seq transcripts), BRAKER1 (v. 1.6) (132) (--fungus, --softmasking), GeneMark-ES (205) (--fungus, --soft

1000, --min_gene_prediction 90), AUGUSTUS (v. 3.2) (124,133,206) (default parameters), and GeneMark-ET (v. 4.30) (207) (--fungus, --soft_mask 1000, --min_gene_prediction 90) were each run. CodingQuarry, GeneMark-ES, and GeneMark-ET self-train. BRAKER1 is a pipeline that trains AUGUSTUS using predictions made by GeneMark-ET and then runs AUGUSTUS using intron hints. The training parameters generated by the BRAKER1 pipeline were also used to run AUGUSTUS a second time without RNA-seq derived hints, that is, as an *ab initio* predictor. The longest open-reading frame beginning a methionine within each Cufflinks assembled transcript was also recorded as a type of gene prediction for benchmarking.

4.4.2 Benchmarking predictions

Predicted genes were compared to the genuine annotation of known effector genes and the prediction success of each effector defined as correct, missed, or partial. The prediction was defined as "correct" where the predicted coding sequence exactly matched the genuine annotation. A gene was defined as "missed" where no predicted coding sequence overlapped it (in the same frame). "Partial" was used to describe where the predicted gene did not exactly match the correctly annotated version but there was in-frame overlap between the predicted gene coding sequence and the correctly annotated gene coding sequences. For transcripts reconstructed by Cufflinks, a "correct" reconstructed transcript was taken to be where all exons of the coding sequence of the gene were covered by the exons of an assembled transcript and the transcript-to-genome alignment delineated intron boundaries in agreement with those in the coding sequence annotation. In many cases transcripts of adjacent loci are merged into a single transcript due to having overlapping UTRs. In these cases the transcript assembly was still described as "correct". An assembled transcript was described as "partial" where the coding sequence has some coverage by a transcript, but not all exons/introns were supported. Finally, a gene was "missed" by the RNA-seq derived transcripts where no assembled transcript overlapped the coding sequence of the gene.

4.4.3 CodingQuarry's "pathogen" mode

CodingQuarry was extended to attempt to predict effector-like genes that are missed by either the normal run-mode of CodingQuarry or other prediction methods. This new "pathogen" run mode, referred to from here on as CodingQuarry-PM (https://sourceforge.net/projects/codingquarry/), uses an effector-suited gene model to predict genes in regions annotated as intergenic sequence. This run mode can either be used after the standard CodingQuarry run mode or with a set of externally derived annotations as input. By this method, CodingQuarry-PM does not attempt to correct incorrectly predicted gene models, but rather specifically addresses the problem of genes that are completely missed. In addition to having utility in searching for genes missed in prior annotation efforts, this method is suited to where annotations are transferred from a reference isolate to the genome of novel pathogen isolate. As different pathogen isolates of the same species are known to have differing effector complements, this method can be used to predict effectors that are present only in the non-reference isolate.

CodingQuarry uses a hidden Markov model method to predict genes, described in detail in the Methods section of Chapter 3. In creating a gene model to better describe genes encoding secreted, effector-like proteins, additional states were introduced to model the signal peptide, summarised in Figure 3 A. In this model, the signal peptide was allowed to cover either part of the first coding exon of a gene, or in a multi-exon gene, the first exon and part of the second exon (see examples in Figure 3 B). Predictions are not restricted to secreted genes only, however the states used describe the coding sequence are designed to be better suited to genes encoding a protein with a signal peptide. As such, genes predicted in this way should not be assumed to be secreted and secretion prediction should follow on translated protein sequences as part of downstream analysis.

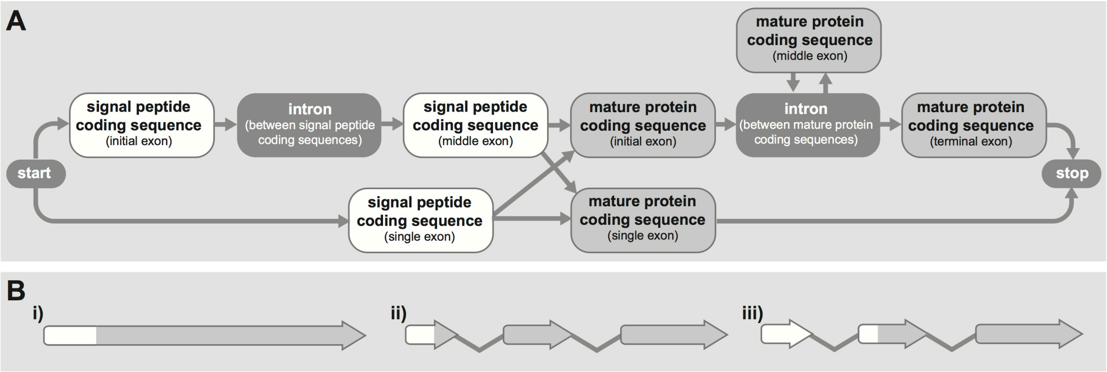

A
start
signal peptide coding sequence (initial exon)
intron (between signal peptide coding sequences)
signal peptide coding sequence (middle exon)
mature protein coding sequence (initial exon)
mature protein coding sequence (middle exon)
intron (between mature protein coding sequences)
mature protein coding sequence (terminal exon)
stop
signal peptide coding sequence (single exon)
mature protein coding sequence (single exon)
B
i)
ii)
iii)

Figure 3 A: A flow chart summarising the states used by CodingQuarry to predict effector like genes is shown in A. B: the possible positioning of a signal peptide predicted by CodingQuarry can be within a single exon gene (i), the first exon of a multiple exon gene (ii), or extending to within the second exon of a multiple exon gene (iii).

To further suit this model to effector-like genes that are missed by other prediction methods, the content models of the coding sequence corresponding to the mature peptide was trained to be typical of a high-cysteine sequence with atypical codon usage. The challenge in doing this is that the set of genes that are predicted to be secreted, high in cysteine, and have atypical codon use is too small to use for CodingQuarry's usual method of training, which is based on the frequency of nucleotide hexamers. In circumventing this problem, nucleotide hexamer frequencies were estimated from the codon usage of secreted genes with atypical codon usage and the amino acid frequencies of high-cysteine mature peptide sequences. High-cysteine secreted genes were considered to be those with a percentage cysteine in the upper quartile for the predicted secretome (that is, the top 25%). Secreted genes with atypical codon usage were considered to be those with a codon adaptation index within the lower quartile of the secretome (that is, the lowest 25%). The codon adaptation index (CAI) was calculated with respect to the codon usage frequencies of the set of genes that would be used for gene model training in a normal CodingQuarry run, that is, complete, non-overlapping genes over 600 nucleotides in length. While the CAI is usually calculated in reference to a set of highly expressed genes, here the set of genes typically used for training were used as a reference. This was deliberate, with the intension of identifying genes with a codon usage atypical when compared to the training set and thus likely to be poorly described by normal parameters. The codon usage frequencies as derived from the low CAI secreted gene set and the amino acid frequencies from predicted signal peptides and high-cysteine mature peptide sequences were used to estimate nucleotide hexamer frequencies for high-cysteine, low CAI genes. For example, to estimate the frequency of CTTATT starting in the 0^{th} reading frame encoding a leucine followed by an isoleucine, the frequency of isoleucine and leucine from the training set and the codon usage frequency of CTT and ATT are used

$$f(\mathrm{CTTATT}) \approx f(\mathrm{CTT|L})f(\mathrm{ATT|I})f(\mathrm{LI})$$

where *f*(CTTATT) is the estimated nucleotide hexamer frequency, *f*(CTT|L) is frequency of CTT relative to the other codons that encode leucine, *f*(ATT|I) is the usage frequency of ATT relative to the other codons that encode isoleucine, and *f*(LI) is the frequency of a leucine followed by and isoleucine. For parameters for the signal peptide sequence, due to the training set consisting of a small overall amount of sequence *f*(IL) was estimated as the product of the frequencies of *f*(L) and *f*(I), which are calculated from the translation of the signal peptide sequences. For parameters used to model a high-cysteine mature peptide, *f*(LI) was be calculated from the mature peptide part of high-cysteine predicted to be secreted genes. In both cases, *f*(CTT|L) and *f*(ATT|I) are calculated from the set of low CAI genes. As such the estimate of *f*(CTTATT) combines the amino acid use from one set with the codon usage from another. This estimation method is used for every possible hexamer, with those containing in-frame stop codons (TAA, TAG, or TGA) set to frequency 0. The Markov chain parameters used for CodingQuarry's HMM gene prediction are then derived directly from these hexamer frequencies.

CodingQuarry-PM gene predictions were restricted to a minimum mature peptide length of 60 bp and the maximum exon length was set at 1,500 bp. For exon states adjoining the signal peptide, lengths were trained from the exon lengths of secreted genes. For exons not adjoining the signal peptide, values were trained as per CodingQuarry's normal exon length training. The signal peptide length model was trained from the lengths of predicted signal peptides, smoothed using a 6-nucleotide window, and restricted to a minimum length of 30 bp and a maximum length of 90 bp. The length model of a signal peptide split over multiple exons was set to have minimum length of 1 for each exon.

CodingQuarry's standard prediction method uses predictions made directly from transcript sequences to improve prediction accuracy. To allow CodingQuarry-PM to also benefit from RNA-seq information, where RNA-seq data is present predictions are initially made from transcripts that do not already contain predicted coding sequence. As described previously (Chapter 3 (198)), single exon predictions and coding sequence predictions that could be incomplete are discarded, and these loci are re-predicted along with additional predictions from genome sequence. When running CodingQuarry-PM, repeats were hard masked to prevent an excess of false positive predictions in repetitive regions.

4.4.4 Assessing the success of CodingQuarry's pathogen prediction mode

CodingQuarry's normal RNA-seq driven prediction was run on *L. maculans*, *P. fulva*, *M. grisea*, and *F. oxysporum* as described in the benchmarking section of the methods. Protein sequences were extracted from the predicted gene set from a normal CodingQuarry run (excluding the dubious set). SignalP (v. 4.1) (208) was used to predict the secretome from each of these sets. The predicted gene set was then provided as input to CodingQuarry-PM with the predicted location of signal peptides, thus using the "pathogen" prediction mode described in the previous section. Using this method, CodingQuarry-PM attempts to predict genes from sequences designated as intergenic in the results of the previous prediction run. This run mode of using CodingQuarry followed by CodingQuarry-PM has been automated in a script supplied in the latest CodingQuarry release.

In an extended assessment of CodingQuarry-PM, publically available annotations were downloaded for nine pathogen species (Table 5) and gene entries describing the location of known effectors were removed from the general feature format (GFF) files. Translated protein sequences were extracted using the GFF and genome sequence. SignalP (v. 4.1) was run, and the predicted location of signal peptides was recorded. Annotations (minus known effectors) and predicted signal peptide locations were input to CodingQuarry-PM, and predictions made in regions annotated as intergenic.

In assessing the success of CodingQuarry's pathogen mode of prediction the success of CodingQuarry at predicting known effectors was recorded and the set of additional predictions was analysed. The protein sequences of predictions made by CodingQuarry's pathogen mode were extracted and searched for homology against NCBI's nr (199) using blastp (BLAST+ v. 2.2.31) (209) and those with hits with an e-value less than 10^{-4} were recorded. While this is a low stringency e-value cut-off, effectors are known to rapidly diversify meaning homology to genes in other species if often weak. This e-value cut-off has been previously employed in searching for effector homologues (95). Hmmscan (HMMER 3.1b2 (210)) was used to search CodingQuarry-PM predicted proteins for homology to protein family (Pfam) domains using the Pfam library (v. 29.0 (96), automatic cut-off determined using the parameter –cut-ga). Where RNA-seq data was used (that is, for *P. fulva*, *M. grisea*, *F. oxysporum*, and *L. maculans*), the proportion of predicted coding sequence with RNA-seq read coverage was recorded using BEDTools (v. 2.17) (211) coverageBED. Genes with >80% coding sequence coverage by RNA-seq reads were considered to be RNA-seq supported.

To further assess the success of CodingQuarry-PM predictions, protein sets where searched for homology to AntiFam (212) domains and Repbase (213) entries that may be indicative of false positive predictions. As with the Pfam domain searches, hmmscan (HMMER 3.1b2 (210)) was used to search CodingQuarry-PM predicted proteins for homology to AntiFam domains (part of Pfam v. 29.0 (96) release, automatic cut-off determined using the parameter –cut-ga). Tblastn (BLAST+ v. 2.2.31)

(209) was used to search for homology between CodingQuarry-PM predictions and the fungal portion of Repbase (213) (v. 21.01) using an e-value cut-off of 10^{-4}.

4.5 Results

4.5.1 Benchmarking existing gene prediction against known effectors

The success of each of the gene prediction software at predicting known *M. oryzae*, *P. fulva*, *L. maculans*, and *F. oxysporum* effectors is displayed in Table 3. Across all gene prediction methods (excluding CodingQuarry-PM), the most common type of error in predicting effectors was that they were completely missed.

Ab initio methods tested include AUGUSTUS (v. 3.2, run without hints), GeneMark-ES, and GeneMark-ET (v. 4.30). GeneMark-ES self-trains without any data additional to the genome sequence, whereas GeneMark-ET trains with the aid of intron locations derived from RNA-seq read to genome alignment, but the gene predictions are essentially *ab initio*. AUGUSTUS (v. 3.2) was run without intron/exon hints (thus *ab initio*) but using the parameters trained with the BRAKER1 pipeline, which leverages RNA-seq for training using GeneMark-ET. AUGUSTUS (v. 3.2) run with RNA-seq hints (resulting from the BRAKER1 pipeline) are discussed later. Strikingly, 13 out of 32 (40%) of the tested effector genes are completely missed by each of the GeneMark tools, and 17 out of 32 (53%) are missed by AUGUSTUS (v. 3.2). A completely missed gene (that is, a gene mistaken for intergenic sequence) was the most common mode of prediction error. Incorrect intron / exon boundaries were only seen in five of the GeneMark predictions and four of the AUGUSTUS predictions. Although GeneMark-ET used RNA-seq data to assist in training and has been demonstrated to have a better overall accuracy than GeneMark-ES (207), no differences in effector gene prediction success were noted in this study between these two methods.

Table 3 Prediction of known fungal effector genes by gene prediction software

Species	Isolate	Genome reference	Effector	Size (aa)	Introns	Signal peptide length (aa)	Cysteines (mature)	% cysteines (mature)	CAI*	Cuff.	ORF	GM-ES	GM-ET	AUG.	BRA.	CQ	CQ+D	CQ-PM
Fusarium oxysporum f. sp. *lycopersici*	4287	(60)	*Avr3(Six1)*	284	1	21	8	3.0	0.77	M	M	C	C	C	C	P	P	P
			Six2	232	1	20	8	3.8	0.78	M	M	C	C	C	C	C	C	C
			SIX6	215	1	16	8	4.0	0.79	M	M	P	P	P	C	P	P	P
			Six5	119	4	17	7	6.9	0.80	M	M	C	C	M	C	C	C	C
			Avr2/SIX3	163	1	19	3	2.1	0.81	C	C	P	P	P	P	P	P	P
Leptosphaeria maculans	v23.1.3	(17)	*AvrLm4-7*	143	2	21	8	6.6	0.70	P	P	M	M	M	M	M	M	C
			AvrLm6	144	4	20	6	4.8	0.71	M	M	M	M	M	M	M	M	C
			AvrLmJ1	141	1	19	7	5.7	0.73	C	C	M	M	M	M	M	M	M
			AvrLm1	205	2	22	1	0.5	0.74	P	P	P	P	M	M	P	P	P
			AvrLm2	232	1	19	8	3.8	0.76	C	C	C	C	M	M	C	C	C
Magnaporthe oryzae	70-15	(85)	*Avr-Pita*	224	4	17	9	4.3	0.54	M	M	M	M	M	M	M	M	M
			Avr-Pik	113	1	21	3	3.3	0.60	M	M	M	M	M	M	M	M	M
			Pwl2	145	1	21	2	1.6	0.62	M	M	M	M	M	M	M	M	C
			Pwl3	138	1	21	0	0	0.62	M	M	C	C	C	C	C	C	C
			Bas1	115	1	22	0	0	0.63	M	M	M	M	M	M	M	M	C
			Bas107	132	1	19	2	1.8	0.65	M	M	C	C	C	C	C	C	C
			Avr-Piz-T	108	1	18	4	4.5	0.67	M	M	M	M	M	M	M	M	C
			Bas4	102	1	21	8	9.9	0.73	M	M	C	C	C	C	C	C	C
			Bas2	102	2	19	6	7.2	0.74	C	M	C	C	C	C	C	C	C
			MC69	54	2	17	3	8.1	0.80	C	M	C	C	M	M	C	C	C
			Bas3	113	3	20	10	10.7	0.82	P	P	C	C	C	C	P	P	P
			Avr-Pi9	91	2	18	6	8.2	0.82	C	M	C	C	P	C	C	C	C
			Slp1	162	2	16	6	4.1	0.93	M	M	C	C	C	C	C	C	C
Passalora fulva	CBS 131901	(22)	*Avr2*	78	2	20	8	13.8	0.66	C	P	M	M	M	C	M	C	C
			Arv5	103	5	22	10	12.3	0.69	C	C	M	M	M	C	M	C	C
			Avr9	63	2	23	6	15.0	0.69	C	C	M	M	M	P	M	C	C
			Ecp5	115	3	19	6	6.2	0.71	C	C	M	M	M	C	M	C	C
			Avr4	135	1	18	8	6.8	0.75	C	C	C	C	C	C	C	C	C
			Avr4E	120	1	22	6	6.1	0.78	P	M	M	M	M	M	M	M	C
			Ecp2	165	2	18	4	2.7	0.78	C	C	C	C	C	C	C	C	C
			Ecp4	119	2	19	6	6.0	0.84	C	C	P	P	P	C	C	C	C
			Ecp6	228	3	18	9	4.3	0.85	C	C	C	C	C	C	C	C	C
			Ecp1	96	3	19	2	2.6	0.89	C	C	P	P	M	M	C	C	C

Tested prediction methods / software from left are: Cufflinks (v. 2.1.1) assembled transcripts (Cuff.), longest ORF from within an assembled transcript, GenMark-ES (GM-ES), GeneMark-ET v. 4.30 (GM-ET), AUGUSTUS v. 3.2 *ab initio* (AUG.), BRAKER1 v. 1.6 (BRA.), CodingQuarry v. 2.0 (CQ), CodingQuarry v. 2.0 including secreted dubious predictions (CQ+D), CodingQuarry's v. 2.0 newly described pathogen run method (CQ-PM) (see methods section for further details). Entries in red, labelled "M" correspond to where no part of the coding sequence of the effector gene is predicted, green entries labelled "P" where a part of the coding sequence is predicted, and blue entries labelled "C" where the entire coding sequence is correctly predicted (see Methods for further details).

One coding sequence prediction method reliant on transcript assembly is where an open reading frame is taken from the transcript sequence. The longest open-reading frame from each Cufflinks transcript was annotated to assess whether this is a successful strategy for annotating effectors with possible utility where effectors are not predicted by other gene prediction methods. This is clearly highly dependent on the transcript being correctly reconstructed. Of the 16 assembled transcript sequences, the longest ORF within the transcript corresponded to the correct effector coding sequence in 12 cases (75% success rate). In 3 cases (*Bas2*, *MC69*, and *Avr-Pi9* of *M. oryzae*) the longest ORF did not correspond to the effector coding sequence, and the effector was therefore missed by this approach. In the cases of *Bas2* and *AvrPi9*, this is due to the assembled transcripts being merged to with the adjacent genes due to overlapping 3' and 5' UTRs — a common problem in fungal transcript reconstruction and gene prediction (198,214,215). The longest ORF therefore corresponds to one of the adjacent coding sequences in each case. In the case of *MC69*, the assembled transcript contains a larger ORF with no blast hits (to NCBI's nr database), which is likely to be a spurious open reading frame rather than a coding sequence.

AUGUSTUS (v. 3.2), run with RNA-seq hints through the BRAKER1 pipeline, shows a higher success rate at predicting effectors than where it was run without RNA-seq hints. Of the 17 cases where effectors are missed by the *ab initio* predictions made by AUGUSTUS, 3 are correctly predicted through the use of hints. A further 2 effectors which were partially predicted by the *ab initio* run of AUGUSTUS are correctly predicted when using hints. Prediction improvement at a particular gene locus by using AUGUSTUS hints relies on RNA-seq data at that specific locus, and as such many of the effectors that are not well supported by RNA-seq do not benefit from this method. However, from the eight effectors that are missed by AUGUSTUS *ab initio* and have strong RNA-seq support, 3 are correctly predicted and one partially predicted when using hints. As such, where an effector gene was supported by RNA-seq data, using the hints run mode of AUGUSTUS substantially improve the prediction success rate.

The existing (published) run mode of CodingQuarry uses RNA-seq data to train and incorporates predictions made directly from transcript sequences to improve the accuracy of intron/exon boundaries and prevent false joining of adjacent coding sequences. The standard CodingQuarry (v. 2.0) run missed 13 effectors, which equates to 40% of the tested effectors. With the incorporation of RNA-seq supported genes from CodingQuarry's dubious set, the number of missed effectors dropped by 4. Of 4 effectors missed by CodingQuarry that have a correctly reconstructed RNA-seq transcript, all 4 are captured in CodingQuarry's dubious set. The full set of these dubious predictions is large and likely to contain many false positive predictions, although the subset of these predictions that are also predicted to be secreted is typically small and is reasonable to include in downstream analysis. For example, the "dubious" set of predictions on *P. fulva* contains 1,322 predictions, however just 37 of these are predicted to be secreted 4 of which are known effectors. As with

AUGUSTUS hints predictions, these predictions made by CodingQuarry are heavily reliant on RNA-seq support at effector loci. In the case of CodingQuarry's "dubious" predictions, the gene must have a complete assembled transcript to be predicted.

Consistent with expectations that atypical codon usage may affect the success of gene prediction on effector genes, several of the lower CAI effector (CAI assessed with respect to coding sequence annotations ≥ 600 nucleotides long) such as *M. oryzae* effectors *Avr-Pita*, *Avr-Pik*, and *Pwl2* were consistently missed by gene predictors. There were also some examples of very high-cysteine effectors being problematic for gene predictors, such as *P. fulva* effectors Avr5, Avr9, and Avr2 that are each over 10% cysteine. In some effectors, the signal peptide accounts for a high proportion of the overall coding sequence, meaning it has a strong impact on the overall amino acid composition of the coding sequence. This is most extreme in *P. fulva's* Avr9, where the predicted signal peptide accounts for 36% of the full protein length.

We note a strong reliance on RNA-seq in predicting effector genes. Out of the sixteen tested effector genes that have a correctly assembled Cufflinks transcript, which can be viewed as evidence of strong RNA-seq support, all sixteen are correctly predicted by at least one of the discussed methods. This is very positive when considering that annotations are often arrived as through a combination of different prediction strategies. Conversely, of the 16 effectors with partial or no RNA-seq support, only 7 — less than half — are correctly predicted by at least one prediction method. This result excludes CodingQuarry's newly presented "pathogen" mode, which is discussed in the next section.

4.5.2 CodingQuarry's pathogen mode – CodingQuarry-PM

CodingQuarry's newly presented "pathogen" prediction mode was run as a subsequent step to a normal CodingQuarry run to attempt to predicted missed effector-like genes. The success of CodingQuarry-PM (v. 2.0) at predicting known effectors when run subsequent to CodingQuarry and using RNA-seq data has been recorded in the final column of Table 3. As this method does not aim to correct existing predictions, the 5 partially correct effector gene predictions made by the standard CodingQuarry run mode remain unchanged. However, using this method greatly increased the success rate of effector prediction by predicting genes that are otherwise missed. Of the 32 effectors tested, 24 were correctly predicted and just 3 were missed by CodingQuarry-PM. This number is low when compared to the 9 effectors missed by CodingQuarry with "dubious" set predictions and the 13 effectors missed by BRAKER1 (v. 1.6) and the standard CodingQuarry.

A summary of the additional predictions made by CodingQuarry-PM following a standard RNA-seq driven CodingQuarry run is shown in Table 4. The number of predictions unique to CodingQuarry-PM, that is, not overlapping predictions made by any of the other gene prediction tools, ranged from 155 (*P. fulva*) up to 249 (*M. oryzae*). Where CodingQuarry-PM predictions agreed with predictions made by another method, this agreement lends support to the gene model being correct. On the other hand, gene models unique to CodingQuarry-PM are valuable in that these may be genuine genes that are not predicted by other methods. RNA-seq and homology support of such unique genes lends support to them being genuine coding sequence predictions rather than false positives. CQ-PM *L. maculans* predictions had the highest rate of RNA-seq support, with 84% supported by RNA-seq data. *P. fulva* predictions had the next highest level of RNA-seq support with 70% of CQ-PM predictions RNA-seq supported. Lower levels of RNA-seq support were seen among the *M. oryzae* and *F. oxysporum* CQ-PM predictions, at 47% and 35% respectively. For each species, the majority of uniquely CodingQuarry-PM predictions were supported by RNA-seq or homology evidence (Figure 4 and Table 4). The number of predicted genes unique to CodingQuarry-PM and with no homology or RNA-seq support was low, at 34, 9, 24, and 8 for *P. fulva*, *L. maculans*, *M. oryzae*, and *F. oxysporum* respectively. None of the CodingQuarry-PM predictions had homology to AntiFam domains, supporting the predicted genes are genuine coding sequences. The number of predictions showing homology to Repbase was also low, with the highest number (of 11) seen in *M. oryzae*, one in *L. maculans* and *F. oxysporum,* and none seen in *P. fulva*. Weak homology to a Repbase entry does not necessarily indicate that a predicted gene is a false positive prediction, however the low numbers seen in this category are consistent with CodingQuarry-PM predictions being genuine coding sequences.

CodingQuarry-PM (v. 2.0) was then used to predict genes from the intergenic regions of 9 phytopathogen species with publically available genome sequences and annotations. Known effectors were removed from the annotations, and the ability of CodingQuarry-PM to *ab initio* predict effectors was compared to the standard CodingQuarry *ab initio* run mode. Of the 44 known effectors tested, 16 are missed by CodingQuarry's standard *ab initio* mode, compared to just 4 missed by CodingQuarry-PM. Two effectors that are partially predicted by CodingQuarry are correctly predicted by CodingQuarry-PM.

CodingQuarry-PM makes a great improvement to the *ab initio* predict effectors (Table 5), yet we see that some errors still prevail. Some of these are due to inherent inaccuracies in HMM gene prediction. For example, the exon boundaries of *SnToxA* are incorrectly predicted by both methods due to it containing an intron with a non-canonical splice site. The *L. maculans* effector *AvrLm1* has an intron after the first nucleotide of the starting methionine (that is, splitting the ATG) making it an impossible *ab initio* prediction for CodingQuarry (and many other gene predictors) as the ATG start is modelled as a single state. The prediction from transcript sequence strategy employed by CodingQuarry (198) and also used by CodingQuarry-PM may reduce such inaccuracies at RNA-seq supported loci. Some effectors are still missed despite the targeted efforts to correctly predict them, although the number of these is greatly reduced.

Table 4 A summary of additional genes predicted by CodingQuarry-PM, run subsequent to the standard CodingQuarry RNA-seq driven prediction mode.

Species	Isolate	No. of additional predictions	No. predicted to be secreted (SignalP v. 4.1)	Average length (aa)	Average number of cysteines per gene	No. of unique predictions	No. with blast hit to NCBI's nr (evalue 10^{-4})	No. with RNA-seq support	No. with Pfam domain	No. with AntiFam domain	No. with homology to Repbase entry (evalue 10^{-4})
P. fulva	CBS 131901	267	176	124	6.7	155	111	191	21	0	0
L. maculans	v23.1.3	283	161	110	4.6	172	177	238	35	0	1
M. oryzae	70-15	290	194	103	4.3	246	204	137	13	0	11
F. oxysporum	4287	306	149	153	6.4	103	268	108	37	0	1

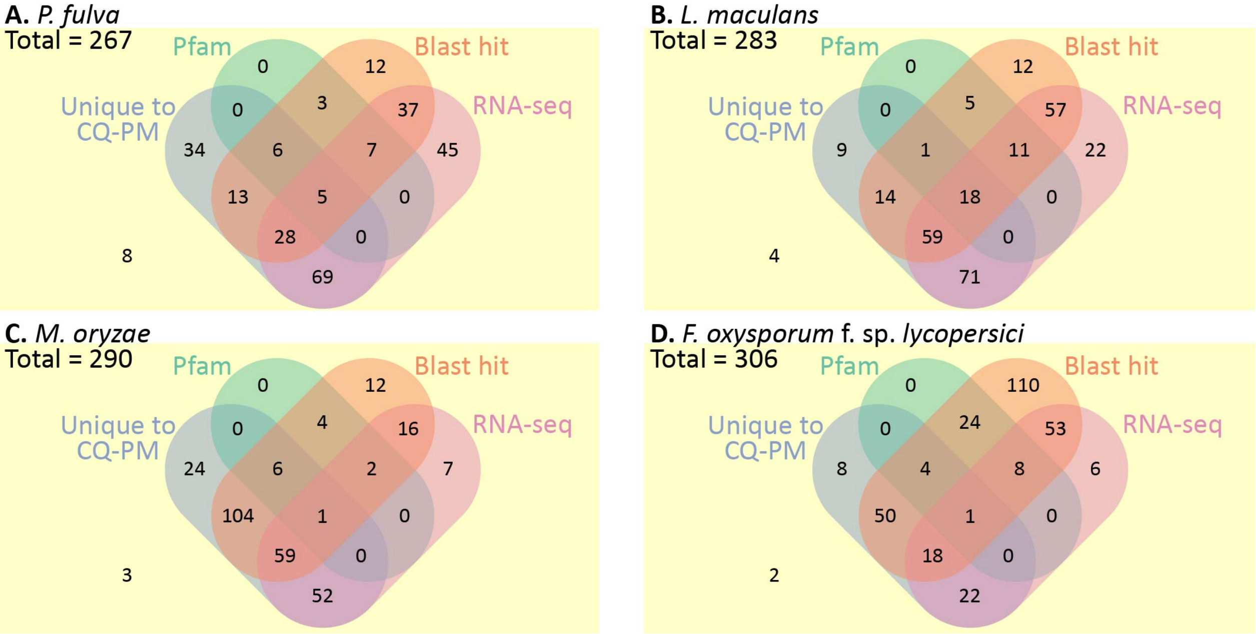


Figure 4 Venn diagrams summarising the gene predictions made by CodingQuarry's pathogen mode (CodingQuarry-PM, part of v. 2.0 of CodingQuarry).

Table 5 Ab initio prediction of effectors by CodingQuarry and CodingQuarry-PM (v. 2.0)

Species	Isolate	Genome reference	Effector	CQ (*ab initio*)	CQ-PM (*ab initio*)
Ustilago maydis	521	(216)	*Tin2*	P	C
			Cmu1	C	C
			Pit2	P	C
			Pep1	C	C
			See1	C	C
Parastagonospora nodorum	SN15	(8)	*SnTox1*	M	C
			SnTox3	C	C
			SnToxA	P	P
Verticillium dahliae	VdLs.17	(116)	*Vdlsc1*	C	C
Ustilago hordei	Uh4857-4	(217)	*UhAvr1*	M	C
Magnaporthe oryzae	70-15	(85)	*Bas1*	M	C
			Bas2	C	C
			Bas3	C	C
			Bas4	C	C
			Bas107	C	C
			MC69	C	C
			AvrPi9	C	C
			AvrPiz-t	M	C
			AVR-Pik	M	C
			AVR-Pita	M	M
			Pwl2	M	C
			Pwl3	C	C
			slp1	M	M
Passalora fulva	CBS 131901	(22)	*Avr2*	M	C
			Arv5	M	C
			Avr9	M	M
			Ecp5	M	P
			Avr4	C	C
			Avr4E	M	C
			Ecp2	C	C
			Ecp4	P	P
			Ecp6	M	M
			Ecp1	P	C
Leptosphaeria maculans	v23.1.3	(17)	*AvrLm4-7*	M	C
			AvrLm6	M	C
			AvrLmJ1	P	P
			AvrLm1	P	P
			AvrLm2	C	C
Fusarium oxysporum f. sp. *lycopersici*	4287	(218)	*Avr2/SIX3*	C	C
			Avr3(Six1)	C	C
			Six2	C	C
			SIX6	C	C
			Six5	C	C
Pyrenophora tritici-repentis	BFP-ToxAC	(12)	*ToxA*	P	P

The full set of genes predicted by CodingQuarry-PM for each species is shown in Table 6. The number of CodingQuarry-PM predicted genes varied greatly from species to species – from 61 additional genes (*U. hordei*) up to 726 genes (*F. oxysporum* f. sp. *lycopersici*). This differs to the results from the use of CodingQuarry-PM following the RNA-seq driven run of CodingQuarry, where the number of genes predicted is less variable across species. This is likely due to the annotations used for training being derived from different methods. This influences the quality of gene model training that CodingQuarry can achieve. Furthermore, the stringency of the original annotation affects the number of genes that remain unannotated and have the potential to be predicted by CodingQuarry-PM.

Not all CodingQuarry-PM predictions are predicted to be secreted by subsequent analysis. This is as to be expected, as the model used does not require predictions to be secreted, but rather favours their prediction. In seven of the ten species, more than a third of the CodingQuarry-PM predicted proteins were also predicted to be secreted (by SignalP v. 4.1). The highest proportion of CodingQuarry-PM predicted genes predicted to be secreted was seen in *P. fulva* at 66%. *P. fulva* predictions also had the highest average cysteine content, at 6%.

