Bioinformatic genome analysis of the necrotrophic wheat-pathogenic fungus *Phaeosphaeria nodorum* and related Dothideomycete fungi.

James Kyawzwar Hane

Bachelor of Molecular Biology (Hons.), University of Western Australia

This thesis is presented for the degree of Doctor of Philosophy

2011

I declare that this thesis is my own account of my own work and contains as its main content work which has not been previously submitted for any degree at any tertiary institution.

....

James Kyawzwar Hane

It is a truth universally acknowledged that a single Ph. D. student in possession of a whole genome sequence must be in want of a stiff drink. Of all the genomes in all the fungi in all the world, *ToxA* horizontally transfers out of mine. To lose 11 kilobases may be regarded as a misfortune; to lose 280 kilobases looks like carelessness! It was the best of times, it was the worst of times; it was the age of genomics, it was the age of bioinformatics; it was the epoch of next-generation sequencing, it was the epoch of data analysis; it was the season of transcriptomics, it was the season of proteogenomics; it was the spring of mesosynteny, it was the winter of repeat-induced point mutation; we had everything before us, we had nothing before us; we were all going directly to Curtin, but I was going the other way. All this happened, more or less.

In my younger and more vulnerable years my supervisor gave me some advice that I've been turning over in my mind ever since: "Everything's got a moral, if only you can find it". I don't believe there's an atom of meaning in it. It would be so nice if something made sense for a change. Sometimes I've believed as many as six impossible things before breakfast. Granted: I am a student of a research institution; my supervisor is watching me, he never lets me out of his sight; there's a peephole in the door, and my supervisor's expertise is the shade of biochemistry that can never see through a bioinformatics type like me.

Oh! it is absurd to have a hard and fast rule about what one should read and what one shouldn't. More than half of scientific literature depends on what one shouldn't read. The truth is rarely pure and never simple. Scientific research would be very tedious if it were either, and scientific literature a complete impossibility! I have always been of opinion that a man who desires to get a Ph. D. should know either everything or nothing. I do not approve of anything that tampers with natural ignorance. Ignorance is like a delicate exotic fruit; touch it and the bloom is gone. The whole theory of modern education is radically unsound. Fortunately in Murdoch, at any rate, education produces no effect whatsoever. If it did, it would prove a serious danger to the upper management, and probably lead to acts of violence in Bush Court.

Of all the things that drive men to research, the most common disaster, I've come to learn, is women. Here's looking at you, kid. We'll always have Perth. You must remember this - maybe not today, maybe not tomorrow, but soon and for the rest of your life.

I have never begun a thesis with more misgiving. I write this sitting in the kitchen sink. I don't write accurately - anyone can write accurately - but I write with wonderful expression. Read the directions and directly you will be directed in the right direction. When I use a word, it means just what I choose it to mean – neither more nor less. I never travel without my thesis. One should always have something sensational to read in the train. Whether I shall turn out to be the hero of my own life, or whether that station will be held by anybody else, these pages must show.

Abstract

Phaeosphaeria nodorum (anamorph: *Stagonospora nodorum*) is the causal agent of Stagonospora nodorum blotch (SNB, syn. glume blotch) in wheat. *P. nodorum* is estimated to cause up to 31% wheat yield loss worldwide. Within Australia it is the primary pathogen of wheat and is estimated to cause losses of $108 million per annum. The genome assembly of *P. nodorum* was sequenced in 2005 and was the first species in the class Dothideomycetes, a significant fungal taxon containing several major phytopathogens, to be publically released. The *P. nodorum* genome database has since evolved from basic sequence data into a powerful resource for studying the SNB host-pathogen interaction and advancing the understanding of fungal genome structure. The genes of *P. nodorum* have been annotated to a high level of accuracy and now serve as a model dataset for comparative purposes. *P. nodorum* gene annotations have been refined by a combination of several techniques including manual curation, orthology with related species, expressed sequence tag (EST) alignment, and proteogenomics. Analysis of the repetitive DNA in the *P. nodorum* genome lead to the development of software for the analysis of repeat-induced point mutation (RIP), a fungal-specific genome defence mechanism, which was a major improvement upon previous methods. Comparative genomics between *P. nodorum* and related species has highlighted a novel pattern of genome sequence conservation between filamentous fungi called ‘mesosynteny’ and has lead to the development of novel 'genome finishing' strategies.

Acknowledgements

First and foremost, I would like to express my deep gratitude to my principal supervisor Richard Oliver, whose guidance has been invaluable. Richard provides his students with exceptional opportunities to develop and grow. Many thanks also to my supervisors Barbara Howlett and Thierry Rouxel who have been extremely helpful with constructive criticism and expert advice.

I am indebted to the GRDC, without which I would not have had the financial means to pursue a scientific career, and to iVEC and the NCI which provide excellent high-performance computing facilities for research purposes at no charge. I would also like to express my gratitude to Karam Singh, who has recently hired me to work at CSIRO.

I would like to thank my friends and colleagues at the ACNFP for their professional support, expertise and friendship – I learnt a lot from all of you. KC, Zippy, Rohan, Peter, Lars, Kasia, Ormonde, Joel, Christian and Eva – my waistline will miss our squash games. Rob, Maryn, Ramisah, Francis, Judith, Julie, Madeline, Emma, Nola, Megan and Simon – I enjoyed working with you.

Many thanks to my father Zor and step-mother Teeni, my mother Kimberley, my brothers Michael and Alex and my grandmother Gwenda, who have been supportive and understanding while I was finishing my Ph. D. Last but not least, I would like to thank Angela from the bottom of my heart for her support and for not letting me give up.

Table of contents

Chapter 1: Introduction

Whole-genome sequencing rapidly accelerates scientific discovery. The knowledge of an organism's genome content allows life-scientists to predict its inherent biological processes. Whole-genome sequencing is also the foundation upon which additional layers of information can be juxtaposed. Transcriptomic, proteomic and metabolomic data can be built upon a whole genome sequence to fully develop a complete research model of an organism. This thesis project outlines bioinformatic work applied to the whole-genome assembly of the necrotrophic wheat pathogen *Phaeosphaeria nodorum*. It is presented as a series of peer-reviewed publications (Chapters 2-9) and a book chapter (Chapter 10).

Phaeosphaeria nodorum (anamorph *Stagonospora nodorum*, syn. *Leptosphaeria nodorum*, syn. *Septoria nodorum*) is the causal agent of Stagonospora nodorum blotch (SNB, syn. glume blotch) in wheat. *P. nodorum* is widely distributed [1] and has been estimated to cause up to 31% in yield losses worldwide [2]. Within Western Australia it is the primary wheat pathogen and is estimated to cause upwards of $108 million in yield losses annually [3]. *P. nodorum* is a member of the class Dothideomycetes which contains many agriculturally and economically significant species [4] . Whole genome sequencing of Dothideomycetes was undertaken relatively late compared with other classes of Fungi. Prior to the commencement of this Ph. D. project in February 2007, 49 fungal species had been sequenced and publically released within its phylum, the Ascomycota (Table 1) *P. nodorum* was at the time the only member of the class Dothideomycete for which a whole-genome assembly was sequenced and publically released (www.broadinstitute.org). The sequenced Ascomycetes were predominantly yeasts, Aspergilli and saprophytic species. Only two plant pathogens had been sequenced at this time: *Magnaporthe orzyae*, which causes rice blast [5]; and *Fusarium* *graminearum* which causes wheat head blight [6].

Some genomic resources for *P.nodorum* were available prior to the release of the genome assembly. These included 86 characterised gene sequences, 3 transposons and 1197 expressed-sequence tags (ESTs). The 3 characterised transposon sequences of *P. nodorum* were observed to functionally active [7]. Of the 1197 ESTs, 213 had been from oleate-induced libraries [8] and the remaining 984 ESTs were generated from libraries grown on wheat-cell [9]. Functions of characterised genes were inferred from observations of gene knock-out mutant phenotypes (for references refer to: Table 2). These included genes involved in cell wall degradation (*AreA*, *Bgl1*, *Snp1*, *Snp2*, *Snp3*, *Snodxyl*), virulence (*Als1*, *Gap1*, *Mak2*, *Mls1*, *OdcI*, *SP1*), fungicide resistance (*TubA*), sporulation (*CpkA*), cell wall structure (*Chs1*, *Chs2*) and metabolic processes (*LeuA*, *Gpd1*, *Glo1*, *Mpd1*, *NiiI*, *NiaI*).

In 2004 the ACNFP commissioned the BROAD institute (www.broadinstitute.org) to sequence, assemble [10] and auto-annotate [11] the genome of *P. nodorum*. This work was completed and publically released in 2005 (http://www.broadinstitute.org/annotation/genome/stagonospora_nodorum.2/Home.html). A total of 16597 genes were been predicted based on a training set of 317 genes, which were based on a combination of existing EST data and alignments of highly conserved fungal genes. Using this initial assembly, Friesen et al. discovered evidence for the lateral transfer of a host-specific toxin (*ToxA*) which was previously found only in *Pyrenophora tritici-repentis* [12]. Lowe *et al.* later developed more comprehensive EST libaries, with 10751 ESTs sequenced from *P. nodorum*-infected wheat lesions and 10752 ESTs derived from oleate-induced *P. nodorum* grown in culture [13]. This formed the foundation for the work performed during this Ph. D. project.

Table 1: Fungal whole genome sequencing projects listed in the NCBI Genome Project database prior to February 2007.

