School of Science

Comparative Genomics of *Parastagonospora* and *Pyrenophora* species

Robert Andrew Syme

This thesis is presented for the Degree of

Doctor of Philosophy

of

Curtin University

May 2015

Declaration

To the best of my knowledge and belief this thesis contains no material previously published by any other person except where due acknowledgment has been made.

This thesis contains no material which has been accepted for the award of any other degree or diploma at any university

Signature: ……………………………………………

Date: 2015-05-19

Table of Contents

Chapter 1 | Introduction

Rise of Non-Model Organisms

Early efforts in the field of molecular biology were focused on the general properties of genetics and molecular systems. The organisms used as subjects of study were often chosen because of logistical or technical characteristics such as genetic tractability, phenotypic stability and low generation times. Establishing key molecular techniques in the 1970s and 80s such as transformation in a handful of model organisms such as *Saccharomyces cerevisiae* (Beggs 1978, Hinnen, Hicks et al. 1978), *Neurospora crassa* (Case, Schweizer et al. 1979), *Schizosaccharomyces pombe* (Beach and Nurse 1981), and *Aspergillus nidulans* (Ballance, Buxton et al. 1983, Tilburn, Scazzocchio et al. 1983) were important breakthroughs. However, genome characteristics that made these species appealing as a laboratory workhorses - such as the replication origin sequences and small centromeres in *S. cerevisiae*, were often difficult points of difference when it came to translating knowledge and techniques to other organisms.

Gradual decreases in the cost of nucleotide sequencing and generalisation of molecular techniques allowed a shift towards biological research that favoured investigation of representative organisms much closer to the specific species of interest (if not the direct species itself), rather than a model organism proxy. Traditional model organisms have not been supplanted in this transition and the resources invested in these systems continue to pay dividends, but a class of 'neo-model' organisms has also developed. Small research groups with limited resources can now apply genomic, transcriptomic and/or proteomic techniques to previously non-model organisms. This pattern of expanded academic focus is broadly observed in fungi (Dujon, Sherman et al. 2004, Loftus, Fung et al. 2005, Hane, Lowe et al. 2007), invertebrates (Zhang, Fang et al. 2012, Simakov, Marletaz et al. 2013), plants (Bennetzen, Schmutz et al. 2012, Brenchley, Spannagl et al. 2012, Consortium 2012, Guo, Zhang et al. 2013), insects (Consortium 2012, You, Yue et al. 2013), fish(Jones, Grabherr et al. 2012), birds (Ellegren, Smeds et al. 2012, Shapiro, Kronenberg et al. 2013), primates (Prüfer, Munch et al. 2012) and other mammals (Qiu, Zhang et al. 2012, Zhang, Cowled et al. 2013). Fungi have been particularly attractive targets because many species are amenable to laboratory handling, are causative agents of disease (Brown, Denning et al. 2012, Fisher, Henk et al. 2012), and have genomes with smaller sizes and reduced complexity (Noble and Andrianopoulos 2013) compared to most higher eukaryotes.

The rise in 'neo-model' organisms has been particularly valuable in systems where important genetic mechanisms are not well conserved across taxa or are idiosyncratic to one or a few species.

Consequently, many exciting questions about population structure, host specific pathology, or the evolutionary relationship between host and pathogen can now be answered through the direct study of a species of interest.

Effector Biology and Discovery

Early investigations into the genetic components of bacterial pathogenicity revealed that a single locus of an avirulent *Pseudomonas syringae* strain was sufficient to convert a virulent strain into an avirulent strain (Staskawicz, Dahlbeck et al. 1984). Inoculation of strains that produced this 'avirulence factor' would trigger a hypersensitive response in some soybean cultivars. It would be discovered later that the production of an 'avirulence factor' could *increase* virulence when infecting a host cultivar lacking a corresponding resistance (R) gene. This has come to be termed a 'gene-for-gene' interaction (Flor 1942). The term 'effector' was borrowed from medical literature (Birnbaumer 1992), and is also now widely used (Figure 1) to avoid the potential confusion caused by using the term 'avirulence gene' to describe an element that is simultaneously a factor that promotes virulence in some hosts and avirulence in others. The more general term also allows inclusion of a large class of genes with products that have huge diversity of function – an effector can be defined as any molecule produced by the pathogen that has an effect on its interaction with the host.

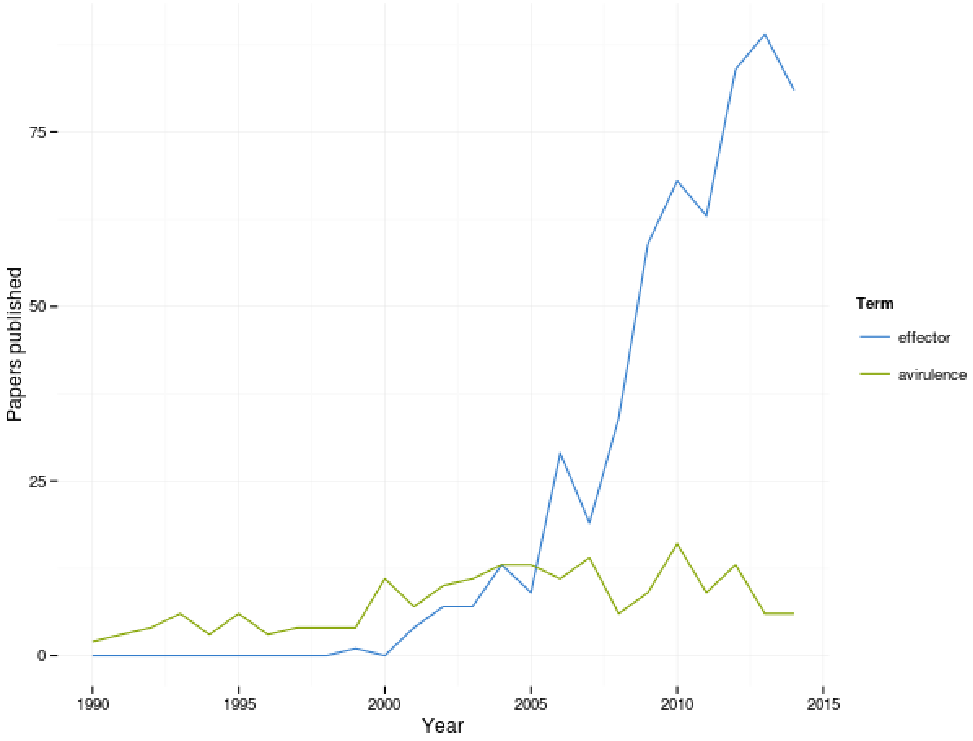


Figure 1: Usage of the terms "effector" and "avirulence" in titles and abstracts from MPMI, PLoS Pathogens, and Fungal Genetics and Biology from 1990 to 2015

The knowledge that a pathogen's effector must be externalised and then mobilised into host cells was used as a basis to uncover bacterial effectors. Early discoveries were made during investigation of protein products delivered via the bacterial type III secretion system (T3SS) (Abramovitch, Anderson et al. 2006). Proteins utilising T3SSs modulate host defences and/or induce disease responses via diverse enzymatic processes, including ubiquitin-like protease (Roden, Eardley et al. 2004), protein phosphatase (Espinosa, Guo et al. 2003), and cysteine protease (Shao, Merritt et al. 2002) activity. Bacterial research established a simple molecular basis for plant pathogenicity, but pathogen-derived molecules that modulate host defence mechanisms are certainly not limited to the prokaryotic kingdom. Effector-host interactions are an important part of many fungal and oomycete pathosystems.

Fungal Phytopathology

Successful fungal infection requires that the fungus first gain entry into the plant tissue. The stomata can provide an easy natural pore for fungal access, but more direct host cell penetration of leaf cuticles is possible via the specialised infections cells appressoria (Ryder and Talbot 2015). Once inside the cell, the fungus needs to negotiate a complex interaction with host defence mechanisms. Basal plant immune responses detect well conserved molecular features characteristic of pathogen presence, called pathogen-associated molecular patterns (PAMPs) or microbe-associated molecular patterns (MAMPs). A class of leucine-rich repeat (LRR) containing transmembrane receptor kinases and receptor-like proteins collectively called pattern recognition receptors (PRRs) detect MAMPs in the apoplastic space. Detection of MAMPs by a PRR can rapidly initiate basal defence responses including the production of reactive oxygen species (ROS) and compounds with antimicrobial function such as chitinases, proteinases and nonproteinaceous antimicrobial molecules (Macho and Zipfel 2014). PRRs can also be triggered by damage associated molecular patterns (DAMPs) (Albert 2013). Successful deployment of the MAMP-triggered defences is described as MAMP-triggered immunity (MTI) (Zipfel 2008), and is often a successful strategy for defence against pathogens without host-specific adaptations (Uma, Rani et al. 2011). Circumvention of MTI requires intervention by the pathogen. The type of intervention can be grouped into a number of broad strategies. Effectors can also act to mask the pathogen's presence so as to avoid or reduce the MTI response. The Slp1 effector from *Magnaporthe oryzae* binds the MAMP chitin and reduces its availability for detection by the host (Mentlak, Kombrink et al. 2012). If the pathogen is unable to prevent MTI, effectors can help guard against the basal defence mechanisms. The biotroph *Cladosporium fulvum* proteins Avr4 and Avr2 do so by providing protection from host chitinases and proteases, respectively (van den Burg, Harrison et al. 2006, van Esse, van't Klooster et al. 2008). Effectors can supress host immune response, such as the *Ustilago maydis* effector Cmu1. Chorismate is a metabolite required for the production of salicylic acid (SA), a key phytohormone used to induce cell death. Cmu1 is a chorismate mutase, catalysing the conversion of chorismate to prephenate and in doing so, redirects this important SA synthesis metabolic intermediate towards an alternative biochemical fate (Djamei, Schipper et al. 2011). The effector Pep1, again in *Ustilago maydis*, also acts to supress host immune response by inhibition of POX12 in maize, which is required for the generation of reactive oxygen species.

When MTI is overcome by fungal effectors, the plant requires a new strategy to develop resistance. Receptors for specific effector recognition rather than the broadly conserved patterns of microbial presence compose an effective second tier to the plant's immune surveillance. Detection of a biotrophic fungal effector allows the plant to mount a hypersensitive response resulting in localised cell death which inhibits the growth of biotrophic pathogens. The resistance conferred by the presence of an effector/receptor pairing is described as effector-triggered immunity (ETI).

Inverse Gene-for-Gene Effectors

In contrast to the gene-for-gene system observed in biotrophic pathogens, many necrotrophs produce effectors that do not function to evade or interrupt the MTI response, but instead operate to trigger an apoptotic response in the host (Lorang, Kidarsa et al. 2012). In an 'inverse gene-for-gene' system, infection requires the production of an effector by the pathogen and the presence of a corresponding susceptibility allele in the host.

Effector delivery to the host can occur by secretion into the apoplastic space between host cells or by translocation into the host cell where they can interact directly with intracellular components (Birch, Rehmany et al. 2006). Many oomycete effectors include a well conserved RXLR domain within 60 amino acids of the protein N-terminus. The domain is not required for effector activity (Bos, Kanneganti et al. 2006), but it mediates the translocation of the effector into the host cell (Kale 2012). Just as T3SS was used as the key to uncover effectors in bacterial systems, the RXLR motif was used to identify oomycete effectors. High levels of RXLR motif conservation amongst the oomycetes allowed for genome-wide searches that return hundreds of RXLR effectors in a single genome (Tyler, Tripathy et al. 2006).

Effectors of fungi also interact with intracellular components in the host, but no domain as clear and conserved as RXLR has been identified. Functional RXLR variants that bound components of the plant cell plasma membrane were observed in some fungal effectors (Kale 2012), but the diversity even in confirmed RXLR-like domains was too high to be useful for screening whole genomes for candidates. As fungal effectors were uncovered (Baker, Kroken et al. 2006, Gout, Fudal et al. 2006, Walton 2006, Staats, van Baarlen et al. 2007, Liu, Faris et al. 2009), it became apparent that they shared very little sequence identity and were very rarely homologous. Effector discovery strategies from oomycetes would need to be modified to be successful in fungal pathosystems. The common characteristics of some fungal effectors were their genomic context, population structure, and evolutionary history rather than the presence of conserved effector domains. Genes such as Avr1-9 in *L. maculans* (Gout, Fudal et al. 2006), and effectors *ToxA* (Friesen, Stukenbrock et al. 2006) and *Tox3* (Liu, Faris et al. 2009) in *P. nodorum* are all located near repetitive sequences. *NIP1* in *R. secalis* (Schürch, Linde et al. 2004), *Avr1-9* in *L. maculans* (Gout, Fudal et al. 2006), and NEP1-like effector in *B. cinerea* (Staats, van Baarlen et al. 2007) are under positive selection. It was characteristics such as these that became popular for fungal effector identification rather than homology or sequence conservation. Effector discovery has been greatly aided by genome sequencing of organisms that are not considered traditional model systems as it allows observation of the genomic context of potential effector genes.

A stark exception to this rule of effector identification by means other than sequence identity was the discovery of *ToxA* in *P. nodorum* (Friesen, Stukenbrock et al. 2006). *ToxA* was already established as an effector in *P. tritici-repentis*, and the possession of the full genome sequence of both allowed Friesen et al. to identify a near-perfect homolog in *P. nodorum*. Patterns of sequence diversity in populations of the two species suggests that *ToxA* gene and flanking sequence had been laterally transferred from *P. nodorum* to *P. tritici-repentis*. Subsequent instances of effector discovery in *P. nodorum* were made possible by the availability of the genome sequence. *Tox1* was identified by assessing each of the predicted proteins against a set of criteria common to effectors. Top candidates from this ranking system were assessed for activity via expression in a yeast expression system. *Tox3* was identified by matching proteomic fragments from an extracellular sample to the reference genome assembly. The recent rise in available genome sequences has expanded the opportunities for *in-silico* prediction of effector candidates. New techniques such as predictions that do not rely on a priori assumptions about the characteristics of effectors (Saunders, Win et al. 2012, Sperschneider, Gardiner et al. 2013) are made possible by the wealth of genome data available for comparison.

Dothideomycete Genomics

The Dothideomycetes are a large class of fungi that include many phytopathogens infecting a broad range of hosts including many economically important crops. The orders *Pleosporales* and *Capnodiales* contain the largest number of phytopathogens (Ohm, Feau et al. 2012), and are the most well studied and well sequenced clades.

The first published Dothideomycete genome was that of *P. nodorum* in 2007 (Hane, Lowe et al. 2007), establishing the organism as a research focus for necrotrophic fungal pathologens. The *Pyrenophora teres f. teres* genome was published in 2010 (Ellwood, Liu et al. 2010), followed by *Leptosphaeria maculans* (Rouxel, Grandaubert et al. 2011) and *Mycosphaerella graminicola* (Goodwin, M'Barek et al. 2011) in 2011. *Dothistroma septosporum* (de Wit, Van Der Burgt et al. 2012), *Cladosporium fulvum* (de Wit, Van Der Burgt et al. 2012), *Macrophomina phaseolina* (Islam, Haque et al. 2012), *Zymoseptoria ardabiliae* (Stukenbrock, Christiansen et al. 2012) and *Zymoseptoria pseudotritici* (Stukenbrock, Christiansen et al. 2012) were published in 2012 as well as a large-scale comparison of 18 Dothideomycete species accompanied by the release of 14 new genome sequences (Ohm, Feau et al. 2012). The Joint Genome Institute's "MycoCosm" fungal sequencing portal currently hosts 482 fungal genome sequencing projects.

At the commencement of this PhD, *P. nodorum* had a draft genome assembly for a single representative isolate, and neither *P. avenae* nor *Pyrenophora teres f. teres* had genomic resources available. The work presented in this thesis leverages the rapidly expanding array of techniques and technology available to researchers working in molecular biology to more fully understand the genome dynamics, evolutionary history and molecular arsenal of these agronomically important fungal species. The assembly of a *P. teres f. teres* genome and comparison assembly allowed the phylogenetic comparison with other *Pyrenophora* spp. (Chapter 3), and now 12 *P. avenaria* genome sequences are available for comparison (Chapter 6). The *P. nodorum* genomic infrastructure has grown to include multiple omics resources aggregated from multiple studies leading to a refined reference genome sequence, carefully annotated gene models, RNA-seq transcription data and 25 alternate genome assemblies (Chapters 2, 5, and 6). The advancing technology and tools available for the *P. nodorum* and *P. avenaria* pathosystems have provided opportunities for intra-species introspection as well as comparisons to other species in the Dothideomycetes. The comparisons have yielded insight into the evolutionary history of pathogen and host (Chapter 4), insights into the mechanisms of genome evolution (Chapter 6), and repeated rounds of effector prediction (Chapters 2 and 6).

References

Abramovitch, R. B., J. C. Anderson and G. B. Martin (2006). "Bacterial elicitation and evasion of plant innate immunity." Nature Reviews Molecular Cell Biology **7**(8): 601-611.

Albert, M. (2013). "Peptides as triggers of plant defence." Journal of experimental botany: ert275.

Baker, S. E., S. Kroken, P. Inderbitzin, T. Asvarak, B.-Y. Li, L. Shi, O. C. Yoder and B. G. Turgeon (2006). "Two polyketide synthase-encoding genes are required for biosynthesis of the polyketide virulence factor, T-toxin, by *Cochliobolus heterostrophus*." Molecular plant-microbe interactions **19**(2): 139-149.

Ballance, D., F. Buxton and G. Turner (1983). "Transformation of *Aspergillus nidulans* by the orotidine-5′-phosphate decarboxylase gene of *Neurospora crassa*." Biochemical and biophysical research communications **112**(1): 284-289.

Beach, D. and P. Nurse (1981). "High-frequency transformation of the fission yeast Schizosaccharomyces pombe."

Beggs, J. D. (1978). "Transformation of yeast by a replicating hybrid plasmid."

Bennetzen, J. L., J. Schmutz, H. Wang, R. Percifield, J. Hawkins, A. C. Pontaroli, M. Estep, L. Feng, J. N. Vaughn and J. Grimwood (2012). "Reference genome sequence of the model plant Setaria." Nature biotechnology **30**(6): 555-561.

Birch, P. R., A. P. Rehmany, L. Pritchard, S. Kamoun and J. L. Beynon (2006). "Trafficking arms: oomycete effectors enter host plant cells." Trends in microbiology **14**(1): 8-11.

Birnbaumer, L. (1992). "Receptor-to-effector signaling through G proteins: roles for βγ dimers as well as α subunits." Cell **71**(7): 1069-1072.

Bos, J. I., T. D. Kanneganti, C. Young, C. Cakir, E. Huitema, J. Win, M. R. Armstrong, P. R. Birch and S. Kamoun (2006). "The C‐terminal half of *Phytophthora infestans* RXLR effector AVR3a is sufficient to trigger R3a‐mediated hypersensitivity and suppress INF1‐induced cell death in *Nicotiana benthamiana*." The Plant Journal **48**(2): 165-176.

Brenchley, R., M. Spannagl, M. Pfeifer, G. L. Barker, R. D'Amore, A. M. Allen, N. McKenzie, M. Kramer, A. Kerhornou and D. Bolser (2012). "Analysis of the bread wheat genome using whole-genome shotgun sequencing." Nature **491**(7426): 705-710.

Brown, G. D., D. W. Denning, N. A. Gow, S. M. Levitz, M. G. Netea and T. C. White (2012). "Hidden killers: human fungal infections." Science translational medicine **4**(165): 165rv113-165rv113.

Case, M. E., M. Schweizer, S. R. Kushner and N. H. Giles (1979). "Efficient transformation of *Neurospora crassa* by utilizing hybrid plasmid DNA." Proceedings of the National Academy of Sciences **76**(10): 5259-5263.

Consortium, H. G. (2012). "Butterfly genome reveals promiscuous exchange of mimicry adaptations among species." Nature **487**(7405): 94-98.

Consortium, T. G. (2012). "The tomato genome sequence provides insights into fleshy fruit evolution." Nature **485**(7400): 635-641.

de Wit, P. J., A. Van Der Burgt, B. Ökmen, I. Stergiopoulos, K. A. Abd-Elsalam, A. L. Aerts, A. H. Bahkali, H. G. Beenen, P. Chettri and M. P. Cox (2012). "The genomes of the fungal plant pathogens *Cladosporium fulvum* and *Dothistroma septosporum* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry." PLoS genetics **8**(11): e1003088.

Djamei, A., K. Schipper, F. Rabe, A. Ghosh, V. Vincon, J. Kahnt, S. Osorio, T. Tohge, A. R. Fernie and I. Feussner (2011). "Metabolic priming by a secreted fungal effector." Nature **478**(7369): 395-398.

Dujon, B., D. Sherman, G. Fischer, P. Durrens, S. Casaregola, I. Lafontaine, J. De Montigny, C. Marck, C. Neuvéglise and E. Talla (2004). "Genome evolution in yeasts." Nature **430**(6995): 35-44.

Ellegren, H., L. Smeds, R. Burri, P. I. Olason, N. Backström, T. Kawakami, A. Künstner, H. Mäkinen, K. Nadachowska-Brzyska and A. Qvarnström (2012). "The genomic landscape of species divergence in Ficedula flycatchers." Nature **491**(7426): 756-760.

Ellwood, S. R., Z. Liu, R. A. Syme, Z. Lai, J. K. Hane, F. Keiper, C. S. Moffat, R. P. Oliver and T. L. Friesen (2010). "A first genome assembly of the barley fungal pathogen Pyrenophora teres f. teres." Genome Biol **11**(11): R109.

Espinosa, A., M. Guo, V. C. Tam, Z. Q. Fu and J. R. Alfano (2003). "The *Pseudomonas syringae* type III-secreted protein HopPtoD2 possesses protein tyrosine phosphatase activity and suppresses programmed cell death in plants." Molecular microbiology **49**(2): 377-387.

Fisher, M. C., D. A. Henk, C. J. Briggs, J. S. Brownstein, L. C. Madoff, S. L. McCraw and S. J. Gurr (2012). "Emerging fungal threats to animal, plant and ecosystem health." Nature **484**(7393): 186-194.

Flor, H. (1942). "Inheritance of pathogenicity in *Melampsora lini*." Phytopathology **32**(653): e69.

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

Goodwin, S. B., S. B. M'Barek, B. Dhillon, A. H. J. Wittenberg, C. F. Crane, J. K. Hane, A. J. Foster, T. A. J. van der Lee, J. Grimwood, A. Aerts, J. Antoniw, A. Bailey, B. Bluhm, J. Bowler, J. Bristow, A. van der Burgt, B. Canto-Canché, A. C. L. Churchill, L. Conde-Ferràez, H. J. Cools, P. M. Coutinho, M. Csukai, P. Dehal, P. de Wit, B. Donzelli, H. C. van de Geest, R. C. H. J. van Ham, K. E. Hammond-Kosack, B. Henrissat, A. Kilian, A. K. Kobayashi, E. Koopmann, Y. Kourmpetis, A. Kuzniar, E. Lindquist, V. Lombard, C. Maliepaard, N. Martins, R. Mehrabi, J. P. H. Nap, A. Ponomarenko, J. J. Rudd, A. Salamov, J. Schmutz, H. J. Schouten, H. Shapiro, I. Stergiopoulos, S. F. F. Torriani, H. Tu, R. P. de Vries, C. Waalwijk, S. B. Ware, A. Wiebenga, L. H. Zwiers, R. P. Oliver, I. V. Grigoriev and G. H. J. Kema (2011). "Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis." PLoS Genetics **7**(6).

Gout, L., I. Fudal, M. L. Kuhn, F. Blaise, M. Eckert, L. Cattolico, M. H. Balesdent and T. Rouxel (2006). "Lost in the middle of nowhere: the AvrLm1 avirulence gene of the Dothideomycete *Leptosphaeria maculans*." Molecular microbiology **60**(1): 67-80.

Guo, S., J. Zhang, H. Sun, J. Salse, W. J. Lucas, H. Zhang, Y. Zheng, L. Mao, Y. Ren and Z. Wang (2013). "The draft genome of watermelon (*Citrullus lanatus*) and resequencing of 20 diverse accessions." Nature genetics **45**(1): 51-58.

Hane, J. K., R. G. Lowe, P. S. Solomon, K.-C. Tan, C. L. Schoch, J. W. Spatafora, P. W. Crous, C. Kodira, B. W. Birren and J. E. Galagan (2007). "Dothideomycete–plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*." The Plant Cell Online **19**(11): 3347-3368.

Hinnen, A., J. B. Hicks and G. R. Fink (1978). "Transformation of yeast." Proceedings of the National Academy of Sciences **75**(4): 1929-1933.

Islam, M. S., M. S. Haque, M. M. Islam, E. M. Emdad, A. Halim, Q. M. M. Hossen, M. Z. Hossain, B. Ahmed, S. Rahim and M. S. Rahman (2012). "Tools to kill: genome of one of the most destructive plant pathogenic fungi Macrophomina phaseolina." BMC genomics **13**(1): 493.

Jones, F. C., M. G. Grabherr, Y. F. Chan, P. Russell, E. Mauceli, J. Johnson, R. Swofford, M. Pirun, M. C. Zody and S. White (2012). "The genomic basis of adaptive evolution in threespine sticklebacks." Nature **484**(7392): 55-61.

Kale, S. D. (2012). "Oomycete and fungal effector entry, a microbial Trojan horse." New Phytologist **193**(4): 874-881.

Liu, Z., J. D. Faris, R. P. Oliver, K. C. Tan, P. S. Solomon, M. C. McDonald, B. A. McDonald, A. Nunez, S. Lu, J. B. Rasmussen and T. L. Friesen (2009). "SnTox3 acts in effector triggered susceptibility to induce disease on wheat carrying the Snn3 gene." PLoS Pathogens **5**(9): e1000581.

Loftus, B. J., E. Fung, P. Roncaglia, D. Rowley, P. Amedeo, D. Bruno, J. Vamathevan, M. Miranda, I. J. Anderson and J. A. Fraser (2005). "The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*." Science **307**(5713): 1321-1324.

Lorang, J., T. Kidarsa, C. Bradford, B. Gilbert, M. Curtis, S.-C. Tzeng, C. Maier and T. Wolpert (2012). "Tricking the guard: exploiting plant defense for disease susceptibility." Science **338**(6107): 659-662.

Macho, A. P. and C. Zipfel (2014). "Plant PRRs and the activation of innate immune signaling." Molecular cell **54**(2): 263-272.

Mentlak, T. A., A. Kombrink, T. Shinya, L. S. Ryder, I. Otomo, H. Saitoh, R. Terauchi, Y. Nishizawa, N. Shibuya and B. P. Thomma (2012). "Effector-mediated suppression of chitin-triggered immunity by Magnaporthe oryzae is necessary for rice blast disease." The Plant Cell **24**(1): 322-335.

Noble, L. M. and A. Andrianopoulos (2013). "Fungal genes in context: genome architecture reflects regulatory complexity and function." Genome biology and evolution **5**(7): 1336-1352.

Ohm, R. A., N. Feau, B. Henrissat, C. L. Schoch, B. A. Horwitz, K. W. Barry, B. J. Condon, A. C. Copeland, B. Dhillon and F. Glaser (2012). "Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi." PLoS Pathogens **8**(12): e1003037.

Prüfer, K., K. Munch, I. Hellmann, K. Akagi, J. R. Miller, B. Walenz, S. Koren, G. Sutton, C. Kodira and R. Winer (2012). "The bonobo genome compared with the chimpanzee and human genomes." Nature **486**(7404): 527-531.

Qiu, Q., G. Zhang, T. Ma, W. Qian, J. Wang, Z. Ye, C. Cao, Q. Hu, J. Kim and D. M. Larkin (2012). "The yak genome and adaptation to life at high altitude." Nature genetics **44**(8): 946-949.

Roden, J., L. Eardley, A. Hotson, Y. Cao and M. B. Mudgett (2004). "Characterization of the *Xanthomonas* AvrXv4 effector, a SUMO protease translocated into plant cells." Molecular plant-microbe interactions **17**(6): 633-643.

Rouxel, T., J. Grandaubert, J. K. Hane, C. Hoede, A. P. van de Wouw, A. Couloux, V. Dominguez, V. Anthouard, P. Bally and S. Bourras (2011). "Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations." Nature Communications **2**: 202.

Ryder, L. S. and N. J. Talbot (2015). "Regulation of appressorium development in pathogenic fungi." Current opinion in plant biology **26**: 8-13.

Saunders, D. G., J. Win, L. M. Cano, L. J. Szabo, S. Kamoun and S. Raffaele (2012). "Using hierarchical clustering of secreted protein families to classify and rank candidate effectors of rust fungi." PLoS One **7**(1): e29847.

Schürch, S., C. C. Linde, W. Knogge, L. F. Jackson and B. A. McDonald (2004). "Molecular population genetic analysis differentiates two virulence mechanisms of the fungal avirulence gene NIP1." Molecular plant-microbe interactions **17**(10): 1114-1125.

Shao, F., P. M. Merritt, Z. Bao, R. W. Innes and J. E. Dixon (2002). "A Yersinia effector and a *Pseudomonas* avirulence protein define a family of cysteine proteases functioning in bacterial pathogenesis." Cell **109**(5): 575-588.

Shapiro, M. D., Z. Kronenberg, C. Li, E. T. Domyan, H. Pan, M. Campbell, H. Tan, C. D. Huff, H. Hu and A. I. Vickrey (2013). "Genomic diversity and evolution of the head crest in the rock pigeon." Science **339**(6123): 1063-1067.

Simakov, O., F. Marletaz, S.-J. Cho, E. Edsinger-Gonzales, P. Havlak, U. Hellsten, D.-H. Kuo, T. Larsson, J. Lv and D. Arendt (2013). "Insights into bilaterian evolution from three spiralian genomes." Nature **493**(7433): 526-531.

Sperschneider, J., D. M. Gardiner, J. M. Taylor, J. K. Hane, K. B. Singh and J. M. Manners (2013). "A comparative hidden Markov model analysis pipeline identifies proteins characteristic of cereal-infecting fungi." BMC genomics **14**(1): 807.

Staats, M., P. van Baarlen, A. Schouten, J. A. van Kan and F. T. Bakker (2007). "Positive selection in phytotoxic protein-encoding genes of *Botrytis* species." Fungal Genetics and Biology **44**(1): 52-63.

Staskawicz, B. J., D. Dahlbeck and N. T. Keen (1984). "Cloned avirulence gene of *Pseudomonas syringae* pv. glycinea determines race-specific incompatibility on Glycine max (L.) Merr." Proceedings of the National Academy of Sciences **81**(19): 6024-6028.

Stukenbrock, E. H., F. B. Christiansen, T. T. Hansen, J. Y. Dutheil and M. H. Schierup (2012). "Fusion of two divergent fungal individuals led to the recent emergence of a unique widespread pathogen species." Proceedings of the National Academy of Sciences **109**(27): 10954-10959.

Tilburn, J., C. Scazzocchio, G. G. Taylor, J. H. Zabicky-Zissman, R. A. Lockington and R. W. Davies (1983). "Transformation by integration in *Aspergillus nidulans*." Gene **26**(2): 205-221.

Tyler, B. M., S. Tripathy, X. Zhang, P. Dehal, R. H. Jiang, A. Aerts, F. D. Arredondo, L. Baxter, D. Bensasson and J. L. Beynon (2006). "*Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis." Science **313**(5791): 1261-1266.

Uma, B., T. S. Rani and A. R. Podile (2011). "Warriors at the gate that never sleep: non-host resistance in plants." Journal of plant physiology **168**(18): 2141-2152.

van den Burg, H. A., S. J. Harrison, M. H. Joosten, J. Vervoort and P. J. de Wit (2006). "Cladosporium fulvum Avr4 protects fungal cell walls against hydrolysis by plant chitinases accumulating during infection." Molecular Plant-Microbe Interactions **19**(12): 1420-1430.

van Esse, H. P., J. W. van't Klooster, M. D. Bolton, K. A. Yadeta, P. van Baarlen, S. Boeren, J. Vervoort, P. J. de Wit and B. P. Thomma (2008). "The Cladosporium fulvum virulence protein Avr2 inhibits host proteases required for basal defense." The Plant Cell **20**(7): 1948-1963.

Walton, J. D. (2006). "HC-toxin." Phytochemistry **67**(14): 1406-1413.

You, M., Z. Yue, W. He, X. Yang, G. Yang, M. Xie, D. Zhan, S. W. Baxter, L. Vasseur and G. M. Gurr (2013). "A heterozygous moth genome provides insights into herbivory and detoxification." Nature Genetics **45**(2): 220-225.

Zhang, G., C. Cowled, Z. Shi, Z. Huang, K. A. Bishop-Lilly, X. Fang, J. W. Wynne, Z. Xiong, M. L. Baker and W. Zhao (2013). "Comparative analysis of bat genomes provides insight into the evolution of flight and immunity." Science **339**(6118): 456-460.

Zhang, G., X. Fang, X. Guo, L. Li, R. Luo, F. Xu, P. Yang, L. Zhang, X. Wang and H. Qi (2012). "The oyster genome reveals stress adaptation and complexity of shell formation." Nature **490**(7418): 49-54.

Zipfel, C. (2008). "Pattern-recognition receptors in plant innate immunity." Current opinion in immunology **20**(1): 10-16.

Chapter 2 | Resequencing and Comparative Genomics of *Stagonospora nodorum*: Sectional Gene Absence and Effector Discovery

Attribution Statement

Authors:	**Robert A. Syme**, James K Hane, Timothy L. Friesen, Richard P. Oliver
Citation:	Resequencing and comparative genomics of Stagonospora nodorum: sectional gene absence and effector discovery. RA Syme, JK Hane, TL Friesen, RP Oliver - G3: Genes|Genomes|Genetics, 2013

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

The Ph.D. candidate (Robert A. Syme) made the following contributions to this chapter:

- Genome Assembly of strains SN4 and SN79 and generation of summary statistics for strains SN15, SN4, and SN79
- Gene calling on strains SN4 and SN79
- Read mapping for strains SN4 and SN79
- Comparative analyses including generation of figures and tables
- Effector gene prediction
- Co-writing of manuscript with RPO and JKH.

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

Signature: ………………………………………………

Date: 2015-05-19

I, Richard Oliver, certify that this attribution statement is an accurate record of Robert Syme's contribution to the research presented in this chapter.

Signature: ………………………………………………

Date: 2015-05-20

Resequencing and Comparative Genomics of *Stagonospora nodorum*: Sectional Gene Absence and Effector Discovery

Robert Andrew Syme,* James K. Hane,† Timothy L. Friesen,‡ and Richard P. Oliver*,1
*Australian Centre for Necrotrophic Fungal Pathogens, Curtin University, Department of Environment and Agriculture, Bentley WA 6845, Australia, †Molecular Pathology and Plant Pathology Laboratory, Centre for Environment and Life Sciences, CSIRO, Floreat WA 6014, Australia, and ‡U.S. Department of Agriculture, Agricultural Research Service, Cereal Crops Research Unit, Fargo, North Dakota 58102-2765

ABSTRACT *Stagonospora nodorum* is an important wheat (*Triticum aestivum*) pathogen in many parts of the world, causing major yield losses. It was the first species in the large fungal Dothideomycete class to be genome sequenced. The reference genome sequence (SN15) has been instrumental in the discovery of genes encoding necrotrophic effectors that induce disease symptoms in specific host genotypes. Here we present the genome sequence of two further *S. nodorum* strains (Sn4 and Sn79) that differ in their effector repertoire from the reference. Sn79 is avirulent on wheat and produces no apparent effectors when infiltrated onto many cultivars and mapping population parents. Sn4 is pathogenic on wheat and has virulences not found in SN15. The new strains, sequenced with short-read Illumina chemistry, are compared with SN15 by a combination of mapping and de novo assembly approaches. Each of the genomes contains a large number of strain-specific genes, many of which have no meaningful similarity to any known gene. Large contiguous sections of the reference genome are absent in the two newly sequenced strains. We refer to these differences as "sectional gene absences." The presence of genes in pathogenic strains and absence in Sn79 is added to computationally predicted properties of known proteins to produce a list of likely effector candidates. Transposon insertion was observed in the mitochondrial genomes of virulent strains where the avirulent strain retained the likely ancestral sequence. The study suggests that short-read enabled comparative genomics is an effective way to both identify new *S. nodorum* effector candidates and to illuminate evolutionary processes in this species.

KEYWORDS
Fungal comparative genomics
Fungal effectors
Phytopathogen

Stagonospora nodorum (teleomorph: *Phaeosphaeria nodorum*; *syn. Septoria*) is the causal agent of the wheat (*Triticum aestivum*) diseases Stagonospora nodorum blotch and glume blotch (Solomon *et al.* 2006a,b; Oliver *et al.* 2012). *S. nodorum* is an economically important pathogen in many parts of the world, causing major yield losses and degradation of grain quality (Murray and Brennan 2009; Tan *et al.* 2009).

The genome sequence of a reference strain (SN15) was shotgun-sequenced using Sanger technology and assembled into 107 nuclear scaffolds and the complete mitochondrial genome (Hane *et al.* 2007). The nuclear scaffolds totaled 38 Mb and included approximately 5% long interspersed repeats. Of the 15,980 genes initially predicted to be present in the SN15 nuclear assembly, 12,382 have had their annotation manually corrected or presence validated (Bringans *et al.* 2009; Casey *et al.* 2010; Ipcho *et al.* 2012).

A major impact stemming from the genome sequence was the identification of necrotrophic effector genes (previously called host-specific or host-selective toxins). These studies revealed that disease symptoms and severity differed in a cultivar- and isolate-specific manner. It therefore has become a priority to sequence more isolates to determine the extent of genetic variation. Necrotrophic effectors have been characterized in *S. nodorum* by a combination of approaches that typically start with the mapping of disease quantitative trait loci (QTL). More than 20 distinct disease resistance/susceptibility QTL have been

doi: 10.1534/g3.112.004994
Manuscript received November 28, 2012; accepted for publication April 1, 2013

Supporting information is available online at http://www.g3journal.org/lookup/suppl/doi:10.1534/g3.112.004994/-/DC1
[1]Corresponding author: ACNFP, Curtin University, Room 304.109, Department of Environment and Agriculture, GPO Box U1987, Perth, WA, Australia, 6845. E-mail: richard.oliver@curtin.edu.au

mapped in wheat and the number continues to grow (Friesen *et al.* 2007; Abeysekara *et al.* 2009; Faris and Friesen 2009; Friesen *et al.* 2009; Chu *et al.* 2010; Abeysekara *et al.* 2012; Crook *et al.* 2012; Liu *et al.* 2012; Oliver *et al.* 2012). The isolation of effector-active proteins either by direct purification from culture filtrates or by expression of candidate effector genes in microbial systems allows the mapping of sensitivity loci in structured wheat populations. Co-location of effector sensitivity loci and disease QTL has been performed in five cases. Three of the effectors have been cloned and fully characterized. These are *SnToxA*, *SnTox1*, and *SnTox3* (Friesen *et al.* 2006; Liu *et al.* 2009; Liu *et al.* 2012). Each of these is a small, secreted, and cysteine-rich protein with important disulfide bonds (Luderer *et al.* 2002; Liu *et al.* 2009). In addition, two of the effector genes (*SnToxA* and *SnTox3*) were found to be adjacent to repetitive DNA or to scaffold termini.

Different strains of *S. nodorum* produce different suites of effectors. Effector gene absence is a commonly observed allele in the three well-known effectors produced by *S. nodorum*. Haplotype analysis of world-wide populations of *S. nodorum* shows 15% of isolates lack *SnTox1* (Liu *et al.* 2012), 40% of isolates lack *SnTox3* (Liu *et al.* 2009), and 60% of isolates lack *SnToxA*, (Stukenbrock and Mcdonald 2007). Some strains produce no apparent effectors. Effector-deficient strains are avirulent on wheat (Friesen *et al.* 2006; Liu *et al.* 2009; Zhang *et al.* 2011; Liu *et al.* 2012) but cause disease on related grass weeds.

The three known effectors exist in a number of variant forms expressed from up to 13 different alleles. Differences in the protein sequence modulate effector activity (Tan *et al.* 2012). The sequence differences carry the hallmark of accelerated evolution as determined by elevated dN/dS ratios (Stukenbrock and Mcdonald 2007).

The related wheat pathogen *Pyrenophora tritici-repentis* also produces a battery of effectors. The majority of isolates produce a version of ToxA, here referred to as *PtrToxA* (Ciuffetti et al. 1997; Friesen *et al.* 2006). We presented several lines of evidence that a *ToxA* gene in *S. nodorum* was laterally transferred to *P. tritici-repentis* shortly before 1941 (Friesen *et al.* 2006). The *SnToxA* sequence in *S. nodorum* was present on a scaffold of 32.4 kb. A central section of 11 kb containing *SnToxA* was highly homologous and colinear with the *PtrToxA* containing sequence. The 11 kb also contained a gene for a DNA transposase. Repetitive DNA was found on either side of the 11 kb whereupon an easily recognized level of homology was lost. This finding suggested that the *ToxA* gene was transferred along with at least 11 kb of DNA and possibly as part of a transposon (Friesen *et al.* 2006). The publication of the *P. tritici-repentis* Pt-1c genome (Manning *et al.* 2013) and the sequences of strains of both species that lack *ToxA* allows us to examine this phenomenon in more detail.

In this study we have used second-generation sequencing to determine the DNA sequence of two further strains of *S. nodorum*. One strain—Sn4—was isolated from wheat in the United States. It is known to produce a different complement of effectors to SN15. The second strain, Sn79-1087 (hereafter Sn79), was isolated from the grassy weed *Agropyron* (Friesen *et al.* 2006). The Sn79 strain is essentially avirulent on wheat and Sn79 culture filtrates produce no reactions on a diverse panel of wheat cultivars (Figure 1).

Comparison of the genome sequences of these two additional *S. nodorum* strains was undertaken to identify further effectors present in Sn4 or SN15 but absent in Sn79 and to explore the genetic differences in each strain that underpins their differing pathogenicity profiles.

MATERIALS AND METHODS

Strains, infection, DNA preparation, and sequencing

S. nodorum strains Sn4 (Zhang *et al.* 2011) and Sn79 (Friesen *et al.* 2006) were grown in Fries3 with 30 mM glutamate for 1 wk. Spores were harvested and used to infect detached leaves of wheat cultivars Kulm and Selkirk as previously described (Friesen *et al.* 2006). For DNA extraction, mycelium was harvested with a milk filter and cells lysed with a bead crusher. The concentrations of DNA extracted using the CTAB protocol (Doyle and Doyle 1987) were measured to be 364 ng/mL for the Sn4 extraction and 223 ng/mL for Sn79.

Mapping

Raw reads were filtered using one round of cutadapt v1.0 (Martin 2011) using a Phred quality threshold of 28, trimming known Illumina adapters from the 3′ end. Trimmed reads were mapped to the reference sequence using BWA 0.5.7 (Li and Durbin 2009), allowing for three gap openings and 21-bp gap extensions per read to allow for large indels. Relatively loose mapping parameters were chosen to reduce the probability of false positive low-coverage regions. Reads mapping around identified indels were locally realigned with GATK 1.2.4 (McKenna *et al.* 2010) using the RealignerTargetCreator and IndelRealigner modules to minimize the likelihood of false single-nucleotide polymorphisms (SNPs).

Mapping depth overviews were visualized as Hilbert curves using the HilbertVis v1.15 (Anders 2009), Rsamtools 1.0, and Bioconductor 2.6 (Gentleman *et al.* 2004) packages. Details of the Hilbert curve generation are available in Supporting Information, Figure S1. Regions of low coverage and consecutive gene absence were calculated using the genomeCoverageBed and intersectBed utilities of BEDTools 2.12.0 (Quinlan and Hall 2010). Regions of low coverage were defined as parts of the reference genome with mapped read depth of 5 or less. Scaffold gaps of poly-'N' in the reference sequence were removed from the list of low coverage regions as no reads can map to these loci.

Assembly

In addition to mapping the reads from Sn4 and Sn79 back to the reference SN15 genome, each genome was assembled *de novo* with Velvet v1.0.10 (Zerbino and Birney 2008) using VelvetOptimizer 2.2.0 to find the assembly with optimal N50 length.

Mitochondrial sequence

The mitochondrial genomes for Sn4 and Sn79 were considered separately to the nuclear genome. Sequences for the new mitochondrial genomes were inferred from the consensus of reads mapped by BWA to the SN15 mitochondrial genome. Two instances of manual gap removal and contig joining were required to finish the assembly. The velvet assembly was cleaned of mitochondrial sequence by aligning the velvet contigs to the SN15 reference with Mauve 2 using default parameters and removing contigs that aligned with the mitochondrial sequence. The removed mitochondrial contigs were manually checked for major structural changes not present in BWA-mapped consensus to validate the consensus assembly.

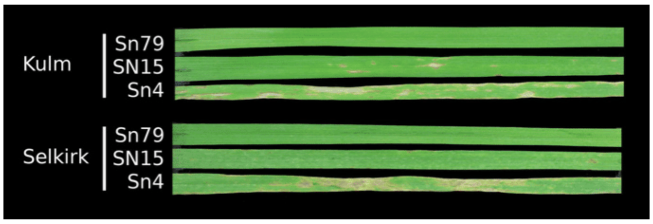


Figure 1 The wheat varieties Kulm and Selkirk infected with *S. nodorum* strains Sn79, SN15, and Sn4.

Annotation

Draft gene models were generated with MAKER2 (Cantarel *et al.* 2008) using evidence from a SNAP (version 2010-07-28) hidden Markov model generated from the validated SN15 gene structures, a Genemark ES 2.3 hidden Markov model (Ter-Hovhannisyan *et al.* 2008) trained on each of the new genome sequences, and mRNA sequences from the set of validated SN15 genes. Complete predicted proteins were clustered with the validated SN15 reference proteins using OrthoMCL 2.0.2 (Li *et al.* 2003), using an mcl expansion parameter of 1.5 and a maximum evalue of 1e-5 for the blastp search v2.2.27+.

Repeats

Known *Stagonospora* repeats (Hane *et al.* 2007; Hane and Oliver 2008; Hane and Oliver 2010); transposons Molly [Genbank:AJ488502.1], Pixie [Genbank:AJ488503.1], and Elsa [Genbank:AJ277966.1]; and simple repeats were identified using RepeatMasker v3.2.8 and crossmatch (from phrap/cross_match/swat ver 1.090518) (Smit *et al.* 1996-2004) in the Sn4 and Sn79 genome sequences using default parameters.

Repeat-induced point mutation (RIP) analysis

RIP is a fungal-specific process in which duplicated sequences are mutated during meiosis (Selker *et al.* 1987; Selker 1990). In Pezizomycotina, RIP modifies CpN dinucleotides and changes are predominantly in the direction CpA to TpA (Cambareri *et al.* 1989; Clutterbuck 2011). RIPCAL identifies regions likely to have undergone RIP by alignment of similar sequences and scanning for nucleotide changes that are characteristic of RIP (Hane and Oliver 2010). Thus, RIPcal can be used to identify both the presence and direction of a RIP-like sequence change. *S. nodorum* SN15 scaffolds were identified that showed sequence similarity to the *ToxA* region in *P. tritici-repentis* by blat v34 (Kent 2002) using parameters minIdentity = 70 and minScore = 200. For each blat matched region, the nucleotide sequence was extracted from each species' scaffold and aligned with ClustalW v2.1 (Larkin *et al.* 2007). A 250-bp nonoverlapping sliding window was moved across the concatenated alignments and RIP-like mutations of CpN → TpN were identified from *Pyrenophora* to *Stagonospora* and then from *Stagonospora* to *Pyrenophora* using RIPCAL.

Phylogenetic analysis

S. nodorum Sn4, Sn79, SN15, *P. tritici-repentis*, and *L. maculans* protein sequences for the AFTOL2 (Celio *et al.* 2006) genes *CDC47*, *KRR1*, *CCT4*, *GCD6*, *GPI8*, *MCM3*, *COX15*, *RAD3*, *DBP3*, *NBP35*, *CHC1*, and *RBG2* were identified in the resequenced strains using blast. Homologous genes from the three strains were aligned using MUSCLE 3.6 (Edgar 2004) using default parameters. The alignments were then manually trimmed and concatenated to provide regions present in all five genomes. The concatenated alignments were used to construct phylogenetic trees using the neighbor joining algorithm provided by Geneious (Drummond *et al.* 2011), using the Jukes-Cantor model for genetic distance and *L. maculans* as an outgroup. A consensus tree from 20 replicates was built using a support threshold of 50%.

Effector candidates

Each of the validated SN15 proteins was assessed against a series of normative assumptions about the characteristics of proteinaceous effectors (Table 1). Small proteins (≤30 kD, calculated using BioRuby 1.4.2 (Goto *et al.* 2010) were scored. Mean cysteine content as a percentage of total protein length was calculated and proteins were scored if they had cysteine percentage greater than one standard deviation from the mean. Proteins were scored if they were encoded by genes within 5 kb of repeats. Repetitive sequences are likely to result in contig breaks in the final assembly, so scaffold ends were treated as equivalent to repeats for this scoring step. Proteins were scored if they had no blast hits to the National Center for Biotechnology Information nonredundant protein database (NR) after we excluded *S. nodorum* proteins, using a minimum e-value cutoff of 1e-20. Proteins were scored if WolfPSort v0.2 (Horton *et al.* 2007) predicted the protein to be located extracellularly, or SignalP 3.0 (Bendtsen *et al.* 2004) identified a signal peptide in the sequence. WolfPSort was run with "fungi" as the organism source. SignalP was run with default parameters. Proteins with coding sequence coverage by Sn4 reads of at least 20% were scored, and proteins with coding sequence coverage by Sn79 reads <20% were scored. Evidence of positive selection for each gene was calculated by taking Sn4 SNPs identified in coding regions and measuring dN/dS using the CNFGTR model (Yap *et al.* 2010) as implemented in PyCogent 1.5.1 (Knight *et al.* 2007).

RESULTS

Cultivar response

We chose two strains for resequencing. Sn79 was isolated from Agropyron and produces no significant disease on any wheat cultivar tested (Figure 1). Infiltration of Sn79 culture filtrate into wheat cultivars produced no reaction but transformation of ToxA, Tox1, or Tox3 into Sn79 rendered the strain capable of causing disease and necrosis on wheat lines carrying sensitivity loci for the respective effector (Friesen *et al.* 2006; Liu *et al.* 2009, 2012). Sn4 is a virulent wheat isolate that is known to produce additional effectors compared with SN15 (Faris *et al.* 2011). Figure 1 shows the more pronounced necrosis in response to Sn4 than both SN15 and Sn79 on wheat cultivars Kulm and Selkirk which suggests the presence of effectors in Sn4 that are absent from SN15.

Table 1 Criteria used to identify effector candidates

Criteria	*S. nodorum sn15* Proteins That Match the Criteria
In silico analysis	
Protein size < 30 kDa	7069
Cysteine rich[a]	1817
Within 5 kb of repeats or scaffold ends	9119
No blasts matches in nr (evalue < 1e-5)	2650
Predicted to be secreted by WolfPSort or SignalP	2788
Resequencing data	
Present, but modified coding sequence in sn4.	6507
Absent in Sn79 (read coverage of coding sequence <20%)	1943
Evidence of positive selection pressure	729

[a] Proteins with cysteine composition percentage more than one SD above the mean.

Sequencing

DNA sequencing of 35-bp and 75-bp single-end Illumina read libraries yielded 7,097,630 and 13,783,290 reads, respectively, for strain Sn4 and 6,685,623 reads and 15,615,628 reads for Sn79. Assuming the newly sequenced genome sizes are equal to that of SN15 assembly (37.2 Mbp), the sequencing data provided 34-38x coverage.

The reads for each new strain were assembled using Velvet final contigs are available from NCBI as bioproject accessions PRJNA170815 (Sn4) and PRJNA170816 (Sn79). A summary of the genome assembly and mapping statistics is given in Table 2. As expected, the Illumina-sequenced genomes were more fragmented than the Sanger-sequenced reference strain. Commonly reported metrics of assembly contiguity are N50 and L50 (N50 length). The N50 is the smallest number of sequences that contain 50% of the total assembly length and the L50 is the length of the smallest sequence within this subset. The fragmentation of the short-read assemblies has resulted in L50s of 22 kbp and 17 kbp for Sn4 and Sn79 compared with 1.045 Mbp for SN15 and contig counts of 2559 and 3132 for Sn4 and Sn79 compared to 107 scaffolds for SN15. The assembly sizes of Sn4 and Sn79 are comparable (34.6 Mbp and 33.8 Mbp) but smaller than SN15 (37.2 Mbp) because of reduced representation of their repetitive DNA content.

The process of *de novo* assembly with short reads is likely to collapse highly similar repeat units and assemble only divergent repeat copies as well as single copy regions. Fragments of known *S. nodorum* repeats (Hane *et al.* 2007; Hane and Oliver 2008, 2010), were identified in the new assemblies. The short-read assemblies contain 8.5% (Sn4) and 8.4% (Sn79) of the total repeat content of SN15 (data not shown). All but one of the SN15 repeat families was detected as fragments in the new assemblies. The exception was the telomere-associated repeat R22, which was not detected in the Sn4 assembly. The SN15 genome contains 35.48 Mbp of nonrepetitive nuclear DNA, comparable with the assembly sizes of Sn4 (34.6 Mbp) and Sn79 (33.8 Mbp). We therefore tentatively conclude that all three genomes have similar genome sizes.

Gene calling

We have accumulated substantial evidence for the validation of 12,383 genes in SN15 (Bringans *et al.* 2009; Casey *et al.* 2010; Ipcho *et al.* 2012) and demoted 3598 genes to “nonvalidated” status (Hane *et al.* 2007). Gene calling by GeneMark-ES v2 (Ter-Hovhannisyan *et al.* 2008) in the new assemblies predicted 14,391 and 14,352 genes in Sn4 and Sn79, respectively. The genes in Sn4 and Sn79 have average feature sizes comparable with the validated SN15 genes and much larger than the discarded SN15 genes (Table 2). Compared with experimentally validated gene models in SN15, the gene models in Sn4 and Sn79 have a slightly larger mean protein size (437aa and 434aa compared with 423aa), a larger median protein size (367aa and 366aa compared with 347aa), fewer exons/gene (2.45 compared with 2.65), and longer mean exon length (515 bp and 510 bp compared with 408 bp). These findings indicate that the new assemblies and gene calls are of comparable accuracy to the SN15 reference and are likely better than the set of low-confidence unvalidated reference genes.

The phylogenetic relationships of the three strains were investigated by comparing 15 single-copy or low-copy genes used routinely for phylogenetic analysis. These genes were aligned together with those of *P. tritici-repentis* and *L. maculans*, which constituted an outgroup. The phylogenetic tree indicates that each new strain is closely related to SN15 and thus a *bona fide S. nodorum* sequence. It also indicated that the two wheat pathogens are more closely related to each other than either is to the grass pathogen Sn79 (Figure 2). Tree topologies of the individual genes agreed with the tree constructed from the sequence concatenation (data not shown). The three strains are equally distant from both species in the outgroup. This finding is consistent with recent molecular analyses of Pleosporales that place *Pyrenophora*, *Leptosphaeria* and *Phaeosphaeria* (*Stagonospora*) in three distinct families within the suborder Pleosporinae (Zhang *et al.* 2009). Previous authors had placed *S. nodorum* in the *Leptosphaeria* genus.

Table 2 Genome overview. The resequenced genome assemblies are rich in genic regions

	S. nodorum SN15[a]		*S. nodorum* Sn4	*S. nodorum* Sn79
Nuclear assembly				
Sequencing technology	Sanger		Illumina	Illumina
Assembled size, Mbp	37.2		34.6	33.8
L50, kbp	1,045		22	17
N50 (scaffolds or contigs)	13		499	613
Max, kbp	2532		113	88
Scaffolds >1kb	107		2559	3132
G+C content, %	50.5		51.7	51.9
	Validated[b]	Nonvalidated		
Predicted protein coding gene number	12,382	3598	14,391	14,352
Mean protein size (amino acids)	423	232	437	434
Median protein size (amino acids)	347	170	367	366
Mean exons/gene	2.65	2.30	2.45	2.45
Mean exon length, bp	480	306	515	510
Median exon length, bp	278	170	293	296
Gene density (genes per 10 kb)	3.3	–	4.2	4.2
Mitochondrial genome				
Size, kbp	49.8	–	49.8	42.4
G+C content, %	29.4	–	29.4	26.9
Gene content	22	–	22	19

[a] Hane *et al.* 2007.
[b] Genes validated by expressed sequence tagging, proteogenomics, and microarray evidence.

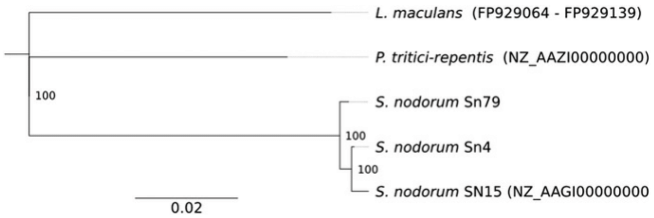

L. maculans (FP929064 - FP929139)
P. tritici-repentis (NZ_AAZI00000000)
100
S. nodorum Sn79
100
S. nodorum Sn4
100
S. nodorum SN15 (NZ_AAGI00000000
0.02

Figure 2 Phylogenetic tree constructed from an alignment of the AFTOL proteins CDC47, KRR1, CCT4, GCD6, GPI8, MCM3, COX15, RAD3, DBP3, NBP35, CHC1, and RGB2. Genbank accessions are shown in brackets and consensus support (%) shown at tree nodes.

Comparative genomics

The presence and absence of regions in the SN15 genome was analyzed in strains Sn4 and Sn79 via a combination of read-mapping and *de novo* assembly. *De novo* assembly was important for the detection of loci that were present in Sn4 or Sn79 but absent in SN15 and thus not detectable by read-mapping.

The new reads were mapped to the genome by BWA. The average mapped read depth in coding regions were 23.2X (Sn4) and 26.2X (Sn79) but was highly variable. Repetitive regions were unreliably accounted for by this procedure, so we only consider low copy number, mainly genic regions here.

Mapping of the reads of each strain to the reference assembly and calculating coverage of each gene's coding sequence suggested that 180 SN15 genes were absent from Sn4 and 367 SN15 genes were absent from Sn79 (Table 3) where genes with <5% coding region coverage are defined as absent. Inspection of the pattern of missing genes suggested that in many cases contiguous stretches of genes were absent. Quantification of this phenomenon showed that 57% (Sn4) and 76% (Sn79) of absent genes have a neighboring gene that is also absent. We refer to the contiguous runs of missing genes as sectional gene absence (SGA). The degree of SGA was much greater in Sn79 in both frequency and size. The largest SGA from Sn4 included only five genes, and the largest from Sn79 included 51 genes.

Genes sectionally absent from Sn79 include all the genes on scaffold 44 (48 genes) and scaffold 45 (51 genes, the largest SGA), and the polyketide synthase gene SNOG_07866 [NCBI Gene: 5975086], which is absent in a cluster of 11 genes at the end of scaffold 11.