As in the RNA-seq driven CodingQuarry-PM results, none of the CodingQuarry-PM predictions showed homology to AntiFam, and numbers of predicted genes with homology to Repbase entries was generally low. Excepting *P. tritici-repentis*, in which 21.9% of CodingQuarry-PM predictions had homology to Repbase, the proportion of CodingQuarry-PM predictions with homology to a Repbase entry was less than 10%. These results support that the majority of CodingQuarry-PM predictions are genuine coding sequences. *P. fulva*, *M. oryzae*, *L. maculans*, and *F. oxysporum* predictions with homology to Repbase while infrequent, where more common than when CodingQuarry-PM was run as an RNA-seq driven predictor following CodingQuarry. This may be due to the RNA-seq driven version of CodingQuarry being able to establish a higher quality training set than is achieved by training from publically available annotations. *P. nodorum* has a very high quality annotation — the result of manual curation incorporating RNA-seq and proteomics evidence (10) — and less than 0.5% of CodingQuarry-PM predictions that trained on this annotation show homology to a Repbase entry.

In all cases excepting *L. maculans*, over 50% of the CodingQuarry-PM predictions had homology to a sequence record in GenBank's nr protein database (e-value < 10^{-4}) (Table 6). This is a good indicator that the CodingQuarry-PM prediction set contains genuine genes, rather than being false positive predictions of non-coding open-reading frames. Nineteen per cent of *L. maculans* CodingQuarry-PM predicted proteins were supported by a blast hit — the lowest of all the species — but *L. maculans* had the highest proportion of Pfam domain supported proteins at 44%. The presence of putative proteins with homology to Pfam domains also supports that these may be genuine proteins. Of particular interest to effector prediction, 7 chitin recognition Pfam domain containing proteins were identified in *P. nodorum*, all of which were predicted to be secreted. A chitin recognition protein domain was also identified in a predicted protein from each of *L. maculans*, *V. dahliae*, *P. fulva*, and *P. tritici-repentis.* Chitin binding has been associated with some effectors that function primarily to inhibit plant recognition of PAMPs (33,219). Two LysM domain proteins were identified amongst *F. oxysporum* f. sp. *lycopersici* predictions, as well as one LysM domain protein within *P. tritici-repentis.* LysM domains have been found within some known effectors (33,220).

Table 6 CodingQuarry-PM (v. 2.0) predictions made in intergenic regions of publically available, annotated phytopathogen genomes

Species	Isolate	No. of additional predictions	No. predicted to be secreted	Average length (aa)	Average number of cysteines per gene	No. with blast hit to NCBI's nr (evalue 10^{-4})	No. with Pfam domain	No. with AntiFam domain	No. with homology to Repbase entry (evalue 10^{-4})
U. maydis	521	234	41	359	4.2	187	93	0	1
P. nodorum	SN15	288	130	221	6.0	250	69	0	1
V. dahliae	VdLs.17	156	58	175	4.5	118	28	0	12
U. hordei	Uh4857-4	61	8	179	3.4	36	12	0	3
M. oryzae	70-15	304	133	229	4.6	243	82	0	22
P. fulva	CBS 131901	198	128	151	7.6	105	33	0	13
L. maculans	v23.1.3	209	72	331	5.8	39	93	0	0
F. oxysporum f. sp. *lycopersici*	4287	726	248	236	5.3	670	205	0	19
P. tritici-repentis	BFP-ToxAC	333	126	319	6.5	251	107	0	73

4.6 Discussion

In testing existing gene prediction software at predicting known effectors, a high proportion of known effectors were shown to be missed by gene prediction software. This accuracy is well below the genome wide accuracy of gene prediction (132,198). The failure of prediction methods at predicting most effectors indicates that parameters that are successfully able to predict the majority of genes may match effectors poorly. Effectors were predicted better where loci were RNA-seq supported, with BRAKER1 and CodingQuarry (including "dubious" set predictions) performing better at these loci.

We demonstrate a novel method for predicting effector-like genes (CodingQuarry-PM), whereby most effectors are successfully predicted *ab initio* or using RNA-seq through the use of parameters suited to secreted, cysteine-rich, low CAI sequences. This mode can be used to mine "intergenic" genome regions for putative genes either after standard CodingQuarry prediction or to supplement an existing annotation from another source. The utility of this novel prediction method is reliant on it making useful predictions beyond known effectors. For example, predicting thousands of false positive genes may result in the correct prediction of effectors, but these predictions are less useful if there are is a surplus of false positive predictions. The predictions made by CodingQuarry-PM showed a significant proportion had homology to on or more proteins in NCBI's nr.

In this study, the issue of missed genes was addressed, rather than attempting to correct gene models that are incorrectly predicted. As such, errors in either the externally derived annotations or the primary run of CodingQuarry will not be resolved by using CodingQuarry-PM. In some cases, CodingQuarry-PM makes a correct prediction where CodingQuarry predicts a gene partially (*U. maydis* effectors *Tin2* and *Pit2*). A method for selecting the correct model could address in future research. Even so when predicting genes, a background level of *ab initio* errors are inevitable using even the most accurate methods.

RNA sequencing has become commonplace to assist in gene prediction, the importance of which is emphasized by results showing improved effector prediction at RNA-seq supported loci. Ever lowering sequencing costs mean there are fewer barriers to the acquisition of this data, although in pathogen population studies, RNA-seq is not always gathered for non-reference isolates, and *ab initio* prediction is relied upon for non-reference isolates. Based on the results presented in this study, CodingQuarry-PM greatly improves the chances of predicting novel effectors in these isolates. Given the low cost of RNA-seq data, we recommend that when genome sequencing non-reference pathogen isolates RNA-seq is likely to be highly beneficial in locating novel effectors. Even so, it is possible that important pathogenicity genes such as effectors may not be expressed under the conditions used for RNA sequencing. CQ-PM predictions showed the highest rate of RNA-seq support in *L. maculans* (84% of predictions). *L. maculans* was also the only pathogen tested that had RNA-seq libraries specified as from *in planta* experiments. This result supports the importance of *in planta* RNA-seq in capturing the expression of pathogenicity related genes. Sequencing cost can be a limiting factor in deciding the number of infection time-points to sample, and even *in planta* RNA-seq experiments may miss the expression of some effectors. Furthermore, at early stages of infection it can be difficult to obtain sufficient sequencing coverage of fungal mRNAs, which are drowned out by high levels of plant-derived mRNAs. These experimental and practical issues highlight the importance of being able to make *ab initio* predictions.

Another useful type of data in gene discovery is proteomic data, which is aligned to a reference library of protein sequences thus providing evidence of translation. This relies on the reference library containing the protein sequence evidenced in the mass spectrometry data. For the discovery of novel genes, a 6-frame translation of the genome can be used. A limitation of this is that it increases the likelihood of spurious hits and that peptide alignments across intron boundaries will be missed. Providing a 6-frame transcriptome translation can be an improvement on this, but this relies on the locus of interest being correctly reconstructed. CodingQuarry-PM can now be used to supplement protein libraries to be used for proteomics. This is likely to be useful were secreted or effector active protein fractions are sent for proteomic analysis. Furthermore, the use of proteomic data may assist in the experimental validation of CodingQuarry-PM predictions.

The prediction of effectors is a common research goal in the study of pathogens and of sufficient importance to industry to warrant special attention. Numerous studies have looked at addressing the problem of selecting effector candidates from larger lists of proteins (9,87–90,221). More recently, other bioinformatic processes that affect effector prediction accuracy, such as secretion prediction and assembly, have been recognised as important to overall effector prediction accuracy (113,222). This study is — to our knowledge — the first report of a gene prediction software tool targeted to the prediction of effector-like gene loci from genome sequence. CodingQuarry-PM is therefore useful in the study of phytopathogens, where effector-like small, secreted, cysteine-rich proteins are of paramount importance to downstream analysis.

Chapter 5 Overview of genomic and bioinformatic resources for Zymoseptoria tritici

5.1 Attribution statement

Title:	Overview of genomic and bioinformatic resources for Zymoseptoria tritici
Authors:	Alison C Testa, Richard P Oliver, and James K Hane
Citation:	Fungal Genetics and Biology 79 (2015) 13–16

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 Ph. D. candidate (ACT) made the following contributions to this chapter:

- ACT wrote the software and performed all bioinformatic analysis
- ACT and JKH co-wrote the manuscript

The following contributions were made by co-authors:

- JKH co-wrote the manuscript
- RPO reviewed the manuscript

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

________________________________ ________________

Alison C Testa Date

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

________________________________ ________________

Richard P Oliver Date

(Principal supervisor and co-author)

________________________________ ________________

James K Hane (co-author) Date

Fungal Genetics and Biology 79 (2015) 13–16

Contents lists available at ScienceDirect

Fungal Genetics and Biology

journal homepage: www.elsevier.com/locate/yfgbi

Overview of genomic and bioinformatic resources for *Zymoseptoria tritici*

Alison Testa[a], Richard Oliver[a], James Hane[a,b,*]

[a] Centre for Crop and Disease Management, Curtin University, Perth, WA, Australia
[b] Curtin Institute for Computation, Curtin University, Perth, WA, Australia

ARTICLE INFO

Article history:
Received 5 January 2015
Revised 10 April 2015
Accepted 13 April 2015

Keywords:
Zymoseptoria tritici
Genome
Transcriptome
Bioinformatics
Pathogenomics

ABSTRACT

Zymoseptoria tritici (syn. *Mycosphaerella graminicola*, *Septoria tritici*) is a haploid fungus belonging to the class Dothideomycetes. It is the causal agent of septoria leaf blotch – one of the world's most significant diseases of wheat. Here we review the genomic and bioinformatic resources that have been generated for *Z. tritici*. These include the whole-genome reference assembly for isolate IPO323, genome resequencing of alternate isolates, mitochondrial genome sequences, transcriptome sequences and expression data, and annotations of gene structure and function. We also highlight important advances in our fundamental knowledge of genome evolution and its effects on adaptation and pathogenicity in *Z. tritici* that have been facilitated by these resources.

1. The nuclear genome

An early landmark paper studied the electrophoretic karyotype of 7 isolates of *Z. tritici*, which estimated 14–16 chromosomes ranging from 330 kb to 3.5 Mb (McDonald and Martinez, 1991). Chromosome length and number polymorphisms were also observed between isolates, with chromosome presence/absence variation (PAV) observed exclusively for the two shortest chromosome bands. Subsequent generation of a genetic map from a cross between parent isolates IPO323 and IPO94269 predicted a higher number of chromosomes at 23 linkage groups (Kema et al., 2002). Further analysis of this mapping population revealed that at least 8 of the smaller linkage groups – corresponding to 8 dispensable chromosomes ranging from 0.39 to 0.77 Mb – were not required for saprophytic growth i.e. were dispensable or accessory chromosomes (ACs) (Wittenberg et al., 2009). The distinction between core chromosomes (CCs) and ACs was an important milestone for *Z. tritici* genomics.

Z. tritici was the first species of the Dothideomycetes – a fungal class of high agricultural significance (Hane et al., 2011b; Ohm et al., 2012) – to have had the genome of a representative isolate (IPO323) (Kema and van Silfhout, 1997) sequenced beyond the typical 'draft genome' status, achieving near complete chromosome sequences spanning from telomere to telomere (Goodwin et al., 2011) (Table 1). The IPO323 reference genome assembly is 39.69 Mb in length, representing 21 chromosomes. At least half of the genome is contained within the six largest assembled sequences (that is, an N50 of 6), with the sixth largest sequence having a length of 2.67 Mb. Its chromosomes have been designated into CCs and ACs, numbered 1–13 and 14–21 respectively.

In addition to chromosomal polymorphisms across the ACs, the *Z. tritici* genome has also been observed to exhibit at least two other forms of genomic plasticity. The repetitive content of the IPO323 genome reference is 12.26% (Ohm et al., 2012) and its repeats exhibit the hallmarks of repeat-induced point mutation (RIP) (Goodwin et al., 2011) – a fungal specific type of mutation that targets repetitive DNA and randomly converts cytosine to thymine bases (Hane and Oliver, 2008, 2010; Hane et al., 2015). In comparisons of whole-genome sequences across species belonging to the sub-phylum Pezizomycotina, the *Z. tritici* genome also exhibits a typical "mesosyntenic" conservation pattern, that is, the relative order and orientation of genes are reshuffled within homologous chromosomes due to frequent intra-chromosomal recombinations (Croll et al., 2013; Hane et al., 2011a). The combined effects of AC variability, mesosynteny and RIP may have significant implications for genome evolution and pathogenicity in *Z. tritici*.

Due to the completeness of its genome assembly, *Z. tritici* IPO323 has been a key point of reference in several comparative genomics studies, particularly across the Dothideomycetes (Hane et al., 2011b; Morais do Amaral et al., 2012; Ohm et al., 2012). Observations of mesosynteny made from genome sequence

* Corresponding author at: CCDM Bioinformatics, Centre for Crop and Disease Management, Department of Environment and Agriculture, Curtin University, Australia. Tel.: +61 8 9266 1726. Postal address: Department of Environment & Agriculture, Curtin University, GPO Box U1987, Perth, Western Australia 6845, Australia.

E-mail address: James.Hane@curtin.edu.au (J. Hane).

http://dx.doi.org/10.1016/j.fgb.2015.04.011
1087-1845/© 2015 The Authors. Published by Elsevier Inc.

alignments of different fungal species with *Z. tritici* can also be applied to fungal genome finishing. Whole-genome alignments of draft assemblies of other Pezizomycotina species with the *Z. tritici* IPO323 reference assembly have previously been used to predict the co-location of scaffolds on the same chromosome (i.e. synteny) (Goodwin et al., 2011). Significant insight has also been gleaned within the species *Z. tritici* itself through comparisons of alternate isolates, revealing differential selection pressures between CCs and ACs (Stukenbrock et al., 2010) and highlighting isolate-specific patterns of chromosome presence and absence (Goodwin et al., 2011; Croll et al., 2013) (see also McDonald et al., 1991).

2. The mitochondrial genome

The mitochondrial genomes (mtDNA) of two *Z. tritici* isolates, the reference isolate IPO323 and strain STBB1 (Table 1), have been sequenced in full (Torriani et al., 2008). The *Z. tritici* mtDNA [Nucleotide: EU090238] is 43,960 bp in length, indicating that it has been relatively untouched by invading intronic endonucleases (as compared with *Leptosphaeria maculans* and *Pyrenophora tritici-repentis*) (Manning et al., 2013; Rouxel et al., 2011). The mtDNA contains 15 protein-coding genes, large and small ribosomal subunits, 27 tRNAs and 8 unknown open-reading frames. Protein-coding genes residing on the mtDNA include *atp6*, *atp8*, *atp9*, *cox1-3*, *cytb*, *nad1-6*, *nad4L* and *RNA-Pol*. The presence and relative order of these genes is not highly conserved with closely related species of the Dothideomycetes (Aguileta et al., 2014; Hane et al., 2011b). However within the species *Z. tritici* across multiple isolates there was relatively little sequence variability (Zhan et al., 2003). Polymorphic sequences were identified between the mtDNA of the two sequenced isolates and used to generate markers for PCR screening across numerous isolates – an important genomic resource for studying the emergence and distribution of fungicide resistance. The mtDNA sequence data has also been a useful tool for tracking the history of the co-evolution of *Z. tritici* and its wheat host (Torriani et al., 2011; Zhan et al., 2004).

3. Gene function

Annotations of gene structure and function for *Z. tritici* IPO323 (Goodwin et al., 2011) are downloadable from JGI Mycocosm (Grigoriev et al., 2013) and EnsemblFungi (Kersey et al., 2014) (Table 1). As part of the initial genome survey study, potential carbohydrate-active enzymes (CAZymes) encoded by the gene content of *Z. tritici* were annotated, highlighting a distinctive reduction in cellulose and other cell-wall degrading enzymes relative to other plant pathogens (Goodwin et al., 2011). This observation supported a model of “stealth pathogenesis” for *Z. tritici*, in which CAZyme activity is reduced – with a compensatory shift toward protein degradation (Goodwin et al., 2011) – in conjunction with other adaptations for avoidance of triggering host-defenses (Lee et al., 2014). Morais do Amarai et al. subsequently used the gene annotations generated by Goodwin et al. to bioinformatically predict genes encoding secreted proteins and assign additional functional annotations. The predicted secretome (Morais do Amaral et al., 2012) comprises 492 proteins, with 321 possessing some level of functional annotation and 171 with no functional annotation.

4. Gene expression

An early transcriptome resource for *Z. tritici* was generated by Kema et al., comprising Expressed Sequence Tag (EST) libraries for IPO323 from 7 *in vitro* and 3 *in planta* growth conditions (Kema et al., 2008) (Table 1). Limitations around capture of fungal transcripts in mixed plant samples were overcome by purifying via hybridisation against IPO323 genomic DNA. A total of 27,007 ESTs were clustered into 9190 unigene loci, totaling 5.2 Mb. Though a comparable number of ESTs were typically sequenced in similar fungal genomics studies of the day, a very high representation of loci was achieved due to the high number of libraries from diverse growth conditions.

Recently, high-coverage whole-transcriptome expression data for the *Z. tritici*-wheat interaction has been facilitated by next-generation sequencing (RNA-Seq) (Table 1). Yang et al. have generated 18 Gbp of raw RNA-Seq data, comparing expression across the infection and necrotrophic growth stages of *Z. tritici* at 4, 10 and 13 days post-infection (dpi), as well as the corresponding host responses in *Triticum aestivum* [BioProject: 196595] (Yang et al., 2013). This study expanded upon the predicted secretome of Morais do Amarai et al., highlighting 313 genes of IPO323 that were highly expressed during early and late infection. A second RNA-Seq study by Kellner et al., generating 117.8 Gbp of RNA-Seq data, followed up with an in-depth comparison of *Z. tritici* expression during infection of two hosts: *T. aestivum* and *Brachypodium distachyon* [NCBI GEO: GSE54874] (Kellner et al., 2014). Differential expression patterns were observed across core and accessory chromosomes of *Z. tritici*. Kellner et al. observed no host-specific gene expression on accessory chromosomes during early infection and 25 wheat-specific genes during late infection (13 dpi). Notably, neither study had used the wealth of exon structure information afforded by RNA-Seq to re-annotate the initial set of gene predictions presented by Goodwin et al. (2011) (which relied heavily on *in silico* gene predictions supplemented with limited EST and homology supporting data). Consequently, despite considerable accumulation of knowledge of conditional gene-expression, a reasonable level of error in the prediction of gene loci, their exon structure and their translated sequences has persisted in the gene-based datasets of *Z. tritici* for a few years. However, RNA-seq data generated in a recent transcriptome study (Rudd et al., 2015) has been used in-house by the same research group to improve the accuracy of gene annotations in *Z. tritici* (annotations available from EnsemblFungi) (Kersey et al., 2014). The study itself by Rudd et al. has provided further insight into the wheat-*Z. tritici* interaction. RNA-Seq was used to profile expression at five infection time points (symptomless growth: 1 and 4 dpi; host-cell death: 9 and 14 dpi; and asexual sporulation: 21 dpi), detecting expression of at least 80% of predicted loci and identifying >3000 *Z. tritici* genes and >7000 wheat genes as differentially expressed during infection. Variation in the contribution of chromosomes to changes in expression was observed, with CCs exhibiting the highest overall and most significant changes in expression. Conversely ACs exhibited minimal gene expression and few differentially expressed genes. Genes encoding candidate effector proteins were found to be up-regulated *in planta* and expression data was used provide additional supporting evidence for the “top” candidates. Incorporating the above with profiles of differential metabolites over the same time course, Rudd et al. also observed that the switch from latent to necrotrophic growth and reproduction coincided with the activation of plant defense responses and a switch in *Z. tritici*’s nutrient source from lipids and fatty acid stores to carbohydrate metabolism.

5. Genome mutation, adaptation and pathogenicity

ACs have been described by Croll and McDonald as a “cradle for adaptive evolution”. This refers to their potential for retention of higher levels of mutation over time, due to low impact on fitness and correspondingly lower selective pressures (Croll and McDonald, 2012). A two-speed rate of evolution across CCs and

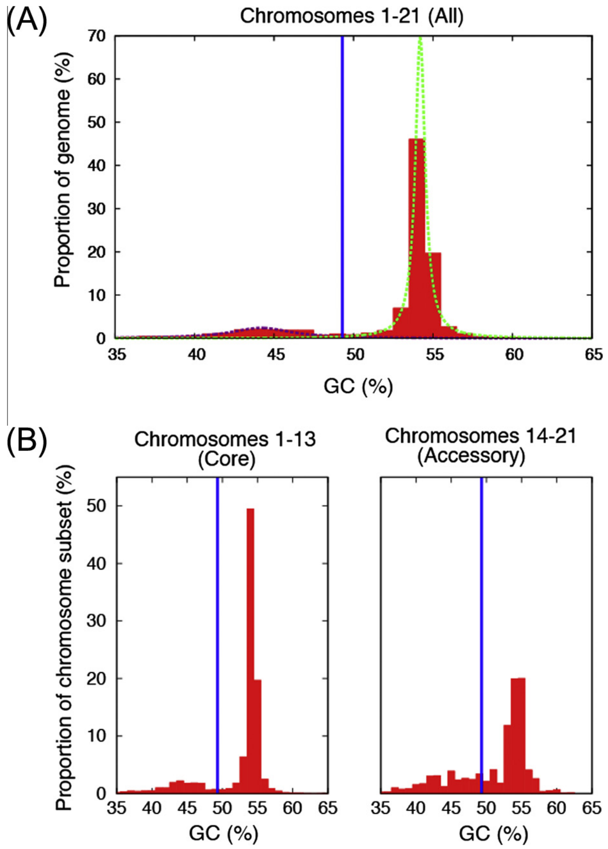


Fig. 1. Histograms of the GC content of segmented regions of the *Z. tritici* IPO323 genome assembly, and the relative proportion of the genome vs GC content. (A) The dotted lines show the two components of the Cauchy mixture model fit to the data. The solid vertical line shows the GC content selected to categorize segments, corresponding to the intersection between the two components of the mixture model. (B) Relative proportions of the *Z. tritici* IPO323 CC and AC chromosome subsets corresponding to ranges of GC content from 35% to 65% (red), illustrating depleted GC content within accessory chromosomes. The solid vertical line (blue) shows the GC content selected to categorize segmented regions of the genome as AT-rich or GC-equilibrated.

ACs allows pathogens like *Z. tritici* to retain its core gene content governing its viability and metabolism, whilst at the same time innovating and mutating genes rapidly in response to host-defences and control strategies. Plant pathogenicity genes have previously been observed to reside on the ACs of other plant-pathogenic fungi species, including *F. oxysporum* f. sp. *lycopersici* (Ma et al., 2010), *F. solani* (Coleman et al., 2009). Despite this, no *Z. tritici* pathogenicity genes have been mapped to ACs (Goodwin et al., 2011) with recent work reporting eight QTLs, all of which mapped to CCs (Mirzadi Gohari et al., 2015).

Z. tritici ACs are richer in repetitive DNA relative to CCs and also undergo recombination (including mesosyntenic rearrangements) more frequently (Croll et al., 2013). The higher repetitive DNA content of ACs, combined with active RIP, can also lead to the rapid mutation of non-repetitive genes flanking RIP-targeted repeats. In the related Dothideomycete plant pathogen *Leptosphaeria maculans* – which is well known for widespread distribution of AT-rich regions (syn. AT-rich isochores) within its genome – the "leakage" of RIP into neighboring non-repetitive has been demonstrated to accelerate the rate of non-synonymous mutation in avirulence genes arranged in a repeat-proximal configuration (Fudal et al., 2009; Hane et al., 2015; Van de Wouw et al., 2010). The *Z. tritici* genome has a similar abundance of AT-rich regions across its genome, particularly within its ACs, which may also have contributed to gene innovation and diversification over time. To assess the prevalence of AT-rich regions in *Z. tritici*, the genome was recursively segmented into regions of differing GC content using Jensen-Shannon divergence (Bernaola-Galván et al., 1996; Elhaik et al., 2010a, 2010b) (stopping criteria: minimum segment length of 1000 bp, t-test (5% significance level) on adjacent average GC values, similar to Oliver et al. (2004). Segmented genome regions were classified according to their constituent GC content. A mixture of two Cauchy distributions was fit to the data using expectation maximization, which identified two peaks centered at 54.2% and 44.2% GC. This allowed a GC boundary to be defined at 49.3% GC in order to distinguish "GC-equilibrated" from "AT-rich" DNA regions (Fig. 1A, Supplementary Data File 1). The AT-rich regions are substantially shorter than GC-equilibrated regions, with average lengths of 10.8 kbp and 48.5 kbp respectively. AT-rich regions were observed on all chromosomes and comprised 18.2% of the total genome length, however the proportion differed between ACs to CCs with AT-rich regions comprising 31.9% in ACs and 16.3% in CCs (Fig. 1B). The majority of annotated genes reside within GC-equilibrated regions, with an overall gene density of 334 genes/Mbp. In contrast, 91 genes have ⩾50% of their length within AT-rich regions and are comparatively sparsely distributed at 12.6 genes/Mbp. Studies of related Dothideomycete plant pathogens, e.g. *Leptosphaeria maculans* and *Passalora fulva*, contain many secreted and/or effector-like genes associated with similar AT-rich regions (de Wit et al., 2012; Rouxel et al., 2011; Van de Wouw et al., 2010). As such, although genes within AT-rich regions contribute little to the overall gene content or gene-expression of *Z. tritici*, they may still play some role in pathogenicity and adaptability.

Table 1
Summary of sequence and bioinformatic resources available for *Z. tritici*.

Resource	Repository	Reference
EST libraries (IPO323)	EMBL (multiple accessions)	Kema et al. (2008)
Nuclear genome (IPO323)	NCBI Nucleotide: ACPE00000000.1	Goodwin et al. (2011)
Mitochondrial genome (IPO323 & STBB1)	NCBI Nucleotide: EU090238	Torriani et al. (2011, 2008)
Gene annotation (IPO323)	JGI (www.jgi.doe.gov)	Goodwin et al. (2011), Grigoriev et al. (2013)
	EnsemblFungi	Kersey et al. (2014)
Predicted secretome (IPO323)	n/a	Morais do Amaral et al. (2012)
Gene expression: RNA-Seq data for host and pathogen during wheat infection	NCBI BioProject: 196595	Yang et al. (2013)
	NCBI BioProject: 278138	Rudd et al. (2015)
Gene expression: RNA-Seq data for wheat infection vs *B. distachyon* infection	NCBI GEO: GSE54874	Kellner et al. (2014)
Mapped QTLs and effector candidate list (IPO323)	n/a	Mizardi Gohari et al. (2015)
Resequencing	n/a	Croll et al. (2013)_ McDonald et al. (1991)

6. Conclusion

The sum of sequence and bioinformatic resources generated over the last decade for *Z. tritici* (Table 1) has significantly advanced knowledge of this fungal pathogen and its host-interactions. Foremost among these has been the generation of the whole-genome sequence – a solid foundation upon which multiple layers of additional information have been juxtaposed.

Various bioinformatic techniques have been applied to predicted gene regions and their protein products with a view to focussing on those with properties relevant to pathogenicity. This has been complemented by analysis of gene expression during infection, which has highlighted generic and host-specific pathogenicity genes in *Z. tritici*. The chromosome-length genome sequence has itself been a powerful tool in understanding genome mutation mechanisms influencing adaptation in *Z. tritici* and assessing the "genomic context" of its putative pathogenicity genes.

Appendix A. Supplementary material

Supplementary data associated with this article can be found, in the online version, at http://dx.doi.org/10.1016/j.fgb.2015.04.011.

References

Aguileta, G. et al., 2014. High variability of mitochondrial gene order among fungi. Genome Biol. Evol. 6, 451–465.

Bernaola-Galván, P. et al., 1996. Compositional segmentation and long-range fractal correlations in DNA sequences. Phys. Rev. E 53, 5181.

Coleman, J.J. et al., 2009. The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. PLoS Genet. 5, e1000618.

Croll, D., McDonald, B.A., 2012. The accessory genome as a cradle for adaptive evolution in pathogens. PLoS Pathog. 8, e1002608.

Croll, D. et al., 2013. Breakage-fusion-bridge cycles and large insertions contribute to the rapid evolution of accessory chromosomes in a fungal pathogen. PLoS Genet. 9, e1003567.

de Wit, P.J. et al., 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 Genet. 8, e1003088.

Elhaik, E. et al., 2010a. Comparative testing of DNA segmentation algorithms using benchmark simulations. Mol. Biol. Evol. 27, 1015–1024.

Elhaik, E. et al., 2010b. Identifying compositionally homogeneous and nonhomogeneous domains within the human genome using a novel segmentation algorithm. Nucleic Acids Res. 38, e158-e158.

Fudal, I. et al., 2009. Repeat-induced point mutation (RIP) as an alternative mechanism of evolution toward virulence in *Leptosphaeria maculans*. Mol. Plant Microbe Interact. 22, 932–941.

Goodwin, S.B. et al., 2011. Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genet. 7, e1002070.

Grigoriev, I.V. et al., 2013. MycoCosm portal: gearing up for 1000 fungal genomes. Nucl. Acids Res., gkt1183

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

Hane, J.K., Oliver, R.P., 2010. *In silico* reversal of repeat-induced point mutation (RIP) identifies the origins of repeat families and uncovers obscured duplicated genes. BMC Genom. 11, 655.

Hane, J.K. et al., 2011a. A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. Genome Biol. 12, R45.

Hane, J.K. et al., 2011b. Genomic and comparative analysis of the class dothideomycetes. Evolution of Fungi and Fungal-Like Organisms. Springer, pp. 205–229.

Hane, J.K. et al., 2015. Repeat-Induced Point Mutation: A Fungal-Specific, Endogenous Mutagenesis Process. Genetic Transformation Systems in Fungi, vol. 2. Springer, pp. 55–68.

Kellner, R. et al., 2014. Expression profiling of the wheat pathogen *Zymoseptoria tritici* reveals genomic patterns of transcription and host-specific regulatory programs. Genom. Biol. Evol. 6, 1353–1365.

Kema, G.H., van Silfhout, C.H., 1997. Genetic variation for virulence and resistance in the Wheat-*Mycosphaerella graminicola* Pathosystem III. Comparative seedling and adult plant experiments. Phytopathology 87, 266–272.

Kema, G.H. et al., 2002. A combined amplified fragment length polymorphism and randomly amplified polymorphism DNA genetic kinkage map of *Mycosphaerella graminicola*, the septoria tritici leaf blotch pathogen of wheat. Genetics 161, 1497–1505.

Kema, G.H. et al., 2008. Large-scale gene discovery in the septoria tritici blotch fungus *Mycosphaerella graminicola* with a focus on *in planta* expression. Mol. Plant Microbe Interact. 21, 1249–1260.

Kersey, P.J. et al., 2014. Ensembl Genomes 2013: scaling up access to genome-wide data. Nucleic Acids Res. 42, D546–D552.

Lee, W.S. et al., 2014. *Mycosphaerella graminicola* LysM effector-mediated stealth pathogenesis subverts recognition through both CERK1 and CEBiP homologues in wheat. Mol. Plant Microbe Interact. 27, 236–243.

Ma, L.J. et al., 2010. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature 464, 367–373.

Manning, V.A. et al., 2013. Comparative genomics of a plant-pathogenic fungus, *Pyrenophora tritici-repentis*, reveals transduplication and the impact of repeat elements on pathogenicity and population divergence. G3: Genes| Genomes| Genetics 3, 41–63.

McDonald, B., Martinez, J., 1991. Chromosome length polymorphisms in a *Septoria tritici* population. Curr. Genet. 19, 265–271.

Mirzadi Gohari, A., 2015. Effector discovery in the fungal wheat pathogen *Zymoseptoria tritici*. Mol. Plant. Pathol. http://dx.doi.org/10.1111/mpp.12251.

Morais do Amaral, A. et al., 2012. Defining the predicted protein secretome of the fungal wheat leaf pathogen *Mycosphaerella graminicola*. PLoS ONE 7, e49904.

Ohm, R.A. et al., 2012. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. PLoS Pathog. 8, e1003037.

Oliver, J.L. et al., 2004. IsoFinder: computational prediction of isochores in genome sequences. Nucleic Acids Res. 32, W287–W292.

Rouxel, T. et al., 2011. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. Nature Commun. 2, 202.

Rudd, J., et al., 2015. Transcriptome and metabolite profiling the infection cycle of *Zymoseptoria tritici* on wheat (*Triticum aestivum*) reveals a biphasic interaction with plant immunity involving differential pathogen chromosomal contributions, and a variation on the hemibiotrophic lifestyle definition. Plant Physiol., pp. 114.255927.

Stukenbrock, E.H. et al., 2010. Whole-genome and chromosome evolution associated with host adaptation and speciation of the wheat pathogen *Mycosphaerella graminicola*. PLoS Genet. 6, e1001189.

Torriani, S.F. et al., 2008. Intraspecific comparison and annotation of two complete mitochondrial genome sequences from the plant pathogenic fungus *Mycosphaerella graminicola*. Fungal Genet. Biol. 45, 628–637.

Torriani, S.F. et al., 2011. Evolutionary history of the mitochondrial genome in *Mycosphaerella* populations infecting bread wheat, durum wheat and wild grasses. Mol. Phylogenet. Evol. 58, 192–197.

Van de Wouw, A.P. et al., 2010. Evolution of linked avirulence effectors in *Leptosphaeria maculans* is affected by genomic environment and exposure to resistance genes in host plants. PLoS Pathog. 6, e1001180.

Wittenberg, A.H. et al., 2009. Meiosis drives extraordinary genome plasticity in the haploid fungal plant pathogen *Mycosphaerella graminicola*. PLoS ONE 4, e5863.

Yang, F. et al., 2013. Transcriptional reprogramming of wheat and the hemibiotrophic pathogen *Septoria tritici* during two phases of the compatible interaction. PLoS ONE 8, e81606.