Year	Center/Consortium	Phylum	Class	Species	Strain/isolate	Project Type	Genome Project ID	Taxonomy ID
1996	Welcome Trust Sanger Institute	Ascomycota	Saccharomycetes	*Saccharomyces cerevisiae*	S288c	first	13838	559292
2001	Stanford University	Ascomycota	Saccharomycetes	*Candida albicans*	SC5314	first	10701	237561
2001	Genoscope	Microsporidia	Incertae sedis	*Encephalitozoon cuniculi*	GB-M1	first	13833	284813
2002	S. pombe European Sequencing Consortium (EUPOM)	Ascomycota	Schizosaccaromycetes	*Schizosaccharomyces pombe*	972h-	first	13836	284812
2003	Broad Institute	Ascomycota	Eurotiomycetes	*Aspergillus nidulans*	FGSC A4	first	130	227321
2003	Microbia	Ascomycota	Eurotiomycetes	*Aspergillus terreus*	ATCC 20542	first	187	285217
2003	Genome Sequencing Center (GSC) at Washington University (WashU) School of Medicine	Ascomycota	Saccharomycetes	*Saccharomyces bayanus*	623-6C	alternate strain	1443	226231
2003	Broad Institute	Ascomycota	Saccharomycetes	*Saccharomyces bayanus*	MCYC623	first	1441	226127
2003	Genome Sequencing Center (GSC) at Washington University (WashU) School of Medicine	Ascomycota	Saccharomycetes	*Saccharomyces castellii*	NRRL Y-12630	alternate strain	1444	226301
2003	Genome Sequencing Center (GSC) at Washington University (WashU) School of Medicine	Ascomycota	Saccharomycetes	*Saccharomyces kluyveri*	NRRL Y-12651	first	1445	226302
2003	Genome Sequencing Center (GSC) at Washington University (WashU) School of Medicine	Ascomycota	Saccharomycetes	*Saccharomyces kudriavzevii*	IFO1802	first	1442	226230
2003	Broad Institute	Ascomycota	Saccharomycetes	*Saccharomyces mikatae*	IFO1815	first	374	226126
2003	Genome Sequencing Center (GSC) at Washington University (WashU) School of Medicine	Ascomycota	Saccharomycetes	*Saccharomyces mikatae*	IFO1815	alternate strain	9601	226126
2003	Broad Institute	Ascomycota	Saccharomycetes	*Saccharomyces paradoxus*	NRRL Y-17217	first	1440	226125
2003	International Gibberella zeae Genomics Consortium	Ascomycota	Sordariomycetes	*Gibberella zeae*	PH-1	first	13839	229533
2003	International Rice Blast Genome Consortium	Ascomycota	Sordariomycetes	*Magnaporthe oryzae*	70-15	first	13840	242507
2003	Broad Institute	Ascomycota	Sordariomycetes	*Neurospora crassa*	OR74A	first	13841	367110

2003	Broad Institute	Basidiomycota	Agaricomycetes	*Coprinopsis cinerea okayama*	7#130	first	1447	240176
2003	Broad Institute	Basidiomycota	Tremellomycetes	*Cryptococcus neoformans var. grubii*	H99	first	411	235443
2003	Broad Institute	Basidiomycota	Ustilaginomycetes	*Ustilago maydis*	521	first	1446	237631
2004	Broad Institute	Ascomycota	Eurotiomycetes	*Coccidioides immitis*	RS	first	12883	246410
2004	Zoologisches Institut der University Basel, Switzerland	Ascomycota	Saccharomycetes	*Ashbya gossypii*	ATCC 10895	first	13834	284811
2004	Genolevures	Ascomycota	Saccharomycetes	*Candida glabrata*	CBS138	first	13831	284593
2004	Genolevures	Ascomycota	Saccharomycetes	*Debaryomyces hansenii*	CBS767	first	13832	284592
2004	Genolevures	Ascomycota	Saccharomycetes	*Kluyveromyces lactis*	NRRL Y-1140	first	13835	284590
2004	Broad Institute	Ascomycota	Saccharomycetes	*Kluyveromyces waltii*	NCYC 2644	first	10734	262981
2004	Genolevures	Ascomycota	Saccharomycetes	*Yarrowia lipolytica*	CLIB122	first	13837	284591
2004	DOE Joint Genome Institute	Basidiomycota	Agaricomycetes	*Phanerochaete chrysosporium*	RP-78	first	135	273507
2004	Stanford University	Basidiomycota	Tremellomycetes	*Cryptococcus neoformans var. neoformans*	B-3501A	alternate strain	12386	283643
2005	ACNFP, Broad Institute	Ascomycota	Dothideomycetes	*Phaeosphaeria nodorum*	SN15	first	13754	321614
2005	Broad Institute	Ascomycota	Eurotiomycetes	*Ajellomyces capsulatus*	NAm1	first	12654	339724
2005	J. Craig Venter Institute	Ascomycota	Eurotiomycetes	*Aspergillus clavatus*	NRRL1	first	15664	344612
2005	J. Craig Venter Institute	Ascomycota	Eurotiomycetes	*Aspergillus flavus*	NRRL3357	first	13284	332952
2005	J. Craig Venter Institute	Ascomycota	Eurotiomycetes	*Aspergillus fumigatus*	Af293	first	131	330879
2005	NITE	Ascomycota	Eurotiomycetes	*Aspergillus oryzae*	RIB40	first	20809	510516
2005	Broad Institute	Ascomycota	Eurotiomycetes	*Aspergillus terreus*	NIH2624	alternate strain	15631	341663
2005	TIGR	Ascomycota	Eurotiomycetes	*Neosartorya fischeri*	NRRL181	first	15672	331117
2005	Broad Institute	Ascomycota	Eurotiomycetes	*Uncinocarpus reesii*	1704	first	15634	336963
2005	Syngenta Biotechnology, Inc.	Ascomycota	Leotiomycetes	*Botryotinia fuckeliana*	B05.10	first	15632	332648
2005	Broad Institute	Ascomycota	Leotiomycetes	*Sclerotinia sclerotiorum*	1980 UF-70	first	15530	665079
2005	Broad Institute	Ascomycota	Saccharomycetes	*Candida tropicalis*	MYA-3404	first	13675	294747
2005	Broad Institute	Ascomycota	Saccharomycetes	*Clavispora lusitaniae*	ATCC 42720	first	12753	306902
2005	Broad Institute	Ascomycota	Saccharomycetes	*Pichia guilliermondii*	ATCC 6260	first	12729	294746

2005	Broad Institute	Ascomycota	Saccharomycetes	*Saccharomyces cerevisiae*	RM11-1a	alternate strain	13674	285006
2005	Stanford University	Ascomycota	Saccharomycetes	*Saccharomyces cerevisiae*	YJM789	alternate strain	13304	307796
2005	The Genome Sequencing Platform, The Genome Assembly Team	Ascomycota	Sordariomycetes	*Chaetomium globosum*	CBS 148.51	first	12795	306901
2005	Broad Institute	Ascomycota	Sordariomycetes	*Gibberella moniliformis*	7600	first	15553	334819
2005	DOE Joint Genome Institute	Ascomycota	Sordariomycetes	*Trichoderma reesei*	QM6a	first	15571	431241
2005	Broad Institute	Basidiomycota	Tremellomycetes	*Cryptococcus bacillisporus*	R265	first	13691	294750
2005	TIGR	Basidiomycota	Tremellomycetes	*Cryptococcus neoformans var. neoformans*	JEC21	alternate strain	13856	214684
2005	Broad Institute	Zygomycota	Incertae sedis	*Rhizopus oryzae*	RA 99-880	first	13066	246409
2006	Baylor College of Medicine	Ascomycota	Eurotiomycetes	*Ascosphaera apis*	USDA-ARSEF 7405	first	17285	392613
2006	DSM, The Netherlands	Ascomycota	Eurotiomycetes	*Aspergillus niger*	CDS 513.88	first	19275	425011
2006	Broad Institute	Ascomycota	Eurotiomycetes	*Coccidioides immitis*	H538.4	alternate strain	17355	396776
2006	Broad Institute	Ascomycota	Eurotiomycetes	*Coccidioides immitis*	RMSCC 2394	alternate strain	17713	404692
2006	Broad Institute	Ascomycota	Saccharomycetes	*Candida albicans*	WO-1	alternate strain	16373	294748
2006	Broad Institute	Ascomycota	Saccharomycetes	*Lodderomyces elongisporus*	NRRL YB-4239	first	12899	379508
2006	Broad Institute	Ascomycota	Schizosaccaromycetes	*Schizosaccharomyces japonicus*	yFS275	first	13640	402676
2006	North Carolina State University (NCSU)	Ascomycota	Sordariomycetes	*Magnaporthe oryzae*	70-15	resequencing	16061	242507
2006	Broad Institute	Chytridiomycota	Chytridiomycetes	*Batrachochytrium dendrobatidis*	JEL423	first	13653	403673

Table 2: Summary of genomic resources for *P. nodorum*, available from NCBI Protein and Nucleotide databases, prior to February 2007.

		NCBI Protein		NCBI Nucleotide		
Description	**Gene name**	**#Records**	**Accessions**	**#Records**	**Accessions**	**Reference**
Whole genome sequencing project						
BROAD Strain SN15 annotated hypothetical proteins (v1)		16586*	EAT76043.1-EAT92628.1			[12]
BROAD Strain SN15 genome assembly scaffolds (v1)		2	AAT11122.1, AAT84078.1	109	CH445325-CH445394, CH959327-CH959365	[12]
Characterised genes						
delta-aminolevulinic acid synthase	*Als1*	1	AAZ95011.1	1	DQ167577.1	
AreA protein	*AreA*	1	AAP30890.1	1	AY135715.1	[14]
beta-glucosidase	*Bgl1*	4	AAT95381.1, AAT95382.1, AAT95383.1, AAT95384.1	4	AY683617.1, AY683618.1, AY683619.1, AY683620.1	[15]
chitin synthase	*Chs1, Chs2*	2	CAB41508.1, CAB41532.1	2	AJ133695.2, AJ133696.1	[16]
chloride channel protein	*ClC-SN1*			1	X78582.1	[17]
Ca/Cm-dependent protein kinase A	*CpkA*	1	ABD59786.1	1	DQ397887.1	[18]
G-alpha subunit	*Gap1*	1	AAQ94737.1	1	AY327542.1	[19]
glyoxylase I	*Glo1*	1	AAT73077.1	1	AY576607.1	[20]
glyceraldehyde 3-phosphate dehydrogenase	*Gpd1*	6	AAQ62909.1, AAQ62910.1, AAQ62911.1, AAQ62913.1, CAB72263.1, Q9P8C0.1	1	AJ271155.1, AY364460.1, AY364461.1, AY364462.1, AY364464.1	[21]
tri-functional histidine biosynthesis protein	*His*	5	ABC40952.1, ABL63152.1, ABL63160.1, ABL63161.1, ABL63164.1	5	DQ312266.2, EF030691.1, EF030699.1, EF030700.1, EF030703.1	[22]
3-isopropylmalate dehydrogenase	*LeuA*	1	CAB72262.1	1	AJ271154.1	[23]
MAP kinase	*Mak2*	1	AAX63387.1	1	AY847792.1	[24]
Mannitol dehydrogenase	*Mdh1*	1	AAX14688.1	1	AY788902.1	[25]
Malate synthase	*Mls1*	1	AAS91580.1	1	AY508881.1	[26]
Mannitol 1-phosphate dehydrogenase	*Mtd1*	2	AAT11122.1, AAT84078.1	2	AY547308.1, AY587541.1	[27]
Nitrate reductase	*Nia1*	2	CAA08857.1, CAA74005.1	2	AJ009827.1, Y13654.1	[28, 29]
Nitrite reductase	*Nii1*	2	CAA08856.1, CAA08858.1	2	AJ009826.1, AJ009827.1	[28, 29]
Ornithine decarboxylase	*Odc1*	1	CAB56523.1	1	AJ249387.1	[30]
RNA polymerase II second largest subunit	*RPB2*	8	ABB86549.1, ABF56194.1, ABF56195.1, ABF56196.1,	8	DQ278491.1, DQ499803.1, DQ499804.1, DQ499805.1,	[31]