This pattern of nonrandom SGA prompted us to explore ways to graphically display this finding. Very long read-depth data can be conveniently visualized as a space filling curve that attempts to preserve proximity during its transformation to an increased number of dimensions. Figure 3 shows Hilbert plots generated using R with the HilbertVis (Anders 2009) and Bioconductor (Gentleman *et al.* 2004) packages (see Figure S1 for more information on Hilbert plots). Only scaffold 2 is shown; scaffolds 1–12 are shown in the supplementary data. White sections in Figure 3A represent those sections of SN15 scaffold 2 that are annotated as coding sequence, whereas B and C represent the mapped read density of Sn4 and Sn79, respectively. White regions in B and C are parts of SN15 scaffold 2 that are covered by Sn4 and Sn79. Large dark boxes 1−3 indicate regions of SN15 scaffold 2 that have no reads mapped from Sn4 and Sn79 and are likely to be absent from the newly sequenced strains.

Assembly of the reads *de novo* allowed us to identify genes not present in the SN15 reference assembly and to estimate the core *S. nodorum* gene set. Clustering the predicted proteins from the three genomes using OrthoMCL gave a core cluster number estimate of 10,464 (Figure 4). The clusters showed a larger conserved gene set between SN15 and Sn4 (430) than between SN15 and Sn79 (246) consistent with the phylogenetic analysis. The large set shared between Sn79 and Sn4 is likely an overestimate as a result of these strains' shared gene calling procedure. As gene models are improved, the number of genes in this set is likely to fall.

The 10,464 protein clusters in the conserved *S. nodorum* core include many of the products essential to metabolism. Of the protein clusters present in all three strains, 87% had blast hits in NR. A markedly lower percentage of the strain-specific genes (21–50%) had similarity to protein sequences in NR.

Expansion in gene copy number was detected in eight clusters (Table 4). Of the clusters, five contained genes with higher copy numbers in SN15, and three clusters contained genes with greater copy numbers in Sn4. No clusters contained genes expanded in the non-pathogenic strain Sn79. Four of the eight expanded clusters contained genes with no blast hits (e value ≤ 1e–20).

Effector comparison

S. nodorum Sn79 produces none of the three known *S. nodorum* effectors: *SnTox1*, *SnToxA*, and *SnTox3* (Friesen *et al.* 2006; Liu *et al.* 2009, 2012). Effector-containing regions in SN15 were compared with the corresponding regions in Sn4 and Sn79. These known *S. nodorum* effectors are all absent from the Sn79 assembly (Figure 5 and Figure S2). In the case of *SnTox1*, the effector gene and 2 kb of upstream intergenic sequence was absent from Sn79. *SnTox3* was absent from Sn79 together with the three upstream genes (SNOG_08982 to SNOG_08985; NCBI 5976184 to 5976187). The absent genes included a predicted protein-disulfide isomerase, a short-chain dehydrogenase, and a gene without blast hits. The repetitive regions downstream of *SnTox3* gene did not assemble in either Sn4 or Sn79.

SnToxA is found on scaffold 55 together with just three more called genes, including a transposase. The remainder of the scaffold comprises sections of gene-free single copy DNA and repetitive elements. The single-copy regions, including a contiguous stretch of 11 kb, were present in Sn4 albeit on different contigs. The 11-kb region was absent in Sn79, but some of the repetitive elements were present on short contigs (Figure 5). The presence of *ToxA* in some strains of both *S. nodorum* and *P. tritici-repentis* is suggestive of lateral gene transfer. We earlier reported that the colinear region of 11 kb found in both SN15 and Sn4 is present in *P. tritici-repentis* with very

Table 3 ***S. nodorum* SN15 genes deemed absent in the newly sequenced strains**

	S. nodorum Sn4	*S. nodorum Sn79*
SN15 genes absent in the other strains (<5% exon coverage)	180	367
SN15 genes absent from other strain in sections	103 (23 with informative blast)	279 (74 with informative blast)
Largest sectional gene absence (gene count)	10	51
Number of SN15 sections absent from other strain	40	56

Illumina reads from Sn4 and Sn79-1087 were mapped to the SN15 scaffolds and genes with <5% coding sequence coverage were identified as absent. Few genes are absent in isolation, and many are missing from Sn79 in large sections.

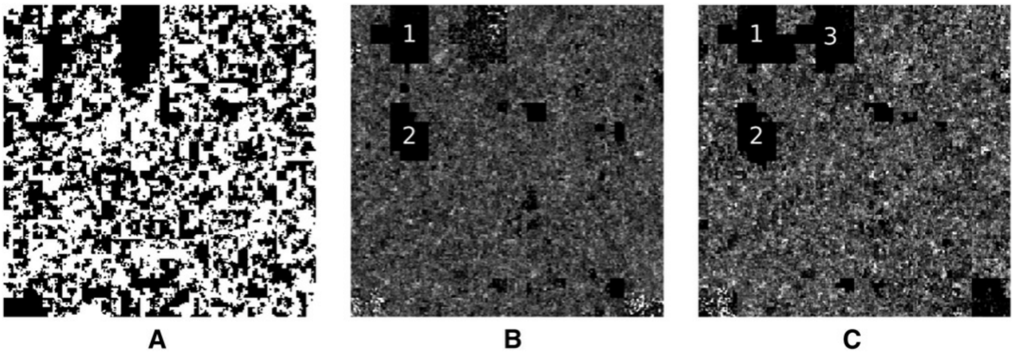


Figure 3 Hilbert plots of coding sequence and read density of SN15 scaffold 2. (A) White regions correspond to coding regions. *S. nodorum* Sn4 (B) and Sn79 (C) mapped read density on SN15 scaffold 2. Lighter shades correspond to higher read density. Regions 1 (45−46 kbp) and 2 (31 kbp) are absent from both Sn4 and Sn79. Region 3 shows a 48-kbp region absent from Sn79.

high levels of similarity (Friesen *et al.* 2006). The assembly of *P. tritici-repentis* Pt-1c using Sanger sequencing aided by optical mapping has recently been published (Manning *et al.* 2013). *PtrToxA* is present on scaffold 4/chromosome 6 of the Pt-1c assembly. Alignment of scaffold 4 to *C. heterostrophus* indicated that a region of 156 kb including *PtrToxA* was absent in the corn pathogen but the homologous flanking regions were present in both species (Manning *et al.* 2013). A series of SN15 scaffolds [68, 55 (including *SnToxA*), 51, 46, 64, and 73] appear to be colinear compared with *P. tritici-repentis* (Figure 6). Regions of colinearity were interspersed with repeated sequences inserted into the SN15 genome. Scaffold 46 has 4 such colinear regions interspersed with five repetitive sections. The total length of the colinear DNA in *P. tritici-repentis* was 72 kb, corresponding to a 350-kb region in SN15 that was inflated by multiple transposon invasions. Read mapping of Sn4 to the SN15 genome indicated that most of the repetitive regions, the 11 kb *SnToxA* region, and a gene free single copy region on scaffolds 68 and 51 were present in Sn4. In contrast, a single copy region on scaffold 51 with two annotated genes, and the single copy regions on scaffolds 46, 64, and 73 were absent in Sn4. In Sn79 all single copy regions were absent and only traces of the repetitive DNA could be found.

The 72 kb of colinear DNA contains 21 genes in *P. tritici-repentis* but only seven genes in SN15. The hallmarks of RIP have been abundantly found in the SN15 genome but only sparsely in the *P. tritici-repentis* genome (Hane and Oliver 2008, 2010; Manning *et al.* 2013). We extracted and aligned 46 kb of matching regions from the 72-kb *ToxA* region shown in Figure 6 and scanned for dinucleotide SNP mutations characteristic of RIP (Figure 7). The dinucleotide changes are predominantly CpA->TpA mutation, which is characteristic of RIP in Pezizomycotina species. Across most of the region, changes are consistent with the *S. nodorum* sequence having been subject to RIP. These changes were most likely in the form of RIP leakage. We have previously observed that RIP-like changes, initiated from interspersed repeat regions, are also found in neighboring single-copy regions (Hane and Oliver 2010; Van de Wouw *et al.* 2010). In contrast, one region upstream of the ToxA gene on scaffold 55 had changes characteristic of RIP in the Pt-1c sequence. Subsequences of this region were found in multiple copies in Pt-1c genome, which suggests that the transferred region has been susceptible to RIP in PTR since its lateral transfer, but the majority has remained in single copy and thus has evaded RIP mutation.

Mating type locus

The mating type idiotype of SN15 is *Mat1-1* and of Sn4 and Sn79 is *Mat1-2*. These idiomorphs have previously been sequenced (Bennett *et al.* 2003). Sn4 and Sn79 both contained copies of the *MAT1-2* idiomorph (Figure 8). The mating type genes show perfect sequence identity to published sequences. There is a 19-bp region present in the Sn79 *Mat1-2* locus that is absent in both the Sn4 and sn436GA98 sequences.

Effector candidate selection

The resequencing data were combined with data from existing SN15 gene models to predict likely effector proteins, based on a set of normative assumptions about the qualities of effector molecules and their genes both in *S. nodorum* (Friesen *et al.* 2006; Liu *et al.* 2009, 2012) and other fungal pathogens (De Wit *et al.* 2009) (Table 1). Some criteria were assessed before this sequencing project. These include size, cysteine content, proximity to repetitive elements, and prediction of secretion. Addition of data from this resequencing project allowed us to include criteria reporting gene absence or difference between the three strains.

Because Sn79 lacks detectable wheat effectors, we sought genes that were absent in this strain or highly divergent compared to SN15. Genes with <20% of their coding sequence covered by Sn79 reads (1943 genes) were scored in this criterion. The three known effector genes exist in different isoforms in different strains (Friesen *et al.* 2006; Liu *et al.* 2009, 2012). To exploit this, we sought genes that were present in Sn4 but contained sequence variation compared withSN15; a total of 6507 genes fit this criterion. Sequence comparison of many effector genes shows evidence of accelerated evolution. We therefore screened for genes with elevated dN/dS ratios when comparing Sn4 and SN15. A total of 729 genes showed evidence of high dN/dS.

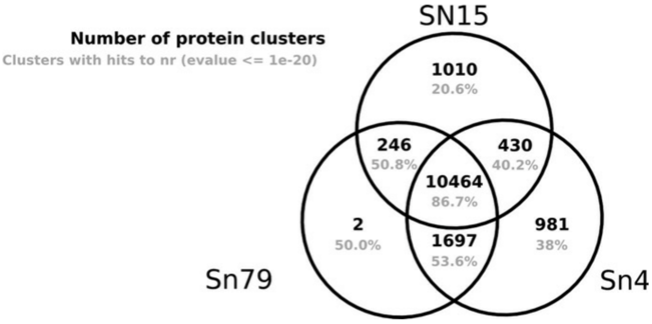


Figure 4 Protein ortholog clusters between the three *S. nodorum* strains. Proteins were grouped into clusters using orthoMCL (Li *et al.* 2003). Clusters were regarded as "nr-unique" if none of its members had blast hits to the nr database at an e-value cutoff of 1e-20.

Table 4 Gene family expansion in the three *S. nodorum* strains

Cluster id	No. SN15 Proteins	No. Sn4 Proteins	No. Sn79 Proteins	Function
stago_10024	8	0	0	No blast hits with evalue ≤1e-20
stago_10031	1	4	2	Phosphoribosylformylglycinamidine synthase (EC:6.3.5.3)
stago_10129	1	3	1	Conserved with unknown function
stago_10171	4	0	0	No blast hits with evalue ≤1e-20
stago_10172	3	0	0	No blast hits with evalue ≤1e-20
stago_10388	3	0	0	No blast hits with evalue ≤1e-20
stago_10389	3	0	0	Conserved with unknown function
stago_10390	0	3	0	Putative retrotransposon protein

There was no evidence of gene expansion in the non-pathogenic strain Sn79. NB cluster id numbers do not correspond to gene identifiers.

Each of the criteria matched between 729 and 9119 genes, providing a filter that reduced 12,382 gene models to 159 candidates that matched at least six of the criteria (Table 5). Known effectors *SnTox1* and Sn*ToxA* were ranked first and equal third according to these criteria (Table S1).

Proteins from the Sn4 genome not observed in SN15 were also scored for likely effector properties including small size (≤30 kD), predicted to be secreted, and cysteine-rich. Of the predicted proteins present in the 981 clusters unique to Sn4, four fulfill all three of the criteria given previously (File S2).

Mitochondrial genomes

The mitochondrial genomes of Sn4 and Sn79 were considered separately to the nuclear genome. Sequences for the new mitochondria were assembled by taking the consensus sequence of the BWA-mapped reads of the new strains and removing mitochondrial sequences from the Velvet-assembled contigs.

Alignment of the mitochondrial genomes of Sn4 and SN15 showed entirely conserved gene content and order with no indels larger than 3 bp. Two sections of the SN15 and Sn4 mitochondrial genomes were absent in the Sn79 genome (Figure S3). These missing sections included an intronic endonuclease and two downstream open reading frames of unknown function. The total mitochondrial sequence discrepancy was 7.43 kb absent from Sn79 but present in Sn4 and SN15.

DISCUSSION

Fungal plant pathogenic species are notoriously variable in their morphology and growth characteristics; pathology studies have traditionally focused on differences in virulence and for some species these have been formalized via the definition of *forma specialis*, pathotypes or races. Exploration of the genetics of the virulence differences has proved a fruitful way to understand the molecular bases of virulence properties. Genome sequences of fungal plant pathogens have been obtained for a handful of species using Sanger-based technologies (Oliver *et al.* 2012). The advent of second-generation sequencing technologies has opened up the possibility of sequencing different strains and thereby directly measuring genetic differences.

Early studies of *S. nodorum* readily identified significant genotypic differences in strains based on expressed sequence tag, restriction fragment length polymorphism, and chromosome-length polymorphism studies (Oliver *et al.* 2012) but these differences were not related to virulence properties. It was not until the discovery of necrotrophic effectors that a rational basis for differentiating strains was elaborated (Friesen *et al.* 2008; Tan *et al.* 2010; Oliver *et al.* 2012). Proteins purified from culture filtrates of different isolates induced necrosis in different wheat cultivars. Sensitivity to the proteins mapped to QTL for disease susceptibility that differed between isolates. One strain, isolated from a wild grass, produced neither observable effectors nor disease on wheat. These strains and the new necrotrophic effector rationale provided sufficient incentive for resequencing.

The process of generating second generation sequencing data for two further strains was affordable but presented significant challenges to analyze the data. It was clear that the new strains genomes differed markedly from the reference SN15 strain. These differences were explored by mapping reads onto the reference and by examining *de novo* assemblies.

Both methods showed that many genes differed between the strains or were entirely absent. The mapping data suggests that Sn4/Sn79 lacked at least 180 of 367 genes, respectively, compared with SN15 (Table 3). In addition 6507 Sn4 genes differed in nucleotide coding sequence (Table 1). By any criteria, these are very large strain to strain differences which stretch the definition and technology of resequencing.

The gene complement differences were not randomly located in the genome but were predominantly in the form a contiguous sections

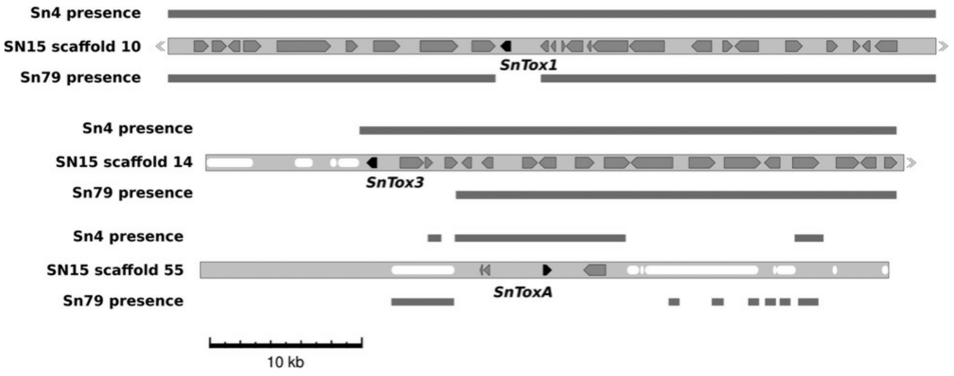


Figure 5 Effector context in three *Stagonospora* strains. Black arrows represent effector genes, gray arrows represent other genes, and white lozenges represent repetitive sequence. Gray boxes above and below the SN15 scaffold sections represent sequence present in the assemblies of the resequenced strains. The sequence surrounding *SnTox1* is present in all three strains. *SnTox3* is absent from Sn79 in a section that includes four genes. The entire *SnToxA* region is absent in Sn79.

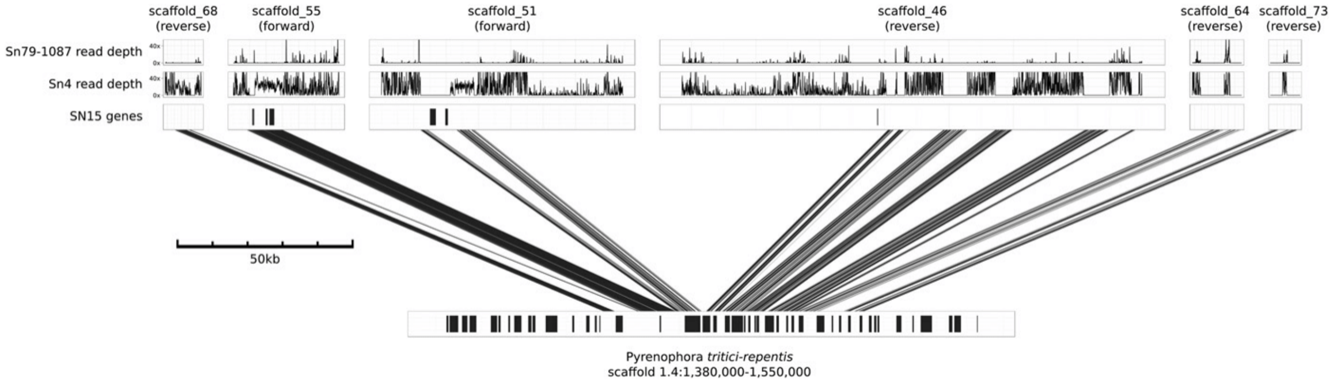


Figure 6 Regions around *ToxA* in *P. tritici-repentis* show homology to other SN15 scaffolds. The scaffolds are linked here with blat hits >75% identity. The total size of the expanded predicted laterally transferred DNA region is at least 72 kb.

containing two to 50 or more absent genes. We refer to this pattern as SGA. These genes include a high proportion without clear homologs in other species. One region absent in Sn79 on scaffold_11:124,000-127,500 comprises a cluster of genes that could well function in synthesizing a secondary metabolite involved in pathogenicity on wheat.

This pattern may resemble the presence in *M. graminicola* and *Fusarium oxysporum* of entire chromosomes that are unevenly harbored by different strains (Ma *et al.* 2010; Goodwin *et al.* 2011). These species contain dispensable chromosomes with relatively low gene density, a high frequency of genes without significant matches to known genes and high repetitive element content. In *S. nodorum* strain-to-strain SGA sometime comprised entire scaffolds. Pulsed field gel and genetic mapping evidence suggests that there are about 20 chromosomes in *S. nodorum* (Cooley and Caten 1991; Malkus *et al.* 2009), whereas the SN15 assembly contains 107 scaffolds. The lack of a finished SN15 genome assembly means that it is not currently possible to determine the chromosomal location of the SGA regions.

Clustering the predicted protein datasets of each strain can estimate the core proteins conserved between them. This approach gave a minimum conserved gene set of 10,464 protein clusters present in all three strains and an additional 430 conserved between both wheat pathogenic strains.

Mitochondrial sequence

The mitochondrial genomes of Ascomycete fungi contain a well-conserved set of genes but a variable amount of additional DNA (Hane

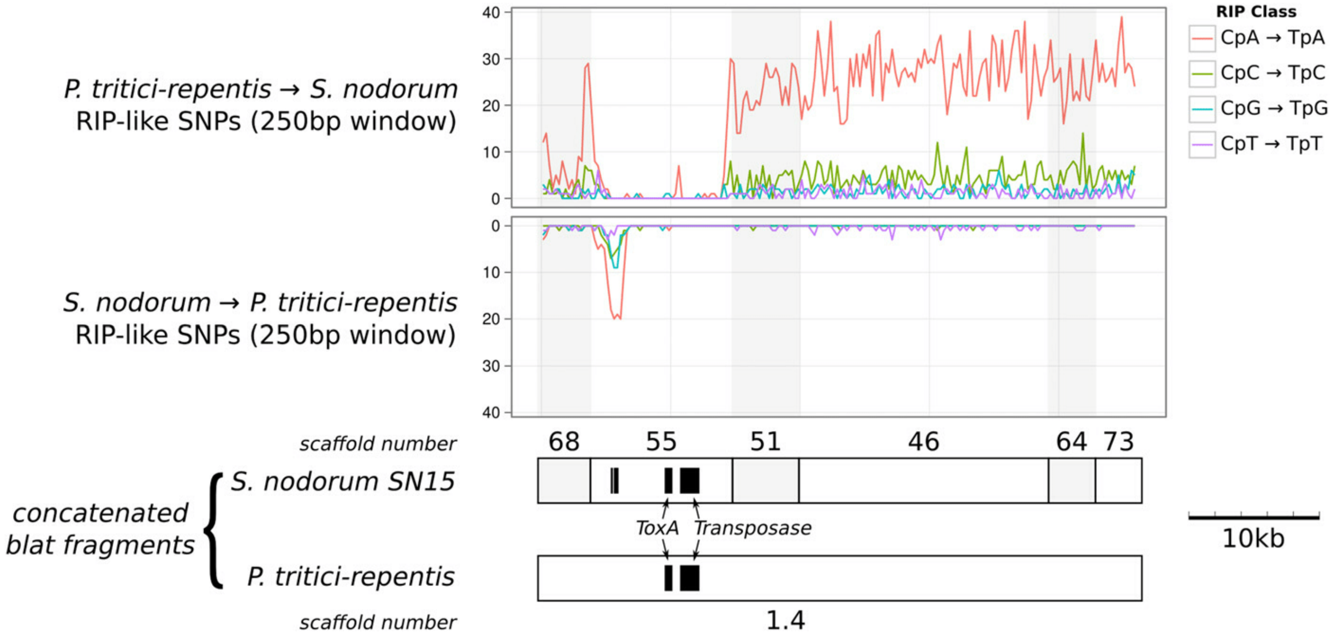


Figure 7 Regions of similarity around ToxA in *P. tritici-repentis* and *S. nodorum* SN15 identified by blat (Figure 6) were concatenated and aligned. RIP irreversibly converts cytosine bases to thymine, therefore comparisons between two sequences have directionality. The frequency of CpN to TpN mutations across 250-bp nonoverlapping windows has been separated by direction into those with CpN dinucleotides of *S. nodorum* being converted into TpN in *P. tritici-repentis* (top) and vice-versa (bottom). CpA to TpA changes that are associated with RIP are indicated by the red line. The location of genes is indicated by black boxes.

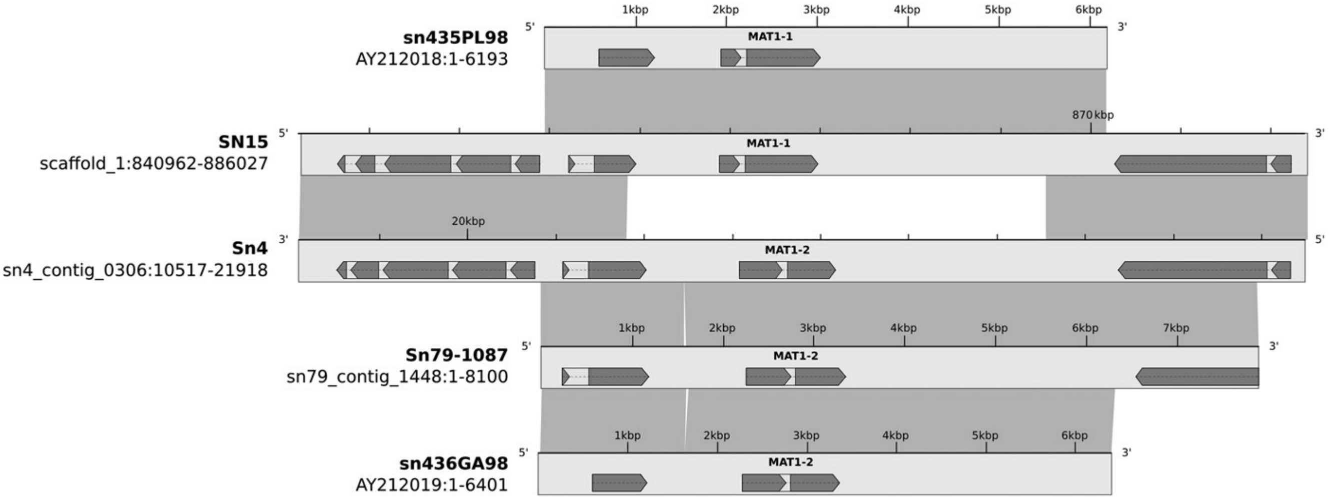


Figure 8 The MAT locus in 5 *Stagonospora* strains showing the two idiomorphs. Strains sn435PL98 and SN15 are mating type 1 (*MAT1-1*) and strains Sn4, Sn79, and sn436GA98 are mating type 2 (*MAT1-2*). Light gray boxes represent nucleotide sequence, and dark gray boxes connect regions with sequence similarity.

et al. 2011). This additional DNA can add 120 kb to the core ca. 40 kb of the genome (Rouxel *et al.* 2011). This additional DNA includes genes or unknown function and intronic genes encoding homing endonucleases. The mitochondrial genome of SN15 is 49,761 bp and includes three open reading frames (ORFs) of unknown function (ORF 1, 2, and 3) and four intronic endonucleases. The mitochondrial genome of Sn4 was found to be essentially identical, but the Sn79 genome differed significantly in one region (Figure S3) lacking the intron found in *atp6* and a region containing ORF 1 and 2. This pattern is consistent with the insertion of ORF3 and three introns into the ancestor of all three strains and the subsequent insertions of the *atp6* intron and the region containing ORF 1 and 2 into the lineage containing Sn4 and SN15. This scenario is consistent with the phylogeny based on nuclear DNA sequences (Figure 2). The detection of the *atp6* intron and the region containing ORF 1 and 2 provides a polar and datable marker for evolutionary studies of the fungus.

Effector prediction

The main reason for carrying out this study was to facilitate the identification of effectors from the fungal genomes. Experience from a range of pathogens has generated a set of criteria that can be used to select effector candidates for experimental validation. These criteria (Tables 1 and 5) were used to filter and rank 12382 genes so that they can be efficiently prioritized for functional characterization. This process predicted the subsequently verified effector *SnTox1* that conforms to all eight criteria (Liu *et al.* 2012). *SnTox3* and *SnToxA* match 5 and 6 of the criteria respectively indicating that we still have an imperfect knowledge of effectors' properties.

Context of effector genes

All three known effector genes were absent from Sn79. Comparisons of the three genomes can be used to infer hypotheses about the evolutionary history of these genes (Figure 5) albeit the short contig lengths limit the conclusions. *SnTox1* is absent from Sn79 but the flanking sequences are homologous in the other two strains. *SnTox3* is absent from Sn79 along with a section of at least three genes. Both of these scenarios are consistent with a horizontal gene transfer hypothesis (Oliver *et al.* 2008) although gene loss from Sn79 is also a possibility; the source of the genes is unknown.

Multiple studies provide evidence supporting recent horizontal transfer for *ToxA* from *S. nodorum* to *P. tritici-repentis* (Friesen *et al.* 2006) and earlier from an unknown source into *S. nodorum* (Stukenbrock and McDonald 2007). The new data in this paper and the *P. tritici-repentis* genome sequence (Manning *et al.* 2013) allow us to speculate on the history of this critical region. Sn79 lacks *ToxA* along with a long section (at least 72 kb) of single copy DNA. This finding suggests that *ToxA* was transferred into the ancestor of both wheat pathogenic strains along with at least this length of DNA (Figure 6). Subsequently, in some strains such as SN15, this section has been invaded by transposons and subjected to RIP (Figure 7). The most parsimonious hypothesis for the source and structure of the *ToxA* region in *P. tritici-repentis* is an *S. nodorum* strain that had not yet suffered this transposon invasion and RIP. The recent acquisition, the homothallic nature of *P. tritici-repentis* and the scarcity of RIP have resulted in the monomorphism of *PtrToxA* (Tan *et al.* 2012) and the retention of genes lost to RIP in *S. nodorum.* The presence of both *ToxA*-expressing and nonexpressing strains of both species in different proportions in different parts of the world (Stukenbrock and McDonald 2007) can be accounted for by asexual propagation, sexual crossing and differential selection by wheat cultivars depending on whether they express the recognition gene *Tsn1* (Oliver *et al.* 2009; Antoni *et al.* 2010; Faris *et al.* 2010). Confirmation and elaboration of these hypotheses must await the assembly of a larger sample of genome sequences from both species.

Table 5 Counts of proteins matching the eight criteria from Table 3

Minimum Effector Criteria Matched	No. sn15 Proteins
8	1
7	10
6	159
5	904
4	3158

The criteria are highly selective for the top effector candidates.

ACKNOWLEDGMENTS

We thank Bruce McDonald, Eva Stukenbrock, Viola Manning, and Lynda Ciuffetti for many insightful discussions. This research was supported by the Australian Grains Research and Development Corporation. The work was supported by iVEC through the use of advanced computing resources located at iVEC@Murdoch and by an award under the Merit Allocation Scheme on the NCI National Facility at the Australian National University.

LITERATURE CITED

Abeysekara, N. S., T. L. Friesen, B. Keller, and J. D. Faris, 2009 Identification and characterization of a novel host-toxin interaction in the wheat-*Stagonospora nodorum* pathosystem. Theor. Appl. Genet. 120: 117–126.

Abeysekara, N. S., J. D. Faris, S. Chao, P. E. McClean, and T. L. Friesen, 2012 Whole-genome QTL analysis of *Stagonospora nodorum* blotch resistance and validation of the SnTox4-Snn4 interaction in hexaploid wheat. Phytopathology 102: 94–102.

Anders, S., 2009 Visualization of genomic data with the Hilbert curve. Bioinformatics 25: 1231–1235.

Antoni, E. A., K. Rybak, M. P. Tucker, J. K. Hane, P. S. Solomon *et al.*, 2010 Ubiquity of ToxA and absence of ToxB in Australian populations of *Pyrenophora tritici*-repentis. Australas. Plant Pathol. 39: 63–68.

Bendtsen, J. D., H. Nielsen, G. von Heijne, and S. Brunak, 2004 Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol. 340: 783–795.

Bennett, R. S., S. H. Yun, T. Y. Lee, B. G. Turgeon, E. Arseniuk *et al.*, 2003 Identity and conservation of mating type genes in geographically diverse isolates of *Phaeosphaeria nodorum*. Fungal Genet. Biol. 40: 25–37.

Bringans, S., J. K. Hane, T. Casey, K. C. Tan, R. Lipscombe *et al.*, 2009 Deep proteogenomics; high throughput gene validation by multidimensional liquid chromatography and mass spectrometry of proteins from the fungal wheat pathogen *Stagonospora nodorum*. BMC Bioinformatics 10: 301.

Cambareri, E. B., B. C. Jensen, E. Schabtach, and E. U. Selker, 1989 Repeat-induced G-C to A-T mutations in Neurospora. Science 244: 1571–1575.

Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross *et al.*, 2008 MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res. 18: 188–196.

Casey, T., P. S. Solomon, S. Bringans, K. C. Tan, R. P. Oliver *et al.*, 2010 Quantitative proteomic analysis of G-protein signalling in *Stagonospora nodorum* using isobaric tags for relative and absolute quantification. Proteomics 10: 38–47.

Celio, G. J., M. Padamsee, B. T. M. Dentinger, R. Bauer, and D. J. McLaughlin, 2006 Assembling the fungal tree of life: constructing the structural and biochemical database. Mycologia 98: 850–859.

Chu, C.-G., J. D. Faris, S. Xu, and T. L. Friesen, 2010 Genetic analysis of disease susceptibility contributed by the compatible Tsn1-SnToxA and Snn1-SnTox1 interactions in the wheat-Stagonospora nodorum pathosystem. Theor. Appl. Genet. 129: 1451–1459.

Ciuffetti, L. M., R. P. Tuori, and J. M. Gaventa, 1997 A single gene encodes a selective toxin causal to the development of tan spot of wheat. Plant Cell 9: 135–144.

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

Cooley, R. N., and C. E. Caten, 1991 Variation in electrophoretic karyotype between strains of *Septoria nodorum*. Mol. Gen. Genet. 228: 17–23.

Crook, A. D., T. L. Friesen, Z. H. Liu, P. S. Ojiambo, and C. Cowger, 2012 Novel necrotrophic effectors from *Stagonospora nodorum* and corresponding host genes in winter wheat germplasm in the Southeastern U.S. Phytopathology . 102: 498–505.

De Wit, P. J. G. M., R. Mehrabi, H. A. Van Den Burg, and I. Stergiopoulos, 2009 Fungal effector proteins: past, present and future. Review Mol. Plant Pathol. 10: 735–747.

Doyle, J., and J. Doyle, 1987 A rapid DNA isolation procedure for small quantities of fresh leaf tissue. Phytochem. Bull. 19: 11–15.

Edgar, R. C., 2004 MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5: 113.

Faris, J. D., and T. L. Friesen, 2009 Reevaluation of a tetraploid wheat population indicates that the Tsnl-ToxA interaction is the only factor governing *Stagonospora nodorum* blotch susceptibility. Phytopathology 99: 906–912.

Faris, J. D., Z. Zhang, H. Lu, S. Lu, L. Reddy *et al.*, 2010 A unique wheat disease resistance-like gene governs effector-triggered susceptibility to necrotrophic pathogens. Proc. Natl. Acad. Sci. USA 107: 13544–13549.

Faris, J. D., Z. Zhang, J. B. Rasmussen, and T. L. Friesen, 2011 Variable expression of the *Stagonospora nodorum* effector SnToxA among isolates is correlated with levels of disease in wheat. Mol. Plant Microbe Interact. 24: 1419–1426.

Friesen, T. L., E. H. Stukenbrock, Z. Liu, S. Meinhardt, H. Ling *et al.*, 2006 Emergence of a new disease as a result of interspecific virulence gene transfer. Nat. Genet. 38: 953–956.

Friesen, T. L., S. W. Meinhardt, and J. D. Faris, 2007 The *Stagonospora nodorum*-wheat pathosystem involves multiple proteinaceous host-selective toxins and corresponding host sensitivity genes that interact in an inverse gene-for-gene manner. Plant J. 51: 681–692.

Friesen, T. L., J. D. Faris, P. S. Solomon, and R. P. Oliver, 2008 Host-specific toxins: effectors of necrotrophic pathogenicity. Cell. Microbiol. 10: 1421–1428.

Friesen, T. L., C. G. Chu, Z. H. Liu, S. S. Xu, S. Halley *et al.*, 2009 Host-selective toxins produced by *Stagonospora nodorum* confer disease susceptibility in adult wheat plants under field conditions. Theor. Appl. Genet. 118: 1489–1497.

Gentleman, R. C., V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling *et al.*, 2004 Bioconductor: open software development for computational biology and bioinformatics. Genome Biol. 5: R80.

Goodwin, S. B., S. B. M'Barek, B. Dhillon, A. H. J. Wittenberg, C. F. Crane *et al.*, 2011 Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genet. 7: 1002070.

Goto, N., P. Prins, M. Nakao, R. Bonnal, J. Aerts *et al.*, 2010 BioRuby: Bioinformatics software for the Ruby programming language. Bioinformatics (Oxford, England). 26: 2617–2619.

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

Hane, J. K., and R. P. Oliver, 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, J. K., R. G. Lowe, P. S. Solomon, K. C. Tan, C. L. Schoch *et al.*, 2007 Dothideomycete plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*. Plant Cell 19: 3347–3368.

Hane, J. K., A. Williams, and R. P. Oliver, 2011 Genomic and comparative analysis of the class Dothideomycetes, pp. 205–226 in *The Mycota*, edited by S. Poggeler, and J. Wostemeyer. Springer-Verlag, Berlin.

Horton, P., K. J. Park, T. Obayashi, N. Fujita, H. Harada *et al.*, 2007 WoLF PSORT: protein localization predictor. Nucleic Acids Res. 35: W585–587.

Ipcho, S. V. S., J. K. Hane, E. A. Antoni, D. Ahren, B. Henrissat *et al.*, 2012 Transcriptome analysis of *Stagonospora nodorum*: gene models, effectors, metabolism and pantothenate dispensability. Mol. Plant Pathol. 13: 531–545.

Kearse, M, R. Moir, A. Wilson, S. Stones-Havas, M. Cheung *et al.*, 2012 Geneious Basic: an integrated and extendable desktop software platform for the organization and analysis of sequence data. Bioinformatics. 28: 1647–1649.

Kent, W. J., 2002 BLAT - The BLAST-like alignment tool. Genome Res. 12: 656–664.

Knight, R., P. Maxwell, A. Birmingham, J. Carnes, J. G. Caporaso *et al.*, 2007 PyCogent: a toolkit for making sense from sequence. Genome Biol. 8: R171.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan *et al.*, 2007 Clustal W and Clustal X version 2.0. Bioinformatics 23: 2947–2948.

Li, H., and R. Durbin, 2009 Fast and accurate short read alignment with Burrows-Wheeler transform. Bioinformatics 25: 1754–1760.

Li, L., C. J. Stoeckert Jr, and D. S. Roos, 2003 OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13: 2178–2189.

Liu, Z., J. D. Faris, R. P. Oliver, K. C. Tan, P. S. Solomon *et al.*, 2009 SnTox3 acts in effector triggered susceptibility to induce disease on wheat carrying the Snn3 gene. PLoS Pathog. 5: e1000581.

Liu, Z., Z. Zhang, J. D. Faris, R. P. Oliver, R. Syme *et al.*, 2012 The cysteine rich necrotrophic effector SnTox1 produced by *Stagonospora nodorum* triggers susceptibility of wheat lines harboring Snn1. PLoS Pathog. 8: 1002467.

Luderer, R., M. J. D. De Kock, R. H. L. Dees, P. De Wit, and M. Joosten, 2002 Functional analysis of cysteine residues of ECP elicitor proteins of the fungal tomato pathogen *Cladosporium fulvum*. Mol. Plant Pathol. 3: 91–95.

Ma, L. J., H. C. Van Der Does, K. A. Borkovich, J. J. Coleman, M. J. Daboussi *et al.*, 2010 Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature 464: 367–373.

Malkus, A., Q. Song, P. Cregan, E. Arseniuk, and P. P. Ueng, 2009 Genetic linkage map of *Phaeosphaeria nodorum*, the causal agent of stagonospora nodorum blotch disease of wheat. Eur. J. Plant Pathol. 124: 681–690.

Manning, V. A., I. Pandelova, B. Dhillon, L. J. Wilhelm, S. B. Goodwin *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 (Bethesda) 3: 41–63.

Martin, M., 2011 Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet J. **17:** North America, 17, May. 2011. Available at: http://journal.embnet.org/index.php/embnetjournal/article/view/200. Accessed: April 23, 2013.

McKenna, A. H., M. Hanna, E. Banks, A. Sivachenko, K. Cibulskis *et al.*, 2010 The Genome Analysis Toolkit: a MapReduce framework for analyzing next-generation DNA sequencing data. Genome Res. 20: 1297.

Murray, G. M., and J. P. Brennan, 2009 Estimating disease losses to the Australian wheat industry. Australas. Plant Pathol. 38: 558–570.

Oliver, R. P., P. S. Solomon, J. Hane, M. Ferguson-Hunt, B. A. McDonald *et al.*, 2008 Multiple host-specific toxins, lateral gene transfer and gene loss in the evolution of cereal Pleosporalean pathogens, in *9th International Congress of Plant Pathology*, Torino, Italy.

Oliver, R. P., K. Rybak, P. S. Solomon, and M. Ferguson-Hunt, 2009 Prevalence of ToxA-sensitive alleles of the wheat gene *Tsn1* in Australian and Chinese wheat cultivars. Crop Pasture Sci. 60: 348–352.

Oliver, R. P., T. L. Friesen, J. D. Faris, and P. S. Solomon, 2012 *Stagonospora nodorum*: from pathology to genomics and host resistance. Annu. Rev. Phytopathol. 50: 23–43.

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

Rouxel, T., J. Grandaubert, J. Hane, C. Hoede, A. van de Wouw *et al.*, 2011 The compartmentalized genome of *Leptosphaeria maculans*: diversification of effectors within genomic regions affected by repeat induced Point mutations. Nat. Commun. 2: art. 202.

Selker, E. U., 1990 Premeiotic instability of repeated sequences in *Neurospora crassa*. Annu. Rev. Genet. 24: 579–613.

Selker, E. U., E. B. Cambareri, B. C. Jensen, and K. R. Haack, 1987 Rearrangement of duplicated DNA in specialized cells of *Neurospora*. Cell 51: 741–752.

Smit, A., R. Hubley, and P. Green, 1996–2004 RepeatMasker Open-3.0. Available at: www.repeatmasker.com. Accessed: April 23, 2013.

Solomon, P. S., R. G. T. Lowe, K. C. Tan, O. D. C. Waters, and R. P. Oliver, 2006a *Stagonospora nodorum*: cause of stagonospora nodorum blotch of wheat. Mol. Plant Pathol. 7: 147–156.

Solomon, P. S., T. J. G. Wilson, K. Rybak, K. Parker, R. G. T. Lowe *et al.*, 2006b Structural characterisation of the interaction between *Triticum aestivum* and the dothideomycete pathogen *Stagonospora nodorum*. Eur. J. Plant Pathol. 114: 275–282.

Stukenbrock, E. H., and B. A. McDonald, 2007 Geographical variation and positive diversifying selection in the host-specific toxin SnToxA. Mol. Plant Pathol. 8: 321–332.

Tan, K. C., R. D. Trengove, G. L. Maker, R. P. Oliver, and P. S. Solomon, 2009 Metabolite profiling identifies the mycotoxin alternariol in the pathogen *Stagonospora nodorum*. Metabolomics 5: 330–335.

Tan, K. C., R. P. Oliver, P. S. Solomon, and C. S. Moffat, 2010 Proteinaceous necrotrophic effectors in fungal virulence. Funct. Plant Biol. 37: 907–912.

Tan, K.-C., M. Ferguson-Hunt, K. Rybak, O. D. C. Waters, W. A. Stanley *et al.*, 2012 Quantitative variation in effector activity of ToxA Isoforms from *Stagonospora nodorum* and Pyrenophora tritici-repentis. Mol. Plant Microbe Interact. 25: 515–522.

Ter-Hovhannisyan, V., A. Lomsadze, Y. O. Chernoff, and M. Borodovsky, 2008 Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res. 18: 1979–1990.

Van de Wouw, A. P., A. J. Cozijnsen, J. K. Hane, P. C. Brunner, B. A. McDonald *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.

Yap, V. B., H. Lindsay, S. Easteal, and G. Huttley, 2010 Estimates of the effect of natural selection on protein-coding content. Mol. Biol. Evol. 27: 726–734.

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

Zhang, Y., C. L. Schoch, J. Fournier, P. W. Crous, J. de Gruyter *et al.*, 2009 Multi-locus phylogeny of Pleosporales: a taxonomic, ecological and evolutionary re-evaluation. Stud Mycol 64: 85–102S105.

Zhang, Z., T. L. Friesen, S. S. Xu, G. Shi, Z. Liu *et al.*, 2011 Two putatively homoeologous wheat genes mediate recognition of SnTox3 to confer effector-triggered susceptibility to *Stagonospora nodorum*. Plant J. 65: 27–38.

Communicating editor: N. D. Young

Chapter 3 | A first genome assembly of the barley fungal pathogen Pyrenophora teres f. teres.

Attribution Statement

Authors: Simon R. Ellwood, Zhaohui Liu, **Robert A. Syme**, Zhibing Lai, James K. Hane, Felicity Keiper, Caroline S. Moffat, Richard P. Oliver, Timothy L. Friesen

Citation: A first genome assembly of the barley fungal pathogen Pyrenophora teres f. teres. SR Ellwood, Z Liu, RA Syme, Z Lai, JK Hane, F Keiper, CS Moffat, RP Oliver, TL Friesen - Genome Biology, 2010

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

The Ph.D. candidate (Robert A. Syme) made the following contributions to this chapter:

- Genome assembly, annotation and data management
- Assignment of functional terms to predicted peptides
- Analysis of protein homology
- Validation of genome by comparison to BAC sequences

Other contributions:

- SRE analysed the data, wrote the manuscript, and performed initial laboratory SSR genetic marker validation
- JKH provided scripts for SSR design
- SRE wrote the manuscript
- CSM assisted with comparisons of predicted Pyrenophora proteins
- ZL screened for polymorphic SSR markers on genetic mapping population parents and conducted genotyping, genetic map construction, and electrophoretic karyotyping
- FK contributed STMS markers
- ZL undertook the AFLP genotyping and the cytological karyotyping
- RPO and TLF contributed to the design of the project and provided assistance in finalizing the manuscript prior to publication

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

Signature: ……………………………………………

Date: 2015-05-19

I, Richard Oliver, certify that this attribution statement is an accurate record of Robert Syme's contribution to the research presented in this chapter.

Signature: ……………………………………………

Date: 2015-05-20

Genome Biology

RESEARCH Open Access

A first genome assembly of the barley fungal pathogen *Pyrenophora teres* f. *teres*

Simon R Ellwood[1*], Zhaohui Liu[2], Rob A Syme[1], Zhibing Lai[2], James K Hane[3], Felicity Keiper[4], Caroline S Moffat[5], Richard P Oliver[1] and Timothy L Friesen[2,6]

Abstract

Background: *Pyrenophora teres* f. *teres* is a necrotrophic fungal pathogen and the cause of one of barley's most important diseases, net form of net blotch. Here we report the first genome assembly for this species based solely on short Solexa sequencing reads of isolate 0-1. The assembly was validated by comparison to BAC sequences, ESTs, orthologous genes and by PCR, and complemented by cytogenetic karyotyping and the first genome-wide genetic map for *P. teres* f. *teres*.

Results: The total assembly was 41.95 Mbp and contains 11,799 gene models of 50 amino acids or more. Comparison against two sequenced BACs showed that complex regions with a high GC content assembled effectively. Electrophoretic karyotyping showed distinct chromosomal polymorphisms between isolates 0-1 and 15A, and cytological karyotyping confirmed the presence of at least nine chromosomes. The genetic map spans 2477.7 cM and is composed of 243 markers in 25 linkage groups, and incorporates simple sequence repeat markers developed from the assembly. Among predicted genes, non-ribosomal peptide synthetases and efflux pumps in particular appear to have undergone a *P. teres* f. *teres*-specific expansion of non-orthologous gene families.

Conclusions: This study demonstrates that paired-end Solexa sequencing can successfully capture coding regions of a filamentous fungal genome. The assembly contains a plethora of predicted genes that have been implicated in a necrotrophic lifestyle and pathogenicity and presents a significant resource for examining the bases for *P. teres* f. *teres* pathogenicity.

Background

Net blotch of barley (*Hordeum vulgare*) is caused by *Pyrenophora teres* Drechsler (anamorph *Drechslera teres* [Sacc.] Shoem.). *P. teres* is an ascomycete within the class Dothideomycetes and order Pleosporales. This order contains plant pathogens responsible for many necrotrophic diseases in crops, including members of the genera *Ascochyta*, *Cochliobolus*, *Pyrenophora*, *Leptosphaeria* and *Stagonospora*. Net blotch is a major disease worldwide that causes barley yield losses of 10 to 40%, although complete loss can occur with susceptible cultivars in the absence of fungicide treatment [1]. In Australia the value of disease control is estimated at $246 million annually with average direct costs of $62 million annually, making it the country's most significant barley disease [2].

Net blotch exists in two morphologically indistinguishable but genetically differentiated forms: *P. teres* f. *teres* (net form of net blotch, NFNB) and *P. teres* f. *maculata* (spot form of net blotch, SFNB) [3,4]. These forms have been proposed as distinct species based on the divergence of *MAT* sequences in comparison to *Pyrenophora graminea* [4]. Additionally, it has been suggested that limited gene flow may occur between the two forms [5,6]. As their names indicate, the two forms show different disease symptoms. NFNB produces lattice-like symptoms, in which necrosis develops along leaf veins with occasional transverse striations. SFNB displays more discrete, rounded lesions, often surrounded by a chlorotic zone. NFNB and SFNB may both be present in the same region but with one form prevailing in individual locales. NFNB has historically been regarded as the

* Correspondence: srellwood@gmail.com
[1]Department of Environment and Agriculture, Curtin University, Kent Street, Bentley, Perth, Western Australia 6102, Australia
Full list of author information is available at the end of the article

more significant of the two diseases, but in recent years there have been reports of SFNB epidemics, notably in regions of Australia and Canada [7,8].

Only recently have researchers begun to focus on the molecular and genetic aspects of *P. teres* pathogenesis and host-pathogen interactions. NFNB is known to produce non-host selective low molecular weight compounds that cause chlorosis on barley leaves [9]. Both forms also produce phytotoxic proteinaceous effectors in culture [10,11]. It has been suggested that these effectors are responsible for the brown necrotic component of the disease symptoms on susceptible cultivars. Host resistance to *P. teres* appears to conform to the gene-for-gene model [12]. Both dominant and recessive resistance loci have been reported that are genetically distinct. These are host genotype, form, and isolate specific, and occur along with multigenic/quantitative resistance on each of the barley chromosomes [13,14].

Little is known at the molecular level about the mechanisms of *P. teres* pathogenicity, with neither the mechanism of virulence nor host resistance known. A genome assembly offers a powerful resource to assist the dissection of virulence mechanisms by providing suites of genetic markers to characterize and isolate genes associated with virulence and avirulence via map-based cloning. It also enables potential effector candidate genes to be identified from partially purified active fractions in conjunction with mass spectrometry peptide analysis. The sequencing and assembly of fungal genomes to date have relied primarily on Sanger sequencing with read lengths of 700 to 950 bp. Several newer sequencing technologies are now available that are orders of magnitude less expensive, although currently they exhibit shorter read lengths. These include Roche/454 pyrosequencing (400 to 500 bp) and Illumina/Solexa sequencing (currently up to 100 bp). Recent improvements, including paired-end sequencing (reads from each end of longer DNA fragments) and continuing increases in read lengths should make the *de novo* assembly of high quality eukaryotic genomes possible.

Filamentous fungal genomes are relatively small and contain a remarkably consistent number of genes. Their genomes range in size from 30 to 100 Mbp and contain 10,000 to 13,000 predicted genes [15]. Their reduced complexity and small size relative to most eukaryotes makes them amenable to assessing the suitability of new sequencing technologies. These technologies have recently been described in the assembly of the filamentous fungus *Sordaria macrospora* [16], which involved a hybrid assembly of Solexa 36-bp reads and 454 sequencing. The objectives of this study were to assemble the genome of *P. teres* f. *teres* based on Solexa sequencing chemistry only, to validate the assembly given the short read lengths (in this study, 75-bp paired ends), and to provide initial characterization of the draft genome. We have complemented the assembly with the first cytogenetic visualization and genome-wide genetic map for this species.

Results

The genome of *P. teres* f. *teres* isolate 0-1 was sequenced using Illumina's Solexa sequencing platform with paired-end 75-bp reads. The Solexa run in a single flow cell yielded over 833 Mbp of sequence data, or approximately 20 times coverage of the final assembly length. Optimal kmer length in the parallel assembler Assembly By Short Sequences (ABySS) v. 1.0.14 [17] occurred at k = 45 and *n* = 5. This yielded a N_{50} where 50% of the assembly is contained in the largest 408 scaffolds and an L_{50} whereby 50% of the genome is contained in scaffolds of 26,790 bp or more. The total assembly size was 41.95 Mbp. Summary statistics of the assembly are presented in Table 1.

The Solexa sequencing reads that were used for the *P. teres* f. *teres* 0-1 genome assembly have been deposited in the NCBI sequence read archive [GenBank: SRA020836]. This whole genome shotgun project assembly has been deposited at DDBJ/EMBL/GenBank under the accession [GenBank: AEEY00000000]. The version described in this paper [GenBank: AEEY01000000] is the first version. Note NCBI does not accept contigs less than 200 bp in whole genome submissions, unless such sequences are important to the assembly, for example, they contribute to scaffolds or are gene coding regions. In addition, all scaffold nucleotide sequences, predicted coding region nucleotide sequences, and translated amino acid sequences are provided in Additional files 1, 2, and 3, respectively.

Both the initial contigs (composed of unpaired reads) and the scaffolds contained a large number of short sequences. In total there were 147,010 initial contigs with an N_{50} of 493 and an L_{50} of 22,178 bp. This

Table 1 *Pyrenophora teres* f. *teres* genome assembly key parameters

Parameter	Value
Size (bp)	41,957,260
G + C percentage	48
Predicted protein coding genes ≥100 amino acids	11,089
Predicted protein coding sequences ≥50 amino acids	11,799
Conserved proteins[a]	11,031
Unique hypothetical proteins	766
Percent complete	97.57
Mean gene size (bp)	1411
Mean exon size (bp)	557
Mean number of exons per gene	2.53

[a]Significant at an *e*-value cutoff of $\leq 10^{-5}$.

compared with a total of 146,737 scaffolds. The majority of initial contigs (140,326 of 147,010) were 200 bp or less, and were shared with the scaffold file. Such short contigs are a result of reads from repetitive regions. In AbySS, where highly similar repetitive regions occur, a 'bubble' removal algorithm simplifies the repeats to a single sequence. Thus, short isolated 'singletons' occur that were not assembled into scaffolds. Gene rich, more complex regions of the genome were represented by 6,684 scaffolds containing over 80% of the assembled sequences.

The assembly contains 11,799 predicted gene models of 50 amino acids or more. Most of the predicted genes (93.5%) were conserved within other species and of these conserved genes, 45.2% showed very high homology with a BLASTP *e*-value of 0. As a further confirmation of the success in capturing gene-rich regions, the percentage of complete genes (genes with defined start and stop codons) was 97.57%.

To validate the assembly over relatively large distances, the assembly was compared to two Sanger sequenced BACs, designated 8F17 and 1H13. Direct BLASTN [18] against assembly scaffolds showed that complex or regions with a high GC content assembled effectively (Figure 1). BAC 1H13 contains several low-complexity regions containing repetitive sequences, in which Solexa reads were over-represented and where only short scaffold assemblies are evident (Additional file 4).

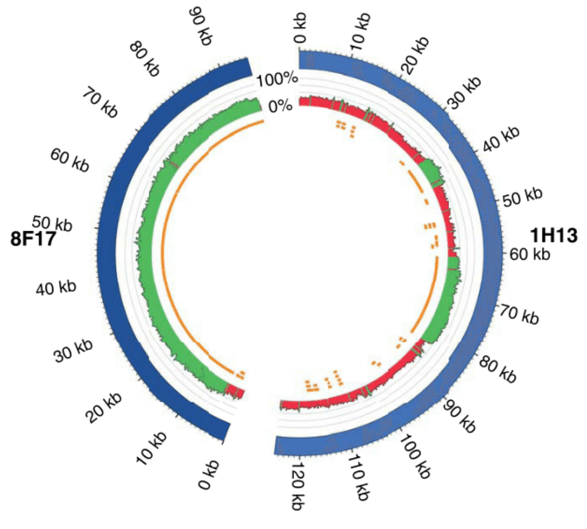


Figure 1 Comparison of the *P. teres* f. *teres* Solexa assembly with Sanger-sequenced BACs using CIRCOS [69]. BACs 8F17 and 1H13 are represented in blue. Percent GC is shown in the middle track with regions >40% shown in green and regions <40% shown in red. The inner track shows assembly scaffold BLASTN hits to the BACs.

To validate the assembly over short distances of moderately low complexity, and to provide a resource for genetic mapping and genetic diversity studies, we created a set of simple sequence repeats (SSRs). Motif repeats ranged in size from 34 bp with 100% identity and 0% indels to 255 bp with 64% identity and 1% indels. We examined the amplification of a subset (75) of the primer pairs and all gave unambiguous single bands and robust amplification. Primer characteristics and amplicon sizes for the 75 SSRs are provided in Additional file 5. The markers also readily amplified single bands in an isolate of *P. teres* f. *maculata*, albeit with slightly lower efficiency in 20% of the reactions. As a demonstration of their utility, three markers that were polymorphic between *P. teres* f. *teres* and f. *maculata* were used to fingerprint eight randomly selected isolates of each form (Table 2). Markers $(ACA)_{18}$-34213 and $(CTG)_{19}$-61882 were highly polymorphic in *P. teres* f. *teres* and f. *maculata*, respectively, with eight and five alleles. Form-specific diagnostic band sizes are evident

Table 2 Inter-form amplification of genome assembly-derived simple sequence repeat markers

Isolate	Marker[a]		
	$(ACA)_{18}$-34213	$(CAT)_{13}$-49416	$(CTG)_{19}$-61882
P. teres f. *teres*			
Cad 1-3	161	230	177
Cor 2	206	242	180
Cun 1-1	200	230	177
Cun 3-2	215	230	180
NB100	182	230	177
OBR	197	242	180
Stir 9-2	185	228	177
Won 1-1	256	242	177
Number of alleles	8	3	2
P. teres f. *maculata*			
WAC10721	197	230	196
WAC10981	149	221	189
WAC11177	149	218	189
WAC11185	149	221	189
Cad 6-4	149	221	196
Mur 2	149	221	186
NFR	149	221	199
SG1-1	149	221	190
Number of alleles	2	3	5

Examples of allele sizes from three SSRs are shown for eight randomly selected *P. teres* f. *teres* and *P. teres* f. *maculata* isolates. [a]Includes SSR motif, template copy number in subscript, and numeric identifier.

from the data, but with overlap in the ranges of allele sizes of each form for $(CAT)_{13}$-49416, and for $(ACA)_{18}$-34213 at 197 bp.

In addition to the above assembly validations, we compared 50 randomly selected non-homologous ESTs against the assembly to determine their presence; 49 gave unambiguous matches, with the highest *e*-value cutoff $<10^{-80}$, and one gave no hit. This orphan EST showed no BLASTX similarity to any sequence in GenBank and might be regarded as a library contaminant. Forty-seven (96%) of the remaining ESTs were predicted by GeneMark.

Electrophoretic and cytological karyotyping of *P. teres* f. *teres*

To estimate the genome size of *P. teres* f. *teres* by pulsed-field gel electrophoresis (PFG), isolate 0-1 was examined and compared to isolate 15A. Isolate 0-1 showed at least seven chromosome bands as indicated in Figure 2, with estimated sizes of 6.0, 4.9, 4.7, 3.9, 3.6, 3.4, and 3 Mbp. The brightness of the band at 6.0 Mbp indicated the presence of at least two chromosomes, and was further resolved into bands of 5.8 and 6.2 Mbp on a second longer electrophoresis run (image not shown). The relative brightness of the 3.4 Mbp band indicates two and possibly three chromosomes are co-migrating. The smallest band visible in Figure 2 is less than 1 Mbp and is most likely mitochondrial DNA. Thus, there is a minimum of nine and as many as eleven chromosomes present in isolate 0-1. This gave an estimated genome size of between 35.5 and 42.3 Mbp. Isolate 15A shows conspicuous differences in the lengths of the chromosomes for intermediate sized bands (greater than 3 Mbp and less than 6 Mbp), and appears to have two bands around 3 Mbp.