Zhan, J. et al., 2003. The global genetic structure of the wheat pathogen *Mycosphaerella graminicola* is characterized by high nuclear diversity, low mitochondrial diversity, regular recombination, and gene flow. Fungal Genet. Biol. 38, 286–297.

Zhan, J. et al., 2004. Evidence for natural selection in the mitochondrial genome of *Mycosphaerella graminicola*. Phytopathology 94, 261–267.

Chapter 6 A comprehensive survey of AT-rich regions in fungal genomes

6.1 Attribution statement

Title: A comprehensive survey of AT-rich regions in fungal genomes

Authors: Alison C Testa, Richard P Oliver, and James K Hane

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 Ph. D. candidate (ACT) made the following contributions to this chapter:

- ACT wrote the software and conducted the bioinformatics analysis
- ACT summarised the data and generated the figures and tables
- ACT wrote the manuscript

The following contributions were made by co-authors:

- JKH and ACT conceived the study
- JKH reviewed and refined the manuscript
- RPO reviewed the manuscript

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

________________________________ ________________

Alison C Testa Date

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

________________________________ ________________

Richard P Oliver Date
(Principal supervisor and co-author)

________________________________ ________________

James K Hane (co-author) Date

OcculterCut: A Comprehensive Survey of AT-Rich Regions in Fungal Genomes

Alison C. Testa[1],*, Richard P. Oliver[1], and James K. Hane[1,2]

[1]Department of Environment & Agriculture, Centre for Crop and Disease Management, Curtin University, Perth, Australia

[2]Curtin Institute for Computation, Curtin University, Perth, Australia

*Corresponding author: E-mail: 13392554@student.curtin.edu.au.

Accepted: May 14, 2016

Abstract

We present a novel method to measure the local GC-content bias in genomes and a survey of published fungal species. The method, enacted as "OcculterCut" (https://sourceforge.net/projects/occultercut, last accessed April 30, 2016), identified species containing distinct AT-rich regions. In most fungal taxa, AT-rich regions are a signature of repeat-induced point mutation (RIP), which targets repetitive DNA and decreases GC-content though the conversion of cytosine to thymine bases. RIP has in turn been identified as a driver of fungal genome evolution, as RIP mutations can also occur in single-copy genes neighboring repeat-rich regions. Over time RIP perpetuates "two speeds" of gene evolution in the GC-equilibrated and AT-rich regions of fungal genomes. In this study, genomes showing evidence of this process are found to be common, particularly among the Pezizomycotina. Further analysis highlighted differences in amino acid composition and putative functions of genes from these regions, supporting the hypothesis that these regions play an important role in fungal evolution. OcculterCut can also be used to identify genes undergoing RIP-assisted diversifying selection, such as small, secreted effector proteins that mediate host-microbe disease interactions.

Key words: fungi, genome evolution, two-speed genome, repeat-induced point mutation, isochore.

Introduction

The fungal kingdom contains many highly specialized organisms of interest to the agriculture, food, and medical industries. Specialization and adaptation are keys to the success of many fungi, with accelerated evolution capabilities increasingly recognized as facilitating these processes. For plant pathogens in particular, rapid evolution to overcome chemical control measures and host resistance is essential to their survival. Genome sequences of fungi—small but nonetheless eukaryotic—have revealed many features that challenge the conventional paradigm of genomic stability and contribute to their ability to evolve.

Even before whole genome sequences of fungi were available, genetic variability in the form of dispensable chromosomes (Tzeng et al. 1992) and chromosome length polymorphisms (Zolan 1995) was observed using pulsed-field gel electrophoresis. *Saccharomyces cerevisiae* was the first fungus to have its genome sequence published (Goffeau et al. 1996). Genomics of the filamentous fungi began some years later with the genome sequences of model organisms *Neurospora crassa* in 2003 (Galagan et al. 2003) and *Aspergillus nidulans* in 2005 (Galagan et al. 2005). An avalanche of sequencing projects followed to the point where hundreds of fungal genomes, representing a range of different lifestyles and taxa, are now publically available. These data have enabled further studies of genomic variability and adaptability in fungi relating to multiple mechanisms for sex (Heitman et al. 2013), chromosome length polymorphism (Zolan 1995), horizontal gene transfers (Hane et al. 2011; Gardiner et al. 2012; Sun et al. 2013; Dhillon et al. 2015), repeats and transposable elements (Spanu 2012), conditional dispensability of DNA segments or whole chromosomes (Coleman et al. 2009; Croll and McDonald 2012; Croll et al. 2013), and the conservation and breakdown of synteny and co-linearity (Hane et al. 2011).

The GC-content of genomic DNA has historically been of broad interest in the life sciences. In fungi, broad variation in DNA GC-content (38–63%) was observed from melting-temperature and buoyant-density measurements (Storck 1965) long before whole genome sequence data became available. Genome GC-content is now reported as one of the basic attributes of a genome assembly, confirming wide variation in the GC-content of fungal genomes. For example, *Pneumocystis jirovecii*—which causes severe lung infections in immunocompromised humans—has a genome GC-content of just 29.5% (Cissé et al. 2013). In contrast, the wood degrading fungus *Phanerochaete chrysosporium* has a much higher genome GC-content of 57% (Martinez et al. 2004). GC-content variation also exists within genomes, a property that was first observed in warm-blooded vertebrates (Bernardi et al. 1985). The human genome was observed to be composed of a mosaic of long stretches (typically > 300 kb) of DNA homogeneous in base composition (Bernardi 2000). These regions were termed "isochores" and can be grouped into families based on their GC-content. The presence of isochores has been documented in many species, including some fungi (Costantini et al. 2013). However, in fungal genomics the term "isochore" or "AT-isochore" has sometimes been used to refer specifically to sequence regions with markedly depleted GC-content (AT-rich). As this study focuses on these fungal AT-rich regions rather than what is traditionally termed an isochore, we refer to them as "AT-rich regions" from here on. AT-rich regions appear to differ from traditional isochores in their length and suspected origin (Rouxel et al. 2011) and do not necessarily have the homogeneous base composition implicit when using the terminology isochore. Fungal genomes with large proportions of AT-rich regions exhibit a distinctive bimodal pattern of GC-content bias. Observations of higher evolutionary rates in repeat rich genome compartments of filamentous plant pathogens (Raffaele et al. 2010), coupled with evidence of host jumps (Raffaele et al. 2010) and rapid adaptation to crop resistance (Fudal et al. 2009; Van de Wouw et al. 2010), have given rise to the concept of "two-speed" genome evolution. This concept describes a genome in which gene content has been compartmentalized into two types of sequence regions: regions containing the essential or "core" genome and the variable genome, often characterized by a higher density of repetitive elements and in some cases AT-rich sequence.

One mechanism by which AT-rich regions can occur is repeat-induced point mutation (RIP), a process specific to fungi and primarily considered to act as a defense against transposon propagation. RIP was initially observed in the saprobic Ascomycete *N. crassa* (Selker et al. 1987) and a cytosine methyltransferase homologue (*rid*) gene was shown to be essential for RIP (Freitag and Williams 2002). RIP has been identified experimentally in *Leptosphaeria maculans* (Idnurm and Howlett 2003; Fudal et al. 2009; Van de Wouw et al. 2010; Rouxel et al. 2011*), Fusarium graminearum* (Cuomo et al. 2007), *Magnaporthe oryzae* (Nakayashiki et al. 1999; Ikeda and Nakayashiki 2002; Dean et al. 2005; Farman 2007), and *Podospora anserina* (Graïa et al. 2001) with in silico evidence supporting RIP activity in many more species within the subphylum Pezizomycotina (Hane and Oliver 2008, 2010; Clutterbuck 2011; Goodwin et al. 2011) and some species within the Basidiomycota (Horns et al. 2012). RIP occurs during heterokaryon formation prior to meiosis, targeting repetitive DNA above a certain length (Watters et al. 1999) and identity (Cambareri et al. 1991) (400 bp and 80% in *N. crassa*), with irreversible transitions from cytosine to thymine bases (i.e., C to T). RIP has also been observed to leak beyond repetitive DNA (Irelan et al. 1994) into nearby single copy and often genic regions, in some cases mutating genes with known roles in pathogenicity (Fudal et al. 2009; Van de Wouw et al. 2010). Within genome assemblies, the observable impact of RIP is the depletion of GC-content within, and to a lesser extent nearby, repeats. Over time, the GC-content of RIP-affected sequence becomes distinct from nonRIP affected regions and can be described as "AT-rich".

Leptosphaeria maculans was the first published fungal genome reported to have a significant proportion of distinctly AT-rich regions, accounting for approximately one-third of the assembly (Rouxel et al. 2011). AT-rich regions within *L. maculans* were found to have a few genes but many transposons and showed strong evidence of RIP. While it has been suggested that *L. maculans* is unusual in its high component of AT-rich regions (Raffaele and Kamoun 2012; Lo Presti et al. 2015), several other studies have identified AT-rich regions in fungal genomes including *Blastomyces dermatitidis* (Clutterbuck 2011) (note: genome unpublished), *Passalora fulva* (de Wit et al. 2012), multiple *Epichloë* spp. (Schardl et al. 2013), and *Zymoseptoria tritici* (Croll et al. 2013; Testa, Oliver, et al. 2015). Furthermore, Clutterbuck (2011) found widespread sequence-based evidence of RIP in the examination of 49 filamentous Ascomycetes (subphylum Pezizomycotina), suggesting that bimodal GC bias and high AT-rich content may be common across this large taxon.

Within the discipline of plant pathology, interest in AT-rich regions has been fuelled by observations of genes encoding avirulence/effector-like proteins within or close to AT-rich regions (e.g., *L. maculans* genes *AvrLm6*, *AvrLm4-7* and *AvrLm1*; Gout et al. 2006; Fudal et al. 2009). Effector proteins are typically small, secreted proteins and play an important role in interactions with the host plant. The frequent proximity of effector genes to repetitive regions is well documented (Lo Presti et al. 2015). It has been proposed that pathogenic fungi with effector genes in or near AT-rich RIP hotspots have been selected by evolution as they have an advantageous mechanism by which to rapidly lose or modify these genes, avoid recognition by host defenses and thus repeatedly overcome newly deployed resistance genes (Oliver 2012).

Despite the clear motivations for investigating AT-rich regions, several obstacles limit our understanding of this interesting genome feature. The first is in identifying and defining AT-rich regions. In past studies AT-rich regions within fungi have generally been identified using a "moving window" approach that reports the GC-content within a series of windows over the entire length of the genome. The disadvantage of such an approach is the uncertainty in region boundaries. Another method by which AT-rich regions have been identified is by annotating repeats and recording their GC-content, but this method does not account for repeats that have been degraded (e.g., by RIP) to the point where they are not recognized by repeat-detection software (Hane and Oliver 2010).

Several approaches exist to identify isochore-like regions have been previously described. A highly successful method uses the Jensen–Shannon divergence (Elhaik, Graur, Josic 2010). In isochore studies, genome segments are classified by GC-content into isochore families L1 ($<37\%$ GC), L2 (37–42% GC), H1 (42–47% GC), H2 (47–52% GC) and H3 ($>52\%$ GC) (Oliver et al. 2001; Costantini et al. 2013). AT-rich regions in *L. maculans* were reported as 34% GC (Rouxel et al. 2011), compared with the 44% GC-content of AT-rich regions within *Z. tritici* (Testa, Oliver, et al. 2015). This variation in closely related species indicates that broadly applied and arbitrary GC-content categories may not be suitable for investigating AT-rich regions within fungal genomes. This highlights the need for a systematic method to measure GC-content distributions in fungal genomes and define different regions. Furthermore, small-scale studies often differ in their method of identifying and analyzing AT-rich regions, making comparisons difficult. This relates to the second obstacle—that although there have been detailed studies of a few individual fungi and wider surveys of repetitive elements and RIP within fungal genomes, we lack a taxonomically broad survey and understanding of AT-rich regions across the fungi.

In overcoming these obstacles, we present: firstly, a reproducible method and software tool, OcculterCut, which facilitates AT-content analysis in genomes; and secondly, a survey of the AT-content of >500 published fungal genomes. We identify species in which RIP (or other means of mutagenesis with similar effects) has led to significant proportions of AT-rich regions. By extension, this has predicted several species in which RIP-mediated "two-speed" genome evolution is likely to have significantly influenced their roles in plant association. OcculterCut also returns information about the proximity of genes to AT-rich regions, making it a useful tool for identifying genes likely to be under the influence of RIP-mediated "hypermutation", such as candidate avirulence/effector genes or other genes involved in rapid evolution and specialization.

Materials and Methods

Collecting Genomes for Survey

Published fungal genome sequences were obtained from multiple sources, a full list of which is provided in the Supplementary Material online (supplementary table S1, Supplementary Material online). As assembled genome sequence(s) could potentially be contaminated with AT-rich mitochondrial sequences, a filtering process was carried out on all genomes surveyed. Firstly, a database of fungal mitochondrial sequences (mtDNAs) was downloaded from NCBI's Organelle Genome resource (NCBI Resource Coordinators 2014) (supplementary table S6, Supplementary Material online—accessions of sequences used in mtDNA screen set). BLASTn (Altschul et al. 1990) was used to search for alignments between scaffolds in the surveyed genomes and the mtDNA database ($e \leq 1e\text{-}10$ and $\geq 75\%$ identity). Scaffold coverage of mtDNA matches was calculated via BEDtools (version 2.17) (Quinlan and Hall 2010) coverageBed and scaffolds were removed if 50% or more of the scaffold length was covered by alignments (see supplementary table S7, Supplementary Material online, matches to mtDNA screen). A bash script for mtDNA filtering has been included in the OcculterCut release files (available from https://sourceforge.net/projects/occultercut, last accessed April 30, 2016).

The lifestyle of the surveyed species was documented in order to reveal possible links between AT-rich regions and fungal lifestyle. The broad lifestyle categories used were saprobe, pathogen, and symbiont. Symbionts were separated into plant symbionts, which account for the majority of symbionts surveyed, and other symbionts. Pathogens were separated into plant pathogens and animal pathogens, and plant pathogens were further separated into obligate biotrophs, nonobligate biotrophs, hemibiotrophs, and necrotrophs as in (Spanu 2012). Sources of information about fungal lifestyle are cited in supplementary table S1, Supplementary Material online. We note that in plant pathology the reliability of such lifestyle categories has become a contentious issue for many species, and in such cases, we aimed to identify the recent consensus in the literature.

Identifying AT-Rich Regions with OcculterCut

The GC-contents of genome assemblies included in this survey were evaluated using a procedure that we have presented as a software tool, OcculterCut. The steps employed by OcculterCut in segmenting genome sequence and identifying AT-rich and GC-equilibrated regions are described below.

Genome Segmentation

Assembled genome sequence is segmented into regions of differing GC-content using the Jensen–Shannon divergence (D_{JS}), based on the methods described in past isochore studies (Bernaola-Galván et al. 1996; Elhaik, Graur, Josic 2010; Elhaik, Graur, et al. 2010). A border is moved along the query sequence and at each position the Jensen–Shannon divergence is calculated. At the position where the Jensen–Shannon divergence maximized, the sequence is split into two proposed subsequences, and the split is retained providing certain conditions are met (fig. 1*A*). This process continues recursively until a proposed split is rejected. The conditions that decide whether a split is rejected are based on the size of the segments to the left and right of a potential split and whether the

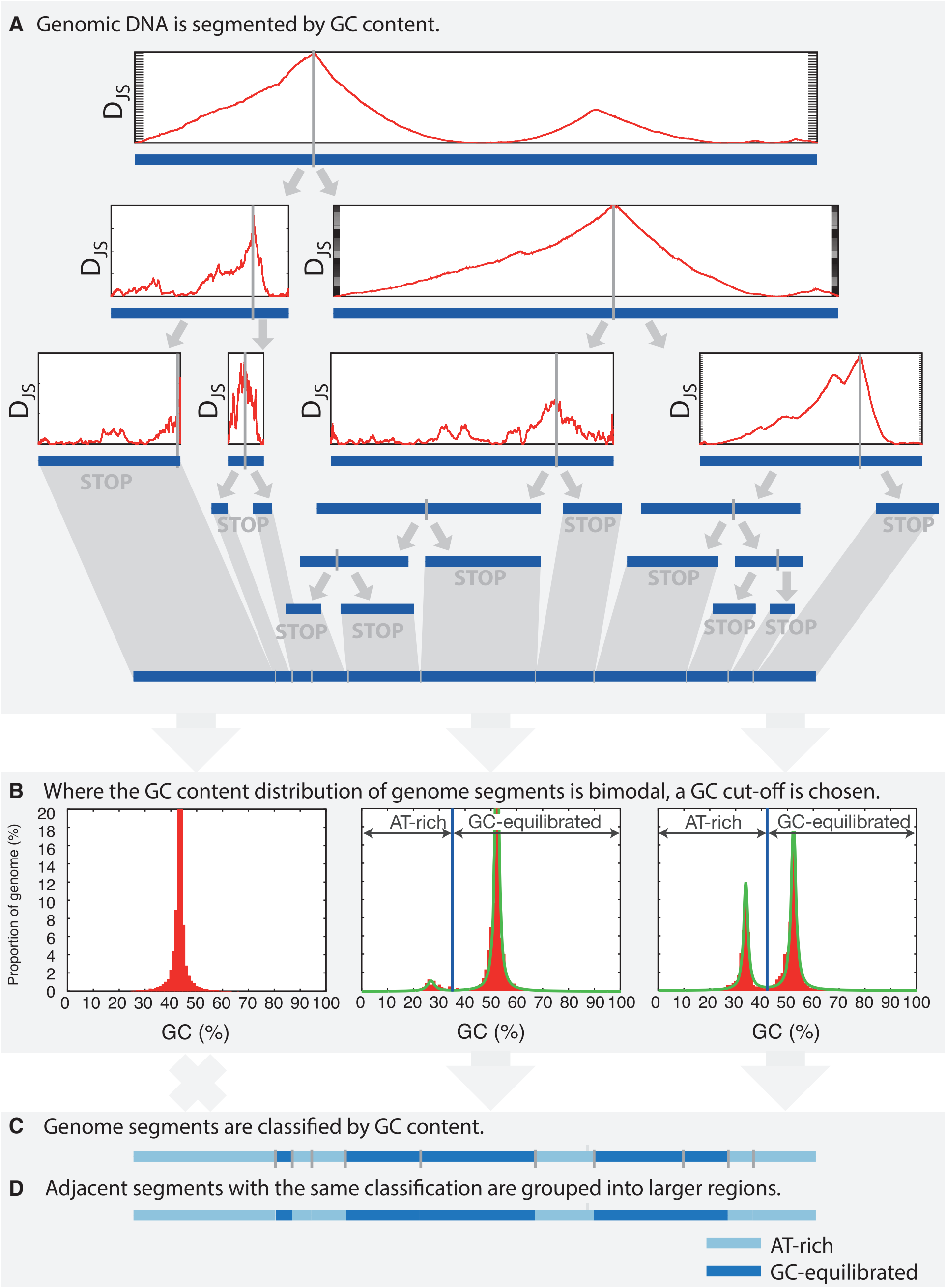

A Genomic DNA is segmented by GC content.
D_JS
STOP
B Where the GC content distribution of genome segments is bimodal, a GC cut-off is chosen.
Proportion of genome (%)
0 2 4 6 8 10 12 14 16 18 20
0 10 20 30 40 50 60 70 80 90 100
GC (%)
AT-rich
GC-equilibrated
C Genome segments are classified by GC content.
D Adjacent segments with the same classification are grouped into larger regions.
AT-rich
GC-equilibrated

Fig. 1.—The basic steps employed by OcculterCut in annotating AT-rich and GC-equilibrated genome regions. (*A*) DNA sequence is recursively split into segments of differing GC-content. The point at which a sequence is split into two segments is chosen to be where the Jensen–Shannon divergence statistic (D_{JS} on the *y*-axis of cartoon plots in (*A*) is maximized. (*B*) AT-rich and GC-equilibrated genome segments are identified, and a GC-content cut-off is selected. (*C*) Genome segments are categorized as either AT-rich or GC-equilibrated, and the genome segments are grouped into broader regions.

potential split would result in two adjacent segments with a statistically significant difference in GC-contents. In this study, the minimum segment length of 1 kb was set. To assess the statistical significance of the difference between GC-contents either side of the split, the left-hand and right-hand segments are divided into nonoverlapping 300 bp sections and the GC-content of each section is recorded. Similar to Oliver et al. (2004), a *t*-test is then used to decide whether the GC-contents of the left-hand sections differ significantly from the GC-contents of the right-hand sections.

Segment Classification

The genome segmentation process results in genome sequence designated into segments of differing GC-content. In assessing the AT-rich region content of a genome the next step is to determine whether AT-rich genome segments are present and, if present, to categorize segments and as either AT-rich or GC-equilibrated. This requires an assessment of the GC-content distribution of the genome segments. The GC-content of each genome segment, including un-segmented contigs <1 kb, and the proportion of the genome taken up by that particular segment (segment length/genome assembly size) is recorded. These data can be visualized as a plot of the GC-content of the genome segments against the proportion of the genome (see examples in figs. 1*B* and 2). A mixture of two Cauchy distributions is fit to the data. That is, the data are assumed to be of the form:

$$\begin{aligned} f(x;\omega,x_{01},\gamma_2) &= \omega f_1(x;x_{01},\gamma_1) + (1-\omega)f_2(x;x_{02},\gamma_2) \\ &= \frac{\omega}{\pi}\frac{\gamma_1}{(x-x_{01})^2+\gamma_1^2} + \frac{1-\omega}{\pi}\frac{\gamma_2}{(x-x_{02})^2+\gamma_2^2} \end{aligned}$$

where $0 \leq \omega \leq 1$ and describes the weight of each peak, x_{01} and x_{02} describe the *x*-axis position (GC-content) of each peak and γ_1 and γ_2 relate to the peak widths. Examples of Cauchy distributions fit to genome GC-content distributions are shown in figure 1*B*. The Cauchy distribution mixture model of the GC-content distribution (as described above) is fit to the genome segment GC-content data by determining the values ω, x_{01}, γ_1, x_{02}, and γ_2 using expectation maximization. Whether genome segments can be grouped into AT-rich and GC-equilibrated regions and the GC-content cut-off used to define each region type are based on this Cauchy distribution mixture model.

Categorization of genome segments as AT-rich or GC-equilibrated is not carried out in all cases; many genomes do not have AT-rich regions and the plot of the GC-content is unimodal (figs. 1*B* and 2*A–C*). In some cases, the plot of the GC-content may suggest multiple peaks, but these overlap too much to allow the segments to be classified reliably as from one or the other. The decision to categorize the segments is therefore based on the %GC separation between the two peaks (the difference between x_{01} and x_{02}), the existence of a local minimum in the Cauchy distribution mixture model between the two peaks, and the confidence with which segments can be classified. The minimum GC-content peak separation is set at a default 5%. This filters out cases where the GC-content distribution is unimodal or the peak separation of is too small to be of interest. Segment classification is always carried out where peak separation is ≥10% GC and a minimum in the Cauchy distribution mixture model can be identified between the two peaks. In cases where the peak separation is between 5% and 10%, region grouping is only carried out where a minimum can be identified between the peaks and 75% of the segments in each group can be classified with 75% or better confidence.

Where classification of genome segments is carried out, segments are classified as AT-rich or GC-equilibrated according to whether they have a GC-content above or below a cut-off GC value set as the local minimum between the two peaks in the Cauchy distribution mixture model (fig. 1*B*). Adjacent genome segments with the same classification are then merged into single, larger regions (fig. 1*C* and *D*). To allow the user to explore different grouping of genome segments, the Cauchy fit-to-data can be disabled in favor of the user specifying two or more GC-content intervals on which to group the genome segments.

Outputs

The presented software, OcculterCut, automates the described genome segmentation and segment classification steps and outputs a number of files containing the results of these steps. A brief description of the content of the OcculterCut outputs is given here and detailed description of output file names and content is given in the instruction manual that accompanies OcculterCut.

A General Feature Format (GFF) (Wellcome Trust Sanger Institute 2015) containing the genomic coordinates of genome segments resulting from the GC-content segmentation is output. This is accompanied by a text file containing a list of GC-content intervals (from 0% to 100% GC in 1% intervals) and the proportion of the genome covered by genome segments with a GC-content within each interval. These data can be plotted with GC-content on the *x*-axis and the proportion of the genome on the *y*-axis to produce a plot of the GC-content distribution (see examples in fig. 2). These data are returned for all genomes, regardless of whether the genome is bimodal or not.

In cases where the genome is found to be bimodal, the parameters of the Cauchy mixture fit-to-data (see the segment classification section) are returned in a text file along with the GC-content cut-off used to define AT-rich and GC-equilibrated regions. A GFF containing the genomic coordinates of AT-rich and GC-equilibrated regions is also included (fig. 1*C*). The user can choose to supply a GFF of gene locations when running OcculterCut, in which case a summary of the distance from each gene to the closest AT-rich region or scaffold end is returned. This feature may be particularly useful to phytopathogen researchers interested in identifying effector gene candidates. Single, di- and tri-nucleotide frequencies from AT-rich and GC-equilibrated regions are also returned as text files, allowing the user to assess sequence biases.

Analyzing AT-Rich Regions and Their Gene Content

In assemblies found to have an AT-rich region component ≥5%, AT-rich and GC-equilibrated regions were compared. This included looking at the length of each of the region types,

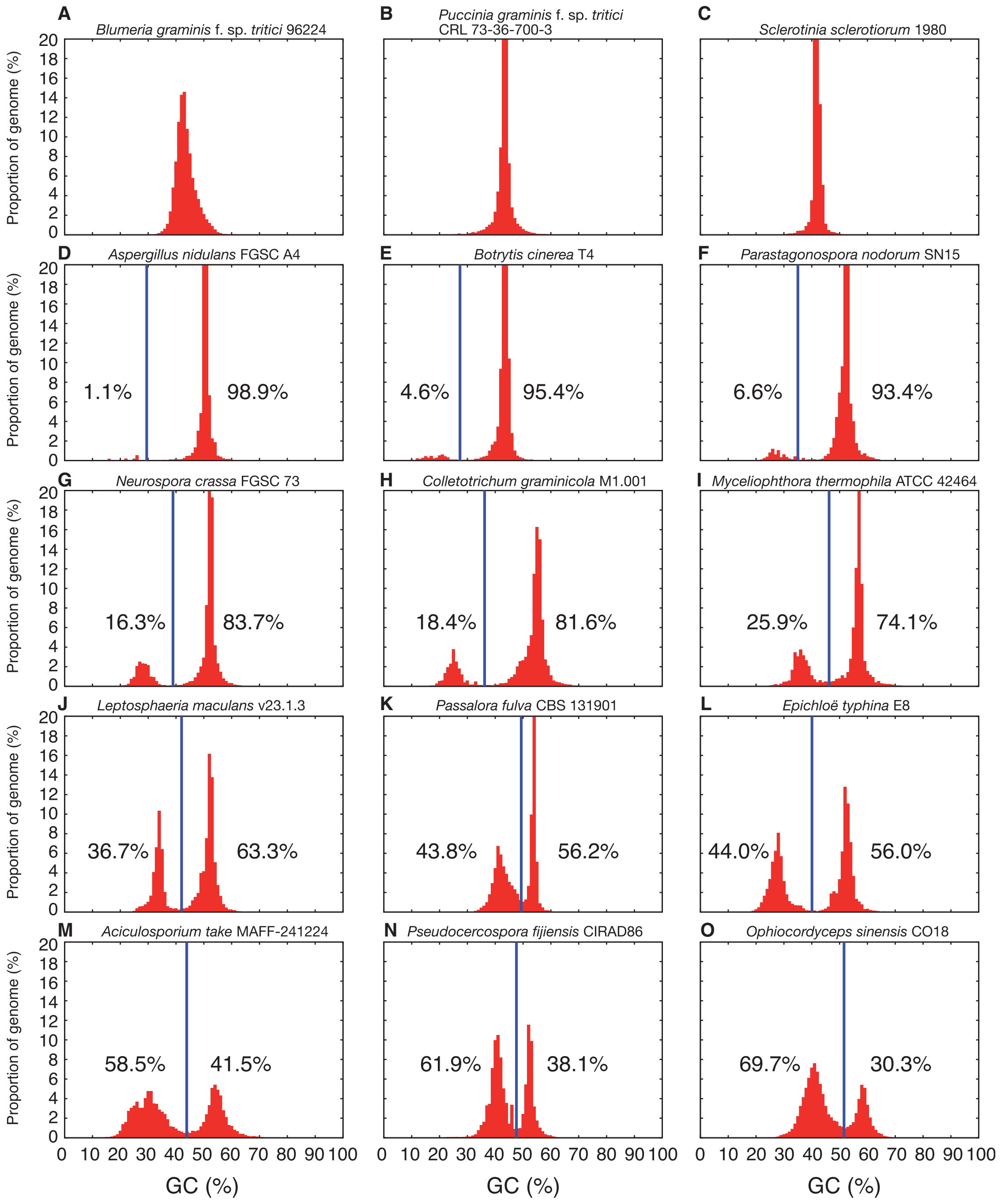


Fig. 2.—Example GC-content plots of 15 surveyed fungal genomes are displayed, arranged from (*A*) to (*O*) ordered in increasing AT-rich region composition. Diversity in the GC-content of peaks, their shape, spacing, and height can be observed. Vertical blue lines show the GC cut-off chosen by OcculterCut and used to classify genome segments into distinct AT-rich and GC-equilibrated region types. The percentage values shown either side of the vertical blue lines indicate the percentage of the genome classified as AT-rich (left) and GC-equilibrated (right).

di-nucleotide frequencies within each region type, and gene content. Additional analysis of the difference between coding sequences from genes within AT-rich and GC-equilibrated regions was conducted on a subset of species. For this purpose, incomplete coding sequences (lacking a methionine at the start), coding sequences containing an in-frame stop codon, and coding sequences <90 nucleotides (corresponding to a protein length <30 amino acids) were excluded from the analysis. This was done in an attempt to remove false positive or incorrect coding sequences from the analysis that may bias results. An additional filtering step was carried out to ensure the coding sequences being analyzed were not false positive predictions made on transposons sequences. TBLASTn (Camacho et al. 2009) was used to search for alignments between the Repbase (Jurka et al. 2005) fungal database and translated protein sequences from each genome. Where one or more of the alignments had e-value(s) $<10^{-5}$ and covered 50% or more of the query protein sequence, the protein sequence and corresponding nucleotide sequence were removed from the sets used for analysis. Finally, analysis was carried out only where 50 or more genes with coding sequences meeting the above criteria could be identified in the AT-rich regions of a particular genome assembly. For calculations of gene density the full set of annotations was included regardless of the above criteria.

When comparing the amino acid composition of proteins from genes within and outside AT-rich regions, the frequency of each amino acid was calculated for each gene from the different sets. A Mann–Whitney *U* test ($P=0.05$) was used to compare the distributions of frequency values between the different sets. Codon frequencies for each amino acid were calculated within and outside AT-rich regions on the full set of coding sequence rather than a per gene basis.

Hmmscan (Eddy 2011) was used to identify Pfam domains in proteins within and outside AT-rich regions (automatic cutoff determined using the parameter—cut-ga) using the Pfam library version 28.0 (Finn et al. 2014). A Fisher's exact test (two sided, implemented in R) was then used to compare the number proteins found to contain a particular Pfam domain in relation to the number of proteins found to contain a different Pfam domain in protein sets within and outside AT-rich regions. An additional comparison was made between proteins with a Pfam domain hit and with no Pfam domain hit. A *P* value of 0.05 was used, with a Bonferroni correction for the total number of tests for significance carried out for each species.

The number of secreted genes, as predicted by SignalP v. 4.1 (Petersen et al. 2011), was recorded for the sets of genes within and outside AT-rich regions for each species. A Fisher's exact test (two sided, *P* value 0.05, implemented in R) was used to compare the number of proteins that were predicted to be secreted in relation to the number of proteins that were not predicted to be secreted in protein sets within and outside AT-rich regions.

Results

OcculterCut—A Tool for AT-Rich Region Analysis

OcculterCut was used to apply GC-content genome segmentation (see fig. 1 and Materials and Methods) to over 500 published fungal genomes (see supplementary table S1, Supplementary Material online, for list of genomes). Plotting the proportion a given genome accounted for by genome segments with a GC-content within 1% intervals (from 0% to 100%) offered a way of visualizing the GC-content distribution of each genome (see examples in fig. 2). Such plots of the surveyed genomes revealed diversity in peak GC-content(s) and distribution shapes and spreads (fig. 2). Heat map summary plots of the GC-content of segmented genomes of the species surveyed are displayed next to dendograms generated according to taxonomic classifications in figures 3 and 4. In most cases, GC-content distributions could be classified as unimodal (having a single peak, fig. 2*A*–*C*) or bimodal (two peaks, see fig. 2*D*–*O*). OcculterCut carried out this classification automatically (see Materials and Methods and fig. 1*B* and *C*, examples shown in fig. 2), and in genomes where distinctly AT-rich segments were present, the genome segments were grouped into broader AT-rich and GC-equilibrated region types.

The majority (~63%) of surveyed genomes had unimodal GC-content distributions, with variation in the mode, peak width and shape observed between species (fig. 2*A*–*C*). GC-content plots of genomes known from previous studies to have distinct AT-content patterns showed two clearly separate peaks (fig. 2, *L. maculans* [J], *P. fulva* [K]). Additional genomes with bimodal GC-content distributions (see fig. 2*D*–*O*) confirm the efficacy of the GC-content genome segmentation method in distinguishing these regions. Where the GC-content of a segmented genome was bimodal (fig. 2*D*–*O*), OcculterCut selected a local minima between the peaks to divide the distribution and group genome segments into broader categories: AT-rich and GC-equilibrated. The selected GC-content cut-offs, shown as vertical blue lines in figure 2*D*–*O* and listed in supplementary table S2, Supplementary Material online, varied from species to species. *Harpophora oryzae* (R5-6-1) (Xu et al. 2014) had the highest GC-content AT-rich region component at 48% and *Monacrosporium haptotylum* (CBS 200.50) (Meerupati et al. 2013) the lowest at 12.3% (fig. 4). This wide range of 35.7% strongly supports the need to define AT-rich and GC-equilibrated regions case-by-case, based on comparisons within each genome, rather than by broadly applying pre-defined GC-content cut-offs.