			ABF56197.1, ABF56198.1, ABF56199.1, ABF56200.1		DQ499806.1, DQ499807.1, DQ499808.1, DQ499809.1	
SNF1 protein	*Snf1*	1	AAR02440.1	1	AY349155.1	[32]
Putative beta-1,4-xylanase	*Snodxyl*	1	CAB53543.1	1	AJ249197.1	[33]
Trypsin-like protease	*Snp1*	1	AAC61777.1	1	AF092435.1	[17]
Aspartic protease	*Snp2*	1	AAP30888.1	1	AY135713.1	[17]
Subtilisin-like protease	*Snp3*	1	AAP30889.1	1	AY135714.1	[17]
Snodprot1	*Sp1*	2	O74238.1, AAC26870.1	1	AF074941.1	[34]
Host-selective toxin	*ToxA*	1	ABD85141.1	1	DQ423483.1	[35]
Beta-tubulin	*TubA*	12	1912290A, AAB25800.1, AAV53378.1, AAV53379.1, AAV53381.1, AAV53382.1, AAV53383.1, AAV53384.1, AAV53385.1, AAV53386.1, AAV83496.1, P41799.2	10	AY786331.1, AY786332.1, AY786334.1, AY786335.1, AY786336.1, AY786337.1, AY786338.1, AY786339.1, AY823526.1, S56922.1	[36, 37]
ATP-dependent RNA helicase		18	Q0TXZ2.1, Q0U397.2, Q0U6X2.1, Q0U7S9.1, Q0U8V9.1, Q0UCB9.1, Q0UG00.1, Q0UHM7.1, Q0UMB6.1, Q0UMB9.1, Q0UN57.2, Q0UR48.1, Q0UU86.1, Q0UWA6.1, Q0UWC8.1, Q0UY62.1, Q0UZ59.1, Q0V1Z7.1			[38]
Di/tri peptide transporter 2 gene		1	AAO31597.1	1	AY187281.1	[39]
Mating type 1 protein		3	AAK09389.1, AAL69553.1, AAO31740.1	3	AF322008.1, AY072933.1, AY212018.1	[40, 41]
Mating type 2 protein		2	AAL69554.1, AAO31742.1	2	AY072934.1, AY212019.1	[40, 41]
Transcriptome						
cDNA library of oleate-induced*P. nodorum* strain SN15 grown in vitro				213	DR046162.1-DR046373.1, DR074925.1	[8]
cDNA library of *P. nodorum* strain SN15 grown on wheat cell walls				984	DR045178.1-DR046161.1	[9]
Transposons						
Pixie		1	CAD32689.1	1	AJ488503.1	[7]
Molly		1	CAD32687.1	1	AJ488502.1	[7]
Elsa		1	CAB91877.1	1	AJ277966.1	[7]

The goal of this Ph. D. project was to survey the genome content of *P. nodorum* and generate a resource upon which additional information could be juxtaposed. This fundamental work is presented in chapter 2, which summarises the predicted gene-coding regions, repetitive DNA, pathogenicity-related and secondary metabolite genes, mitochondrial genome, transcriptomics and comparative genomics of *P. nodorum*. As additional information became available, new hypotheses were formed and investigated. These investigations can be broadly divided into three main areas of inquiry: analysis of fungal repetitive DNA (Chapters 3-6); curation and characterisation of *P. nodorum* gene content (Chapter 7); and comparative genomics of Dothideomcyetes (Chapters 8-9).

Analysis of the repetitive DNA content of *P. nodorum* and a fungal-specific phenomenon known as repeat-induced point mutation (RIP) is outlined in Chapter 3. Chapter 3 also introduces a new bioinformatics software tool called RIPCAL, which was specifically designed for the analysis of RIP in whole-genome assemblies. Analysis of RIP is continued in Chapter 4, which describes an addition to RIPCAL, deRIP. DeRIP is a bioinformatic method capable of accurately predicting the sequences of repetitive elements before they were subject to RIP mutation. This allowed further characterisation of the role and origin of the repetitive DNA content of *P. nodorum*. During the course of this research project, additional Dothideomycete genome assemblies became available. The genome assembly of *Leptosphaeria maculans*, which causes blackleg on canola and other brassicas [42], was completed in 2006 and made available to us through collaboration with researchers at the French National Institute for Agricultural Research (http://www.international.inra.fr/). RIPCAL and deRIP were applied to studying the effects of RIP on pathogenicity effectors in *Leptosphaeria maculans*, which is presented in Chapter 5. While RIP is triggered by repetitive DNA sequences, it can also leak into non-repetitive regions and rapidly generate sequence diversity. Evidence presented in Chapter 5 suggests that *L. maculans* exploits RIP to rapidly adapt certain effector proteins involved in pathogenicity. RIPCAL and DeRIP were subsequently employed to quantify RIP and characterise the repetitive DNA content of *L. maculans* in Chapter 6.

The *P. nodorum* gene content was initially predicted by the BROAD using a training set of 317 genes. Chapter 2 includes details of changes to this set of genes based on EST support for 2695 *P. nodorum* genes. In addition to transcriptomic-based curation of predicted gene models, in Chapter 7 we also apply a proteogenomic approach. Proteogenomics involves the matching of peptide masses determined by mass-spectrometry to translated sequences of open-reading frames within the genome assembly. Genes which had previously been incorrectly annotated or not predicted by computation methods were corrected. As a complementary technique, proteogenomics verified similar numbers of gene predictions as did EST-support (2134 vs 2695) and provided evidence for the correction of 662 genes.

In addition to *L. maculans*, five other Dothideomycete genome assemblies and gene prediction datasets became available during this Ph. D. project. *Pyrenophora tritici-repentis*, *Mycosphaerella fijiensis* and *Mycosphaerella graminicola* data were publically released in 2007 and *Alternaria brassicola* and *Cochliobolus heterostrophus* data were released the following year. This presented an opportunity to explore genomic commonalities and differences within the Dothideomycetes.

Sequence comparisons at a whole-genome scale revealed a novel type of synteny between these Dothideomycete species. Chapter 8 explores this novel type of sequence conservation, which we call mesosynteny. In this study, mesosynteny was found to be particularly evident in the Dothideomycetes but also generally observed between all species of the sub-phylum Pezizomycotina. Chapter 8 introduces mesosynteny and reflects upon its potential mode of action and implications for the evolution of fungal genomes. In Chapter 9 mesosynteny is also investigated in relation to *M. graminicola*. The body of work presented in Chapters 2-9 is summarised in Chapter 10, which was written for the Mycota volume XIV close to the end of the research project. It summarises published genomic analysis of Dothideomycete species, primarily of *P. nodorum*, but also of *L. maculans* and *M. graminicola*, with special attention given to comparisons between these three species. Genome curation, RIP, mesosynteny and functional comparative genomics analyses are recapitulated in Chapter 10 in the context of comparative genomics of Dothideomcyetes.

In summary, the genomic analysis of *P. nodorum* genome was a significant milestone in the study of Dothdideomycetes and plant pathogenicity. The sequencing and initial survey of its genome rapidly advanced knowledge of *P. nodorum* and related species and opened up new areas of research such as deRIP and mesosynteny which have broad relevance to the field of mycology.

References

1. Solomon PS, Lowe RG, Tan KC, Waters OD, Oliver RP: **Stagonospora nodorum: cause of stagonospora nodorum blotch of wheat**. *Molecular plant pathology* 2006, **7**(3):147-156.
2. Bhathal JS, Loughman R, Speijers J: **Yield reduction in wheat in relation to leaf disease from yellow (tan) spot and *septoria nodorum* blotch.** *Eur J Plant Pathol* 2003, **109**:435–443.
3. Murray GM, Brennan JP: **Estimating disease losses to the Australian wheat industry.** *Aust Plant Pathol* 2009, **38**:558-570.
4. Schoch CL, Crous PW, Groenewald JZ, Boehm EW, Burgess TI, de Gruyter J, de Hoog GS, Dixon LJ, Grube M, Gueidan C *et al*: **A class-wide phylogenetic assessment of Dothideomycetes**. *Studies in mycology* 2009, **64**:1-15S10.
5. Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, Thon M, Kulkarni R, Xu JR, Pan H *et al*: **The genome sequence of the rice blast fungus *Magnaporthe grisea***. *Nature* 2005, **434**(7036):980-986.
6. Cuomo CA, Guldener U, Xu JR, Trail F, Turgeon BG, Di Pietro A, Walton JD, Ma LJ, Baker SE, Rep M *et al*: **The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization**. *Science (New York, NY* 2007, **317**(5843):1400-1402.
7. Rawson JM: **Transposable Elements in the Phytopathogenic Fungus *Stagonospora nodorum.* PhD Dissertation**. Birmingham, UK: University of Birmingham; 2000.
8. Lee RC, Oliver RP: ***S. nodorum* oleate-induced cDNAs**. In: *unpublished.* Perth, Australia: ACNFP, Murdoch University; 2005.
9. Bindschedler LV, Cooper RM, Thomas SW, Madrid MP, Oliver RP: **cDNA library of *Phaeosphaeria nodorum* grown on Wheat cell walls**. In: *unpublished.* Perth, Australia: ACNFP, Murdoch University; 2005.
10. Batzoglou S, Jaffe DB, Stanley K, Butler J, Gnerre S, Mauceli E, Berger B, Mesirov JP, Lander ES: **ARACHNE: a whole-genome shotgun assembler**. *Genome research* 2002, **12**(1):177-189.
11. Engels R, Yu T, Burge C, Mesirov JP, DeCaprio D, Galagan JE: **Combo: a whole genome comparative browser**. *Bioinformatics (Oxford, England)* 2006, **22**(14):1782-1783.
12. Hane JK, Lowe RG, Solomon PS, Tan KC, Schoch CL, Spatafora JW, Crous PW, Kodira C, Birren BW, Galagan JE *et al*: **Dothideomycete plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum***. *The Plant cell* 2007, **19**(11):3347-3368.
13. Lowe RGT: **Sporulation of *Stagonospra nodorum*.** Perth, Australia: Murdoch University; 2006.
14. Bindschedler LV, Sanchez P, Cooper RM: **Secreted proteases and depolymerases from *Stagonospora nodorum* during infection of wheat** *unpublished data submitted to NCBI* 2002.
15. Reszka E, Arseniuk E, Malkus A: **A New Biotype of *Phaeosphaeria* sp. of Uncertain Affinity Causing Stagonospora Leaf Blotch Disease in Cereals in Poland.** *Plant Disease* 2006, **90**(1):113.
16. Howard K, Foster SG, Cooley RN, Caten CE: **Gene disruption as a method of fungicide target validation**. In: *Proceedings 15th Long Ashton International*

Symposium – Understanding Pathosystems: A Focus on Septoria: 5–17 September 1997; Long Ashton, UK; 1997: 49.