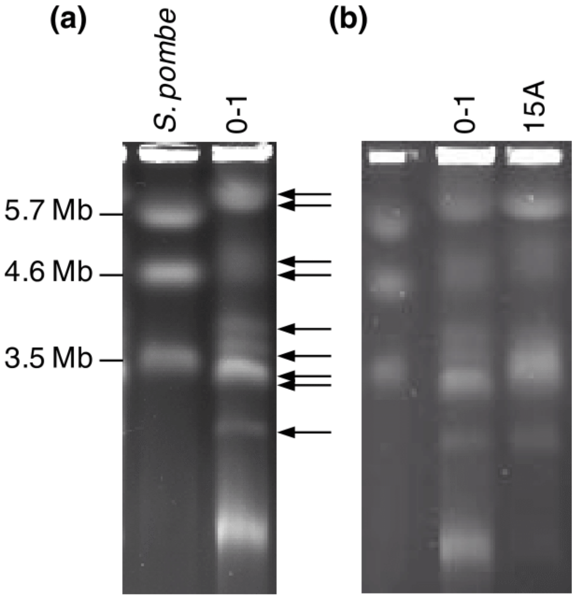


Figure 2 CHEF (clamped homogenous electric fields) separations of *P. teres* f. *teres* chromosomes. **(a)** Electro-karyotypes of isolate 0-1 with nine chromosomal bands indicated. **(b)** Chromosome level polymorphisms between isolates 0-1 and 15A.

Cytological karyotyping of isolate 0-1 using the germ tube burst method (GTBM) is depicted in Figure 3. Most of the discharged nuclei (above 90%) were observed at interphase (Figure 3a) where the chromosomes exist in the form of chromatin and are enclosed by the nuclear membrane. Of the remaining 10%, most of the chromosomes were either in early metaphase or clumped and entangled together, making it difficult to distinguish chromosomes (Figure 3b). In a few nuclei, condensed metaphase chromosomes were spread out sufficiently and we were able to count at least nine chromosomes (highlighted in Figure 3c). The four largest chromosomes are longer than or equal to 2 μm. The remainder depicted are smaller, but likely to be longer than 1 μm. The four largest chromosomes likely correspond to the four bands shown in PFG electrophoresis that have sizes greater than 3.9 Mbp.

Gene content

The genome assembly as a whole contains many predicted genes that have been implicated in pathogenicity. Genes encoding efflux pumps have roles in multidrug and fungicide resistance and toxic compound exclusion. For example, the ABC1 transporter in *Magnaporthe grisea* protects the fungus against azole fungicides and the

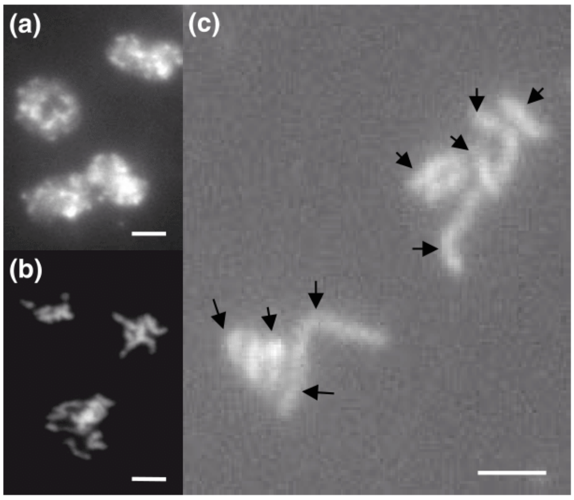


Figure 3 Visualization of *P. teres* f. *teres* chromosomes using the germ tube burst method (GTBM). **(a)** Nuclei at interphase. **(b)** Nuclei at early metaphase. **(c)** Condensed metaphase chromosomes with nine larger chromosomes indicated. Scale bars = 2 μm.

rice phytoalexin sakuranetin [19]. These genes are especially prevalent, with 79 homologues including representatives of the ATP-binding cassette (ABC), major facilitator, and multi antimicrobial extrusion protein superfamilies. Proteins encoded by other notable gene family members are the highly divergent cytochrome P450 s [20], which are involved in mono-oxidation reactions, one member of which has been shown to detoxify the antimicrobial pea compound pisatin [21]; the siderophores, which contribute to iron sequestration and resistance to oxidative and abiotic stresses but which also have essential roles in protection against antimicrobials and formation of infection structures [22,23]; and the tetraspanins, which are required for pathogenicity in several plant pathogenic fungi, one of which is homologous to the newly uncovered Tsp3 family [24].

Genome-specific expansion of non-orthologous gene families

Cluster analysis of *P. teres* f. *teres* genes in OrthoMCL [25] against the closely related Dothideomycetes species for which genomes and/or ESTs have been made publicly available (*Pyrenophora tritici-repentis*, *Cochliobolus heterostrophus*, *Stagonospora nodorum*, *Leptosphaeria maculans*, *Mycosphaerella graminicola*, together with two *Ascochyta* spp. sequenced in-house, *Ascochyta rabiei* and *Phoma medicaginis* (Ramisah Mod Shah and Angela Williams, personal communication) was used to reveal *P. teres* f. *teres*-specific expansion of gene families. The largest group of these were new members of class I and II transposable elements (Figure 4). Class I transposable elements are retrotransposons that use a RNA intermediate and reverse transcriptase to replicate, while class II transposons use a transposase to excise and reinsert a copy. In total, 36 clusters of new class I and II transposable elements are present in the assembly.

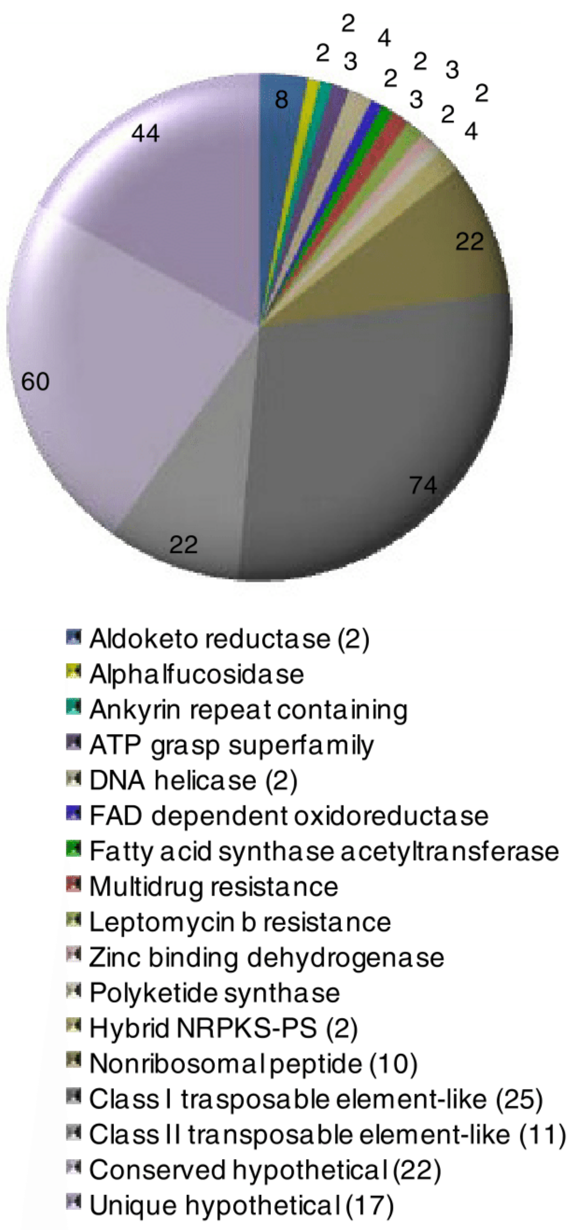


Figure 4 Expanded *P. teres* f. *teres* gene clusters. The number of non-orthologous and paralogous genes in each class of genes (as defined by OrthoMCL [25]) is shown at the end of each chart slice and the number of clusters greater than 1 is given in the key.

A prominent feature of expanded gene families in *P. teres* f. *teres* is a substantial expansion in specialized multi-functional enzymes known as non-ribosomal peptide synthetases (NRPSs) and polyketide synthases (PKSs) that produce secondary metabolites. The non-orthologous NRPSs are present in 10 clusters of 22 genes. NRPSs catalyze the production of cyclic peptides to form a diverse range of products, including antibiotics and siderophores, and are known to be phytotoxic [26]. Among plant pathogenic Pleosporales fungi, HC toxin from *Cochliobolus carbonum* [27] and AM toxin from *Alternaria alternata* [28] are notable examples. Also evident are hybrid NRPS-PKSs [29] in two clusters of four genes. PKSs produce polyketides in a manner similar to fatty acid biosynthesis. In fungi, better known polyketides are the mycotoxins fumonisin and autofusarin, and the phytotoxin cercosporin [30]. Hybrid NRPS-PKSs occur where PKS and NRPS modules coexist and add to the complexity of secondary metabolites. Most of the remaining non-orthologous gene clusters include homologues to genes involved with secondary metabolism and signaling. Investigations into the functional significance of these genes may provide new insights into the requirements of this pathogen. Also present are six non-orthologous genes encoding antibiotic and multi-drug resistance proteins that may have a role against toxic plant compounds. Indeed, the *P. teres* f. *teres* assembly as a whole contains ten genes with homology to ABC drug transporters.

Secreted proteins

Comparisons between plant pathogenic ascomycetes *S. nodorum* and *M. grisea* with the saprophyte *Neurospora crassa* [31,32] have both shown the expansion of secreted gene families consistent with their roles as plant pathogens. *P. teres* f. *teres* contains a large number of genes (1,031) predicted to be secreted by both WolfPSORT [33] and SignalP [34]. A significant proportion of these genes in *P. teres* f. *teres* (85%) are homologous with *P. tritici repentis*, as might be expected given their close phylogenetic relationship. This contrasts with 54% of the predicted genes in *S. nodorum* for which no phylogenetically close relative was sequenced [32]. Of the remaining genes, a small number (1.6%) show strongest homology to species outside the Pleosporales, while 6% are unique to *P. teres* f. *teres* isolate 0-1 with no functional annotation. These genes may include genes that have been laterally transferred.

In Blast2GO [35,36], 61.6% of the predicted genes were annotated with Gene Ontology (GO) terms. GO annotations are limited to well characterized genes but they do provide a useful overview. A large proportion of predicted genes encode proteins associated with plant cell wall and cutin degradation, presumably to degrade plant tissue during necrotrophic growth. Most are protein and carbohydrate hydrolases, together with carbohydrate binding proteins that target various polysaccharides (Table 3). For example, there are nine and seven predicted gene products with homology to cellulose binding proteins and cellulases, respectively, and five and four predicted gene products with homology to cutin binding proteins and cutinases, respectively. Predicted proteins annotated with the GO term ‘pathogenesis’ include homologues of glycosyl hydrolases, cutinase precursors, surface antigens, and a monoxygenase related to maackiain detoxification protein from *Nectria haematococca* [37].

Table 3 Common GO terms associated with genes predicted to be secreted

GO identifier	Description	Number of genes
Biological process		
GO:0006508	Proteolysis	42
GO:0055114	Oxidation reduction	25
GO:0043581	Mycelium development	23
GO:0051591	Response to cAMP	16
GO:0045493	Xylan catabolic process	14
GO:0009405	Pathogenesis	9
GO:0034645	Macromolecule biosynthesis	8
GO:0044248	Cellular catabolic process	7
GO:0021700	Developmental maturation	7
GO:0006139	Nucleic acid metabolism	7
GO:0050794	Regulation of cellular process	7
GO:0006629	Lipid metabolic process	7
GO:0019222	Metabolic regulation	6
GO:0016998	Cell wall catabolic process	6
GO:0034641	Nitrogen metabolism	6
GO:0030245	Cellulose catabolic process	6
GO:0006032	Chitin catabolic process	6
GO:0006979	Response to oxidative stress	6
GO:0009847	Spore germination	6
GO:0007154	Cell communication	5
GO:0006464	Protein modification process	5
Molecular function		
GO:0016787	Hydrolase activity	193
GO:0043167	Ion binding	84
GO:0016491	Oxidoreductase activity	73
GO:0048037	Cofactor binding	36
GO:0000166	Nucleotide binding	36
GO:0030246	Carbohydrate binding	26
GO:0046906	Tetrapyrrole binding	16
GO:0001871	Pattern binding	14
GO:0016740	Transferase activity	13
GO:0016829	Lyase activity	9
GO:0005515	Protein binding	6
GO:0016874	Ligase activity	6
GO:0016853	Isomerase activity	6

Terms are filtered for ≥5 members; molecular function GO terms are limited to GO term level 3

Marker development and linkage map construction

A total of 279 amplified fragment length polymorphisms (AFLPs) were generated that were polymorphic between the mapping population parents 15A and 0-1 using 96 primer combinations of 8 *Mse*I primers and 12 *Eco*RI primers (Additional file 6). On average, each pair produced approximately three polymorphic AFLPs. We identified a total of 68 polymorphic SSRs for genetic mapping; 44 from the genome assembly sequence, 20 from sequence tagged microsatellite site (STMS) markers [38], and 4 from ESTs (Additional file 5). In addition to AFLPs and SSRs, five random amplified polymorphic DNA markers associated with *AvrHar* [39] and the mating type locus were genotyped across 78 progeny from the 15A × 0-1 cross. All markers were tested for segregation ratio distortion; 69 (19%) were significantly different from the expected 1:1 ratio at P = 0.05, of which 32 were distorted at P = 0.01.

The genetic map was initially constructed with a total of 354 markers composed of 279 AFLPs, 68 SSRs, 5 random amplified polymorphic DNA markers, and a single mating type locus marker. The markers were first assigned into groups using a minimum LOD (logarithm of the odds) threshold of 5.0 and a maximum θ = 0.3. We excluded 111 markers from the map because they had a LOD <3 by RIPPLE in MAPMAKER [40]. The final genetic map was composed of 243 markers in 25 linkage groups, with each linkage group having at least 3 markers. The map spans 2,477.7 cM in length, with an average marker density of approximately one marker per ten centiMorgans (Figures 5 and 6). Individual linkage groups ranged from 24.9 cM (LG25) to 392.0 cM (LG1), with 3 and 35 markers, respectively. Three of the linkage groups had a genetic distance greater than 200 cM and 10 linkage groups had genetic distances of less than 50 cM, leaving 12 medium-sized linkage groups ranging between 50 and 200 cM. Other than a 30-cM gap on LG2.1, the markers are fairly evenly distributed on the linkage groups without obvious clustering. Linkage groups 2.1 and 2.2 are provisionally aligned together in Figure 5 as they may represent a single linkage group. This association is based on forming a single linkage group at LOD = 2, and by comparative mapping of SSR scaffold sequences with the *P. tritici-repentis* assembly (data not shown). The mating type locus mapped to linkage group LG4, and except for six of the small linkage groups, each linkage group has at least one SSR marker, which may allow comparisons to closely related genome sequences.

Discussion

This is the first wholly Illumina-based assembly of an ascomycete genome and the third assembly to be reported for a necrotrophic plant pathogenic ascomycete [31,32]. As might be expected, the *P. teres* f. *teres* genome assembly demonstrates that the short paired-end reads can be used to effectively capture higher complexity gene-containing regions. The assembly was validated by comparison to BAC sequences, ESTs and by direct amplification of predicted sequences across SSRs. Based on the published assemblies for the phytopathogens *M. grisea* and *S. nodorum* [31,32], the number of predicted genes in *P. teres* f. *teres* is similar (11,089 versus 11,109 and 10,762, for genes larger than 100 amino acids or *S. nodorum* version 2 gene models, respectively). Gene prediction algorithms, even when trained on ESTs from the species in question, are unlikely to correctly predict all coding regions in more complex genomes, and in some instances require further corroborating data from approaches such as proteomics and mass-spectrometry [41]. Thus, the true number of genes may be less dependent on the assembly *per se* and gene models may be further adjusted, concatenated or introduced.

The inevitable corollary of an assembly based on short paired-end reads is that low-complexity regions (containing low GC content, simple microsatellites and repetitive DNA) are under-represented. As a consequence, the assembly is composed of a large number of singleton contigs that are inappropriate for estimating the genomic proportions of such regions. To support the minimum estimate of the genome size based on the assembly, and to provide basic information on chromosome composition, we conducted PFG and GTBM karyotyping. From the PFG results, we concluded that *P. teres* f. *teres* most likely contains a minimum of 9 chromosomes but with band intensities suggesting 11 chromosomes is possible. This provided an estimated genome size of at least 35.5 Mbp and an upper value of 42.3 Mbp. Clumping and co-migration of bands is a common phenomenon in PFG, as shown, for example, by Eusebio-Cope *et al.* [42]. Resolution of co-migrating bands requires techniques such as Southern blotting [43] and fluorescence *in situ* hybridization [44] for accurate discrimination. However, the cytological karyotyping correlated with the PFG results in depicting at least nine chromosomes. An upper estimate of nine chromosomes was postulated for *P. teres* by Aragona *et al.* [45], although that study did not identify which *P. teres* form was examined, and the technique used gave poor resolution of bands between 4.5 and >6 Mbp. Overall, the total assembly size in this study correlates with the higher estimate by elecrophoretic karyotyping and indicates a genome of at least 42 Mbp. This is somewhat larger than the Pleosporales assemblies reported to date for *Cochliobolus heterostrophus* (34.9 Mbp; Joint Genome Institute), *P. tritici-repentis* (37.8 Mbp; NCBI) and *S. nodorum* (37.1 Mbp [32]).

An expansion in genome size compared to other Pleosporales might be explained by the presence in the assembly of new classes of transposable elements and large numbers of novel repeats (over 60, although these data are incomplete due to poor assembly of degraded regions and therefore have not been shown). These in turn may also explain the large PFG chromosomal level polymorphisms between the two isolates examined here and the relatively large genetic map. Chromosomal level polymorphisms are a feature of some ascomycetes [46]. Among plant pathogenic fungi, there is growing evidence that host-specificity genes and effectors are located in or next to transposon-rich regions [31,47]. This provides opportunities for horizontal acquisition, duplication and further diversification to generate new, species-specific genetic diversity or, where they are recognized as an avirulence gene, to be lost, a process that may also aid host range expansion. The contribution of transposons in *P. teres* f. *teres* pathogenicity has yet to be determined, although we have preliminary data showing that the avirulence gene *AvrHar* is associated with transposon repeats on the second largest chromosome. There is no evidence in *P. teres* f. *teres* for small chromosomes <2 Mbp, as in *N. haematococca* and *A.*

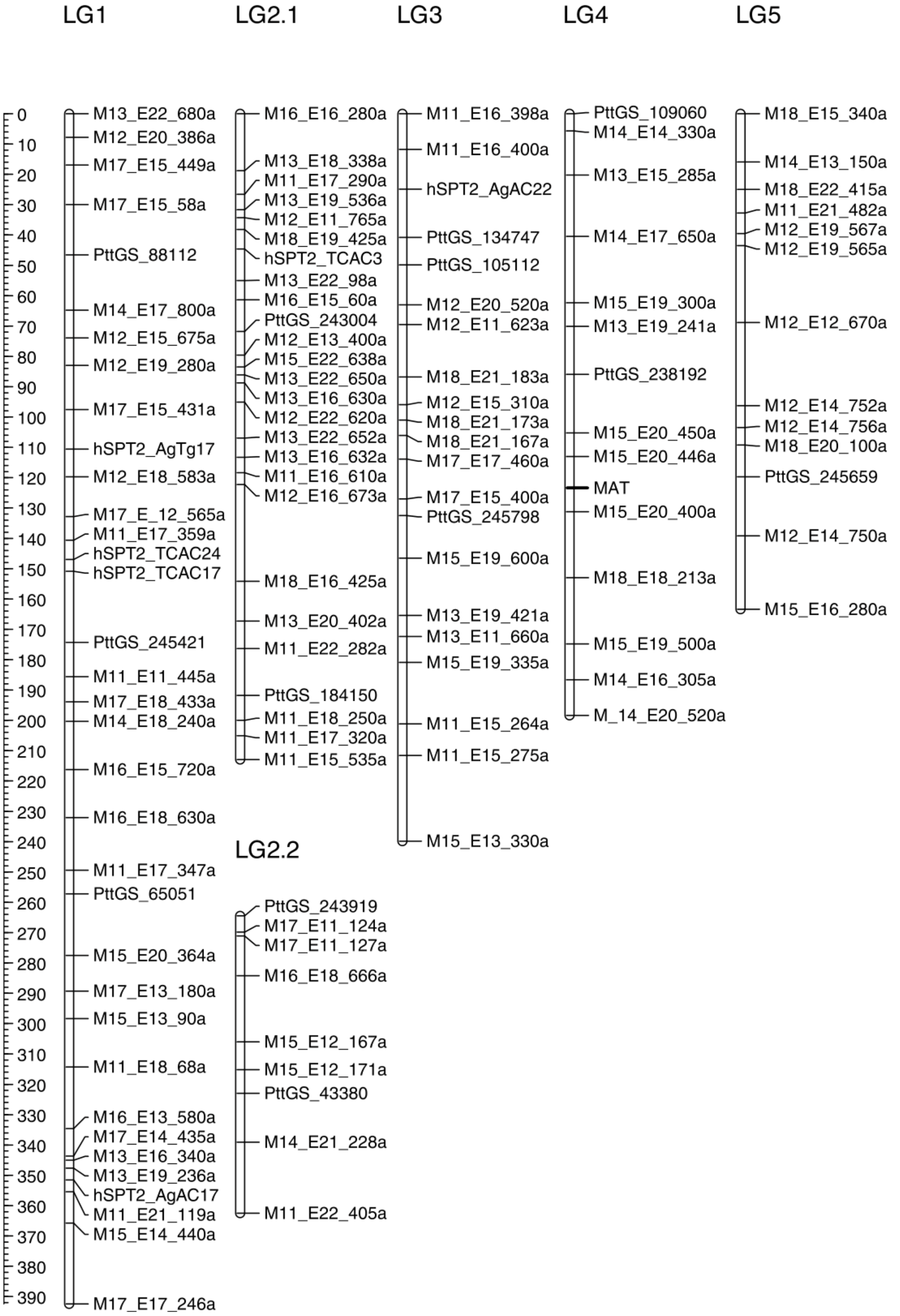


Figure 5 Genetic linkage map of *P. teres* f. *teres*. Linkage groups are drawn with genetic distance in cM on the scale bar to the left and are ordered according to their genetic length. AFLP markers are indicated by the *Mse*I (M) and *Eco*RI (E) primer combination (Additional file 6), followed by the size of the marker. SSR markers were developed from three sources: ESTs, STMSs and the genome assembly, prefixed PtESTSSR_, hSPT2_, and PttGS_, respectively. The mating type locus (MAT) is depicted in bold on linkage group 4.

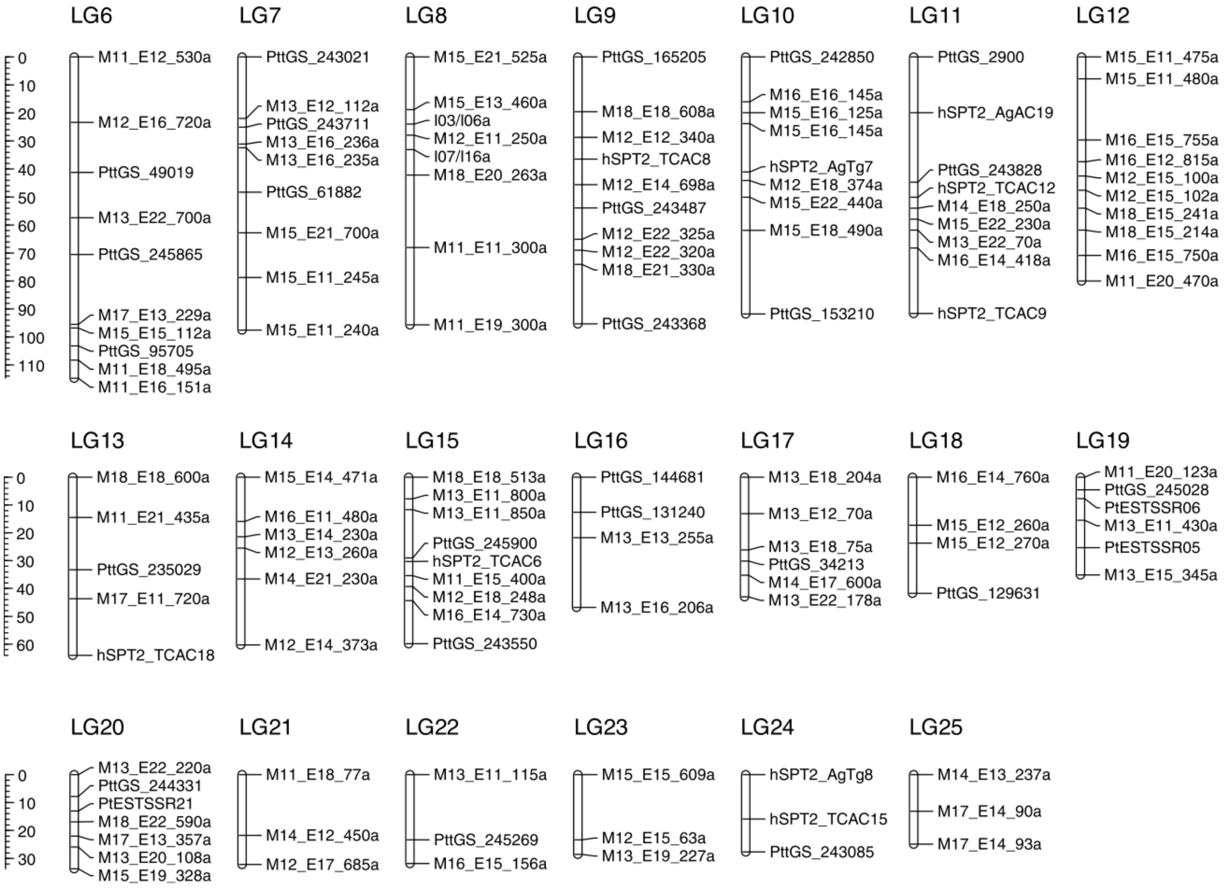


Figure 6 Genetic linkage map of *P. teres* f. *teres* continued from Figure 5. Linkage groups are drawn with genetic distance in cM on the scale bar to the left and are ordered according to their genetic length.

alternate, where they confer host-specific virulence [48,49], and in *Fusarium oxysporum*, where they have been demonstrated to be mobile genetic elements conferring virulence to non-pathogenic strains [50].

The analysis of the gene content of the genome assembly shows that it shares many of the characteristics of similar plant pathogenic fungi, and strong homology to most genes from *P. tritici-repentis*. These include highly diverse proteins involved in host contact, signal transduction, secondary metabolite production and pathogenesis. Secreted proteins are of particular interest to plant pathologists since they represent the key interface of host-pathogen interactions, notably avirulence proteins and effectors. These are key components of inducing disease resistance and promoting disease, while expressed effector proteins offer tangible discriminating resistance assay tools in a variety of breeding programs. This is because fungal necrotrophic disease is the sum of the contribution of individual effectors [51,52] and single, purified effectors give a qualitative response when infiltrated into leaves. However, effector genes often encode small, cysteine-rich proteins with little or no orthology to known genes. Examples include *Avr2* and *Avr4* in *Cladosporium fulvum*, *Avr3* in *F. oxysporum* (reviewed in [53]), *ToxA* and *ToxB* in *P. tritici repentis* [54,55] and *SnToxA* and *SnTox3* in *S. nodorum* [56,57]. Identifying candidate effectors in the genome assembly in conjunction with genetic mapping, functional studies and proteomic approaches will in future aid their isolation.

We provide the first genetic linkage map of *P. teres* f. *teres*. The total length is nearly 2,500 cM, longer than that reported for other ascomycete fungal pathogens; 1,216 cM for *M. graminicola* [58], 1,329 cM for *Cochliobus sativus* [59], and 900 cM for *M. grisea* [60]. However, a genetic map of 359 loci for the powdery mildew fungus *Blumeria graminis* f. sp. *hordei*, an obligate biotrophic pathogen of barley, covered 2,114 cM [61]. The length of the genetic map of *P. teres* f. *teres* may be a function of the relatively large genome size and the presence of large numbers of recombinogenic repetitive elements. This is paralleled by a greater number of linkage groups (25) compared to the estimated number of chromosomes that may also be suggestive of interspersed tracts of repetitive DNA.

The genetic map and karyotyping data will be instrumental in a final assembly of the *P. teres* f. *teres* genome, as they will allow scaffolds to be orientated and tiled onto linkage groups. A combination of the genome assembly and the genetic map provides an invaluable resource to identify potential effector candidate genes from phytotoxic protein fractions in conjunction with mass spectrometry peptide analysis. Genetically characterized SSRs provided in this study will also provide an important resource for the community in comparative mapping, gene-flow and genetic diversity studies. Further validation, assembly of low-complexity sequence regions, and genome annotation are now underway using proteomic approaches and 454 pyrosequencing. The priority now is to fully understand the mechanism of pathogenicity in *P. teres* f. *teres* in order to achieve a solution to control this pathogen.

Conclusions

This study demonstrates that the successful assembly of more complex and gene-rich regions of a filamentous fungus is possible using paired-end Solexa sequencing. The approach provides a cost-effective means of directly generating marker resources that would previously have been prohibitively expensive with modest research funding. At 42 Mbp or more, the genome of *P. teres* f. *teres* 0-1 is larger by comparison to closely related Pleosporales members, and has a correspondingly large genetic map. The genome is dynamic, in that different isolates show obvious chromosomal level differences, while fractionated linkage groups and the length of the genetic map also suggest an abundance of repetitive DNA. In common with other plant pathogens, *P. teres* f. *teres* contains a rich diversity of predicted genes, notably protein and carbohydrate hydrolases, efflux pumps, cytochrome P450 genes, siderophores, tetraspanins, non-ribosomal peptide synthetases and polyketide synthases, and a complex secretome that can be attributed to its lifestyle. Non-ribosomal peptide synthetases and efflux pumps in particular appear to have undergone a *P. teres* f. *teres*-specific expansion of non-othologous gene families. The assembly presented provides researchers with an excellent resource to further examine net blotch pathogenicity and plant-microbe interactions in general.

Materials and methods

Origin of *P. teres* isolates

The NFNB isolate sequenced in this study, 0-1, was originally collected in Ontario, Canada [39]. Isolate 15A (10-15-19), the opposite parental isolate used to develop a mapping population, was collected from Solano County, California [62]. The remaining NFNB isolates (Cad 1-3, Cor 2, Cun 1-1, Cun 3-2, NB100, OBR, Stir 9-2, and Won 1-1) were collected in Western Australia by S Ellwood in the 2009 barley growing season. SFNB isolates WAC10721, WAC10981, WAC11177, and WAC11185 were obtained from the Department of Agriculture and Food, Western Australia (3, Baron Hay Court, South Perth, Western Australia 6151); isolates Cad 6-4, Mur 2, NFR, and SG1-1 were collected in Western Australia by S Ellwood during 2009.

Electrophoretic and cytological karyotyping

Protoplasting and pulsed-field gel electrophoresis

Chromosome size and number were analyzed for North American NFNB isolates; 0-1 and 15A, previously used to develop a genetic cross for identifying avirulence genes [39,63]. Fungal protoplasts were prepared using a protocol established for *S. nodorum* as described by Liu *et al.* [56] with some modifications. Briefly, conidia were harvested from 7-day fungal cultures and inoculated into 60 ml liquid Fries medium in 250 ml Erlenmeyer flasks. After growth at 27°C in a shaker (100 rpm) for 48 h, the fungal tissue was then homogenized in a Waring blender and re-inoculated into 200 ml liquid Fries medium in 500 ml Erlenmeyer flasks. The fungus was grown under the same growth conditions for 24 h. Mycelium was harvested by filtering through two layers of Miracloth, washed thoroughly with water and finally with mycelial wash solution (MWS: 0.7 M KCl and 10 mM $CaCl_2$). Around 2 g (wet weight) of mycelial tissue was then transferred into a Petri dish (100 × 20 mm) containing 40 ml filter-sterilized protoplasting solution containing 40 mg/ml β-d-glucanase, 0.8 mg/ml chitinase, and 5 mg/ml driselase (Interspex Product Inc., San Mateo, CA, USA) in MWS. The Petri dish was shaken at 70 rpm at 28°C for at least 5 h. Protoplasts were filtered through four layers of Miracloth and pelleted by centrifugation at 2,000 × *g* for 5 minutes at room temperature, followed by another wash with MWS and pelleting. Protoplasts were resuspended in MWS to a final concentration of 2×10^8 protoplasts/ml and mixed with an equal volume of 2% low melting temperature agarose (Bio-Rad Laboratories, Hercules, CA, USA) dissolved in MWS. Agarose plugs were made by pipetting 80 μl of the mixture into plug molds (Bio-Rad Laboratories). Once solidified, plugs were placed in 20 ml Proteinase K reaction buffer containing 100 mM EDTA (pH 8.0), 1% N-lauroyl sarcosine, 0.2% sodium deoxycholate and 1 mg/ml Proteinase K (USBiological, Swampscott, MA, USA) at 50°C for 24 h. Plugs were washed four times in 10 mM Tris pH 8.0 and 50 mM EDTA for 1 h with gentle agitation, then stored in 0.5 M EDTA (pH 8.0) at 4°C. PFG was performed on a Bio-Rad CHEF Mapper system. Separation of chromosomes in the 1 to 6 Mb range was carried out in 1.0× TAE at 14°C using 0.8% Low EEO agarose gel (USBiological). Run time was 72 h at 2 V/cm (70 V) with a 20- to 40-minute switch time ramp at an angle of 106°.

Spore germination and germ tube burst cytological karyotyping

Conidia were washed with water from 7-day cultures grown on V8 potato dextrose agar (V8PDA) plates, filtered through two layers of miracloth and centrifuged at 3,000 × *g* for 5 minutes. Conidia were washed twice with potato dextrose broth and re-suspended in this with a final concentration of 4 × 10^5 spores/ml. Approximately 400 μl of spore suspension was placed onto slides coated with poly-L-lysine (Sigma-Aldrich Corp., St Louis, MO, USA) and covered by a 22 × 40 mm piece of parafilm to keep moist. All slides were kept in a sealed plastic box at room temperature for 3 h, and then moved to the fridge for cold treatment overnight. Slides were dipped in H_2O to carefully remove the covers and then placed in a methanol/acetic acid (22:3) solution overnight to fix fungal tissue. The slides were flame dried to burst cells and release chromosomes. Slides were stained for 5 minutes in the dark with 1 μg/ml 4',6-diamidino-2-phenylindole (DAPI; Sigma-Aldrich) and 1 μg/ml Flourescent Brightener 28 (Sigma-Aldrich) in anti-fade mounting solution. Slides were examined and photographed using a Zeiss Axioplan 2 epiflourescent microscope.

Genome sequence acquisition

Whole shotgun genome sequencing

DNA of *P. teres* f. *teres* isolate 0-1 was extracted using a Biosprint DNA Plant Kit and a BioSprint 15 automated workstation (Qiagen, Hilden, Germany). Genomic sequencing was performed on a Solexa sequencing platform at the Allan Wilson Centre (Massey University, Palmerston North, New Zealand). DNA preparation, cluster formation, primer hybridization and DNA amplification reactions were according to the manufacturer's recommended protocol [64]. DNA sequencing was performed using 75-bp paired-end reads of randomly sheared 200-bp fragments in a single flow cell. Data were pre-filtered in Illumina's Pipeline v.1.4 and IPAR v.1.3. Reads failing a 'chastity' filter of 0.6 were discarded. The steps described below for genome scaffold assembly, annotation and analysis were performed on the iVEC advanced computing facilities [65].

Paired-end scaffold assemblies

Single (split pairs) and paired-end reads were assembled using ABySS v.1.0.14 [17]. In addition to the read filtering described above, ABySS removes reads containing ambiguous characters (Ns). The optimal sequence kmer (overlap) length was determined by incrementally adjusting the kmer length by 4 bp and graphing the number of contigs against L_{50} for a given kmer length. The optimal kmer length occurred where N_{50} was minimal and L_{50} was largest as visualized by R [66]. N_{50} is a weighted median statistic such that 50% of the entire assembly is contained in the number of contigs or scaffolds equal to or greater than this value, while L_{50} is the length of the scaffold that separates the half of the assembled genome from the remainder of smaller scaffolds, if the sequences are ordered by size.

Annotation and analysis

Protein coding sequences were identified with GeneMark-ES v.2 [67]. GeneMark uses a self-training algorithm optimized for features of fungal gene organization by incorporating an enhanced intron submodel to accommodate sequences with and without branch point sites. GeneMark compares favorably with the accuracy of gene finders that employ supervised training based on cDNA sequences.

Annotation of predicted proteins was conducted with the following tools. A mirror of the NCBI database at iVEC, together with publicly available fungal protein sequence files not present at NCBI, was interrogated by BLASTP [18]. Blast2GO v.2.4.2 [35,36], which incorporates GO, KEGG maps, InterPro and Enzyme Codes was used with default parameters for functional annotation. *De novo* annotation of PFAM domains was performed using HMMER v.2.3.2 [68]. HMMER searches for homologues of protein sequences and implements methods using probabilistic models called 'profile hidden Markov models'. To detect orthologous genes, we used OrthoMCL [25] by BLAST to the NCBI non-redundant database with an *e*-value cutoff of $\leq 10^{-5}$. OrthoMCL is a genome-scale algorithm for grouping protein sequences between species based on BLAST similarity that was used to identify species-specific expanded gene families. Subcellular localization of proteins and secretion signals were identified with Wolf PSort [33] and SignalP v.3.0 [34] using default parameters and selection of the appropriate organism type.

Genome assembly validation

Assembly comparison with Sanger-sequenced BACs

To validate the assembly over a larger scale, BLASTN [18] was used to compare the assembly against two NFNB 0-1 BACs, designated 8F17 and 1H13, sequenced and assembled by The Genome Center (Washington University, St Louis, MO, USA). The data were visualized with CIRCOS [69]. To establish if all regions of the BACs were covered by Solexa sequencing, raw reads were mapped to the BACs with the Burrows-Wheeler Aligner [70] and visualized using R and the ggplot2 package [66,71].

SSR primer design and PCR amplification

Short tandem repeats or SSRs (also known as microsatellites) were identified by scanning the genome assembly with Tandem Repeat Finder v.4 [72] for a minimum of ten tandem repeats from 2 to 7 bp. Primers were designed using Primer3 [73] using parameters designed to minimize secondary structures, with a GC content >40%, and an optimum melting temperature of 58 to 60°C, for amplicons in a size range of 150 to 400 bp. The primers were assayed using single-spored *P. teres* isolates collected from different sites in Western Australia. DNA extraction and PCR amplification using the Multiplex Ready Technique were performed as described previously [74,75]. Allele sizing was performed using GeneMapper v.3.7 (Applied Biosystems, Foster City, CA, USA).

EST library preparation, sequencing, and assembly comparison

Total RNA was extracted from isolate 0-1 using fungal mycelium tissue grown in liquid Fries medium for 4 days. The RNA was extracted with TRIZOL (Invitrogen, Carlsbad, CA, USA) following the manufacturer's instructions. EST library construction and sequencing was conducted by The Genome Center. To investigate the presence of ESTs in the assembly and the efficiency of GeneMark predictions, unique EST sequences were BLASTN searched against the assembly. BLASTN hits were then compared against the location of GeneMark predicted coding regions with BEDtools [76].

Marker development and genetic linkage map construction

Lai *et al*. [63] used a subset of AFLPs to identify markers associated with fungal avirulence on the barley lines 'Harbin' and 'Prato' on two linkage groups. That study used a segregating population of 78 progeny from a cross between NFNB isolates 15A and 0-1. The AFLP markers were generated based on the technique of Vos *et al*. [77] and employed 96 primer combinations containing *Eco*RI and *Mse*I restriction sites. In this study, all available AFLPs from the 96 primer combinations were used to develop a comprehensive genetic map (Additional file 6). In addition, we incorporated polymorphic STMSs developed from microsatellite libraries by Keiper *et al*. [38], together with SSRs from EST sequences and the genome assembly herein. SSR PCR amplification and population genotyping were performed as described previously [38,78,79]. In addition, the mating type locus was assayed using primers Pt5 and Pt7 that amplify the *P. teres* HMG box [80].

Linkage map construction was performed with MAPMAKER v.2.0 for Macintosh as described by Liu *et al*. [78]. A minimum LOD value of 5.0 and a maximum θ = 0.3 were used to establish the linkage groups. For each linkage group, the most plausible order of markers was determined using commands 'FIRST ORDER' and 'RIPPLE', and markers with low confidence (LOD <3.0 for RIPPLE) were excluded from the map. All markers were tested for fitness of a 1:1 segregation ratio using *Qgene* [81]. The genetic map was drawn with the software program MapChart v.2.1 [82].

Additional material

Additional file 1: *P. teres* f. *teres* isolate 0-1 scaffold assembly nucleotide sequences.

Additional file 2: *P. teres* f. *teres* isolate 0-1 predicted coding region nucleotide sequences.

Additional file 3: *P. teres* f. *teres* isolate 0-1 predicted coding region translated amino acid sequences.

Additional file 4: Solexa read coverage of BACs 1H13 and 8F17.

Additional file 5: Characteristics of 75 genome assembly-derived SSRs and those polymorphic SSRs used in the *P. teres* f. *teres* 01 × 15A genetic map construction.

Additional file 6: AFLP di-nucleotide selective primer extensions and their codes.

Abbreviations

ABC, ATP-binding cassette; AFLP, amplified fragment length polymorphism; BAC, bacterial artificial chromosome; bp, base pair; DDBJ, DNA Data Bank of Japan; EST, expressed sequence tag; GO, Gene Ontology; GTBM, germ tube burst method; LOD, logarithm of the odds; MWS, mycelial wash solution; NCBI, National Center for Biotechnology Information; NFNB, net form of net blotch; NRPS, non-ribosomal peptide synthetase; PFG, pulsed-field gel electrophoresis; PKS, polyketide synthase; SFNB, spot form of net blotch; SSR, simple sequence repeat; STMS, sequence-tagged microsatellite site.

Acknowledgements

The authors would like to thank Dr Steven Xu and Dr CC Jan for their assistance in chromosomal microscopy. This research was funded by the Grains Research and Development Corporation under project UMU00025.

Author details

[1]Department of Environment and Agriculture, Curtin University, Kent Street, Bentley, Perth, Western Australia 6102, Australia. [2]Department of Plant Pathology, North Dakota State University, Fargo, North Dakota 58105, USA. [3]CSIRO Plant Industry, Centre for Environment and Life Sciences, Private Bag 5, Wembley, Western Australia 6913, Australia. [4]South Australian Research and Development Institute, Waite Institute, Adelaide, South Australia 5064, Australia. [5]Division of Health Sciences, Murdoch University, Murdoch Drive, Perth, Western Australia 6150, Australia. [6]USDA-ARS Cereal Crops Research Unit, Northern Crop Science Laboratory, 1307 18th Street North, Fargo, North Dakota 58105, USA.

Authors' contributions

SRE analyzed the data, wrote the manuscript, and performed initial laboratory SSR genetic marker validation. RAS provided informatics expertise in software implementation, scripting and primary data analysis (including for genome assembly, gene prediction, annotation of predicted peptides, and homology searches). JKH provided scripts to facilitate SSR design and data elucidation. CSM assisted with comparisons of predicted *Pyrenophora* proteins. ZL screened for polymorphic SSR markers on genetic mapping population parents and conducted genotyping, genetic map construction, and electrophoretic karyotyping. FK contributed STMS markers. ZL undertook the AFLP genotyping and the cytological karyotyping. RPO and TLF contributed to the design of the project and provided assistance in finalizing the manuscript prior to publication.

Received: 25 June 2010 Revised: 21 October 2010
Accepted: 10 November 2010 Published: 10 November 2010

References

1. Mathre DE: *Compendium of Barley Diseases.* 2 edition. St Paul MN, American Phytopathological Society; 1997.

2. Murray GM, Brennan JP: **Estimating disease losses to the Australian barley industry.** *Aust Plant Pathol* 2010, **39**:85-96.
3. Smedegård-Petersen V: ***Pyrenophora teres* f. *maculata* f. nov. and *Pyrenophora teres* f. *teres* on barley in Denmark.** *Kgl Vet Landbohojsk Arsskr* 1971, 124-144.
4. Rau D, Attene G, Brown A, Nanni L, Maier F, Balmas V, Saba E, Schäfer W, Papa R: **Phylogeny and evolution of mating-type genes from *Pyrenophora teres*, the causal agent of barley 'net blotch' disease.** *Curr Genet* 2007, **51**:377-392.
5. Campbell GF, Lucas JA, Crous PW: **Evidence of recombination between net- and spot-type populations of *Pyrenophora teres* as determined by RAPD analysis.** *Mycol Res* 2002, **106**:602-608.
6. Leisova L, Kucera L, Minarikova V, Ovesna J: **AFLP-based PCR markers that differentiate spot and net forms of *Pyrenophora teres*.** *Plant Pathol* 2005, **54**:66-73.
7. McLean MS, Howlett BJ, Hollaway GJ: **Spot form of net blotch, caused by *Pyrenophora teres* f. *maculata*, is the most prevalent foliar disease of barley in Victoria, Australia.** *Aust Plant Pathol* 2010, **39**:46-49.
8. Tekauz A: **Characterisation and distribution of pathogenic variation in *Pyrenophora teres* f. *teres* and *P. teres* f. *maculata* from western Canada.** *Can J Plant Pathol* 1990, **12**:141-148.
9. 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**:154-162.
10. 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**:907-915.
11. Sarpeleh A, Wallwork H, Tate ME, Catcheside DE, Able AJ: **Initial characterisation of phytotoxic proteins isolated from *Pyrenophora teres*.** *Physiol Mol Plant Pathol* 2008, **72**:73-79.
12. Flor HH: **Current status of the gene-for-gene concept.** *Annu Rev Phytopathol* 1971, **9**:275-296.
13. McLean MS, Howlett BJ, Hollaway GJ: **Epidemiology and control of spot form of net blotch (*Pyrenophora teres* f. *maculata*) of barley: a review.** *Crop Pasture Sci* 2009, **60**:303-315.
14. Liu Z, Ellwood SR, Oliver RP, Friesen TL: ***Pyrenophora teres*: profile of an increasingly damaging barley pathogen.** *Mol Plant Pathol* 2010.
15. **Properties of Eukaryotic Genome Sequencing Projects.** [http://www.ncbi.nlm.nih.gov/genomes/leuks.cgi].
16. Nowrousian M, Stajich JE, Chu M, Engh I, Espagne E, Halliday K, Kamerewerd J, Kempken F, Knab B, Kuo H-C, Osiewacz HD, Poggeler S, Read ND, Seiler S, Smith KM, Zickler D, Kuck U, Freitag M: ***De novo* assembly of a 40 Mb eukaryotic genome from short sequence reads: Sordaria macrospora, a model organism for fungal morphogenesis.** *PLoS Genet* 2010, **6**:e1000891.
17. Simpson JT, Wong K, Jackman SD, Schein JE, Jones SJM, Birol A: **ABySS: A parallel assembler for short read sequence data.** *Genome Res* 2009, **19**:1117-1123.
18. Altschul S, Madden T, Schaffer A, Zhang J, Zhang Z, Miller W, Lipman D: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
19. Del Sorbo G, Schoonbeek H-j, De Waard MA: **Fungal transporters involved in efflux of natural toxic compounds and fungicides.** *Fungal Genet Biol* 2000, **30**:1-15.
20. Deng J, Carbone I, Dean R: **The evolutionary history of Cytochrome P450 genes in four filamentous Ascomycetes.** *BMC Evol Biol* 2007, **7**:30.
21. Maloney AP, VanEtten HD: **A gene from the fungal plant pathogen *Nectria haematococca* that encodes the phytoalexin-detoxifying enzyme pisatin demethylase defines a new cytochrome P450 family.** *Mol Gen Genet* 1994, **243**:506-514.
22. Idnurm A, Howlett BJ: **Pathogenicity genes of phytopathogenic fungi.** *Mol Plant Pathol* 2001, **2**:241-255.
23. Hof C, Eisfeld K, Welzel K, Antelo L, Foster AJ, Anke H: **Ferricrocin synthesis in *Magnaporthe grisea* and its role in pathogenicity in rice.** *Mol Plant Pathol* 2007, **8**:163-172.
24. Lambou K, Tharreau D, Kohler A, Sirven C, Marguerettaz M, Barbisan C, Sexton A, Kellner E, Martin F, Howlett B, Orbach M, Lebrun M-H: **Fungi have three tetraspanin families with distinct functions.** *BMC Genomics* 2008, **9**:63.
25. Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS: **OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups.** *Nucleic Acids Res* 2006, **34**:D363-368.
26. Walton JD: **Host-selective toxins: agents of compatibility.** *Plant Cell* 1996, **8**:1723-1733.
27. Walton JD: **HC-toxin.** *Phytochemistry* 2006, **67**:1406-1413.
28. Johnson RD, Johnson L, Itoh Y, Kodama M, Otani H, Kohmoto K: **Cloning and characterization of a cyclic peptide synthetase gene from *Alternaria alternata* apple pathotype whose product is involved in AM-toxin synthesis and pathogenicity.** *Mol Plant Microbe Interact* 2000, **13**:742-753.
29. Silakowski B, Kunze B, Müller R: **Multiple hybrid polyketide synthase/non-ribosomal peptide synthetase gene clusters in the myxobacterium *Stigmatella aurantiaca*.** *Gene* 2001, **275**:233-240.
30. Daub ME, Ehrenshaft M: **The photoactivated *Cercospora* toxin cercosporin: contributions to plant disease and fundamental biology.** *Annu Rev Phytopathol* 2000, **38**:461-490.
31. Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, Thon M, Kulkarni R, Xu J-R, Pan H, Read ND, Lee Y-H, Carbone I, Brown D, Oh YY, Donofrio N, Jeong JS, Soanes DM, Djonovic S, Kolomiets E, Rehmeyer C, Li W, Harding M, Kim S, Lebrun M-H, Bohnert H, Coughlan S, Butler J, Calvo S, Ma L-J, *et al*: **The genome sequence of the rice blast fungus *Magnaporthe grisea*.** *Nature* 2005, **434**:980-986.
32. Hane JK, Lowe RGT, Solomon PS, Tan K-C, Schoch CL, Spatafora JW, Crous PW, Kodira C, Birren BW, Galagan JE, Torriani SFF, McDonald BA, Oliver RP: **Dothideomycete plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*.** *Plant Cell* 2007, **19**:3347-3368.
33. Horton P, Park K-J, Obayashi T, Nakai K: **Protein subcellular localization prediction with WoLF PSORT.** In *Proceedings of the 4th Annual Asia-Pacific Bioinformatics Conference APBC06: 13-16 Feb 2006; Taipei, Taiwan* Edited by: Jiang T, Yang U-C, Chen Y-PP, Wong L 2006, 39-48.
34. Emanuelsson O, Brunak S, von Heijne G, Nielsen H: **Locating proteins in the cell using TargetP, SignalP and related tools.** *Nat Protoc* 2007, **2**:953-971.
35. Conesa A, Gotz S, Garcia-Gomez JM, Terol J, Talon M, Robles M: **Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research.** *Bioinformatics* 2005, **21**:3674-3676.
36. Gotz S, Garcia-Gomez JM, Terol J, Williams TD, Nagaraj SH, Nueda MJ, Robles M, Talon M, Dopazo J, Conesa A: **High-throughput functional annotation and data mining with the Blast2GO suite.** *Nucleic Acids Res* 2008, **36**:3420-3435.
37. Covert S, Enkerli J, Miao V, VanEtten H: **A gene for maackiain detoxification from a dispensable chromosome of *Nectria haematococca*.** *Mol Gen Genet* 1996, **251**:397-406.
38. Keiper FJ, Grcic M, Capio E, Wallwork H: **Diagnostic microsatellite markers for the barley net blotch pathogens, *Pyrenophora teres* f. *maculata* and *Pyrenophora teres* f. *teres*.** *Aust Plant Pathol* 2008, **37**:428-430.
39. Weiland JJ, Steffenson BJ, Cartwright RD, Webster RK: **Identification of molecular genetic markers in *Pyrenophora teres* f. *teres* associated with low virulence on 'Harbin' barley.** *Phytopathology* 1999, **89**:176-181.
40. Lander ES, Green P, Abrahamson J, Barlow A, Daly MJ, Lincoln SE, Newburg L: **MAPMAKER: An interactive computer package for constructing primary genetic linkage maps of experimental and natural populations.** *Genomics* 1987, **1**:174-181.
41. Bringans S, Hane J, Casey T, Tan K-C, Lipscombe R, Solomon P, Oliver R: **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**:301.
42. Eusebio-Cope A, Suzuki N, Sadeghi-Garmaroodi H, Taga M: **Cytological and electrophoretic karyotyping of the chestnut blight fungus *Cryphonectria parasitica*.** *Fungal Genet Biol* 2009, **46**:342-351.
43. Talbot NJ, Salch YP, Ma M, Hamer JE: **Karyotypic variation within clonal lineages of the rice blast fungus, *Magnaporthe grisea*.** *Appl Environ Microbiol* 1993, **59**:585-593.
44. Taga M, Murata M, VanEtten HD: **Visualization of a conditionally dispensable chromosome in the filamentous Ascomycete *Nectria haematococca* by fluorescence *in situ* hybridization.** *Fungal Genet Biol* 1999, **26**:169-177.

45. Aragona M, Montigiani M, Porta-Puglia A: **Electrophoretic karyotypes of the phytopathogenic *Pyrenophora graminea* and *P. teres*.** *Mycol Res* 2000, **104**:853-857.
46. Mehrabi R, Taga M, Kema GHJ: **Electrophoretic and cytological karyotyping of the foliar wheat pathogen *Mycosphaerella graminicola* reveals many chromosomes with a large size range.** *Mycologia* 2007, **99**:868-876.
47. Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H, Faris JD, Rasmussen JB, Solomon PS, McDonald BA, Oliver RP: **Emergence of a new disease as a result of interspecific virulence gene transfer.** *Nat Genet* 2006, **38**:953-956.
48. Miao V, Covert S, VanEtten H: **A fungal gene for antibiotic resistance on a dispensable ("B") chromosome.** *Science* 1991, **254**:1773-1776.
49. Harimoto Y, Hatta R, Kodama M, Yamamoto M, Otani H, Tsuge T: **Expression profiles of genes encoded by the supernumerary chromosome controlling AM-toxin biosynthesis and pathogenicity in the apple pathotype of *Alternaria alternata*.** *Mol Plant Microbe Interact* 2007, **20**:1463-1476.
50. Ma L-J, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-J, Di Pietro A, Dufresne M, Freitag M, Grabherr M, Henrissat B, Houterman PM, Kang S, Shim W-B, Woloshuk C, Xie X, Xu J-R, Antoniw J, Baker SE, Bluhm BH, Breakspear A, Brown DW, Butchko RAE, Chapman S, Coulson R, Coutinho PM, Danchin EGJ, Diener A, Gale LR, Gardiner DM, Goff S, *et al*: **Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*.** *Nature* 2010, **464**:367-373.
51. Friesen TL, Meinhardt SW, Faris JD: **The *Stagonospora nodorum*-wheat pathosystem involves multiple proteinaceous host-selective toxins and corresponding host sensitivity genes that interact in an inverse gene-for-gene manner.** *Plant J* 2007, **51**:681-692.
52. Ciuffetti LM, Manning VA, Pandelova I, Betts MF, Martinez JP: **Host-selective toxins, Ptr ToxA and Ptr ToxB, as necrotrophic effectors in the *Pyrenophora tritici-repentis*-wheat interaction.** *New Phytologist* 2010, **187**:911-919.
53. Stergiopoulos I, de Wit PJGM: **Fungal effector proteins.** *Annu Rev Phytopathol* 2009, **47**:233-263.
54. Tuori RP, Wolpert TJ, Ciuffetti LM: **Purification and immunological characterization of toxic components from cultures of *Pyrenophora tritici-repentis*.** *Mol Plant Microbe Interact* 1995, **8**:41-48.
55. Martinez JP, Ottum SA, Ali S, Francl LJ, Ciuffetti LM: **Characterization of the *ToxB* Gene from *Pyrenophora tritici-repentis*.** *Mol Plant Microbe Interact* 2001, **14**:675-677.
56. Liu Z, Faris JD, Oliver RP, Tan K-C, Solomon PS, McDonald MC, McDonald BA, Nunez A, Lu S, Rasmussen JB, Friesen TL: **SnTox3 acts in effector triggered susceptibility to induce disease on wheat carrying the *Snn3* gene.** *PLoS Pathog* 2009, **5**:e1000581.
57. Friesen T, Chu CG, Liu Z, Xu S, Halley S, Faris J: **Host-selective toxins produced by *Stagonospora nodorum* confer disease susceptibility in adult wheat plants under field conditions.** *Theor Appl Genet* 2009, **118**:1489-1497.
58. Kema GHJ, Goodwin SB, Hamza S, Verstappen ECP, Cavaletto JR, Van der Lee TAJ, de Weerdt M, Bonants PJM, Waalwijk C: **A combined amplified fragment length polymorphism and randomly amplified polymorphism DNA genetic linkage map of *Mycosphaerella graminicola*, the *Septoria tritici* leaf blotch pathogen of wheat.** *Genetics* 2002, **161**:1497-1505.
59. Zhong S, Steffenson BJ, Martinez JP, Ciuffetti LM: **A molecular genetic map and electrophoretic karyotype of the plant pathogenic fungus *Cochliobolus sativus*.** *Mol Plant Microbe Interact* 2002, **15**:481-492.
60. Nitta N, Farman ML, Leong SA: **Genome organization of *Magnaporthe grisea*: integration of genetic maps, clustering of transposable elements and identification of genome duplications and rearrangements.** *Theor Appl Genet* 1997, **95**:20-32.
61. Pedersen C, Rasmussen SW, Giese H: **A genetic map of *Blumeria graminis* based on functional genes, avirulence genes, and molecular markers.** *Fungal Genet Biol* 2002, **35**:235-246.
62. Steffenson BJ, Webster RK: **Pathotype diversity of *Pyrenophora teres f. teres* on barley.** *Phytopathology* 1992, **82**:170-177.
63. Lai Z, Faris JD, Weiland JJ, Steffenson BJ, Friesen TL: **Genetic mapping of *Pyrenophora teres* f. *teres* genes conferring avirulence on barley.** *Fungal Genet Biol* 2007, **44**:323-329.
64. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL: **Accurate whole human genome sequencing using reversible terminator chemistry.** *Nature* 2008, **456**:53-59.
65. **iVEC.** [http://www.ivec.org/].
66. R Development Core Team: *R: a Language and Environment for Statistical Computing* Vienna: R Foundation for Statistical Computing; 2010.
67. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M: **Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training.** *Genome Res* 2008, **18**:1979-1990.
68. Eddy S: *HMMER User's Guide. Biological Sequence Analysis Using Profile Hidden Markov Models* , 2.3.2 2003.
69. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, Horsman D, Jones SJ, Marra MA: **Circos: An information aesthetic for comparative genomics.** *Genome Res* 2009, **19**:1639-1645.
70. Li H, Durbin R: **Fast and accurate short read alignment with Burrows-Wheeler transform.** *Bioinformatics* 2009, **25**:1754-1760.
71. Wickham H: *ggplot2: Elegant Graphics for Data Analysis.* 2 edition. New York: Springer; 2009.
72. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** *Nucleic Acids Res* 1999, **27**:573-580.
73. Rozen S, Skaletsky H: **Primer3 on the WWW for general users and for biologist programmers.** *Methods Mol Biol* 2000, **132**:365-386.
74. Nourollahi K, Javannikkhah M, Naghavi MR, Lichtenzveig J, Okhovat SM, Oliver RP, Ellwood SR: **Genetic diversity and population structure of *Ascochyta rabiei* from the western Iranian Ilam and Kermanshah provinces using MAT and SSR markers.** *Mycol Prog* 2010.
75. Hayden M, Nguyen T, Waterman A, Chalmers K: **Multiplex-ready PCR: A new method for multiplexed SSR and SNP genotyping.** *BMC Genomics* 2008, **9**:80.
76. Quinlan AR, Hall IM: **BEDTools: a flexible suite of utilities for comparing genomic features.** *Bioinformatics* 2010, **26**:841-842.
77. Vos P, Hogers R, Bleeker M, Reijans M, Lee Tvd, Hornes M, Friters A, Pot J, Paleman J, Kuiper M, Zabeau M: **AFLP: a new technique for DNA fingerprinting.** *Nucleic Acids Res* 1995, **23**:4407-4414.
78. Liu ZH, Anderson JA, Hu J, Friesen TL, Rasmussen JB, Faris JD: **A wheat intervarietal genetic linkage map based on microsatellite and target region amplified polymorphism markers and its utility for detecting quantitative trait loci.** *Theor Appl Genet* 2005, **111**:782-794.
79. Zhong S, Leng Y, Friesen TL, Faris JD, Szabo LJ: **Development and characterization of expressed sequence tag-derived microsatellite markers for the wheat stem rust fungus *Puccinia graminis* f. sp. *tritici*.** *Phytopathology* 2009, **99**:282-289.
80. Arie T, Christiansen SK, Yoder OC, Turgeon BG: **Efficient cloning of *Ascomycete* mating type genes by PCR amplification of the conserved *MAT*HMG box.** *Fungal Genet Biol* 1997, **21**:118-130.
81. Nelson JC: **QGENE: software for marker-based genomic analysis and breeding.** *Mol Breed* 1997, **3**:239-245.
82. Voorrips RE: **MapChart: software for the graphical presentation of linkage maps and QTLs.** *J Hered* 2002, **93**:77-78.

doi:10.1186/gb-2010-11-11-r109
Cite this article as: Ellwood *et al.*: A first genome assembly of the barley fungal pathogen *Pyrenophora teres* f. *teres*. *Genome Biology* 2010 **11**:R109.