There was a broad range in the proportion of each genome containing AT-rich regions—from ~1% to AT-rich region components over 50%. AT-rich regions are not present in all genomes with a high repeat content, with known examples of repeat rich genomes showing clearly unimodal GC-content distributions (e.g., *Blumeria* spp. and *Puccinia* spp. genomes fig. 2*A* and *B* and supplementary table S1, Supplementary Material online). A total of 79 genome assemblies from 61 distinct species were identified as being composed of ≥ 5% AT-rich regions (supplementary table S1, Supplementary Material online). Although *L. maculans* is perhaps the best known case of a fungal genome interspersed with AT-rich regions (Gout et al. 2006; Fudal et al. 2009; Van de Wouw et al. 2010; Rouxel et al. 2011; Ohm et al. 2012), we note 14 genomes with higher AT-rich region contents than *L. maculans* and many more with comparable AT-rich region content (supplementary table S2, Supplementary Material online).

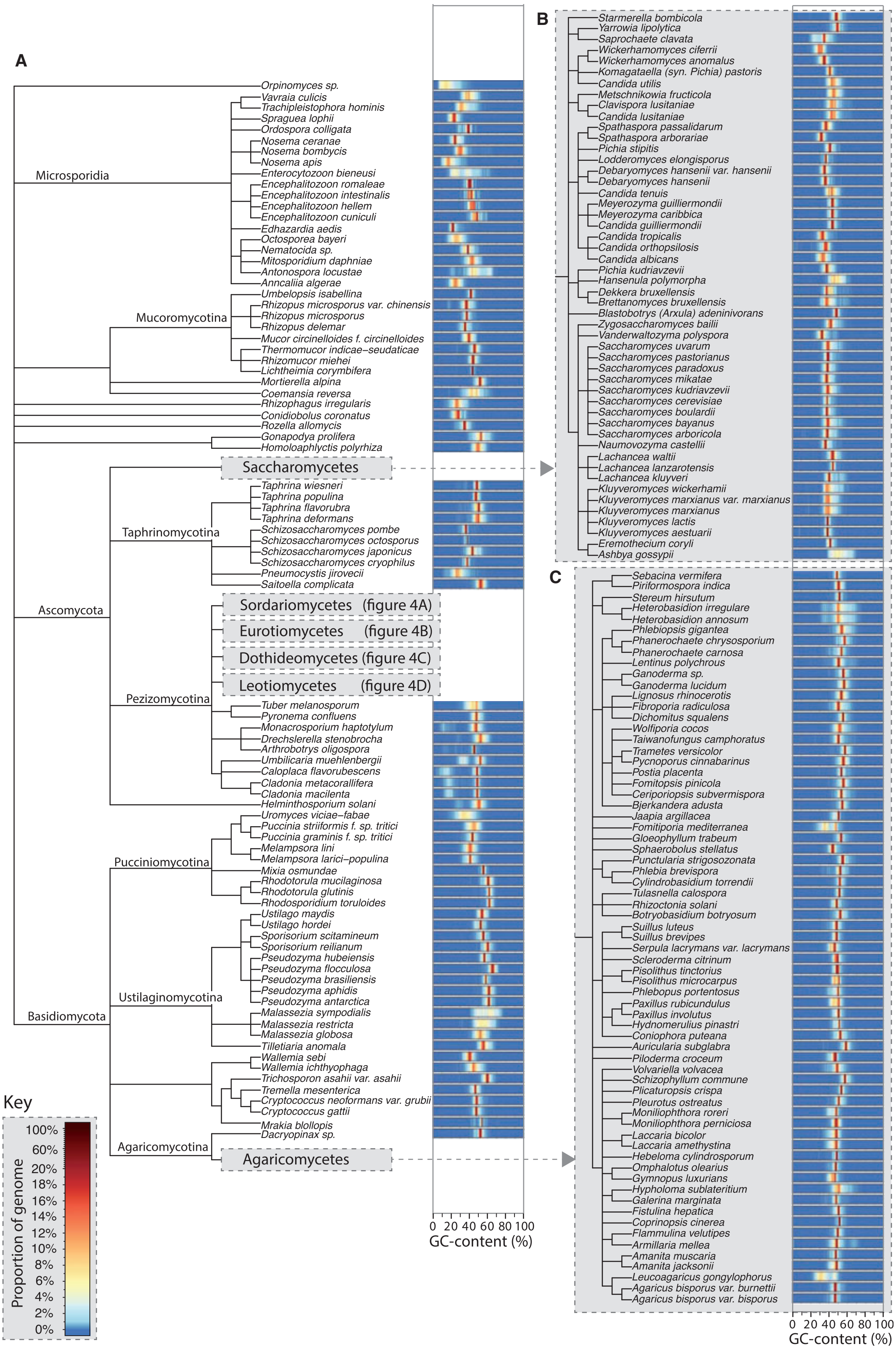


Fig. 3.—(*A*) Clade trees showing the taxonomic relationship between the fungal species included in this study. To the right of each species' name, there is a bar displaying a heat map plot of the GC-content distribution of segments of each genome, where different colors represent varying genome proportions (see key). Classes containing high numbers of sequenced genomes have been displayed separately in panels (*B*) (Saccharomycetes) and (*C*) (Agaricomycetes), and in figure 4 (Sordariomycetes, Eurotiomycetes, Dothideomycetes and Leotiomycetes).

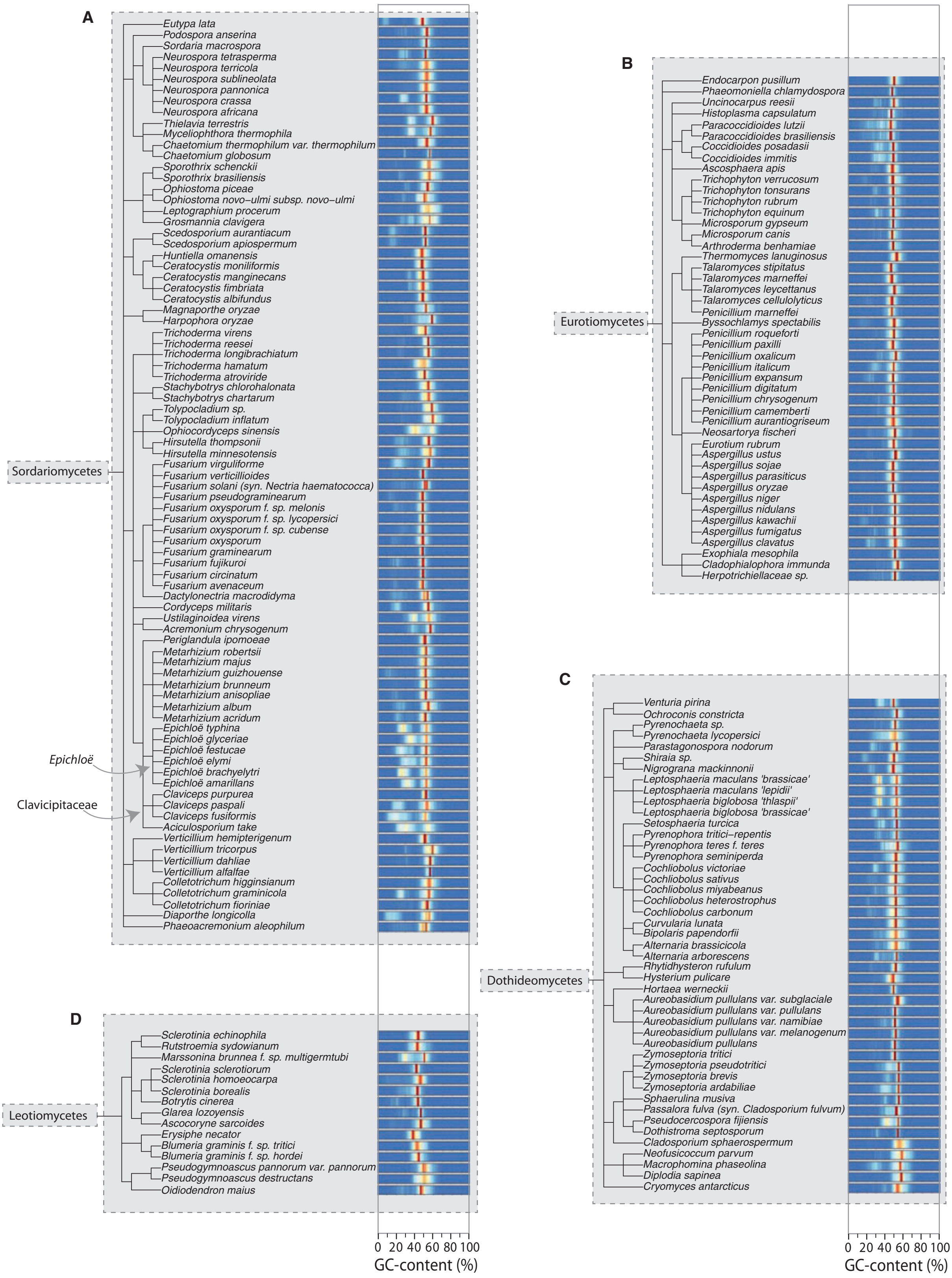


Fig. 4.—Clade trees showing the taxonomic relationship between fungal species belonging the classes Sordariomycetes (*A*), Eurotiomycetes (*B*), Dothideomycetes (*C*) and Leotiomycetes (*D*) within the subphylum Pezizomycotina. These fit into the broader taxonomic tree displayed in figure 3*A*. To the right of each species' name, there is a bar displaying a heat map plot of the GC-content of segments of each genome, where different colors represent varying proportions of the genome with each percentage GC (see fig. 3 key).

In four species, AT-rich regions were in higher proportions than GC-equilibrated regions: *Leucoagaricus gongylophorus* (Ac12) (Aylward et al. 2013) with 76.6%, *Ophiocordyceps sinensis* (CO18) (Hu et al. 2013) with 69.7%, *Aciculosporium take* (MAFF-241224) (Schardl et al. 2013) with 58.5%, and *Fomitiporia mediterranea* (LYAD-421 SS1) (Floudas et al. 2012) with 57.8%. Despite the dominant component of each of these genomes being AT-rich, where gene annotations were available for these species (*L. gongylophorus*, *A. take*, and *F. mediterranea*) the majority of the gene content was still attributed to GC-equilibrated genome regions.

In six of the surveyed 500+ genomes, manual inspection of the GC-content plots of genomes motivated their removal from the set of AT-rich region genomes. The GC-content plots of *Armillaria mellea* (DSM 3731) (Collins et al. 2013), *Nosema bombycis* (CQ1) (Pan et al. 2013), and *Lachancea kluyveri* (NRRL Y-12651) (Cliften et al. 2003) showed evidence of segments of DNA, amounting to small components of the overall genome, with a higher GC-content than the rest of the genome. *Armillaria mellea* and *Candida orthopsilosis* (MCO456) (Pryszcz et al. 2014) contained small percentages of 0–1% GC-content segments that may be an artefact of assembly. *Malassezia sympodialis* (ATCC 42132) (Gioti et al. 2013) (fig. 3*A*, see Ustilaginomycotina for *M. sympodialis* GC-content heat map) appeared to have an unusually broad, albeit unimodal, GC-content distribution. In response to these examples, we have included a feature to allow the manual input of one or more GC-content boundaries for the categorization of genome segments into genome regions. This feature allows unusual and exceptional cases of genome GC-content distribution to be studied.

Taxonomic Distribution of Genomes with AT-Rich Regions

Figure 5*A* shows the percentage of genomes with various AT-rich genome contents for each subphylum surveyed, showing that this genome type is most common within the Pezizomycotina. The majority of genomes with ≥ 5% AT-rich region content are from the subphylum Pezizomycotina with only six exceptions: four Agaricomycetes (*L. gongylophorus, F. mediterranea, Paxillus rubicundulus, Moniliophthora roreri*), one Microsporidia (*Enterocytozoon bieneusi*), and one Saccharomycotina (*Saprochaete clavata*). Heat-map plots of GC-content distributions arranged alongside dendograms generated according to taxonomic classifications (figs. 3 and 4) also show this trend. Additionally, clusters of species with AT-rich regions are clearly visible in figures 3 and 4 (e.g., the family Clavicipitaceae within the class Sordariomycetes).

AT-Rich Regions and Fungal Lifestyle

Classifications of fungal species into broad lifestyle categories (saprobe, pathogen, and symbiont) are displayed next to each surveyed species in supplementary table S1, Supplementary Material online. Pathogens and symbionts are further classified based on whether they have a plant host, and plant pathogens are additionally classified into obligate biotrophs, nonobligate biotrophs, hemibiotrophs, and necrotrophs. The

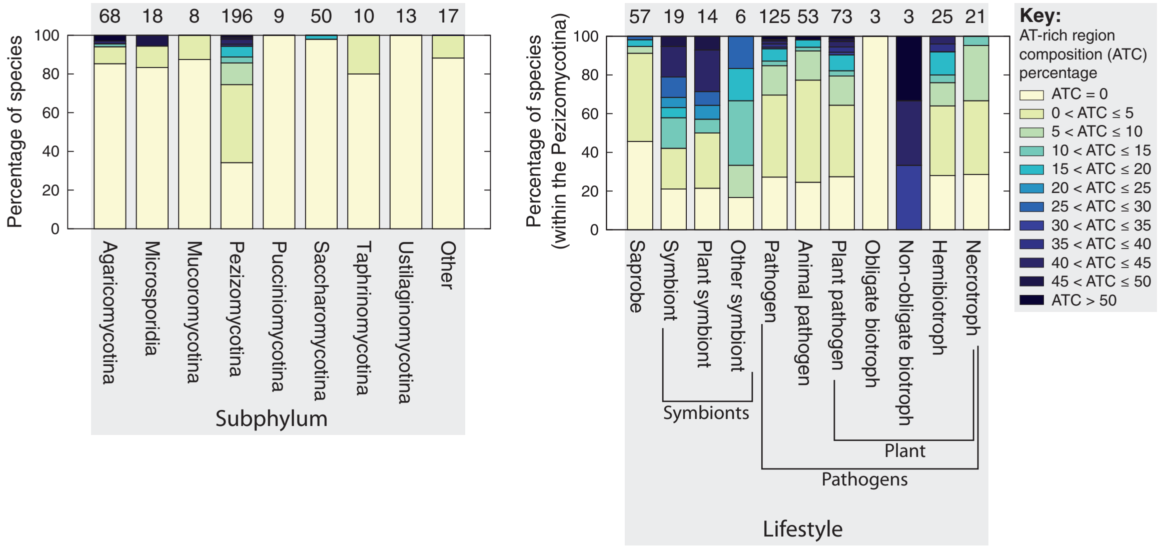


FIG. 5.—A summary of the number of species surveyed for each subphylum is shown (left-hand plot), where within each bar the colours describe the proportion of surveyed genomes within each subphyla with an AT-rich content in a particular range (see key far right). Subphyla with less than five surveyed species have been grouped into the entry "Other". A similar plot (right-hand plot) is shown where species (from within the Pezizomycotina) have been classified according to their lifestyle.

distribution of fungal lifestyles in the different subphylum varied. The percentage of AT-rich genomes with various AT-rich genome contents are displayed according to their lifestyle categories in figure 5*B*. Within 196 surveyed Pezizomycotina species, 125 (~64%) were pathogens, compared with 9 out of 68 (~13%) of the Agaricomycotina species and 10 out of 50 (20%) of the Saccharomycotina species. As discussed in the previous section, AT-rich bimodal genomes were found more commonly in the Pezizomycotina subphylum. To prevent these taxonomic biases from influencing our assessment of AT-rich regions in relation to fungal lifestyle, the results shown in figure 5*B* are restricted to Pezizomycotina species.

Attributes of AT-Rich Regions Compared with GC-Equilibrated Regions

The subset of genomes with a $\geq$ 5% proportion of AT-rich regions was further analyzed (see Materials and Methods) to assess the role of RIP in AT-rich region formation and compare AT-rich regions between species. For each of these species, the frequency of each of the 16 possible dinucleotides was calculated from AT-rich region genome sequences and GC-equilibrated genome sequences (supplementary table S3, Supplementary Material online). Dinucleotide frequencies have commonly been employed in previous studies to ascertain the genomic impact of RIP (Hane and Oliver 2008; Clutterbuck 2011). The percentage differences between dinucleotide frequencies within AT-rich and GC-equilibrated regions are shown in a series of plots (fig. 6), with each different colored bar representing a different species. In all cases, the dinucleotide pairs TpA, ApT, TpT and ApA were higher in AT-rich regions, as expected and reflecting their lower GC-content. There was however a much larger difference in the frequency of the TpA dinucleotide, the primary product of RIP in the Pezizomycotina (Clutterbuck 2011) (fig. 6*A*), when compared with the other low GC dinucleotides ApT, TpT and ApA (fig. 6*Biii*). This is a strong indicator of RIP activity in these species.

While analysis and comparison of the lengths of AT-rich regions can be impeded by relative differences in assembly quality, we were able to observe variation in the size of AT-rich regions accurately in higher quality genomes. *Coccidioides immitis* (RS) (Sharpton and Stajich 2009), *Thielavia terrestris* (NRRL 8126) (Berka et al. 2011), *Myceliophthora thermophile* (ATCC 42464) (Berka et al. 2011), *N. crassa* (OR74A) (Galagan et al. 2003), *Z. tritici* (IPO323) (Goodwin et al. 2011), *Cordyceps militaris* (CM01) (Zheng et al. 2011), and *L. maculans* "brassicae" (v23.1.3) (Rouxel et al. 2011) all have assemblies with <50 scaffolds and we consider these to be of "good quality". Among these, *L. maculans* had the longest average AT-rich region length at 31.7 kb and *Z. tritici* the shortest at 10.7 kb. GC-equilibrated regions were also longer within *L. maculans* than within *Z. tritici*, indicating that the *Z. tritici* genome is interrupted with AT-rich regions more frequently than *L. maculans*, despite having a smaller overall AT-rich genome component. Similarly *N. crassa, M. thermophila* and *C. immitis* have comparable average AT-rich region lengths, with differing overall AT-rich region composition and correspondingly different higher GC-region average lengths.

Gene Content

Mutations accumulated during AT-rich region formation could influence the properties of coding sequences in those regions. All observed AT-rich regions are gene-sparse when compared with GC-equilibrated regions (table 1). Indeed, many of the surveyed genomes contained only a very small number of genes within AT-rich regions, making meaningful comparisons between sets of genes within and outside AT-rich regions difficult. After quality filtering the sets of genes (see Materials and Methods), 19 species were identified with >50 genes within AT-rich regions and thus suitable for further analysis (supplementary table S4, Supplementary Material online). Comparisons between coding sequences within AT-rich and GC-equilibrated regions confirmed differences in amino acid usage, codon usage, and putative function between these sets as well as identifying common trends across the selected species.

Percentage differences in the average amino acid content of coding sequences from genes within AT-rich regions compared with those in GC-equilibrated regions are shown per species in figure 8. Statistically significant differences (Mann–Whitney *U* tests, *P* value $\leq$ 0.05) in amino acid content were found in all species surveyed, with common trends across different species observable (fig. 8). If these differences occurred due to coding sequences from GC-equilibrated regions being subjected to nonsynonymous RIP-like C to T and G to A transitions, we can form expectations about how the resulting coding sequences in AT-rich regions would differ in amino acid composition. As an aid to understanding this, possible nonsynonymous changes (dN) are summarized in figure 7. In addition to illustrating how C to T and G to A transitions can alter one amino acid to another, this network highlights nodes which are absolute start and end points. Glycine, alanine and proline (G, A, and P) may undergo C to T and G to A changes to other amino acids, but can never be created by these mutations. As such, coding sequences subjected to C to T and G to A transitions are expected to have lower levels of these amino acids. Conversely, phenylalanine, tyrosine, lysine, isoleucine and asparagine (F, Y, K, I, and N) could be expected to increase in number. This appears as expected in the plots shown in figure 8. Both nonsynonymous and synonymous mutations within coding sequences can disrupt codon usage. The proportion of each codon encoding each of the amino acids was compared within and outside AT-rich regions within figure 9. In most cases codon usage in coding sequences within AT-rich regions favored codon choices higher in A and T than coding sequences in GC-equilibrated

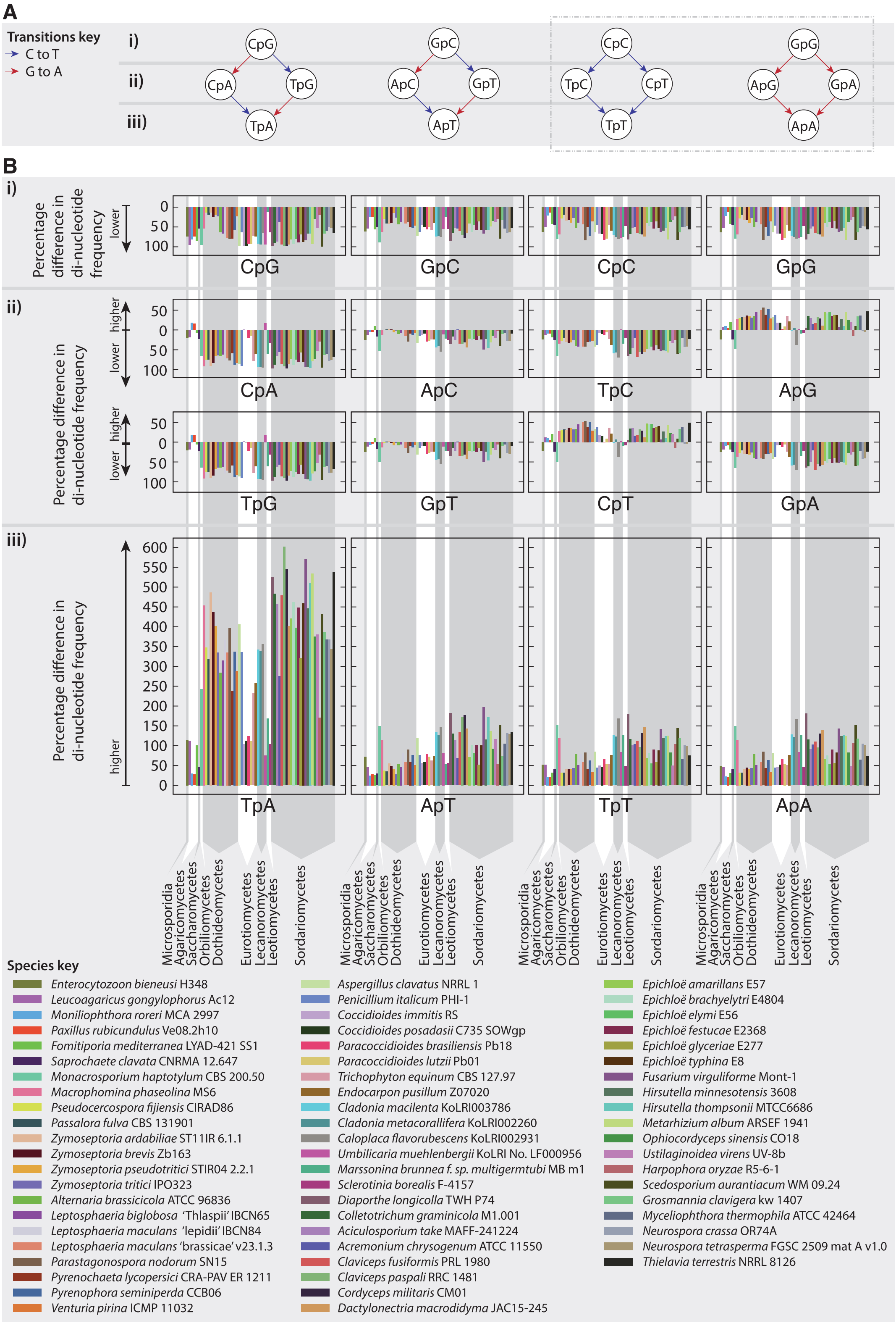


Fig. 6.—Possible changes to dinucleotide pairs are described in a series of graphs (*A*). Dinucleotides are represented by nodes, and arrows represent possible changes resulting from *C* to *T* (blue) and *G* to *A* (red) transitions. Dinucleotides are organized into bands marked (i), (ii), and, (iii) based on the number of mutable sites they have (2, 1 and 0, respectively). Below each graph, corresponding plots (*B*) of percentage differences between dinucleotide frequencies within and outside AT-rich regions are shown for species with 5% or greater AT-rich region content. Percentage differences above *y* = 0 correspond to higher values within AT-rich regions and below *y* = 0 correspond to lower values within AT-rich regions (see axis labels). Each species is represented by a different color as shown in the species key and arranged according to taxonomy. Vertical grey bands group species by class, with class names marked at the bottom of (*B*).

Table 1
Attributes of Coding Sequences from Genome Regions within AT-Rich and GC-Equilibrated Regions

Species	AT-Rich Region Gene Content						GC-Equilibrated Gene Content						Genome Source
	Gene Density (Genes Per Mb)	No. Genes (Total)	No. Genes (Filtered)[a]	Average Length (Amino Acids)	No. Secreted	Per cent Secreted	Gene Density (Genes Per Mb)	No. Genes (Total)	No. Genes (Filtered)[a]	Average Length (Amino Acids)	No. Secreted	Per cent Secreted	
Enterocytozoon bieneusi (H348)	865	1,616	1,549	313.2	40	2.58	1,010	2,016	1,341	184.3	36	2.68	JGI (Akiyoshi et al. 2009)
Leucoagaricus gongylophorus (Ac12)	1.9	148	66	186.8	3	4.54	222	5,272	3,610	421.6	194	5.37	JGI (Aylward et al. 2013)
Moniliophthora roreri (MCA2997)	188	1,473	1,050	248.9	33	3.14	353	15,677	15,128	433.0	1,660	10.97	NCBI (Meinhardt et al. 2014)
Fomitiporia mediterranea (LYAD-421 SS1)	53.2	1,947	1,614	332.1	112	6.93	351	9,386	8,836	464.9	681	7.70	JGI (Floudas et al. 2012)
Pseudocercospora fijensis (CIRAD86)	25.7	1,179	1,039	392.1	70	6.73	421	11,928	10,514	439.2	752	7.15	JGI
Passalora fulva (CBS 131901)	19.8	528	93	152.4	11	11.70	395	13,589	13,380	449.7	1,134	8.47	JGI (de Wit et al. 2012)
Leptosphaeria maculans (v23.1.2)	11.3	186	181	141.1	28	15.46	432	12,283	12,272	422.8	1,042	8.49	JGI (Rouxel et al. 2011)
Coccidioides immitis (RS)	65.2	319	249	140.8	3	1.20	399	9,591	9,537	445.4	537	5.63	JGI (Sharpton and Stajich 2009)
Paracoccidioides lutzii (Pb01)	34.4	202	193	195.5	8	4.14	330	8,934	8,841	448.7	386	4.36	Broad (Desjardins et al. 2011)
Marssonina brunnea f. sp. *Multigermtubi* (Mb m1)	15.4	303	251	220.2	49	19.52	302	9,724	9,634	505.1	922	9.57	JGI (Zhu et al. 2012)
Aciculosporium take (MAFF-241224)	17.9	616	223	89.2	7	3.05	364	8,899	8,213	443.6	532	6.47	NCBI (Schardl et al. 2013)
Claviceps fusiformis (PRL 1980)	93.9	2,338	272	85.6	9	3.20	324	8,955	8,479	450.6	671	7.91	NCBI (Schardl et al. 2013)
Epichloë festucae (E2368)	37.8	400	130	76.3	3	2.23	355	8,573	8,032	477.0	613	7.63	NCBI (Schardl et al. 2013)
Epichloë glyceriae (E277)	37.8	765	388	162.3	15	3.81	408	12,755	10,329	456.0	754	7.29	NCBI (Schardl et al. 2013)
Ustilaginoidea virens (UV-8b)	10.6	144	135	270.9	5	3.70	322	8,282	8,271	467.8	607	7.33	NCBI (Berka et al. 2011)
Myceliophthora thermophila (ATCC42464)	15.9	160	64	169.9	2	3.12	312	8,950	8,613	487.8	722	8.38	JGI (Zhang et al. 2014)
Neurospora crassa (FGSC 73)	41.1	249	56	112.4	3	5.35	341	11,723	11,190	447.6	910	8.13	JGI (Baker et al. 2015)
Neurospora tetrasperma (FGSC 2509)	71.3	254	62	225.3	6	9.67	308	10,938	10,287	466.6	805	7.82	JGI (Ellison et al. 2011)
Neurospora tetrasperma (FGSC 2508)	38.5	160	64	229.0	4	6.25	292	10,220	9,741	484.8	797	8.18	JGI (Ellison et al. 2011)

Note.—The number of genes, average gene length, and number of genes predicted to be secreted is shown for each species. Species listed are those with 50 or more protein sequences from each region type (AT-rich and GC-equilibrated) after quality filters were applied to remove incomplete genes and transposons (see Materials and Methods and supplementary table S4, Supplementary Material online).

[a] Number of genes (filtered) and subsequent values relate post-quality filtering gene sets (see Materials and Methods).

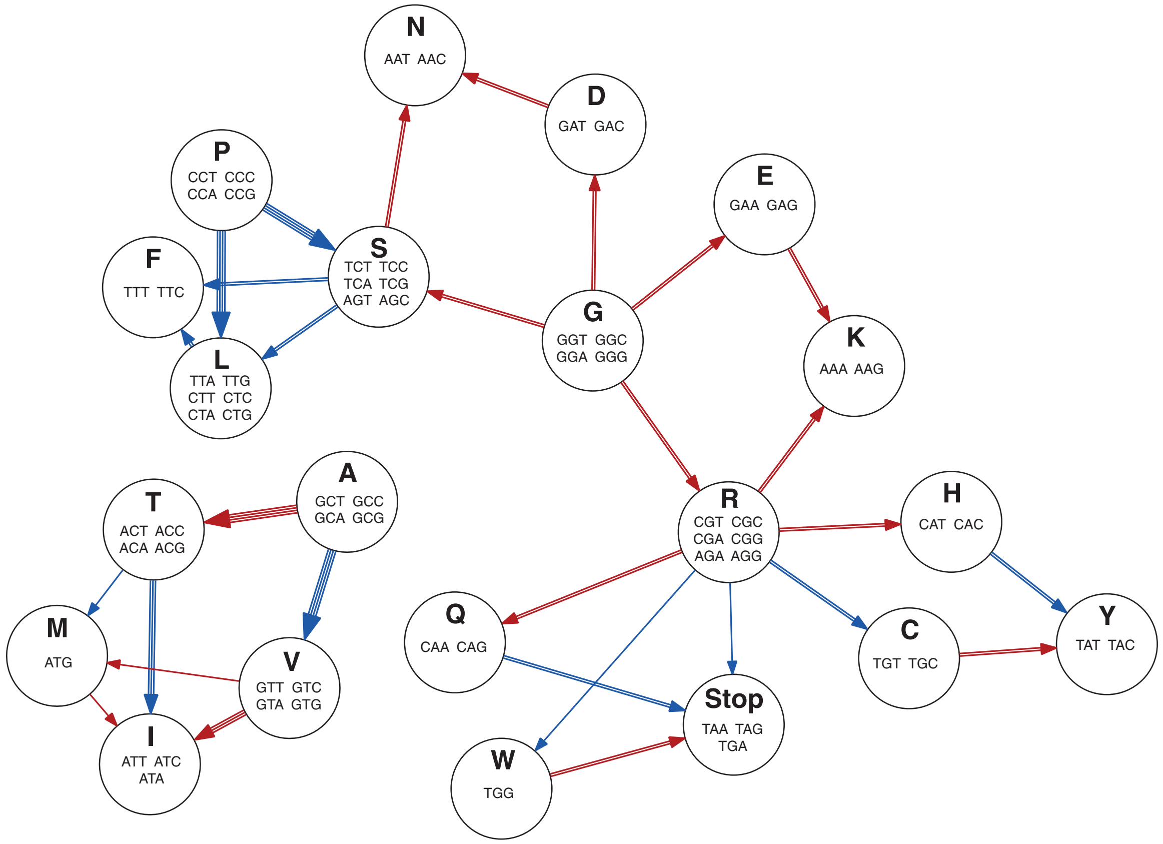


FIG. 7.—Possible amino acid changes that can results from nonsynonymous *C* to *T* (blue) or *G* to *A* (red) transitions. The circular graph nodes represent the amino acids, and the edges represent a mutation resulting in a nonsynonymous change. An edge comprised of multiple lines indicates where multiple different codons of an amino acid can undergo the same amino acid change.

regions. The particularly strong percentage increases in codon usage were seen in codons containing TpA dinucleotides, which have been observed to be a typical product of RIP in noncoding regions.