17. Borsani G, Rugarli EI, Taglialatela M, Wong C, Ballabio A: **Characterization of a human and murine gene (*CLCN3*) sharing similarities to voltage-gated chloride channels and to a yeast integral membrane protein**. *Genomics* 1995, **27**(1):131-141.
18. Solomon PS, Rybak K, Trengove RD, Oliver RP: **Investigating the role of calcium/calmodulin-dependent protein kinases in *Stagonospora nodorum***. *Molecular microbiology* 2006, **62**(2):367-381.
19. Solomon PS, Tan KC, Sanchez P, Cooper RM, Oliver RP: **The disruption of a Galpha subunit sheds new light on the pathogenicity of *Stagonospora nodorum* on wheat**. *Mol Plant Microbe Interact* 2004, **17**(5):456-466.
20. Solomon PS, Oliver RP: **Functional characterisation of glyoxalase I from the fungal wheat pathogen *Stagonospora nodorum***. *Current genetics* 2004, **46**(2):115-121.
21. Ueng PP, Reszka E, Chung KR, Arseniuk E: **Comparison of glyceraldehyde-3-phosphate dehydrogenase genes in *Phaeosphaeria nodorum* and *P. avenaria* species.** *Plant Pathology* 2003, **12**(255):0032-0862.
22. Malkus A, Chung KR, Chang CJ, Ueng PP: **The Tri-functional Histidine Biosynthesis Gene (*his*) in Wheat Stagonospora Nodorum Blotch Pathogen, *Phaeosphaeria nodorum*.** *Plant Pathology Bulletin* 2006, **15**:55-62.
23. Cooley RN, Monk TP, McLoughlin SB, Foster SG, Dancer JE: **Gene disruption and biochemical characterisation of 3-isopropylmalate dehydrogenase from *Stagonospora nodorum***. *Pest Management Science* 1999, **55**(3):364-367.
24. Solomon PS, Waters OD, Simmonds J, Cooper RM, Oliver RP: **The Mak2 MAP kinase signal transduction pathway is required for pathogenicity in *Stagonospora nodorum.*** *Current genetics* 2005, **48**(1):60-68.
25. Solomon PS, Waters OD, Jorgens CI, Lowe RG, Rechberger J, Trengove RD, Oliver RP: **Mannitol is required for asexual sporulation in the wheat pathogen *Stagonospora nodorum* (glume blotch).** *Biochem J* 2006, **399**(2):231-239.
26. Solomon PS, Lee RC, Wilson TJ, Oliver RP: **Pathogenicity of *Stagonospora nodorum* requires malate synthase**. *Molecular microbiology* 2004, **53**(4):1065-1073.
27. Solomon PS, Tan KC, Oliver RP: **Mannitol 1-phosphate metabolism is required for sporulation in planta of the wheat pathogen *Stagonospora nodorum***. *Mol Plant Microbe Interact* 2005, **18**(2):110-115.
28. Cutler SB, Cooley RN, Caten CE: **Cloning of the nitrate reductase gene of Stagonospora (Septoria) nodorum and its use as a selectable marker for targeted transformation**. *Current genetics* 1998, **34**(2):128-137.
29. Howard K, Foster SG, Cooley RN, Caten CE: **Disruption, replacement, and cosuppression of nitrate assimilation genes in Stagonospora nodorum**. *Fungal Genet Biol* 1999, **26**(2):152-162.
30. Bailey A, Mueller E, Bowyer P: **Ornithine decarboxylase of Stagonospora (Septoria) nodorum is required for virulence toward wheat.** *J Biol Chem* 2000, **275**(19):14242-14247.
31. Malkus A, Chang PF, Zuzga SM, Chung KR, Shao J, Cunfer BM, Arseniuk E, Ueng PP: **RNA polymerase II gene (RPB2) encoding the second largest protein subunit in *Phaeosphaeria nodorum* and *P. avenaria***. *Mycological research* 2006, **110**(Pt 10):1152-1164.

32. Shaikh SK, Bailey AM: **SNF1 gene of *Stagonospora nodorum***. In: *unpublished.* Bristol, UK: Biological Sciences, Bristol University; 2003.

33. Emami K: **PCR-based characterization of fungal xylanase genes**. In: *unpublished.* Newcastle, UK: Biological And Nutritional Sciences, University Of Newcastle Upon Tyne; 1999.

34. Hall N, Keon JPR, Hargreaves JA: **A homologue of a gene implicated in the virulence of human fungal diseases is present in a plant fungal pathogen and is expressed during infection.** *Physiological and Molecular Plant Pathology* 1999, **55**(1):69-73.

35. 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**. *Nature genetics* 2006, **38**(8):953-956.

36. Cooley RN, Caten CE: **Molecular analysis of the Septoria nodorum beta-tubulin gene and characterization of a benomyl-resistance mutation**. *Mol Gen Genet* 1993, **237**(1-2):58-64.

37. Malkus A, Reszka E, Chang CJ, Arseniuk E, Chang PF, Ueng PP: **Sequence diversity of beta-tubulin (tubA) gene in Phaeosphaeria nodorum and P. avenaria**. *FEMS microbiology letters* 2005, **249**(1):49-56.

38. Kuramae EE, Robert V, Echavarri-Erasun C, Boekhout T: **Cophenetic correlation analysis as a strategy to select phylogenetically informative proteins: an example from the fungal kingdom.** *BMC Evolutionary Biology* 2007, **7**:134.

39. Solomon PS, Thomas SW, Spanu P, Oliver RP: **The utilisation of di/tripeptides by *Stagonospora nodorum* is dispensable for wheat infection.** *Physiol Mol Plant Pathol* 2004, **63**(4):191-199.

40. Bennett RS, Yun SH, Lee TY, Turgeon BG, Arseniuk E, Cunfer BM, Bergstrom GC: **Identity and conservation of mating type genes in geographically diverse isolates of Phaeosphaeria nodorum**. *Fungal Genet Biol* 2003, **40**(1):25-37.

41. Dai Q, Arseniuk E, Cui K, Ueng PP: **Genetic segregation and sexuality in *Phaeosphaeria nodorum.*** In: *unpublished.* Beltsville, MD, USA: Molecular Plant Pathology Laboratory, USDA-ARS; 2000.

42. Howlett BJ, Idnurm A, Pedras MSC: ***Leptosphaeria maculans*, the causal agent of blackleg disease of Brassicas.** *Fungal Genet Biol* 2001, **33**:1–14.

Chapter 2: Attribution Statement

Title: **Dothideomycete–Plant Interactions Illuminated by Genome Sequencing and EST Analysis of the Wheat Pathogen *Stagonospora nodorum.***

Authors: **James K. Hane**, Rohan G.T. Lowe, Peter S. Solomon, Kar-Chun Tan, Conrad L. Schoch, Joseph W. Spatafora, Pedro W. Crous, Chinappa Kodira, Bruce W. Birren, James E. Galagan, Stefano F.F. Torriani, Bruce A. McDonald and Richard P. Oliver

Citation: *The Plant Cell* 19:3347-3368 (2007)

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 (JKH) made the following contributions to this chapter:

- Assisted SFFT with mitochondrial genome annotation in regions requiring correction to initial base-calling and reassembly.
- Analysis of repetitive DNA.
- Mapping of EST sequences, manual curation of EST-supported genes and EST-based prediction of version 2 gene annotations.
- Assigned predicted function (gene ontology) to gene annotations.
- Assisted RGTL with statistical analysis of relative EST expression levels.
- Characterisation, manual curation and analysis of polyketide-synthase (PKS) and non-ribosomal peptide synthase (NRPS) genes.
- Characterisation and analysis of hydrophobin genes.
- Comparison of pathogenicity-related gene functions and comparative genomics of the predicted secretomes of selected sequenced fungi.
- Co-writing of manuscript with RPO.

The following contributions were made by co-authors:

- RGTL constructed single and paired-end EST libraries for sequencing and performed statistical analysis of relative EST expression levels (Figure 9).
- PSS performed qPCR experiments.
- KCT performed comparisons between conserved microsyntenic clusters of genes (Figure 5).
- CLS, JWS and PWC performed phylogenetic analyses and contributed to sections of the text pertaining to phylogenetics.
- CK, BB and JEG represent staff at the BROAD institute, who sequenced and assembled the *Stagonospora nodorum* whole-genome assembly and predicted version 1 gene annotations.
- SFFT and BAM assembled and annotated the mitochondrial genome of *Stagonospora nodorum*.
- RPO compiled and wrote the manuscript.
- All authors read and approved the manuscript.

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

--   --

James Hane (Ph. D. candidate)   Date

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

--   --

Richard Oliver (Principal supervisor)   Date

The Plant Cell, Vol. 19: 3347–3368, November 2007, www.plantcell.org © 2007 American Society of Plant Biologists

RESEARCH ARTICLES

Dothideomycete–Plant Interactions Illuminated by Genome Sequencing and EST Analysis of the Wheat Pathogen *Stagonospora nodorum* [W] [OA]

James K. Hane,[a,1] Rohan G.T. Lowe,[a,1] Peter S. Solomon,[a] Kar-Chun Tan,[a] Conrad L. Schoch,[b] Joseph W. Spatafora,[b] Pedro W. Crous,[c] Chinappa Kodira,[d] Bruce W. Birren,[d] James E. Galagan,[d] Stefano F.F. Torriani,[e] Bruce A. McDonald,[e] and Richard P. Oliver[a,2]

[a] Australian Centre for Necrotrophic Fungal Pathogens, Murdoch University, WA 6150, Australia
[b] Department of Botany and Plant Pathology, Oregon State University, Corvallis, Oregon 97331
[c] Centraalbureau voor Schimmelcultures, 3508 AD Utrecht, The Netherlands
[d] The Broad Institute, Cambridge, Massachusetts 02141-2023
[e] Plant Pathology, Institute of Integrative Biology, LFW B16 8092 Zurich, Switzerland

Stagonospora nodorum **is a major necrotrophic fungal pathogen of wheat (*Triticum aestivum*) and a member of the Dothideomycetes, a large fungal taxon that includes many important plant pathogens affecting all major crop plant families. Here, we report the acquisition and initial analysis of a draft genome sequence for this fungus. The assembly comprises 37,164,227 bp of nuclear DNA contained in 107 scaffolds. The circular mitochondrial genome comprises 49,761 bp encoding 46 genes, including four that are intron encoded. The nuclear genome assembly contains 26 classes of repetitive DNA, comprising 4.5% of the genome. Some of the repeats show evidence of repeat-induced point mutations consistent with a frequent sexual cycle. ESTs and gene prediction models support a minimum of 10,762 nuclear genes. Extensive orthology was found between the polyketide synthase family in *S. nodorum* and *Cochliobolus heterostrophus*, suggesting an ancient origin and conserved functions for these genes. A striking feature of the gene catalog was the large number of genes predicted to encode secreted proteins; the majority has no meaningful similarity to any other known genes. It is likely that genes for host-specific toxins, in addition to ToxA, will be found among this group. ESTs obtained from axenic mycelium grown on oleate (chosen to mimic early infection) and late-stage lesions sporulating on wheat leaves were obtained. Statistical analysis shows that transcripts encoding proteins involved in protein synthesis and in the production of extracellular proteases, cellulases, and xylanases predominate in the infection library. This suggests that the fungus is dependant on the degradation of wheat macromolecular constituents to provide the carbon skeletons and energy for the synthesis of proteins and other components destined for the developing pycnidiospores.**

INTRODUCTION

Stagonospora nodorum (syn. *Phaeosphaeria*) is a major pathogen of wheat (*Triticum aestivum*) in most wheat-growing areas of the world. It is the major cause of losses due to foliar pathogens in Western Australia and north central and northeastern North America (Solomon et al., 2006a), and until the 1970s, it was the major foliar necrotrophic pathogen in Europe (Bearchell et al., 2005). Infection begins when spores (ascospores or asexual pycnidiospores) land on leaf tissue (Bathgate and Loughman, 2001; Solomon et al., 2006b). The spores rapidly germinate to produce hyphae that invade the leaf, using hyphopodia to gain entry to epidermal cells or by growing directly through stomata (Solomon et al., 2006b). The hyphae rapidly colonize the leaves and begin to produce pycnidia in 7 to 10 d. The infection has been divided into three metabolic phases (Solomon et al., 2003a). The first phase is penetration of the host epidermis and is fuelled predominantly by lipid stores in the spores (Solomon et al., 2004a); the second is proliferation throughout the interior of the leaf, involves toxin release (Liu et al., 2006), and uses host-derived simple carbohydrate sources (Solomon et al., 2004a); the final phase produces the new spores, but so far the metabolic requirements for pycnidiation are unclear.