Chapter 4 | Evolution of three Pyrenophora cereal pathogens: Recent divergence, speciation and evolution of non-coding DNA

Attribution Statement

Authors: Simon R. Ellwood, **Robert A. Syme**, Caroline S. Moffat, Richard P. Oliver

Citation: Evolution of three Pyrenophora cereal pathogens: recent divergence, speciation and evolution of non-coding DNA. SR Ellwood, RA Syme, CS Moffat, RP Oliver - Fungal Genetics and Biology, 2012

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

The Ph.D. candidate (Robert A. Syme) made the following contributions to this chapter:

- Genome assembly and annotation with SRE
- Authorship of code for sequence analysis required for calculation of divergence times

Other contributions:

- SRE analysed the data, generated figures and tables, and wrote the manuscript.
- RPO and TLF contributed to the design of the project and provided assistance in finalizing the manuscript prior to publication.

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

Signature: ……………………………………………

Date: 2015-05-19

I, Richard Oliver, certify that this attribution statement is an accurate record of Robert Syme's contribution to the research presented in this chapter.

Signature: ……………………………………………

Date: 2015-05-20

Fungal Genetics and Biology 49 (2012) 825–829

Contents lists available at SciVerse ScienceDirect

Fungal Genetics and Biology

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

Evolution of three *Pyrenophora* cereal pathogens: Recent divergence, speciation and evolution of non-coding DNA

Simon R. Ellwood *, Rob A. Syme, Caroline S. Moffat, Richard P. Oliver

Department of Environment and Agriculture, Curtin University, Kent Street, Bentley, Perth, Western Australia 6102, Australia

ARTICLE INFO

Article history:
Received 20 April 2012
Accepted 19 July 2012
Available online 28 July 2012

Keywords:
Necrotroph
Coevolution
Neutral selection
Nucleotide transitions and transversions

ABSTRACT

Three of the most important fungal pathogens of cereals are *Pyrenophora tritici-repentis*, the cause of tan spot on wheat, and *Pyrenophora teres* f. *teres* and *Pyrenophora teres* f. *maculata*, the cause of spot form and net form of net blotch on barley, respectively. Orthologous intergenic regions were used to examine the genetic relationships and divergence times between these pathogens. Mean divergence times were calculated at 519 kya (±30) between *P. teres* f. *teres* and *P. teres* f. *maculata*, while *P. tritici-repentis* diverged from both *Pyrenophora teres* forms 8.04 Mya (±138 ky). Individual intergenic regions showed a consistent pattern of co-divergence of the *P. teres* forms from *P. tritici-repentis*, with the pattern supported by phylogenetic analysis of conserved genes. Differences in calculated divergence times between individual intergenic regions suggested that they are not entirely under neutral selection, a phenomenon shared with higher Eukaryotes. *P. tritici-repentis* regions varied in divergence time approximately 5–12 Mya from the *P. teres* lineage, compared to the separation of wheat and barley some 12 Mya, while the *P. teres* f. *teres* and *P. teres* f. *maculata* intergenic region divergences correspond to the middle Pleistocene. The data suggest there is no correlation between the divergence of these pathogens the domestication of wheat and barley, and show *P. teres* f. *teres* and *P. teres* f. *maculata* are closely related but autonomous. The results are discussed in the context of speciation and the evolution of intergenic regions.

1. Introduction

Pyrenophora species are ascomycetes within the class Dothideomycetes. Several are important pathogens of cereals, notably *Pyrenophora tritici-repentis* on wheat and *Pyrenophora teres* f. *teres* and *Pyrenophora teres* f. *maculata* on barley. *P. tritici-repentis* (PTR) is the cause of tan spot, while *Pyrenophora teres* causes net blotch. These species are necrotrophs, causing cell death and feeding off the nutrients released, and have become recognised diseases only over the last century. The two barley forms closely resemble each other morphologically and are distinguished by their disease symptoms. *P. teres* f. *teres* (PTT) produces net form of net blotch (NFNB), while *P. teres* f. *maculata* (PTM) produces spot form of net blotch (SFNB). NFNB is typified by elongated lesions, where necrosis develops along leaf veins with occasional transverse striations. SFNB displays more ovoid lesions, often surrounded by a chlorotic zone.

The relationship of the two forms of *P. teres* to each other has remained unclear despite several DNA-based studies. These have suggested that PTT and PTM are closely related but autonomous and divergent genetic groups (Bakonyi and Justesen, 2007; Bogacki et al., 2010; Lehmensiek et al., 2010; Rau et al., 2007, 2003). However, PTT and PTM can be artificially hybridised to create progeny that are morphologically intermediate and are genetically stable (Campbell and Crous, 2003). Two studies have also suggested infrequent hybridization may occur naturally (Campbell et al., 2002; Leišova et al., 2005). The uncertainty in these relationships has been due to the small amount of genomic DNA sequence information available and limitations in the marker types deployed. Long DNA sequences are often necessary to obtain a robust phylogenetic tree while some marker techniques, notably RAPDS, are anonymous in their DNA sequence content and context and are unreliable for inferring hybridization events unless isolated and sequenced (e.g. Van De Zande and Bijlsma, 1995).

Wheat and barley are both believed to have been domesticated in the Fertile Crescent around 10,000 years ago (Badr et al., 2000; Heun et al., 1997; Özkan et al., 2002). Since then, widespread intensive cultivation and the development of genetically homogeneous crops has led to strong selective pressure on natural fungal populations or the emergence of entirely new pathogens (Brunner et al., 2007; Friesen et al., 2006; Ma et al., 2010; Stukenbrock et al., 2007). The history of the *Pyrenophora* diseases in this study and their relationship to their hosts is poorly understood, with the exception of PTR, which only recently became a serious pathogen on wheat through the horizontal transfer of the effector ToxA from *Stagonospora nodorum* (Friesen et al., 2006). Prior to this PTR was recorded as occasional and insignificant pathogen on wheat and

* Corresponding author. Fax: +61 8 9266 2021.
E-mail address: srellwood@gmail.com (S.R. Ellwood).

1087-1845 Crown Copyright © 2012 Published by Elsevier Inc. Open access under CC BY-NC-ND license.
http://dx.doi.org/10.1016/j.fgb.2012.07.003

other grasses (Nisikado, 1929). An obvious lack of fossil records for fungal pathogens means indirect dating techniques are needed to set molecular clocks and species diversification dates. The most reliable DNA sequences to base species divergence on are neutral regions where the nucleotide substitution rates are constant, as gene coding sequences, regulatory elements and introns may both be subject to selective constraints (Hare and Palumbi, 2003; Nielsen et al., 2004). In fungi, Oberhaensli et al. (2011) used intergenic regions and transposable elements to show a convergence between the evolution of powdery mildew and host speciation, while Martin et al. (2010) used inactive transposable elements to date the insertion age of truffle LTRs.

The advent of next generation sequencing technologies has made genome sequencing of fungal species rapid and affordable (Haridas et al., 2011; Nowrousian et al., 2010). Their relatively compact and simple genomes compared to most eukaryotes are suited to assembling the gene content using short sequencing reads. In this study we sequenced a representative of PTR and PTM, and, in combination with a pre-existing assembly for PTT (Ellwood et al., 2010), we examined the genetic relationships, divergence, and host co-existence between these pathogens.

2. Material and methods

2.1. Genome sequencing

PTR isolate WAC11137 was acquired from the Department of Agriculture and Food, Western Australia (3, Baron Hay Court, South Perth, Western Australia 6151). PTM isolate SG1-1 was collected in Western Australia by S. Ellwood during 2009. One hundred base pair sequencing was performed on a Solexa GAII sequencing platform (Bentley et al., 2008), using 300 bp paired-end libraries with a minimum 40 times genome coverage. Preparation of randomly sheared DNA, cluster formation, primer hybridization and DNA amplification reactions were according to the manufacturer's recommended protocol. Data was pre-filtered by Illumina's data pipeline. Removal of any contaminating DNA adaptors, tags, and trimming of bases with an Illumina quality score of less than 30 was performed using the FASTX-toolkit v 0.0.13.

PTR WAC11137 and PTM SG1-1 were assembled with Velvet v 1.1.02 (Zerbino and Birney, 2008). The optimal kmer (sequence overlap) length to construct the assemblies was determined by incrementally adjusting the kmer by 4 bp. The optimal kmer length was selected where N_{50} was minimal and L_{50} was maximal. L_{50} defined here is the length of the smallest N_{50} contig, where N_{50} is the minimum number of contigs required to represent 50% of the genome. Protein coding sequences were identified with GeneMark-ES v.2 (Ter-Hovhannisyan et al., 2008). GeneMark-ES uses a self-training algorithm optimized for features of fungal gene organization and incorporates an enhanced intron submodel to accommodate sequences with and without branch point sites. The Geneious (Drummond et al., 2011) plugin Phobos (Mayer, 2010) was used to identify short tandem repeats or microsatellites. Larger interspersed repeats and transposon-like elements were predicted *de-novo* using RepeatScout v 1.0.5 (Price et al., 2005). These were combined with characterized transposable elements from a closely related Dothideomycete, *S. nodorum* (Hane and Oliver, 2008), and their locations together with low complexity regions delineated with RepeatMasker 3.3.0 (Smit et al., 1996–2011).

2.2. Calculation of divergence time

PTR, PTM, and PTT (Ellwood et al., 2010) genomic scaffolds with large intergenic regions were compared by BLASTN (Altschul et al., 1997). Matching scaffolds delimited by orthologous genes were initially aligned in the Mauve genome viewer within Geneious, using the progressive alignment option. Intergenic regions were selected that were at least 1.5 kbp from the nearest predicted flanking gene, which resided on separate scaffolds, and which lacked interspersed repeats. As a second precaution to ensure regions were not coding, they were queried against NCBI. Each region was aligned using MUSCLE (Edgar, 2004). The number of transitions and transversions between each pathogen were counted with a BioRuby script (Goto et al., 2010). Indels were removed from the alignment length and hence the total number of informative sites. Evolutionary distance per site (K), divergence time, and standard errors were calculated using the Kimura 2-parameter model and the equations detailed in Kimura (1980), using an average substitution rate for fungi of 8.8×10^{-9} per site per year (Kasuga et al., 2002). FASTA alignments of intergenic regions used to calculate divergence times are presented in Supplementary file 1.

2.3. Multigene phylogeny

Phylogenetic relationships between PTR, PTM, PTT and *Cochliobolus heterostrophus* were inferred from five concatenated orthologous genes; actin (*act1*), β-tubulin (*tub2*), cytochrome P450 14α-demethylase (*cyp51A*), translation elongation factor-1α (*EF-1α*) and glyceraldehyde-3-phosphate dehydrogenase (*G3PD*). These regions were selected based on the number of informative sites among genes commonly used to infer phylogeny and correctly predicted between the different assemblies. *C. heterostrophus* isolate C5 v 2 sequences were obtained from the DOE Joint Genome Institute (http://genome.jgi-psf.org/programs/fungi/index.jsf, Barbara Turgeon, pers. comm.). Concatenated sequences were aligned with MUSCLE and analysed by Metropolis coupled Markov chain Monte Carlo analyses with Mr Bayes (Huelsenbeck and Ronquist, 2001) within Geneious with *C. heterostrophus* selected as the outgroup. Unconstrained branch lengths were used with default parameters (gamma distribution approximated with four categories and a proportion of invariable sites, HKY substitution model, subsampling every 200th generation and a burn in length of 1,100,000). Branch bootstrap support values were obtained from a Neighbour Joining tree with 10,000 bootstrap replicates. MEGA v 5 (Tamura et al., 2011) was used to portray trees. New gene sequences used in this study have been deposited in GenBank under accessions JQ314397-JQ314406. Aligned concatenated sequences are provided in Supplementary file 2.

3. Results and discussion

Neutral substitution rates vary across gene loci, in part due to low numbers of scorable sites per gene, a particular issue in closely related taxa, and to differences in diversifying and purifying selection. We therefore selected intergenic regions on which to base divergence times between PTR, PTM, and PTT. As no fossil evidence for these species exists to calibrate a molecular clock, a mean substitution rate based on the third codon of protein coding genes in fungi was used, based on data from Kasuga et al. (2002), who showed the range of neutral substitution mutations were surprisingly similar across kingdoms.

A similar approach has previously been used to date divergence in truffles and powdery mildew (Martin et al., 2010; Oberhaensli et al., 2011) using transposable elements. However, in many fungi, transposable elements are subject to repeat-induced point (RIP) mutation, a fungal-specific genome defence mechanism that introduces G/C-to-A/T transition mutations into repeat copies during the sexual stage (Cambareri et al., 1991). In the Dothideomycetes, the process has been identified in *Stagonospora* and *Leptosphaeria*, genera closely related to *Pyrenophora* (Hane et al., 2007; Rouxel et al., 2011), and more recently within *Pyrenophora* itself (James Hane, pers. comm.). In this study, therefore, aligned intergenic regions were identified on separate genomic scaffolds devoid of transposable elements. Transposable elements would not be expected to be prevalent in small DNA fragment short-read genome assemblies, as such repetitive DNA tends not to assemble (Ellwood et al., 2010). To ensure their absence, intergenic regions with interspersed repeats and transposable element-like sequences were identified and excluded from the analyses.

In total 10 randomly selected intergenic regions were identified that fitted the sampling criteria. The total alignment length was 35.5 kbp and the number of aligned sites per region ranged from 2.3 to 4.3 kbp (Table 1). The intergenic regions exhibited a range of divergence dates and the mean divergence dates for the three fungi were calculated as just over 519 kya (±30) between PTM and PTT, and between PTR and both PTM and PTT an average of 8.04 Mya (±138 ky). The average ratio of transitions to transversions between PTM and PTT was 2:1 and between PTR compared to PTM and PTT was 1.53. These ratios are consistent with transitions being generated at a higher frequency than transversions and a bias or movement towards saturation of transitions at higher levels of genetic divergence (Yang and Yoder, 1999).

When divergence times in individual intergenic regions are examined, a consistent pattern of co-divergence of both forms of *P. teres* from PTR is evident but with varying dates between the regions. To validate these results, the phylogeny of five concatenated orthologous genes regions was compared using *C. heterostrophus* as an out group. The total alignment length was 8351 bp with 21 nucleotide changes between PTM and PTT. Tree topology was compared using Mr Bayes inference and Neighbour Joining. Each method produced trees entirely congruent with the divergence time model, with a relatively long branch separating PTR from the closely related *P. teres* forms (Fig. 1).

In PTR, there are at least four groups of intergenic regions showing divergence at approximately 5, 7, 10 and 11 Mya from *P. teres*, while PTM and PTT show divergence dates from approximately 400 kya to 600 kya (Fig. 2). The author proposes two theories to account for these different divergence times. The first is that despite the precautions described in materials and methods in selecting the intergenic regions, a proportion of these in *Pyrenophora* are

Table 1
Divergence time estimates of PTM, PTT and PTR.

Locus	Comparison	Aligned sites	Transitions	Transversions	Divergence time	Evolutionary distance per site (K)	Standard error (σK)
Locus 1							
	PTM/PTT	3847	20	10	445,697	0.008	0.001
	PTT/PTR	3427	222	104	5,839,474	0.103	0.006
	PTM/PTR	3427	227	103	5,919,599	0.104	0.006
Locus 2							
	PTM/PTT	3384	16	10	439,002	0.008	0.002
	PTT/PTR	2933	202	148	7,434,579	0.131	0.007
	PTM/PTR	2930	204	148	7,490,782	0.132	0.007
Locus 3							
	PTM/PTT	2835	28	7	708,753	0.012	0.002
	PTT/PTR	2406	263	137	10,875,296	0.191	0.010
	PTM/PTR	2428	264	136	10,763,935	0.189	0.010
Locus 4							
	PTM/PTT	3329	28	10	654,486	0.012	0.002
	PTT/PTR	3081	172	98	5,332,921	0.094	0.006
	PTM/PTR	3083	177	92	5,313,349	0.094	0.006
Locus 5							
	PTM/PTT	4357	29	18	617,772	0.011	0.002
	PTT/PTR	3580	316	253	10,231,709	0.180	0.008
	PTM/PTR	3592	315	258	10,270,356	0.181	0.008
Locus 6							
	PTM/PTT	3146	16	6	399,523	0.007	0.002
	PTT/PTR	2551	165	115	6,788,031	0.119	0.007
	PTM/PTR	2556	162	113	6,641,673	0.117	0.007
Locus 7							
	PTM/PTT	3255	12	10	385,855	0.007	0.001
	PTT/PTR	2913	194	144	7,207,790	0.127	0.007
	PTM/PTR	2906	192	140	7,087,926	0.125	0.007
Locus 8							
	PTM/PTT	2929	20	12	625,762	0.011	0.002
	PTT/PTR	2369	242	172	11,439,512	0.201	0.010
	PTM/PTR	2378	240	171	11,292,908	0.199	0.010
Locus 9							
	PTM/PTT	2932	21	8	566,371	0.010	0.002
	PTT/PTR	2414	279	158	11,988,192	0.211	0.011
	PTM/PTR	2393	283	160	12,309,594	0.217	0.011
Locus 10							
	PTM/PTT	3359	13	9	373,892	0.007	0.001
	PTT/PTR	2981	155	125	5,725,307	0.101	0.006
	PTM/PTR	2979	162	124	5,864,577	0.103	0.006
Total							
	PTM/PTT	33369	203	100	519,481	0.009	0.001
	PTT/PTR	28653	2209	1453	8,032,881	0.141	0.002
	PTM/PTR	28670	2225	1444	8,046,343	0.142	0.002

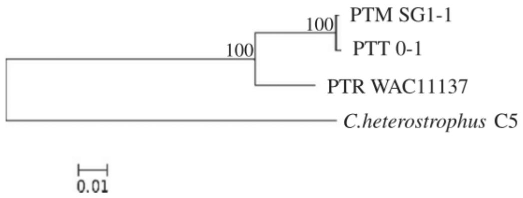


Fig. 1. Phylogram between PTM, PTT and PTR, with *C. heterostrophus* as an outgroup, based on an 8351 bp concatenated alignment of actin, β-tubulin, cyp51A, EF-1α and G3PD. The phylogram represents the consensus posterior output built by MrBayes. Neighbour Joining bootstrap support percentages are indicated at branching points. Scale bar represents substitutions per site.

functionally important and not under neutral selection. Andolfatto (2005) showed in *Drosophila* that over 40% of intergenic nucleotides were evolutionarily constrained relative to synonymous sites, but also around 20% of nucleotides showed a greater than expected between-species divergence or positive selection. The majority of intergenic DNA has no known function but evolutionarily constrained intergenic regions may be due to several factors including the location of regulatory silencing siRNA (Feng et al., 2009), large noncoding RNAs that associate with chromatin modifying complexes and affect gene expression (Khalil et al., 2009), and methylation sites thought to regulate cell context-specific alternative promoters (Maunakea et al., 2010). Positive selection may be explained by a recent study in sticklebacks, where 41% of loci associated with adaptive marine–freshwater evolution mapped entirely to non-coding regions of the genome (Jones et al., 2012). A second theory envisages hybridization between recently diverging lineages still capable of interbreeding and may include lineages predating divergence and unknown intermediary lineages. Rare single hybridization events, clonal (asexual) selection of progeny with favourable combinations of genes, and isolation restricting recombination would allow intergenic regions with different divergence dates to become fixed in a population. The mosaic of divergence times evident in PTR is a complex scenario suggesting several hybridization events and might argue against this mechanism.

This study was partly conducted to determine whether the divergence times coincided with the divergence wheat and barley, the domestication of cereals or some other identifiable historical event. Wheat and barley diverged some 12 Mya (Chalupska et al., 2008) and a lack of knowledge about host specificity before further host speciation and pathogen specialization cannot exclude a possible overlap of pathogen and host divergence. On the other hand, the divergence of PTM and PTT clearly predates the domestication of barley. Unlike in biotrophs, there is no studies to date indicating agriculturally important necrotrophs have co-evolved with their hosts, relying instead on horizontal transfer of genes conferring virulence or by strong selection pressure on local pathogen populations from wild relatives (Brunner et al., 2007; Friesen et al., 2006; Ma et al., 2010; Oliver and Solomon, 2008; Stukenbrock et al., 2007).

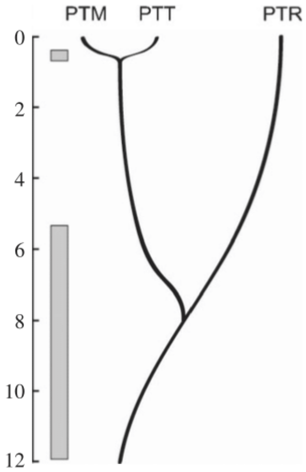


Fig. 2. Model of divergence between PTM, PTT and PTR based on mean intergenic divergence estimates. Grey bars represent the range of individual divergence time estimates for single intergenic regions for each comparison. Vertical scale bar is in million year intervals.

The divergence of PTM and PTT predating the domestication of barley confirms those studies suggesting the two forms of *P. teres* are genetically isolated (Bakonyi and Justesen, 2007; Bogacki et al., 2010; Lehmensiek et al., 2010; Rau et al., 2007, 2003) and might be considered as different species. *Pyrenophora graminea*, the cause of barley stripe, has yet to be sequenced. This species is considered as differentiated with *P. teres* as the magnitude between PTM and PTT (Rau et al., 2007). The mean and all individual intergenic divergence estimates between PTM and PTT fit within the middle Pleistocene of the Quaternary Period. This geological stage was punctuated by significant glaciation eras and long periods of isolation in glacial refuges may have provided conditions for speciation. More recently, isolation and specialization may have continued in different centres of barley diversity and domestication in Asia and the Horn of Africa (Badr et al., 2000; Morrell and Clegg, 2007; Orabi et al., 2007). Literature describing two forms of *P. teres* dates back to the second half of the last century, with Smedegård-Petersen (1971) the first to report SFNB following detection in Denmark. The absence of previous information together with new reports and increases in severity of SFNB in all barley growing regions of the world (reviewed in Ficsor et al., 2010; Liu and Friesen, 2010; McLean et al., 2009) suggest this coexistence is recent, with different centres of barley diversity and domestication combined with modern travel providing a mechanism for the present co-existence. Alternatively, or as part of this process, wild hosts may have facilitated barley colonization. PTT is reported to infect a range of gramineous species in the genera *Aegilops, Agropyron, Elymus, Hordeum, Hordelymus*, and *Stipa* (Brown et al., 1993). Comparable data for PTM has not been collected, but barley colonization by *P. teres* populations in new areas of cultivation, followed by adaption and specialization, is also plausible.

The finding that intergenic regions in *Pyrenophora* are not entirely under neutral selection requires that genome-wide estimates of substitution rates, based on the number of mutations that have occurred in a known number of generations, are needed for more accurate estimates of rates within the regions. However, until such data is available, deploying a constant genome-wide substitution rate allows broad comparisons to be made, with the age of the common ancestor of PTR and *P. teres* being over fifteen times greater than that of the common ancestor PTM and PTT. The role of *Pyrenophora* intergenic regions showing different divergence times can be resolved with more complete genome assemblies and by combining comparative genomic analyses with genetically diverse isolates and population-level data to unravel adaptive differences between populations and species.

Acknowledgments

This research was funded by the Australian Grains Research and Development Corporation under project CUR00012. The authors acknowledge the West Australian computing resources hosted by IVEC (http://www.ivec.org).

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.2012.07.003.

References

Altschul, S. et al., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research 25, 3389–3402.

Andolfatto, P., 2005. Adaptive evolution of non-coding DNA in *Drosophila*. Nature 437, 1149–1152.

Badr, A. et al., 2000. On the origin and domestication history of barley (*Hordeum vulgare*). Molecular Biology and Evolution 17, 499–510.

Bakonyi, J., Justesen, A.F., 2007. Genetic relationship of *Pyrenophora graminea*, *P. teres* f. *maculata* and *P. teres* f. *teres* assessed by RAPD analysis. Journal of Phytopathology 155, 76–83.

Bentley, D.R. et al., 2008. Accurate whole human genome sequencing using reversible terminator chemistry. Nature 456, 53–59.

Bogacki, P. et al., 2010. Genetic structure of South Australian *Pyrenophora teres* populations as revealed by microsatellite analyses. Fungal Biology 114, 834–841.

Brown, M.P. et al., 1993. Host range of *Pyrenophora teres* f. *teres* isolates from California. Plant Disease 77, 942–947.

Brunner, P.C. et al., 2007. The origin and colonization history of the barley scald pathogen *Rhynchosporium secalis*. Journal of Evolutionary Biology 20, 1311–1321.

Cambareri, E.B. et al., 1991. Recurrence of repeat-induced point mutation (RIP) in *Neurospora crassa*. Genetics 127, 699–710.

Campbell, G.F., Crous, P.W., 2003. Genetic stability of net x spot hybrid progeny of the barley pathogen *Pyrenophora teres*. Australasian Plant Pathology 32, 283–287.

Campbell, G.F. et al., 2002. Evidence of recombination between net- and spot-type populations of *Pyrenophora teres* as determined by RAPD analysis. Mycological Research 106, 602–608.

Chalupska, D. et al., 2008. Acc homoeoloci and the evolution of wheat genomes. Proceedings of the National Academy of Sciences 105, 9691–9696.

Drummond, A.J., et al., 2011. Geneious v 5.4. <http://www.geneious.com/>.

Edgar, R., 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5, 113.

Ellwood, S. et al., 2010. A first genome assembly of the barley fungal pathogen *Pyrenophora teres* f. *teres*. Genome Biology 11, R109.

Feng, J. et al., 2009. Coding DNA repeated throughout intergenic regions of the *Arabidopsis thaliana* genome: evolutionary footprints of RNA silencing. Molecular Biosystems 5, 1679–1687.

Ficsor, A. et al., 2010. First report of spot form of net blotch of barley caused by *Pyrenophora teres* f. *maculata* in Hungary. Plant Disease 94, pp. 1062–1062.

Friesen, T.L. et al., 2006. Emergence of a new disease as a result of interspecific virulence gene transfer. Natural Genetic 38, 953–956.

Goto, N. et al., 2010. BioRuby: bioinformatics software for the ruby programming language. Bioinformatics, 1–3.

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

Hane, J.K. et al., 2007. Dothideomycete plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*. Plant Cell 19, 3347–3368.

Hare, M.P., Palumbi, S.R., 2003. High intron sequence conservation across three mammalian orders suggests functional constraints. Molecular Biology and Evolution 20, 969–978.

Haridas, S. et al., 2011. A biologist's guide to de novo genome assembly using next-generation sequence data: a test with fungal genomes. Journal of Microbiological Methods 86, 368–375.

Heun, M. et al., 1997. Site of Einkorn wheat domestication identified by DNA fingerprinting. Science 278, 1312–1314.

Huelsenbeck, J.P., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17, 754–755.

Jones, F.C. et al., 2012. The genomic basis of adaptive evolution in threespine sticklebacks. Nature 484, 55–61.

Kasuga, T. et al., 2002. Estimation of nucleotide substitution rates in Eurotiomycete fungi. Molecular Biology and Evolution 19, 2318–2324.

Khalil, A.M. et al., 2009. Many human large intergenic noncoding RNAs associate with chromatin-modifying complexes and affect gene expression. Proceedings of the National Academy of Sciences 106, 11667–11672.

Kimura, M., 1980. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide substitutions. Journal of Molecular Evolution 23, 111–120.

Lehmensiek, A. et al., 2010. Population structure of South African and Australian *Pyrenophora teres* isolates. Plant Pathology 59, 504–515.

Leišova, L. et al., 2005. Genetic diversity of *Pyrenophora teres* isolates as detected by AFLP analysis. Journal of Phytopathology 153, 569–578.

Liu, Z.H., Friesen, T.L., 2010. Identification of *Pyrenophora teres* f. *maculata*, causal agent of spot type net blotch of barley in North Dakota. Plant Disease 94, pp. 480–480.

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

Martin, F. et al., 2010. Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. Nature 464, 1033–1038.

Maunakea, A.K. et al., 2010. Conserved role of intragenic DNA methylation in regulating alternative promoters. Nature 466, 253–257.

Mayer, C., 2010. Phobos v 3.3.11. <http://www.rub.de/spezzoo/cm/cm_phobos.htm>.

McLean, M.S. et al., 2009. Epidemiology and control of spot form of net blotch (*Pyrenophora teres* f. *maculata*) of barley: a review. Crop and Pasture Science 60, 303–315.

Morrell, P.L., Clegg, M.T., 2007. Genetic evidence for a second domestication of barley (*Hordeum vulgare*) east of the fertile crescent. Proceedings of the National Academy of Sciences 104, 3289–3294.

Nielsen, C.B. et al., 2004. Patterns of intron gain and loss in fungi. PLoS Biology 2, e422.

Nisikado, Y., 1929. Studies on the *Helminthosporium* diseases of Gramineae in Japan. Ber Ohara Inst Landw Forsch 4, 111–126.

Nowrousian, M. et al., 2010. *De novo* assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. PLoS Genetics 6, e1000891.

Oberhaensli, S. et al., 2011. Comparative sequence analysis of wheat and barley powdery mildew fungi reveals gene colinearity, dates divergence and indicates host-pathogen co-evolution. Fungal Genetics and Biology 48, 327–334.

Oliver, R.P., Solomon, P.S., 2008. Recent fungal diseases of crop plants: is lateral gene transfer a common theme? Molecular Plant-Microbe Interactions 21, 287–293.

Orabi, J. et al., 2007. The horn of Africa as a centre of barley diversification and a potential domestication site. Theoretical and Applied Genetics 114, 1117–1127.

Özkan, H. et al., 2002. AFLP analysis of a collection of tetraploid wheats indicates the origin of Emmer and hard wheat domestication in southeast Turkey. Molecular Biology and Evolution 19, 1797–1801.

Price, A.L. et al., 2005. De novo identification of repeat families in large genomes. Bioinformatics 21, i351–358.

Rau, D. et al., 2003. Population genetic structure of *Pyrenophora teres* Drechs. the causal agent of net blotch in Sardinian landraces of barley (*Hordeum vulgare* L.). Theoretical and Applied Genetics 106, 947–959.

Rau, D. et al., 2007. Phylogeny and evolution of mating-type genes from *Pyrenophora teres*, the causal agent of barley 'net blotch' disease. Current Genetics 51, 377–392.

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

Smedegård-Petersen, V., 1971. *Pyrenophora teres* f. *maculata* f. nov. and *Pyrenophora teres* f. *teres* on barley in Denmark. Aarsskrift Kongelige Veterinear of Landbohojskole, 124–144.

Smit, A.F.A., et al., 1996–2011. RepeatMasker Open-3.0. <http://www.repeatmasker.org>.

Stukenbrock, E.H. et al., 2007. Origin and domestication of the fungal wheat pathogen *Mycosphaerella graminicola* via sympatric speciation. Molecular Biology and Evolution 24, 398–411.

Tamura, K. et al., 2011. MEGA5: molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Molecular Biology and Evolution. http://dx.doi.org/10.1093/molbev/msr121.

Ter-Hovhannisyan, V. et al., 2008. Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. Genome Research 18, 1979–1990.

Van De Zande, L., Bijlsma, R., 1995. Limitations of the RAPD technique in phylogeny reconstruction in *Drosophila*. Journal of Evolutionary Biology 8, 645–656.

Yang, Z., Yoder, A.D., 1999. Estimation of the transition/transversion rate bias and species sampling. Journal of Molecular Evolution 48, 274–283.

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

Chapter 5 | Comprehensive annotation of the *Parastagonospora nodorum* reference genome using next-generation genomics, transcriptomics and proteogenomics

Attribution Statement

Authors: **Robert A. Syme**, Kar-Chun Tan James K. Hane, Kejal Dodhia, Thomas Stoll, Eiko Furuki, Simon R. Ellwood, Angela Williams, Yew-Foon Tan, Alison Testa, Jeffrey J. Gorman, Richard P. Oliver

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

The Ph.D. candidate (Robert A. Syme) made the following contributions to this chapter:

- All data management, server build and systems administration for web applications, databases and version control
- Corrections to genome assembly
- Manual annotation of genes on 88 of 91 scaffolds
- All analyses between genomes and annotations
- Generation of all figures and tables
- Manuscript writing with editing from RPO, JKH and KCT

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

Signature: ………………………………………………

Date: 2015-05-19

I, Richard Oliver, certify that this attribution statement is an accurate record of Robert Syme's contribution to the research presented in this chapter.

Signature: ………………………………………………

Date: 2015-05-20

Background

The cost of DNA sequencing has decreased to the point where it no longer represents a significant hindrance to obtaining a genomic assembly (Chain, Grafham et al. 2009). Despite the ease with which raw reads are procured, obtaining an accurately assembled and annotated eukaryotic genome remains a significant challenge. Genome assembly can be hampered by errors arising from the sequencing chemistry which can introduce incorrect bases or by repetitive regions, which can lead to truncated contigs and a fragmented assembly. Genes and other features are typically annotated using homology-based methods or are predicted *ab initio*. Experimental gene validation techniques are required to complement *in silico* methods to obtain high quality gene model annotations.

Parastagonospora nodorum [Teleomorph: *Phaeosphaeria* (Hedjar.) syn. *Leptosphaeria nodorum (*Müll.*),* syn. *Septoria nodorum* (Berk.), syn. *Stagonospora nodorum* (Berk.)] is a filamentous Ascomycete and member of the Dothideomycetes, a taxonomic class that consists of several agriculturally-damaging phytopathogens (Murray and Brennan 2009, Crook, Friesen et al. 2012, Stergiopoulos, Collemare et al. 2013). *P. nodorum* causes the wheat disease septoria nodorum blotch (SNB syn. glume blotch) (Oliver, Tan et al. In Press) and is responsible for substantial yield losses in many regions around the world. As part of the infection process, the fungus produces an arsenal of proteinaceous effectors that induce tissue necrosis and/or chlorosis on hosts expressing the corresponding susceptibility gene (Tan, Oliver et al. 2010). Analysis of the *P. nodorum* / *Triticum* pathosystem has revealed the necrotrophic effectors SnToxA (Friesen, Stukenbrock et al. 2006), SnTox1 (Liu, Faris et al. 2004), SnTox3 (Liu, Faris et al. 2009), SnTox4 (Abeysekara, Friesen et al. 2009) and SnTox6 (Gao, Faris et al. 2015). The presence of undiscovered effectors in *P. nodorum* is evident by observation of disease symptoms in wheat cultivars challenged with culture filtrate from the reference strain devoid of known effectors (Tan, Waters et al. 2014) and culture filtrate from other *P. nodorum* populations (Crook, Friesen et al. 2012). In addition to effectors, *P. nodorum* genes involved in primary metabolism, secondary metabolism, and signal transduction have been studied. Characterised metabolic enzymes include malate synthase (Solomon, Lee et al. 2004), δ-aminolevulinic acid synthase (Solomon, Jörgens et al. 2006), pantoate-β-alanine ligase (Ipcho, Hane et al. 2012), mannitol 2-dehydrogenase (Solomon, Waters et al. 2006), mannitol 1-phosphate dehydrogenase (Solomon, Tan et al. 2005), and trehalose 6-phosphate synthase (Lowe, Lord et al. 2009). *P. nodorum* signal transduction and regulatory loci that have been studied in depth include the transcription factor *StuA* (IpCho, Tan et al. 2010), a MAP kinase (Solomon, Waters et al. 2005), the calcium/calmodulin-dependent protein kinases Cpk1/Cpk2/Cpk3 (Solomon, Rybak et al. 2006), G-protein subunits Gα (Solomon, Tan et al. 2004), Gβ and Gγ (Gummer, Trengove et al. 2012), and putative short-chain dehydrogenases (Tan, Heazlewood et al. 2008, Casey, Solomon et al. 2010) revealed to be necessary for the formation of the mycotoxin altenariol and sporulation (Tan, Trengove et al. 2009).

The first published Dothideomycete whole genome assembly was of *P. nodorum* strain SN15. The original sequence was obtained in 2004 using 1 kb, 4 kb, and 40 kb Sanger shotgun sequenced paired-end reads assembled as 37.1 Mb of nuclear DNA in 107 scaffolds and the complete 49.8 kbp mitochondrial genome (Hane, Lowe et al. 2007). Initial gene-structure annotation relied heavily on automated methods, but was subsequently revised after analysis of proteogenomic (Bringans, Hane et al. 2009), and microarray data (Ipcho, Hane et al. 2012) (Figure 1) to give a total of 10761 gene models with a mean exon count of 2.6, mean CDS length of 1400 bp, mean intergenic distance of 1685 bp, and a mean intron length of 91 bp. In addition to the 10761 gene models with some experimental support, an additional 1621 low-confidence genes (total count of 12382) were incorporated in analysis by Syme, Hane et al. (2013) to minimise the possibility of missing potential effectors. Repetitive sequence comprised 4.52% of the genome in 5 subtelomeric repeat classes, 1 ribosomal DNA repeat and 20 transposon or transposon-like clusters (Hane, Lowe et al. 2007). Repeat-induced point (RIP) mutations in repeat instances were subsequently reversed *in-silico* to allow classification of the repeat X26 as a RecQ helicase, R25 as a pseudogene, and repeats X3 and X8 as members of the same ancestral class (Hane and Oliver 2010).

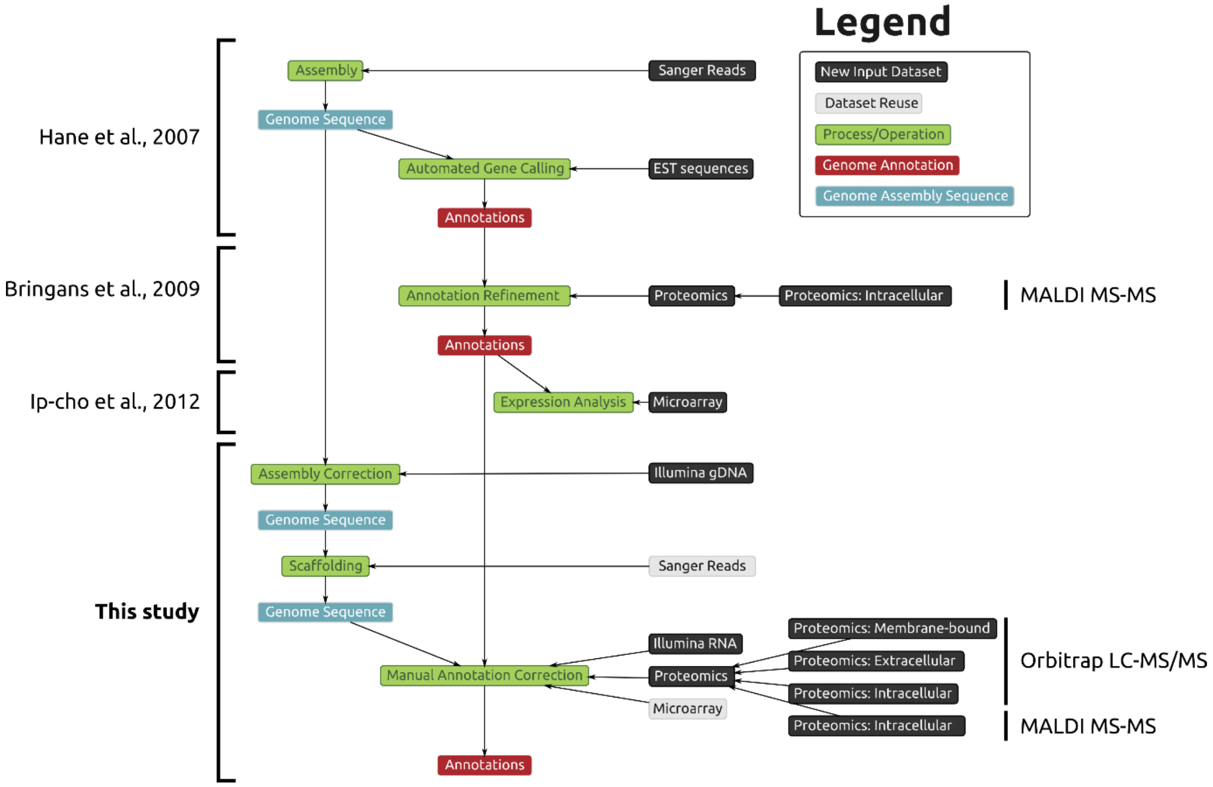


Figure 1: Overview of *P. nodorum* reference genome and annotation history.

The initial assembly was found to contain a homolog of the *Pyrenophora tritici-repentis* necrotrophic effector ToxA, providing evidence of a horizontal gene transfer event from *P. nodorum* to *P. tritici-repentis (Friesen 2006)*. The ToxA-containing transfercon was initially estimated to be 11 kbp, but it has been suggested that at least 72 kbp was transferred including *P. nodorum* sequence corresponding to scaffolds 68, 55, 51, 46, 64, and 73 (Syme, Hane et al. 2013).

The genomics resources available to *P. nodorum* researchers were expanded to include the genomes of two more strains - one isolated from the grass Agropyron, unable to infect wheat and a wheat pathogen known to produce a different suite of effectors to the SN15 reference strain. In comparing the three strains, the analysis by Syme, Hane et al. (2013) included 1621 lower confidence genes to the 10761 genes from Bringans, Hane et al. (2009) to minimise the possibility of missing potential effector loci, bringing the total number of putative genes used in that comparison to 12382. Clustering of the predicted proteomes from the three strains revealed a core set of 10464 conserved proteins and 2421 proteins exclusive to strains able to infect wheat (Syme, Hane et al. 2013).

The accuracy and completeness of a genome assembly can be improved by the addition of new sequencing data. Error characteristics and shortcomings of one sequencing technique may be overcome by a complementary chemistry (Shendure and Ji 2008, Zhang, Chiodini et al. 2011). The long read lengths available from Sanger are useful to resolve repetitive regions and provide the large-scale structural assembly whereas the depth and accuracy of Illumina short-reads delivers the ability to correct remaining SNPs and small insertions or deletions (Chevreux, Wetter et al. 1999).

Comparison of the chromosomes of filamentous ascomycetes has shown that related chromosomes tend to conserve gene content, but with shuffled gene order (Hane, Rouxel et al. 2011). The resulting syntenic patterns are described as mesosynteny and can be explained by frequent chromosomal inversions but infrequent translocations (Hane, Rouxel et al. 2011, Ohm, Feau et al. 2012).When describing the process, Hane, Rouxel et al. (2011) also suggested that mesosynteny may resolve the order and orientation of scaffolds in a fragmented genome assembly and thereby identify groups of scaffolds that comprise a single chromosome. The utility of this technique is most obvious when a finished genome can be used to improve the fragmented assembly of a closely related species or strain.

Homing endonucleases are mobile genetic elements with highly specific DNA nuclease activity. The elements are encoded inside the introns of other functional genes (the 'host gene'). If the translated endonuclease encounters a copy of the host gene without the insertion, it will introduce a double stranded cut at a specific point (Belfort and Perlman 1995, Belfort and Roberts 1997).

The coding sequence for the endonuclease can be introduced into the cut site during repair. While most homing endonucleases are purely selfish elements, some have been co-opted into performing biologically helpful functions for the host, such as a mating-type switch in yeast (Jin, Binkowski et al. 1997).

In this study, we report extensive correction of SNP and indel base-calling errors in the *P. nodorum* SN15 reference assembly, the closing of assembly gaps, extensive automated and manual gene annotation, and improvements to the functional characterisation of gene models. The new experimental data comprises RNA sequencing, DNA sequencing, and multiple sets of proteomic data which were used to inform comprehensive manual curation of gene models. Using these complementary approaches, we have generated an improved genome assembly, gene and protein datasets set and have re-predicted potential pathogenicity effector genes of *P. nodorum* with higher accuracy. These bioinformatic resources represent a substantial knowledge-base that will support future research in plant pathology.

Methods

Fungal culture

P. nodorum SN15 was maintained on V8-PDA medium. For the induction of extracellular and intracellular proteins, 1 x 10^6 *P. nodorum* SN15 spores were grown in Fries broth (Liu, Faris et al. 2004). For genomic DNA, RNA and protein extraction experiments involving the intracellular and cell-wall/membrane sub-proteomes, 1 x 10^6 *S. nodorum* SN15 spores were grown in minimal medium broth for 3 days (Solomon, Tan et al. 2004). The mycelium was harvested and freeze-dried prior to further manipulations.

Genomic DNA extraction and Illumina sequencing

P. nodorum SN15 genomic DNA was extracted using a modified high-salt cetyltrimethylammonium bromide (CTAB) protocol (Clarke 2009). Briefly, freeze-dried mycelia were ground to a fine powder using a chilled mortar and pestle. Genomic DNA was extracted using an extraction buffer that consisted of 100 mM Tris, 50 mM EDTA, 2M NaCl, 0.4% (v/v) β-mercaptoethanol, 2% (w/v) polyvinylpyrrolidone and 2% (w/v) CTAB. The genomic DNA was subjected to phenol/chloroform extraction, ethanol precipitation and washes. A paired-end library with an average insert size of 439 bp and read lengths of 100 bp was generated from SN15 genomic DNA and used for sequencing. Sequencing of the genomic DNA was carried by the Australian Genome Research Facility (Melbourne, Australia) using an Illumina HiSeq 2000 (Illumina, CA, USA).

RNA extraction and Illumina sequencing

P. nodorum SN15 total RNA was extracted using the Trizol reagent (Invitrogen, CA, USA) and DNase-treated. PCR was used to check that the sample is free of genomic DNA (Tan, Heazlewood et al. 2008). RNA sequencing was carried out by Macrogen (Seoul, South Korea) using an Illumina HiSeq 2000 platform to generate 100 bp paired-end reads.

Raw Illumina sequencing reads were inspected with FastQC (Andrews 2010). Adapter sequence and low quality ends were removed with Cutadapt v1.0 (Martin 2011). Parameters and run details are available in appendix A5-1.

Proteomic datasets

The extracellular proteome was extracted as described by Vincent et al. (Vincent, Tan et al. 2012), using a modified TCA/acetone protein precipitation procedure. Briefly. Proteins from the extracellular culture filtrate were precipitated, collected by centrifugation and washed with 100% acetone. The protein pellet was subsequently air-dried at room temperature and suspended in 20 mM Tris pH 7. Residual TCA was progressively removed by dialysis of the suspension using D-Tube™ Dialyzer Maxi, MWCO 3.5 kDa (Novagen, Darmstadt, Germany) in several changes of 20 mM Tris pH 7 at 4°C for 48 hrs. Solubilised proteins were retained and stored at -80°C until further manipulation.

The intracellular proteome was extracted as previously described by the authors of this study (Tan, Heazlewood et al. 2008). Briefly, intracellular proteins from mechanically ground freeze-dried mycelia were solubilised in 20 mM Tris-Cl pH 7 and de-salted using a PD10 chromatography column (GE Healthcare, UK). Solubilised proteins were retained and stored at -80°C until required.

To facilitate cell wall/membrane proteome extraction, freeze-dried fungal mycelia were ground with a mortar and pestle and washed three times with 20 mM Tris-Cl pH 7 to release and remove soluble intracellular proteins. The pellet was then washed three times with 0.1 M Na_2CO_3 to further remove soluble and peripherally-attached proteins. The pellet was then resuspended in 20 mM Tris-Cl pH 7 and subjected to 3 cycles of slow freeze and thaw to further break up the cellular material. Membrane-bound proteins were extracted using two methods.

Extraction Procedure 1 (EP1): One hundred milligrams of membrane enriched pellet was extracted with 2% (w/v) SDS, 100 mM EDTA and 50 mM DTT in 100 mM Tris/HCl (pH7.8) by vortexing and boiling for 5 minutes followed by 5 minutes on ice (based on methods presented in Meijer et al., 2006 and Feiz et al., 2006).

Extraction Procedure 2 (EP2): One hundred milligrams of membrane enriched pellet was extracted with 2% (w/v) SDS, 7 M urea, 2 M thiourea and 50 mM DTT in 125 mM triethylammonium bicarbonate (TEAB, pH 8.5) by vortexing and sonication for 15 minutes in an ice-coeld sonication bath followed by resting for 30 minutes on ice. Vortexing and sonication steps were repeated. Subsequent sample processing for suspensions derived from 'Extraction Procedure' EP1 and EP2 were identical. Suspensions were centrifuged at 16,000 x g for 5 minutes (4°C) and the supernatants removed. Pellets were washed twice with either 100 mM Tris/HCl (pH 7.8) for EP1 or 100 mM TEAB (pH 8.5) for EP2. Respective supernatants were pooled, centrifuged at 20,000 x g for 15 minutes (4°C) and collected for further processing. Proteins were precipitated from supernatants by the addition 100% TCA (tricholoacetic acid) to 20% (v/v) and incubated on ice for 30 minutes. Protein precipitates were harvested by centrifugation at 20,000 x g for 10 minutes (4°C). Pellets were washed twice with 90% (v/v) acetone and centrifuged each time as before. Protein pellets were briefly dried under a gentle stream of nitrogen and used immediately. The final pellets were re-suspended in 45 µL of EP2 extraction buffer (without DTT) and 5 µL of 1 M TEAB (pH 8.5) by repeated vortexing and incubating the tubes for 10 min in an ice cold sonication bath. Samples were centrifuged at 20,000 x g for 10 minutes (4°C) and supernatants collected for further processing. Protein concentration of all samples was determined using the 2D-Quant kit (GE Healthcare) according to the manufacturer's 'Standard procedure' protocol. CWM proteins were digested without prior fractionation.

Intracellular and extracellular proteins were separated into 24 fractions, based on their isoelectric point using OFFGEL fractionation (Agilent 3100 Offgel fractionators) followed by trypsin-digestion and LCMS analysis of peptides. Offgel separations were performed using high resolution pH 3-10 separation kits (Agilent) wieth 1 mg of protein per strip as described previously (Hastie, Headlam et al. 2012). CWM proteins were digested without any fractionation. Recovered Offgel fractions and CWM preparations were reduced by addition of tris (2-carboxyethyl) phosphine to 22 mM and alkylated by addition of iodoacetamide to 122 mM. Reduced and alkylated proteins were co-precipitated overnight with 1 ug of trypsin by addition of ice-cold methanol. The recovered protein pellet was proteolytically digested in two steps with a second addition of 1 µg trypsin as described in Hastie et al. Peptides were separated on a C18 reversed-phase column and data collected using a Hybrid LTQ Orbitrap mass spectrometer (Thermo Fischer Scientific, Bremen, Germany). Offgel fraction digests were separated by reverse phase capillary HPLC using a Prominence nano HPLC system (Shiumadzu, Kyoto, Japan), Vydac Everest C18 5 µm 150 µm x 150 mm column and analysed on an LTQ Orbitrap XL as described in Morrison *et al* (Morrison, Hastie et al. 2012). The separation gradient was 2-30% B over 78 minutes (A- 0.1% (v/v) aqueous formic acid; B-80% (v/v) ACN, 0.1% (v/v) aqueous formic acid) followed by a 95% B wash step, with a total run time 110 minutes. Data was acquired in the Orbitrap XL as described by Hastie (Hastie, Headlam et al. 2012).

CWM digests were loaded onto a Reprosil aq C18 3 µm, 120 A, 300 µm trap (SGE pn-2222066) at 30µL/min in 2% (v/v) acetonitrile 0.1% (v/v) aqueous formic acid for 3.5 minutes at 50 ºC, then switched in-line with an analytical column (15 cm x 75 µm fused silica, self-packed) reprosil aq C18 2.4um (pn-r124.aq batch-9756, Dr. Maisch GmbH) using a flow rate of 1µL/min and 98% solvent A (0.1%(v/v) aqueous formic acid) , 2% solvent B (80%(v/v) ACN, 0.1% (v/v) aqueous formic acid). Peptides were separated at 50 °C using a sequence of linear gradients: to 7% B over 3.5 minutes; to 35% B over 166.5 minutes; to 45% B over 10 minutes; to 95% B over 10 minutes and then holding the column at 95% B for 10 minutes.

Eluate from the analytical columns was introduced into the LTQ-Velos Orbitrap throughout the entire run via a Nanospray Flex Ion Source (Thermo Fisher Scientific) and a 30 µm inner diameter uncoated silica emitter (New Objective). Typical spray voltage was 1.4 kV with no sheath, sweep or auxiliary gases used. The heated capillary temperature was set to 250ºC. The LTQ- Velos Orbitrap ETD was controlled using Xcalibur 2.2 software (Thermo Fisher Scientific) and operated in a data-dependent acquisition mode to automatically switch between Orbitrap-MS and ion trap-MS/MS as described previously.

These spectra were then searched using the tide search engine (Diament and Noble 2011) implemented in the crux toolkit (Park, Klammer et al. 2008) with specifications as follows: spectra mapped against: 6-frame translations of both the new and the old genome assemblies and the set of predicted protein sequences from both the new and the old annotations. The search parameters used were: variable modifications, oxidation (M); and deamidation (NQ); fixed modification, carbamidomethyl (C); peptide tolerance, 20 ppm; MS/MS tolerance: ±0.8 Da; Digestion enzyme: trypsin; maximum missed cleavages: 1. Peptide-spectrum matches were refined using Percolator (Käll, Canterbury et al. 2007), again as implemented in the crux toolkit.

For 1D-LC MALDI MS/MS analysis of the SN15 extracellular proteome, SN15 trypsin-digested peptides were resuspended in 20 µl of 2% acetonitrile and 0.05% trifluoroacetic acid. Peptides were loaded onto a C18 PepMap100, 3 mm column (Dionex, CA, USA) through the Ultimate 3000 nano HPLC system (Dionex, CA, USA). Mass spectrometry analysis was carried out on a 4800 MALDI TOF/TOF Analyser as previously described (Casey, Solomon et al. 2010). These spectra were also searched using the tide search engine (Diament and Noble 2011) with specifications: variable modifications, oxidation (M); fixed modification, carbamidomethyl (C) and other parameters and post-processing as above.

Conflicts with existing annotations were identified where proteomic spectra searched against the six-frame translation of the genome mapped into intergenic regions, intronic annotations, coding regions in the wrong frame.

Improvements to the SN15 Genome Assembly

SNP and indel errors in the *P. nodorum* SN15 assembly sequence (Hane, Lowe et al. 2007) were corrected by MIRA (v3.4.1.1) (Chevreux, Wetter et al. 1999), using its mapping algorithm to assemble Illumina gDNA reads onto the pre-existing scaffolds. The original Sanger-sequenced reads were also re-mapped to the corrected assembly using BWA v0.7.3a-r367 (Bringans, Hane et al. 2009).

Groups of putative scaffold linkage groups were predicted by comparison to *Pyrenophora tritici*-repentis (Manning, Pandelova et al. 2013) using the synteny-based cumulative binomial test described by Hane, Rouxel et al. (2011).

In order to assess the outcomes of genome sequence and gene annotation corrections, various diagnostic tests were performed. Changes made to the corrected genome were calculated with the dnadiff tool distributed with MUMmer (Kurtz, Phillippy et al. 2004). Improvements of WGS read mapping to the corrected assembly were calculated by alignment with BWA v0.7.5a-r405 using the default parameters and summary statistics calculated with Picard v1.9.4 (Simakov, Marletaz et al. 2013)(Simakov, Marletaz et al. 2013)(Simakov, Marletaz et al. 2013). Improvements of RNA read mapping to the corrected assembly were calculated by alignment with TopHat v2.0.12 (Kim, Pertea et al. 2013) and summary statistics calculated with Picard v1.9.4 and from the TopHat reports.

Improvements to the Genome Annotations

Errors in *P. nodorum* SN15 gene annotations were corrected using a combination of supporting data from RNA-seq and proteogenomic peptide alignments to the corrected assembly. RNA-seq reads were mapped to the corrected genome using TopHat v2.0.8 (Kim, Pertea et al. 2013). Manual correction of gene models and was performed using WebApollo (Lee, Helt et al. 2013). JBrowse (Skinner, Uzilov et al. 2009), through WebApollo was used to visualise the various omics data sources that informed the manual correction.