In all but one case (*E. bieneusi*), genes of the analyzed species within AT-rich regions were of a shorter average length than those outside (table 1). This may be due to mutations causing nonsynonymous changes to certain codons encoding arginine, tryptophan, and glutamine to stop codons (fig. 7), thus shortening the open reading frame (Hane et al. 2015). This possibility is supported by trends of lower levels of these amino acids in proteins within AT-rich regions (fig. 8). Higher levels of secreted proteins in AT-rich regions were previously noted in studies of *L. maculans* (Rouxel et al. 2011), however, we did not find this was a common theme among the species listed in table 1. Only *L. maculans* and *M. brunnea* f. sp. *multigermtubi* were found to have a significantly enriched (by Fisher's exact test, *P* value ≤ 0.05) complement of secreted proteins within AT-rich regions. Indeed, we note several species were significantly depleted in secreted proteins within their AT-rich regions (*C. fusiformis, E. festucae, E. glyceriae, C. immitis*, and *M. roreri*). A summary of enriched and depleted Pfam domains in genes/proteins within AT-rich regions compared with those outside (supplementary table S5, Supplementary Material online) did not highlight any notable enriched Pfam domains for most species. However, most species were enriched in proteins lacking conserved Pfam domains (by Fisher's exact test, *P* ≤ 0.05)

Discussion

Links between RIP, bimodal genomes, transposon activity, and the evolution of fungal genomes have been made previously. Presented here is a facile method for determining the presence of AT-rich genome regions and a systematic survey of published fungal genomes. We show that far from unusual or unique, bimodal genomes are common in the fungal kingdom. Furthermore, for the first time, we have been able to compare how this genome type manifests in a diverse set of species and compare the gene content of AT-rich regions between species.

AT-Rich Region Formation

Our results are consistent with the hypothesis that AT-rich regions are generally formed by transposon invasion followed by mutation by RIP. The higher frequency of species with

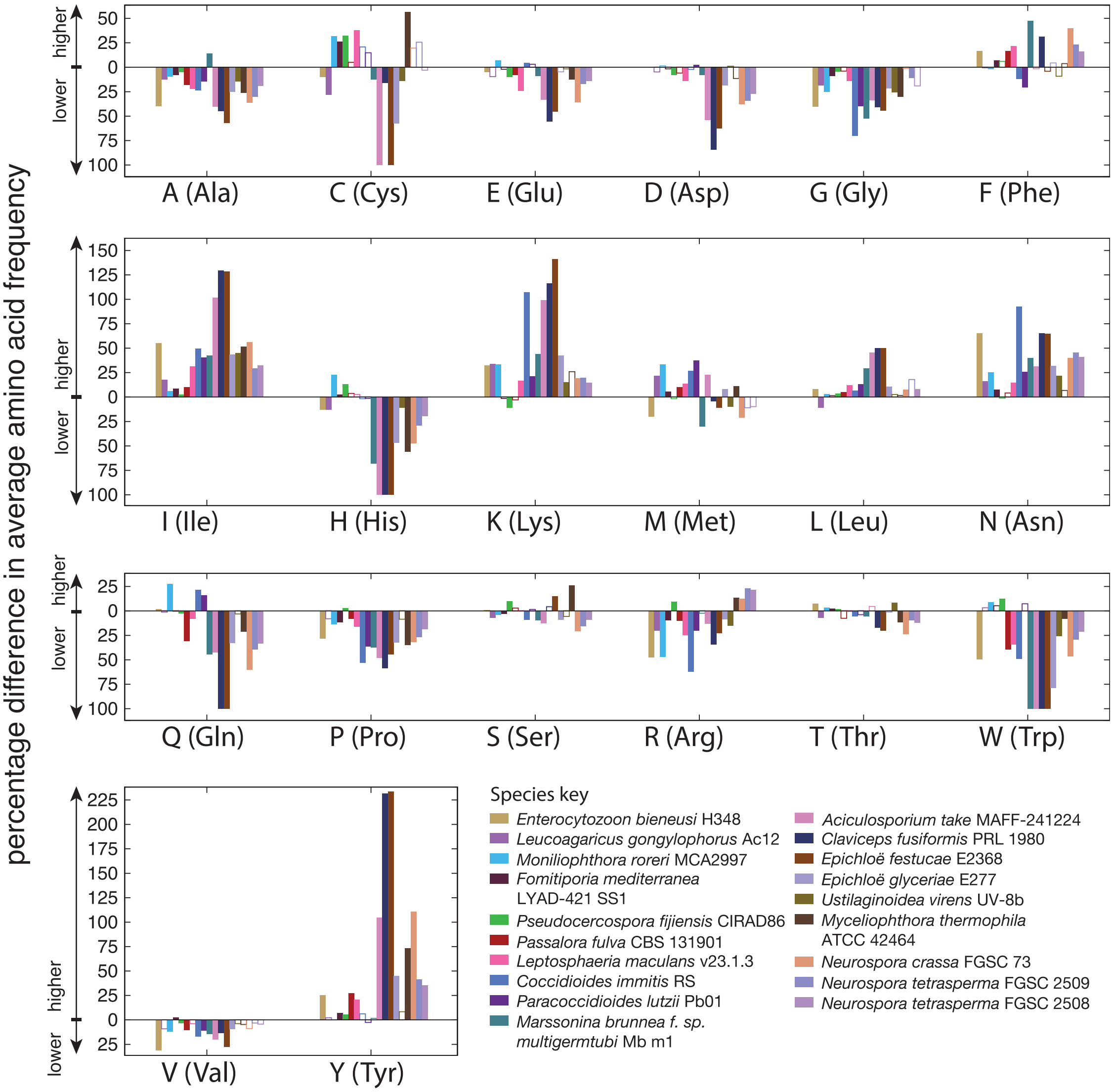


FIG. 8.—Percentage differences between average amino acid content of genes within and outside AT-rich regions are shown. Different colors correspond to different fungal species as per the key (bottom right). Bars that have a solid fill color show where the difference in amino acid frequency is statistically significant (Mann–Whitney *U* test, $P = 0.05$). Percentage differences above $y = 0$ correspond to higher values within AT-rich regions and below $y = 0$ correspond to lower values within AT-rich regions (see axis labels).

AT-rich genome regions within the Pezizomycontina (fig. 5) conforms to the taxonomic distribution of RIP, as previously reported by Clutterbuck (2011) and supports the role of RIP in the formation of these regions. The observed TpA dinucleotide bias in most species with AT-rich regions surveyed provides further evidence of this. Clusters of some closely related species harboring AT-rich regions may indicate that AT-rich regions either developed prior to speciation or that closely related species share a common predisposition for the development of AT-rich regions.

Several species known to be RIP competent were not found to have bimodal genomes. For example, in silico evidence supports RIP activity in *Penicillium roqueforti* (Ropars et al. 2012), and whereas small amounts of AT-rich genome segments are visible in the GC-content distribution, there is not a distinct AT-rich peak in the GC-content distribution. Previously reported in silico evidence also supports RIP-like activity in obligate biotrophs *Melampsora laricis-populina* and *Puccinia graminia* (Horns et al. 2012), from the subphylum Pucciniomycotina, albeit with different dinucleotide biases to those typically seen within the Pezizomycotina. Both these species show unimodal distributions. Therefore, while RIP appears to be necessary for the formation of AT-rich regions, evidence of RIP activity within a genome does not necessarily mean the genome is bimodal. The evidence of RIP can also be due to historical RIP and RIP may no longer be an active process. The

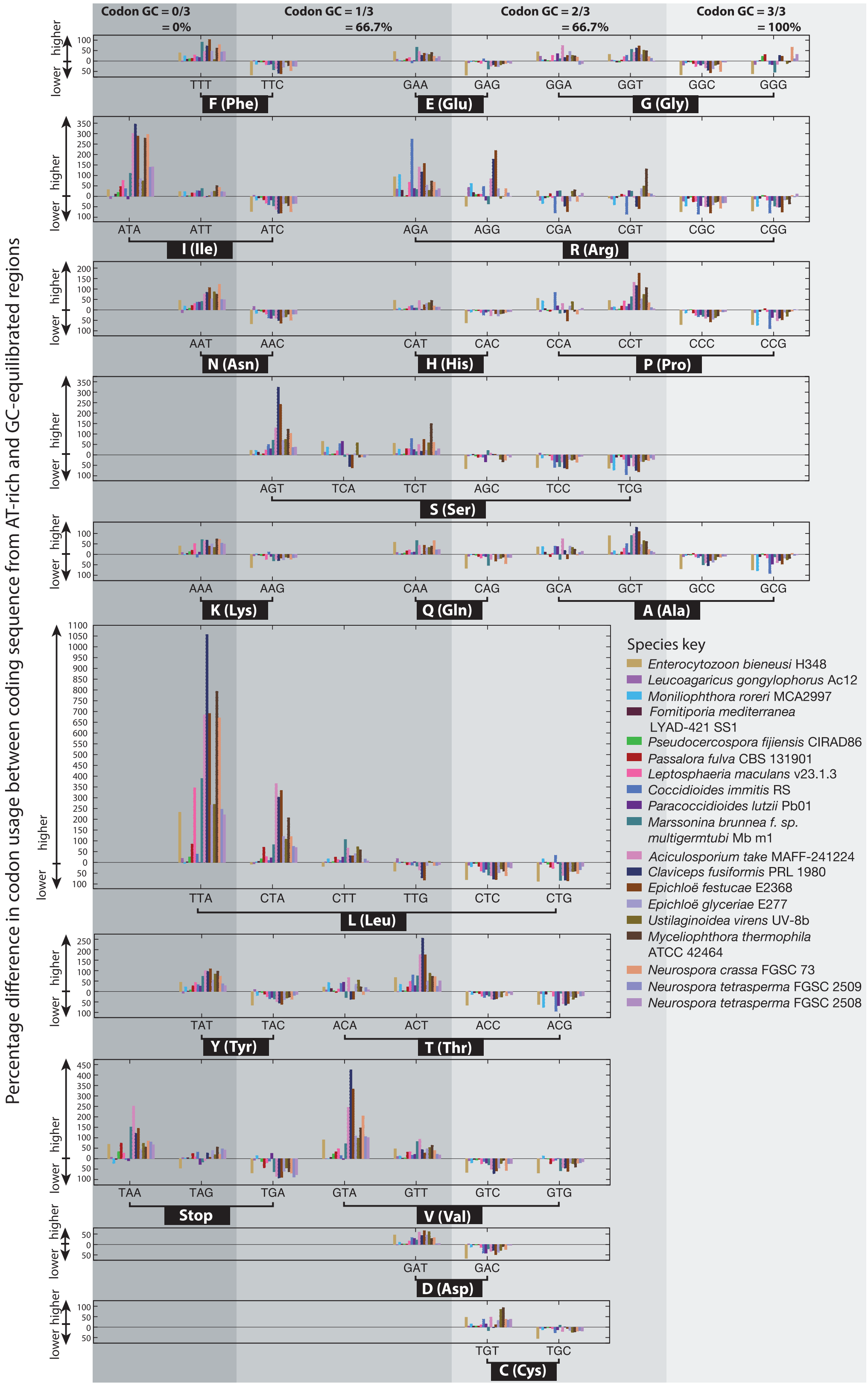


Fig. 9.—Percentage differences between codon usage frequencies in coding sequences within and outside AT-rich regions are shown. Percentage differences above $y = 0$ correspond to higher values within AT-rich regions and below $y = 0$ correspond to lower values within AT-rich regions (see axis labels). Each species is represented by different color as per the species key. Codons are grouped according to the amino acid they encode (see brackets and labels on the *x*-axis). The horizontal position of each plot is based on the GC-content of the codons, with low GC-content codons positioned to the left and higher GC-content codons to the right. Vertical grey bands and labels at the top of the figure show the codon GC-content.

presence of repeats is also required for AT-rich region formation though RIP activity, however without subsequent RIP these sequences are not necessarily distinguishable from the rest of the genome by their GC-content.

A number of additional factors that could affect AT-rich genome formation are not investigated in this survey. RIP occurs around the time of meiosis and as such the sexual cycle of the fungi is relevant to the presence of RIP affected sequences. We note that species thought to be predominantly asexual—*Verticillium dahliae*, *Magnaporthe oryzae*, *Fusarium oxysporum*, and most *Metarhizium* spp. (Taylor et al. 1999; Hu et al. 2014)—contained relatively low proportions of AT-rich regions (supplementary table S1, Supplementary Material online). Time between sexual cycles (and RIP cycles) could relate to how much repeats propagate throughout the genome prior to RIP deactivation. Reproductive frequency (both sexual and asexual) could also influence selection, as the burden of replicating a large genome increases with more frequent reproductive cycles. It is beyond the scope of the current study to gather detailed data on each species, not least because many sequenced species have not yet been studied in detail. Other factors affecting AT-rich region formation that are not investigated in this survey could relate to the efficiency of RIP and the level of activity of other transposon defence mechanisms such as DNA methylation (Miura et al. 2001) and RNA interference (Buchon and Vaury 2006; Cerutti and Casas-mollano 2006; Chung et al. 2008). The genomes we see are those that have survived and as such selection plays a part in the amount of AT-rich sequence we see.

Links to Evolution

Transposon activity contributes to genetic variability and diversification. RIP is primarily considered to be a fungal defence mechanism against the proliferation of repeats. In this survey, we saw a range of AT-rich region genome contents—i.e., the overall proportion of the genome comprised of AT-rich regions—from genomes comprised of ~1% AT-rich regions, to extreme cases where AT-rich regions dominated the genome. Most examples of genomes identified with AT-rich regions were at the lower end of the spectrum where RIP has prevented the spread of repeats and could be thought of as evidence of selection favoring genome stability. At the higher end of the spectrum, we see some genomes have been extensively invaded by repeats prior to their deactivation by RIP, resulting bloated genomes with large components of AT-rich regions. Transposon proliferation and subsequent RIP has the potential to give a burst of diversification, provided both by the transposons and the RIP mutations, followed by a return to relative stability. Within the surveyed Pezizomycotina, bimodal genomes were similarly common within the genomes of saprobes, pathogens, and symbionts. However, in the genomes of symbionts and pathogens there was a shift toward higher components of AT-rich regions (fig. 5*B*).

An important consideration in our interpretation of these data is the bias in the subset of fungi that have been sequenced. Fungal species are not selected for study at random, but rather because of their industrial relevance or experimental tractability. For example, model organisms such as saprobic Pezizomycotina species *N. crassa* and *A. nidulans* were selected for stability. Many plant pathogens are sequenced because of their economic relevance and as such this group is biased toward pathogens that are particularly successful and thus more frequently employing effective evolutionary strategies. Further biases exist where sequencing has been motivated by the desire to conduct comparative genomics research, manifesting in clusters of genome sequences of closely related organisms.

It is not yet possible to know if a truly random sampling of the fungal kingdom would show the same trends seen here, however there are some reasons why high components of AT-rich regions might be more common in symbionts and pathogens. In both the symbionts and the pathogens, it appears to be species with a plant host that are largely responsible for this trend (fig. 5*B*). Both plant symbionts and pathogens need to evade host defences and are therefore under different selection pressures to saprobes. For example, plant pathogens frequently have an evolutionary history of host jumps and overcoming host resistance or control measures. Many of these fungi are cosmopolitan—they travel the world and are exposed to a variety of host resistance genes and fungicides—and only the fit survive. Genome dynamism in these species is key to their survival. This could explain why bursts of diversification that could come about with AT-rich region formation would be a successful evolutionary strategy in plant pathogens and thus a common theme to their genomes.

The GC-content distributions of the eight surveyed obligate biotroph plant pathogen species (*Blumeria graminis* f. sp. *hordei*, *Blumeria graminis* f. sp. *tritici*, *Erysiphe necator*, *M. larici-populina*, *Melampsora lini*, *P. graminis* f. sp. *tritici*, *Puccinia striiformis* f. sp. *tritici* and *Uromyces viciae-fabae*) were clearly unimodal. Considering the taxonomic distribution of RIP and that just three of the nine are Pezizomycotina species (*B. graminis* f. sp. *hordei*, *B. graminis* f. sp. *tritici* and *E. necator*), we cannot say whether unimodal genomes are common theme in obligate biotrophs. Only three nonobligate biotroph species from the Pezizomycotina were surveyed (*A. take, P. fulva*, and *U. virens* with 58.5%, 43.8%, and 34.6% AT-rich region composition, respectively) and as such we cannot reliably compare nonobligate biotrophs with the other plant pathogen groupings. Obligate biotrophs are known to have large, repeat-bloated genomes (Spanu 2012) with the set of assemblies included in this survey ranging in size from a large 82 Mb (*B. graminis* f. sp. *tritici* 96224; Wicker et al. 2013) to an enormous 210 Mb (*U. viciae-fabae* I2; Link et al. 2014). In fact, the genome of *B. graminis* f. sp. *tritici* was estimated to be 120 Mb (Wicker et al. 2013) and *U. viciae-fabae* between 330 and 379 Mb (Link et al. 2014), however due to difficulties assembling repetitive DNA the assemblies are significantly smaller. It has previously been suggested that in the evolution of obligate biotrophs a relaxation of constraints on transposon activity was advantageous as it allowed genetic variability. Notably, the genes necessary for RIP are absent from *Blumeria* spp., despite being common among other Pezizomycotina. Future sequencing projects may shed light on whether this is a common trend in obligate biotrophs and clarify any differing trends in the nonobligate biotrophs.

Although we have larger sets of hemibiotrophic and necrotrophic species within the Pezizomycotina that are more suitable for comparison (25 and 21, respectively), the distinction between these groups is contentious. Many of these pathogens have at one time or another been referred to in the published literature as both necrotrophs and hemibiotrophs. Hemibiotrophs have a longer latent phase than necrotrophs, but the exact length of latent phase that distinguishes these two states is not clearly defined. Within the surveyed Pezizomycotina, higher AT-rich region components were more common among the hemibiotrophic plant pathogens than the necrotrophs (fig. 5*B*, see supplementary table S1, Supplementary Material online, for individual species). Three species out of the 25 surveyed hemibiotrophic Pezizomycotina (*L. maculans* "brassicae", *Marssonina brunnea* f. sp. *multigermtubi*, and *Pseudocercospora fijiensis*) had bimodal genomes with >30% AT-rich region content, whereas the highest AT-rich region content seen among the 21 necrotrophic Pezizomycotina was 11.8% (*Macrophomina phaseolina* MS6; Islam et al. 2012). These differences could relate to the length the latent phase of infection, although it is also possible that additional or unknown characteristics not documented in this survey may explain these differences.

As discussed in our introduction, coding sequences can be affected by RIP mutations either by "leakage" of RIP into neighboring nonrepeat regions or by being themselves duplicated and thus directly targeted by RIP. Our results support this, showing evidence that proteins within AT-rich regions have an amino acid composition consistent with them evolving under these conditions. The differences in amino acid frequency in coding sequences within AT-rich regions could support the idea that AT-rich regions contribute to the evolution of proteins with novel sequence and function. On the other hand, some nonsynonymous changes may be tolerated if they occur without affecting protein function. This may occur where amino acids are altered to those with similar properties.

Bioinformatic Challenges and Consequences

We note that several of the genomes with the highest AT-rich component also have high numbers of scaffolds (supplementary table S2, Supplementary Material online). This is likely due to the difficulty in assembling repetitive sequences, possibly compounded by amplification biases toward GC-equilibrated regions when using some sequencing platforms (Elhaik, Graur, Josic 2010; Ross et al. 2013). We placed no restrictions on genome assembly quality when selecting genomes for this survey, and it is likely that some low coverage short read assemblies have failed to capture AT-rich regions. We refer to the reader to the recent paper by Thomma et al. (2015) which highlights the benefits of generating more complete genomes and closing gaps in assemblies. As genomic resources are improved it will be interesting to see how our understanding of repetitive genome regions and AT-rich regions improves.

As demonstrated, AT-rich regions can have a distinctly lower GC-content than other genome regions and have different dinucleotide compositional biases. Results showing different amino acid usage and codon biases in AT-rich regions demonstrate that coding sequences in these regions also differ from the rest of the genome. These factors all affect gene prediction, which in many cases relies on patterns in short (typically 5 or 6 nucleotide) sequences of DNA. Ab initio gene prediction training is typically carried out on a training set of genes, which may result in parameters that are a poor match to sequences in AT-rich regions. Existing fungal specific approaches to gene prediction (Reid et al. 2014; Testa, Hane, Ellwood, et al. 2015) may offer a platform for addressing these issues.

Leveraging AT-Rich Region Annotation to Advance Fungal Bioinformatics and Crop Protection

This study has revealed the true extent of AT-rich regions across the fungal kingdom, with special emphasis on the Pezizomycotina. Theoretical analysis and analysis of the gene content of AT-rich regions has highlighted a tell-tale amino acid bias in RIP affected proteins. This phenylalanine, tyrosine, lysine, isoleucine and asparagine enriched and glycine, alanine and proline depleted (FYKIN-enriched, GAP-depleted) signature could be used to screen for effectors. Searching for FYKIN-enriched GAP-depleted proteins has the potential to locate candidate effectors that have evolved under RIP conditions but no longer have an AT-rich genomic context.

Knowledge of the locations of AT-rich regions within fungal genomes also allows additional metadata to be associated with gene loci, describing their genomic context and potential to accumulate RIP mutations through leakage. There are only a handful established examples of pathogenicity-related avirulence genes in fungi that conform to this pattern, therefore this method may yet highlight many new candidate pathogenicity genes. To this end, we present the software OcculterCut (available from https://sourceforge.net/projects/occultercut, last accessed April 30, 2016) that is capable of replicating the analyses presented herein as well as reporting gene annotations within and near AT-rich regions. Finally, current gene prediction methods typically rely on training based on the overall gene set or existing homologs, both of which are ill suited to accurate prediction of highly unique gene sets residing in AT-rich regions and/or with specialized roles in plant pathogen interactions. We anticipate that this new knowledge will open up avenues for the prediction of nonstandard genes, particularly fungal effector-like genes, which will be the subject of our continued investigations.

Supplementary Material

Supplementary tables S1–S7 are available at *Genome Biology and Evolution* online (http://www.gbe.oxfordjournals.org/).

Acknowledgments

An Australian Postgraduate Award (APA), funded by the Australian Federal Government, and a Grains Research Scholarship, funded by the Grains Research & Development Corporation (GRS10564), supported the studentship of A.C.T. Many thanks to Megan Meates for giving the article a final read through and polish. This work was funded by the Grains Research & Development Corporation.

Literature Cited

Akiyoshi DE, et al. 2009. Genomic survey of the non-cultivatable opportunistic human pathogen, *Enterocytozoon bieneusi*. PLoS Pathog. 5:e1000261.

Altschul SF, Gish W, Miller W, Lipmann DJ. 1990. Basic local alignment search tool. J Mol Biol. 215:403–410.

Aylward FO, et al. 2013. *Leucoagaricus gongylophorus* produces diverse enzymes for the degradation of recalcitrant plant polymers in leaf-cutter ant fungus gardens. Appl Environ Microbiol. 79:3770–3778.

Baker SE, et al. 2015. Draft genome sequence of *Neurospora crassa* strain FGSC'73. Genome Announcements 3:2012–2013.

Berka RM, et al. 2011. Comparative genomic analysis of the thermophilic biomass-degrading fungi *Myceliophthora thermophila* and *Thielavia terrestris*. Nat Biotechnol. 29:922–927.

Bernaola-Galván P, Román-Roldán R, Oliver J. 1996. Compositional segmentation and long-range fractal correlations in DNA sequences. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Top. 53:5181–5189.

Bernardi G. 2000. Isochores and the evolutionary genomics of vertebrates. Gene 241:3–17.

Bernardi G, et al. 1985. The mosaic genome of warm-blooded vertebrates. Science 228:953–958.

Buchon N, Vaury C. 2006. RNAi: a defensive RNA-silencing against viruses and transposable elements. Heredity 96:195–202.

Camacho C, et al. 2009. BLAST+: architecture and applications. BMC Bioinform. 10:421.

Cambareri EB, Singer MJ, Selker EU. 1991. Recurrence of repeat-induced point mutation. Genetics April(127):699–710.

Cerutti H, Casas-mollano JA. 2006. On the origin and functions of RNA-mediated silencing: from protists to man. Curr Genet. 50:81–99.

Chung W-J, Okamura K, Martin R, Lai EC. 2008. Endogenous RNA interference provides a somatic defense against *Drosophila* transposons. Curr Biol. 18:795–802.

Cissé OH, Pagni M, Hauser PM. 2013. De novo assembly of the pneumocystis jirovecii genome from a single bronchoalveolar lavage fluid specimen from a patient. mBio 4(1):1–4.

Cliften P, et al. 2003. Finding functional features in Saccharomyces genomes by phylogenetic footprinting. Science 301:71–77.

Clutterbuck AJ. 2011. Genomic evidence of repeat-induced point mutation (RIP) in filamentous ascomycetes. Fungal Genet Biol. 48:306–326.

Coleman JJ, et al. 2009. The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. PLoS Genet. 5:e1000618.

Collins C, et al. 2013. Genomic and proteomic dissection of the ubiquitous plant pathogen, *Armillaria mellea*: toward a new infection model system. J. Proteome Res. 12:2552–2570.

Costantini M, Alvarez-Valin F, Costantini S, Cammarano R, Bernardi G. 2013. Compositional patterns in the genomes of unicellular eukaryotes. BMC Genomics 14:755.

Croll D, McDonald BA. 2012. The accessory genome as a cradle for adaptive evolution in pathogens. PLoS Pathog. 8:e1002608.

Croll D, Zala M, McDonald BA. 2013. Breakage-fusion-bridge cycles and large insertions contribute to the rapid evolution of accessory chromosomes in a fungal pathogen. PLoS Genet. 9:e1003567.

Cuomo CA, et al. 2007. The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. Science 317:1400–1403.

de Wit PJGM, et al. 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 Genet. 8:e1003088.

Dean RA, Talbot NJ, Ebbole DJ. 2005. The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature 434:980–986.

Desjardins CA, et al. 2011. Comparative genomic analysis of human fungal pathogens causing paracoccidioidomycosis. PLoS Genet. 7:e1002345.

Dhillon B, et al. 2015. Horizontal gene transfer and gene dosage drives adaptation to wood colonization in a tree pathogen. Proc Natl Acad Sci U S A. 112:3451–3456.

Eddy SR. 2011. Accelerated profile HMM searches. PLoS Comput Biol. 7:e1002195.

Elhaik E, Graur D, Josic K. 2010. Comparative testing of DNA segmentation algorithms using benchmark simulations. Mol Biol Evol. 27:1015–1024.

Elhaik E, Graur D, Josić K, Landan G. 2010. Identifying compositionally homogeneous and non-homogeneous domains within the human genome using a novel segmentation algorithm. Nucleic Acids Research 38:e158.

Ellison CE, et al. 2011. Massive changes in genome architecture accompany the transition to self-fertility in the filamentous fungus *Neurospora tetrasperma*. Genetics 189:55–69.

Farman ML. 2007. Telomeres in the rice BLAST fungus *Magnaporthe oryzae*: the world of the end as we know it. FEMS Microbiol Lett. 273:125–132.

Finn RD, et al. 2014. Pfam: the protein families database. Nucleic Acids Res. 42:D222–D230.

Floudas D, et al. 2012. The Paleozoic origin of enzymatic lignin decomposition reconstructed from 31 fungal genomes. Science 336:1715–1719.

Freitag M, Williams RL, Kothe GO, Selker EU. 2002. A cytosine methyltransferase homologue is essential for repeat-induced point mutation in *Neurospora crassa*. PNAS 99:8802–8807.

Fudal I, et al. 2009. Repeat-induced point mutation (RIP) as an alternative mechanism of evolution toward virulence in *Leptosphaeria maculans*. Mol Plant Pathol. 22:932–941.

Galagan JE, et al. 2003. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422:859–868.

Galagan JE, et al. 2005. Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. Nature 438:1105–1115.

Gardiner DM, et al. 2012. Comparative pathogenomics reveals horizontally acquired novel virulence genes in fungi infecting cereal hosts. PLoS Pathog. 8:e1002952.

Gioti A, et al. 2013. Genomic insights into the atopic eczema-associated skin commensal yeast *Malassezia sympodialis*. mBio. 4:1–16.

Goffeau A, et al. 1996. Life with 6000 genes. Science 274:546–567.

Goodwin SB, et al. 2011. Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genet. 7:e1002070.

Gout L, et al. 2006. Lost in the middle of nowhere: the AvrLm1 avirulence gene of the Dothideomycete *Leptosphaeria maculans*. Mol Microbiol. 60:67–80.

Graïa F, et al. 2001. Genome quality control: RIP (repeat-induced point mutation) comes to *Podospora*. Mol Microbiol. 40:586–595.

Hane JK, Oliver RP. 2008. RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. BMC Bioinform. 9:478.

Hane JK, Oliver RP. 2010. In silico reversal of repeat-induced point mutation (RIP) identifies the origins of repeat families and uncovers obscured duplicated genes. BMC Genomics 11:655.

Hane JK, et al. 2011. A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. Genome Biol. 12:R45.

Hane JK, Williams AH, Taranto AP, Solomon PS, Oliver RP. 2015. Genetic transformation systems in fungi. Fungal Biol. 2:55–68.

Heitman J, Sun S, James TY. 2013. Evolution of fungal sexual reproduction. Mycologia 105:1–27.

Horns F, Petit E, Yockteng R, Hood ME. 2012. Patterns of repeat-induced point mutation in transposable elements of basidiomycete fungi. Genome Biol Evol. 4:240–247.

Hu X, et al. 2013. Genome survey uncovers the secrets of sex and lifestyle in caterpillar fungus. Chin Sci Bull. 58:2846–2854.

Hu X, et al. 2014. Trajectory and genomic determinants of fungal-pathogen speciation and host adaptation. Proc Natl Acad Sci U S A. 111:16796–16801.

Idnurm A, Howlett BJ. 2003. Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete *Leptosphaeria maculans*. Fungal Genet Biol. 39:31–37.

Ikeda K-i, et al. 2002. Repeat-induced point mutation (RIP) in *Magnaporthe grisea*: implications for its sexual cycle in the natural field context. Mol Microbiol. 45:1355–1364.

Irelan JT, Hagemann AT, Selker EU. 1994. High frequency repeat-induced point mutation (RIP) is not associated with efficient recombination. Genetics 138:1093–1103.

Islam MS, et al. 2012. Tools to kill: genome of one of the most destructive plant pathogenic fungi *Macrophomina phaseolina*. BMC Genomics 13:493.

Jurka J, et al. 2005. Repbase update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 110:462–467.

Link T, Seibel C, Voegele RT. 2014. Early insights into the genome sequence of *Uromyces fabae*. Front Plant Sci. 5:587.

Lo Presti L, et al. 2015. Fungal effectors and plant susceptibility. Annu Rev Plant Biol. 66:513–545.

Martinez D, et al. 2004. Genome sequence of the lignocellulose degrading fungus *Phanerochaete chrysosporium* strain RP78. Nat Biotechnol. 22(6):695–700.

Meerupati T, et al. 2013. Genomic mechanisms accounting for the adaptation to parasitism in nematode-trapping fungi. PLoS Genet. 9:e1003909.

Meinhardt LW, et al. 2014. Genome and secretome analysis of the hemibiotrophic fungal pathogen, *Moniliophthora roreri*, which causes frosty pod rot disease of cacao: mechanisms of the biotrophic and necrotrophic phases. BMC Genomics 15:164.

Miura A, et al. 2001. Mobilization of transposons by a mutation abolishing full DNA methylation in *Arabidopsis*. Nature 411:212–214.

Nakayashiki H, Nishimoto N, Ikeda K., Tosa Y, Mayama S. 1999. Degenerate MAGGY elements in a subgroup of *Pyricularia grisea*: a possible example of successful capture of a genetic invader by a fungal genome. Mol Gen Genet. 261:958–966.

NCBI Resource Coordinators 2014. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 43:6–17.

Ohm RA, et al. 2012. Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. PLoS Pathog. 8:e1003037.

Oliver JL, Bernaola-Galván P, Carpena P, Román-Roldán R. 2001. Isochore chromosome maps of eukaryotic genomes. Gene 276:47–56.

Oliver JL, Carpena P, Hackenberg M, Bernaola-Galván P. 2004. IsoFinder: computational prediction of isochores in genome sequences. Nucleic Acids Res. 32:W287–W292.

Oliver R. 2012. Genomic tillage and the harvest of fungal phytopathogens. New Phytol. 196:1015–1023.

Pan G, et al. 2013. Comparative genomics of parasitic silkworm microsporidia reveal an association between genome expansion and host adaptation. BMC Genomics 14:186.

Petersen TN, Brunak S, von Heijne G, Nielsen H. 2011. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 8(10):785–786.

Pryszcz LP, Németh T, Gácser A, Gabaldén T. 2014. Genome comparison of Candida orthopsilosis clinical strains reveals the existence of hybrids between two distinct subspecies. Genome Biol Evol. 6:1069–1078.

Quinlan AR, Hall IM. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics 26:841–842.

Raffaele S, Kamoun S. 2012. Genome evolution in filamentous plant pathogens: why bigger can be better. Nat Rev Microbiol. 10:417–430.

Raffaele S, et al. 2010. Genome evolution following host jumps in the irish potato famine pathogen lineage. Science 330:1540–1544.

Reid I, et al. 2014. SnowyOwl: accurate prediction of fungal genes by using RNA-Seq and homology information to select among ab initio models. BMC Bioinform. 15:229.

Ropars J, et al. 2012. Sex in cheese: evidence for sexuality in the fungus *Penicillium roqueforti*. PloS One 7:e49665.

Ross MG, et al. 2013. Characterizing and measuring bias in sequence data. Genome Biol. 14:R51.

Rouxel T, et al. 2011. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by repeat-induced point mutations. Nat Commun. 2:202.

Schardl CL, et al. 2013. Plant-symbiotic fungi as chemical engineers: multi-genome analysis of the clavicipitaceae reveals dynamics of alkaloid loci. PLoS Genet. 9:e1003323.

Selker EU, Cambareri EB, Jensen BC, Haack KR. 1987. Rearrangement of duplicated DNA in specialized cells of neurospora. Cell 51: 741–752.

Sharpton TJ, et al. 2009. Comparative genomic analyses of the human fungal pathogens Coccidioides and their relatives. Genome Res. 19:1722–1731.

Spanu PD. 2012. The genomics of obligate (and nonobligate) biotrophs. Annu Rev Phytopathol. 50:91–109.