The infection epidemiology follows a polycyclic pattern with repeated cycles of both asexual (Bathgate and Loughman, 2001) and sexual (B.A. McDonald, unpublished data) infection throughout the growing season. New rounds of infection are initiated by rain-splash of pycnidiospores and wind dispersal of ascospores. Eventually, wheat heads become infected, causing the glume blotch symptom. In Mediterranean climates, the fungus oversummers in senescent straw and stubble. Seed transmission can be important but is readily controlled by fungicides. Infected stubble harbors the pseudothecia that produce airborne (Bathgate

[1] These authors contributed equally to this work.
[2] Address correspondence to roliver@murdoch.edu.au.
The author responsible for distribution of materials integral to the findings presented in this article in accordance with the policy described in the Instructions for Authors (www.plantcell.org) is: Richard P. Oliver (roliver@murdoch.edu.au).
[W] Online version contains Web-only data.
[OA] Open Access articles can be viewed online without a subscription.
www.plantcell.org/cgi/doi/10.1105/tpc.107.052829

and Loughman, 2001), heterothallic (Bennett et al., 2003) ascospores to reinitiate the infection in the following growing season. Epidemiological and population genetic evidence suggest that the fungus undergoes meiosis in the field regularly and that gene flow is global (Stukenbrock et al., 2006).

S. nodorum is a member of the Dothideomycetes class of filamentous fungi. This is a newly recognized major class that broadly replaced the long-recognized loculoascomycetes (Winka and Eriksson, 1997) and includes the causal organisms of many economically important plant diseases; notable examples are black leg (*Leptosphaeria maculans*), southern maize (*Zea mays*) leaf blight (*Cochliobolus heterostrophus*), barley (*Hordeum vulgare*) net blotch (*Pyrenophora teres*), apple scab (*Venturia inaequalis*), black sigatoka (*Mycosphaerella fijiensis*) of banana (*Musa* spp), wheat leaf blotch (*M. graminicola*), tomato (*Solanum lycopersicum*) leaf mold (*Passalora fulva*), and Ascochyta blight of many legume species (*Ascochyta* spp). The taxon also includes large numbers of saprobic species occurring on substrates ranging from dung to plant debris, a few species associated with animals, and several lichenized species (Del Prado et al., 2006).

Many plant pathogens in this group produce host-specific toxins (Wolpert et al., 2002). These molecules interact with specific host factors to produce disease symptoms only in selected genotypes of specific plant hosts. Well-known examples are found in *Alternaria alternata*, *Cochliobolus heterostrophus*, and *Pyrenophora tritici-repentis*, while species in the genera *Mycosphaerella*, *Corynespora*, and *Stemphylium* are also thought to produce host-specific toxins (Agrios, 2005). Proteinaceous host-specific toxins have recently been shown to be important virulence determinants in *S. nodorum* (Liu et al., 2004a, 2004b), including one whose gene is thought to have been interspecifically transferred to the wheat tan spot pathogen, *P. tritici-repentis* (Friesen et al., 2006). Only one example of a host-specific toxin has been identified outside of the Dothideomycetes (Wolpert et al., 2002). Non-host-specific toxins produced by species in this group include cercosporin (*Cercospora*) and solanopyrone (*Ascochyta*) but are also produced by many other fungal taxa (Agrios, 2005).

S. nodorum is an experimentally tractable organism, which is easily handled in defined media, was one of the first fungal pathogens to be genetically manipulated (Cooley et al., 1988), and has been a model for fungicide development (Dancer et al., 1999). Molecular analysis of pathogenicity determinants is aided by facile tools for gene ablation and rapid in vitro phenotypic screens, and thus far, a small number of genes required for pathogenicity have been identified (Cooley et al., 1988; Bailey et al., 1996; Bindschedler et al., 2003; Solomon et al., 2003b, 2004a, 2004b, 2005, 2006a). It has thus emerged as a model for dothideomycete pathology.

Whole-genome sequences have been described for a handful of fungal saprobes and pathogens (Galagan et al., 2003, 2005; Jones et al., 2004; Dean et al., 2005; Kamper et al., 2006). Here, we present an initial analysis of the genome sequence of the dothideomycete *S. nodorum*. Gene expression studies using EST libraries from axenic mycelium and heavily infected wheat leaves complement the genome sequence and provide a broad-based analysis of the genomic basis of infection by *S. nodorum*.

RESULTS

Acquisition and Analysis of the Genome Sequence

The genome sequence was obtained using a whole-genome shotgun approach. Approximately 10× sequence coverage was obtained as paired-end reads from plasmids of 4 and 10 kb plus 40-kb fosmids. Reads were assembled using Arachne (Jaffe et al., 2003), forming 496 contigs totaling 37,071 kb. The contig N50 was 179 kb, meaning an average base in the assembly lies within a contig of at least 179 kb. Greater than 98% of the bases in the assembly have a consensus quality score of ≥40, corresponding to the standard error rate of fully finished sequence. The contigs are connected in 109 scaffolds spanning 37,202 kb with a scaffold N50 of 1.05 Mb. More than 50% of the genome is contained in the 13 largest scaffolds. Within the scaffolds, only ~140 kb is estimated to lie within sequence gaps. Thus, in terms of representation, contiguity, and sequence accuracy, the draft assembly is of high quality. The mitochondrial genome comprises a circle of 49,761 bp (GenBank accession number EU053989). It replaces two of the auto-assembled scaffolds, 52 and 65. The nuclear genome is therefore assembled in 107 scaffolds of total length 37,164,227 bp.

General features of the assembly are described in Table 1. As detailed below, the genomic sequencing was complemented by sequencing EST libraries. One library was constructed from axenically grown mycelium, and after removing vector, poly(A) sequences, poor-quality sequences, and sequences <100 bp, there were 7750 remaining ESTs. Of these, 97.6% mapped to the genome assembly via Sim4, and 1.4% mapped to unassembled reads. This indicates that the assembly has achieved good coverage of the genome. In all, seven EST contigs and 13 singletons clustered to the unassembled reads.

Analysis of Repetitive Elements

Repetitive elements were identified de novo by identifying sequence elements that existed in 10 or more copies, were greater than 200 bp, and exhibited >65% sequence identity. The analysis revealed the presence of 25 repeat classes. Table 2 lists the general features of the repeats (see Supplemental Data Set 1 online). Only three, Molly, Pixie, and Elsa, had been detected earlier (GenBank accession numbers AJ277502, AJ277503, and AJ277966, respectively). Molly, Pixie, X15, and R37 show

Table 1. The Assembly of the *S. nodorum* Nuclear Genome Sequence

Scaffold count	107
Total/bp	37,164,227
Coverage	~10×
G + C %	50.3%
Scaffold minimum length/bp	2,005
Scaffold maximum length/bp	2,531,949
Scaffold median length/bp	32,376
Scaffold mean length/bp	347,329
Gaps/number	387
Total length of gaps/bp	146,025
Max length of gap/bp	11,204
Median length of gap/bp	325

Table 2. Features of Repetitive Element Classes Found in the Nuclear Genome

Repeat Name	Class	Count	Full Length (bp)	Occupied (bp)	Percentage of Genome	RIP[a]
Subtelomeric						
R22	Telomeric repeat	23	678	12,565	0.03	No
X15	Telomeric repeat/transposon or remnant	37	6,231	89,126	0.24	No
X26	Telomeric repeat	38	4,628	77,020	0.21	No
X35	Telomeric repeat	19	1,157	14,393	0.04	No
X48	Telomeric repeat	22	265	5,631	0.02	No
Ribosomal						
Y1	rDNA repeat	113	9,358	40,1890	1.08	No
Other						
Elsa	Transposon	17	5,240	34,287	0.09	Yes
Molly	Transposon	40	1,862	50,453	0.14	Yes
Pixie	Transposon	28	1,845	39,148	0.11	No
R10	Unknown	59	1,241	44,018	0.12	No
R25	Unknown	23	3,320	44,860	0.12	No
R31	Unknown	23	3,031	40,267	0.11	No
R37	Transposon or remnant	98	1,603	104,915	0.28	No
R38	Unknown	25	358	8,556	0.02	No
R39	Unknown	29	2,050	35,758	0.10	No
R51	Unknown	39	833	25,538	0.07	No
R8	Unknown	48	9,143	275,643	0.74	No
R9	Transposon or remnant	72	4,108	163,739	0.44	No
X0	Unknown	76	3,862	145,268	0.39	No
X11	Transposon or remnant	36	8,555	128,638	0.35	No
X12	Unknown	29	2,263	25,369	0.07	No
X23	Unknown	29	685	11,679	0.03	No
X28	Unknown	30	1,784	32,002	0.09	No
X3	Unknown	213	9,364	463,438	1.24	No
X36	Unknown	10	512	5,067	0.01	No
X96	Unknown	14	308	4,321	0.01	No
Sum					4.52	

[a] Evidence from the alignment that the repeat has been subjected to RIP mutation.

sequence characteristics of inverted terminal repeat–containing transposons, while Elsa, R9, and X11 appear to be retrotransposons. Repetitive elements were found individually throughout the genome but were often found in clusters spanning several kilobases.

Telomere-associated repeats were identified by searching for examples of the canonical telomere repeat TTAGGG at the termini of auto-assembled scaffolds. Physically linked repetitive sequences were then analyzed for association with the TTAGGG sequence repeats. Between 19 and 38 copies of telomere-associated repeats were found in the assembly.

Repeat-induced point (RIP) mutation is a fungal-specific genome-cleansing process that detects repeated DNA at meiosis and introduces C-to-T mutations into the copies (Cambareri et al., 1989). Using the parameters defined for *Magnaporthe grisea* (Dean et al., 2005), we identified RIP-like characteristics in several of the repeat classes (Table 2; see Supplemental Data Set 1 online). The transposons Molly and Elsa were the most clearly affected classes. None of the telomere-associated repeats displayed RIP characteristics.