RNA sequencing reads were aligned to the genome, and the gene models identified by Bringans et al (Bringans, Hane et al. 2009) were checked to ensure they matched all introns supported by 5 or more RNA-seq reads. Introns were introduced or removed from the annotations to match the

RNA-seq data. New genes were annotated where transcription levels exceeded 5x when a suitable open reading frame (ORF) could be found. Gene annotations were split when the RNA-seq depth dropped to 0 and the concatenated protein's blast hits showed two moieties of hit coverage.

RNA-seq depth was also used to correct events where an open reading frame occurred inside the intron of another gene. These events were identified by large changes in read depth at a single locus. For each intronic insertion annotation, the translated region of the splice site skipping over the internal ORF was checked for consistency with blast results and with InterProScan-predicted domains that spanned that splice site.

Exported and cleaned gff3 and fasta files were checked into git version control for distributed backup, sharing and review (https://github.com/robsyme/Parastagonospora_nodorum_SN15). Genome-wide support for gene annotations was summarised according to evidence type, requiring 80% coverage of coding sequence length and 5 X coverage for RNA-seq support, peptides mapping within the coding region for proteogenomic support and four or more for microarray probes showing with expression levels at or above the cut-off determined by Ipcho, Hane et al. (2012).

All gene annotations were manually reviewed and curated using the WebApollo platform, checking for consistency with RNA-seq, proteomics, microarray, blast hits against nr and conserved protein domain structures. Matches to conserved protein domains identified from translated gene models using InterProScan v5.8-49.0 (Jones, Binns et al. 2014) were compared between previously published and corrected datasets. Each protein set was submitted to dbCAN (Yin, Mao et al. 2012) for CAZyme enzyme family identification. GO functional annotations assigned by InterProScan were analysed for functional enrichment of the new protein set using the Fisher's test implemented in the goatools package (Haibao Tang 2015).

Annotation and Comparison with Alternate Strains

P. nodorum strains SN4 and SN79 were re-annotated using Maker v2.31.8. Evidence supplied to Maker included the updated SN15 protein set and ab-initio predictions from the ab-initio mode of gene predictor CodingQuarry (Testa, Hane et al. 2015) using parameters generated from training on the updated SN15 annotations.

The predicted protein set from the three *P. nodorum* strains were clustered using ProteinOrtho v5.11 using the synteny option. Execution of many parts of the analysis including the ProteinOrtho clustering were aided by GNU parallel (Tange 2011) and BioRuby scripts and gems (Goto, Prins et al. 2010, Bonnal, Aerts et al. 2012).

Results

Genome Assembly Sequence Correction

The genome of *P. nodorum* SN15 was re-sequenced using 100 bp paired-end Illumina libraries yielding 11.0 Gbp of raw sequence data equivalent to approximately 290x coverage. Short-reads were reassembled using the MIRA mapping algorithm to resolve or remove 37,501 Ns and correct 12,911 SNPs, 1005 deletions, and 16,820 insertions (Table 1).

The genome annotations as described by Bringans, Hane et al. (2009) were supplied as input to the MIRA assembly so that gene coordinates and identifiers could be preserved despite the correction of insertions and deletions to the underlying assembly.

Table 1: Summary statistics comparing changes to the genome assembly and annotation. Correction of the original assembly with deep sequencing with short reads increases the number of reads that align to the genome, reduces the number of unknown bases and allows for new scaffold joins. An insertion corrected is a deletion of erroneous sequence from the original assembly and a deletion corrected is the insertion of sequence missing from the original assembly. 37501 base pairs of unknown sequence have been resolved in the corrected genome sequence. WGS and RNA read mismatch rate is the rate of bases mismatching the reference for all bases aligned to the reference sequence. WGS and RNA indel rate is the number of short insertions/deletions seen in reads / total aligned bases. The percentage of reads aligned in pairs is the percentage of reads whose mate pair was also aligned to the reference.

Description	*Before*	*After*	*Change*
Number of nuclear scaffolds	108	91	-17
SNP changes	n/a	n/a	12911
Single bp insertion corrected	n/a	n/a	16820
Single bp deletion corrected	n/a	n/a	1005
Ns count	164388	126887	-37501
WGS Reads mapping to genome >=q20 (bp)	93867773	94594136	726363
WGS read mismatch rate (%)	0.5623	0.4851	-0.0772
WGS indel rate (%)	0.0615	6.2e-03	-0.0553
WGS reads aligned in pairs (%)	99.6402	99.6427	2.5e-3
RNA Reads mapping to genome >=q20 (bp)	5872361103	10842396864	4970035761
RNA indel rate (%)	0.0348	0.0043	-0.0305
RNA reads aligned in pairs (%)	95.0119	96.1274	1.1155

The corrected genome sequence allowed for an additional 726 kb of DNA reads to be mapped. Similarly, an additional 4,970 Mb of RNA reads were able to be mapped to the corrected assembly compared to the uncorrected assembly. The reads mapped with lower rates of mismatch (0.4851% for DNA), and insertions/deletions (0.0062% for DNA and 0.0043% for RNA). The number of reads mapping in concordant pairs increased to 99.6% for DNA and 96.1% for RNA (Table 1).

Proteomic mass spectral matches from extracellular, membrane-bound and intracellular protein fractions were pooled. Matches isolated by more than 200 bp from another match were discarded as likely false-positives. Existing annotations were checked for reading-frame consistency with the remaining spectral matches and new proteins were annotated or existing annotations extended where spectral search results fell outside the coding regions.

Sanger-sequenced reads from 4 and 40 kb libraries were aligned to the corrected assembly and paired-end information was used to join and orient scaffolds (Table 2). We identified read-supported scaffold pairings and orientation by filtering Sanger reads where each read in a pair mapped to a different scaffold, where each of the pairs mapped at only one position in the genome, and where each of the pairs mapped within 40 kb of the scaffold ends. We excluded scaffold joins where multiple read pairs suggested conflicting pairs or orientation, leaving only unambiguous joins. This process linked 16 scaffolds.

Scaffolds 76, 92, and 106 were identified by blast as misassembled high-identity matches (>95%) to the mitochondrial genome sequence and were excluded from the nuclear genome assembly.

The repeat content of the new assembly was reassessed. Subtelomeric repeats R22 and X48 (Hane, Lowe et al. 2007) are modestly expanded in the corrected assembly, but repeat content remains largely unchanged (Appendix A5-2).

Table 2: Additional scaffolds joined by long insert libraries after correction. The indicated orientation is relative to the original assembly scaffolds.

Left scaffold	*Right scaffold*	*Orientation*
scaffold_2	scaffold_107	← →
scaffold_7	scaffold_105	→ →
scaffold_8	scaffold_26	→ →
scaffold_17	scaffold_36	→ →
scaffold_18	scaffold_77	← ←
scaffold_20	scaffold_49	→ →
scaffold_29	scaffold_48	← ←
scaffold_54	scaffold_64	← ←
scaffold_60	scaffold_72	← ←
scaffold_28	scaffold_61	← →
scaffold_29	scaffold_85	→ →

scaffold_33	scaffold_17	← →
scaffold_51	scaffold_55	← →

Gene Model Correction Summary

After genome corrections, there were 13,563 predicted nuclear genes (Table 3), of which 866 are new genes at new loci and 1936 are confirmed genes that had been regarded as doubtful in earlier revisions. New genes are numbered starting at 30,001.

In total, 12,143 (89%) genes in the current list possess some form of experimental support (Figure 2). Microarray probe intensity supported the transcription of 9961 loci (Figure 2). RNA-seq supported the exon structure of 10544 gene models, including 299 loci with at least one alternatively spliced transcript, bringing the total number of predicted protein-coding genes to 13,949. 8,366 existing genes have had their protein sequence modified, 1936 previously deprecated loci have been reinstated, and 866 new genes were introduced when the previous genome annotation had incorrectly split genes (55 occurrences, Appendix A5-3), joined genes (356 occurrences, Appendix A5-3) or where there was no previous annotation (455 occurrences) (Table 3, Appendix A5-5). Four intronic insertion events were annotated where an open reading frame occurred within another gene (SNOG_30297, SNOG_30841, SNOG_14322, and SNOG_16073) (Figure 3). Blasting the inserted endonuclease protein sequences to the NCBI non-redundant protein database returns only hits to fragments of loci annotated as the host gene.

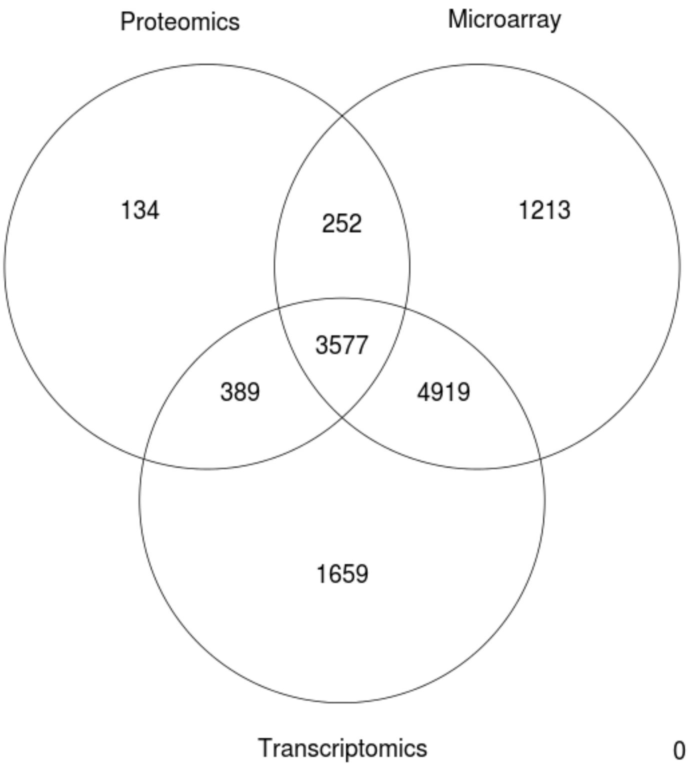


Figure 2: Sources of evidence for corrected annotations. Using a variety of omics sources to guide genome annotation gave 12143 annotations some level of experimental support. The remaining annotations are supported by non-experimental sources such as the presence of conserved domains or homology to genes in other species.

Table 3: Summary statistics showing results to changes to key gene model metrics from Bringans et al to the updated set of annotations.

	Bringans et al	Corrected
Gene model count	10761	13563
Average exon count	2.6	2.5
Average CDS length (bp)	1400.6	1372.3
Intergenic distance mean (bp)	1685	1011
Intergenic distance std dev (bp)	2590	2063
Intron length mean (bp)	91.3	66.6
Models with peptide support	2665	4352
Models with peptide conflict	150	0
Genes with alternative transcripts	0	299
PKS genes	19	24
Interproscan		
Proteins with Pfam domains	11111	13245
Proteins with Gene3D domains	11181	13180
Proteins with SignalP predictions	1029	1476
Proteins with SMART predictions	4160	5400
Proteins with ProSite Profiles	4115	5279
Promoted genes	-	1936
New or promoted genes	-	866
Correcting split genes	-	55
Correcting joined genes	-	356
At new loci	-	455
CAZyme Family		
Auxiliary Activity	122	139
Carbohydrate-Binding Module Family	64	110
Carbohydrate Esterase Family	142	174
Dockerin	1	1
Glycoside Hydrolase Family	264	280
Glycosyl Transferase Family	96	105
Polysaccharide Lyase	10	10

Clustering the 13,949 SN15 proteins with the re-annotated SN4 and SN79 (proteome sizes 13,899 and 13,746) derived from the improved reference annotation reveals a core *P. nodorum* protein set of 11,849 clusters (Figure 4).

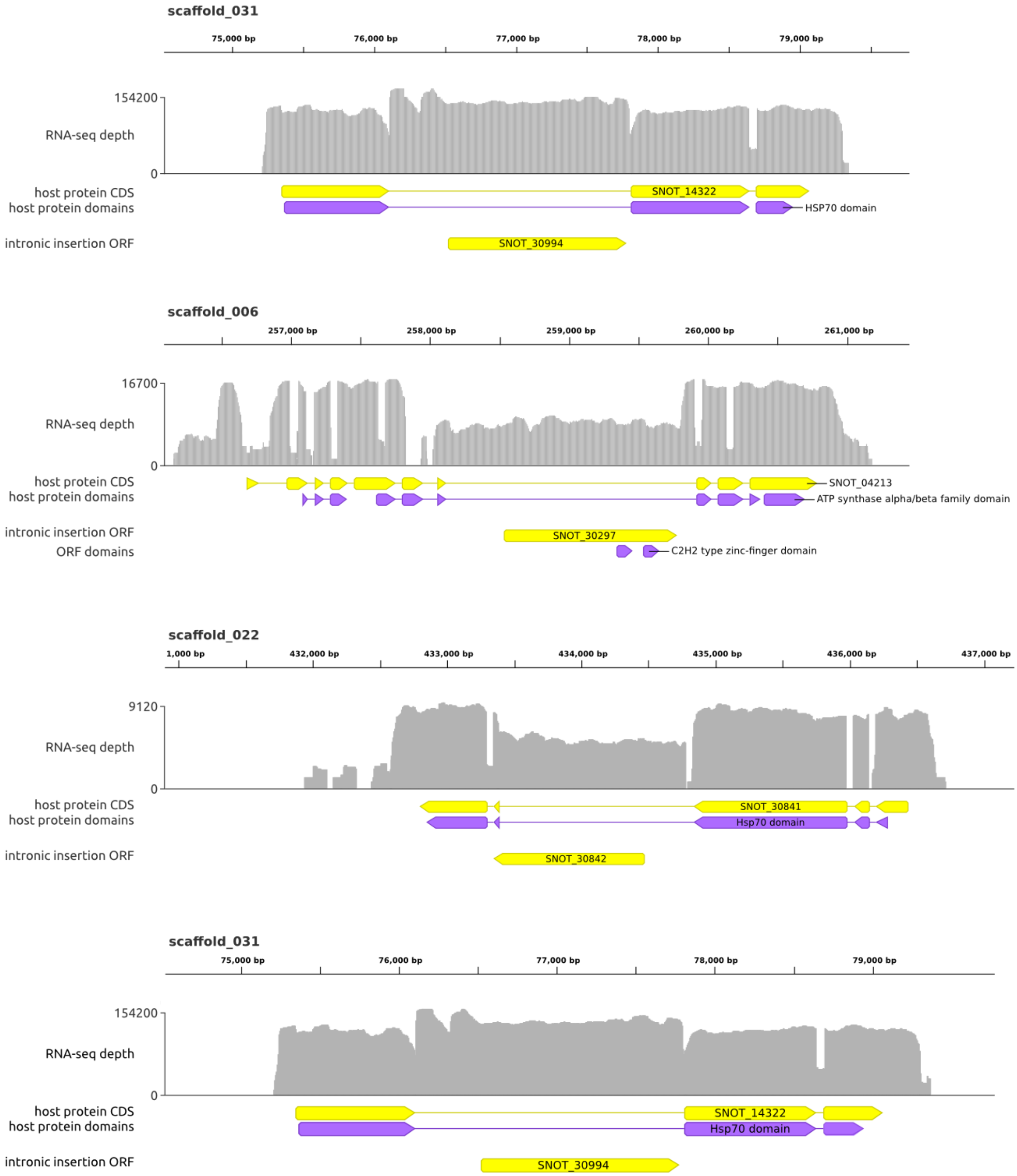


Figure 3: CDS annotation, RNA-seq depth and PFam domains of the four genes with targeted ORF intronic insertions. RNA-seq depth is shown on a log scale. Three of the four genes hosting insertions encode HSP70 proteins.

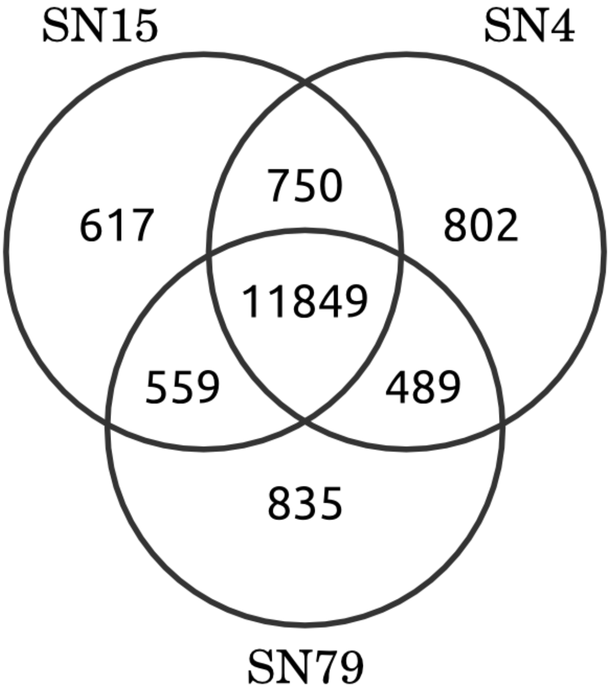


Figure 4: Protein cluster membership across the three sequenced *P. nodorum* strains showing 11,849 core conserved protein clusters. Orthologous clusters were derived from predicted proteomes with sizes 13,949, 13,899 and 13,746 for SN15, SN4, and SN79 respectively.

Functional Annotation Improvements

Comparison of each predicted protein to their top BLAST hit not belonging to the *Parastagonospora* genus reveals the new annotation set to be more concordant with annotations in other species (Figure 5). In particular, we observe a dramatic shift from shorter annotations to longer annotations that represent a higher proportion of the length of their best-matching homolog. Manual correction has eliminated occurrences of conflict between the predicted protein set and the mapped location of proteomic spectra (Table 3).

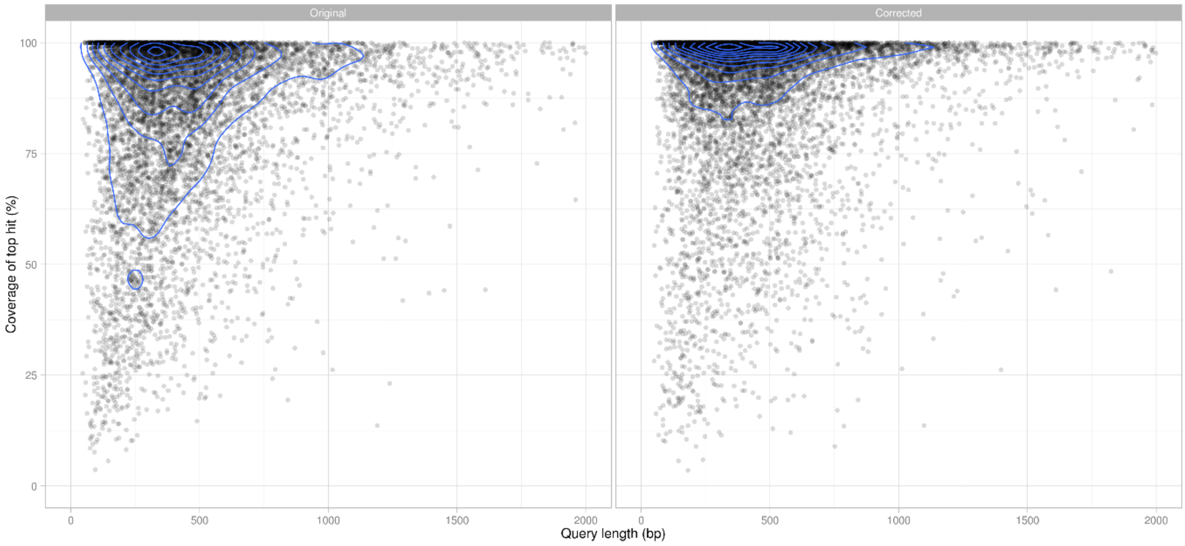


Figure 5: Coverage of the top blast hit for each predicted protein. Contours of the kernel density estimate are shown in blue. Only proteins smaller than 2000 amino acids are shown. The change from the original set (left) to the corrected set (right) is characterised by a general upwards shift towards more complete coverage of each protein's top hit.

Compared to the gene annotations from Bringans, Hane et al. (2009), the new set includes 1784 more proteins with predicted Pfam domains (Finn, Bateman et al. 2013), 1897 more with Gene3D domains (Lees, Lee et al. 2014), and 354 more with SignalP - predicted signal peptides (Bendtsen, Nielsen et al. 2004) (Table 3). CAZyme classifications show an increase in the number of proteins belonging to the carbohydrate-binding module (46), carbohydrate esterase (32), glycoside hydrolase (16), and glycosyl transferase (9) families (Table 3).

Genes and Domains of Interest

Known *P. nodorum* effectors Sn*ToxA*, *SnTox1*, and *SnTox3* are not homologous but do share common characteristics. They are small (13 kDa, 10 kDa and 17 kDa respectively), contain signal peptides to target the protein to the secretory system and have a high number of cysteine residues which may form disulphide bridges that help maintain protein stability once secreted. Their genes are positioned close to repeats. It has been suggested that effector proximity to repeats may expose them to an elevated level of mutation due to leakage of the RIP process outside truly repetitive sequence (Rouxel, Grandaubert et al. 2011). The known *P. nodorum* effectors are absent from the SN79 strain, and are highly expressed early in infection (Ipcho, Hane et al. 2012). The 866 proteins annotated at new loci are a set enriched for elements that have the properties we expect of necrotrophic effectors. The newly annotated proteins have products with a higher average cysteine content than the unchanged or modified proteins (Figure 6). Of the 54 proteins in the corrected set with more than 4% cysteine content, 16 are from genes at previously unannotated loci, and 51 have no blast hits to nr (Table 4). The corrected set revealed 187 extra proteins with blast hits to entries in the PHIbase pathogen-host interaction database (Urban, Pant et al. 2014) that are experimentally shown to influence pathogenicity (Table 5). Included among the cysteine-rich genes at new loci is a putative degraded copy of *P. nodorum* effector Tox1 (Table 4, Appendix A5-8).

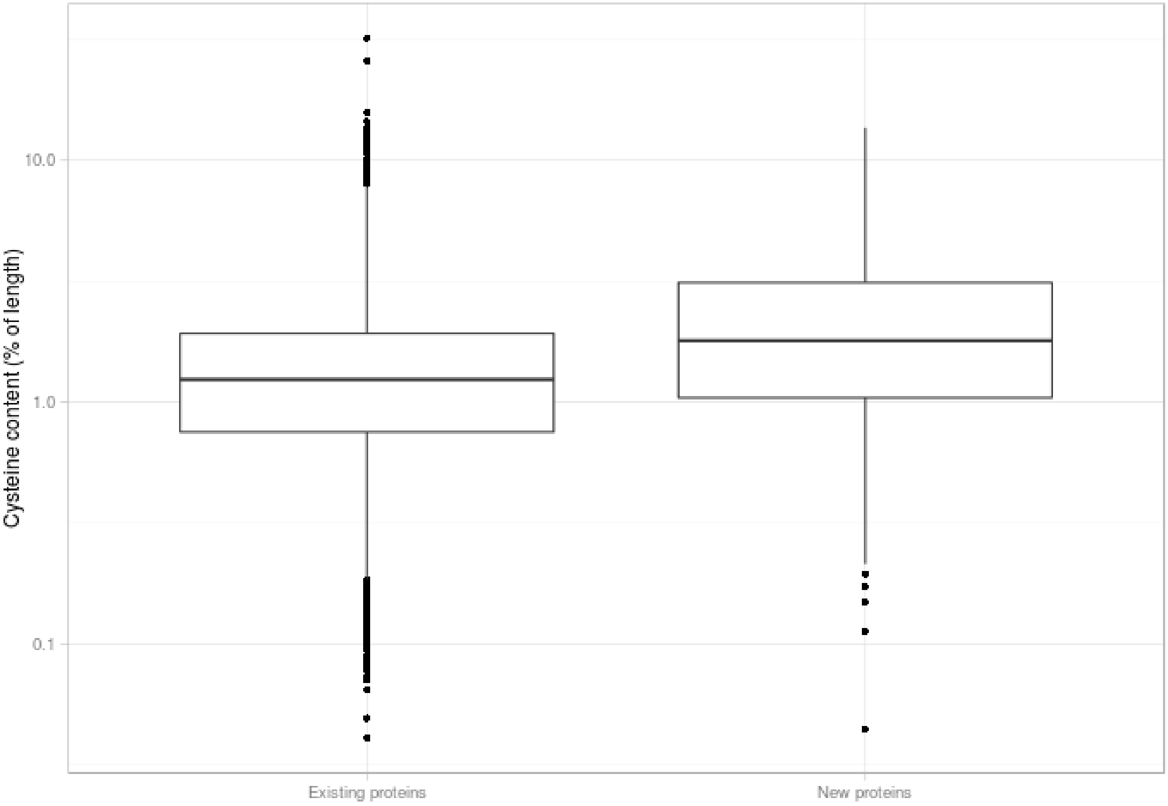


Figure 6: Comparison of the distribution of protein cysteine content between genes found at previously unannotated loci (right) and all other genes (left). Horizontal bars in boxplot are mean, 1st quartile and 3rd quartile. Outliers greater than 1.5 x interquartile range are shown as points. Proteins at newly annotated loci, unannotated by automated methods are more likely to be cysteine rich.

Table 4: Cysteine-rich genes annotated at previously unannotated loci.

Gene name	Protein length	Cysteine count	Cysteine percentage	Blast hits
SNOG_30077	66	9	13.6	No
SNOG_30525	74	10	13.5	No
SNOG_30316	94	11	11.7	No
SNOG_30335	70	8	11.4	No
SNOG_30888	53	6	11.3	No
SNOG_30837	56	6	10.7	No
SNOG_30741	355	37	10.4	Carbohydrate-binding
SNOG_30253	58	6	10.3	No
SNOG_30352	79	8	10.1	No
SNOG_30019	60	6	10	No
SNOG_30451	62	6	9.7	Fungal hypothetical genes
SNOG_30925	104	10	9.6	No
SNOG_30828	84	8	9.5	No
SNOG_30466	84	8	9.5	Tox1
SNOG_30530	76	7	9.2	No
SNOG_30989	55	5	9.1	No

Effectors and other components of pathogenicity are likely to be members of the set of 2169 protein clusters present in at least one wheat pathogen but absent from the avirulent SN79 strain (Figure 4).

Table 5: New hits to PHIbase. 187 proteins had blast hits to PHIbase in the corrected set, but not in the original set. Of these new PHIbase targets, 26 were at new loci and 161 were from genes previously demoted or gene whose correction had revealed the blast hit.

PHIbase Class	New proteins	Modified proteins
Mutants are lethal	2	3
Mutants have reduced virulence	20	120
Mutants show mixed results	1	17
Effector (plant avirulence determinant)	0	2
Mutants have lost pathogencity	3	11
Mutants have increased virulence	0	4
Chemistry target (unknown phenotype)	0	4

All but one of the polyketide synthase (PKS) genes have had their gene structure modified (Table 6, Appendix A5-6). The modified protein models were used by Chooi, Muria-Gonzalez et al. (2014) to identify 24 PKS genes with 1 type III PKS, 1 hybrid non-ribosomal peptide synthetase/PKS, 1 partially reducing PKS, 7 non-reducing PKSs, and 14 highly reducing PKSs. Two extra proteins with putative pathogenicity domains HCE2 and Ricin-type beta-trefoil lectin are uncovered in the new protein set (Table 8, Appendix A5-7). Pfam domains with an increased representation in the new protein set include DNA-binding domains (117), transcription factors (51) and chitin-binding sequence (21) (Table 8).

Table 6: The known PKS complement of *P. nodorum*. Only one PKS gene (SNOG_06676) annotation remained unmodified after assembly and annotation correction.

Gene name	PKS Type	Annotation modifications made
SNOG_09622	Type III PKS	Two intron/exon boundaries shifted
SNOG_00308	Hybrid nonribosomal peptide synthetase/PKS	One intron removed
SNOG_00477	Partially reducing PKS	Exon extended at 5' end by 195 bp
SNOG_02561	Highly reducing PKS	Two exons added at 5' end, original intron removed
SNOG_04868	Highly reducing PKS	2 SNP changes and a 1 bp insertion removed. Insertion removal caused frameshift and false intron removed
SNOG_05791	Highly reducing PKS	Six exons added at 5' end, two other exons extended
SNOG_06676	Highly reducing PKS	No changes made
SNOG_07866	Highly reducing PKS	One exon extended, one intron removed
SNOG_09623	Highly reducing PKS	One exon reduced, one exon extended
SNOG_11066	Highly reducing PKS	One extended, three new exons introduced, exons 6, 7 and 8 joined (introns removed)
SNOG_11076	Highly reducing PKS	Nine 1 bp insertions removed, Nine intron/exon boundaries shifted
SNOG_11272	Highly reducing PKS	Two 1 bp insertions removed, one intron removed, three intron/exon boundaries shifted
SNOG_12897	Highly reducing PKS	Two introns removed, two exons added at 5' end
SNOG_13032	Highly reducing PKS	Four insertion events corrected, two introns removed, one exon added at 5' end
SNOG_14927	Highly reducing PKS	One exon added, three intron/exon boundaries shifted
SNOG_15965	Highly reducing PKS	One exon added, three intron/exon boundaries shifted

SNOG_09490	Highly reducing PKS	5 insertions and 6 SNPs corrected, two introns removed, two exons added 5 intron/exon boundaries shifted
SNOG_06682	Non-reducing PKS	Two 1 bp insertions corrected, two introns removed, four intron/exon boundaries shifted
SNOG_07020	Non-reducing PKS	Two introns removed, one intron/exon boundary shifted
SNOG_11981	Non-reducing PKS	Two insertions removed, two introns removed, one intron added, and two intron/exon boundaries shifted
SNOG_15829	Non-reducing PKS	One insertion removed, two exons added at 5' end, two introns removed, three intron/exon boundaries shifted
SNOG_08274	Non-reducing PKS	Two exons added at 5' end, four intron/exon boundaries shifted
SNOG_08614	Non-reducing PKS	One intron removed, two exons added at 5' end, three intron/exon boundaries shifted
SNOG_09932	Non-reducing PKS	Two exons added, three exons removed, and four intron/exon boundaries shifted

Table 7: Changes in the number of annotations with conserved protein domains.

Domain ID	Domain name	Original Protein Count	Corrected Protein Count
Putative pathogenicity-related domains			
PF14856	HCE2 - Pathogen effector; putative necrosis-inducing factor	1	2
PF00652	Ricin-type beta-trefoil lectin	0	1
Top 5 domains with increased hits			
IPR001138	Zn(2)-C6 fungal-type DNA-binding domain	89	206
IPR007219	Transcription factor domain	91	142
IPR000719	Protein kinase domain	120	157
IPR001810	F-box domain	49	77
IPR001002	Chitin-binding	14	35

Discussion

The completeness and accuracy of an organism's reference genome sequence and its gene annotations directly influence the validity of computational and reverse genetics-based downstream functional studies. This is especially relevant in plant pathology, for which considerable research efforts are invested into predicting and functionally characterising putative effector genes from genomic datasets. Identification of effectors and subsequent effector-assisted breeding programs have been an important contribution to crop protection against pathogens (Vleeshouwers and Oliver 2014). Screening of potential lines with a purified effector negates or diminishes the need for more costly and time-consuming infection assays and field trials. Analysis based on protein sequence such as effector prediction or functional annotation rely on accurate gene models, and by extension, assembly sequence.

For example, insertion and deletion errors in the underlying assembly sequence can force automated gene calling software to introduce erroneous intron features in order to extend an open-reading frame. This can lead to an inflated exon count (Table 3) and interrupt blast and/or protein domain matches, which can impair assignment of biologically relevant functional terms to genes. The insertion of an intronic endonuclease into a gene can cause nested open reading frames, which will not be correctly annotated by current automated gene calling software. The inserted open reading frame and can lead to the confusion of gene function annotation by the introduction of conserved domains that do not relate to the function of the host gene.

A number of corrections have been made to the *P. nodorum* SN15 genome assembly, reducing the number of nuclear scaffolds from 107 to 91. SNP and indel removal facilitated by the addition of the Illumina data allowed a re-evaluation of the long-range paired end Sanger read data which, in turn, permitted the confident joining of 8 pairs of scaffolds. Eight scaffolds were joined that exhibited mesosyntenic relationships (Table 8). Scaffolds 8 and 26, for example both show mesosyntenic similarity to scaffold 4 on *P. tritici-repentis* (Figure 12). Joining scaffolds adds to our knowledge the genomic context of particular regions of the genome, including the transfercon harbouring the necrotrophic effector ToxA. Confirmation of the scaffold 55/51 join predicted by mesosyntenic pairings and by homology to the ToxA region in *P. tritici-repentis* (Syme, Hane et al. 2013) lends support to the theory of an expanded 72kb transfercon and subsequent repeat invasion in *P. nodorum*. Pulsed field gel electrophoresis has been previously used to resolve between 14 and 19 chromosomes from different *P. nodorum* isolates (Cooley and Caten 1991). Hence, a substantial number of gaps still remained unresolved in the current assembly.

Table 8: Scaffold joins predicted by mesosyntenic relations and validated by long insert Sanger reads. The indicated orientation is relative to the original assembly scaffolds.

Left scaffold	*Right scaffold*	*Orientation*
scaffold_8	scaffold_26	→ →
scaffold_29	scaffold_48	← ←
scaffold_37	scaffold_48	→ →
scaffold_51	scaffold_55	← ←

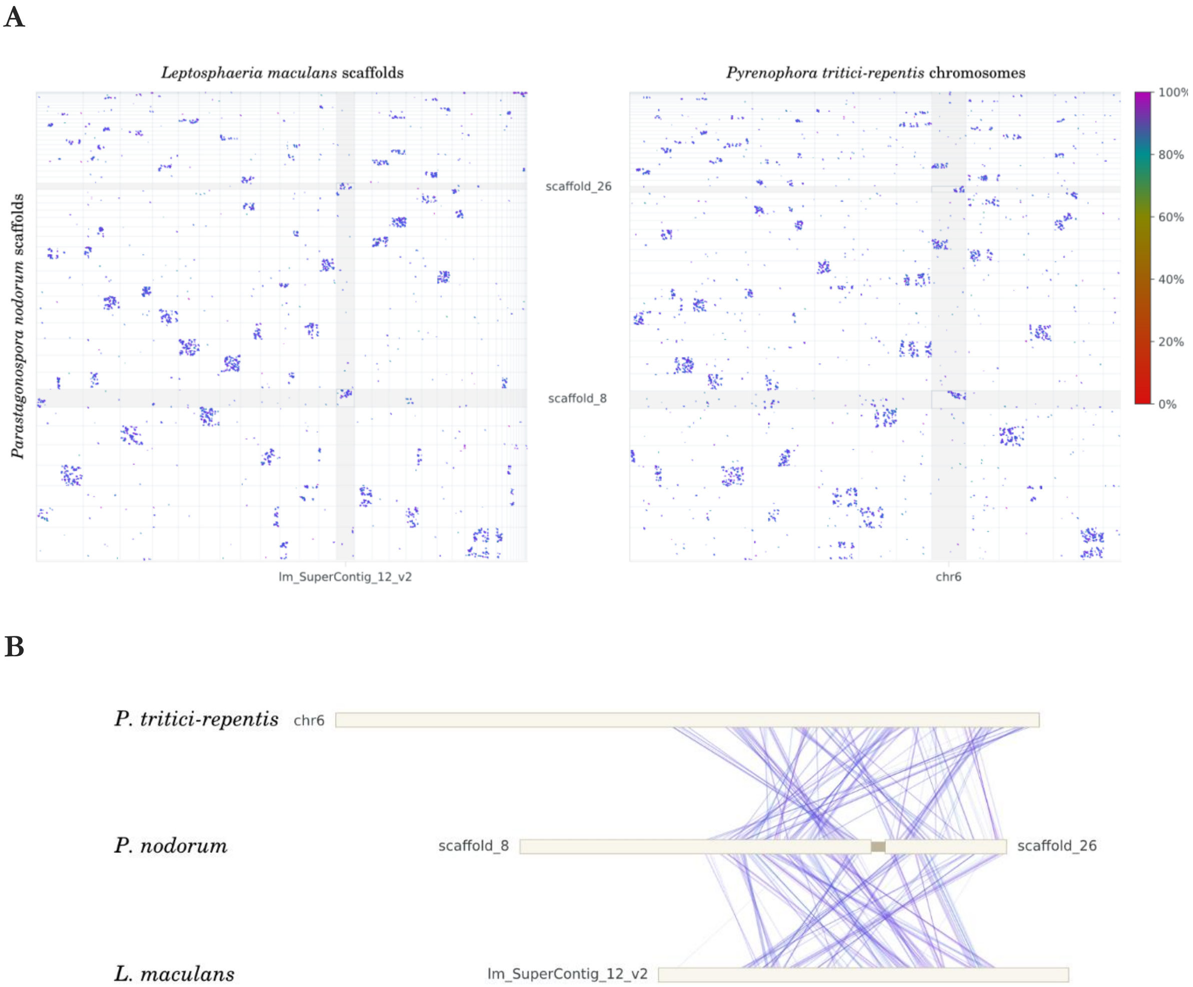


Figure 7: (A) Dotplot of promer matches between the original, unjoined *P. nodorum* scaffolds with *L. maculans* and *P. tritici-repentis*. *P. nodorum* scaffolds 8 and 26 are highlighted in grey. Scaffolds/chromosomes are highlighted in the alternate strains where they have mesosyntenic matches to *P. nodorum* scaffolds 8 and 26. In both *P. tritici-repentis* and *L. maculans*, scaffolds 8 and 26 have mesosyntenic matches to only one scaffold in the alternate strain which supports the join event. (B) Nucmer matches between the new *P. nodorum* scaffold 8 composed by joining the original scaffolds 8 and 26.

In addition to improvements in existing genes, manual annotation has also uncovered genes at new loci. Many of these new genes are small and cysteine-rich (Figure 4) with few blast hits to nr - characteristics of proteins involved in pathogenicity (Syme, Hane et al. 2013) and are effector candidates. Further evidence that these are relevant effectors could be obtained by determining whether they are expressed *in planta*.

The errors in the *P. nodorum* assembly sequence and its genome annotations are not unusual for a genome project of its age, assembly strategy and sequencing history. Similar fungal genome projects lacking 'multi-omics'-based evidence may therefore harbour undiscovered annotation and sequencing errors, adversely affecting the accuracy of its genome analysis and the accuracy of comparative genomics studies in which they have been used.

We present an integrated analysis of multiple genomic, transcriptomic and proteomic datasets and their application to the improvement of the genome assembly and gene annotations of the fungal pathogen *P. nodorum* SN15. Experimental approaches undertook in this study can readily be applied to other biological systems to refine gene models and assist in the assembly of uncompleted genomes. We anticipate that others establishing fungal genome projects would similarly benefit from the techniques described in this study.

Acknowledgements

This work was supported through access to facilities managed by Bioplatforms Australia and funded by the Australian Government National Collaborative Research Infrastructure Strategy and Education Investment Fund Super Science Initiative and Grains Research and Development Corporation research grant CUR00012. *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 and* by iVEC through the use of advanced computing resources located at the Pawsey Supercomputing Centre. RAS received funding through scholarships from GRDC (GRS10061) and Curtin University.

References

Abeysekara, N. S., T. L. Friesen, B. Keller and J. D. Faris (2009). "Identification and characterization of a novel host–toxin interaction in the wheat–*Stagonospora nodorum* pathosystem." Theoretical and applied genetics **120**(1): 117-126.

Andrews, S. (2010). "FastQC: A quality control tool for high throughput sequence data." Reference Source.

Belfort, M. and P. S. Perlman (1995). "Mechanisms of intron mobility." Journal of Biological Chemistry **270**(51): 30237-30240.

Belfort, M. and R. J. Roberts (1997). "Homing endonucleases: keeping the house in order." Nucleic acids research **25**(17): 3379-3388.

Bendtsen, J. D., H. Nielsen, G. von Heijne and S. Brunak (2004). "Improved prediction of signal peptides: SignalP 3.0." J Mol Biol **340**(4): 783-795.

Bonnal, R. J., J. Aerts, G. Githinji, N. Goto, D. MacLean, C. A. Miller, H. Mishima, M. Pagani, R. Ramirez-Gonzalez and G. Smant (2012). "Biogem: an effective tool based approach for scaling up open source software development in bioinformatics." Bioinformatics.

Bringans, S., J. K. Hane, T. Casey, K. C. Tan, R. Lipscombe, P. S. Solomon and R. P. Oliver (2009). "Deep proteogenomics; high throughput gene validation by multidimensional liquid chromatography and mass spectrometry of proteins from the fungal wheat pathogen *Stagonospora nodorum*." BMC Bioinformatics **10**: 301.

Casey, T., P. S. Solomon, S. Bringans, K. C. Tan, R. P. Oliver and R. Lipscombe (2010). "Quantitative proteomic analysis of G-protein signalling in *Stagonospora nodorum* using isobaric tags for relative and absolute quantification." Proteomics **10**(1): 38-47.

Casey, T., P. S. Solomon, S. Bringans, K. C. Tan, R. P. Oliver and R. Lipscombe (2010). "Quantitative proteomic analysis of G-protein signalling in *Stagonospora nodorum* using isobaric tags for relative and absolute quantification." Proteomics **10**: 38-47.

Chain, P., D. Grafham, R. Fulton, M. Fitzgerald, J. Hostetler, D. Muzny, J. Ali, B. Birren, D. Bruce and C. Buhay (2009). "Genome project standards in a new era of sequencing." Science (New York, NY) **326**(5950).

Chevreux, B., T. Wetter and S. Suhai (1999). Genome sequence assembly using trace signals and additional sequence information. German Conference on Bioinformatics.

Chooi, Y.-H., M. J. Muria-Gonzalez and P. S. Solomon (2014). "A genome-wide survey of the secondary metabolite biosynthesis genes in the wheat pathogen *Parastagonospora nodorum*." Mycology **5**(3): 192-206.

Clarke, J. D. (2009). "Cetyltrimethyl ammonium bromide (CTAB) DNA miniprep for plant DNA isolation." Cold Spring Harbor protocols **2009**: pdb prot5177.

Cooley, R. N. and C. E. Caten (1991). "Variation in electrophoretic karyotype between strains of *Septoria nodorum*." Mol Gen Genet **228**(1-2): 17-23.

Crook, A., T. Friesen, Z. Liu, P. Ojiambo and C. Cowger (2012). "Novel necrotrophic effectors from *Stagonospora nodorum* and corresponding host sensitivities in winter wheat germplasm in the southeastern United States." Phytopathology **102**(5): 498-505.

Diament, B. J. and W. S. Noble (2011). "Faster SEQUEST searching for peptide identification from tandem mass spectra." Journal of proteome research **10**(9): 3871-3879.

Finn, R. D., A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm and J. Mistry (2013). "Pfam: the protein families database." Nucleic acids research: gkt1223.

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

Gao, Y., J. D. Faris, Z. Liu, Y. Kim, R. A. Syme, R. P. Oliver, S. S. Xu and T. L. Friesen (2015). "Identification and characterization of the SnTox6-Snn6 interaction in the *Parastagonospora nodorum*-wheat pathosystem." Molecular Plant-Microbe Interactions(ja).

Goto, N., P. Prins, M. Nakao, R. Bonnal, J. Aerts and T. Katayama (2010). "BioRuby: Bioinformatics software for the Ruby programming language." Bioinformatics (Oxford, England).

Gummer, J. P., R. D. Trengove, R. P. Oliver and P. S. Solomon (2012). "A comparative analysis of the heterotrimeric G-protein Gα, Gβ and Gγ subunits in the wheat pathogen *Stagonospora nodorum*." BMC microbiology **12**(1): 131.

Haibao Tang, B. P., Aurelien Naldi, Patrick Flick, Jeff Yunes, Kenta Sato, Chris Mungall (2015). "Goatools." from https://github.com/tanghaibao/goatools.

Hane, J. K., R. G. Lowe, P. S. Solomon, K.-C. Tan, C. L. Schoch, J. W. Spatafora, P. W. Crous, C. Kodira, B. W. Birren and J. E. Galagan (2007). "Dothideomycete–plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*." The Plant Cell Online **19**(11): 3347-3368.

Hane, J. K., R. G. Lowe, P. S. Solomon, K. C. Tan, C. L. Schoch, J. W. Spatafora, P. W. Crous, C. Kodira, B. W. Birren, J. E. Galagan, S. F. Torriani, B. A. McDonald and R. P. Oliver (2007). "Dothideomycete plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*." The Plant Cell **19**(11): 3347-3368.

Hane, J. K. and R. P. Oliver (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, J. K., T. Rouxel, B. J. Howlett, G. H. Kema, S. B. Goodwin and R. P. Oliver (2011). "A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi." Genome Biol **12**(5): R45.

Hastie, M. L., M. J. Headlam, N. B. Patel, A. A. Bukreyev, U. J. Buchholz, K. A. Dave, E. L. Norris, C. L. Wright, K. M. Spann and P. L. Collins (2012). "The human respiratory syncytial virus nonstructural protein 1 regulates type I and type II interferon pathways." Molecular & Cellular Proteomics **11**(5): 108-127.

Ipcho, S. V., J. K. Hane, E. A. Antoni, D. Ahren, B. Henrissat, T. L. Friesen, P. S. Solomon and R. P. Oliver (2012). "Transcriptome analysis of *Stagonospora nodorum*: gene models, effectors, metabolism and pantothenate dispensability." Molecular plant pathology **13**(6): 531-545.

IpCho, S. V., K.-C. Tan, G. Koh, J. Gummer, R. P. Oliver, R. D. Trengove and P. S. Solomon (2010). "The transcription factor StuA regulates central carbon metabolism, mycotoxin production, and effector gene expression in the wheat pathogen *Stagonospora nodorum*." Eukaryotic cell **9**(7): 1100-1108.

Ipcho, S. V. S., J. K. Hane, E. A. Antoni, D. Ahren, B. Henrissat, T. L. Friesen, P. S. Solomon and R. P. Oliver (2012). "Transcriptome analysis of Stagonospora nodorum: Gene models, effectors, metabolism and pantothenate dispensability." Molecular Plant Pathology **13**(6): 531-545.

Jin, Y., G. Binkowski, L. D. Simon and D. Norris (1997). "Ho endonuclease cleaves MAT DNA in vitro by an inefficient stoichiometric reaction mechanism." Journal of Biological Chemistry **272**(11): 7352-7359.

Jones, P., D. Binns, H.-Y. Chang, M. Fraser, W. Li, C. McAnulla, H. McWilliam, J. Maslen, A. Mitchell and G. Nuka (2014). "InterProScan 5: genome-scale protein function classification." Bioinformatics **30**(9): 1236-1240.

Käll, L., J. D. Canterbury, J. Weston, W. S. Noble and M. J. MacCoss (2007). "Semi-supervised learning for peptide identification from shotgun proteomics datasets." Nature methods **4**(11): 923-925.

Kim, D., G. Pertea, C. Trapnell, H. Pimentel, R. Kelley and S. L. Salzberg (2013). "TopHat2: accurate alignment of transcriptomes in the presence of insertions, deletions and gene fusions." Genome Biol **14**(4): R36.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu and S. L. Salzberg (2004). "Versatile and open software for comparing large genomes." Genome biology **5**(2): R12.

Lee, E., G. A. Helt, J. T. Reese, M. C. Munoz-Torres, C. P. Childers, R. M. Buels, L. Stein, I. H. Holmes, C. G. Elsik and S. E. Lewis (2013). "Web Apollo: a web-based genomic annotation editing platform." Genome Biol **14**: R93.

Lees, J. G., D. Lee, R. A. Studer, N. L. Dawson, I. Sillitoe, S. Das, C. Yeats, B. H. Dessailly, R. Rentzsch and C. A. Orengo (2014). "Gene3D: Multi-domain annotations for protein sequence and comparative genome analysis." Nucleic acids research **42**(D1): D240-D245.

Liu, Z., J. Faris, S. Meinhardt, S. Ali, J. Rasmussen and T. Friesen (2004). "Genetic and physical mapping of a gene conditioning sensitivity in wheat to a partially purified host-selective toxin produced by *Stagonospora nodorum*." Phytopathology **94**(10): 1056-1060.

Liu, Z., J. D. Faris, R. P. Oliver, K. C. Tan, P. S. Solomon, M. C. McDonald, B. A. McDonald, A. Nunez, S. Lu, J. B. Rasmussen and T. L. Friesen (2009). "SnTox3 acts in effector triggered susceptibility to induce disease on wheat carrying the Snn3 gene." PLoS Pathogens **5**(9): e1000581.

Liu, Z. H., J. D. Faris, S. W. Meinhardt, S. Ali, J. B. Rasmussen and T. L. Friesen (2004). "Genetic and physical mapping of a gene conditioning sensitivity in wheat to a partially purified host-selective toxin produced by *Stagonospora nodorum*." Phytopathology **94**(10): 1056-1060.

Lowe, R. G., M. Lord, K. Rybak, R. D. Trengove, R. P. Oliver and P. S. Solomon (2009). "Trehalose biosynthesis is involved in sporulation of *Stagonospora nodorum*." Fungal Genetics and Biology **46**(5): 381-389.

Manning, V. A., I. Pandelova, B. Dhillon, L. J. Wilhelm, S. B. Goodwin, A. M. Berlin, M. Figueroa, M. Freitag, J. K. Hane and B. Henrissat (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**(1): 41-63.

Martin, M. (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads." EMBnet. journal **17**(1): pp. 10-12.

Morrison, B. J., M. L. Hastie, Y. S. Grewal, Z. C. Bruce, C. Schmidt, B. A. Reynolds, J. J. Gorman and J. A. Lopez (2012). "Proteomic comparison of mcf-7 tumoursphere and monolayer cultures." PloS one **7**(12): e52692.

Murray, G. M. and J. P. Brennan (2009). "Estimating disease losses to the Australian wheat industry." Australasian Plant Pathology **38**: 558-570.

Ohm, R. A., N. Feau, B. Henrissat, C. L. Schoch, B. A. Horwitz, K. W. Barry, B. J. Condon, A. C. Copeland, B. Dhillon and F. Glaser (2012). "Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi." PLoS Pathogens **8**(12): e1003037.

Oliver, R. P., K. C. Tan and C. S. Moffat (In Press). "Necrotrophic Pathogens on Wheat " Encyclopedia of Food Grain Science.

Park, C. Y., A. A. Klammer, L. Käll, M. J. MacCoss and W. S. Noble (2008). "Rapid and accurate peptide identification from tandem mass spectra." Journal of proteome research **7**(7): 3022-3027.

Rouxel, T., J. Grandaubert, J. K. Hane, C. Hoede, A. P. van de Wouw, A. Couloux, V. Dominguez, V. Anthouard, P. Bally and S. Bourras (2011). "Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations." Nature Communications **2**: 202.

Shendure, J. and H. Ji (2008). "Next-generation DNA sequencing." Nature biotechnology **26**(10): 1135-1145.

Simakov, O., F. Marletaz, S.-J. Cho, E. Edsinger-Gonzales, P. Havlak, U. Hellsten, D.-H. Kuo, T. Larsson, J. Lv and D. Arendt (2013). "Insights into bilaterian evolution from three spiralian genomes." Nature **493**(7433): 526-531.

Skinner, M. E., A. V. Uzilov, L. D. Stein, C. J. Mungall and I. H. Holmes (2009). "JBrowse: a next-generation genome browser." Genome research **19**(9): 1630-1638.

Solomon, P., O. Waters, C. Jorgens, R. Lowe, J. Rechberger, R. Trengove and R. Oliver (2006). "Mannitol is required for asexual sporulation in the wheat pathogen *Stagonospora nodorum* (glume blotch)." Biochem. J **399**: 231-239.

Solomon, P. S., C. I. Jörgens and R. P. Oliver (2006). "δ-Aminolaevulinic acid synthesis is required for virulence of the wheat pathogen *Stagonospora nodorum*." Microbiology **152**(5): 1533-1538.

Solomon, P. S., R. C. Lee, T. Wilson and R. P. Oliver (2004). "Pathogenicity of *Stagonospora nodorum* requires malate synthase." Molecular microbiology **53**(4): 1065-1073.

Solomon, P. S., K. Rybak, R. D. Trengove and R. P. Oliver (2006). "Investigating the role of calcium/calmodulin-dependent protein kinases in *Stagonospora nodorum*." Molecular microbiology **62**(2): 367-381.

Solomon, P. S., K.-C. Tan and R. P. Oliver (2005). "Mannitol 1-phosphate metabolism is required for sporulation in planta of the wheat pathogen *Stagonospora nodorum*." Molecular Plant-Microbe Interactions **18**(2): 110-115.

Solomon, P. S., K.-C. Tan, P. Sanchez, R. M. Cooper and R. P. Oliver (2004). "The disruption of a Gα subunit sheds new light on the pathogenicity of Stagonospora nodorum on wheat." Molecular Plant-Microbe Interactions **17**(5): 456-466.

Solomon, P. S., K. C. Tan, P. Sanchez, R. M. Cooper and R. P. Oliver (2004). "The disruption of a Gα subunit sheds new light on the pathogenicity of *Stagonospora nodorum* on wheat." Molecular Plant-Microbe Interactions **17**(5): 456-466.

Solomon, P. S., O. D. Waters, J. Simmonds, R. M. Cooper and R. P. Oliver (2005). "The Mak2 MAP kinase signal transduction pathway is required for pathogenicity in *Stagonospora nodorum*." Current genetics **48**(1): 60-68.

Stergiopoulos, I., J. Collemare, R. Mehrabi and P. J. De Wit (2013). "Phytotoxic secondary metabolites and peptides produced by plant pathogenic Dothideomycete fungi." FEMS microbiology reviews **37**(1): 67-93.

Syme, R. A., J. K. Hane, T. L. Friesen and R. P. Oliver (2013). "Resequencing and comparative genomics of *Stagonospora nodorum*: Sectional gene absence and effector discovery." G3: Genes| Genomes| Genetics **3**(6): 959-969.

Tan, K.-C., J. L. Heazlewood, A. H. Millar, G. Thomson, R. P. Oliver and P. S. Solomon (2008). "A signaling-regulated, short-chain dehydrogenase of *Stagonospora nodorum* regulates asexual development." Eukaryotic cell **7**(11): 1916-1929.

Tan, K.-C., O. D. Waters, K. Rybak, E. Antoni, E. Furuki and R. P. Oliver (2014). "Sensitivity to three *Parastagonospora nodorum* necrotrophic effectors in current Australian wheat cultivars and the presence of further fungal effectors." Crop and Pasture Science **65**(2): 150-158.

Tan, K. C., J. L. Heazlewood, A. H. Millar, G. Thomson, R. P. Oliver and P. S. Solomon (2008). "A signaling-regulated, short-chain dehydrogenase of Stagonospora nodorum regulates asexual development." Eukaryotic Cell **7**(11): 1916-1929.

Tan, K. C., R. P. Oliver, P. S. Solomon and C. S. Moffat (2010). "Proteinaceous necrotrophic effectors in fungal virulence." Functional Plant Biology **37**: 907-912.

Tan, K. C., R. D. Trengove, G. L. Maker, R. P. Oliver and P. S. Solomon (2009). "Metabolite profiling identifies the mycotoxin alternariol in the pathogen *Stagonospora nodorum*." Metabolomics: 1-6.

Tange, O. (2011). "Gnu parallel-the command-line power tool." The USENIX Magazine **36**(1): 42-47.

Testa, A. C., J. K. Hane, S. R. Ellwood and R. P. Oliver (2015). "CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts." BMC genomics **16**(1): 170.

Urban, M., R. Pant, A. Raghunath, A. G. Irvine, H. Pedro and K. E. Hammond-Kosack (2014). "The Pathogen-Host Interactions database (PHI-base): additions and future developments." Nucleic acids research: gku1165.

Vincent, D., K. C. Tan, L. Cassidy, P. S. Solomon and R. P. Oliver (2012). "Proteomic techniques for plant-fungal interactions." Methods In Molecular Biology **835**: 75-96.

Vleeshouwers, V. G. and R. P. Oliver (2014). "Effectors as tools in disease resistance breeding against biotrophic, hemibiotrophic, and necrotrophic plant pathogens." Molecular Plant-Microbe Interactions **27**(3): 196-206.

Yin, Y., X. Mao, J. Yang, X. Chen, F. Mao and Y. Xu (2012). "dbCAN: a web resource for automated carbohydrate-active enzyme annotation." Nucleic acids research **40**(W1): W445-W451.

Zhang, J., R. Chiodini, A. Badr and G. Zhang (2011). "The impact of next-generation sequencing on genomics." Journal of Genetics and Genomics **38**(3): 95-109.

Chapter 6 | Pan-*Phaeosphaeria* Comparative Genome Analysis – effector prediction and genome evolution

Attribution Statement

Authors: **Robert A. Syme**, James K Hane, Megan C. McDonald, Kar-Chun Tan, Kasia Clarke, Bruce A. McDonald, Richard P. Oliver

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

The Ph.D. candidate (Robert A. Syme) made the following contributions to this chapter:

- DNA purification and library preparation for all samples except WAC8410
- Genome assembly and annotation of all strains except SN15
- All data analysis
- Generation of all figures and tables
- Manuscript authorship with editing from RPO, JKH, and KCT

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

Signature: ……………………………………………

Date: 2015-05-19

I, Richard Oliver, certify that this attribution statement is an accurate record of Robert Syme's contribution to the research presented in this chapter.

Signature: ……………………………………………

Date: 2015-05-20

Abstract

Important questions about fungal pathogenicity and genome evolution are answerable by comparative genomics within and between species. Here we expand the genomic catalogue of *Parastagonospora nodorum* by the sequencing and assembly of 22 strains and *P. avenae* with 12 strains. Genomic comparisons within and between the two species reveal accessory elements under strong positive selection in *P. nodorum*. The *Phaeosphaeria* pan-genome is found to include an expansion in paralogs of the necrotrophic effector SnTox3, including in *P. avenae* group Pat5 but not in other *P. avenae* groups. Regions with an elevated density of genes under positive selection were observed in *P. nodorum* adjacent to repetitive sequences. Finally, presence/absence variation across all strains were combined with measures of positive selection to update the list of putative *P. nodorum* effectors.

Introduction

The *Parastagonospora* pathosystems

Parastagonospora (teleomorph: *Phaeosphaeria*) *nodorum* (Berk.) is an economically important necrotrophic fungal pathogen that causes septoria nodorum blotch (SNB) in wheat (*Triticum aestivum*) (Solomon, Lowe et al. 2006) and also a model organism for the fungal order Pleosporales and for necrotrophic phytopathogenicity (Oliver and Solomon 2010, Tan, Oliver et al. 2010, Oliver, Friesen et al. 2012). Significant experimental resources are available for *P. nodorum*, including a high-quality genome assembly of the reference strain SN15 (Hane, Lowe et al. 2007, Syme, Tan et al. 2016), microarray analyses of gene expression (Ipcho, Hane et al. 2012), proteomics and proteogenomics, metabolomic profiling, and genome resequencing of two contrasting *P. nodorum* strains: Sn4 and Sn79-1087 (Syme, Hane et al. 2013). SN15 and Sn4 are highly aggressive isolates on wheat whereas Sn79-1087, hereafter referred to as Sn79, was isolated from *Agropyron,* is unable to establish SNB on wheat, and has served as a useful negative control for comparative genomics in a disease context (Friesen, Stukenbrock et al. 2006). Here we extend these resources to include several additional isolates of *P. nodorum* as well as closely species within the *Parastagonospora* genus. *Parastagonospora avenae* (teleomorph: *Phaeosphaeria avenaria*) is a species associated with SNB-like symptoms in various Poaceae (Quaedvlieg, Verkley et al. 2013). *P. avenae* was further divided into two *formae speciales*; *P. avenaria* f. sp. *avenaria* (*Paa*) infects oats (*Avena* spp.) and *P. avenaria* f. sp. *triticea* (*Pat*) infects wheat and some other grasses (Cunfer 2000, Ueng, Dai et al. 2003, Malkus, Reszka et al. 2005). Restriction fragment length polymorphism patterns were used to further differentiate Pat strains into distinct subgroups named Pat1-Pat6 (Ueng and Chen 1994, Ueng, Cunfer et al. 1995, Ueng, Subramaniam et al. 1998).