Storck R. 1965. Nucleotide composition of nucleic acids of fungi. J Bacteriol. 91(1):227–230.

Sun B-F, et al. 2013. Multiple interkingdom horizontal gene transfers in pyrenophora and closely related species and their contributions to phytopathogenic lifestyles. PloS One 8:e60029.

Taylor JW, Jacobson DJ, Fisher MC. 1999. The evolution of asexual fungi: reproduction, speciation and classification. Annu Rev Phytopathol. 37:197–246.

Testa A, Oliver R, Hane J. 2015. Overview of genomic and bioinformatic resources for *Zymoseptoria tritici*. Fungal Genet Biol. 79:13–16.

Testa AC, Hane JK, Ellwood SR, Oliver, Richard P. 2015. CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts. BMC Genomics 16:170.

Thomma BPHJ, et al. 2015. Mind the gap; seven reasons to close fragmented genome assemblies. Fungal Genet Biol. 90:24–30.

Tzeng TH, Lyngholm LK, Ford CF, Bronson CR. 1992. A restriction fragment length polymorphism map and electrophoretic karyotype of the fungal maize pathogen *Cochliobolus heterostrophus*. Genetics. 130(1):81–96.

Van de Wouw AP, et al. 2010. Evolution of linked avirulence effectors in *Leptosphaeria maculans* is affected by genomic environment and exposure to resistance genes in host plants. PLoS Pathog. 6:e1001180.

Watters MK, Randall, Thomas A, Margolin BS, Selker EU, Stadier DR. 1999. Action of repeat-induced point mutation on both strands of a duplex and on tandem duplications of various sizes in neurospora. Genetics 153(2):705–714.

Wellcome Trust Sanger Institute. 2015. GFF (General Feature Format) specifications document. [cited 2016 Apr 30]. Available from: http://www.sanger.ac.uk/resources/software/gff/.

Wicker T, et al. 2013. The wheat powdery mildew genome shows the unique evolution of an obligate biotroph. Nat Genet. 45:1092–1096.

Xu X-H, et al. 2014. The rice endophyte *Harpophora oryzae* genome reveals evolution from a pathogen to a mutualistic endophyte. Sci Rep. 4:5783.

Zhang Y, et al. 2014. Specific adaptation of *Ustilaginoidea virens* in occupying host florets revealed by comparative and functional genomics. Nat Commun. 5:3849.

Zheng P, et al. 2011. Genome sequence of the insect pathogenic fungus *Cordyceps militaris*, a valued traditional Chinese medicine. Genome Biol. 12:R116.

Zhu S, et al. 2012. Sequencing the genome of *Marssonina brunnea* reveals fungus-poplar co-evolution. BMC Genomics 13:382.

Zolan ME. 1995. Chromosome-length polymorphism in fungi. Microbiol Rev. 59:686–698.

Associate editor: Davide Pisani

Chapter 7 *Pyrenophora teres* f. *teres* and *Pyrenophora teres* f. *maculata* genome assemblies and effector prediction

7.1 Attribution statement

Title:	*Pyrenophora teres* f. *teres* and *Pyrenophora teres* f. *maculata* genomics and effector prediction
Authors:	Alison C Testa, Simon R Ellwood, Julie Lawrence, and Richard P Oliver

This thesis chapter was the result of collaborative laboratory based work and bioinformatics analysis. As such, not all work contained within this chapter can be attributed to the Ph. D. candidate.

The Ph. D. candidate (ACT) made the following contributions to this chapter:

- ACT conducted the bioinformatics analysis
- ACT wrote the chapter, excepting the sections describing DNA and RNA extractions
- ACT did not conduct any of the wet laboratory work

The following contributions were made by co-authors:

- SRL wrote the DNA and RNA extraction descriptions
- DNA and RNA extractions and all other lab work was carried out by SRE and JL
- RPO, SRE, and JKH reviewed the manuscript

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

________________________________ ________________
Alison C Testa Date

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

________________________________ ________________
Richard P Oliver Date
(Principal supervisor and co-author)

________________________________ ________________
Simon R Ellwood Date
(Principal supervisor and co-author)

7.2 Abstract

Pyrenophora teres f. *teres* (*Ptt*) and *Pyrenophora teres* f. *maculata (Ptm)* are fungal pathogens causing net-form of net-blotch and spot-form of net-blotch on susceptible barley cultivars. In this study, the genome of each of the pathogens was sequenced using PacBio single-molecule real-time (SMRT) sequencing technology, resulting in 56.7 Mbp and 42.0 Mbp assemblies for *Ptt* and *Ptm* respectively. RNA-seq driven gene prediction methods were used to predict the gene content of each species, with particular attention given to the inclusion of small, secreted proteins of interest to downstream effector candidate prediction work. Finally, the genomic context and protein sequences of each of the species was analysed and candidate effectors selected. These high-quality genomic resources and the assessment of candidate effectors will form a strong basis for continued research into these pathogens.

7.3 Background

Barley is Australia's second largest crop (2) as well as being an important crop worldwide, and is used for animal feed, in malting, and grain for human consumption. In Australia, barley pathogens are estimated to cause average annual yield losses of 252 million dollars or 19.6% of the average annual barley crop value (2). Among the most important pathogens of barley are the fungal pathogen species *Pyrenophora teres* f. *teres* (*Ptt*) and *Pyrenophora teres* f. *maculata* (*Ptm*).

Ptt and *Ptm* cause net-form of net-blotch and spot-form of net-blotch respectively. These species have generally been described as necrotrophs in the literature (14,15,46), however each has a latent phase of infection and could be considered a hemibiotroph. Disease symptoms of the net-form appear as necrotic lines running parallel along the leaf veins, with perpendicular necrosis occasionally joining these longer lines to form a net-like pattern. In contrast, spot form appears as oval or circular necrotic regions surrounded by yellowing chlorotic leaf tissues, similar to the lesions caused by *Pyrenophora tritici-repentis (Ptr)* on wheat (13). *Ptt* and *Ptm* are classed as 'forms', but may be considered as distinct species, with hybridization between the two species noted as either unusual or absent under field conditions (223). Phylogenetic analysis of orthologous intergenic regions in *Ptt*, *Ptm*, and *Ptr* estimated the two barley infecting species diverged from the wheat infecting *Ptr* about 8.04 M years ago, while *Ptm* and *Ptt* diverged from each much more recently at around 519 k years ago (14).

The first Pyrenophora species to be sequenced was *Ptr*, using Sanger technology with the assembly assisted by optical mapping, resulting in a compact 37.8 Mbp assembly contained within 47 scaffolds (5,12). The first genome assembly of *Ptt* was published in 2010, assembled from short paired-end read data from the Solexa platform (15). The resulting assembly was 41.95 Mb and allowed the identification of 11,799 putative genes. Analysis of the gene content found non-ribosomal peptide synthetases and efflux pumps to have undergone *Ptt* specific gene expansion. A more recent addition to the set of sequenced *Pyrenophora* spp. has been the seed pathogen *Pyrenophora semeniperda* (224). To date a genome sequence of *Ptm* has not been published.

Evidence has shown that the barley-*Ptt* pathosystem operates, at least in part, as necrotrophic effector triggered susceptibility (46), similar to that seen in the wheat *Ptr* and wheat-*Parastagonospora nodorum* interaction (87,152). There is also evidence of dominant resistance, and these species may therefore also harbour avirulence effectors (13). Despite this, no effectors have as yet been cloned from *Ptt* or *Ptm*. Bioinformatic analysis of genomic resources can assist in identifying putative effectors, which can then be further investigated by heterologous expression and infiltration into cultivars of the host plant. The bioinformatic components to this approach typically involve assembling a pathogen genome sequence, predicting the gene content, and selecting a set of candidate effector genes. Effector candidates are typically chosen based on properties observed from the set of known fungal effectors, such as their size, cysteine content, genomic context, expression profile, and evidence of positive selection (86–88). Recently, EffectorP, a machine learning fungal effector predictor, has been developed to assess the protein sequences of secreted proteins and predict putative effectors (89). The selection of effector candidates relies on the genome assembly and gene predictions, and as such establishing high-quality genomic resources for a pathogen species is an important preliminary step in predicting the effector gene content.

PacBio single-molecule real-time (SMRT) sequencing technology has recently emerged as a low cost sequencing technology that achieves long DNA reads. These long reads are capable of spanning repetitive regions that cannot be reconstructed when assembling short read data alone. Furthermore, this technology does not have a GC-content bias, and has been shown to achieve consistent coverage of AT-rich sequence (112). This is particularly relevant to Dothideomycete genome where AT-rich genome regions are common (Chapter 6) due to the activity of repeat-induced point mutation (RIP). RIP is a mechanism by which repeat regions are peppered with C to T transitions on both strands, resulting in genome sequence that is AT-rich (64,65). The potential for PacBio reads to be sufficiently long to traverse repetitive regions while maintaining consistent coverage over AT-rich RIP degraded repeats makes this technology an attractive choice for improving assembly quality. The frequent close proximity of known effectors to repetitive and in some cases AT-rich genome regions means reducing assembly gaps and fragmentation is of particular importance to downstream effector gene prediction (113).

After genome assembly, gene prediction is the first step in assessing the gene content of an organism and forms the basis for downstream analysis, including the prediction of effectors. RNA-seq has become a popular as a low cost sequencing technology that can be used to improve gene predictions (10,119,121,225), with alignments of reads and/or transcripts to the genome providing evidence of gene loci and intron boundaries. Adding the utility of RNA-seq data, many gene prediction tools and pipelines incorporate RNA-seq data to improve gene predictions (129,132,198). RNA-seq data has been used to annotate numerous fungal pathogen species (10,119) and the identification of novel small, cysteine-rich proteins has been reported (10). As such, the use of RNA-seq data in gene prediction has the potential to greatly benefit downstream effector prediction.

7.4 Methods

7.4.1 DNA library preparation and extraction (short-read)

Isolates were grown in Fries 2 media until sufficient biomass was achieved. Due to a high polysaccharide content, DNA was extracted using a modified high salt cetyltrimethylammonium bromide (CTAB) procedure incorporating 2M NaCl and based on that that described by Michiels et al. (226). Samples were extracted with an equal volume of chloroform/isoamyl (24:1) treated with 50 µg RNase, and extracted with chloroform/isoamyl again. DNA was precipitated with 5M NH_4Ac and precipitated at room temperature overnight with 0.7 volume isopropanol. Pelleted DNA was dissolved in TE and the further treated with buffer RLT (Qiagen). Finally, the DNA was extracted using chloroform/isoamyl followed by a 2x volumes EtOH precipitation.

7.4.2 RNA library preparation and extraction (short-read)

RNA tissue samples were taken at 3 days, 1 week and 3 weeks for isolates grown Fries 2 cultures. Samples were also taken from early growth on a mixture of V8 and PDA rapidly growing plates at 2 days) and from older plates (10 days) when conidia start to form. RNA extractions using Trizol (Thermo Fisher Scientific, Waltham, MA, USA) were performed as per the manufacturer's instructions. Any contaminating DNA was removed by DNase treatment (Thermo Fisher Scientific). cDNA was synthesised using a BioRad (Hercules, CA, USA) iScript cDNA kit.

7.4.3 Preparation of DNA for PacBio sequencing

Single molecule PacBio sequencing requires high MW DNA free of contaminants, a challenge with many fungal species and *P. teres* in particular as DNA preparations are typically viscous and coloured. Although DNA yields and A260/280 ratios tend to be good with published protocols, A260/230 ratios are consistently low. We therefore used partial digestion of fungal tissue with hydrolytic enzymes.

P. teres isolates were grown in the dark in Fries 2 with 3g/L sucrose in flasks at 125 rpm at room termperature for 2-3 days or until there was adequate biomass. Mycelia were rinsed in sterile water blended then grown overnight-24 hours as above. The mycelia were harvested by centrifuging at 2000g 10 minutes and rinsed three times with MWS (1M NaCl, 10mM MgCl2, 10mM KH2PO4 pH 5.8). Two grams of mycelia were digested overnight in 40 mL MWS with 3% Extralyse (Laffort, Bordeaux, France) with gentle rocking gently at 4°C. Mycelia harvested and rinsed three times with water before blotting dry on milk filters and blotting paper.

The CTAB protocol described above for short read DNA libraries and based on Michiels et al. (2003) was used with the following changes: Only a single phenol/chloroform/isoamyl (25:24:1) extraction followed by a 5 M NH4Ac and ethanol precipitation was performed after the CTAB incubation step and a single chloroform/isoamyl (24:1) extraction followed by a 5 M NH4Ac and isopropanol precipitation was carried out after RNase treatment to minimise DNA loss and shearing.

7.4.4 Assembly

7.4.4.1 Pyrenophora teres f. teres

PacBio reads were self-corrected and assembled using the PBcR pipeline (parameters: -length 500, -partitions 200, genomeSize=57000000, -sensitive) from Celera Assembler (v. 8.3) (115). Quiver (SMRT pipe) was used to polish the draft assembly generated by PBcR, as recommended in the PBcR manual. DNA Illumina short reads were trimmed for adapter sequences and quality using Trimmomatic (200) (adapters:2:30:10 LEADING:30,TRAILING:30, SLIDINGWINDOW:5:30, MINLEN:35). The resulting short reads were aligned to the polished PacBio assembly using bowtie (v. 2.2.5) (227). Pilon (v. 1.13) (228) (--fix bases,gaps,local,breaks) was then used to correct errors highlighted by the Illumina alignment.

7.4.4.2 Pyrenophora teres f. maculata

The *P. teres f. maculata* assembly was generated in the same manner as described above for *Ptt* using PBcR, Quiver, and Pilon with a smaller genome size set when running PBcR (genomeSize=42000000).

7.4.5 Annotation

7.4.5.1 DNA repeat identification

Repeat identification was based on the description given in the MAKER tutorials (229), and consistent with previously published studies of fungi employing a combination of structure based and *de novo* methods (20,230). The same repeat identification process was conducted independently on *Ptt* and *Ptm* assemblies. An identical analysis of the repetitive content of *Ptr* was also conducted to allow *Ptt* and *Ptm* to be compared to *Ptr* without bias due to differing repeat identification methods. Structure-based repeat identification was used to identify full-length long terminal repeat (LTR), miniature inverted-repeat transposable element (MITE), and terminal-repeat retrotransposons in miniature (TRIM) elements. A filtering process was then used to remove redundancy in this set and select a set of representative sequences. The resulting set of sequences was used to mask the genome before carrying out *de novo* repeat identification. Finally, the repeat libraries from the different repeat finding methods (structure based and *de novo*) were used to identify repeats in assembly. Details of each of these steps are described in detail in the following paragraphs.

MITE sequences were identified using MITE-hunter (231) with default parameters. Putative LTR transposon sequences were identified using LTRharvest (232), firstly searching for those with LTR sequences with 99% identity. Internal features of the putative LTRs were annotated using LTRdigest (233) with the eukaryotic tRNA database GtRNAdb (234). Putative transposons with an annotated polypurine tract (PPT) or primer binding site (PBS) with less than 50% overlap with LTR sequences, situated within 20 bp of the LTR (5' for PPT, 3' for PBS) were retained and others discarded. Sequence similarity between pairs of sequences up and downstream of LTRs can be an indication of a false positive LTR. As such, pairs of regions 50 bp up and downstream of LTRs were aligned and where an alignment had >60% identity over 25 or more nucleotides the putative LTR was discarded. Finally, nested transposons were filtered out by searching for alignments between LTR sequences and the inner regions of transposons, discarding transposons where an LTR aligned with >80% identity over >90% of its length. Representative transposons were selected using an iterative all–by–all blastn (BLAST+ v. 2.2.30) (209) approach. In each iteration blastn was used to find alignments with >80% identity, covering >90% of both the query and reference sequence. The sequence with the most of these alignments was moved to a set of representative sequences, and sequences it aligned to were discarded. This was repeated until either no hits were found (remaining sequences were moved to the set of representative sequences) or all sequences had been moved out of the main set. This process was repeated searching for LTRs with 85% identity, removing putative LTRs with homology to those already identified (>80% identity over 90% length). A library of TRIM sequences was also generated by the same method used to identify LTR-transposons, but with parameters restricting the search to sequence lengths typical of TRIMs.

MITE, TRIM and LTR-transposon libraries were concatenated and used with RepeatMasker (204) to mask the genome. The resulting genome sequence, within which repeats families identified using structure based methods had been masked, was then used as input to the *de novo* repeat identification pipeline RepeatModeler (203), which incorporates k-mer based (using RepeatScout (235)) and homology based (using RECON (203)) repeat prediction. This was done to discover any additional repeat families that were not detected by method relying on the recognition of structural features (such of LTRs) exhibited by members of some repeat families. Finally, the consensus library generated by RepeatModeler was combined with the MITE, TRIM, and LTR-transposon libraries and used with RepeatMasker to locate repeats in the genome.

7.4.5.2 AT-rich region and RIP

AT-rich regions were searched for using in *Ptt*, *Ptm*, and the publically available *Ptr* BFP-ToxAC assembly (12) using OcculterCut (described in Chapter 6). Dinucleotide frequencies calculated from AT-rich genome regions were compared to those calculated from GC-equilibrated genome regions. TpA/ApT RIP indices were calculated and compared for each region type.

7.4.5.3 Genes

Identical gene prediction methods were employed for both *Ptt* and *Ptm.* RNA-seq reads were aligned to the genome assembly using TopHat (v. 2.1.0) (122,201) and assembled using Cufflinks (v. 2.2.1) (202). The "junctions.bed" file output by TopHat was converted to a bed file of intron locations and used as an input to run GeneMark-ET (v. 4.30) (207). A second set of annotations was predicted using CodingQuarry (v. 2.0) (198), using the Cufflinks transcript alignment to self-train and inform predictions. In predicting genes using AUGUSTUS (124,133) a training set of genes was needed to first train AUGUSTUS. A training set of genes was not necessary to run CodingQuarry or GeneMark-ET, as these self-train. A recently released pipeline, BRAKER1 (132), automates the training of AUGUSTUS using GeneMark-ET, however the presented methods were finalised prior to its publication. To select a set of gene models for training, the subset of CodingQuarry predictions with Cufflinks transcript support were compared to GeneMark-ET predictions and a common set separated. Protein sequences were extracted from this set and blastp was used to search for alignments to reviewed Uniprot-KB entries from fungal species. Proteins with an alignment spanning ≥90% of both the query protein and the database entry were separated into a set of high quality genes. This high quality set was used to train AUGUSTUS. AUGUSTUS (v 3.2) (124,133) was run with the resulting parameters and intron hints derived from the TopHat alignment. CodingQuarry-PM (CodingQuarry v. 2.0 pathogen mode, Chapter 4) was run subsequent to CodingQuarry in an attempt to further predict effector-like genes. Gene predictions were run using the soft-masked genome sequence, that is, where repetitive genome regions are indicated by lower-case letters in the genome sequence file.

Predictions made by CodingQuarry, CodingQuarry-PM, GeneMark-ET, and AUGUSTUS and transcript alignments from Cufflinks were input to EVidenceModeler (v. 1.1.1) (128) (weights: Cufflinks = 6, CodingQuarry = 2, AUGUSTUS = 2, and GeneMark-ET = 1). Additional steps were then taken to ensure effector-like genes were included in the final annotation. Predictions made by CodingQuarry, CodingQuarry-PM, AUGUSTUS, and putative coding sequences annotated as the longest open-reading frame within each Cufflinks transcript were compared to the EVidenceModeler gene set using BEDTools (v. 2.17) (211) coverageBED tool. Predicted genes that did not overlap a gene in the EVidenceModeler gene set were separated. Protein sequences were extracted from these gene models, and those that were predicted to be secreted (SignalP v. 4.1 (208)) or had a Pfam domain hit were separated. Conflicting predicted gene structures at loci within this set were manually inspected. Where the Cufflinks assembly showed better support for one of the gene models, it was selected. "Better" support was judged as more complete exon support. Where there was no transcript evidence to assist in the choice of gene model, the longer predicted coding sequence was retained. Upon resolving conflicts, the additional predicted gene set was added to the EVidenceModeler set to form the final set of annotations.

7.4.6 Gene content analysis

Gene IDs and their basic properties (location, number of exons, size) were recorded in a table, alongside a summary of which prediction method had contributed to its prediction. If the gene model predicted by a particular software tool was identical to the gene model at that locus in the final annotation set, this software tool was considered to have contributed to the final annotation. Genes were compared to the Cufflinks derived transcript alignment and this information was documented to allow future users of these gene predictions to assess the reliability of predictions when considering their planned analysis.

Protein sequences from the final gene set were extracted and searched for Pfam (v. 29.0) domains using hmmscan (HMMER 3.1b2 (210), automatic cut-off determined using the parameter --cut-ga). Blastp (BLAST+ v. 2.2.30) (209) was used to identify genes common to *Ptt*, *Ptm*, and *Ptr.* When comparing *Ptm* and *Ptt*, protein sequences were queried against a protein database built from the other species' proteome. A query gene searched (using blastp from BLAST+ v. 2.2.30 (209)) against a protein database showing 90% or greater coverage by 90% or higher amino acid level alignments was considered to be common to both species. The *Ptr* annotation was derived by differing methods to those outlined for *Ptt* and *Ptm*, and as such presence and absence variation when comparing the predicted protein sets could reflect the differences in gene prediction techniques rather than genuine differences in the gene content. To account for this, when making comparisons with *Ptr*, protein sequences were compared to *Ptr* proteins (using blastp) and an additional step of searching for alignments to the *Ptr* genome sequence was carried out using tblastn (BLAST+ v. 2.2.30) (209). Again, to consider a query gene as "present" in a subject database, >90% protein level identity and >90% query sequence coverage by the blast alignments was required.

7.4.7 Effector gene prediction

Genes were assessed according to properties typically associated with effectors and the resulting data was recorded (Supplementary section S2 Tables 1 and 2). The distance from each gene to the closest AT-rich region and repeat was recorded. Coding sequences were extracted, translated into protein sequences, and their size (molecular weight in Daltons and length in amino acids) and cysteine content (as a count and percentage of the protein length) were recorded. Proteins were searched for Pfam domain homology using hmmscan (HMMER 3.1b2 (210), automatic cut-off determined using the parameter --cut-ga). SignalP (v. 4.1) was used to predict proteins with a signal peptide and the signal peptide length was recorded. TMHMM (v. 2.0) (236) was then run on the putative mature peptide sequences output by SignalP. Proteins that were predicted to be secreted by SignalP (v. 4.1) and did not contain a predicted transmembrane domain within their mature peptide sequence were separated and considered as the predicted secretome. This predicted secretome was analysed using EffectorP (89) and resulting probability scores recorded. A cumulative score out of 10 was given to each gene based on the following criteria:

- +2 Strong RNA-seq support
- +1 Partial RNA-seq support
- +2 Predicted to be secreted
- +1 EffectorP score ≥ 0.8
- +1 Within 1 kbp of an AT-rich region
- +1 Within 1 kbp nucleotides of a repetitive element
- +1 Absent or highly diverged in other *P. teres* species
- +1 Absent or highly diverged in *Ptr*
- +1 No Pfam domain OR a Pfam domain present in another known effector.

EffectorP takes into account the protein size and cysteine content as part of its automated machine learning process, and as such these protein attributes have been tabulated but do not contribute to the effector scoring. Strong RNA-seq support was included to ensure that the annotations of genes selected for downstream analysis are accurate, rather than because it indicates a gene may encode an effector. Each of the criteria used in this scoring system have been tabulated (Supplementary S2), allowing different scoring to be applied in future studies as further data comes to light.

7.5 Results

7.5.1 Assembly

A summary of the PacBio SMRT read sequencing output is given in Table 7. Initially, 4 x P4 chemistry and 2 x P5 chemistry SMRT cells were used for each *Ptt* and *Ptm*. A greater total read length for *Ptm* was achieved than *Ptt*. This was thought to be due to differences in the suitability of the library preparation techniques for each of the pathogens. These initial data sets were later supplemented with P6 chemistry reads in order to achieve greater coverage depth necessary for accurate error correction of the reads. Due to earlier observations of the higher read coverage achieved for *Ptm*, more SMRT cells were used for *Ptt* (10 x P6 for *Ptt* compared to 4 x P6 for *Ptm*) to compensate. The distributions of sub-read lengths for P4, P5, and P6 chemistries are shown in Figure 5. *Ptt* Won1-1 assembled into 397 scaffolds with a total sequence of 56.7 Mbp. More than half of the assembly is contained within the 10 largest scaffolds (N50 = 10), with the 10th largest scaffold 1.89 kbp in length (L50 = 1.89 Mbp). *Ptm* SG1-1 assembled into 131 scaffolds with a total sequence of 42.0 Mbp, N50 of 10, and L50 of 1.18 Mbp. Based on the assembly sizes achieved, the PacBio SMRT read coverage was 135X for *Ptt* and 179X for *Ptm*.

Pacbio chemistry	***Ptt***			***Ptm***		
	SMRT cells	Base pairs (Mbp)	Coverage (on 56.7Mbp)	SMRT cells	Base pairs (Mbp)	Coverage (on 42.0Mbp)
P4s	4	1,064	19.8X	4	2,183	51.9X
P5s	2	730	12.9X	2	1,434	34.1X
P6s	10	5,864	103X	4	3,918	93.3X
total	**16**	**7,658**	**135X**	**10**	**7,535**	**179X**

Table 7: A summary of the PacBio sub-reads obtained from sequencing *Ptt* and *Ptm*.

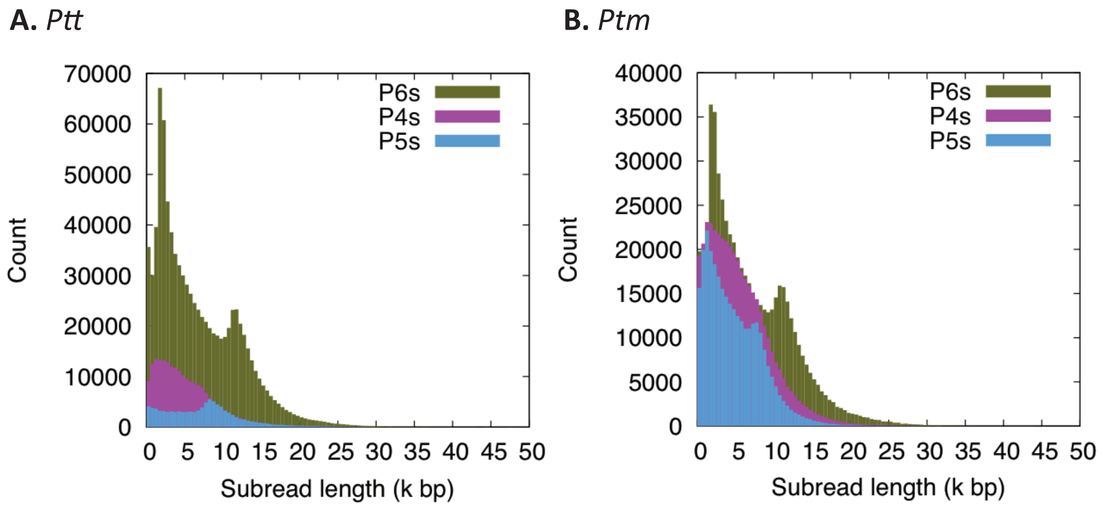


Figure 5 PacBio SMRT sub-read length distributions obtained for A. *Ptt* and B. *Ptm* are shown. Sub-reads derived from each of the different chemistries have been plotted in different colours (P4s – purple, P5s – blue, P6s – green).

7.5.2 Repetitive content

A total of 23.0 Mbp (40.4%) of the *Ptt* assembly was annotated as repetitive, compared to 10.7 Mbp (25.7%) of the *Ptm* assembly. This 14.4 Mbp difference in the repetitive content of the two assemblies accounts for most of the 12.3 Mbp difference in assembly size, giving the two species comparable non-repetitive assembly sizes of 33.7 Mbp (*Ptt*) and 31.3 Mbp (*Ptm*) (Figure 6). These sizes are close to the non-repetitive *Ptr* size of 32.4 Mbp, although *Ptr* had a much smaller repetitive component of 5.4 Mbp (14.3% of the assembly) (Figure 6).

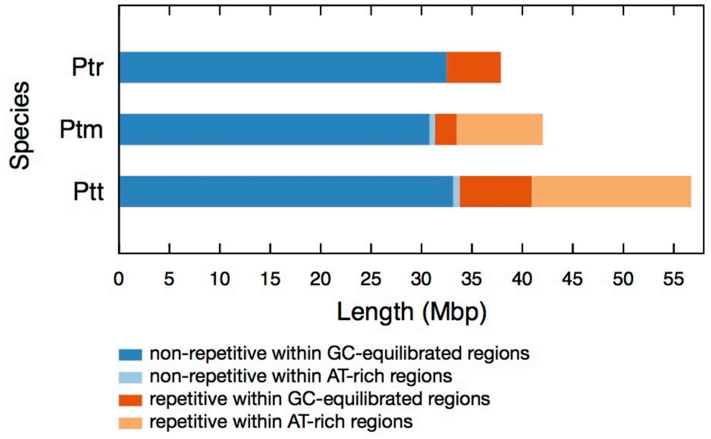


Figure 6: A summary of the length of the *Ptr*, *Ptm*, and *Ptt* assemblies taken up by repetitive (orange) and non-repetitive regions (blue). The overall lengths of repetitive and non-repetitive regions have been further broken down into those within AT-rich genome regions (light blue for non-repetitive regions, light orange for repetitive regions) and those within GC-equilibrated regions (dark blue for non-repetitive regions, dark orange for repetitive regions).

In all 3 assemblies (*Ptt*, *Ptm*, and *Ptr*), Class I LTR retrotransposons accounted for the majority of repetitive sequences (Table 8). In both *Ptt* and *Ptm*, LTR retrotransposon content was predominantly from *Gypsy* superfamily elements, whereas the *Ptr* LTR retrotransposon content was mainly from *Copia* superfamily elements.

Table 8 A summary of the repetitive content of the *Ptt, Ptm* and *Ptr* assemblies

			Total number of families			**Total length (bp)**			**Percentage of repeat content**			**Percentage of genome assembly**		
Class	**Repeat class**	**Repeat type**	***Ptt***	***Ptm***	***Ptr***	***Ptt***	***Ptm***	***Ptr***	***Ptt***	***Ptm***	***Ptr***	***Ptt***	***Ptm***	***Ptr***
Class I	LTR Retrotransposon	Copia	15	9	23	4,193,192	2,294,686	2,512,115	18.3%	21.5%	46.8%	7.4%	5.5%	6.6%
		Gypsy	30	13	9	9,496,045	5,375,888	833,991	41.4%	50.4%	15.5%	16.7%	12.8%	2.2%
total			**44**	**22**	**32**	**13,659,116**	**7,805,494**	**3,345,539**	**59.5%**	**73.1%**	**62.3%**	**24.1%**	**18.6%**	**8.8%**
	Non LTR	LINE/Tad1	2	-	2	252,112	-	121,946	1.1%	-	2.3%	0.4%	-	0.3%
		LINE/RTE-BovB	2	-	-	244,135	-	-	1.1%	-	-	0.4%	-	-
		LINE/R1	1	-	-	26,152	-	-	0.1%	-	-	0.0%	-	-
		LINE (Unkown)	1	-	-	122,898	-	-	0.5%	-	-	0.2%	-	-
		LINE/Penelope	-	1	-	-	20,257	-	-	0.2%	-	-	0.0%	-
		Line/L1	-	-	1	-	-	18,797	-	-	0.4%	-	-	0.0%
total			**6**	**1**	**3**	**645,243**	**20,257**	**140,743**	**2.8%**	**0.2%**	**2.6%**	**1.1%**	**0.0%**	**0.4%**
Class II	DNA transposon	TcMarFot1	8	4	4	970,289	718,752	410,488	4.2%	6.7%	7.6%	1.7%	1.7%	1.1%
		Academ	1	-	-	275,884	-	-	1.2%	-	-	0.5%	-	-
		hAT-Restless	1	1	7	199,380	42,565	540,095	0.9%	0.4%	10.1%	0.4%	0.1%	1.4%
		hAT-Ac	1	1	-	46,847	246,060	-	0.2%	2.3%	-	0.1%	0.6%	-
		TcMarTc1	1	1	1	44,691	24,372	38,624	0.2%	0.2%	0.7%	0.1%	0.1%	0.1%
		TcMar-ISRm11	-	-	2	-	-	33,751	-	-	0.6%	-	-	-
		TcMar-Ant1	-	-	1	-	-	126,589	-	-	2.4%	-	-	-
		MuLE-MuDR	1	-	-	20,313	-	-	0.1%	-	-	0.0%	-	-
		Ginger	-	1	-	-	7,512	-	-	0.1%	-	-	0.0%	-
total			**13**	**8**	**12**	**1,556,126**	**1,034,875**	**1,149,500**	**6.8%**	**9.7%**	**21.4%**	**2.7%**	**2.5%**	**3.0%**
Non-autonomous	TRIMs		4	5	1	154,215	149,084	7,013	0.7%	1.4%	0.1%	0.3%	0.4%	0.0%
	MITEs		35	14	15	2,532,409	224,248	185,420	11.0%	2.1%	3.5%	4.5%	0.5%	0.5%
total			**39**	**19**	**16**	**2,686,100**	**363,013**	**192,432**	**11.7%**	**3.4%**	**3.6%**	**4.7%**	**0.9%**	**0.5%**
Unknown /	Unknown /		43	28	49	4,403,798	1,306,494	912,961	19.2%	12.2%	17.0%	7.8%	3.1%	2.4%

GC-content analysis found the assemblies of *Ptt* and *Ptm* to be bimodal; alternating between blocks of AT-rich, repeat dense, gene sparse regions and gene dense GC-equilibrated regions. This genome feature — discussed in detail in Chapter 6 — has been reported in some fungal species, most notably *L. maculans*. AT-rich regions were not found in the genome assembly of *Ptr*, which has a clearly unimodal GC-content. AT-rich regions were on average around 24 kbp and 28 kbp in the *Ptt* and *Ptm* assemblies respectively, however both contained much longer AT-rich regions, the longest almost 400 kbp in the *Ptt* assembly and just over 300 kbp in the *Ptm* assembly. In both the *Ptt* and *Ptm* assemblies, the majority (63% and 80% respectively) of repetitive sequence was within AT-rich genome regions (Figure 6). AT-rich regions in Pezizomycotina species are generally a signature of RIP (see Chapter 6), which mutates cytosine nucleotides to thymine nucleotides (C to T). RIP is biased towards the mutation of CpA di-nucleotides, resulting in a high TpA/ApT ratio when compared to regions that are not RIP affected. The ratio of TpA/ApT dinucleotides seen in *Ptt* AT-rich regions was 1.9, compared to a ratio of 0.8 in GC-equilibrated regions, supporting the activity of RIP in AT-rich regions. Evidence of RIP is also seen in the AT-rich regions of *Ptm*, which have a TpA/ApT dinucleotide ratio of 1.9.