Mitochondrial Genome

The mitochondrial genome of *S. nodorum* assembled as a circular molecule of 49,761 bp, with an overall G + C content of 29.4%. It contains the typical genes encoding 12 inner mitochondrial membrane proteins involved in electron transport and coupled oxidative phosphorylation (*nad1-6* and *nad4L*, *cytb*, *cox1-3*, and *atp6*), the 5S ribosomal protein, three open reading frames (ORFs) of unknown function, and genes for the large and small rRNAs (*rnl* and *rns*) (Figure 1). The genes were coded on both DNA strands. The 27 tRNAs can carry all 20 amino acids. All tRNA secondary structures showed the expected cloverleaf form, but tRNA-Thr and tRNA-Phe had nine nucleotides in the anticodon loop instead of the typical seven, and tRNA-Arg2 had 11 nucleotides in its anticodon loop (Lowe and Eddy, 1997).

Gene Content

The initial gene prediction process identified 16,957 gene models of which 16,586 were located on the 107 nuclear scaffolds. A revised gene prediction procedure, using new ESTs (see below) and 795 fully supported and manually annotated gene models, identified 10,762 version 2 nuclear gene models. Of these, 617 genes corresponded to a merging of version 1 genes; in 50 cases, three prior gene models were merged, and in one case, four genes were merged. ESTs that aligned to unassembled reads identified one supported gene (SNOG_20000.2). A total of 5354 version 1 gene models were not supported but did not

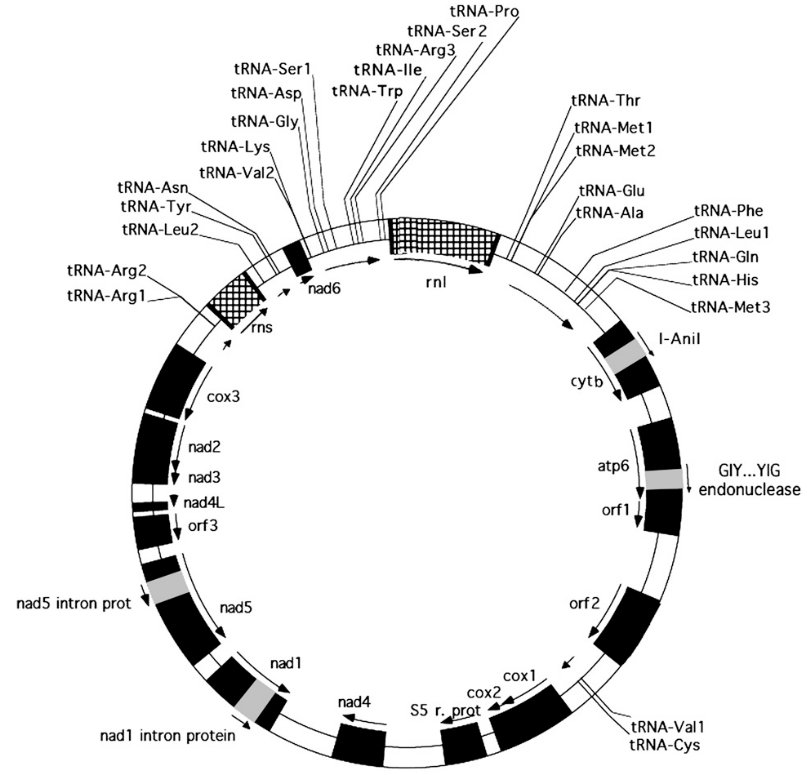


Figure 1. The Structure of the Mitochondrial Genome.

Black segments are exons, hatched segments are rRNA genes, and white segments are introns. The direction of transcription is indicated with the arrows.

conflict with second-round predictions. We therefore conclude that *S. nodorum* contains a minimum of 10,762 nuclear genes of which all but 125 are supported by two gene prediction procedures and 2696 are supported by direct experimental evidence via EST alignment. These genes, with an identification format SNOG_xxxxx.2, were compared with the GenBank nonredundant protein database at the National Center for Biotechnology Information (NCBI). Informative (not hypothetical, predicted, putative, or unknown) BLASTP (Altschul et al., 1990) hits with e-values $<1 \times 10^{-6}$ were found for 7116 genes (see Supplemental Data Set 2 online). It is estimated that at least 46.6% of the nuclear genome is transcribed and 38.8% is translated.

The 5354 gene models without support from the reannotation have unaltered accession numbers as SNOG_xxxxx.1 and are retained for further possible validation and analysis. As 952 of these unsupported gene models have BLASTP hits with e-values $<1 \times 10^{-6}$, we predict that some will be validated as new evidence emerges.

Gene Expression during Infection

Two EST libraries were constructed and analyzed as part of this project. An in vitro library was constructed from axenic fungal mycelium transferred to media with oleate as the sole carbon source; this is referred to as the oleate library. An in planta library was made from bulked sporulating disease lesions on wheat 9, 10, and 11 d after infection (DAI). For both libraries, 5000 random clones were sequenced in both directions. The oleate library was entirely fungal and hence particularly suited for primary genome annotation purposes. The lipid growth media was chosen to replicate the early stages of infection in which lipolysis, the glyoxalate cycle, and gluconeogenesis are thought to be critical metabolic requirements (Solomon et al., 2004a). The in planta library was designed to reveal both plant and fungal genes expressed at a late stage of infection. After trimming and removal of plant ESTs and alignment to the genome assembly, the in planta library formed 1448 and the oleate library formed 1231 unigenes (see Supplemental Data Set 2 online). Only 427 of the unigenes were found in both libraries. Although this represents 19% of the unigenes, it encompasses 46% of the ESTs, showing the genes expressed uniquely in one library are of relatively low expression.

The unigenes obtained from the in planta and oleate libraries were classified according to gene ontology (GO) categories. GO matches were obtained for 851 (59%) of the unigenes found in the in planta and 736 (59%) of those found in the oleate library. Relative numbers of ESTs and genes in different GO classes were compared (Table 3). GO categories were ranked by statistical discrimination between the two libraries, and the top 10 are shown for each of biological process, cellular component, and molecular function.

Coexpression of fungal gene clusters responsible for the synthesis of secondary metabolites (Bok and Keller, 2004) and pathogenicity (Kamper et al., 2006) has been observed. We compared the number of ESTs found in the in planta and oleate libraries to search for clusters of coexpressed genes. Using a window of 10 contiguous genes, we scanned for regions with biased ratios of ESTs from either library. One putatively in planta–induced region stood out. These six genes, SNOG_16151.2 to SNOG_16157.2, were exclusively in the in planta library with 4, 20, 0, 4 10, and 1 EST clones, respectively. The genes have best hits to a major facilitator superfamily transporter, a phytanoyl-CoA dioxygenase, a CocE/NonD hydrolase, and a salicylate hydroxylase and are next to a transcription factor. Such a cluster may be involved in the degradation of phytols, phenylpropanoids, and catechols either for nutritional purposes, providing trichloroacetic acid intermediates via the β-ketoadipate pathway or to detoxify wheat defense compound(s). It is intriguing that overexpression of the thiolase in the β-ketoadipate pathway in *L. maculans* markedly reduced pathogenicity (Elliott and Howlett, 2006). Experiments to test these ideas in *S. nodorum* are in progress.

DISCUSSION

The estimated genome size of *S. nodorum* is 37.2 Mbp, which is significantly larger than the 28 to 32 Mbp previously estimated by pulsed-field gel analysis (Cooley and Caten, 1991). Electrophoretic karyotypes have proven to be unreliable estimators of total genome size in many fungal species (Orbach et al., 1988; Orbach, 1989; Talbot et al., 1993). In the case of *S. nodorum*, this discrepancy may be a consequence of comigration of chromosomal bands on pulsed-field gels leading to an underestimation of genome size (Cooley and Caten, 1991). The electrophoretic karyotype was found to be highly variable between strains, even when isolated from a single ascus, consistent with the generalization that many fungal species have plastic genome structures (Zolan, 1995). The revised genome size estimate is comparable to other sequenced filamentous fungi, such as the rice pathogen *M. grisea*, which is currently estimated to be 41.6 Mbp, and the nonpathogen *Neurospora crassa* now estimated at 39.2 Mbp.

Phylogenetic Analysis

The genome sequence has confirmed the phylogenetic placement of *S. nodorum* in the class Dothideomycetes. Along with the Eurotiomycetes (containing *Aspergillus* and human pathogens such as *Histoplasma*), Lecanoromycetes (the majority of the lichenized species), Leotiomycetes (containing numerous endophytes and the plant pathogen *Sclerotinia*), and Sordariomycetes (with *Neurospora*, *Magnaporthe*, and *Colletotrichum* spp), the Dothideomycetes is now recognized as a major clade of the filamentous Ascomycota (James et al., 2006). Phylogenetic analyses of full fungal genomes and large-scale taxon sampling agree with the placement of *S. nodorum* and point to the rapid divergence of these major classes of ascomycetes (Robbertse et al., 2006; Spatafora et al., 2006). The tree in Figure 2 is a focused sampling of the largest classes of the Ascomycota with an emphasis on plant pathogenic species and those with genome sequences. The Dothideomycetes is supported as a single class and represented by samples of five of the nine currently proposed orders (Eriksson, 2006; Schoch et al., 2006). *S. nodorum* is placed in the Pleosporales, a large order containing >100 genera and several thousand species, many of which are important plant pathogens. The Pleosporaceae family contains *Alternaria*, *Cochliobolus*, and *Pyrenophora* (Kodsueb et al., 2006) and is closely related to the clade containing the genera *Leptosphaeria* and *Phaeosphaeria* (*Stagonospora*). Other orders in the Dothideomycetes include the Dothideales and the Capnodiales (Schoch et al., 2006), which contains the pathogen genera *Mycosphaerella* and *Cladosporium*.

Repeated Elements in the Nuclear Genome

Prior to this study, only one unpublished study had been made of repetitive elements in the *S. nodorum* genome (Rawson, 2000). The de novo analysis of repeats is likely to become a feature of future genome sequencing projects as organisms with little prior molecular work are selected. The total amount of repetitive DNA in the *S. nodorum* nuclear genome is estimated at 4.5%. This compares with 9.5% in *M. grisea*. Some of these repeats could be associated with telomeres. Between 19 and 38 copies of telomere-associated repeats were found in the assembly. These numbers accord well with the 14 to 19 chromosomes visualized by pulsed-field electrophoresis (Cooley and Caten, 1991).

Studies of *S. nodorum* life cycle have indicated that it undergoes regular sexual crossing, particularly in areas with Mediterranean-style climates with the associated need to over-summer as ascospores (Bathgate and Loughman, 2001; Stukenbrock et al., 2006). The relatively low content of repetitive DNA is consistent with this observation. The prevalence of clear cases of RIP further confirms that meiosis is a frequent event. RIP has been found to varying degrees in other fungal genomes. It is notable that the closely related *L. maculans* has previously been shown experimentally to exhibit RIP (Idnurm and Howlett, 2003).