Many of the recent studies of *Parastagonospora* spp. are oriented around an elusive class of genes that encode effector molecules that interact with the host to determine the outcome of specific host-pathogen interactions (Vleeshouwers and Oliver 2014). A class of effectors called necrotrophic effectors (NEs) interacts with proteins encoded by dominant host susceptibility genes to form an inverse gene-for-gene interaction (Oliver and Solomon 2010, Tan, Oliver et al. 2010) in which the presence of both partners leads to a compatible interaction characterized by plant cell necrosis. *Parastagonospora* spp. were already shown to produce several necrotrophic effectors (Friesen, Meinhardt et al. 2007) that are thought to maximise the likelihood of interacting with a corresponding susceptibility protein in the host. So far, three well characterised NEs have been identified in *P. nodorum*: *SnToxA* (Friesen, Stukenbrock et al. 2006), *SnTox1* (Liu, Faris et al. 2004, Liu, Zhang et al. 2012), and *SnTox3* (Friesen, Zhang et al. 2008, Liu, Faris et al. 2009). These three effectors are already used in breeding programs to accelerate development of disease-resistant cereal cultivars (Vleeshouwers and Oliver 2014), but several NEs have not yet been identified to the gene level, including *SnTox2* (Friesen, Meinhardt et al. 2007), *SnTox4* (Abeysekara, Friesen et al. 2009), *SnTox5* (Chu, Xu et al. 2012), *SnTox6* (Gao, Faris et al. 2015), and *SnTox7* (Shi, Friesen et al. 2015). While continued laboratory testing may yield new effectors, continual advances in genome sequencing technologies and bioinformatics methods may also improve effector discovery. This study provides enhancements to the *P. nodorum* SN15 reference strain assembly and its gene annotations, but also explores features of the *Parastagonospora* fungal genomes that are relevant for effector discovery, including repeat-induced point mutation (RIP), mesosynteny, presence-absence variation, and diversifying selection.

Repeat-induced point mutation (RIP) is a fungal-specific form of mutation that targets repetitive sequences and introduces cytosine to thymine (C→T) transitions, or the reverse complement G→A. In the filamentous Ascomycota (syn. Pezizomycotina) where RIP is observed, there is a strong bias for mutations at cytosine bases adjacent to adenine (CpA → TpA). It was suggested that RIP provides a mechanism of genome defence against transposon invasion, by disabling transposable elements (TEs) through introduction of premature stop codons into their open reading frames and/or through silencing of the RIP-mutated sequence through further DNA methylation (Galagan and Selker 2004, Hane and Oliver 2008, Clutterbuck 2011, Hane 2015, Hane, Williams et al. 2015).

RIP is common in many Pezizomycotina species (Testa 2016), and has been linked to effector mutation in *Leptosphaeria maculans* (Fudal, Ross et al. 2009, Van De Wouw and Howlett 2011). Although RIP primarily targets repetitive DNA, RIP mutations may also encroach into nonrepetitive flanking regions (Fudal, Ross et al. 2009, Van de Wouw, Cozijnsen et al. 2010). Genes that are proximal to repeats may become extensively mutated during successive rounds of RIP, generating variants encoding products that are no longer recognised by host-defences. Pathosystems in which host-recognition of a pathogen is controlled by an avirulence (AVR) gene in the pathogen and corresponding resistance (R) gene in the plant are said to conform to the gene-for-gene model of host-pathogen interaction (Flor 1971). In gene-for-gene systems, alteration, inactivation or loss of an AVR gene can benefit the pathogen by allowing it to escape avirulence-mediated host recognition. For example, the avirulence genes *AvrLm1*, *AvrLm6*, and *AvrLm4-7* in the *Leptosphaeria maculans* genome are located in close proximity to highly repetitive regions. RIP-like polymorphisms have been observed in these loci for many isolates (van de Wouw et al., 2010) and a gradient of RIP-like mutations was observed to be inversely correlated with distance from the repetitive sequences. The fungal genes most likely to be affected by RIP leakage can be identified by annotation of AT-rich regions with the OcculterCut tool (Testa, 2016).

In early chromosomal mapping studies, the term synteny was defined to describe two loci known to exist on the same chromosome but not yet demonstrated to be genetically linked by recombination frequency (Renwick, 1971). This semantic distinction was necessary because of the technical difficulty of establishing linkage, particularly for larger plant chromosomes (Novitski and Blixt, 1978). As genome sequencing became commonplace and the genetic maps of individual chromosomes became more densely populated, the term synteny came to describe stretches of collinear loci where gene order and orientation is preserved. In a panmictic population prior to a speciation event, any given chromosome will likely share gene order and orientation, preserving both synteny and co-linearity. After speciation, sister species undergoing independent evolutionary processes can diverge, with gene gain/loss/duplication or chromosomal rearrangements slowly degrading the degree of synteny and co-linearity among the sister species unless selection favors retention of synteny.

A number of different classes of synteny have been proposed. Co-linearity of thousands of genes, observable at the whole-chromosomal scale is often described as macrosynteny. Comparisons of chromosomes between related plant and animal species typically detect a macrosyntenic relationship over comparatively large evolutionary distance. The genomes of cereal, legume or vertebrate species often display conserved macrosynteny (Pennacchio 2003, Cannon, Sterck et al. 2006). Even in the absence of macrosynteny, the linear arrangement of small groups of genes (as few as 2-5, called microsynteny) can be preserved across large phylogenetic distances (Chen, SanMiguel et al. 1997). The evolutionary pressure to preserve the order of these loci can be attributed, in at least some cases, to blocks of coregulated genes, such as those involved in production of secondary metabolites (Engström, Sui et al. 2007, McGary, Slot et al. 2013).

The first fungal species to be sequenced displayed no macrosynteny and remarkably little microsynteny among species in different families and in some cases even at the genus level (Goffeau, Barrell et al. 1996, Galagan, Henn et al. 2005). It was only after a significant number of Pezizomycotina species were sequenced that a syntenic pattern emerged. Comparisons of four species within the Dothideomycetes using a dot-plot approach identified a novel form of synteny that Hane et al (2011) termed mesosynteny. The pattern is characterised by frequent shuffling of gene content within a given chromosome, but infrequent translocations among chromosomes. Observations of mesosynteny have been confined mainly to the Pezizomycotina (Ohm, Feau et al. 2012), though it appears that similar structural rearrangements may occur in the *Agaricomycetes* (Hane, Rouxel et al. 2011, Hane, Anderson et al. 2014). The mechanisms underlying mesosynteny remain unknown, but simulations of repeated rounds of intra-chromosomal inversions produce a recognisably mesosyntenic pattern (Ohm, Feau et al. 2012). Two genome features were proposed to be associated with mesosyntenic breakpoints: simple sequence repeats (SSRs) (Ohm, Feau et al. 2012) and transposable elements (Grandaubert, Schoch et al. 2013). Observations of mesosyntenic patterns are limited to the filamentous Ascomycetes, and most prominent in the Pezizomycotina, Dothideomycetes, and possibly Agaricomycetes - taxon that include many important plant pathogens. Despite contributing to the plasticity of fungal genomes, the effect of mesosynteny on plant pathogenicity is still unknown.

Recent resequencing studies have repeatedly shown that variation in gene content, or presence/absence variation (PAV) genotypes are more common in fungi than was previously thought (McDonald, Oliver et al. 2013, Gao, Faris et al. 2015, Golicz, Martinez et al. 2015). PAV have been observed at two distinct scales: at the gene cluster level (Plissonneau, Stürchler et al. 2016) and at whole (or partial) chromosome level (Ma, Van Der Does et al. 2010, Goodwin, M'Barek et al. 2011). These PAV patterns are particularly relevant in the context of plant pathogenicity because they may highlight variable effector loci when applied across multiple isolates of a single species with a range of virulence phenotypes. Sectional absences of small groups of genes were previously observed in comparisons between the reference strain SN15 and alternate strains of *P. nodorum* (Syme, Hane et al. 2013). The effectors *SnTox1* and *SnTox3* are notably absent from the wheat-avirulent strain SN79 in small 2 kb and 4 kb stretches respectively, whereas *SnToxA* is part of a much larger 72 kb absence in Sn79 (Syme, Hane et al. 2013). There is not a consistent pattern to the size of pathogenicity-related PAV. For each known effector, the pattern of PAV genotypes in field populations of *P. nodorum* and *P. avenaria* varies, which may indicate multiple, independent horizontal gene transfer (HGT) events (McDonald, Oliver et al. 2013). Notably, there does not appear to be a significant fitness penalty incurred by the pathogen harbouring an effector when growing on a host that lacks the corresponding sensitivity gene (McDonald, Oliver et al. 2013). It is also possible that genes residing in genomic regions rich in repetitive DNA may be more prone to loss due to a higher frequency of mesosyntenic recombination and subsequent increased likelihood of imperfect matching between homologous regions of sister chromatids.

The genomes of some pathogenic species are compartmentalised into regions with "two speeds" of evolution: "core" gene content - which tends to be under purifying selection pressures and well-conserved - and variable gene content. For some species, such as *L. maculans*, these variable regions are interspersed throughout the genome. In others, variable and core gene contents are divided among separate chromosomes. Fungal accessory (syn. supernumerary, dispensable) chromosomes have been observed to be gene sparse (Grandaubert, Bhattacharyya et al. 2015) and exhibit higher rates of mutation, transposable elements (Coleman, Rounsley et al. 2009), positive selection and pathogenicity-associated loci relative to 'core' chromosomes (Miao, Covert et al. 1991, Coleman, Rounsley et al. 2009, Balesdent, Fudal et al. 2013). For example, the accessory chromosomes of *Nectria haematococca* (Coleman, Rounsley et al. 2009) encode genes responsible for antibiotic resistance (Miao, Covert et al. 1991) and pathogenicity (VanEtten, Straney et al. 2001). Genes on accessory chromosomes in *Zymoseptoria tritici* have a much higher rate of non-synonymous substitution and lower rates of synonymous substitution, indicating that their loci are under positive selection (Stukenbrock, Jørgensen et al. 2010, Croll, Lendenmann et al. 2015). Accessory chromosomes in *Fusarium* spp. are also rich in transposable elements and pathogenicity genes. The transfer of two whole accessory chromosomes of *F. oxysporum* was demonstrated to be sufficient to convert a non-pathogenic strain into a pathogen (Ma, Van Der Does et al. 2010). *P. nodorum* scaffold 46 was previously predicted to be dispensable on the grounds that it shares characteristics with *Z. tritici* accessory chromosomes, including a low GC content, high repeat content, low gene density and a low percentage of predicted genes with conserved functional domains (Ohm, Feau et al. 2012).

Recent advances across various aspects of fungal genome evolution and effector gene prediction provide new opportunities to study the intra-species diversity of fungal genomes in the context of pathogenicity. Previous studies of mesosynteny focused on inter-species comparisons with no analysis of genomic structural variation within species. Similarly, previous studies of RIP were restricted to repetitive sequences in a single genome or inter-species comparisons (Clutterbuck 2011, Hane, Anderson et al. 2014). Here we investigate RIP, mesosynteny, and presence-absence variation across several isolates of *P. nodorum* and the closely related species *P. avenaria*. The high resolution data generated by deep re-sequencing highlights idiosyncrasies of each of these variable features in model Pleosporales pathogens and provides novel insights that can be translated to the study of related fungal pathogen species.

Methods

Strain Sampling, DNA extraction and Sequencing

Illumina paired-end libraries were constructed for each strain. *P. nodorum* WAC8410 was sequenced from a TruSeq 500 bp library on an Illumina 2000 to produce 150 bp paired-end reads. All other strains were sequenced from 300 bp NexteraXT libraries on an Illumina 2500 multiplexed over two lanes.

Reference-alignment and *de novo* assembly of alternate strain sequences

Raw Illumina reads were trimmed using cutadapt v1.7.1 (Martin 2011), removing adapter sequence 'CTGTCTCTTATACACATCTCCGAGCCCACGAGAC', removing bases with quality score less than 25, and removing any reads shorter than 50 bp after trimming. PCR duplicate reads were identified using Picard tools (Simakov, Marletaz et al. 2013)(Simakov, Marletaz et al. 2013)(Simakov, Marletaz et al. 2013)(Simakov, Marletaz et al. 2013)(Simakov, Marletaz et al. 2013)(Simakov, Marletaz et al. 2013).

The reference genome used for alignments and other comparative analysis in this study was derived from *P. nodorum* strain SN15 (Hane, Lowe et al. 2007). The latest SN15 genome assembly and gene annotations, which have been curated to a high standard, were sourced from https://github.com/robsyme/Parastagonospora_nodorum_SN15. To further inform the genomic context of variation observed across alternate isolates relative to SN15, repetitive and low-complexity DNA regions were also re-predicted for the updated SN15 assembly de novo as per Hane (Hane, Lowe et al. 2007, Hane and Oliver 2008, Bringans, Hane et al. 2009).

Trimmed reads were aligned to the reference genome using bowtie2 (Langmead and Salzberg 2012) using the "very-sensitive" flag and retaining unaligned reads for later use. Read alignments were sorted and compressed using samtools (Li, Handsaker et al. 2009). Coverage statistics used to inform variant calling parameters were calculated using Picard Tools v1.128 `CollectWgsMetrics` module (Simakov, Marletaz et al. 2013)(Simakov, Marletaz et al. 2013)(Simakov, Marletaz et al. 2013)(Simakov, Marletaz et al. 2013)(Simakov, Marletaz et al. 2013)(Simakov, Marletaz et al. 2013).

Trimmed reads were assembled *de novo* with the SPAdes assembler v.3.5.0 (Bankevich, Nurk et al. 2012) at kmer lengths 21, 33, 55, and 77 using mismatch and short indel correction via BWA mapping (Bringans, Hane et al. 2009).

Assembly quality was assessed against the reference genome and annotations using QUAST v2.3 (Gurevich, Saveliev et al. 2013). Heavily fragmented genomes with fewer than 8000 reference genes present were flagged for exclusion from annotation-based measures of variation due to the likelihood of annotation truncation.

Analysis of sequence variation

Sequence polymorphism across alternate *P. nodorum* and *P. avenaria* strains relative to the SN15 reference genome, were predicted using bowtie2 alignments with the GATK v3.3-0 HaplotypeCaller (DePristo, Banks et al. 2011) using the best practices recommendations (Auwera, Carneiro et al. 2013). For strains with a mean coverage below 5x, the `minReadsPerAlignmentStar` cut-off was reduced to 3 from the default of 5. Variants identified in each strain were pooled and genotyped using GATK `GenotypeGVCF` (DePristo, Banks et al. 2011). However, genome assemblies of the alternate strains were also aligned to the reference genome using nucmer (Kurtz, Phillippy et al. 2004) with default parameters and filtered using delta-filter parameters `-r` and `-q` to exclude alignments to and of repetitive sequences. Subsequently, SNP and indel variants from the genome alignments were extracted using custom scripts (https://github.com/robsyme/bioruby-mummer) to capture additional variation in regions where the alternate strain and reference are too dissimilar to reliably map short reads, preventing GATK variant calling.

Fungi of the Pezizomycotina sub-phylum taxon, to which *Parastagonospora* belongs, typically exhibit a repeat-targeted mutation mechanism with a bias towards CpA dinucleotides that is known as repeat-induced point mutation or RIP. All SNPs generated from the variant calling methods above were classified by RIP class. Sites with C → T SNPs were identified and classified according to the adjacent nucleotide (CpA, CpC, CpG, or CpT). Sites were also classified by mutation direction where ref_C → alt_T suggests the alternate strain is RIP-mutated and ref_T → alt_C suggests the reference is RIP-mutated. The same procedure was applied to SNPs on the reverse strand at G → A SNPs. CpA → TpA and GpT → ApT SNPs were annotated as 'RIP-like'. To determine the extent of RIP across the *P. nodorum* genome, the SN15 reference genome sequence was divided into windows of 5 kbp. Within these windows, we calculated the number of RIP-like alt_CpA → ref_TpA SNPs as a fraction of the number of all SNPs in the window.

Genome Annotation

Manually-curated annotations from the reference genome SN15 were used to train CodingQuarry (Testa, Hane et al. 2015) and these parameters applied to each of the alternate strains. A database of repeats was generated using RepeatModeler v1.0.8 (Smit and Hubley 2010) augmented with full copies of known *P. nodorum* repeats Molly (AJ488502.1), Pixie (AJ488503.1), and Elsa (AJ277966.1). The RepeatModeler repeats were combined with known "DeRIPped" (predicted pre-RIP consensus) (Hane and Oliver 2010) and repbase (Jurka, Kapitonov et al. 2005) repeats using RepeatMasker v4.0.5 (Smit, Hubley et al. 1996-2004). A final set of annotations for each alternate strain was generated using Maker v2.31.8 (Cantarel, Korf et al. 2008) which was provided the reference proteome for tblastn hits, the *de-novo* repeat database (Hane and Oliver 2010) and CodingQuarry (Testa, Hane et al. 2015) predictions. Secretion signals were detected using SignalP v4.1 (Bendtsen, Nielsen et al. 2004) and trans-membrane domains by TMHMM v2.0 (Krogh, Larsson et al. 2001). Secondary metabolite clusters were predicted in non-reference strains by antismash 2.1.1 (Medema, Blin et al. 2011).

Clusters of putatively homologous proteins were identified using ProteinOrtho v5.11 (Lechner, Findeiß et al. 2011) using the synteny flag and a blastp e-value of 1×10^{-5}. Protein cluster membership for SnToxA, SnTox1 and SnTox3 were checked by using each protein sequence to query each of the alternate strain *de-novo* assemblies using tblastn (e-value cut-off 1×10^{-5}).

Scaffold and Gene Presence/Absence

Assemblies from each alternate strain were compared to the reference assembly with nucmer using the default parameters. The coverage of each of the reference scaffolds by nucmer matches to the was calculated using the `genomecov` function of BEDtools (Quinlan and Hall 2010). Homologs of known effectors were extracted from orthologous clusters calculated earlier. Each effector absence was manually confirmed by blasting the reference effector against the alternate strain's genome assembly. A tree of nucleotide matches to SnTox3 was constructed using MrBayes using the default parameters.

Phylogeny

SNP and indel variants identified from alternate strains were applied to 35 reference loci where the locus was present in all strains. Each cluster of loci was aligned using ClustalW v2.1 (Larkin, Blackshields et al. 2007) and clusters were concatenated to create a 30,992 bp alignment with 19,501 (62.9%) identical sites and 93.1% pairwise identity. A phylogenetic tree was calculated using MrBayes (Ronquist and Huelsenbeck 2003) with the JC69 substitution model, popinv rate variation, chain length 200,000, subsampling frequency 200, heated chains 4, burn-in length 10,000, and heated chain temp 0.2. SnTox3 paralogs were detected by searching for the SN15 SNOG_08981 protein sequence in all the available genomes using tblastn. Only hits with e-value less than 1×10^{-10} were retained.

Identification of Positive Selection

The ratio of non-synonymous substitutions to synonymous substitutions (ω) can indicate the presence of diversifying/positive or purifying/negative selection. Coding sequence for each orthologous cluster containing a protein from the reference strain were extracted, translated and aligned using ClustalW (Larkin, Blackshields et al. 2007). Protein truncations due to incorrect annotation in the alternate strains limit detection of positive selection as only codons present in all strains can be considered. Short proteins with lengths more than 1 standard deviation from the mean were excluded from the alignments.

Codon alignments of the nucleotide sequence were generated with the pal2nal v14 (Suyama, Torrents et al. 2006). The M1a and M2a site models were applied to each orthologous cluster to generate a maximum likelihood (ML) estimation of ω. The H_0 model (PAML model M1a) confines codon membership to one of two classes where $\omega < 1$ (purifying selection) or $\omega = 1$ (neutral/drift). The H_1 model (PAML model M2a) extends H_0 to allow codons membership to a third possible class where $\omega > 1$ (positive selection). Loci with sites under positive selection were identified where the χ^2-distributed likelihood ratio of the two models exceeded the 1% significance level (2 degrees of freedom).

Regional patterns of selection pressure were identified by stepping a 100 kbp window over the reference assembly in 1 kbp increments, counting the number of transcripts under positive selection as a percentage of the total number of transcripts in each window.

Effector Candidate Criteria

For each reference transcript sequence, a number of tests were applied and a score assigned to transcripts that passed. Proteins were scored that had a molecular mass less than 30 kDa, had Cys greater than 4%, were less than 5 kb from repetitive sequences longer than 200 bp, did not have tblastn hits to the SN79 genome assembly (e-value less than 1×10^{-30}), were in regions of low density (at most 1 other gene in the 2 kb up and downstream of the transcript), were predicted to be secreted by SignalP v4.1 (Bendtsen, Nielsen et al. 2004), were predicted to be under diversifying selection pressure as described above, were not part of the core proteome, or were not predicted to encode a trans-membrane domain by TMHMM (Krogh, Larsson et al. 2001).

Structural Variation

The locations of repetitive sequences calculated earlier were divided into simple repeats and transposable elements (TEs), sub-divided into class I retroelements and class II DNA transposons. TE classes were further sub-classified for selected notable repeat families. Bowtie2 mapping of alternate strains' reads to the reference genome were used as input to detect structural variation using Delly v0.6.1 (Rausch, Zichner et al. 2012) with default parameters. Sites of translocation and inversion events were extracted and the relative distance form each breakpoint to the nearest repetitive sequences were calculated using the `reldist` function in BEDtools v2.22.1 (Quinlan and Hall 2010, Favorov, Mularoni et al. 2012). The relative distance function is a measure of spatial correlation between two sets of intervals, where an association between the sets A and B would present as a high proportion of B sites at low relative distance to their nearest A site and a low proportion of B sites equidistant from A sites (a relative distance of 0.5). Breakpoint sites were clustered and counted together for the tally when breakpoints from two or more alternate strains were predicted within 1 kb.

Results

Sequencing and Assembly

The estimated genome coverage for each strain, relative to the SN15 reference assembly, ranged between 5x and 64x for strains sequenced using the NexteraXT libraries and 81x for the WAC8410 isolate sequenced using a TruSeq library (Table 1). Strains TN5-1, SC3-1, SnSA95.103 and NOR-4 produced fragmented assemblies, with fewer than 10,000 SN15 reference genes present in the assembly as determined by QUAST analysis (Gurevich, Saveliev et al. 2013) (Appendix A6-1).

Table 1: Sequencing and isolate summary. Isolate IDs are identifiers used in the laboratories in which the strains were first isolated. Some collection dates are not available, shown as '-'.

Isolate ID	*Isolate Source*	*Collection Year*	*Sum length of post-QC reads (Mb)*	*Estimated Coverage*	*Species and forma specialis*
***P. nodorum* strains**					
B2.1b	Iran	2005	686.1	9.2	*P. nodorum*
C1.2a	Iran	2005	886.1	11.9	*P. nodorum*
IR10_9.1a	Iran	2010	528.4	7.1	*P. nodorum*
IR10_2.1a	Iran	2010	665.9	8.9	*P. nodorum*
Sn Cp2052	Denmark	-	1,031.9	13.9	*P. nodorum*
FIN-2	Finland	-	1,998.7	26.9	*P. nodorum*
NOR-4	Norway	-	871.7	11.7	*P. nodorum*
SWE-3	Sweden	-	1,517.8	20.4	*P. nodorum*
BRSn9870	Brazil	-	3,316.9	44.6	*P. nodorum*
Sn99CH 1A7a	Switzerland	-	3,888.2	52.3	*P. nodorum*
SnChi01 40a	China	-	2,796.4	37.6	*P. nodorum*
SnSA95.103	South Africa	-	1,019.0	13.7	*P. nodorum*
SnOre11-1	Oregon, USA	-	4,763.4	64.1	*P. nodorum*
OH03 Sn-1501	Ohio, USA	-	2,229.8	30.0	*P. nodorum*
SNOV92X D1.3	Texas, USA	-	2,524.6	34.0	*P. nodorum*
AR1-1	Arkansas, USA	-	1,986.2	26.7	*P. nodorum*
TN 5-1	Tennessee, USA	-	987.4	13.3	*P. nodorum*
VA 5-2	Virginia, USA	-	2,288.3	30.8	*P. nodorum*
GA9-1	Georgia, USA	-	4,296.5	57.8	*P. nodorum*
MD4-1	Maryland, USA	-	4,325.8	58.2	*P. nodorum*
SC 3-1	South Carolina, USA	-	374.4	5.0	*P. nodorum*
WAC8410	Australia	2010	6,032.4	81.2	*P. nodorum*
***P. avenaria* strains**					
IR10_5.2b	Iran	2010	850.5	11.4	*P. avenaria* f. *sp. tritici 1*
SN11IR_2_1.1	Iran	2010	1,415.3	19.0	*P. avenaria* f. *sp. tritici 4*
82-4841	North Dakota, USA	1982	1,207.7	16.2	*P. avenaria* f. *sp. tritici 5*
83-6011-2	North Dakota, USA	1983	1,136.3	15.3	*P. avenaria* f. *sp. tritici 5*
SN11IR_6_1.1	Iran	2010	2,150.2	28.9	*P. avenaria* f. *sp. tritici 6*
SN11IR_7_2.3	Iran	2010	1,039.4	14.0	*P. avenaria* f. *sp. tritici 6*
Mt. Baker	Washington, USA	2009	585.8	7.9	*P. avenaria*
s258	Netherlands	2005	1,163.2	15.6	*P. avenaria*
H6.2b	Iran	2005	3,030.0	40.8	*P. avenaria* f. *sp. tritici 2*
A1 3.1a	Iran	2005	1,813.1	24.4	*P. avenaria* f. *sp. tritici 2*
Hartney99	Canada	2005	1,132.5	15.2	*P. avenaria* f. *sp. tritici 1*
Jansen #4_55	Canada	2005	480.1	6.5	*P. avenaria* f. *sp. tritici 1*

Table 2: *De-novo* assembly summary statistics of newly sequenced strains. N50 is the smallest number of scaffolds that make up half of the total assembly length. QUAST gene presence counts indicate the number of reference (strain SN15) genes that are covered by nucmer matches. The QUAST gene counts are divided into genes covered completely and genes only partially covered. High sequence variation between SN15 and *P. avenaria* frustrated the nucmer matching, resulting in very low gene counts in those strains. Some genomes such as NOR-4 and SC3-1 and SnSA95.103 have more fragmented genomes and as such were excluded from some downstream analyses.

Isolate ID	*№ Scaffolds*	*Largest scaffold (kb)*	*Total length (Mb)*	*N50 (kb)*	*Whole SN15 gene count by QUAST*	*Partial SN15 gene count by QUAST*
***P. nodorum* strains**						
B2.1b	2906	386.4	37.38	60.3	12628	784
C1.2a	1557	325.3	37.43	80.4	12751	649
IR10_9.1a	3673	140.6	37.28	20.7	11442	1920
IR10_2.1a	2131	229.7	37.50	44.8	12412	997
Sn Cp2052	3026	190.2	37.32	38.2	12310	1148
FIN-2	1381	451.6	38.41	117.9	12952	479
NOR-4	17618	34.2	28.72	1.8	2631	9512
SWE-3	1714	335.5	37.85	58.8	12734	727
BRSn9870	4911	309.9	41.23	52.4	12342	1148
Sn99CH 1A7a	853	521.0	37.90	180.7	13062	361
SnChi01 40a	779	1,268.1	37.88	206.3	13118	332
SnSA95.103	11772	48.0	49.94	6.5	9545	3882
SnOre11-1	748	875.4	37.42	249.6	13103	352
OH03 Sn-	1281	349.9	37.10	85.0	12844	577
SNOV92X	785	524.1	36.66	145.6	13009	442
AR1-1	882	748.5	36.61	135.6	12999	412
TN 5-1	15004	22.0	37.55	3.6	7030	6323
VA 5-2	1115	485.6	36.48	89.1	12787	606
GA9-1	664	938.6	36.53	215.3	13074	356
MD4-1	701	599.9	36.53	191.0	13045	384
SC 3-1	16752	44.2	28.05	2.0	3143	9256
WAC8410	384	1,060.5	40.27	316.8	13223	277
***P. avenaria* strains**						
IR10_5.2b	1681	267.0	35.51	57.3	18	14
SN11IR_2_1.1	5762	168.2	41.54	38.6	12	15
82-4841	2444	218.3	38.53	50.5	21	13
83-6011-2	2367	193.8	37.52	43.9	21	13
SN11IR_6_1.1	1174	737.1	33.51	102.8	3	7
SN11IR_7_2.3	2215	244.5	33.60	44.8	6	12
Mt. Baker	8309	38.5	34.14	6.2	15	81
s258	4090	183.1	39.49	32.8	16	21
H6.2b	1613	411.9	38.68	74.6	2	6
A1 3.1a	1764	419.9	39.05	68.3	3	6
Hartney99	3381	124.3	36.58	27.6	47	20
Jansen 4_55	10109	83.1	32.06	4.3	20	166

Phylogeny and Distribution of Known Effectors

Strains were sampled from broad geographic collections with a focus on the Fertile Crescent and the United States (Table 1, Figure 1) and included three previously sequenced *P. nodorum* strains, 22 newly sequenced *P. nodorum* strains and 12 newly sequenced *P. avenaria* strains. The *P*. nodorum clade is distinct from the *P. avenaria* species and each Pat group clusters together in the phylogenetic tree (Figure 1). The Pat2 group composed of Iranian isolates H6.2b and A1 3.1a are very similar to each other, but show very high levels of sequence variation relative to the other *P. avenaria* strains (Figure 1).

The SnToxA effector was detected in 12/25 *P. nodorum* strains and was absent from all *P. avenaria* strains. SnTox1 was detected in 18/25 *P. nodorum* strains and 2/12 *P. avenaria* strains in Pat1 and Pat5. SnTox3 was detected in 16/25 *P. nodorum* strains and 2/12 *P. avenaria* strains (17%), both in Pat5 strains (Figure 1, Figure 2). The pattern of effector presence/absence is not concordant with phylogeny, which is a common for effector alleles (McDonald, Oliver et al. 2013).

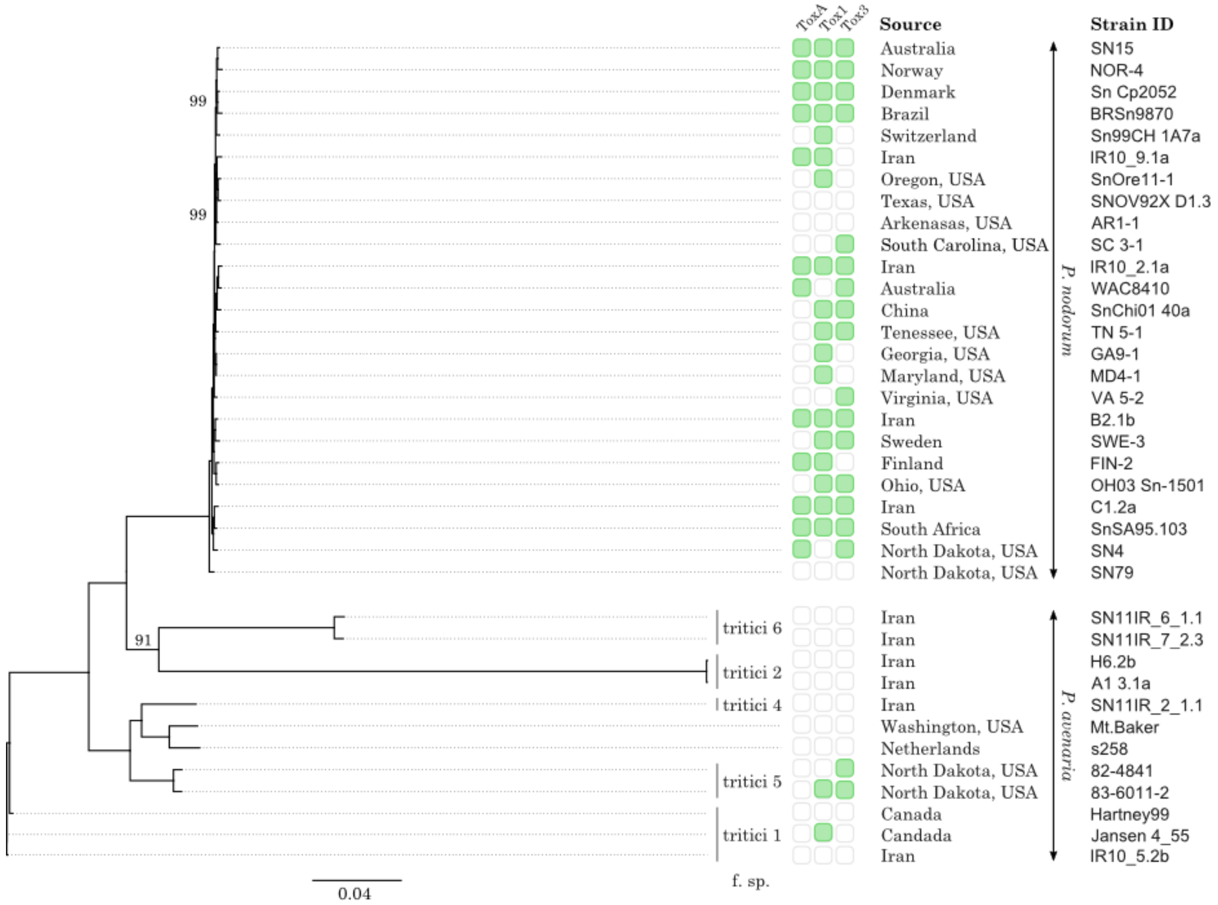


Figure 1: Phylogeny of the *P. nodorum* and *P. avenaria* strains used in this study constructed using MrBayes from 35 loci present in all strains. Branch probabilities are 100% unless indicated otherwise. Green boxes represent presence of an effector in that strain.

Unlike SnToxA and SnTox1, SnTox3-like loci are found in multiple copies in SN15 and other *Phaeosphaeria strains*. The pan-*Phaeosphaeria* SnTox3-like paralogs cluster into five groups (Figure 2). Included in the paralogs are 26 pseudogenes with open reading frames interrupted by premature stop codons and 74 loci that have coding sequences uninterrupted by nonsense mutations. All of the complete genes have secretion signals as predicted by SignalP and all the cysteine residues are conserved. The two Pat5 isolates 82-4841 and 83-6011-2 contain at least three SnTox3-like loci each, but are the only *P. avenaria* isolates to do so. The Pat5 strains each encode a SnTox3-like gene that clusters with the genuine SN15 SnTox3 gene SNOG_08981 (Figure 2). The coding sequence of the Pat5 genes are similar to SnTox3 (Figure 3), but the genomic context of these Pat5 loci is distinct from the *P. nodorum* strains. The genes surrounding SnTox3 on SN15 scaffold 14 are absent in the *P. avenaria* assemblies, including the Pat5 isolates. The region surrounding the Pat5 Tox3-like gene is syntenic to a region in SN15 on a different scaffold to SnTox3. The Tox3-like loci in Pat5 strains have replaced the putative NADPH:quinone reductase gene SNOG_09936 on scaffold 16 (Figure 4).

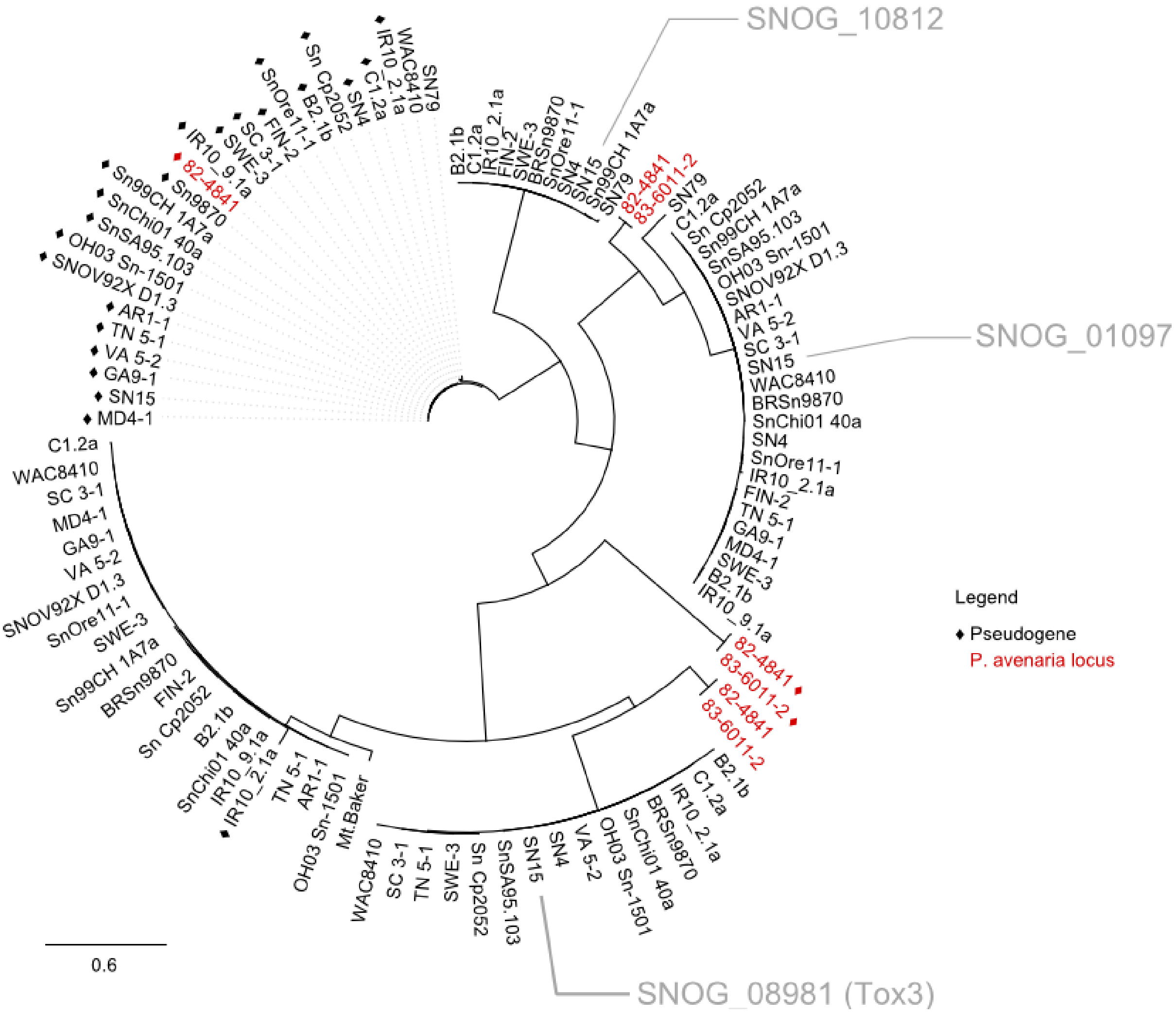


Figure 2: Maximum likelihood phylogenetic tree of Tox3-like paralogs in the pan-*Phaeosphaeria* genome. The two Pat5 strains (highlighted in red) are the only *P. avenaria* isolates with Tox3-like loci, including loci that cluster with the genuine SnTox3 gene (SNOG_08981).

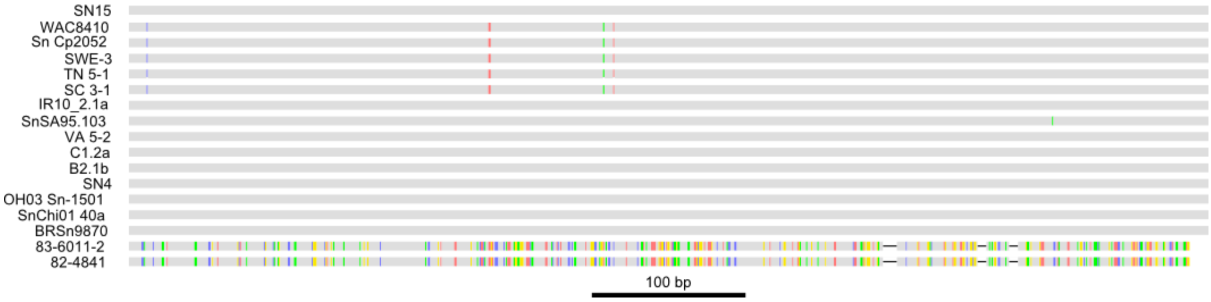


Figure 3: Nucleotide alignment of the loci clustered most closely with SN15 SnTox3 SNOG_08981. Variations to the SN15 sequence are highlighted in colour according to the SNP.

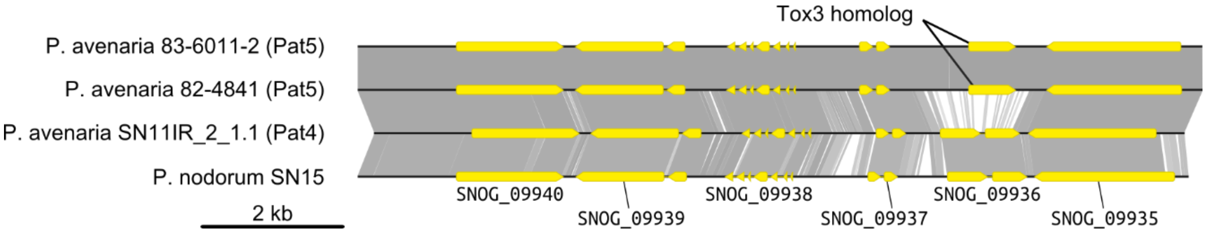


Figure 4: ClustalW alignment of the genomic context of the Tox3-like genes in Pat5 strains with homologous regions in Pat4 and *P. nodorum*. The SnTox3 region on SN15 scaffold 14 is absent from all *P. avenaria* assemblies. A region of *P. nodorum* SN15 scaffold 16 shown here, is syntenic with *P. avenaria* with the exception that a Tox3 homolog replaces SNOG_09936 in Pat5 strains.

Calculations of Presence/Absence by Read Mapping Overestimate Sectional Absence

Two pipelines were used to call variants between the reference and each alternate strain. For *P. nodorum* strains, the mapping/GATK method was able to genotype 90.3% of the genome and the nucmer method was able to find variants over 93.5% of the genome. In *P. avenaria* strains, the mapping/GATK method was able to genotype only 16.2% of the genome and the nucmer method was able to find variants over 61.9% of the genome.

Putative Accessory Chromosome Specific to Virulent Strains

The coverage by nucmer matches of the large SN15 scaffolds by scaffolds from each of the alternate strains is consistent (Figure 5). Scaffold 5 shows a slight decrease in coverage due to the large tandem array of rDNA repeats (Hane and Oliver 2008). The average coverage of the reference scaffolds by the alternate *P. nodorum* strains is 61.8% (Figure 2). Scaffold 44 has 75 genes and 1.9% repetitive sequence. It has 6.7% coverage by SN79, and a mean coverage of 5.7% by the *P. avenaria* strains. Scaffold 45 has 61 genes and 2.4% repetitive sequence. It has 5.0% coverage by SN79 and a mean coverage of 7.7 by the *P. avenaria* strains. Scaffold 46 was predicted to be dispensable by Ohm, Feau et al. (2012). It encodes no genes, has 12.4% repetitive sequence, 25.9% coverage by SN79, and a mean coverage by *P. avenaria* strains of 9.6%. Scaffold 51, which contains SnToxA (Chapter 5) is also absent from the *P. avenaria* strains. *P. avenaria s*trains H6.2b and A1 3.1a are the most dissimilar to SN15 (Figure 1). The dissimilarity of the reference to these isolates has resulted in fewer nucmer matches and lower scaffold coverage. Fragmented and incomplete assemblies from strains SC3-1, SnSA95.103 and NOR-4 (Table 2) were excluded in this analysis.

Abeysekara, N. S., T. L. Friesen, B. Keller and J. D. Faris (2009). "Identification and characterization of a novel host–toxin interaction in the wheat–*Stagonospora nodorum* pathosystem." Theoretical and applied genetics **120**(1): 117-126.

Auwera, G. A., M. O. Carneiro, C. Hartl, R. Poplin, G. del Angel, A. Levy‐Moonshine, T. Jordan, K. Shakir, D. Roazen and J. Thibault (2013). "From FastQ data to high‐confidence variant calls: the genome analysis toolkit best practices pipeline." Current protocols in bioinformatics: 11.10. 11-11.10. 33.

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

Bankevich, A., S. Nurk, D. Antipov, A. A. Gurevich, M. Dvorkin, A. S. Kulikov, V. M. Lesin, S. I. Nikolenko, S. Pham and A. D. Prjibelski (2012). "SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing." Journal of computational biology **19**(5): 455-477.

Bendtsen, J. D., H. Nielsen, G. von Heijne and S. Brunak (2004). "Improved prediction of signal peptides: SignalP 3.0." J Mol Biol **340**(4): 783-795.

Bringans, S., J. K. Hane, T. Casey, K. C. Tan, R. Lipscombe, P. S. Solomon and R. P. Oliver (2009). "Deep proteogenomics; high throughput gene validation by multidimensional liquid chromatography and mass spectrometry of proteins from the fungal wheat pathogen *Stagonospora nodorum*." BMC Bioinformatics **10**: 301.

Cannon, S. B., L. Sterck, S. Rombauts, S. Sato, F. Cheung, J. Gouzy, X. Wang, J. Mudge, J. Vasdewani and T. Schiex (2006). "Legume genome evolution viewed through the Medicago truncatula and Lotus japonicus genomes." Proceedings of the National Academy of Sciences **103**(40): 14959-14964.

Cantarel, B. L., I. Korf, S. M. C. Robb, G. Parra, E. Ross, B. Moore, C. Holt, A. Sánchez Alvarado and M. Yandell (2008). "MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes." Genome research **18**: 188-196.

Chen, M., P. SanMiguel, A. De Oliveira, S.-S. Woo, H. Zhang, R. A. Wing and J. Bennetzen (1997). "Microcolinearity in sh2-homologous regions of the maize, rice, and sorghum genomes." Proceedings of the National Academy of Sciences **94**(7): 3431-3435.

Chu, C., S. S. Xu and J. D. Faris (2012). "SnTox5–Snn5: a novel Stagonospora nodorum effector–wheat gene interaction and its relationship with the SnToxA–Tsn1 and SnTox3–Snn3–B1 interactions." Molecular plant pathology **13**(9): 1101-1109.

Clutterbuck, A. J. (2011). "Genomic evidence of repeat-induced point mutation (RIP) in filamentous ascomycetes." Fungal Genetics and Biology **48**(3): 306-326.

Coleman, J. J., S. D. Rounsley, M. Rodriguez-Carres, A. Kuo, C. C. Wasmann, J. Grimwood, J. Schmutz, M. Taga, G. J. White and S. Zhou (2009). "The genome of Nectria haematococca: contribution of supernumerary chromosomes to gene expansion." PLoS Genet **5**(8): e1000618.

Croll, D., M. H. Lendenmann, E. Stewart and B. A. McDonald (2015). "The impact of recombination hotspots on genome evolution of a fungal plant pathogen." Genetics **201**(3): 1213-1228.

Cunfer, B. M. (2000). "Stagonospora and Septoria diseases of barley, oat, and rye." Canadian journal of plant pathology **22**(4): 332-348.

DePristo, M. A., E. Banks, R. Poplin, K. V. Garimella, J. R. Maguire, C. Hartl, A. A. Philippakis, G. Del Angel, M. A. Rivas and M. Hanna (2011). "A framework for variation discovery and genotyping using next-generation DNA sequencing data." Nature genetics **43**(5): 491-498.

Engström, P. G., S. J. H. Sui, Ø. Drivenes, T. S. Becker and B. Lenhard (2007). "Genomic regulatory blocks underlie extensive microsynteny conservation in insects." Genome research **17**(12): 1898-1908.

Favorov, A., L. Mularoni, L. M. Cope, Y. Medvedeva, A. A. Mironov, V. J. Makeev and S. J. Wheelan (2012). "Exploring massive, genome scale datasets with the GenometriCorr package." PLoS computational biology **8**(5): e1002529.

Friesen, T. L., S. W. Meinhardt and J. D. Faris (2007). "The Stagonospora nodorum‐wheat pathosystem involves multiple proteinaceous host‐selective toxins and corresponding host sensitivity genes that interact in an inverse gene‐for‐gene manner." The Plant Journal **51**(4): 681-692.

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

Friesen, T. L., Z. Zhang, P. S. Solomon, R. P. Oliver and J. D. Faris (2008). "Characterization of the interaction of a novel Stagonospora nodorum host-selective toxin with a wheat susceptibility gene." Plant Physiology **146**(2): 682-693.

Fudal, I., S. Ross, H. Brun, A.-L. Besnard, M. Ermel, M.-L. Kuhn, M.-H. Balesdent and T. Rouxel (2009). "Repeat-induced point mutation (RIP) as an alternative mechanism of evolution toward virulence in Leptosphaeria maculans." Molecular Plant-Microbe Interactions **22**(8): 932-941.

Galagan, J. E., M. R. Henn, L.-J. Ma, C. A. Cuomo and B. Birren (2005). "Genomics of the fungal kingdom: insights into eukaryotic biology." Genome research **15**(12): 1620-1631.

Galagan, J. E. and E. U. Selker (2004). "RIP: the evolutionary cost of genome defense." Trends in Genetics **20**(9): 417-423.

Gao, Y., J. Faris, Z. Liu, Y. Kim, R. Syme, R. Oliver, S. Xu and T. Friesen (2015). "Identification and characterization of the SnTox6-Snn6 interaction in the Parastagonospora nodorum–wheat pathosystem." Molecular Plant-Microbe Interactions **28**(5): 615-625.

Goffeau, A., B. G. Barrell, H. Bussey and R. Davis (1996). "Life with 6000 genes." Science **274**(5287): 546.

Golicz, A. A., P. A. Martinez, M. Zander, D. A. Patel, A. P. Van De Wouw, P. Visendi, T. L. Fitzgerald, D. Edwards and J. Batley (2015). "Gene loss in the fungal canola pathogen Leptosphaeria maculans." Functional & integrative genomics **15**(2): 189-196.

Goodwin, S. B., S. B. M'Barek, B. Dhillon, A. H. J. Wittenberg, C. F. Crane, J. K. Hane, A. J. Foster, T. A. J. van der Lee, J. Grimwood, A. Aerts, J. Antoniw, A. Bailey, B. Bluhm, J. Bowler, J. Bristow, A. van der Burgt, B. Canto-Canché, A. C. L. Churchill, L. Conde-Ferràez, H. J. Cools, P. M. Coutinho, M. Csukai, P. Dehal, P. de Wit, B. Donzelli, H. C. van de Geest, R. C. H. J. van Ham, K. E. Hammond-Kosack, B. Henrissat, A. Kilian, A. K. Kobayashi, E. Koopmann, Y. Kourmpetis, A. Kuzniar, E. Lindquist, V. Lombard, C. Maliepaard, N. Martins, R. Mehrabi, J. P. H. Nap, A. Ponomarenko, J. J. Rudd, A. Salamov, J. Schmutz, H. J. Schouten, H. Shapiro, I. Stergiopoulos, S. F. F. Torriani, H. Tu, R. P. de Vries, C. Waalwijk, S. B. Ware, A. Wiebenga, L. H. Zwiers, R. P. Oliver, I. V. Grigoriev and G. H. J. Kema (2011). "Finished genome of the fungal wheat pathogen *Mycosphaerella graminicola* reveals dispensome structure, chromosome plasticity, and stealth pathogenesis." PLoS Genetics **7**(6).

Grandaubert, J., A. Bhattacharyya and E. H. Stukenbrock (2015). "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: Genes| Genomes| Genetics **5**(7): 1323-1333.

Grandaubert, J., C. Schoch, H. Bohran, B. Howlett, M.-H. Balesdent and T. Rouxel (2013). Transposon analysis and comparative genomics of the Leptosphaeria maculans-L. biglobosa species complex. Dothideomycetes Comparative Genomics workshop 27th Fungal Genetics Conference at Asilomar. 2013-03-122013-03-12, Asilomar.

Gurevich, A., V. Saveliev, N. Vyahhi and G. Tesler (2013). "QUAST: quality assessment tool for genome assemblies." Bioinformatics **29**(8): 1072-1075.

Hane, J. K. (2015). Calculating RIP Mutation in Fungal Genomes Using RIPCAL. Genetic Transformation Systems in Fungi, Volume 2, Springer**:** 69-78.

Hane, J. K., J. P. Anderson, A. H. Williams, J. Sperschneider and K. B. Singh (2014). "Genome sequencing and comparative genomics of the broad host-range pathogen Rhizoctonia solani AG8." PLoS Genet **10**(5): e1004281.

Hane, J. K., R. G. Lowe, P. S. Solomon, K.-C. Tan, C. L. Schoch, J. W. Spatafora, P. W. Crous, C. Kodira, B. W. Birren and J. E. Galagan (2007). "Dothideomycete–plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*." The Plant Cell **19**(11): 3347-3368.

Hane, J. K., R. G. Lowe, P. S. Solomon, K.-C. Tan, C. L. Schoch, J. W. Spatafora, P. W. Crous, C. Kodira, B. W. Birren and J. E. Galagan (2007). "Dothideomycete–plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*." The Plant Cell Online **19**(11): 3347-3368.

Hane, J. K. and R. P. Oliver (2008). "RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences." BMC bioinformatics **9**(1): 478.

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

Hane, J. K. and R. P. Oliver (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, J. K., T. Rouxel, B. J. Howlett, G. H. Kema, S. B. Goodwin and R. P. Oliver (2011). "A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi." Genome Biol **12**(5): R45.

Hane, J. K., A. H. Williams, A. P. Taranto, P. S. Solomon and R. P. Oliver (2015). Repeat-induced point mutation: a fungal-specific, endogenous mutagenesis process. Genetic Transformation Systems in Fungi, Volume 2, Springer**:** 55-68.

Ipcho, S. V. S., J. K. Hane, E. A. Antoni, D. Ahren, B. Henrissat, T. L. Friesen, P. S. Solomon and R. P. Oliver (2012). "Transcriptome analysis of Stagonospora nodorum: Gene models, effectors, metabolism and pantothenate dispensability." Molecular Plant Pathology **13**(6): 531-545.

Jurka, J., V. V. Kapitonov, A. Pavlicek, P. Klonowski, O. Kohany and J. Walichiewicz (2005). "Repbase Update, a database of eukaryotic repetitive elements." Cytogenetic and genome research **110**(1-4): 462-467.

Krogh, A., B. Larsson, G. Von Heijne and E. L. Sonnhammer (2001). "Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes." Journal of molecular biology **305**(3): 567-580.

Kurtz, S., A. Phillippy, A. L. Delcher, M. Smoot, M. Shumway, C. Antonescu and S. L. Salzberg (2004). "Versatile and open software for comparing large genomes." Genome biology **5**(2): R12.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nature methods **9**(4): 357-359.

Larkin, M. A., G. Blackshields, N. P. Brown, R. Chenna, P. A. McGettigan, H. McWilliam, F. Valentin, I. M. Wallace, A. Wilm, R. Lopez, J. D. Thompson, T. J. Gibson and D. G. Higgins (2007). "Clustal W and Clustal X version 2.0." Bioinformatics **23**(21): 2947-2948.

Lechner, M., S. Findeiß, L. Steiner, M. Marz, P. F. Stadler and S. J. Prohaska (2011). "Proteinortho: detection of (co-) orthologs in large-scale analysis." BMC bioinformatics **12**(1): 124.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis and R. Durbin (2009). "The sequence alignment/map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Liu, Z., J. D. Faris, R. P. Oliver, K. C. Tan, P. S. Solomon, M. C. McDonald, B. A. McDonald, A. Nunez, S. Lu, J. B. Rasmussen and T. L. Friesen (2009). "SnTox3 acts in effector triggered susceptibility to induce disease on wheat carrying the Snn3 gene." PLoS Pathogens **5**(9): e1000581.

Liu, Z., Z. Zhang, J. D. Faris, R. P. Oliver, R. Syme, M. C. McDonald, B. A. McDonald, P. S. Solomon, S. Lu and W. L. Shelver (2012). "The Cysteine Rich Necrotrophic Effector SnTox1 Produced by Stagonospora nodorum Triggers Susceptibility of Wheat Lines Harboring Snn1." PLoS Pathogens **8**(1): e1002467.

Liu, Z. H., J. D. Faris, S. W. Meinhardt, S. Ali, J. B. Rasmussen and T. L. Friesen (2004). "Genetic and physical mapping of a gene conditioning sensitivity in wheat to a partially purified host-selective toxin produced by *Stagonospora nodorum*." Phytopathology **94**(10): 1056-1060.

Ma, L.-J., H. C. Van Der Does, K. A. Borkovich, J. J. Coleman, M.-J. Daboussi, A. Di Pietro, M. Dufresne, M. Freitag, M. Grabherr and B. Henrissat (2010). "Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium." Nature **464**(7287): 367-373.

Malkus, A., E. Reszka, C.-J. Chang, E. Arseniuk, P.-F. L. Chang and P. P. Ueng (2005). "Sequence diversity of β-tubulin (tubA) gene in Phaeosphaeria nodorum and P. avenaria." FEMS microbiology letters **249**(1): 49-56.

Martin, M. (2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads." embnet **17**.

McDonald, M. C., R. P. Oliver, T. L. Friesen, P. C. Brunner and B. A. McDonald (2013). "Global diversity and distribution of three necrotrophic effectors in Phaeosphaeria nodorum and related species." New Phytologist **199**(1): 241-251.

McGary, K. L., J. C. Slot and A. Rokas (2013). "Physical linkage of metabolic genes in fungi is an adaptation against the accumulation of toxic intermediate compounds." Proceedings of the National Academy of Sciences **110**(28): 11481-11486.

Medema, M. H., K. Blin, P. Cimermancic, V. de Jager, P. Zakrzewski, M. A. Fischbach, T. Weber, E. Takano and R. Breitling (2011). "antiSMASH: rapid identification, annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences." Nucleic acids research **39**(suppl 2): W339-W346.

Miao, V. P., S. F. Covert and H. D. VanEtten (1991). "A fungal gene for antibiotic resistance on a dispensable (" B") chromosome." Science **254**(5039): 1773.

Ohm, R. A., N. Feau, B. Henrissat, C. L. Schoch, B. A. Horwitz, K. W. Barry, B. J. Condon, A. C. Copeland, B. Dhillon and F. Glaser (2012). "Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi." PLoS Pathogens **8**(12): e1003037.