7.5.3 Gene annotations

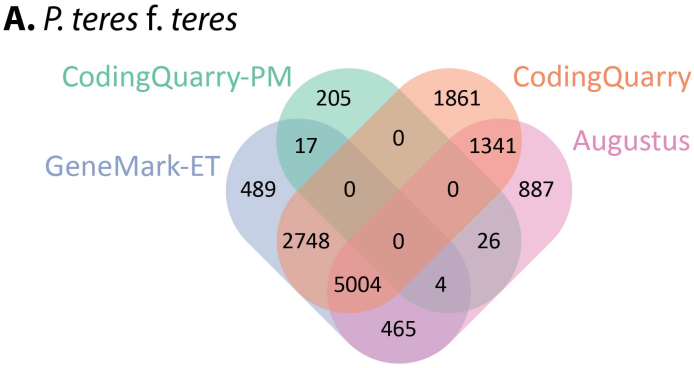
The *Ptt* gene set contained 13,470 genes with an average gene length of 493 amino acids and an average of 2.5 exons per gene. The *Ptm* gene prediction contained 11,813 genes, with a similar average gene length and average number of exons per gene to *Ptt*, at 471 amino acids and 2.6 exons per gene. The higher number of genes predicted in *Ptt* is consistent with the non-repetitive portion of the *Ptt* genome being 2.4 M bp larger than the non-repetitive portion of the *Ptm* genome (Figure 6). Details of the gene content of each species are summarised in Table 9 Summary of *Ptt* and *Ptm* gene content and listed in detail in supplementary section S2.

Table 9 Summary of *Ptt* and *Ptm* gene content

Species	Genome size (Mbp)	No. assembled transcripts (Cufflinks)	No. genes	Average exons per gene	Average gene length (amino acids)	Genes predicted to be secreted	Genes with Pfam domain
Ptt	56.7	14,606	13,470	2.5	493	1,327	8,672
Ptm	42.0	13,055	11,813	2.6	471	1,296	7,692

As described in the methods section, RNA-seq data was utilised in training gene prediction software and informing predictions. In the resulting *Ptt* gene set, 8153 genes had strong support from the transcript assembly alignment, 3,528 had partial support, and a minority of 1789 were unsupported. Similar levels of support were seen in the *Ptm* gene set, with 7115, 3192, and, 1506 genes with strong, partial, and no transcript assembly support respectively.

The bulk of the gene set for each species were output by EVidenceModel, as the result of combining predicted gene sets from CodingQuarry, CodingQuarry-PM, AUGUSTUS, and GeneMark-ET. This totalled 12,829 genes or 95.2% of final gene set for *Ptt* and 11,444 genes or 96.9% of the final gene set for *Ptm*. The agreement of predicted gene models from CodingQuarry, CodingQuarry-PM, GeneMark-ET, and AUGUSTUS is shown in Figure 7. Over 5000 genes in each of the final predicted gene sets for *Ptt* and *Ptm* were predicted identically by each of the gene predictors. In each species, over 80% of the CodingQuarry-PM predicted genes that were included in the final gene set were not predicted by the other gene prediction tools.



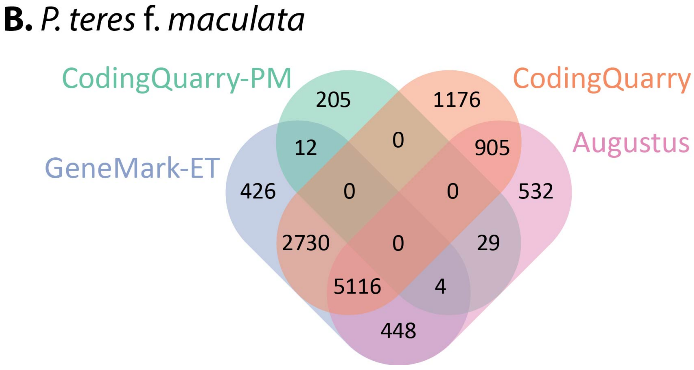


Figure 7 Venn diagrams showing the agreement of the various gene predictors with the final gene set for A *Ptt* and B *Ptm*.

Subsequent measures included predicted genes that were not included in the EVidenceModeler set but were predicted by other gene prediction software, and were predicted to be secreted or had Pfam domain support (see Methods). This resulted in the inclusion of 641 additional genes for *Ptt* and 369 genes for *Ptm.*

7.5.4 Effector prediction

When applied to the *Ptt* gene set, cumulative effector scoring system described resulted in 13 genes scoring 10/10, 96 scoring 9/10, and 233 scoring 8/10. Of the 109 top-scoring (9/10 and 10/10) genes, 62 gene models were part of the set that was added as a supplement to the EVidenceModeler set. Twenty-six of the top-scoring genes or 24% were predicted by CodingQuarry-PM, which is specially designed to predict effector-like genes. These results highlight the importance of ensuring effector like genes are included in annotations, as without the special measures taken to include these genes (see Methods) the list of candidate effectors would be significantly shorter. The set of 233 genes scoring 8/10 contains two proteins that are weak homologous (PttWon1-1.001410 and PttWon1-1.065610) to the *Fusarium oxysporum* effector *SIX7.* Local alignments between the *Ptt* proteins

PttWon1-1.001410 and PttWon1-1.065610 are 163 and 103 amino acids with 32.5% and 33.6% amino acid level identity. Also among the 233 genes scoring 8/10 is a protein sequence present in 12 copies throughout the genome and absent from *Ptm.* While not one of the top-scoring (9/10 or 10/10) effector candidates, the multi-copy nature of this gene is interesting considering the *Ptr* effector *ToxB* is present in multiple copies (173). The protein is 111 amino acids long, contains 3 cysteines, has an EffectorP score of 0.87 and does not have homology to any protein in NCBI's nr.

Of the 11,813 *Ptm* genes, 3 received an "effector" score of 10/10, 28 received a 9/10 score, and 156 received an 8/10 score based on the criteria discussed in the methods section and relating to protein sequence properties and genomic context. Of these 31 top-scoring (scoring 10/10 or 9/10) effector candidates, 12 (39%) were genes that had been added to the EVidenceModeler gene set in attempting to ensure effector-like genes were included in the gene annotation. Of the 31 top-scoring genes, 12 (39%) had been predicted by CodingQuarry-PM, a run mode of CodingQuarry designed to predict effector-like genes that are frequently missed by other gene prediction methods (see Chapter 4). Again, these results support the importance of ensuring effector-like genes are not missed at the gene prediction stage of effector candidate searches. A further 278 genes received a score of 7, among which is a protein that is not present in *Ptt* but shows 51.7% homology to the known *Ptr* effector *ToxB,* with the alignment spanning the full 87 amino acids of *ToxB*.

7.6 Discussion

The assembly and annotation results presented represent a dramatic improvement in the genomic resources available for *Ptt*. In addition to this improved *Ptt* assembly, we have also presented a high quality assembly of the closely related *Ptm*, enabling the presented comparisons of repetitive content and setting the stage for further comparative genomics. Furthermore, these new resources enabled comparisons between the two recently diverged barley pathogens *Ptt* and *Ptm*, and with the closely related wheat pathogen *Ptr*.

The previously published genome sequence of *Ptt* 0-1 was assembled from 75 bp paired end reads from Illumina's Solexa platform. Half of the assembly was contained within 408 scaffolds, as compared to just 10 scaffolds in the assembly presented here. Furthermore, at 56.7 Mbp the new assembly is around 14.7 Mbp larger than the previous assembly. This size difference can be largely attributed to the assembly of repeats and made possible by PacBio SMRT long-reads. This highlights one of the shortcomings of many fungal draft genome assemblies derived from short read sequencing, where repetitive regions are either absent or highly fragmented. This result advocates the use of PacBio SMRT sequencing as an alternative to short-read sequencing where downstream analysis of the repetitive content of the genome is planned.

With our new high-quality assemblies we find notable differences in the transposon content of these three *Pyrenophora* species. Transposon activity has been linked to host-jumps (72), and the higher repetitive content of *Ptm* and *Ptt* when compared to *Ptr* may have been instrumental in the divergence of the barley-infecting *Pyrenophora teres* species from *Ptr*. Expanded genomes and transposon activity has also been linked to the evolution of effectors (50,53). RIP activity in the barley infecting Pyrenophora species *Ptt* and *Ptm* has given them a bimodal GC-content similar to that seen in *L. maculans* and in stark contrast to the unimodal GC-content of *Ptr*. The activity of RIP within AT-rich regions of the *Ptm* and *Ptt* genomes adds another dimension to evolutionary potential of these species.

One of the key research goals in the study of these *P. teres* species is to identify candidate effector proteins, which can be expressed in the lab with the objective of identifying and developing resistant barley varieties. In other plant pathogen species, numerous effectors have been identified close to or within repetitive genome regions, and as such the assembly of these regions is particularly important. In the newly presented assemblies a number of secreted proteins with this genomic context are identified, and a more accurate assessment of this proximity of gene loci to repeats provided. These new insights about the repetitive content of *Ptm* and *Ptt* and additional effector candidate criteria were made possible through the use of PacBio sequencing, without which large portions of the genome would have been fragmented or absent (as was the case in the previously published *Ptt* assembly).

The identification of effector candidates can also be hampered by poor quality gene calls. By using a combination of RNA-seq driven methods, the newly presented annotations can be expected to be far more reliable than the previously published *ab initio* predictions. As such, the assembly and annotation improvements presented here will positively impact future efforts to identify effectors. Furthermore, top scoring effector candidates presented here may can be further refined through the incorporation of additional lines of evidence such as expression data, proteomics, and comparison with additional isolates.

7.6.1 Supporting Information

S2 Table 1: A list of *Ptt* genes with their protein properties and characteristics used to select effector candidates.

S2 Table 2: A list of *Ptm* genes with their protein properties and characteristics used to select effector candidates.

Chapter 8 Conclusion

The bioinformatic prediction of effectors is a multi-step process, involving genome assembly, gene prediction, and the selection of effector candidates. The research presented in this thesis has contributed to effector gene prediction through novel gene prediction techniques, an improved understanding of AT-rich regions and a method for assessing them, and genomic resources and candidate effectors for barley pathogens *Pyrenophora teres* f. *maculata* (*Ptm*) and *Pyrenophora teres* f. *teres* (*Ptt*).

8.1 Gene prediction

A large part of this thesis focused on the gene prediction stage of the typical effector candidate prediction pipeline, which follows genome assembly and precedes the actual selection of putative effectors from larger gene lists. This involved general improvements to fungal gene annotation through the incorporation of RNA-seq data, analysis of the prediction of known fungal effector gene loci, a method to specifically address the problem of effectors that are missed by gene prediction software, and a practical application of gene prediction methods in the annotation of *Ptm* and *Ptt.*

In the first part of this research, fungal gene prediction improvement through the incorporation of RNA-seq transcripts was investigated (198). The method developed, CodingQuarry, capitalises on the high proportion of correctly assembled transcripts seen in fungal RNA-seq experiments by using these both for training and to make coding sequences predictions directly from assembled transcripts sequences. CodingQuarry was also designed mindful of a common error in fungal transcript reconstruction, where adjacent gene loci with overlapping untranslated regions (UTRs) can assemble into a single, merged transcript. Benchmarking against high-quality gene sets within model fungal organisms *S. pombe* and *S. cerevisiae* demonstrated over 90% of genes were predicted correctly — a 4–5% improvement over the next-best competing method tested. Additional advantages were shown in a reduction of adjacent genes being predicted as a single larger gene when compared to AUGUSTUS (237). An additional feature was included to predict putative RNA-seq supported genes where standard gene prediction does not predict a coding sequence. This was included with application to pathogenomics in mind, where additional "dubious" gene predictions could be included in effector candidate lists if they conformed to other effector-like characteristics (for example, are predicted signal peptide).

Approaches combining the results of different gene prediction tools have been shown to achieve higher accuracy than the individual tools alone (128). This relies on one prediction tool making a correct prediction at a locus where another prediction tool fails, meaning a higher overall accuracy can be achieved by selecting the correctly predicted gene model at each locus. An important aspect to CodingQuarry was that it incorporated RNA-seq evidence in a different way to competing software. The result of this was that 95% of the common set of genes predicted by CodingQuarry and AUGUSTUS were correct, and each gene predictor made some correct predictions at loci where the other failed. Furthermore, CodingQuarry is a novel gene prediction tool, rather than one that combines existing tools, meaning that incorporating CodingQuarry into pipelines and combining predictions with those made by other methods has the potential to further improve fungal gene prediction. An avenue of future research in improving fungal gene prediction accuracy could be to catalogue the types of prediction errors different gene prediction software makes and tailor pipelines according to this information. For example, the methodology used in CodingQuarry makes it less likely to merge adjacent gene loci than AUGUSTUS, and fewer of these errors were observed in CodingQuarry predictions when benchmarking on *S. pombe*. As such, a pipeline may benefit from weighting CodingQuarry predictions higher than AUGUSTUS predictions where AUGUSTUS predicts one large gene and CodingQuarry predicts two smaller adjacent genes.

One of the difficulties in attempting to improve fungal gene prediction is having high confidence annotations to benchmark sensitivity and specificity against. Ideally, prediction methods should be benchmarked genome-wide to correctly capture the full range of variation that exists within a genome. Furthermore, gene models used for benchmarking should ideally have a high degree of experimental support. Even in whole genome annotations that are of an overall high quality, experimental support may not exist or its acquisition may be experimentally intractable for certain gene loci. Gene models lacking experimental support are likely to be themselves the result of gene prediction software, which can then bias attempts to benchmark novel and existing software. In avoiding the pitfalls of inaccuracies in benchmarking sets, CodingQuarry predictions were compared to high-quality annotations that exist for model species *Saccharomyces cerevisae* and *Schizosaccharomyces pombe.* It would be extremely beneficial to gene prediction improvement to have high-quality reference pathogen genome annotations, which could be used to further assess and improve CodingQuarry and other gene prediction tools.

Comparing gene prediction results to known effector genes showed that effectors are frequently missed by gene prediction tools. The investigation into the prediction of effector genes using existing gene prediction tools showed that effector genes were frequently missed by gene prediction software. The prediction of effectors was more successful where RNA-seq data supported the gene locus. In addressing the problem of missed effector gene loci, CodingQuarry-PM — a pathogen mode of CodingQuarry that target the prediction of effector-like genes — was presented. CodingQuarry-PM was able to greatly improve this situation, vastly reducing the number of missed effectors, by using a method that benefited from RNA-seq data but was not reliant on it. Future applications of CodingQuarry-PM to fungal pathogen genome sequences are likely to locate novel effector candidates.

An area of future research using CodingQuarry-PM could be use proteomics to validate to predictions. Analysis of proteomics datasets involves aligning peptide weights to locations in a protein sequence database. Matches to CodingQuarry-PM predictions would provide strong evidence that these are genuine genes. Including CodingQuarry-PM predicted protein sequences in protein sequence databases has the potential to identify novel peptide matches that had been previously overlooked. Such results would lend further support to CodingQuarry-PM predictions and may reveal interesting effector candidates.

8.2 AT-rich regions

In Chapter 5 and Chapter 6, AT-rich regions within fungal genomes were investigated, motivated by known examples of effector-encoding genes in close proximity to AT-rich regions. Connections between the activity of repeat-induced point mutation (RIP), transposon activity, and the evolution of pathogenicity genes have been made in previous studies. The work presented in this thesis contributes a succinct method for determining the presence and genomic location of AT-rich regions (OcculterCut) and a systematic survey of their presence and properties in published fungal genomes. The first application of the OcculterCut method was made in describing AT-rich regions within *Z. tritici* (Chapter 5) (183), following which a broad survey of 519 published fungal genomes was conducted (Chapter 6). Finally, OcculterCut was used to determine the AT-rich region content of *P. teres* f. *teres* and *P. teres* f. *maculata* and this information was used as a line of evidence in selecting candidate effectors (Chapter 7).

The survey of the AT-rich region content of published fungal genomes unequivocally demonstrates the ubiquity of AT-rich regions and bimodal genomes, particularly within the Pezizomycotina subphylum and plant-associated fungi therein. This opens up the possibility of investigating candidate effectors from within AT-rich regions in a number of pathogen species, and should alert future sequencing projects to the possibility of seeing this genome type in a newly assembled genome sequences. Community uptake of OcculterCut will be an important aspect to the continued documentation of AT-rich regions and comparisons with the presented survey. It is hoped that the publication of the survey data and software tool presented in Chapter 6, and the fact that OcculterCut is free and trivial to install and use will assist in this regard.

One of the limitations of this study was genome assembly quality and the effect this may have on survey results. It is reasonable to expect that genome assemblies from short read sequencing may poorly describe AT-rich regions. As platforms such as PacBio grow in popularity and draft assemblies from short-read data are updated, we may see differing results to those presented in this survey for some species. Within this thesis this was found to be true of *P. teres* f. *teres*, where the short-read, publically available genome assembly surveyed in Chapter 6 did not contain AT-rich regions but the higher-quality PacBio assembly presented in Chapter 7 had an AT-rich region component of 29%.

The relevance of AT-rich regions to study of phytopathogens is clear given their frequent presence in plant associated Dothideomycete genomes. Additionally, AT-rich regions are arguably relevant to the three main stages of bioinformatic effector prediction: genome assembly, gene prediction, and the selection of effector candidates. The relevance to assembly is due to the difficulty in assembling highly repetitive regions and sequencing biases that result in lower read coverage of AT-rich regions when using some sequencing platforms (112). The demonstrated differences in codon and amino acid usage of genes within AT-rich regions when compared to genes within GC-equilibrated regions (Chapter 6) are relevant to gene prediction within these regions. The proximity of genes to AT-rich regions can be used as a line of evidence when selecting effector candidates from larger lists of genes, based on known examples of effectors locating in AT-rich genome regions (17,66). Furthermore, trends in amino acid and codon usage in coding sequences within AT-rich genome regions may be significant to techniques that assess the amino acid content of effectors, as effectors residing within these regions are likely to differ from other effectors in their amino acid usage.

8.3 *Pyrenophora teres* f. *teres* and *Pyrenophora teres* f. *maculata*

In Chapter 7, improved genomic resources for *Ptt* and the first genomic resources for *Ptm* were presented. The genome assemblies were generated using single-molecule real-time (SMRT) PacBio sequencing technology. Repeats and AT-rich regions were annotated and compared to *P. tritici-repentis*. RNA-seq driven annotations of coding sequences were made, and genes were assessed with respect to criteria typically associated with effectors.

The motivation for using SMRT sequencing was the potential to resolve repetitive regions and generate a more complete assembly than can typically be achieved by using short-read data. The use of SMRT sequencing technology was extremely worthwhile considering the most striking difference between *P. teres* genomes and the related wheat pathogen *P. tritici-repentis* was the large difference in genome size and repeat content, and the presence of distinct AT-rich regions. AT-rich regions were not present in the previously published assembly of *Ptt* 0-1 (15), which is likely due to it being generated from short-read data, and due to the exclusion of small contig fragments from the publically available version of the genome. This further advocates the importance of achieving high-quality genome assemblies for these species.

Predicted gene sets were also presented for *Ptt* and *Ptm*, using a range of RNA-seq driven methods that were then combined to form a final annotation. These annotations are a strong resource for further research into these pathogens. The assessment of the prediction of known effectors in Chapter 4 highlighted the high number of effectors that are missed by gene prediction, motivating additional steps to ensure that secreted genes were included in the final annotation sets. Over 50% of the top-scoring *Ptt* effector candidates and over 30% of the top-scoring *Ptm* effector candidates were added to the predicted gene set through these secretome-targeted gene prediction steps. Together with the results of gene prediction benchmarking on known effectors in Chapter 4, this result strongly supports importance of ensuring effector-like gene loci are included in whole genome annotations. Effector candidates that were the result of prediction by CodingQuarry-PM (Chapter 4) further support the utility of this software in gene prediction where downstream effector candidate selection is planned.

The importance of *Ptt* and *Ptm* to the barley industry means that research into these pathogens is likely to continue and grow. The results presented in this thesis will be of use to further studies, in particular in searching for effectors. Refining effector candidates with additional data such as *in planta* RNA-seq and proteomics of secreted protein fractions is likely to be particularly useful.

8.4 Future directions

Much of the research and review of bioinformatics effector prediction has focussed on the selection of effector candidates from larger gene lists (86,89,90). More recently, the accuracy of other bioinformatic steps that contribute to the overall success of effector prediction has been increasingly recognised and addressed (48,113,222). The bioinformatic prediction of fungal effectors is likely to long remain an active and challenging area of research. It is hoped that the methods presented within this thesis will be adopted and will motivate further efforts to specialise bioinformatic techniques to the prediction of fungal effectors.

References

1. Murray GM, Brennan JP. Estimating disease losses to the Australian wheat industry. Australas Plant Pathol. 2009;38:558–70.
2. Murray GM, Brennan JP. Estimating disease losses to the Australian barley industry. Australas Plant Pathol. 2010;39:85–96.
3. Strange RN, Scott PR. Plant disease: A threat to global food security. Annu Rev Phytopathol. 2005;43:83–116.
4. Murray GM, Brennan JP. The Current and Potential Costs from Diseases of Pulse Crops in Australia. Canberra; 2012.
5. 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.
6. Solomon PSP, Lowe RRGT, Tan K-C, Waters ODC, Oliver RP. *Stagonospora nodorum*: cause of stagonospora nodorum blotch of wheat. Mol Plant Pathol. 2006;7(3):147–56.
7. Oliver RP, Friesen TL, Faris JD, Solomon PS. *Stagonospora nodorum*: From Pathology to Genomics and Host Resistance. Annu Rev Phytopathol. 2012;50(1):23–43.
8. Hane JK, Lowe RGT, 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.
9. 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.
10. Syme RA, Tan K-C, Hane JK, Dodhia K, Stoll T, Hastie M, et al. Comprehensive Annotation of the *Parastagonospora nodorum* Reference Genome Using Next-Generation Genomics, Transcriptomics and Proteogenomics. PLoS One. 2016;11(2):e0147221.
11. Ciuffetti LM, Manning VA, Pandelova I, Faris JD, Friesen TL, Strelkov SE, et al. *Pyrenophora tritici-repentis*: A Plant Pathogenic Fungus with Global Impact. In: Genomics of Plant-Associated Fungi: Monocot Pathogens. 2014. p. 103–22.
12. 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.
13. Liu Z, Ellwood S, Oliver R, Friesen T. *Pyrenophora teres*: profile of an increasingly damaging barley pathogen. Mol Plant Pathol. 2011;12(1):1–19.
14. Ellwood SR, Syme RA, Moffat CS, Oliver RP. Evolution of three *Pyrenophora* cereal pathogens: recent divergence, speciation and evolution of non-coding DNA. Fungal Genet Biol. 2012;49(10):825–9.

15. Ellwood SR, Liu Z, Syme RA, Lai Z, Hane JK, Keiper F, et al. A first genome assembly of the barley fungal pathogen *Pyrenophora teres* f. *teres*. Genome Biol. 2010;11:R109.

16. Rouxel T, Balesdent MH. The stem canker (blackleg) fungus, *Leptosphaeria maculans*, enters the genomic era. Mol Plant Pathol. 2005;6(3):225–41.

17. Rouxel T, Grandaubert J, Hane JK, Hoede C, Wouw P Van De, Couloux A, et al. Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. Nat Commun. 2011;2:202.

18. Goodwin SB, M'barek S Ben, Dhillon B, Wittenberg AHJ, Crane CF, Hane JK, et al. Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLOS Genet. 2011;7(6):e1002070.

19. Kema GHJ, Yu D, Rijkenberg FHJ, Shaw MW, Baayen RP. Histology of the pathogenesis of *Mycosphaerella graminicola* in wheat. Phytopathology. 1996;86(7):777–86.

20. Dhillon B, Gill N, Hamelin RC, Goodwin SB. The landscape of transposable elements in the finished genome of the fungal wheat pathogen *Mycosphaerella graminicola*. BMC Genomics. 2014;15(1):1132.

21. Thomma BPHJ, van Esse HP, Crous PW, DE Wit PJGM. *Cladosporium fulvum* (syn. *Passalora fulva*), a highly specialized plant pathogen as a model for functional studies on plant pathogenic Mycosphaerellaceae. Mol Plant Pathol. 2005;6(4):379–93.

22. de Wit PJGM, van der Burgt A, Ökmen B, Stergiopoulos I, Abd-Elsalam KA, Aerts AL, et al. 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 Genet. 2012;8(11):e1003088.

23. Rouxel T, Balesdent M-H. Avirulence Genes. Encycl Life Sci. 2010;January.

24. 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;66:513–45.

25. De Wit PJGM, Mehrabi R, Van Den Burg HA, Stergiopoulos I. Fungal effector proteins: past, present and future. Mol Plant Pathol. 2009;10(6):735–47.

26. Stergiopoulos I, de Wit PJGM. Fungal effector proteins. Annu Rev Phytopathol. 2009;47:233–63.

27. Dodds PN, Rathjen JP. Plant immunity: towards an integrated view of plant-pathogen interactions. Nat Rev Genet. 2010;11(8):539–48.

28. Weiberg A, Wang M, Lin F-M, Zhao H, Zhang Z, Kaloshian I, et al. Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways. Science (80-). 2013;342(6154):118–23.

29. Bohnert HU. A Putative Polyketide Synthase/Peptide Synthetase from *Magnaporthe grisea* Signals Pathogen Attack to Resistant Rice. Plant Cell. 2004;16(9):2499–513.

30. Garnica DP, Nemri A, Upadhyaya NM, Rathjen JP, Dodds PN. The Ins and Outs of Rust

Haustoria. PLOS Pathog. 2014;10(9):e1004329.

31. Petre B, Kamoun S. How Do Filamentous Pathogens Deliver Effector Proteins into Plant Cells? PLOS Biol. 2014;12(2):e1001801.

32. Marshall R, Kombrink A, Motteram J, Loza-Reyes E, Lucas J, Hammond-Kosack KE, et al. 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 Physiol. 2011 Jun;156(2):756–69.

33. Sánchez-Vallet A, Saleem-Batcha R, Kombrink A, Hansen G, Valkenburg D-J, Thomma BP, et al. Fungal effector Ecp6 outcompetes host immune receptor for chitin binding through intrachain LysM dimerization. Elife. 2013;2:1–16.

34. Bolton MD, van Esse HP, Vossen JH, de Jonge R, Stergiopoulos I, Stulemeijer IJE, et al. The novel *Cladosporium fulvum* lysin motif effector *Ecp6* is a virulence factor with orthologues in other fungal species. Mol Microbiol. 2008;69(1):119–36.

35. Flor H. Inheritance of pathogenicity in *Melampsora lini*. Phytopathology. 1942;32:653–69.

36. Flor H. Inheritance of reaction to rust in flax. J Agric Res. 1947;74:241–62.

37. van Kan JA, van den Ackerveken GFJM, de Wit PJGM. Cloning and characterization of cDNA of avirulence gene avr9 of the fungal pathogen *Cladosporium fulvum*, causal agent of tomato leaf mold. Mol plant-microbe Interact. 1991;4(1):52–9.

38. van den Burg HA, Harrison SJ, Joosten MHAJ, Vervoort J, de Wit PJGM. *Cladosporium fulvum* Avr4 protects fungal cell walls against hydrolysis by plant chitinases accumulating during infection. Mol plant-microbe Interact. 2006;19(12):1420–30.

39. van Esse HP, Van't Klooster JW, Bolton MD, Yadeta KA, van Baarlen P, Boeren S, et al. The *Cladosporium fulvum* virulence protein Avr2 inhibits host proteases required for basal defense. Plant Cell. 2008;20(7):1948–63.

40. Rooney HCE, van't Klooster JW, ver der Hoorn RAL, Joosten MHA, Jones JDG, de Wit PJGM. *Cladosporium* Avr2 Inhibits Tomato Rcr3 Protease Required for Cf-2–Dependent Disease Resistance. Science (80-). 2005;308.

41. Parlange F, Daverdin G, Fudal I, Kuhn M-L, Balesdent M-H, Blaise F, et al. *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. Mol Microbiol. 2009;71(4):851–63.

42. Gout L, Kuhn ML, Vincenot L, Bernard-Samain S, Cattolico L, Barbetti M, et al. Genome structure impacts molecular evolution at the *AvrLm1* avirulence locus of the plant pathogen *Leptosphaeria maculans*. Environ Microbiol. 2007;9(12):2978–92.

43. Joosten MH, Vogelsang R, Cozijnsen TJ, Verberne MC, De Wit PJ. The biotrophic fungus *Cladosporium fulvum* circumvents *Cf-4*-mediated resistance by producing unstable AVR4 elicitors. Plant Cell. 1997;9(3):367–79.

44. Stergiopoulos I, De Kock MJD, Lindhout P, De Wit PJGM. Allelic variation in the effector genes

of the tomato pathogen *Cladosporium fulvum* reveals different modes of adaptive evolution. Mol Plant Microbe Interact. 2007;20(10):1271–83.

45. Oliver RP, Solomon PS. New developments in pathogenicity and virulence of necrotrophs. Curr Opin Plant Biol. 2010;13(4):415–9.

46. Liu Z, Holmes DJ, Faris JD, Chao S, Brueggeman RS, Edwards MC, et al. Necrotrophic effector-triggered susceptibility (NETS) underlies the barley- *Pyrenophora teres* f. *teres* interaction specific to chromosome 6H. Mol Plant Pathol. 2015;16(2):188–200.

47. Schmidt SM, Houterman PM, Schreiver I, Ma L, Amyotte S, Chellappan B, et al. MITEs in the promoters of effector genes allow prediction of novel virulence genes in *Fusarium oxysporum*. BMC Genomics. 2013;14(1):119.

48. Mesarich CH, Griffiths SA, Van Der Burgt A, Ökmen B, Beenen HG, Etalo DW, et al. Transcriptome sequencing uncovers the *Avr5* avirulence gene of the tomato leaf mold pathogen *Cladosporium fulvum*. Mol Plant-Microbe Interact. 2014;27(8):846–57.

49. Amselem J, Lebrun M-H, Quesneville H. Whole genome comparative analysis of transposable elements provides new insight into mechanisms of their inactivation in fungal genomes. BMC Genomics. 2015;16:141.

50. Spanu PD. The Genomics of Obligate (and Nonobligate) Biotrophs. Annu Rev Phytopathol. 2012;50(1):91–109.

51. Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stüber K, et al. Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. Science (80-). 2010;330(6010):1543–6.

52. Amselem J, Cuomo C a, van Kan J a L, Viaud M, Benito EP, Couloux A, et al. Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. PLOS Genet. 2011;7(8):e1002230.

53. Raffaele S, Kamoun S. Genome evolution in filamentous plant pathogens: why bigger can be better. Nat Rev Microbiol. 2012;10(6):417–30.

54. Kang S, Lebrun MH, Farrall L, Valent B. Gain of virulence caused by insertion of a Pot3 transposon in a *Magnaporthe grisea* avirulence gene. Mol plant-microbe Interact. 2001;14(5):671–4.

55. Luderer R, Takken FLW, Wit PJGM de, Joosten MHAJ. *Cladosporium fulvum* overcomes *Cf-2*-mediated resistance by producing truncated AVR2 elicitor proteins. Mol Microbiol. 2002;45(3):875–84.

56. Hu J, Chen C, Peever T, Dang H, Lawrence C, Mitchell T. Genomic characterization of the conditionally dispensable chromosome in *Alternaria arborescens* provides evidence for horizontal gene transfer. BMC Genomics. 2012;13(171):1–13.

57. Sperschneider J, Gardiner DM, Thatcher LF, Lyons R, Singh KB, Manners JM, et al. Genome-Wide Analysis in Three *Fusarium* Pathogens Identifies Rapidly Evolving Chromosomes and Genes Associated with Pathogenicity. Genome Biol Evol. 2015;7(6):1613–27.

58. Ipcho S, Hane J, Antoni EA, Ahren G, Henrissat B, Friesen T, et al. Transcriptome analysis of Stagonospora nodorum: gene models, effectors, metabolism and pantothenate dispensability. Mol Plant Pathol. 2011;1–15.

59. Ajiro N, Miyamoto Y, Masunaka A, Tsuge T, Yamamoto M, Ohtani K, et al. Role of the host-selective ACT-toxin synthesis gene ACTTS2 encoding an enoyl-reductase in pathogenicity of the tangerine pathotype of *Alternaria alternata*. Phytopathology. 2010;100(2):120–6.

60. Ma LL-J, van der Does HC, Borkovich K a, Coleman JJ, Daboussi M-J, Di Pietro A, et al. Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature. 2010;464(7287):367–73.

61. Croll D, McDonald BA. The accessory genome as a cradle for adaptive evolution in pathogens. PLOS Pathog. 2012;8(4):e1002608.