The Mitochondrial Genome

A distinctive feature of fungal mitochondrial genomes is the clustering of tRNA genes (Ghikas et al., 2006), and it is thought that both the tRNA gene content and their placement will be conserved in fungi (Table 4). The 27 tRNA genes clustered into five groups, with the two larger tRNA gene clusters flanking *rnl*, a pattern common to other fungal mitochondrial DNAs (mtDNAs) (Tambor et al., 2006). The tRNA gene cluster 5′ to *rnl* had GDS^{1}WIS^{2}P as a consensus, with its closest sequenced relative *M. graminicola* (GenBank accession number EU090238), while the 3′-downstream consensus was EAFLQHM, having many tRNA genes in the same order found in other Ascomycetes. Variation in the order of tRNAs, such as inversions in *Epidermophyton floccosum* (Ile-Trp) or in *Podospora anserina* and *S.*

Table 3. Comparison of GO Classification of Unigene and Transcript Numbers between the in Planta and Oleate Libraries

Biological Processes

GO Identifier	Description	Loci	In Planta	*Oleate*	P Value[a]
Upregulated in planta					
GO:0006412	Protein biosynthesis	128	710	325	5.97E-19
GO:0045493	Xylan catabolism	24	31	0	6.16E-09
GO:0006508	Proteolysis	78	110	39	8.29E-07
GO:0008643	Carbohydrate transport	58	22	0	1.26E-06
GO:0000272	Polysaccharide catabolism	12	24	1	4.29E-06
GO:0016068	Type I hypersensitivity	13	87	31	9.67E-06
GO:0030245	Cellulose catabolism	16	18	0	1.33E-05
GO:0005975	Carbohydrate metabolism	102	53	16	6.57E-05
GO:0042732	D-xylose metabolism	2	25	4	2.01E-04
GO:0009051	Pentose-phosphate shunt, oxidative branch	2	21	3	4.07E-04
Upregulated oleate					
GO:0007582	Physiological process	24	8	48	1.04E-10
GO:0006629	Lipid metabolism	20	6	40	1.42E-09
GO:0006096	Glycolysis	21	42	92	5.93E-09
GO:0006108	Malate metabolism	4	12	45	5.48E-08
GO:0006099	Tricarboxylic acid cycle	20	50	85	4.11E-06
GO:0015986	ATP synthesis coupled proton transport	26	34	65	6.56E-06
GO:0006183	GTP biosynthesis	1	0	13	1.54E-05
GO:0006228	UTP biosynthesis	1	0	13	1.54E-05
GO:0006241	CTP biosynthesis	1	0	13	1.54E-05
GO:0006334	Nucleosome assembly	12	65	95	3.21E-05

Cellular Components

GO Identifier	Description	Loci	In Planta	*Oleate*	P Value[a]
Upregulated in planta					
GO:0005840	Ribosome	89	567	262	3.14E-15
GO:0015935	Small ribosomal subunit	11	93	33	4.86E-06
GO:0005576	Extracellular region	36	23	1	7.44E-06
GO:0005730	Nucleolus	6	23	2	4.15E-05
GO:0030529	Ribonucleoprotein complex	25	60	19	4.31E-05
GO:0030125	Clathrin vesicle coat	9	7	0	8.86E-03
GO:0016020	Membrane	338	127	122	1.07E-02
GO:0016021	Integral to membrane	401	243	213	1.38E-02
GO:0019867	Outer membrane	7	45	24	1.40E-02
GO:0005874	Microtubule	19	8	1	1.97E-02
Upregulated oleate					
GO:0005829	Cytosol	34	19	50	1.01E-06
GO:0016469	Proton-transporting two-sector ATPase complex	24	27	60	1.41E-06
GO:0000786	Nucleosome	11	65	95	3.21E-05
GO:0043234	Protein complex	33	15	29	1.23E-03
GO:0005739	Mitochondrion	102	138	147	1.62E-03
GO:0005634	Nucleus	288	183	186	1.90E-03
GO:0005777	Peroxisome	14	26	38	3.39E-03
GO:0005746	Mitochondrial electron transport chain	8	36	47	4.40E-03
GO:0045261	Proton-transporting ATP synthase complex, catalytic core F(1)	5	14	23	7.49E-03
GO:0005778	Peroxisomal membrane	3	1	7	8.63E-03

Molecular Functions

GO Identifier	Description	Loci	In Planta	*Oleate*	P Value[a]
Upregulated in planta					
GO:0003735	Structural constituent of ribosome	111	709	328	2.09E-18
GO:0004553	Hydrolase activity, hydrolyzing *O*-glycosyl compounds	83	55	5	4.13E-10
GO:0004252	Ser-type endopeptidase activity	14	43	3	6.92E-09
GO:0005351	Sugar porter activity	64	22	0	1.26E-06
GO:0046556	α-*N*-arabinofuranosidase activity	6	19	0	7.39E-06

(Continued)

Table 3. (continued).

Molecular Functions GO Identifier	Description	Loci	In Planta	*Oleate*	P Value[a]
GO:0030248	Cellulose binding	19	18	0	1.33E-05
GO:0004029	Aldehyde dehydrogenase (NAD) activity	1	15	0	7.85E-05
GO:0050661	NADP binding	6	23	3	1.60E-04
GO:0004185	Ser carboxypeptidase activity	8	13	0	2.56E-04
GO:0004616	Phosphogluconate dehydrogenase (decarboxylating) activity	4	22	3	2.56E-04
Upregulated oleate					
GO:0004459	L-lactate dehydrogenase activity	2	5	44	2.07E-11
GO:0030060	L-malate dehydrogenase activity	2	5	44	2.07E-11
GO:0005498	Sterol carrier activity	3	7	46	1.02E-10
GO:0008415	Acyltransferase activity	17	4	34	4.64E-09
GO:0005506	Iron ion binding	91	89	128	3.72E-06
GO:0004550	Nucleoside diphosphate kinase activity	1	0	13	1.54E-05
GO:0005554	Molecular function unknown	49	40	65	7.95E-05
GO:0046933	Hydrogen-transporting ATP synthase activity, rotational mechanism	24	34	58	8.73E-05
GO:0046961	Hydrogen-transporting atpase activity, rotational mechanism	24	34	58	8.73E-05
GO:0050660	FAD binding	27	13	30	2.85E-04

[a] Probability of significant difference between the libraries (Audic and Claverie, 1997).

nodorum (Met-His), suggest that rearrangements are common in fungal mtDNAs (Table 4).

Kouvelis et al. (2004) argued that gene pairs *nad2-nad3*, *nad1-nad4*, *atp6-atp8*, and *cytb-cox1* would remain joined in ascomycetes with some possible exceptions, as already detected in *M. graminicola* that present only two of these gene pairs coupled. In *S. nodorum*, the *atp8-9* genes were not present, *cytb-cox1* were uncoupled, and the *nad1* and *nad4* genes were on sections of the mtDNA in an inverted orientation. Two *orfs* (*orf1* and *orf2*) had homologous sequences in the in planta EST library. While the *M. graminicola* mtDNA genome had no introns, four intron-encoded genes were found in *S. nodorum*, with high homology to LAGLIDADG-type endonucleases or GIY-YIG–type nucleases (see Supplemental Table 1 online). Intron-encoded proteins have also been reported in *P. anserina* (Cummings et al., 1990) and *Penicillium marneffei* (Woo et al., 2003).

Functional Analysis of Proteins

Our primary goal in obtaining the genome sequence was to reveal genes likely to be involved in pathogenicity. While some genes seem to be specifically associated with pathogenicity in a single organism, other genes and gene families have been generally associated with pathogenicity, albeit they are also found in nonpathogens (see http://www.phi-base.org/about.php; Baldwin et al., 2006). Some of the generic functions are listed (Table 5; see Supplemental Data Set 2 and Supplemental Table 2 online). EST support was found for 59 of the genes, with no statistically significant difference in the numbers of ESTs in the in planta and oleate libraries.

Nonribosomal peptide synthetases (NRPSs) are modular enzymes that synthesize a diverse set of secondary metabolites, including the dothideomycete host-specific toxins AM-toxin, HC-toxin, and victorin from Alternaria and Cochliobolus species (Wolpert et al., 2002). NRPS genes from *C. heterostrophus* have been analyzed in detail (Lee et al., 2005). Comparison of both protein sequences and domain structure of the *C. heterostrophus* NRPS genes (*NPS1-11*) with the eight identified in *S. nodorum* was undertaken (Table 6). Putative orthology was determined by identifying reciprocal best hits among the gene sets. Three pairs were identified. SNOG_02134.2 was linked to *NPS2*, and both appear to be orthologous to the *Aspergillus nidulans* gene *SidC* (Eisendle et al., 2003), which is responsible for the synthesis of ferricrocin, an intracellular iron storage and transport compound involved in protection against iron toxicity (Eisendle et al., 2006). It is likely that SNOG_02134.2 and *NPS2* play similar roles. SNOG_14638.2 was linked to *NPS6*, a ubiquitous gene with a related role in siderophore-mediated iron uptake and oxidative stress protection (Oide et al., 2006). Third, SNOG_14834.2 appears to be directly related to *NPS4*, *Psy1* from *Alternaria brassicae* (Guillemette et al., 2004), and *NPS2* from *Alternaria brassicicola* (Kim et al., 2007). Ab *NPS2* mutants are reduced in virulence, and the gene is predicted to encode a component of the conidial wall. It is likely that SNOG_14834.2 will have a similar generic role. Reciprocal best-hit relationships were observed between a further two SNOG NRPS genes and other fungal genes; SNOG_09081.2 was related to *PesA* (Bailey et al., 1996) and SNOG_14908.2 to the *Salps2* gene from *Hypocrea lixii* (Vizcaino et al., 2006). These genes plus SNOG_14834.2, SNOG_09488.2, and SNOG_01105.2 are all closely related to NPS4, suggesting that this five-gene subfamily is expanded in *S. nodorum*. Interestingly, a further seven NRPS genes from *C. heterostrophus* (*NPS3*, *5*, *8*, *10*, *11*, and *12*) have no obvious ortholog in *S. nodorum*, suggesting expansion of this subfamily in the maize pathogen (Yoder and Turgeon, 2001). Intensive studies in *C. heterostrophus* showed that individual gene knockouts produced altered phenotypes only for *NPS6* (Lee et al., 2005), indicating redundancy of function. The smaller complement of NRPS genes in *M. grisea* (eight) and *S. nodorum* indicate that these organisms may be more fruitful models to study NRPS