Oliver, R. P., T. L. Friesen, J. D. Faris and P. S. Solomon (2012). "Stagonospora nodorum: from pathology to genomics and host resistance." Annual review of phytopathology **50**: 23-43.

Oliver, R. P. and P. S. Solomon (2010). "New developments in pathogenicity and virulence of necrotrophs." Current opinion in plant biology **13**(4): 415-419.

Pennacchio, L. A. (2003). "Insights from human/mouse genome comparisons." Mammalian genome **14**(7): 429-436.

Plissonneau, C., A. Stürchler and D. Croll (2016). "The Evolution of Orphan Regions in Genomes of a Fungal Pathogen of Wheat." mBio **7**(5): e01231-01216.

Quaedvlieg, W., G. Verkley, H.-D. Shin, R. Barreto, A. Alfenas, W. Swart, J. Groenewald and P. Crous (2013). "Sizing up Septoria." Studies in Mycology **75**: 307-390.

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics (Oxford, England) **26**: 841-842.

Rausch, T., T. Zichner, A. Schlattl, A. M. Stütz, V. Benes and J. O. Korbel (2012). "DELLY: structural variant discovery by integrated paired-end and split-read analysis." Bioinformatics **28**(18): i333-i339.

Ronquist, F. and J. P. Huelsenbeck (2003). "MrBayes 3: Bayesian phylogenetic inference under mixed models." Bioinformatics **19**(12): 1572-1574.

Shi, G., T. L. Friesen, J. Saini, S. S. Xu, J. B. Rasmussen and J. D. Faris (2015). "The wheat gene confers susceptibility on recognition of the necrotrophic effector SnTox7." The Plant Genome **8**(2).

Simakov, O., F. Marletaz, S.-J. Cho, E. Edsinger-Gonzales, P. Havlak, U. Hellsten, D.-H. Kuo, T. Larsson, J. Lv and D. Arendt (2013). "Insights into bilaterian evolution from three spiralian genomes." Nature **493**(7433): 526-531.

Smit, A. and R. Hubley (2010). "RepeatModeler Open-1.0." Repeat Masker Website.

Smit, A., R. Hubley and P. Green. (1996-2004). "RepeatMasker Open-3.0.", from http://www.repeatmasker.org.

Solomon, P. S., R. G. Lowe, K. C. TAN, O. D. Waters and R. P. Oliver (2006). "Stagonospora nodorum: cause of stagonospora nodorum blotch of wheat." Molecular plant pathology **7**(3): 147-156.

Stukenbrock, E. H., F. G. Jørgensen, M. Zala, T. T. Hansen, B. A. McDonald and M. H. Schierup (2010). "Whole-genome and chromosome evolution associated with host adaptation and speciation of the wheat pathogen Mycosphaerella graminicola." PLoS Genet **6**(12): e1001189.

Suyama, M., D. Torrents and P. Bork (2006). "PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments." Nucleic acids research **34**(suppl 2): W609-W612.

Syme, R. A., J. K. Hane, T. L. Friesen and R. P. Oliver (2013). "Resequencing and comparative genomics of *Stagonospora nodorum*: Sectional gene absence and effector discovery." G3: Genes| Genomes| Genetics **3**(6): 959-969.

Syme, R. A., K.-C. Tan, J. K. Hane, K. Dodhia, T. Stoll, M. Hastie, E. Furuki, S. R. Ellwood, A. H. Williams and Y.-F. Tan (2016). "Comprehensive annotation of the Parastagonospora nodorum reference genome using next-generation genomics, transcriptomics and proteogenomics." PloS one **11**(2): e0147221.

Tan, K.-C., R. P. Oliver, P. S. Solomon and C. S. Moffat (2010). "Proteinaceous necrotrophic effectors in fungal virulence." Functional Plant Biology **37**(10): 907-912.

Testa, A. C. (2016). "Effector gene prediction from fungal pathogen genome assemblies."

Testa, A. C., J. K. Hane, S. R. Ellwood and R. P. Oliver (2015). "CodingQuarry: highly accurate hidden Markov model gene prediction in fungal genomes using RNA-seq transcripts." BMC genomics **16**(1): 170.

Ueng, P., K. Subramaniam, W. Chen, E. Arseniuk, L. Wang, A. Cheung, G. Hoffmann and G. Bergstrom (1998). "Intraspecific genetic variation of Stagonospora avenae and its differentiation from S. nodorum." Mycological Research **102**(5): 607-614.

Ueng, P. P. and W. Chen (1994). "Genetic differentiation between Phaeosphaeria nodorum and P. avenaria using restriction fragment length polymorphisms." Phytopathology **84**(8): 800-806.

Ueng, P. P., B. M. Cunfer, A. S. Alano, J. D. Youmans and W. Chen (1995). "Correlation between molecular and biological characters in identifying the wheat and barley biotypes of Stagonospora nodorum." Phytopathology **85**(1): 44-52.

Ueng, P. P., Q. Dai, K.-r. Cui, P. C. Czembor, B. M. Cunfer, H. Tsang, E. Arseniuk and G. C. Bergstrom (2003). "Sequence diversity of mating-type genes in Phaeosphaeria avenaria." Current genetics **43**(2): 121-130.

Van de Wouw, A. P., A. J. Cozijnsen, J. K. Hane, P. C. Brunner, B. A. McDonald, R. P. Oliver and B. J. Howlett (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**(11): e1001180.

Van De Wouw, A. P. and B. J. Howlett (2011). "Fungal pathogenicity genes in the age of 'omics'." Molecular Plant Pathology **12**(5): 507-514.

VanEtten, H., D. Straney, S. Covert and H. Kistler (2001). The genetics of Nectria haematococca mating population VI with special emphasis on its conditionally dispensable (CD) chromosomes: a source of habitat specific genes. Summerell BA, Leslie J, Back house D, Bryden WL and Burgess LW (eds) Fusarium: Paul Nelson Memorial Symposium. American Phytophatological Press, St. Paul, MN, USA.

Vleeshouwers, V. G. and R. P. Oliver (2014). "Effectors as tools in disease resistance breeding against biotrophic, hemibiotrophic, and necrotrophic plant pathogens." Molecular Plant-Microbe Interactions **27**(3): 196-206.

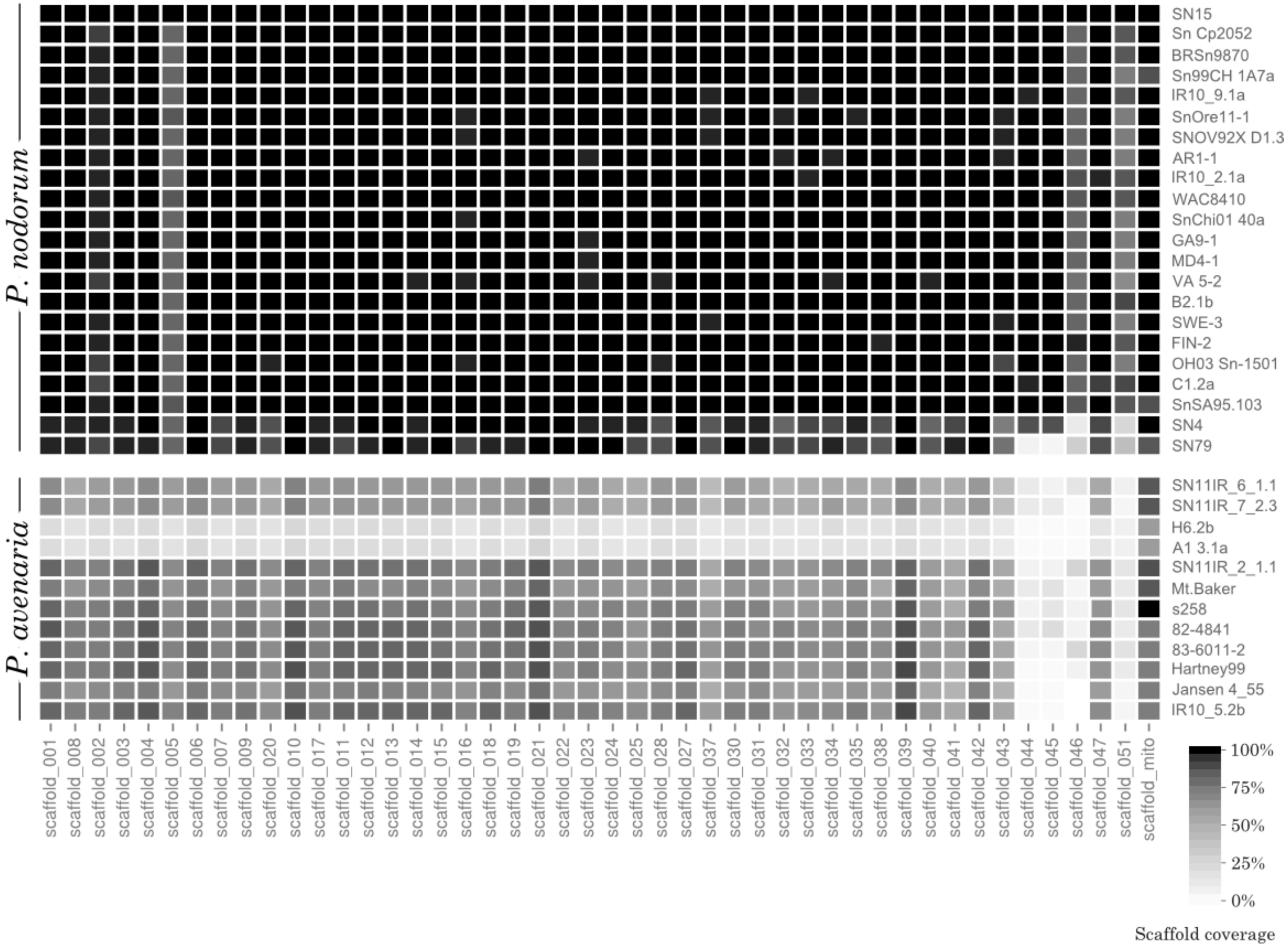


Figure 5: Coverage of reference *P. nodorum* genic scaffolds > 100 kbp by nucmer matches from the alternate strains. Each column corresponds to a reference scaffold (ordered by size), and each row corresponds to one of the alternate strains.

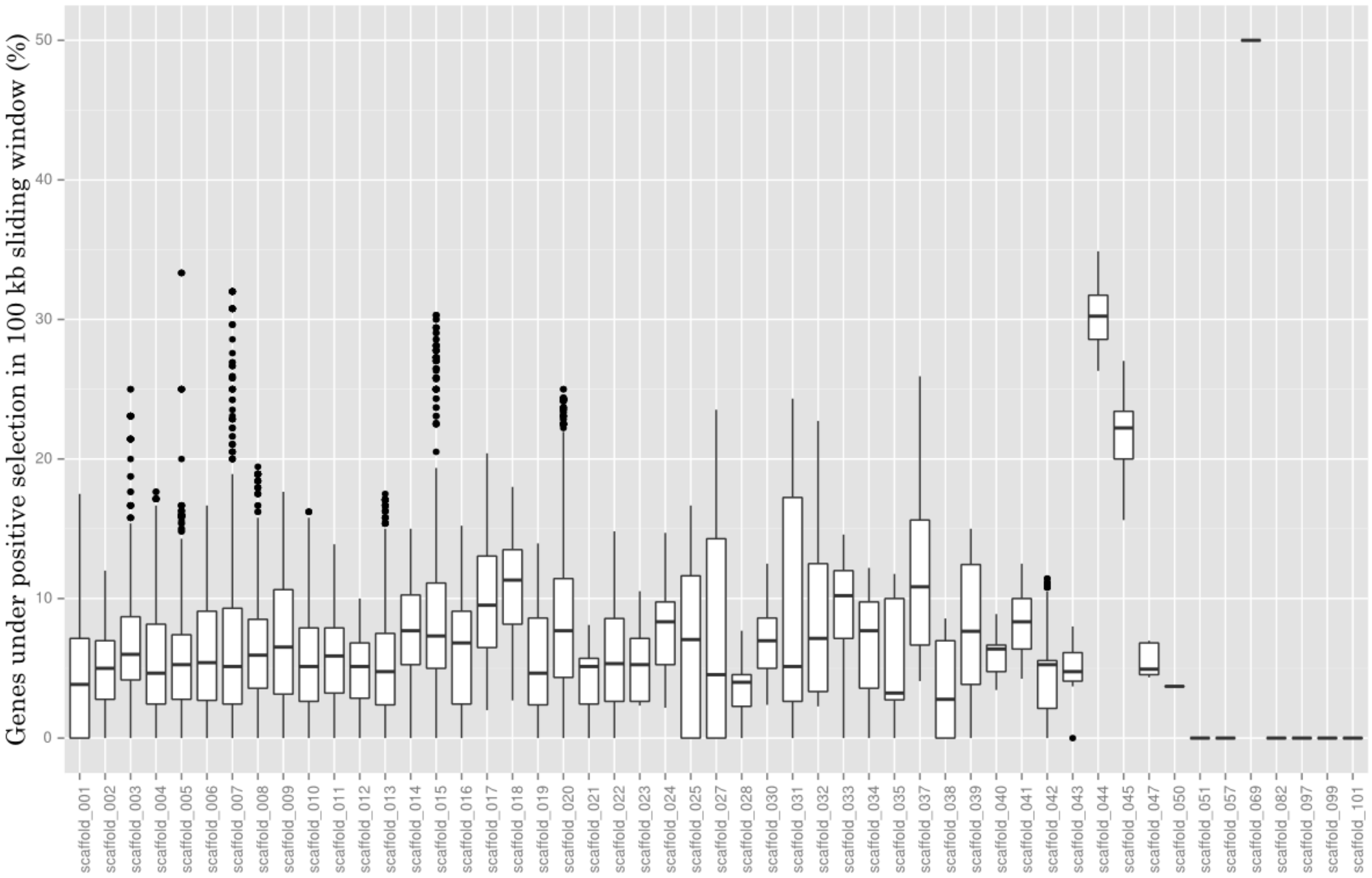


Figure 6: The percentage of genes under positive selection in 100 kbp sliding windows across the *P. nodorum* reference scaffolds. Outliers greater than Q3 + 1.5 x the interquartile range (IQR) or less than Q1 - 1.5 × IQR are plotted as filled circles.

Figure 6 shows the number of genes under a sliding 100 kb window that are under positive selection. A sliding window analysis was chosen rather than taking the average over whole chromosomes so as to avoid introducing bias for small scaffolds. This also allows us to detect regions where a higher than expected proportion of genes are under positive selection. The average percentage of genes under positive selection in genic scaffolds other than 44 and 45 is 6.6% (Figure 6). Scaffolds 44 and 45 are enriched for genes under positive selection for across their length (Figure 6, Figure 10), with 22 (29%) and 14 (23%) genes that show evidence of positive selection respectively.

Islands of positive selection were defined as consecutive windows where at least 20% of the genes in the window were under positive selection. Such islands were detected in scaffolds 7, 15, 20, 44, and 45 (Figures 7-10). The island of positive selection in scaffold 7 (551 kb – 613 kb) is adjacent to a large intergenic gap populated with repeats. The level of RIP-like mutations in SN15 is elevated in the repetitive region, particular when compared to *P. nodorum* strain Sn99CH 1A7a, but the elevated levels of RIP are confined to repetitive regions. The adjacent repetitive region is absent or highly mutated in all *P. avenaria* strains. Scaffold 15 has one region enriched for genes under positive selection at the start of the scaffold (29 kb – 130 kb). The region is also absent from or highly mutated in all *P. avenaria* strains. Scaffold 20 has one region enriched for genes under positive selection (622 kb - 723 kb). The region is adjacent to a repetitive region which has an elevated level of RIP mutations in the reference. The repetitive region is also highly mutated or absent from all *P.* avenaria strains.

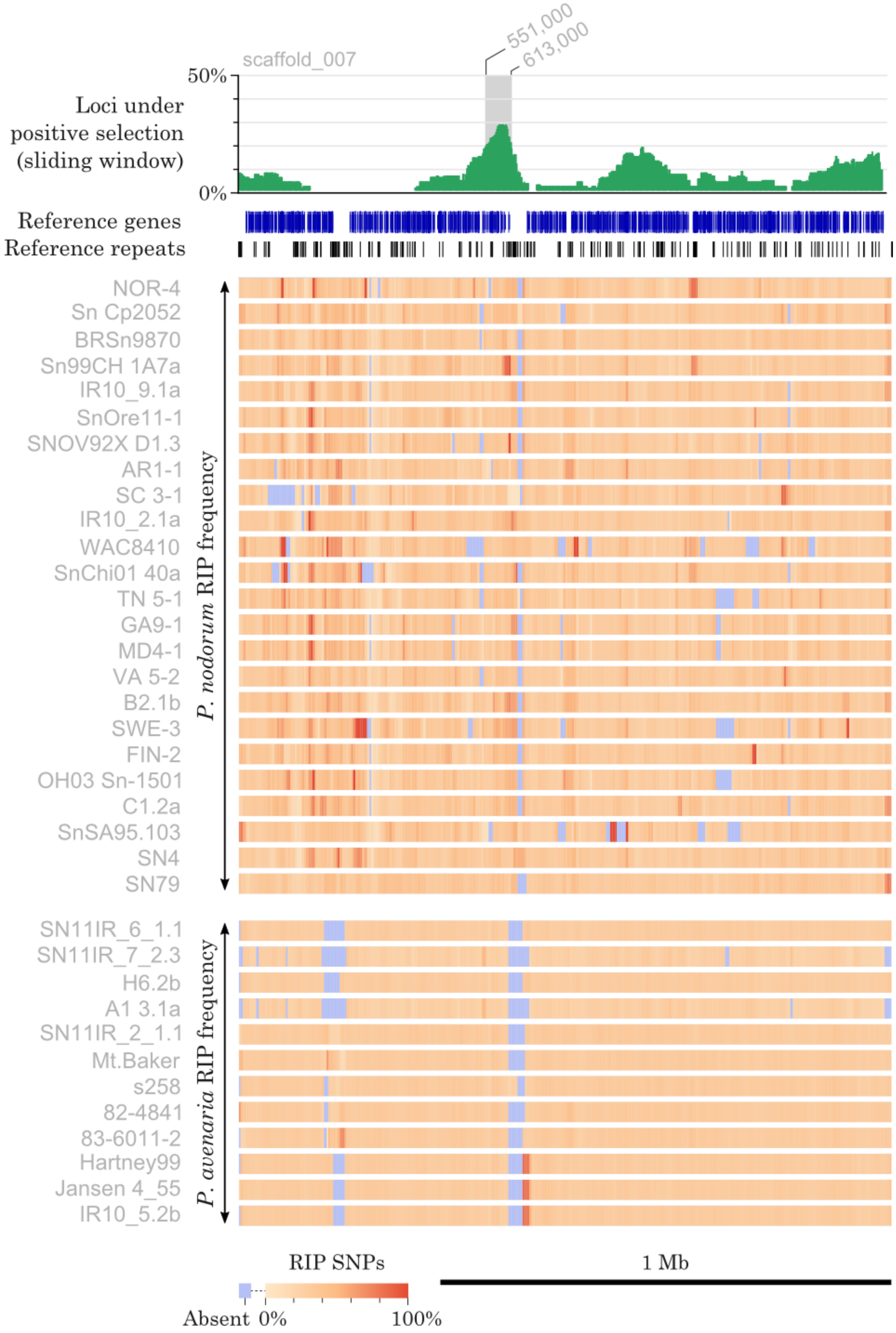


Figure 7: Distribution of loci under positive selection and RIP SNP frequency on scaffold_007. An island of positive selection between positions 551 kb and 613 kb.

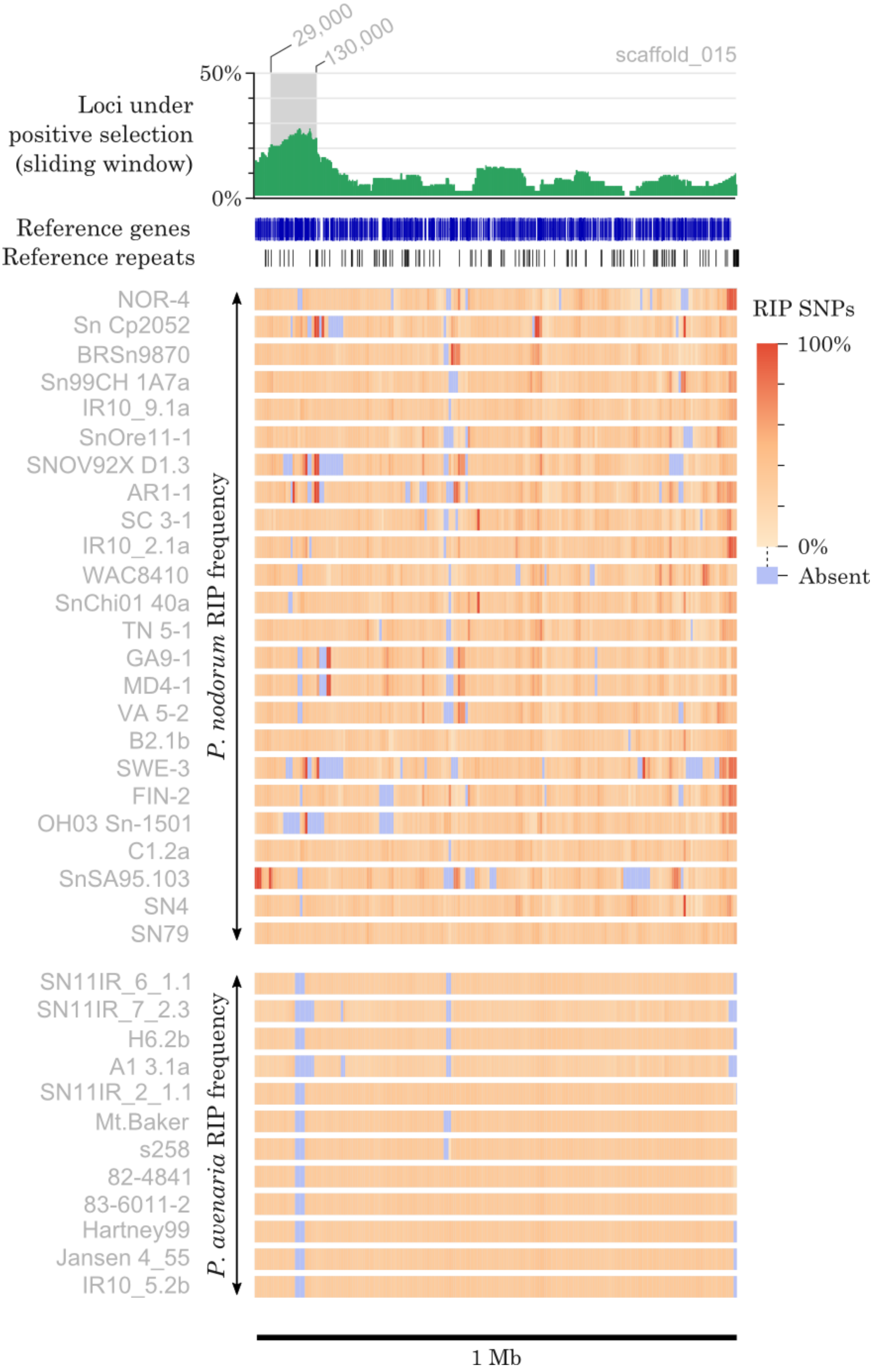


Figure 8: Distribution of loci under positive selection and RIP SNP frequency on scaffold_15. An island of positive selection between positions 29 kb and 130 kb.

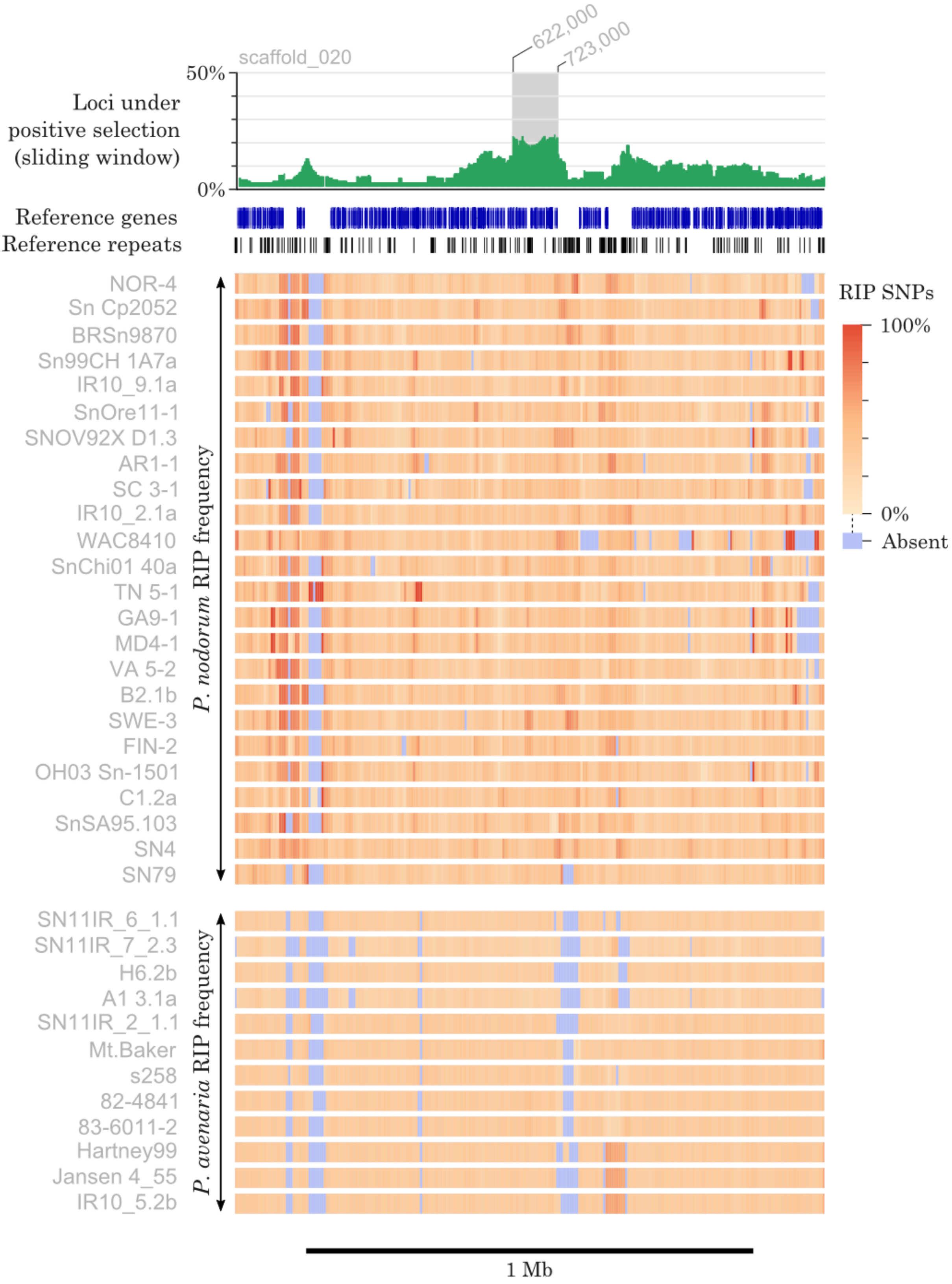


Figure 9: Distribution of loci under positive selection and RIP SNP frequency on scaffold_020. An island of positive selection between positions 622 kb and 723 kb.

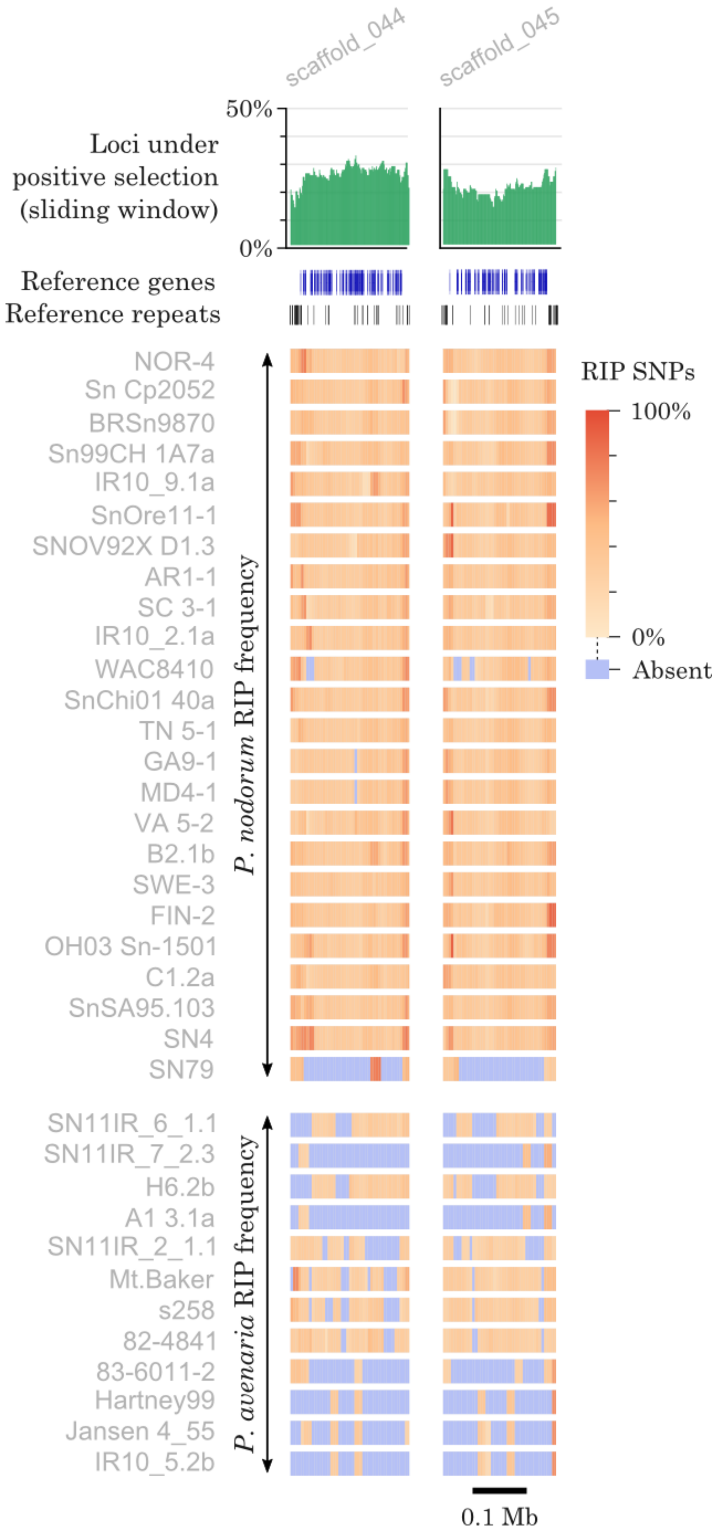


Figure 10: Distribution of loci under positive selection and RIP SNP frequency on scaffolds 44 and 45.

Variable Patterns of Gene Conservation Observed Across Strains

The conservation of genes between strains and species was studied. From the ProteinOrtho orthologue cluster analysis, 4451 protein clusters that were observed in all strains. There are 8128 protein clusters containing members in only one species including 198 proteins observed only in

SN15. The set of 'dispensable' proteins is defined here as proteins that are not species-specific (observed in fewer than 4 isolates) and not well conserved (missing in fewer than 3 isolates). This 'dispensable' set of 6480 proteins contains 4014 SN15 proteins, including all of the known effectors (Table 3, Figure 11). Figure 11 shows the distribution of ProteinOrtho protein clusters by strain and species specificity. The *P. avenaria* are more phylogenetically diverse than the *P. nodorum* clade (Figure 1). This diversity of *P. avenaria* is reflected in the low number (2) of proteins conserved between all *P. avenaria* strains, but absent from *P. nodorum* (Table 3). The unequal diversity between the two species also explains the differing rates of changing density moving away from the well conserved genes at the top-right of Figure 11. There are many more genes conserved in all *P. nodorum* strains and variably absent in *P. avenaria* than there are genes conserved in all *P. avenaria* strains and variably absent in *P. nodorum* (Figure 11).

Table 3: Summary of protein sequence conservation across the *P. nodorum* and *P. avenaria* strains calculated from ProteinOrtho orthologous protein clusters.

Reference protein set	13,563 proteins
Core Phaeosphaeria protein set	
Missing from 0 strains	4451 clusters
Missing from at most 1 strain	7100 clusters
Missing from at most 2 strains	8700 clusters
Core P. nodorum protein set	
Missing from 0 strains	9488 clusters (9595 SN15 proteins)
Missing in fewer than 2 *P. nodorum* strains	12049 clusters (12032 SN15 proteins)
Missing in fewer than 3 *P. nodorum* strains	12688 clusters (12573 SN15 proteins)
Strain-specific protein set	
Observed in only 1 strain	8128 clusters (198 SN15 proteins)
Observed in at most 2 strains	12951 clusters (306 SN15 proteins)
Observed in at most 3 strains	14674 clusters (356 SN15 proteins)
Dispensable protein set (effector-containing set)	
Observed in between 4 and 30 strains (inclusive)	6480 clusters (4014 SN15 proteins)
P. nodorum-specific proteins	
Present in all *P. nodorum* strains, absent from all *P. avenaria*	277
P. avenaria-specific proteins	
Present in all avenaria, absent from all *P. nodorum*	2

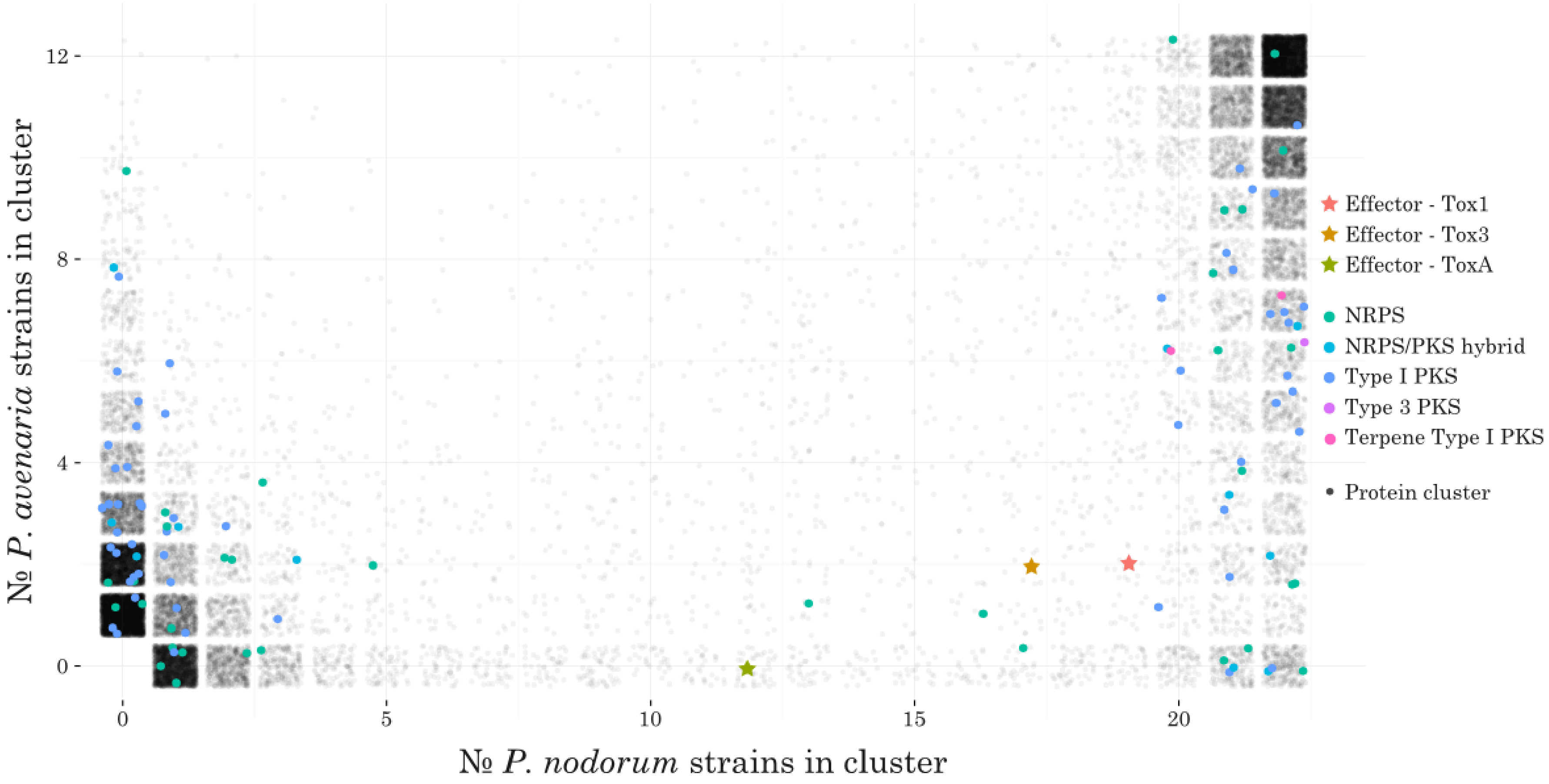


Figure 11: Distribution of protein cluster membership between the two species. Each point represents a group of one or more orthologous proteins. The number of *P. nodorum* strains that have contributed a protein to the cluster determines the x-axis location and the number of *P. avenaria* strains that have contributed a protein to the cluster determines the y-axis location. Core conserved genes with members from all strains are at the top-right and strain-specific genes are at the bottom-left. Known effectors and secondary metabolite synthesis proteins are highlighted.

Mesosyntenic Breakpoints Are Associated with Long Repeats

There were 2169 sites of inversion or translocation identified by comparing the reference assembly to the alternate *P. nodorum* strains. Furthermore, 822 breakpoint sites were predicted by comparison to the *P. avenaria* strains. Combining all predicted breakpoints from all strains resulted in a non-redundant set of 2230 sites. Breakpoint locations for all strains are included in Supplementary data. Mesosyntenic patterns are formed by frequent intra-chromosomal inversions, but the mechanism giving rise to this pattern is unknown. Both Ohm, Feau et al. (2012) and Grandaubert (2013) have looked for features associated with breakpoints in attempts to better understand the phenomenon.

The relative distance metric (Favorov, Mularoni et al. 2012) shows the association between the position of instances of each repeat class and the positions of inversion and translocation. A flat relative distance profile, indicates the repeat class is equally likely to be observed far from breakpoints as they are to be found close to breakpoints. Simple repeats, particularly GC-rich SSRs show flat relative distance profiles. Long repeats such as DNA transposons and LTR retrotransposons show a higher frequency of instances at low relative distance to structural rearrangement breakpoints (Figure 12).

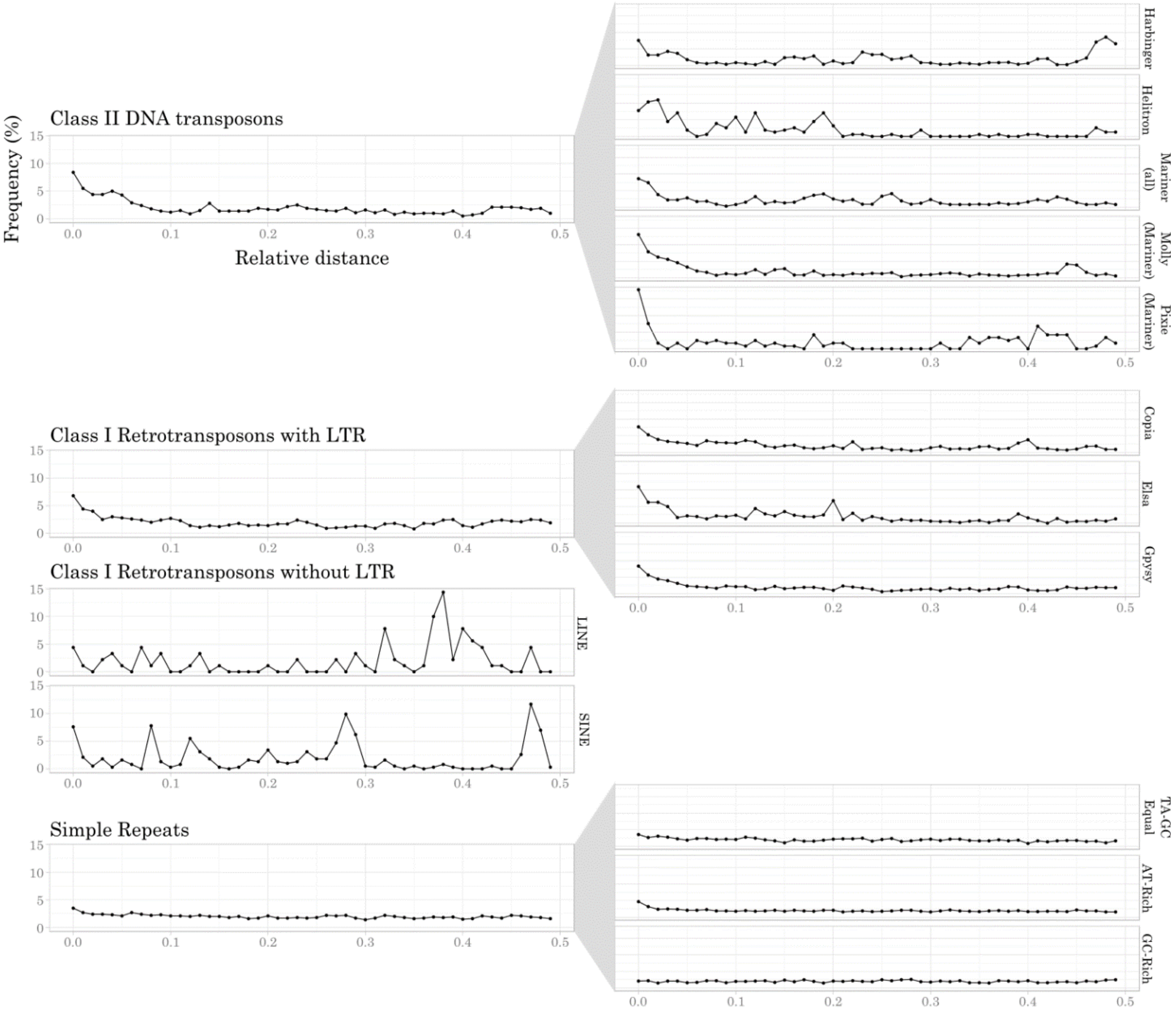


Figure 12: Frequency of relative distance between mesosyntenic breakpoints and repetitive sequence. Relative distance of a repeat instance to the set of breakpoints is the distance between the repeat and the nearest breakpoint divided by the distance between the breakpoints flanking the repeat instance. A repeat equidistant from two breakpoints would have a relative distance of 0.5, and a repeat that overlapped a repeat would have a relative distance of 0. The y-axis is consistent in all plots.

Effector Prediction

Each SN15 reference protein was assessed against a series of criteria defined by the expected or normative characteristics of necrotrophic effectors (Table 5). A positive score was added to 4362 small proteins, 616 cysteine-rich proteins, 3417 proteins encoded by transcripts near repeats, 308 proteins absent from SN79, 2414 proteins encoded in regions with low gene density, 1475 proteins predicted to be secreted and 945 proteins encoded by transcripts predicted to be under positive selection. The known *P.* nodorum effectors have been shown to exhibit scattered presence/absence frequency across populations, with no effector maintained across all strains and no effector found in only one strain (McDonald, Oliver et al. 2013). Negative scores were assigned to 12032 core proteins (missing from at most one *P. nodorum* strain), and 198 strain-specific proteins only found in SN15. Negative scores were also assigned to 2381 proteins predicted to be membrane-bound. Known effectors SnToxA, SnTox1 and SnTox3 all score highly using this system, and nine new loci from the updated manual annotation are among the top candidates (Table 5).

Table 4: Counts of the numbers of SN15 reference proteins that match each effector prediction criteria. Each predicted protein is assessed against each of these criteria and assigned a total score calculated as the sum of the criteria scores as described in methods section.

Criteria	№ Proteins
	Positive Scores
Small – less than 30 kDa	4362
Cysteine-rich – encodes an amino acid with > 4% cysteine residues	616
Near repeats – less than 5 kb from repetitive sequences	3417
Absent from SN79 – no blast hits to the avirulent strain	308
Low gene density – encoded in a region with large intergenic space	2414
Secreted – includes a signal peptide	1475
Diversifying selection – includes codons predicted to be subject to positive selection	945
	Negative Scores
Core Set– Missing in at most one strain	12032
Strain specific – only found in SN15	198
Membrane bound – not predicted to encode a transmembrane domain	2381

Table 5: Top effector candidates with scores greater than 4. Known effectors SnTox1, SnTox3 and SnToxA score highly under the current ranking system. There are 30 effector candidate that score as high as or higher than these three known effectors.

Transcript ID	Score	Effector Name
SNOR_20078	7	SnTox1
SNOR_01124	6	
SNOR_12811	6	
SNOR_30828	6	
SNOR_30343	5	
SNOR_00234	5	
SNOR_01601	5	
SNOR_03715	5	
SNOR_05030	5	
SNOR_05051	5	
SNOR_06079	5	
SNOR_08206	5	
SNOR_08981	5	SnTox3
SNOR_09446	5	
SNOR_09738	5	
SNOR_11828	5	
SNOR_14914	5	
SNOR_16166	5	
SNOR_16243	5	
SNOR_16270	5	
SNOR_16520	5	
SNOR_16571	5	SnToxA
SNOR_20154	5	
SNOR_30026	5	
SNOR_30077	5	
SNOR_30334	5	
SNOR_30466	5	
SNOR_30697	5	
SNOR_30802	5	

SNOR_30973	5

Discussion

Comparative genomics of fungal genomes within a species allows us to add a new dimension to a reference assembly and allow new questions to be asked of the data. Positive selection is only detectable by observation of a population of isolates, and can be used to uncover pathogenicity genes involved in an evolutionary arms race with the host. Stukenbrock, Bataillon et al. (2011) showed positive selection of SSPs by comparison of 12 Z. tritici genomes. Likewise, accessory chromosomes are difficult to predict from a single reference isolate, but are revealed by presence/absence differences between strains. Similarly, intra-specific comparison of eight Colletotrichum graminicola genomes showed evidence of positive selection at pathogenicity-related sequences including putative effector proteins and secondary metabolite biosynthetic enzymes (Rech, Sanz-Martín et al. 2014).

This study highlights various aspects of genomic variation between multiple isolates of *P. nodorum* and the related *P. avenaria*, from globally diverse source locations. *P. nodorum* and *P. avenaria* have common hosts and share many genes, but the two species are very different in their effector repertoire (Figure 1). Important exceptions are the Tox3-like loci identified on *P. avenaria* Pat5 isolates 82-4841 and 83-6011-2. In these strains, multiple Tox3-like loci are present, mirroring the paralogs found in *P. nodorum* strains. The Tox3-like paralog most similar to SN15 SnTox3 (Figure 2 and Figure 3) in Pat5 genomes are located in a region not homologous to the Tox3 region in *P. nodorum* strains (Figure 4).The Tox3-like genes in Pat5 has clearly replaced another gene in a region devoid of repetitive elements (Figure 4). The mechanism for a single-gene swap as seen here is unclear at this time, but it stands in stark contrast to the horizontal gene transfer of multi-kilobase SnToxA region between *P. nodorum* and *P. tritici-repentis (Friesen, Stukenbrock et al. 2006, Syme, Hane et al. 2013).*

Notably, patterns of presence-absence variation are highly prominent between isolates (Figure 5). Scaffolds 44 and 45 are distinct from other scaffolds in their gene density, the number of genes under positive selection, and their pattern of scaffold-level presence/absence among isolates. Reminiscent of the 'two-speed' genome observed in other fungal pathogens (Croll and McDonald 2012), it is likely that these scaffolds constitute part of one or more accessory chromosomes. Their absence from SN79 demonstrates that the genes are not strictly required by the fungus, but their retention in all other wheat-infecting strains and elevated rates of positive selection suggest that they confer some significant advantage. Strong positive selection observed in SN15 scaffolds 44 and 45 suggests that in *P. nodorum* as in other fungal phytopathogens, accessory chromosomes may be used as a 'workshop' for novel genetic material. Scaffold 46 shows similar patterns of presence/absence between strains, but have no annotated genes on which positive selection might act.

Surveying the overall mutation rate across isolates, we observed an oversight in short-read sequencing approaches of the past. Previous genome comparisons between *P. nodorum* species used the depth of reads mapped to the reference genome to infer gene absence in the alternate strain (Syme, Hane et al. 2013). Regions of heavy mutation can prevent reads from mapping to the reference, inflating the count of genes absent in the alternate strain. One solution is to compare the alternate strain's *de novo* assemblies to the reference. Figure 13 shows an illustrative example of a region in the reference genome where reads from *P. nodorum* IR10_2.1a fail to map. The region without mapped reads covers 20 kb and eight reference genes. A protocol that calculates the gene absence from mapping data alone would describe this as an 8 locus sectional absence. However, using nucmer to align *de novo* assembled sequences of this strain to the reference reveals that the absence is only 8 kb long and that only two genes are truly absent from the alternate strain. The surrounding regions are heavily mutated, but are still present.

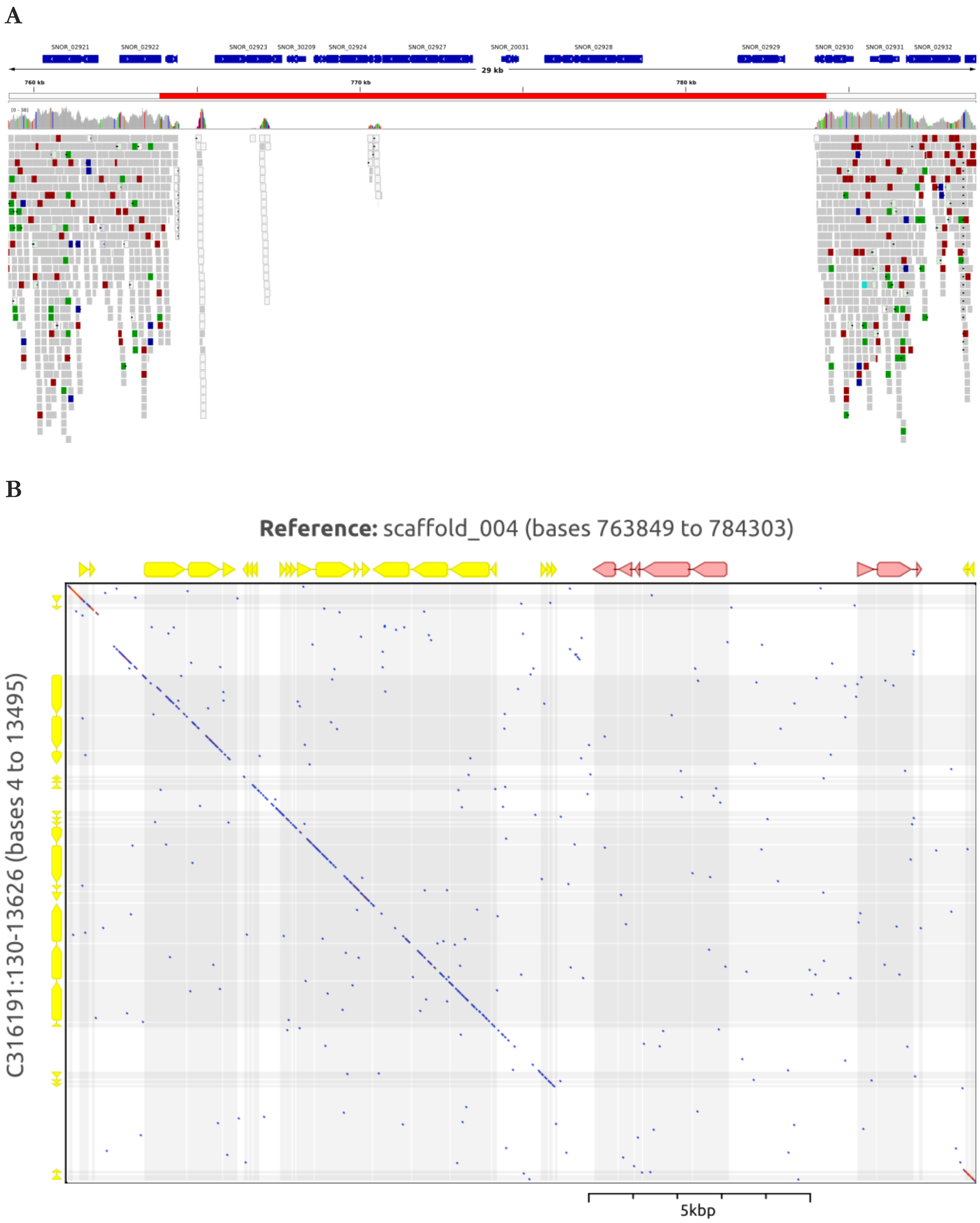


Figure 13: (A) Mapped read depth of a region on scaffold_004 in the SN15 reference assembly shows a putative sectional absence of seven genes. (B) Dotplot of the alternate strain's (*P. nodorum* IR10_2.1a) *de-novo* assembly at the region (marked red in A) shows that only two of the reference genes (marked in pink) are absent in the alternate strain. Highly variable regions around sectional absences can frustrate mapping algorithms leading to an inflated estimation of absent genes.

Whole-genome resequencing and assembly provides the opportunity to calculate powerful codon-based tests of positive selection at almost every locus. Frequency of positive selection allows us to divide the genome into three classes. More than 96% of the 100 kb windows surveyed had fewer than 20% of the genes in that window under positive selection. In a second class, there are small islands of consecutive windows with at least 20% of the genes in the windows under positive selection. Scaffolds 7, 15, and 20 include one 'positive selection island' each (Figure 7,

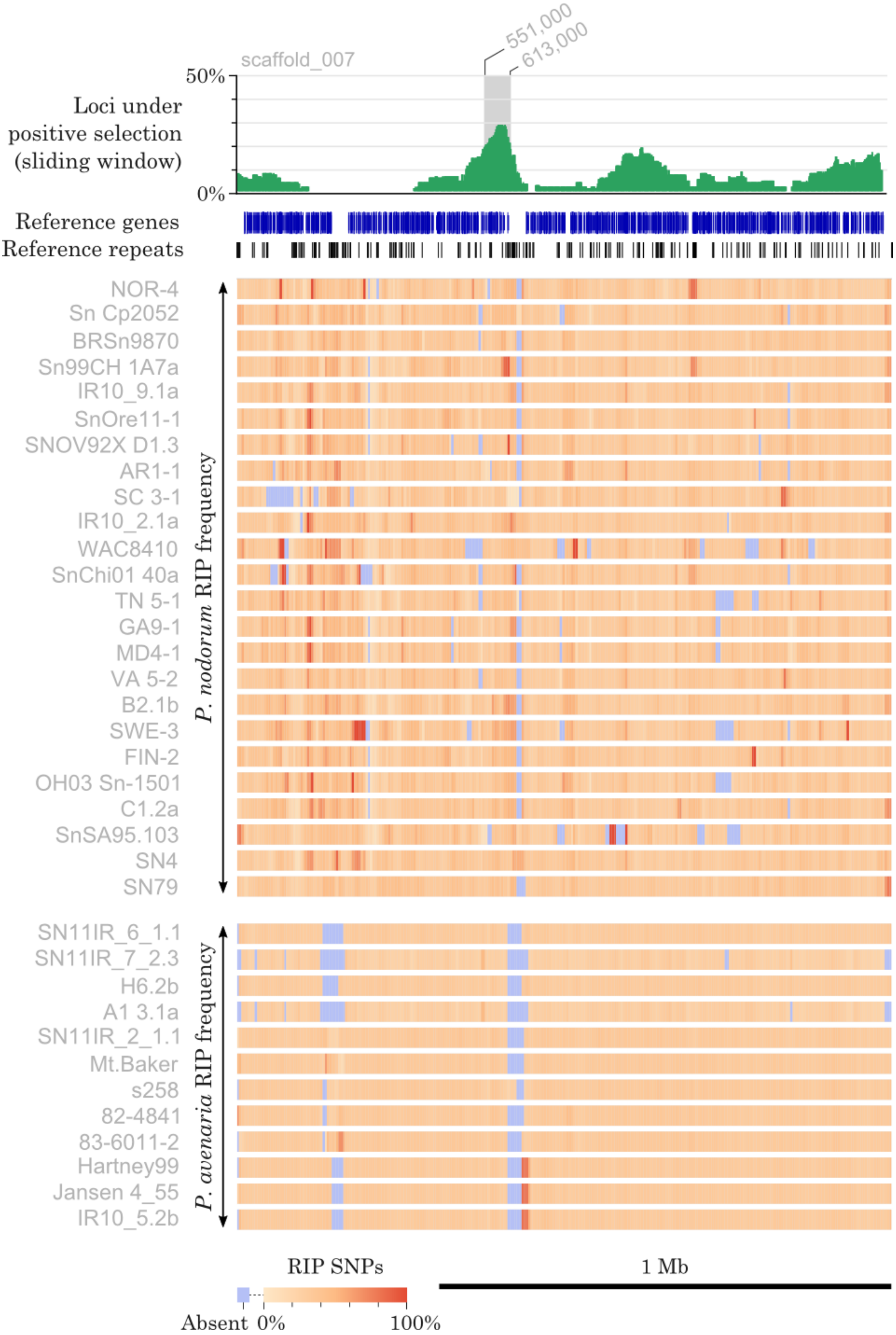


Figure 7 Figure 8, and Figure 9). Lastly, scaffolds 44 and 45 elevated levels of positive selection over their length. Islands of positive selection appear adjacent to repetitive religions or repetitive

scaffold ends (Figure 10). RIP is known to 'leak' from repetitive regions into single-copy genes with the effect of disabling AVR genes (Fudal, Ross et al. 2009, Rouxel, Grandaubert et al. 2011). In the case of AVR, a tolerable level of indiscriminate mutation is beneficial to the pathogen. Any mutation that provides diversity or deletion at the AVR locus would allow the pathogen to grow unhindered on a host with the corresponding resistance gene. However, the potential for RIP to drive positive selection at a non-AVR locus may be less likely, as there would be fewer possible mutations that could confer a fitness advantage. Nevertheless, islands of positive selection adjacent to repetitive regions may suggest some association between RIP and diversity at non-AVR loci in *P. nodorum*.

Syntenic variability was also examined with a focus on the extent and nature of mesosyntenic genome rearrangements. Mesosyntenic conservation of chromosomal content may provide an explanation for the stability of accessory chromosomes, as the low frequency of inter-chromosomal translocations would ensure that the core genome would be unlikely to leak into the accessory chromosome and vice-versa. In *P. nodorum* and *P. avenaria*, long DNA transposons and LTR transposons are frequently found at small relative distance to mesosyntenic breakpoints, and simple repeats are no more likely to be found at low relative distance than high relative distance. These findings support the mechanism suggested by Grandaubert (2013) of mesosyntenic inversions aided by TEs and not SSR (Ohm, Feau et al. 2012). Both this study and Grandaubert's were observing mesosynteny between closely related species, whereas Ohm's comparisons were made across much larger evolutionary distances. As species drift apart, the effects of RIP and normal background mutation may work to obfuscate the association between TEs and mesosyntenic breakpoints.