62. Selker E, Cambareri E, Jensen B, Haack K. Rearrangement of duplicated DNA in specialized cells of Neurospora. Cell. 1987;51:741–52.

63. Watters MK, Randall TA, Margolin BS, Selker EU, Stadier DR. Action of repeat-induced point mutation on both strands of a duplex and on tandem duplications of various sizes in neurospora. Genetics. 1999;153(2):705–14.

64. Cambareri EB, Singer MJ, Selker EU. Recurrence of Repeat-Induced Point Mutation. Genetics. 1991;127:699–710.

65. Clutterbuck AJ. Genomic evidence of repeat-induced point mutation (RIP) in filamentous ascomycetes. Fungal Genet Biol. 2011;48(3):306–26.

66. Gout L, Fudal I, Kuhn M-L, Blaise F, Eckert M, Cattolico L, et al. Lost in the middle of nowhere: the *AvrLm1* avirulence gene of the Dothideomycete *Leptosphaeria maculans*. Mol Microbiol. 2006;60(1):67–80.

67. Van de Wouw AP, Lowe RGT, Elliott CE, Dubois DJ, Howlett BJ. An avirulence gene, *AvrLmJ1*, from the blackleg fungus, *Leptosphaeria maculans*, confers avirulence to *Brassica juncea* cultivars. Mol Plant Pathol. 2014;15(5):523–30.

68. Ghanbarnia K, Fudal I, Larkan NJ, Links MG, Balesdent M-H, Profotova B, et al. Rapid identification of the *Leptosphaeria maculans* avirulence gene *AvrLm2* using an intraspecific comparative genomics approach. Mol Plant Pathol. 2015;16(7):699–709.

69. Grandaubert J, Lowe RGT, Soyer JL, Schoch CL, Van de Wouw AP, Fudal I, et al. Transposable element-assisted evolution and adaptation to host plant within the *Leptosphaeria maculans-Leptosphaeria biglobosa* species complex of fungal pathogens. BMC Genomics. 2014;15:891.

70. Lowe RGT, Cassin A, Grandaubert J, Clark BL, Van de Wouw AP, Rouxel T, et al. Genomes and transcriptomes of partners in plant-fungal-interactions between canola (*Brassica napus*) and two *Leptosphaeria* species. PLoS One. 2014;9(7):e103098.

71. Fudal I, Ross S, Brun H. Repeat-induced point mutation (RIP) as an alternative mechanism of evolution toward virulence in *Leptosphaeria maculans*. Mol Plant-Microbe Interact. 2009;22(8):932–41.

72. Raffaele S, Farrer RA, Cano LM, Studholme DJ, MacLean D, Thines M, et al. Genome Evolution Following Host Jumps in the Irish Potato Famine Pathogen Lineage. Science (80-). 2010;330(December):1540–4.

73. Stukenbrock EH, Jørgensen FG, Zala M, Hansen TT, McDonald BA, Schierup MH. Whole-genome and chromosome evolution associated with host adaptation and speciation of the wheat pathogen *Mycosphaerella graminicola*. PLOS Genet. 2010;6(12):e1001189.

74. Dong S, Raffaele S, Kamoun S. The two-speed genomes of filamentous pathogens: waltz with plants. bioRxiv. Elsevier Ltd; 2015;35:021774.

75. Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H, Faris JD, et al. Emergence of a new disease as a result of interspecific virulence gene transfer. Nat Genet. 2006;38(8):953–6.

76. Coleman JJ, Rounsley SD, Rodriguez-Carres M, Kuo A, Wasmann CC, Grimwood J, et al. The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. PLOS Genet. 2009;5(8):e1000618.

77. Marcet-Houben M, Gabaldón T. Acquisition of prokaryotic genes by fungal genomes. Trends Genet. 2010;26(1):5–8.

78. Cheeseman K, Ropars J, Renault P, Dupont J, Gouzy J, Branca A, et al. Multiple recent horizontal transfers of a large genomic region in cheese making fungi. Nat Commun. 2014;5:2876–85.

79. Soanes D, Richards TA. Horizontal gene transfer in eukaryotic plant pathogens. Annu Rev Phytopathol. 2014;52:583–614.

80. Oliver RP, Solomon PS. Recent Fungal Diseases of Crop Plants: Is Lateral Gene Transfer a Common Theme? Am Phytopathol Soc. 2008;21(3):287–93.

81. McIntosh RA. From Farrer to the Australian Cereal Rust Control Program. Aust J Agric Res. 2007;58(6):550–7.

82. Vleeshouwers VGAA, Oliver RP. Effectors as Tools in Disease Resistance Breeding Against Biotrophic, Hemibiotrophic, and Necrotrophic Plant Pathogens. Mol Plant-Microbe Interact. 2014;27(3):196–206.

83. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, et al. The genome sequence of the filamentous fungus *Neurospora crassa*. Nature. 2003;422(6934):859–68.

84. Galagan JE, Calvo SE, Cuomo C, Ma L-J, Wortman JR, Batzoglou S, et al. Sequencing of Aspergillus nidulans and comparative analysis with *A. fumigatus* and *A. oryzae*. Nature. 2005;438(7071):1105–15.

85. Dean R, Talbot N, Ebbole D. The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature. 2005;434(April):980–6.

86. Sperschneider J, Dodds PN, Gardiner DM, Manners JM, Singh KB, Taylor JM. Advances and Challenges in Computational Prediction of Effectors from Plant Pathogenic Fungi. PLOS Pathog. 2015;11(5):e1004806.

87. 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.

88. 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(1):e29847.

89. Sperschneider J, Gardiner DM, Dodds PN, Tini F, Covarelli L, Singh KB, et al. EFFECTORP: predicting fungal effector proteins from secretomes using machine learning. New Phytol. 2015;December:1–19.

90. 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 Genomics. 2013;14:807.

91. Godfrey D, Böhlenius H, Pedersen C, Zhang Z, Emmersen J, Thordal-Christensen H. Powdery mildew fungal effector candidates share N-terminal Y/F/WxC-motif. BMC Genomics. 2010;11(1):317.

92. Manning VA, Hamilton SM, Karplus PA, Ciuffetti LM. The Arg-Gly-Asp-containing, solvent-exposed loop of Ptr ToxA is required for internalization. Mol plant-microbe Interact. 2008;21(3):315–25.

93. Lu S, Gillian Turgeon B, 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.

94. Shiller J, Van de Wouw AP, Taranto AP, Bowen JK, Dubois D, Robinson A, et al. 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.

95. Stergiopoulos I, van den Burg HA, Okmen B, Beenen HG, van Liere S, Kema GHJ, et al. Tomato Cf resistance proteins mediate recognition of cognate homologous effectors from fungi pathogenic on dicots and monocots. Proc Natl Acad Sci. 2010;107(16):7610–5.

96. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, et al. Pfam: the protein families database. Nucleic Acids Res. 2014;42(Database issue):D222–30.

97. Guyon K, Balagué C, Roby D, Raffaele S. Secretome analysis reveals effector candidates associated with broad host range necrotrophy in the fungal plant pathogen *Sclerotinia sclerotiorum*. BMC Genomics. 2014;15(336):1–18.

98. Yoshida K, Saitoh H, Fujisawa S, Kanzaki H, Matsumura H, Yoshida K, et al. Association genetics reveals three novel avirulence genes from the rice blast fungal pathogen *Magnaporthe oryzae*. Plant Cell. 2009;21(5):1573–91.

99. Zhao S, Fung-Leung WP, Bittner A, Ngo K, Liu X. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. PLoS One. 2014;9(1):e78644.

100. Rudd J, Kanyuka K, Hassani-Pak K, Derbyshire M, Andongabo A, Devonshire J, et al. Transcriptome and metabolite profiling the infection cycle of *Zymoseptoria tritici* on wheat

(Triticum aestivum) reveals a biphasic interaction with plant immunity involving differential pathogen chromosomal contributions, and a variation on the hemibiotro. Plant Physiol. 2015;167(3):1158–85.

101. Mirzadi Gohari A, Ware SB, Wittenberg AHJ, Mehrabi R, Ben M'Barek S, Verstappen ECP, et al. Effector discovery in the fungal wheat pathogen *Zymoseptoria tritici*. Mol Plant Pathol. 2015;16(9):931–45.

102. Bradshaw RE, Guo Y, Sim AD, Kabir MS, Chettri P, Ozturk IK, et al. Genome-wide gene expression dynamics of the fungal pathogen *Dothistroma septosporum* throughout its infection cycle of the gymnosperm host *Pinus radiata*. Mol Plant Pathol. 2015;17(2):210–24.

103. Yajima W, Kav NN V. The proteome of the phytopathogenic fungus *Sclerotinia sclerotiorum*. Proteomics. 2006;6(22):5995–6007.

104. 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.

105. Wang C-I a, Guncar G, Forwood JK, Teh T, Catanzariti A-M, Lawrence GJ, et al. Crystal structures of flax rust avirulence proteins AvrL567-A and -D reveal details of the structural basis for flax disease resistance specificity. Plant Cell. 2007;19(9):2898–912.

106. 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(10):e1005228.

107. Metzker ML. Emerging technologies in DNA sequencing. Genome Res. 2005;15(12):1767–76.

108. Parra G, Bradnam K, Korf I. CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics. 2007 May 1;23(9):1061–7.

109. Parra G, Bradnam K, Ning Z, Keane T, Korf I. Assessing the gene space in draft genomes. Nucleic Acids Res. 2009;37(1):289–97.

110. Simão FA, Waterhouse RM, Ioannidis P, Kriventseva E V., Zdobnov EM. BUSCO: Assessing genome assembly and annotation completeness with single-copy orthologs. Bioinformatics. 2015;31(19):3210–2.

111. Hane JK, Williams AH, Taranto AP, Solomon PS, Oliver RP. Genetic Transformation Systems in Fungi, Volume 2. van den Berg MA, Maruthachalam K, editors. Fungal Biol. Cham; 2015;2:55–68.

112. Ross MG, Russ C, Costello M, Hollinger A, Lennon NJ, Hegarty R, et al. Characterizing and measuring bias in sequence data. Genome Biol. 2013;14(5):R51.

113. Thomma BPHJ, Seidl MF, Shi-Kunne X, Cook DE, Bolton MD, van Kan J a. L, et al. Mind the gap; seven reasons to close fragmented genome assemblies. Fungal Genet Biol. 2015;

114. Salmela L, Rivals E. LoRDEC: accurate and efficient long read error correction. Bioinformatics. 2014;30(24):3506–14.

115. Berlin K, Koren S, Chin C-S, Drake JP, Landolin JM, Phillippy AM. Assembling large genomes

with single-molecule sequencing and locality-sensitive hashing. Nat Biotechnol. 2015;33(6).

116. Thomma BPHJ. Single-Molecule Real-Time Sequencing Combined with Optical Mapping Yields Completely Finished Fungal Genome. MBio. 2015;6(4):1–11.

117. Birney E, Clamp M, Durbin R. GeneWise and Genomewise. Genome Res. 2004;14(5):988–95.

118. Haas BJ, Zeng Q, Pearson MD, Cuomo CA, Jennifer R, Wortman R. Approaches to Fungal Genome Annotation. Mycology. 2011;2(3):118–41.

119. Grandaubert J, Bhattacharyya A, Stukenbrock EH. RNA-seq Based Gene Annotation and Comparative Genomics of Four Fungal Grass Pathogens in the Genus *Zymoseptoria* Identify Novel Orphan Genes and Species-Specific Invasions of Transposable Elements. G3. 2015;5(7):1323–33.

120. Bruno VM, Wang Z, Marjani SL, Euskirchen GM, Martin J, Sherlock G, et al. Comprehensive annotation of the transcriptome of the human fungal pathogen *Candida albicans* using RNA-seq. Genome Res. 2010;20(10):1451–8.

121. Cerqueira GC, Arnaud MB, Inglis DO, Skrzypek MS, Binkley G, Simison M, et al. The *Aspergillus* Genome Database: multispecies curation and incorporation of RNA-Seq data to improve structural gene annotations. Nucleic Acids Res. 2014;42(1):D705–10.

122. Trapnell C, Pachter L, Salzberg SL. TopHat: discovering splice junctions with RNA-Seq. Bioinformatics. 2009;25(9):1105–11.

123. 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.

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

125. Lukashin A V, Borodovsky M. GeneMark. hmm: new solutions for gene finding. Nucleic Acids Res. 1998;26(4):1107–15.

126. Korf I. Gene finding in novel genomes. BMC Bioinformatics. 2004;5:59.

127. Lomsadze A, Ter-Hovhannisyan V, Chernoff YO, Borodovsky M. Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res. 2005;33(20):6494–506.

128. 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.

129. Reid I, O'Toole N, Zabaneh O, Nourzadeh R, Dahdouli M, Abdellateef M, et al. SnowyOwl: accurate prediction of fungal genes by using RNA-Seq and homology information to select among ab initio models. BMC Bioinformatics. 2014;15(1):229.

130. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, et al. MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res.

2008;18(1):188–96.

131. Holt C, Yandell M. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics. BioMed Central Ltd; 2011;12(1):491.

132. Hof K, Lange S, Lomsadze A, Borodovsky M, Stanke M. BRAKER1 : Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. Bioinformatics. 2014;24(2013):2014.

133. Stanke M, Diekhans M, Baertsch R, Haussler D. Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics. 2008;24(5):637–44.

134. Silverstein KAT, Moskal WA, Wu HC, Underwood BA, Graham MA, Town CD, et al. Small cysteine-rich peptides resembling antimicrobial peptides have been under-predicted in plants. Plant J. 2007;51(2):262–80.

135. Brown NA, Antoniw J, Hammond-Kosack KE. The predicted secretome of the plant pathogenic fungus Fusarium graminearum: a refined comparative analysis. PLoS One. 2012;7(4):e33731.

136. Oliver R. Genomic tillage and the harvest of fungal phytopathogens. New Phytol. 2012;196:1015–23.

137. Wolpert TJ, Dunkle LD, Ciuffetti LM. Host-selective toxins and avirulence determinants: What's in a name? Annu Rev Phytopathol. 2002;40(1):251–85.

138. Baker SE, Kroken S, Inderbitzin P, Asvarak T, Li B-Y, Shi L, et al. Two polyketide synthase-encoding genes are required for biosynthesis of the polyketide virulence factor, T-toxin, by *Cochliobolus heterostrophus.* Mol plant-microbe Interact. 2006;19(2):139–49.

139. Yun S-H, Turgeon BG, Yoder OC. REMI-induced mutants of *Mycosphaerella zeae-maydis* lacking the polyketide PM-toxin are deficient in pathogenesis to corn. Physiol Mol Plant Pathol. 1998;52:53–66.

140. Walton JD. HC-toxin. Phytochemistry. 2006;67(14):1406–13.

141. Jr WD, Verlag B, Wolpert TJ, Acklin W, Jaun B, Seibl J, et al. Characterization of victorin C, the major host-selective toxin from *Cochliobolus victoriae*: structure of degradation products. Experientia. 1977;41:1366–70.

142. Nakashima T, Ueno T, Fukami H, Taga T, Masuda H, Osaki K, et al. Isolation and structures of AK-toxin I and II. Host-specific phytotoxic metabolites produced by *Alternaria alternata* Japanese pear pathotype. Agric Biol Chem. 1985;49(3):807–15.

143. Nakatsuka S, Ueda K, Goto T, Yamamoto M, Nishimura S, Kohmoto K. Structure of AF-toxin II, one of the host-specific toxins produced by strawberry pathotype. Tetrahedron Lett. 1986;27(24):2753–6.

144. Okuno T, Ishita Y, Matsumoto T. Characterization of Alternariolide, a host-specific toxin produced by *Alternaria mali roberts*. Chem Lett. 1974;3:635–8.

145. Spassieva SD, Markham JE, Hille J. The plant disease resistance gene Asc-1 prevents

disruption of sphingolipid metabolism during AAL-toxin-induced programmed cell death. Plant J. 2002;32:561–72.

146. Gardner JM, Kono Y, Tatum JH, Suzuki Y, Takeuchi S. Structure of the major component of ACRL-toxins, host-specific pathotoxic compounds produced by *Alternaria citri*. Agric Biol Chem. 1985;49(4):1235–8.

147. Lu S, Faris JD, Sherwood R, Friesen TL, Edwards MC. A dimeric PR-1-type pathogenesis-related protein interacts with ToxA and potentially mediates ToxA-induced necrosis in sensitive wheat. Mol Plant Pathol. 2014;15(7):650–63.

148. Ciuffetti LM, Tuori RP, Gaventa JM. A single gene encodes a selective toxin causal to the development of tan spot of wheat. Plant Cell. 1997;9(2):135–44.

149. Martinez JP, Ottum SA, Ali S, Franci LJ, Ciuffetti LM. Characterization of the ToxB gene from *Pyrenophora tritici-repentis*. Mol plant-microbe Interact. 2001;14(5):675–7.

150. Staats M, van Baarlen P, Schouten A, van Kan JAL, Bakker FT. Positive selection in phytotoxic protein-encoding genes of *Botrytis* species. Fungal Genet Biol. 2007;44(1):52–63.

151. Rohe M, Gierlich A, Hermann H, Hahn M, Schmidt B, Rosahl S, et al. The race-specific elicitor, NIP1, from the barley pathogen, *Rhynchosporium secalis*, determines avirulence on host plants of the *Rrs1* resistance genotype. EMBO J. 1995;14(17):4168–77.

152. Friesen TL, Faris JD, Solomon PS, Oliver RP. Host-specific toxins: effectors of necrotrophic pathogenicity. Cell Microbiol. 2008;10(7):1421–8.

153. Vincent D, Du Fall LA, Livk A, Mathesius U, Lipscombe RJ, Oliver RP, et al. A functional genomics approach to dissect the mode of action of the *Stagonospora nodorum* effector protein SnToxA in wheat. Mol Plant Pathol. 2012 Jun;13(5):467–82.

154. Lorang J, Kidarsa T, Bradford CS, Gilbert B, Curtis M, Tzeng S-C, et al. Tricking the guard: exploiting plant defense for disease susceptibility. Science (80-). 2012;338(6107):659–62.

155. Stergiopoulos I, Collemare J, Mehrabi R, De Wit PJGM. Phytotoxic secondary metabolites and peptides produced by plant pathogenic Dothideomycete fungi. FEMS Microbiol Rev. 2013;37(1):67–93.

156. Nakashima T, Ueno T, Fukami H. Structure elucidation of AK-toxins, host-specific phytotoxic metabolites produced by *Alternaria kikuchiana tanaka*. Tetrahedron Lett. 1996;23(43):4469–72.

157. Tanaka A, Tsuge T. Structural and functional complexity of the genomic region controlling AK-toxin biosynthesis and pathogenicity in the Japanese pear pathotype of *Alternaria alternata*. Mol plant-microbe Interact. 2000;13(9):975–86.

158. Tsuge T, Harimoto Y, Akimitsu K, Ohtani K, Kodama M, Akagi Y, et al. Host-selective toxins produced by the plant pathogenic fungus *Alternaria alternata*. FEMS Microbiol Rev. 2013;37(1):44–66.

159. Mayama S, Bordin A, Morikawa T, Tanpo H, Kato H. Association of avenalumin accumulation with co-segregation of victorin sensitivity and crown rust resistance in oat lines carrying the

Pc-2 gene. Vol. 46, Physiological and Molecular Plant Pathology. 1995. p. 263–74.

160. Lorang JM, Carkaci-Salli N, Wolpert TJ. Identification and characterization of victorin sensitivity in *Arabidopsis thaliana*. Mol Plant Microbe Interact. 2004;17(6):577–82.

161. Lorang JM, Sweat TA, Wolpert TJ. Plant disease susceptibility conferred by a "resistance" gene. Proc Natl Acad Sci. 2007;104(37):11861–6.

162. Johal GS, Briggs SP. Reductase activity encoded by the *HM1* disease resistance gene in maize. Science (80-). 1992;258(5084):985–7.

163. Braun CJ, Siedow JN, Levings CS. Funga1 Toxins Bind to the URF13 Protein in Maize Mitochondria and *Escherichia coli.* Society. 1990;2(February):153–61.

164. Levings III CS, Rhoads DM, Siedow JN. Molecular interactions of *Bipolaris maydis* T-toxin and maize. Can J Bot. 1995;73(S1):483–9.

165. Effertz RJ, Meinhardt SW, Anderson JA, Jordahl JG, Francl LJ. Identification of a Chlorosis-Inducing Toxin from *Pyrenophora tritici-repentis* and the Chromosomal Location of an Insensitivity Locus in Wheat. Phytopathology. 2002;92(5):527–33.

166. Manning VA, Ciuffetti LM. Localization of Ptr ToxA Produced by *Pyrenophora tritici-repentis* Reveals Protein Import into Wheat Mesophyll Cells. Plant Cell. 2005;17:3203–12.

167. Tai Y-S, Bragg J, Meinhardt SW. Functional Characterization of ToxA and Molecular Identification of its Intracellular Targeting Protein in Wheat. Am J Plant Physiol. 2007;2(2):76–89.

168. Manning VA, Hardison LK, Ciuffetti LM. Ptr ToxA interacts with a chloroplast-localized protein. Mol plant-microbe Interact. 2007;20(2):168–77.

169. Sarma GN, Manning VA, Ciuffetti LM, Karplus PA. Structure of Ptr ToxA: an RGD-containing host-selective toxin from *Pyrenophora tritici-repentis*. Plant Cell. 2005;17(11):3190–202.

170. Moffat CS, See PT, Oliver RP. Generation of a *ToxA* knockout strain of the wheat tan spot pathogen *Pyrenophora tritici-repentis*. Mol Plant Pathol. 2014;15(9):918–26.

171. Manning VA, Ciuffetti LM. Necrotrophic Effector Epistasis in the *Pyrenophora tritici-repentis*-Wheat Interaction. PLoS One. 2015;10(4):e0123548.

172. Antoni EA, Rybak K, Tucker MP, Hane JK, Solomon PS, Drenth A, et al. Ubiquity of ToxA and absence of ToxB in Australian populations of *Pyrenophora tritici-repentis*. Australas Plant Pathol. 2010;39(1):63–8.

173. Martinez JP, Oesch NW, Ciuffetti LM. Characterization of the multiple-copy host-selective toxin gene, *ToxB*, in pathogenic and nonpathogenic isolates of *Pyrenophora tritici-repentis*. Mol plant-microbe Interact. 2004;17(5):467–74.

174. Du Fall LA, Solomon PS. The necrotrophic effector SnToxA induces the synthesis of a novel phytoalexin in wheat. New Phytol. 2013;200(1):185–200.

175. Winterberg B, Du Fall LA, Song X, Pascovici D, Care N, Molloy M, et al. The necrotrophic effector protein SnTox3 re-programs metabolism and elicits a strong defence response in susceptible wheat leaves. BMC Plant Biol. 2014;14(1):215.

176. Tan K-C, Phan HP, 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(July):1–9.

177. Sarpeleh A, Tate ME, Wallwork H, Catcheside D, Able AJ. Characterisation of low molecular weight phytotoxins isolated from *Pyrenophora teres*. Physiol Mol Plant Pathol. 2009;73(6):154–62.

178. Sarpeleh A, Wallwork H, Tate ME, Catcheside DE a, Able AJ. Initial characterisation of phytotoxic proteins isolated from *Pyrenophora teres*. Physiol Mol Plant Pathol. 2008;72(1-3):73–9.

179. Sarpeleh A, Wallwork H, Catcheside DE, Tate ME, Able AJ. Proteinaceous Metabolites from *Pyrenophora teres* Contribute to Symptom Development of Barley Net Blotch. Phytopathology. 2007;97(8):907–15.

180. Wittenberg AHJ, van der Lee TAJ, Ben M'barek S, Ware SB, Goodwin SB, Kilian A, et al. Meiosis drives extraordinary genome plasticity in the haploid fungal plant pathogen *Mycosphaerella graminicola*. PLoS One. 2009;4(6):e5863.

181. Rudd JJ, Antoniw J, Marshall R, Motteram J, Fraaije B, Hammond-Kosack K. Identification and characterisation of *Mycosphaerella graminicola* secreted or surface-associated proteins with variable intragenic coding repeats. Fungal Genet Biol. 2010;47(1):19–32.

182. Dean R, Kan J Van, Pretorius ZA, Hammond-Kosack KE, Di Pietro A, Spanu PD, et al. The Top 10 fungal pathogens in molecular plant pathology. Mol Plant Pathol. 2012;13(4):414–30.

183. Testa A, Oliver R, Hane J. Overview of genomic and bioinformatic resources for *Zymoseptoria tritici.* Fungal Genet Biol. 2015;79:13–6.

184. Schürch S, Linde CC, Knogge W, Jackson LF, McDonald BA. Molecular population genetic analysis differentiates two virulence mechanisms of the fungal avirulence gene *NIP1*. Mol plant-microbe Interact. 2004;17(10):1114–25.

185. Aboukhaddour R, Kim YM, Strelkov SE. RNA-mediated gene silencing of *ToxB* in *Pyrenophora tritici-repentis*. Mol Plant Pathol. 2012;13(3):318–26.

186. Amaike S, Ozga JA, Basu U, Strelkov SE. Quantification of *ToxB* gene expression and formation of appressoria by isolates of *Pyrenophora tritici-repentis* differing in pathogenicity. Plant Pathol. 2008;57(4):623–33.

187. McDonald MC, Oliver RP, Friesen TL, Brunner PC, McDonald B a. Global diversity and distribution of three necrotrophic effectors in *Phaeosphaeria nodorum* and related species. New Phytol. 2013;199(1):241–51.

188. Deller S, Hammond-Kosack KE, Rudd JJ. The complex interactions between host immunity and non-biotrophic fungal pathogens of wheat leaves. J Plant Physiol. 2011;168(1):63–71.

189. Inderbitzin P, Asvarak T, Turgeon BG. Six new genes required for production of T-toxin, a polyketide determinant of high virulence of *Cochliobolus heterostrophus* to maize. Mol plant-microbe Interact. 2010;23(4):458–72.

190. Stukenbrock EH, McDonald BA. Geographical variation and positive diversifying selection in the host-specific toxin SnToxA. Mol Plant Pathol. 2007;8(3):321–32.

191. Tan K-C, Ferguson-Hunt M, Rybak K, Waters ODC, Stanley W a, Bond CS, et al. Quantitative variation in effector activity of *ToxA* isoforms from *Stagonospora nodorum* and *Pyrenophora tritici-repentis*. Mol plant-microbe Interact. 2012;25(4):515–22.

192. Oliver RP, Solomon PS. Recent Fungal Diseases of Crop Plants. Am Phytopathol Soc. 2008;21(3):287–93.

193. Nyarko A, Singarapu KK, Figueroa M, Manning VA, Pandelova I, Wolpert TJ, et al. Solution NMR structures of *Pyrenophora tritici-repentis* ToxB and its inactive homolog reveal potential determinants of toxin activity. J Biol Chem. 2014;289(37):25946–56.

194. Oliver R, Lichtenzveig J, Tan K-C, Waters O, Rybak K, Lawrence J, et al. Absence of detectable yield penalty associated with insensitivity to Pleosporales necrotrophic effectors in wheat grown in the West Australian wheat belt. Plant Pathol. 2014;63(5):1027–32.

195. Tan K-C, Waters ODC, Rybak K, Antoni E, Furuki E, Oliver RP. Sensitivity to three *Parastagonospora nodorum* necrotrophic effectors in current Australian wheat cultivars and the presence of further fungal effectors. Crop Pasture Sci. 2014;65(2):150.

196. Chen C, Lian B, Hu J, Zhai H, Wang X. Genome comparison of two *Magnaporthe oryzae* field isolates reveals genome variations and potential virulence effectors. BMC Genomics. 2013;

197. Do JH, Choi D-K, Journal T, Society TM, Vol K. Computational approaches to gene prediction. J Microbiol. 2006;44(2):137–44.

198. 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;16:170.

199. (NCBI Resource Coordinators). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2014;43:6–17.

200. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics. 2014;30(15):2114–20.

201. 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.

202. Trapnell C, Williams BA, Pertea G, Mortazavi A, Kwan G, van Baren MJ, et al. Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. Nat Biotechnol. 2010;28(5):511–5.

203. Smit A, Hubley R. RepeatModeler Open-1.0 [Internet]. Available from: http://www.repeatmasker.org

204. Smit A. RepeatMasker Open-4.0 [Internet]. Available from: http://www.repeatmasker.org

205. Ter-Hovhannisyan V. Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Announc [Internet]. 2008 [cited 2012 Aug 8];1979–90.

Available from: http://gb.cw.com.tw/m2m-0000/genome.cshlp.org/content/18/12/1979.full

206. Stanke M, Waack S. Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics. 2003;19(Suppl 2):ii215–25.

207. Lomsadze A, Burns PD, Borodovsky M. Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. Nucleic Acids Res. 2014;1–8.

208. Petersen TN, Brunak S, von Heijne G, Nielsen H. SignalP 4.0: discriminating signal peptides from transmembrane regions. Nat Methods. 2011;8(10):785–6.

209. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. BMC Bioinformatics. 2009;10:421.

210. Eddy SR. Accelerated Profile HMM Searches. PLOS Comput Biol. 2011;7(10):e1002195.

211. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. Bioinformatics. 2010;26(6):841–2.

212. 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:bas003.

213. Jurka J, Kapitonov V V, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J. Repbase Update, a database of eukaryotic repetitive elements. Cytogenet Genome Res. 2005;110(1-4):462–7.

214. Wang L, Jiang N, Wang L, Fang O, Leach LJ, Hu X, et al. 3' Untranslated Regions Mediate Transcriptional Interference between Convergent Genes Both Locally and Ectopically in *Saccharomyces cerevisiae.* Zhang J, editor. PLOS Genet. 2014;10(1):e1004021.

215. Guida A, Lindstädt C, Maguire SL, Ding C, Higgins DG, Corton NJ, et al. Using RNA-seq to determine the transcriptional landscape and the hypoxic response of the pathogenic yeast *Candida parapsilosis*. BMC Genomics. 2011;12(1):628.

216. Kämper J, Kahmann R, Bölker M, Ma L-J, Brefort T, Saville BJ, et al. Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. Nature. 2006;444(7115):97–101.

217. Laurie JD, Ali S, Linning R, Mannhaupt G, Wong P, Güldener U, et al. Genome comparison of barley and maize smut fungi reveals targeted loss of RNA silencing components and species-specific presence of transposable elements. Plant Cell. 2012;24(5):1733–45.

218. Ma L, Does H Van Der, Borkovich K. Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium.* Nature. 2010;464(7287):367–73.

219. van Esse HP, Bolton MD, Stergiopoulos I, de Wit PJGM, Thomma BPHJ. The chitin-binding *Cladosporium fulvum* effector protein Avr4 is a virulence factor. Mol plant-microbe Interact. 2007;20(9):1092–101.

220. de Jonge R, Thomma BPHJ. Fungal LysM effectors: extinguishers of host immunity? Trends Microbiol. 2009;17(4):151–7.

221. Hane JK, Anderson JP, Williams AH, Sperschneider J, Singh KB. Genome sequencing and comparative genomics of the broad host-range pathogen *Rhizoctonia solani* AG8. PLOS Genet. 2014;10(5):e1004281.

222. 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. Frontiers; 2015;6.

223. Rau D, Attene G, Brown AH, Nanni L, Maier FJ, Balmas V, et al. Phylogeny and evolution of mating-type genes from *Pyrenophora teres*, the causal agent of barley “net blotch” disease. Curr Genet. 2007;51(6):377–92.

224. Soliai MM, Meyer SE, Udall J a, Elzinga DE, Hermansen R a, Bodily PM, et al. De novo genome assembly of the fungal plant pathogen *Pyrenophora semeniperda*. PLoS One. 2014;9(1):e87045.

225. Zhao C, Waalwijk C, de Wit PJGM, Tang D, van der Lee T. RNA-Seq analysis reveals new gene models and alternative splicing in the fungal pathogen *Fusarium graminearum*. BMC Genomics. 2013;14(1):21.

226. Michiels A, Van den Ende W, Tucker M, Van Riet L, Van Laere A. Extraction of high-quality genomic DNA from latex-containing plants. Anal Biochem. 2003;315(1):85–9.

227. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. Nat Methods. 2012;9(4):357–9.

228. 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.

229. Jiang N. Repeat Library Construction - Advanced [Internet]. MAKER Wiki. 2014 [cited 2015 Nov 2]. Available from: http://weatherby.genetics.utah.edu/MAKER/wiki/index.php/Repeat_Library_Construction-Advanced

230. Santana MF, Silva JCF, Mizubuti ESG, Araújo EF, Condon BJ, Turgeon BG, et al. Characterization and potential evolutionary impact of transposable elements in the genome of *Cochliobolus heterostrophus*. BMC Genomics. 2014;15(1):536.

231. Han Y, Wessler SR. MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. Nucleic Acids Res. 2010;38(22):e199–e199.

232. Ellinghaus D, Kurtz S, Willhoeft U. LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. BMC Bioinformatics. 2008;9:18.

233. Steinbiss S, Willhoeft U, Gremme G, Kurtz S. Fine-grained annotation and classification of de novo predicted LTR retrotransposons. Nucleic Acids Res. 2009;37(21):7002–13.

234. Chan PP, Lowe TM. GtRNAdb: a database of transfer RNA genes detected in genomic sequence. Nucleic Acids Res. 2009;37(Database):D93–7.

235. Price AL, Jones NC, Pevzner PA. De novo identification of repeat families in large genomes. Bioinformatics. 2005;21(Suppl. 1):i351–8.

236. Krogh A, Larsson B, von Heijne G, Sonnhammer ELL. Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J Mol Biol. 2001;305(3):567–80.

237. Hoff KJ, Stanke M. WebAUGUSTUS--a web service for training AUGUSTUS and predicting genes in eukaryotes. Nucleic Acids Res. 2013;41(Web Server issue):W123–8.