The addition of new genome assemblies also allows for an expanded set of test criteria for selection of effector candidates. Presence/absence allele frequency in the *P. nodorum* population differs for each effector, and is likely driven by the prevalence of each effector's susceptibility gene in the host where the isolate was sampled (McDonald, Oliver et al. 2013). The common characteristic shared by SnToxA, SnTox1 and SnTox3 is that no effector is present in all populations and no effector is rare (Figure 1). A more accurate set of core and strain-specific proteins (Figure 11, Table 3) allow us to target genes that are neither perfectly conserved nor infrequently occurring. Assessed by the new criteria, SnTox1 is the top-ranked protein, SnTox3 and SnToxA are equal-fifth (Table 6, Appendix A6-4). There are 30 genes that rank as well or better than the known effectors. These top-scoring candidates will be prioritised for purification in a heterologous expression system and screened against wheat lines to test for the effector's ability to produce disease symptoms. Once validated, effector molecules can be applied as tools to accelerate disease resistance breeding programs (Vleeshouwers and Oliver 2014).

Investigations of the *P. nodorum* and *P. avenaria* genomes have allowed us to observe variation at a variety of scales. Large scale variation of genomic and chromosomal structure has demonstrated the association of long repeats with mesosyntenic recombination, and of regions of non-core accessory elements. *De novo* assembly comparison has highlighted the large number of strain-specific loci and extent of presence/absence variation within the two species. At the smallest resolution, the genomic comparisons resolve at fine detail such as the detection of positive selection at individual codons and RIP-like SNPs. Each of these observations contributes to an understanding of the genomic history of these two species and to the prediction of potential effector sequences in *P. nodorum*.

Acknowledgements

This work was supported through access to facilities managed by Bioplatforms Australia and funded by the Australian Government National Collaborative Research Infrastructure Strategy and Education Investment Fund Super Science Initiative and Grains Research and Development Corporation research grant CUR00012. 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 and by iVEC through the use of advanced computing resources located at the Pawsey Supercomputing Centre. RAS received funding through scholarships from GRDC (GRS10061) and Curtin University.

References

Croll, D. and B. A. McDonald (2012). "The accessory genome as a cradle for adaptive evolution in pathogens." PLoS pathogens **8**(4): e1002608.

Favorov, A., L. Mularoni, L. M. Cope, Y. Medvedeva, A. A. Mironov, V. J. Makeev and S. J. Wheelan (2012). "Exploring massive, genome scale datasets with the GenometriCorr package." PLoS computational biology **8**(5): e1002529.

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

Fudal, I., S. Ross, H. Brun, A.-L. Besnard, M. Ermel, M.-L. Kuhn, M.-H. Balesdent and T. Rouxel (2009). "Repeat-induced point mutation (RIP) as an alternative mechanism of evolution toward virulence in Leptosphaeria maculans." Molecular Plant-Microbe Interactions **22**(8): 932-941.

Grandaubert, J. (2013). Génomique comparative et évolutive au sein du complexe d'espèces *Leptosphaeria maculans-Leptosphaeria biglobosa*, Paris 11.

McDonald, M. C., R. P. Oliver, T. L. Friesen, P. C. Brunner and B. A. McDonald (2013). "Global diversity and distribution of three necrotrophic effectors in Phaeosphaeria nodorum and related species." New Phytologist **199**(1): 241-251.

Ohm, R. A., N. Feau, B. Henrissat, C. L. Schoch, B. A. Horwitz, K. W. Barry, B. J. Condon, A. C. Copeland, B. Dhillon and F. Glaser (2012). "Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi." PLoS Pathogens **8**(12): e1003037.

Rech, G. E., J. M. Sanz-Martín, M. Anisimova, S. A. Sukno and M. R. Thon (2014). "Natural Selection on Coding and Noncoding DNA Sequences Is Associated with Virulence Genes in a Plant Pathogenic Fungus." Genome biology and evolution **6**(9): 2368-2379.

Rouxel, T., J. Grandaubert, J. K. Hane, C. Hoede, A. P. van de Wouw, A. Couloux, V. Dominguez, V. Anthouard, P. Bally and S. Bourras (2011). "Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations." Nature Communications **2**: 202.

Stukenbrock, E. H., T. Bataillon, J. Y. Dutheil, T. T. Hansen, R. Li, M. Zala, B. A. McDonald, J. Wang and M. H. Schierup (2011). "The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen Mycosphaerella graminicola and its wild sister species." Genome research **21**(12): 2157-2166.

Syme, R. A., J. K. Hane, T. L. Friesen and R. P. Oliver (2013). "Resequencing and comparative genomics of *Stagonospora nodorum*: Sectional gene absence and effector discovery." G3: Genes| Genomes| Genetics **3**(6): 959-969.

Vleeshouwers, V. G. and R. P. Oliver (2014). "Effectors as tools in disease resistance breeding against biotrophic, hemibiotrophic, and necrotrophic plant pathogens." Molecular Plant-Microbe Interactions **27**(3): 196-206.

Chapter 7 | Conclusion

Model organism platforms

An important by-product of the analysis conducted in this thesis is the generation of omics resources that are used to inform experimental design for other researchers working with these pathogens. The *P. teres f. teres* genome sequence was the first short-read assembly of an ascomycete genome, and showed that short-read assemblies provided valuable insight into fungal genomes. The genome has provided a very effective platform for the design of markers (Chapter 3) and comparative genomics.

The correction of more than 1,000 deletions, 12,000 SNPs and 16,000 small insertions in the *P. nodorum* reference genome SN15 (Chapter 5) improves the accuracy for all subsequent comparative analyses, including those in this thesis. Many important methods rely on the accuracy of assembly sequence and genome annotations. Methods of gene prediction are particularly sensitive to insertion or deletion errors in the underlying genome sequence. Effector prediction is an important example of a method sensitive to annotation accuracy. Previously truncated annotations had prevented the detection of 447 instances of genes encoding signal peptides (Chapter 5). All methods of effector prediction also rely on the gene being annotated. Integration of RNA-seq, proteogenomic and microarray data informed the annotation of 866 new genes. Many of the new genes are small secreted proteins and nine of the top thirty effector candidates are genes are new loci (Chapter 6). The SN15 effector candidate list was an important component in the discovery of Tox1, and we expect that improvements to the method will uncover further necrotrophic effectors.

Inter-species and intra-species comparisons

Comparison of known effector loci in alternate *P. nodorum* strains showed that regions flanking the Tox1 and Tox3 loci in the reference strain are present when the effector is absent. This establishes that these particular effectors are unlikely to be moved together with large number of supporting loci. Sequence similarity of the ToxA region in P. tritici-repentis and scaffolds in *P. nodorum* SN15 provided support for the hypothesis that ToxA was transferred from *P. nodorum* to *P. tritici-repentis* as part of a much larger transfercon than originally expected. This hypothesis also suggests order and orientation information for the putatively homologous SN15 scaffolds. The predicted joining of scaffolds 51 and 55 was validated by the process of SN15 assembly correction (Chapter 5).

A putative accessory chromosome has been uncovered in scaffolds 44 and 45. They exhibit a much higher frequency of genes under positive selection, and a unique pattern of presence/absence between the *P. nodorum* and *P. avenaria* strains. The scaffolds are well conserved in all *P. nodorum* isolates that are able to infect wheat, but are absent from the oat pathogen *P. avenaria* and the *P. nodorum* strain isolated from wild grasses and avirulent on wheat (Chapter 6). Other islands of positive selection were observed outside of these putative accessory elements. Scaffolds 7, 15 and 20 all have regions with a greater density of genes under positive selection (Chapter 6). In each case, the island of positive selection is adjacent to repetitive sequence or a scaffold end. This suggests that repetitive sequence is being used to drive positive selection at some loci. RIP is known to leak out of the target repetitive sequence and affect nearby single-copy sequence. Mutations as a result of RIP leakage have been observed in avirulence genes where disruptive mutations allow the pathogen to avoid host detection by the corresponding R gene. It has not yet been shown that RIP leakage could drive mutations in genes without the loss of their function, but inspection of these regions may demonstrate this.

Positive selection and patterns of presence/absence genotypes across the *P. nodorum* strains were used to augment criteria used for effector detection in the reference strain SN15. The availability of dozens of alleles for each locus provided the sample sizes sufficient large for stronger tests of positive selection, the results of which are fed into the effector prediction pipeline (Chapter 6). Known effectors show an unusual distribution of presence/absence alleles. The vast majority of *P. nodorum* and *P. avenaria* proteins are either strain-specific or well conserved. In contrast, the known effectors are all present in the middle-ground between strain-specificity and core proteome. Each of observations of intra and inter-specific variation were incorporated into the SN15 effector prediction pipeline.

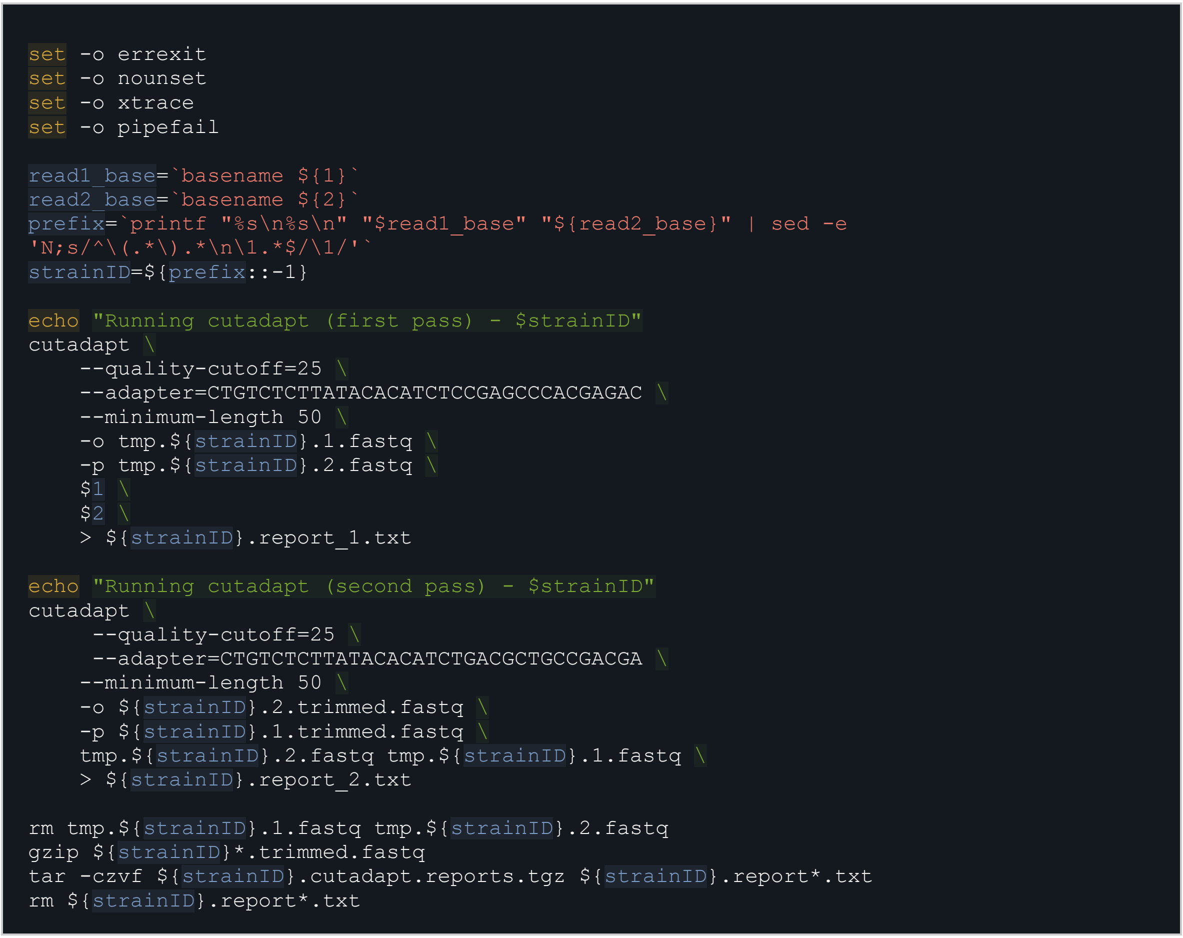
The analysis of the *Pyrenophora* and *Parastagonospora* genomes presented here improve our understanding of the structural genome dynamics of the organisms (Chapter 6), their inter-species evolutionary history (Chapter 4), the characterisation of the pan and core genome (Chapters 3 and 5) by intra-specific comparison and the identification of effector candidates (Chapters 2 and 6). The effector candidate lists presented here have already contributed to the discovery of the *P. nodorum* effector Tox1, and it is expected that others effectors will be uncovered as candidates presented in Chapter 6 are expressed and purified. The research has also significantly expanded the resources available to others working on these pathosystems.

Appendices

Chapter 5 Appendices

A5-1 | Cutadapt Parameters

Cutadapt was run trimming at quality cutoff 25, using known Illumina adapters and discarding trimmed reads where the final length was less than 50:



```
set -o errexit
set -o nounset
set -o xtrace
set -o pipefail

read1_base=`basename ${1}`
read2_base=`basename ${2}`
prefix=`printf "%s\n%s\n" "$read1_base" "${read2_base}" | sed -e
'N;s/^\(.*\).*\n\1.*$/\1/'`
strainID=${prefix::-1}

echo "Running cutadapt (first pass) - $strainID"
cutadapt \
    --quality-cutoff=25 \
    --adapter=CTGTCTCTTATACACATCTCCGAGCCCACGAGAC \
    --minimum-length 50 \
    -o tmp.${strainID}.1.fastq \
    -p tmp.${strainID}.2.fastq \
    $1 \
    $2 \
    > ${strainID}.report_1.txt

echo "Running cutadapt (second pass) - $strainID"
cutadapt \
     --quality-cutoff=25 \
     --adapter=CTGTCTCTTATACACATCTGACGCTGCCGACGA \
    --minimum-length 50 \
    -o ${strainID}.2.trimmed.fastq \
    -p ${strainID}.1.trimmed.fastq \
    tmp.${strainID}.2.fastq tmp.${strainID}.1.fastq \
    > ${strainID}.report_2.txt

rm tmp.${strainID}.1.fastq tmp.${strainID}.2.fastq
gzip ${strainID}*.trimmed.fastq
tar -czvf ${strainID}.cutadapt.reports.tgz ${strainID}.report*.txt
rm ${strainID}.report*.txt
```

A5-2 | Repeat Content

The repeat content did not substantially differ between genome assemblies

Repeat class	Old count	New count	Delta	Delta%
Subtelomeric				
R22	12252	12645	393	3.20764
X15	87189	87249	60	0.068816
X26	76622	77179	557	0.726945
X35	14136	14135	-1	-0.00707414
X48	5377	5593	216	4.01711
Ribosomal				
Y1	400707	400875	168	0.0419259
Other				
ELSA	34285	34319	34	0.0991687
MOLLY	49213	49296	83	0.168655
PIXIE	38612	38635	23	0.059567
R10	43944	43835	-109	-0.248043
R25	43858	44107	249	0.567741
R31	39053	39429	376	0.962794
R37	106258	106392	134	0.126108
R38	8760	8827	67	0.76484
R39	36613	36778	165	0.45066
R51	25640	25863	223	0.869735
R8	277650	277645	-5	-0.00180083
R9	163980	162676	-1304	-0.795219
X0	149537	147850	-1687	-1.12815
X11	126539	126536	-3	-0.00237081
X12	24813	24909	96	0.386894
X23	12354	12348	-6	-0.0485673
X28	28414	28248	-166	-0.584219
X3	464053	463844	-209	-0.045038
X36	5067	5107	40	0.789422
X46	1315	1315	0	0
X96	4320	4311	-9	-0.208333
Sum	2280561	2279946	-615	10.239207

A5-3 | List of genes that are a product of merging to genes

SNOG_30955	SNOG_30946	SNOG_30935	SNOG_30934
SNOG_30930	SNOG_30889	SNOG_30841	SNOG_30798
SNOG_30795	SNOG_30785	SNOG_30783	SNOG_30773
SNOG_30761	SNOG_30760	SNOG_30747	SNOG_30727
SNOG_30715	SNOG_30704	SNOG_30693	SNOG_30688
SNOG_30682	SNOG_30649	SNOG_30643	SNOG_30629
SNOG_30627	SNOG_30611	SNOG_30605	SNOG_30602
SNOG_30583	SNOG_30581	SNOG_30578	SNOG_30574
SNOG_30573	SNOG_30568	SNOG_30557	SNOG_30556
SNOG_30549	SNOG_30546	SNOG_30537	SNOG_30536
SNOG_30534	SNOG_30488	SNOG_30483	SNOG_30473
SNOG_30472	SNOG_30449	SNOG_30372	SNOG_30366
SNOG_30357	SNOG_30342	SNOG_30324	SNOG_30271
SNOG_30247	SNOG_30094	SNOG_30040	

A5-4 | List of genes that are a product of splitting one annotation into two

SNOG_30991	SNOG_30987	SNOG_30986	SNOG_30984
SNOG_30983	SNOG_30982	SNOG_30981	SNOG_30980
SNOG_30979	SNOG_30978	SNOG_30977	SNOG_30975
SNOG_30965	SNOG_30963	SNOG_30961	SNOG_30954
SNOG_30953	SNOG_30951	SNOG_30950	SNOG_30949
SNOG_30947	SNOG_30942	SNOG_30938	SNOG_30933
SNOG_30931	SNOG_30929	SNOG_30928	SNOG_30927
SNOG_30926	SNOG_30924	SNOG_30922	SNOG_30918
SNOG_30917	SNOG_30916	SNOG_30915	SNOG_30914
SNOG_30913	SNOG_30912	SNOG_30911	SNOG_30910
SNOG_30909	SNOG_30904	SNOG_30898	SNOG_30897
SNOG_30896	SNOG_30895	SNOG_30894	SNOG_30892
SNOG_30891	SNOG_30890	SNOG_30886	SNOG_30883
SNOG_30882	SNOG_30879	SNOG_30877	SNOG_30876
SNOG_30875	SNOG_30874	SNOG_30873	SNOG_30872
SNOG_30871	SNOG_30870	SNOG_30868	SNOG_30867
SNOG_30866	SNOG_30865	SNOG_30860	SNOG_30859
SNOG_30857	SNOG_30856	SNOG_30855	SNOG_30853
SNOG_30847	SNOG_30844	SNOG_30843	SNOG_30840
SNOG_30832	SNOG_30830	SNOG_30829	SNOG_30827
SNOG_30826	SNOG_30811	SNOG_30809	SNOG_30808
SNOG_30805	SNOG_30804	SNOG_30800	SNOG_30799
SNOG_30797	SNOG_30796	SNOG_30792	SNOG_30787
SNOG_30786	SNOG_30782	SNOG_30776	SNOG_30770
SNOG_30774	SNOG_30771	SNOG_30768	SNOG_30767
SNOG_30757	SNOG_30756	SNOG_30755	SNOG_30754
SNOG_30753	SNOG_30752	SNOG_30750	SNOG_30749
SNOG_30746	SNOG_30745	SNOG_30744	SNOG_30743
SNOG_30742	SNOG_30740	SNOG_30738	SNOG_30736
SNOG_30735	SNOG_30733	SNOG_30732	SNOG_30730
SNOG_30729	SNOG_30724	SNOG_30720	SNOG_30718
SNOG_30714	SNOG_30713	SNOG_30712	SNOG_30711
SNOG_30709	SNOG_30700	SNOG_30694	SNOG_30690
SNOG_30689	SNOG_30687	SNOG_30681	SNOG_30676
SNOG_30665	SNOG_30663	SNOG_30656	SNOG_30653
SNOG_30652	SNOG_30650	SNOG_30642	SNOG_30640
SNOG_30636	SNOG_30634	SNOG_30626	SNOG_30624
SNOG_30623	SNOG_30622	SNOG_30620	SNOG_30618
SNOG_30617	SNOG_30616	SNOG_30613	SNOG_30606
SNOG_30604	SNOG_30600	SNOG_30593	SNOG_30591
SNOG_30590	SNOG_30587	SNOG_30579	SNOG_30576
SNOG_30575	SNOG_30572	SNOG_30567	SNOG_30562
SNOG_30555	SNOG_30554	SNOG_30552	SNOG_30548
SNOG_30547	SNOG_30541	SNOG_30540	SNOG_30539
SNOG_30527	SNOG_30520	SNOG_30519	SNOG_30514
SNOG_30513	SNOG_30512	SNOG_30511	SNOG_30505
SNOG_30500	SNOG_30496	SNOG_30495	SNOG_30492
SNOG_30491	SNOG_30475	SNOG_30471	SNOG_30469
SNOG_30462	SNOG_30460	SNOG_30457	SNOG_30456
SNOG_30455	SNOG_30454	SNOG_30448	SNOG_30447
SNOG_30446	SNOG_30441	SNOG_30438	SNOG_30431
SNOG_30430	SNOG_30428	SNOG_30426	SNOG_30425
SNOG_30422	SNOG_30421	SNOG_30417	SNOG_30416

SNOG_30412	SNOG_30409	SNOG_30402	SNOG_30401
SNOG_30400	SNOG_30394	SNOG_30393	SNOG_30392
SNOG_30391	SNOG_30389	SNOG_30387	SNOG_30377
SNOG_30376	SNOG_30373	SNOG_30370	SNOG_30368
SNOG_30356	SNOG_30353	SNOG_30351	SNOG_30336
SNOG_30333	SNOG_30330	SNOG_30319	SNOG_30315
SNOG_30313	SNOG_30312	SNOG_30302	SNOG_30299
SNOG_30295	SNOG_30292	SNOG_30291	SNOG_30290
SNOG_30286	SNOG_30282	SNOG_30280	SNOG_30279
SNOG_30277	SNOG_30276	SNOG_30275	SNOG_30274
SNOG_30272	SNOG_30267	SNOG_30263	SNOG_30261
SNOG_30260	SNOG_30257	SNOG_30256	SNOG_30252
SNOG_30246	SNOG_30245	SNOG_30242	SNOG_30241
SNOG_30240	SNOG_30239	SNOG_30238	SNOG_30237
SNOG_30236	SNOG_30231	SNOG_30230	SNOG_30223
SNOG_30218	SNOG_30217	SNOG_30213	SNOG_30212
SNOG_30211	SNOG_30206	SNOG_30202	SNOG_30192
SNOG_30188	SNOG_30185	SNOG_30179	SNOG_30175
SNOG_30992	SNOG_30171	SNOG_30170	SNOG_30169
SNOG_30167	SNOG_30164	SNOG_30158	SNOG_30155
SNOG_30154	SNOG_30153	SNOG_30152	SNOG_30151
SNOG_30149	SNOG_30147	SNOG_30143	SNOG_30139
SNOG_30138	SNOG_30131	SNOG_30130	SNOG_30129
SNOG_30126	SNOG_30125	SNOG_30122	SNOG_30118
SNOG_30116	SNOG_30113	SNOG_30110	SNOG_30109
SNOG_30108	SNOG_30107	SNOG_30106	SNOG_30105
SNOG_30103	SNOG_30097	SNOG_30091	SNOG_30090
SNOG_30089	SNOG_30087	SNOG_30086	SNOG_30084
SNOG_30082	SNOG_30081	SNOG_30080	SNOG_30075
SNOG_30071	SNOG_30069	SNOG_30068	SNOG_30066
SNOG_30062	SNOG_30061	SNOG_30060	SNOG_30059
SNOG_30058	SNOG_30057	SNOG_30056	SNOG_30053
SNOG_30052	SNOG_30050	SNOG_30049	SNOG_30047
SNOG_30045	SNOG_30043	SNOG_30042	SNOG_30041
SNOG_30032	SNOG_30028	SNOG_30027	SNOG_30020
SNOG_30017	SNOG_30016	SNOG_30012	SNOG_30011

A5-5 | Genes at new loci

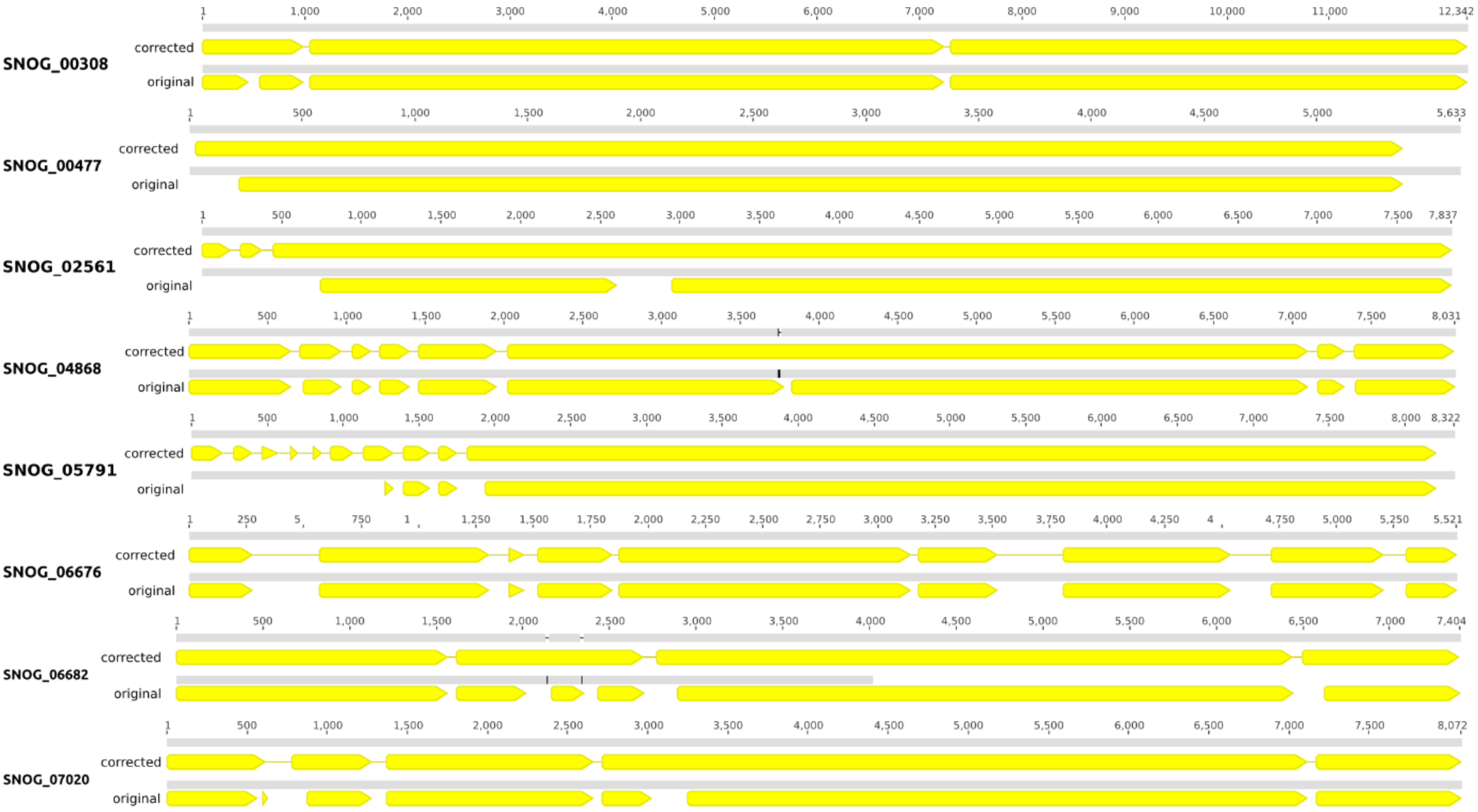
SNOG_30993	SNOG_30990	SNOG_30989	SNOG_30988
SNOG_30976	SNOG_30974	SNOG_30973	SNOG_30972
SNOG_30971	SNOG_30968	SNOG_30967	SNOG_30966
SNOG_30964	SNOG_30962	SNOG_30960	SNOG_30959
SNOG_30957	SNOG_30956	SNOG_30952	SNOG_30948
SNOG_30945	SNOG_30944	SNOG_30941	SNOG_30940
SNOG_30937	SNOG_30936	SNOG_30932	SNOG_30923
SNOG_30921	SNOG_30920	SNOG_30919	SNOG_30908
SNOG_30907	SNOG_30906	SNOG_30905	SNOG_30903
SNOG_30902	SNOG_30901	SNOG_30900	SNOG_30899
SNOG_30994	SNOG_30893	SNOG_30888	SNOG_30887
SNOG_30884	SNOG_30880	SNOG_30878	SNOG_30869
SNOG_30864	SNOG_30863	SNOG_30862	SNOG_30861
SNOG_30858	SNOG_30854	SNOG_30852	SNOG_30851
SNOG_30850	SNOG_30848	SNOG_30846	SNOG_30845
SNOG_30842	SNOG_30838	SNOG_30837	SNOG_30836
SNOG_30835	SNOG_30834	SNOG_30833	SNOG_30831
SNOG_30828	SNOG_30825	SNOG_30824	SNOG_30822
SNOG_30821	SNOG_30819	SNOG_30818	SNOG_30810
SNOG_30807	SNOG_30806	SNOG_30803	SNOG_30802
SNOG_30801	SNOG_30794	SNOG_30793	SNOG_30791
SNOG_30790	SNOG_30788	SNOG_30784	SNOG_30781
SNOG_30780	SNOG_30779	SNOG_30778	SNOG_30777
SNOG_30772	SNOG_30769	SNOG_30766	SNOG_30765
SNOG_30764	SNOG_30763	SNOG_30762	SNOG_30759
SNOG_30758	SNOG_30751	SNOG_30748	SNOG_30741
SNOG_30739	SNOG_30731	SNOG_30728	SNOG_30725
SNOG_30723	SNOG_30722	SNOG_30721	SNOG_30719
SNOG_30717	SNOG_30716	SNOG_30710	SNOG_30708
SNOG_30707	SNOG_30706	SNOG_30705	SNOG_30703
SNOG_30701	SNOG_30699	SNOG_30698	SNOG_30697
SNOG_30696	SNOG_30695	SNOG_30692	SNOG_30691
SNOG_30686	SNOG_30684	SNOG_30679	SNOG_30678
SNOG_30677	SNOG_30675	SNOG_30674	SNOG_30673
SNOG_30672	SNOG_30671	SNOG_30670	SNOG_30669
SNOG_30668	SNOG_30667	SNOG_30666	SNOG_30664
SNOG_30662	SNOG_30661	SNOG_30659	SNOG_30658
SNOG_30657	SNOG_30654	SNOG_30648	SNOG_30647
SNOG_30646	SNOG_30645	SNOG_30637	SNOG_30633
SNOG_30632	SNOG_30631	SNOG_30630	SNOG_30625
SNOG_30621	SNOG_30619	SNOG_30615	SNOG_30614
SNOG_30612	SNOG_30610	SNOG_30609	SNOG_30608
SNOG_30607	SNOG_30603	SNOG_30601	SNOG_30599
SNOG_30597	SNOG_30596	SNOG_30595	SNOG_30594
SNOG_30592	SNOG_30589	SNOG_30588	SNOG_30585
SNOG_30584	SNOG_30582	SNOG_30577	SNOG_30571
SNOG_30569	SNOG_30564	SNOG_30563	SNOG_30560
SNOG_30559	SNOG_30558	SNOG_30553	SNOG_30550
SNOG_30545	SNOG_30544	SNOG_30543	SNOG_30542
SNOG_30532	SNOG_30530	SNOG_30529	SNOG_30528
SNOG_30526	SNOG_30525	SNOG_30524	SNOG_30523
SNOG_30522	SNOG_30518	SNOG_30516	SNOG_30515
SNOG_30510	SNOG_30509	SNOG_30508	SNOG_30507

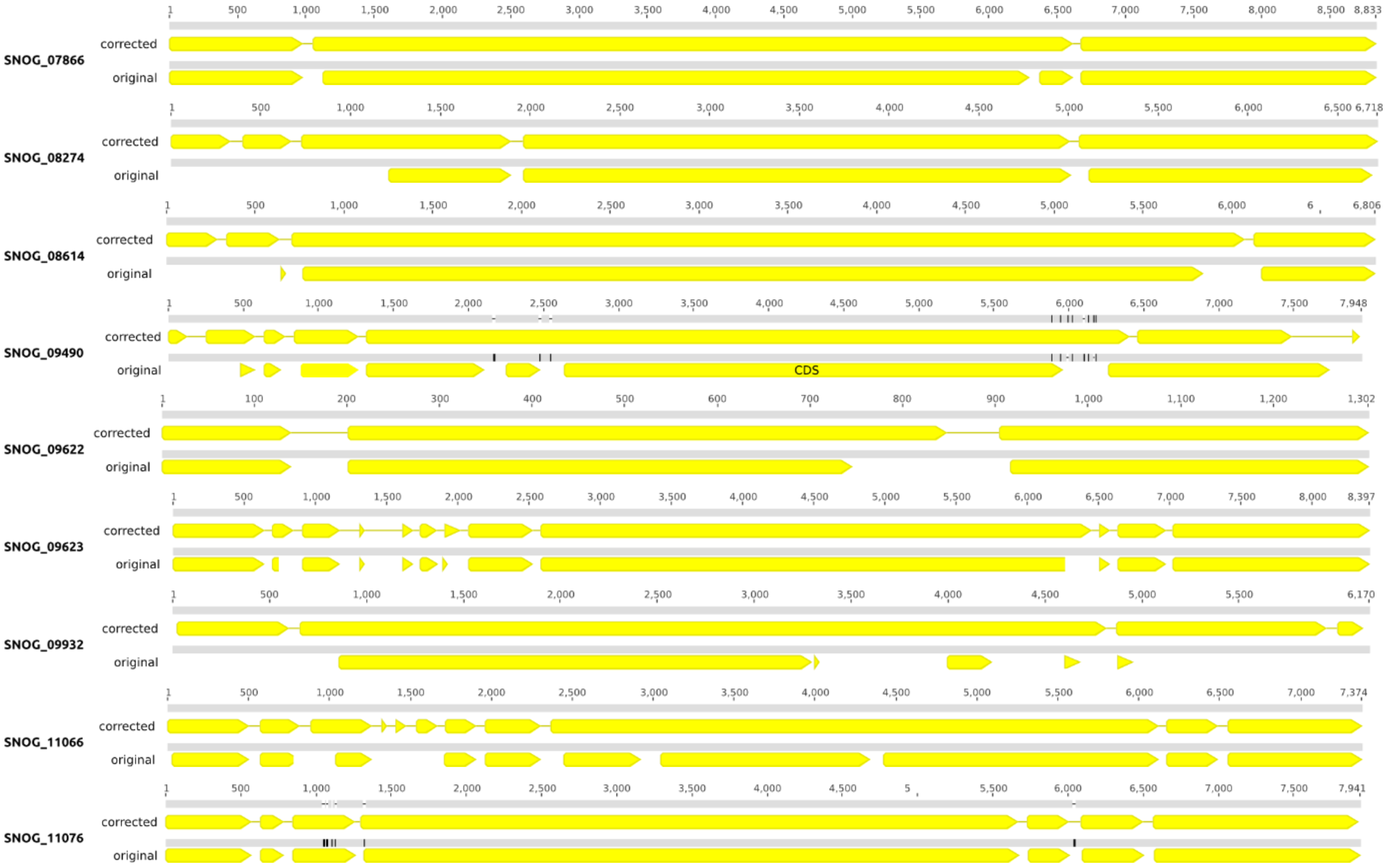
SNOG_30504	SNOG_30503	SNOG_30502	SNOG_30501
SNOG_30499	SNOG_30498	SNOG_30497	SNOG_30494
SNOG_30493	SNOG_30490	SNOG_30489	SNOG_30487
SNOG_30486	SNOG_30485	SNOG_30484	SNOG_30482
SNOG_30481	SNOG_30480	SNOG_30479	SNOG_30478
SNOG_30477	SNOG_30476	SNOG_30474	SNOG_30468
SNOG_30467	SNOG_30466	SNOG_30465	SNOG_30464
SNOG_30463	SNOG_30461	SNOG_30459	SNOG_30453
SNOG_30452	SNOG_30451	SNOG_30450	SNOG_30445
SNOG_30442	SNOG_30440	SNOG_30439	SNOG_30436
SNOG_30434	SNOG_30433	SNOG_30432	SNOG_30429
SNOG_30424	SNOG_30423	SNOG_30420	SNOG_30419
SNOG_30418	SNOG_30415	SNOG_30413	SNOG_30410
SNOG_30407	SNOG_30406	SNOG_30405	SNOG_30403
SNOG_30399	SNOG_30398	SNOG_30396	SNOG_30395
SNOG_30390	SNOG_30388	SNOG_30386	SNOG_30385
SNOG_30384	SNOG_30383	SNOG_30382	SNOG_30380
SNOG_30379	SNOG_30378	SNOG_30374	SNOG_30369
SNOG_30367	SNOG_30365	SNOG_30364	SNOG_30363
SNOG_30362	SNOG_30361	SNOG_30360	SNOG_30359
SNOG_30358	SNOG_30355	SNOG_30354	SNOG_30352
SNOG_30350	SNOG_30349	SNOG_30348	SNOG_30347
SNOG_30346	SNOG_30344	SNOG_30343	SNOG_30341
SNOG_30340	SNOG_30339	SNOG_30337	SNOG_30335
SNOG_30334	SNOG_30332	SNOG_30331	SNOG_30329
SNOG_30328	SNOG_30327	SNOG_30326	SNOG_30325
SNOG_30323	SNOG_30322	SNOG_30321	SNOG_30320
SNOG_30318	SNOG_30317	SNOG_30316	SNOG_30314
SNOG_30311	SNOG_30310	SNOG_30309	SNOG_30308
SNOG_30307	SNOG_30306	SNOG_30305	SNOG_30304
SNOG_30303	SNOG_30300	SNOG_30297	SNOG_30294
SNOG_30289	SNOG_30288	SNOG_30287	SNOG_30285
SNOG_30284	SNOG_30283	SNOG_30281	SNOG_30278
SNOG_30273	SNOG_30270	SNOG_30269	SNOG_30268
SNOG_30266	SNOG_30265	SNOG_30264	SNOG_30258
SNOG_30255	SNOG_30253	SNOG_30250	SNOG_30249
SNOG_30248	SNOG_30244	SNOG_30235	SNOG_30234
SNOG_30233	SNOG_30232	SNOG_30229	SNOG_30228
SNOG_30227	SNOG_30225	SNOG_30224	SNOG_30221
SNOG_30220	SNOG_30219	SNOG_30216	SNOG_30215
SNOG_30214	SNOG_30210	SNOG_30209	SNOG_30208
SNOG_30207	SNOG_30205	SNOG_30204	SNOG_30203
SNOG_30201	SNOG_30200	SNOG_30199	SNOG_30198
SNOG_30197	SNOG_30196	SNOG_30195	SNOG_30194
SNOG_30193	SNOG_30190	SNOG_30186	SNOG_30182
SNOG_30180	SNOG_30178	SNOG_30177	SNOG_30176
SNOG_30174	SNOG_30157	SNOG_30148	SNOG_30145
SNOG_30137	SNOG_30134	SNOG_30132	SNOG_30128
SNOG_30124	SNOG_30123	SNOG_30121	SNOG_30120
SNOG_30119	SNOG_30115	SNOG_30114	SNOG_30112
SNOG_30111	SNOG_30104	SNOG_30102	SNOG_30101
SNOG_30100	SNOG_30099	SNOG_30098	SNOG_30096
SNOG_30095	SNOG_30093	SNOG_30092	SNOG_30085
SNOG_30079	SNOG_30078	SNOG_30077	SNOG_30074

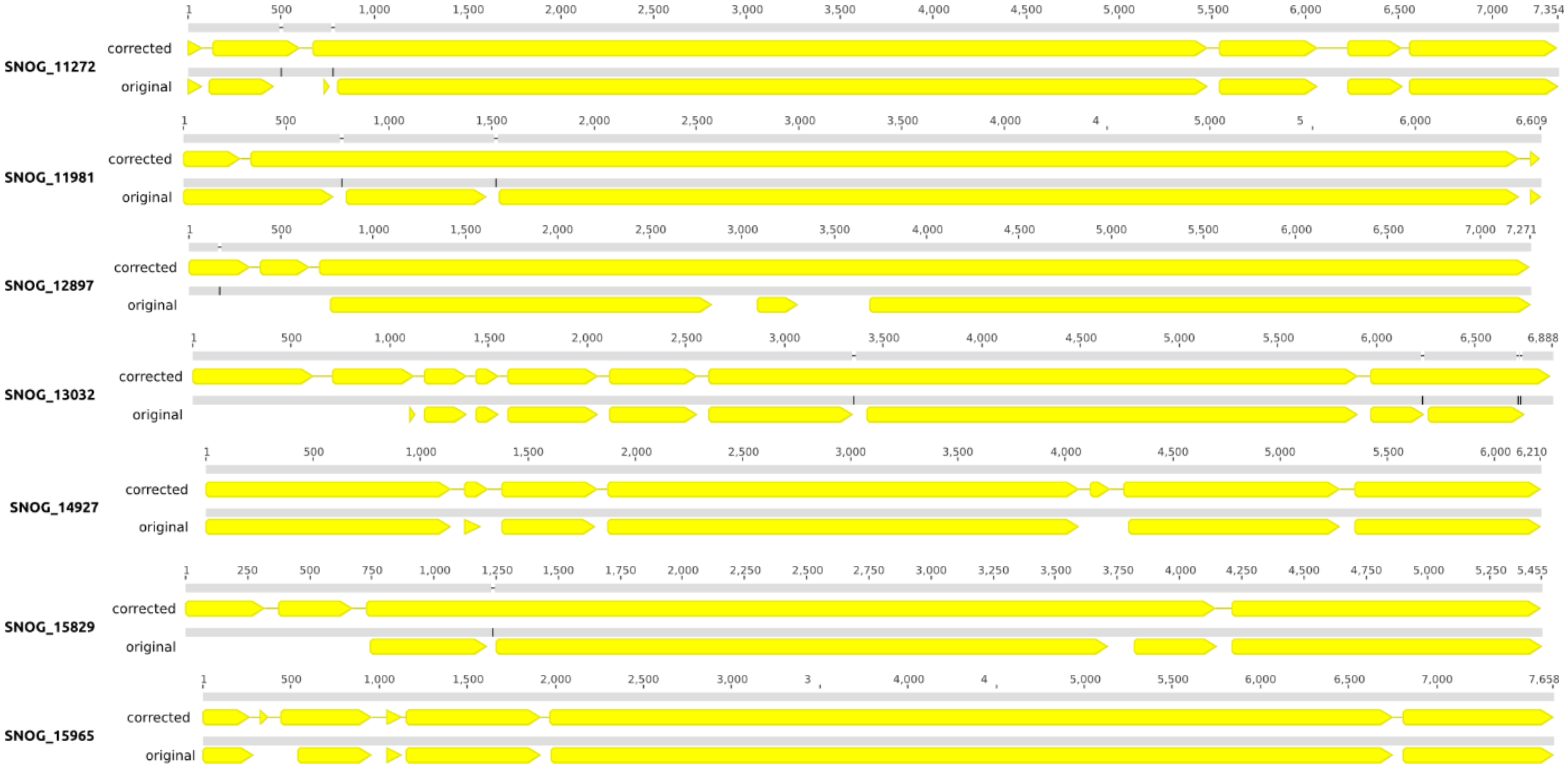
SNOG_30073	SNOG_30072	SNOG_30070	SNOG_30067
SNOG_30065	SNOG_30064	SNOG_30063	SNOG_30054
SNOG_30048	SNOG_30046	SNOG_30044	SNOG_30039
SNOG_30036	SNOG_30035	SNOG_30034	SNOG_30033
SNOG_30031	SNOG_30030	SNOG_30029	SNOG_30026
SNOG_30025	SNOG_30024	SNOG_30022	SNOG_30019
SNOG_30014	SNOG_30013	SNOG_30008	

A5-6 | PKS gene models before and after correction

PKS gene models before and after correction. Coding sequence is shown in yellow. Disagreements between the underlying nucleotide sequences are shown as black regions in the grey bars. Indel errors in the underlying sequence force the gene prediction algorithms to introduce false introns.

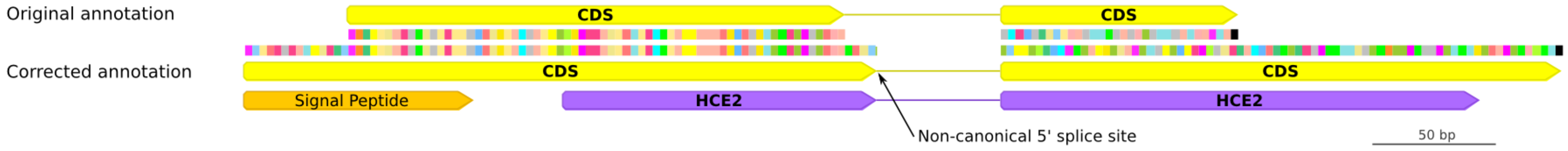



SNOG_07866
corrected
original
1 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5,000 5,500 6,000 6,500 7,000 7,500 8,000 8,500 8,833
SNOG_08274
corrected
original
1 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5,000 5,500 6,000 6,500 6,718
SNOG_08614
corrected
original
1 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5,000 5,500 6,000 6 6,806
SNOG_09490
corrected
original
CDS
1 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5,000 5,500 6,000 6,500 7,000 7,500 7,948
SNOG_09622
corrected
original
1 100 200 300 400 500 600 700 800 900 1,000 1,100 1,200 1,302
SNOG_09623
corrected
original
1 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5,000 5,500 6,000 6,500 7,000 7,500 8,000 8,397
SNOG_09932
corrected
original
1 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5,000 5,500 6,170
SNOG_11066
corrected
original
1 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5,000 5,500 6,000 6,500 7,000 7,374
SNOG_11076
corrected
original
1 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5 5,500 6,000 6,500 7,000 7,500 7,941


SNOG_11272
corrected
original
1 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5,000 5,500 6,000 6,500 7,000 7,354
SNOG_11981
corrected
original
1 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4 5,000 5 6,000 6,609
SNOG_12897
corrected
original
1 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5,000 5,500 6,000 6,500 7,000 7,271
SNOG_13032
corrected
original
1 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5,000 5,500 6,000 6,500 6,888
SNOG_14927
corrected
original
1 500 1,000 1,500 2,000 2,500 3,000 3,500 4,000 4,500 5,000 5,500 6,000 6,210
SNOG_15829
corrected
original
1 250 500 750 1,000 1,250 1,500 1,750 2,000 2,250 2,500 2,750 3,000 3,250 3,500 3,750 4,000 4,250 4,500 4,750 5,000 5,250 5,455
SNOG_15965
corrected
original
1 500 1,000 1,500 2,000 2,500 3,000 3 4,000 4 5,000 5,500 6,000 6,500 7,000 7,658

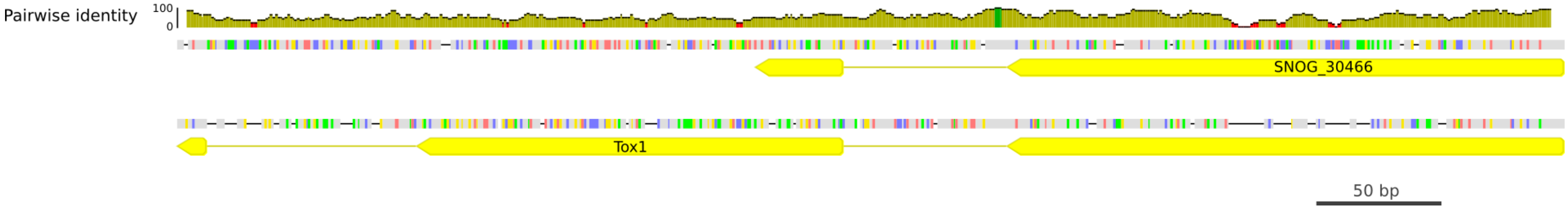
A5-7 | Changes to the annotation of SNOG_11237

Changes to the annotation of SNOG_11237. Coding sequence (CDS) is shown in yellow. Extension of the 5' end of the first exon reveals a signal peptide, and adjustment of the intron boundary causes a frameshift for the second exon revealing the HCE2 domain.



A5-8 | Putative Tox1 pseudogene

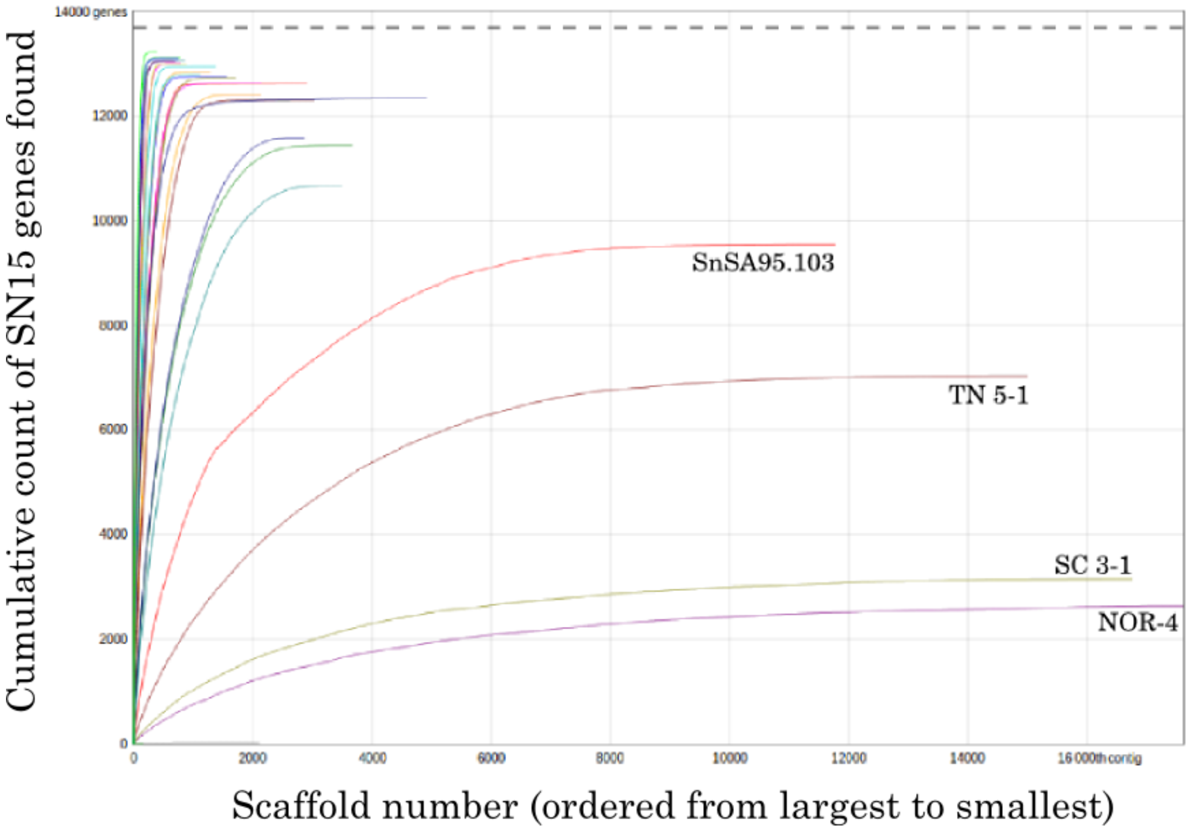
Alignment of the Tox1 gene with the potential paralog SNOG_30466 shows 52.8% pairwise identity. They share the first intron structure, which is confirmed by RNA-seq. SNOG_30466 also has a second intron downstream of the coding region but it is not in the same position as the second (not shown). The gene is likely to be a pseudogenic paralog of the functional copy of Tox1.



Chapter 6 Appendices

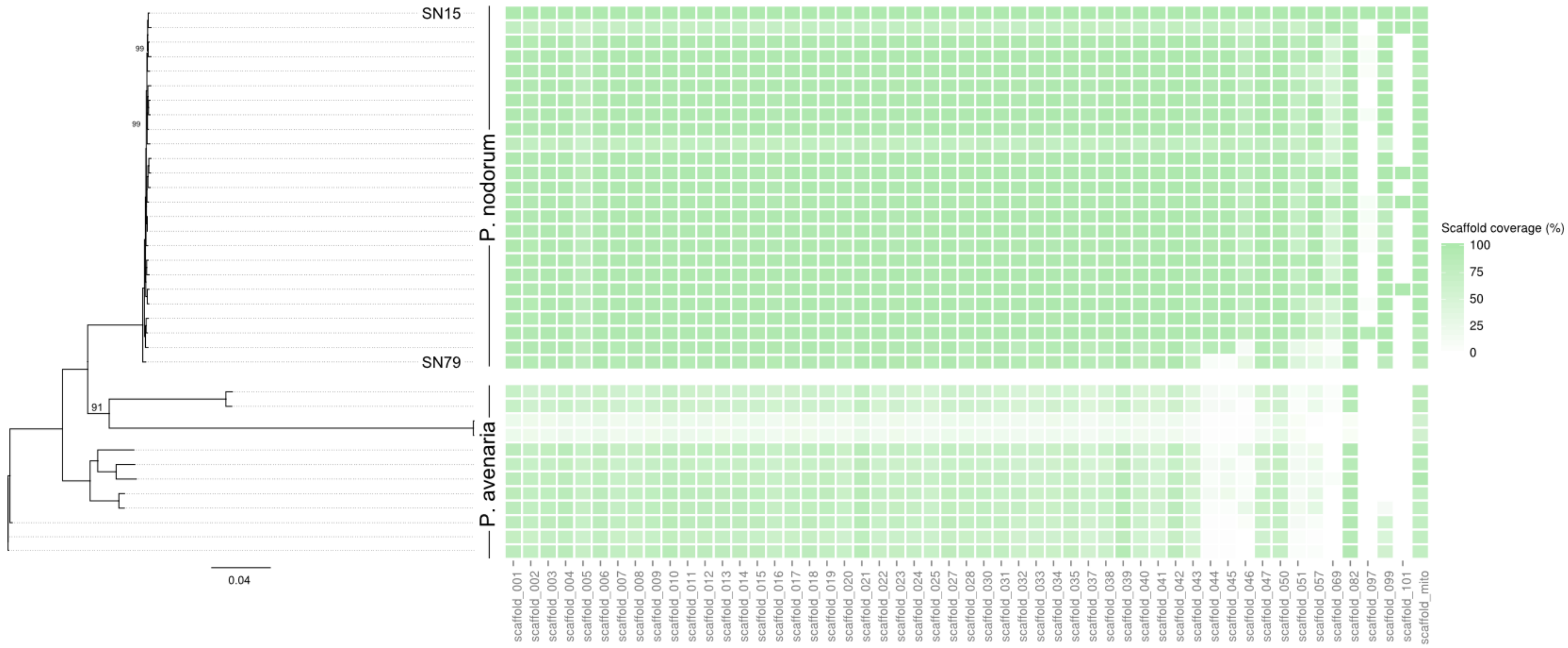
A6-1 | QUAST genome reference gene coverage plot

Three of the lowest coverage strains contain fewer than 8000 of the SN15 reference genes. For the purposes of calculating pan and core genomes, these stains were excluded from the analyses as they would likely contribute too many false negatives.



A6-2 | Coverage of all SN15 genic scaffolds by all altenate strains

The coverage of all genic scaffolds by nucmer matches from all alternate strains. Many of the smaller scaffolds have a high percentage of repeat and very low gene count.



A6-3 | Strain-specific protein counts

Strain name	**Number of strain-specific proteins**
B2.1b	76
C1.2a	59
IR10_9.1a	153
IR10_2.1a	62
IR10_5.2b	210
SN11IR_2_1.1	432
82-4841	256
83-6011-2	246
SN11IR_6_1.1	222
SN11IR_7_2.3	253
Mt. Baker	1043
s258	413
H6.2b	333
A1 3.1a	439
Hartney99	369
Jansen 4_55	1037
Sn Cp2052	79
FIN-2	49
NOR-4	1107
SWE-3	96
BRSn9870	170
Sn99CH 1A7a	27
SnChi01 40a	27
SnSA95.103	549
SnOre11-1	29
OH03 Sn-1501	40
SNOV92X D1.3	28
AR1-1	18
TN 5-1	1097
VA 5-2	41
GA9-1	16
MD4-1	19
SC 3-1	917
SN15	127
SN4	190
SN79	302
WAC8410	184

A6-4 | Top Effector Candidate Predictions

Effector candidate predictions with scores equal to or greater than known effectors ToxA, Tox1 and Tox3. Known effectors are highlighted in orange.

	Secreted	*Absent from SN79*	*Near repeats*	*Low gene density*	*Small molecular mass*	*Membrane-bound*	*Positive selection*	*Cysteine-rich*	*Conserved in P. nodorum*	*Strain-specific*	***Total***
SNOR_20078 (Tox1)	1	1	1	1	1	0	1	1	0	0	**7**
SNOR_01124	1	0	1	1	1	0	1	1	0	0	**6**
SNOR_12811	1	0	1	1	1	0	1	1	0	0	**6**
SNOR_30828	1	1	1	1	1	0	0	1	0	0	**6**
SNOR_30343	1	0	1	1	1	0	0	1	0	0	**5**
SNOR_00234	0	1	0	1	1	0	1	1	0	0	**5**
SNOR_01601	0	1	1	0	1	0	1	1	0	0	**5**
SNOR_03715	0	1	1	1	1	0	0	1	0	0	**5**
SNOR_05030	1	1	0	1	1	0	0	1	0	0	**5**
SNOR_05051	0	1	1	1	1	0	1	0	0	0	**5**
SNOR_06079	1	1	1	1	1	0	0	0	0	0	**5**
SNOR_08206	1	0	1	1	1	0	0	1	0	0	**5**
SNOR_08981 (Tox3)	1	1	1	1	1	0	0	0	0	0	**5**
SNOR_09446	0	0	1	1	1	0	1	1	0	0	**5**
SNOR_09738	1	1	1	0	1	0	0	1	0	0	**5**
SNOR_11828	1	1	1	1	1	0	0	0	0	0	**5**
SNOR_14914	0	1	1	1	1	0	0	1	0	0	**5**
SNOR_16166	0	1	1	1	1	0	0	1	0	0	**5**
SNOR_16243	0	1	1	1	1	0	1	0	0	0	**5**
SNOR_16270	0	1	1	1	1	0	1	0	0	0	**5**
SNOR_16520	1	1	1	0	1	0	1	0	0	0	**5**
SNOR_16571 (ToxA)	1	1	1	1	1	0	0	0	0	0	**5**
SNOR_20154	0	1	1	1	1	0	1	0	0	0	**5**
SNOR_30026	1	0	1	0	1	0	1	1	0	0	**5**
SNOR_30077	1	1	0	1	1	0	0	1	0	0	**5**
SNOR_30334	1	0	1	1	1	0	0	1	0	0	**5**
SNOR_30466	1	1	1	1	1	0	1	1	-2	0	**5**
SNOR_30697	0	1	1	1	1	0	0	1	0	0	**5**
SNOR_30802	1	1	1	0	1	0	0	1	0	0	**5**
SNOR_30973	0	1	0	1	1	0	1	1	0	0	**5**