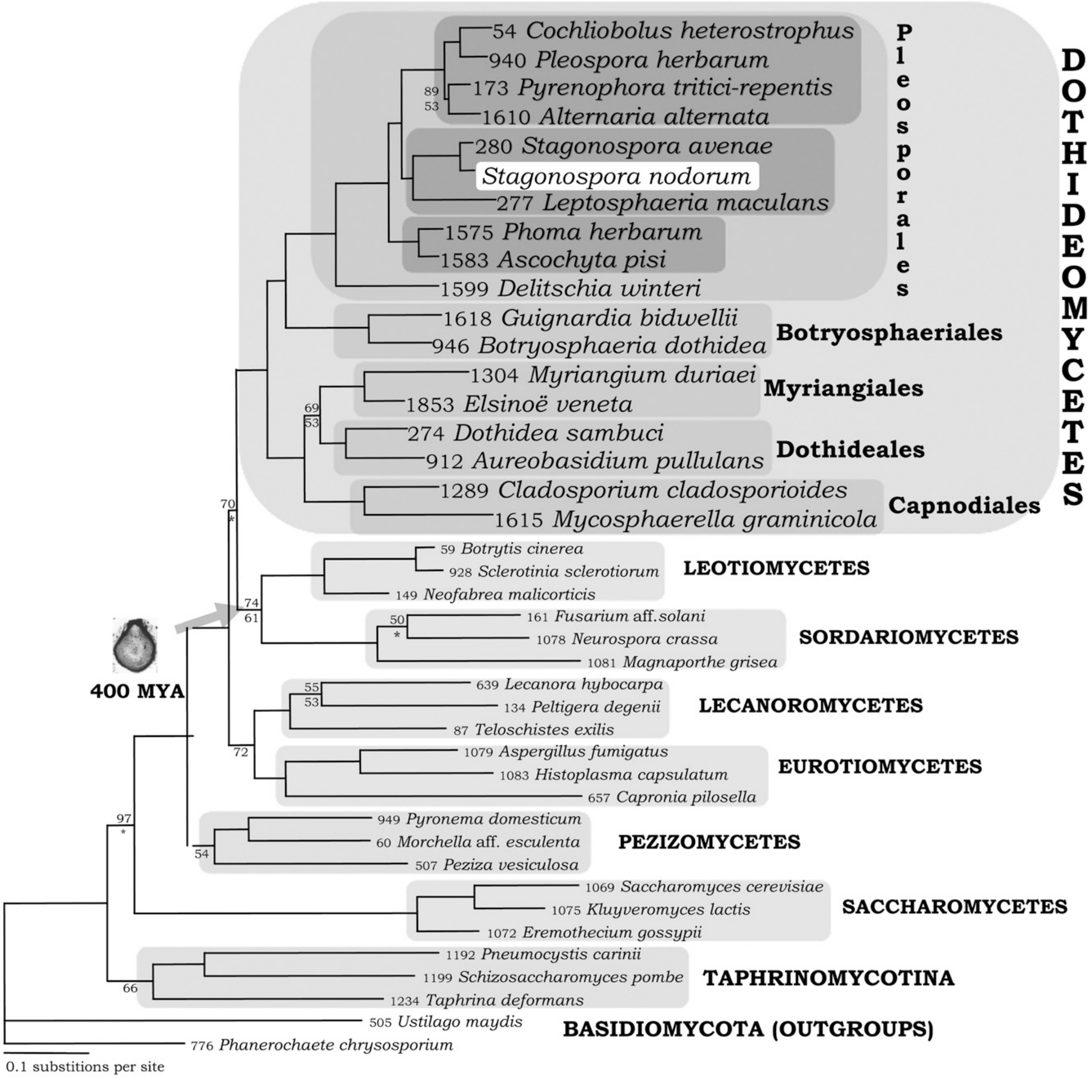


Figure 2. Phylogeny of Ascomycota Focused on Dothideomycetes.

Each species name is preceded by a unique AFTOL ID number (www.aftol.org). The tree is a 50% majority rule consensus tree of 45,000 trees obtained by Bayesian inference. All nodes had posterior probabilities of 100% except where numbers are shown above nodes. Similarly, all nodes had maximum likelihood bootstraps above 80% except where numbers are shown below nodes. Asterisks indicate nodes that were not resolved in >50% of bootstrap trees. Alignment data are provided in Supplemental Data Set 3 online. The gene data used are listed in Supplemental Table 3 online.

Table 4. Comparison of tRNA Gene Clusters Flanking the *rnl* Gene of the mtDNA Genome in Several Ascomycetes[a]

Organism	5′-Upstream Region[b]	rnl	3′-Downstream Region[b]	Accession Number
P. marneffei	RKG^{1}G^{2}DS^{1}WIS^{2}P	*rnl*	TEVM^{1}M^{2}L^{1}AFL^{2}QM^{3}H	AY347307
A. niger	KGDS^{1}WIS^{2}P	*rnl*	TEVM^{1}M^{2}L^{1}AFL^{2}QM^{3}H	DQ217399
M. graminicola	GDS^{1}WIS^{2}P	*rnl*	M^{1}L^{1}EAFL^{2}YQM^{2}HRM^3	EU090238
S. nodorum	VKGDS^{1}WIRS^{2}P	*rnl*	T M^{1}M^{2}EAFLQHM^3	EU053989
E. floccosum	KGDSIWSP	*rnl*	TEVM^{1}M^{2}L^{1}AFL^{2}QM^{3}H	AY916130
H. jecorina	ISWP	*rnl*	TEM^{1}M^{2}L^{1}AFKL^{2}QHM^3	AF447590
P. anserina	ISP	*rnl*	TEIM^{1}L^{1}AFL^{2}Q$\underline{\text{HM}^2}$	X55026

[a] The tRNA gene order of listed organisms is based on GenBank sequences.

[b] Capital letters refer to tRNA genes for the following: R, Arg; K, Lys; G, Gly; D, Asp; S, serine; W, Trp; I, Ile; P, Pro; T, Thr; E, Glu; V, Val; L, Leu; A, Ala; F, Phe; Q, Gln; H, His; Y, Tyr.

The numbers (1 to 3) indicate the presence of more tRNA genes for the same amino acid in the consensus sequence.

function. Furthermore, although toxins have been implicated in the virulence of both pathogens, nonribosomal synthetases cannot be expected to be a major source. No ESTs were found corresponding to these genes. This may be due to their large size or to a generally low level of expression.

Polyketide synthases (PKSs) are a second modular gene family strongly associated with pathogenicity (Kroken et al., 2003; Gaffoor et al., 2005), being responsible for the production of T-toxin from *C. heterostrophus* and PM-toxin from *Mycosphaerella zeae-maydis* (Yun et al., 1998; Baker et al., 2006). *S. nodorum* is predicted to contain 19 PKS genes compared with the 24, 14, and 24 in the pathogens *M. grisea*, *Fusarium graminearum*, and *C. heterostrophus*, respectively, and 28 in the saprobe *A. nidulans* but significantly more than the seven in *N. crassa* (Table 5). Orthology relationships between the well-studied *F. graminearum* and *C. heterostrophus* gene sets and *S. nodorum* are shown in Table 7. Reciprocal best BLAST hits were observed with eight genes from *C. heterostrophus* and five from *F. graminearum*. This close relationship, particularly with *C. heterostrophus*, reflects the close phylogenetic relationship and suggests an ancient origin for the majority of the PKS paralogs (Kroken et al., 2003). SNOG_11981.2 is supported by five ESTs from the in planta library and none from the oleate library, consistent with upregulation during infection and sporulation (no other PKSs have EST support). This gene is orthologous to nonreducing clade 2 PKS genes that are associated with 1,8-dihydroxynaphthalene melanin biosynthesis in *Bipolaris oryzae* and many other pathogens (Kroken et al., 2003; Moriwaki et al., 2004; Amnuaykanjanasin et al., 2005). We showed previously that *S. nodorum* produces melanin from dihydroxyphenlyalanine (Solomon et al., 2004b), for which a PKS would not be necessary. This finding suggests that either *S. nodorum* produces 1,8-dihydroxynaphthalene melanin in addition to dihydroxyphenlyalanine melanin or that SNOG_11981.2 plays another role.

A single PKS-NRPS hybrid protein (SNOG_00308.2.) was identified in the genome of *S. nodorum* as was found for *F. graminearum* (Gaffoor et al., 2005) and *C. heterostrophus* (Lee et al., 2005). *C. heterostrophus NPS7* and SNOG_00308.2 do not appear to be closely related, and it is likely that they evolved independently. By contrast, the hybrid protein of *F. graminearum* (Gz *FUS1*/FG12100) is closely related to SNOG_00308.2.

Table 5. Comparison of Selected Gene Families Identified by PFAM (Release 21) Domains between *S. nodorum* and Latest BROAD Releases of *M. grisea*, *N. crassa*, and *A. nidulans* and Incomplete Data from *C. heterostrophus* and *F. graminearum* (Kroken et al., 2003; Gaffoor et al., 2005; Lee et al., 2005)

	S. nodorum							
Gene Family	Genes with PFAM Match	Genes with EST Clones	In Planta Clones	Oleate Clones	*M. grisea* Release 5	*N. crassa* v3 Assembly 7	*A. nidulans* Release 3	*C. heterostrophus* or *F. graminearum*
G-alpha	4	2	3	0	3	3	3	
Cfem	4	1	17	15	9	6	4	9 Ch 8 Fg
Rhodopsin	2	2	14	8	1	2	1	
Hydrophobin class 1	0	0	0	0	1	1	3	
Hydrophobin class 2	2	0	0	0	2	1	0	
Feruloyl esterase	8	0	0	0	10	1	5	
Cutinase	11	4	4	8	16	3	4	
Subtilisin	11	3	8	2	26	6	3	
Transcription factor	94	25	37	43	97	83	195	
Cytochrome p450	103	20	27	20	115	39	102	
PKS	19	1	3	0	24	7	28	24 Ch 14 Fg
NRPS	8	1	1	0	8	3	13	10 Ch
PKS-NRPS	1	0	0	0	7	0	0	1 Ch

Table 6. Potential Orthologs to *S. nodorum* NRPS Genes

SN15 NRPS	Domain/ Module Structure[a]	Best Hit	Accession Number	Organism	Reciprocal[b]	Percentage of Similarity[c]	Domain/Module Structure[d]	Inferred Function
SNOG_01105.2	TCyAT/CATE/CAT/CAT	*NPS4*	AAX09986	*C. heterostrophus*	No	37.2%	TECAT/CATE/CAT/CATE/C/T/C/T	Unknown
		Abre *Psy1*	AAP78735	*A. brassicae*	No	39.0%	TECAT/CATE/CAT/CATE/C/T/C/T	
SNOG_02134.2	AT/CAT/CAT/C/T/C	*NPS2*	AAX09984	*C. heterostrophus*	Yes	53.8%	AT/CAT/CAT/CATT/C	Iron uptake and oxidative stress protection
		SidC	AAP56239	*A. nidulans*	Yes	40.3%	A/CA/CA/C/T/C	
SNOG_07021.2	ATCT	*NPS1*	AAX09983	*C. heterostrophus*	No	13.8%	AT/CAMT/CAT/C	Unknown
		PesA	CAA61605	*Metarhizium anisopliae*	No	10.2%	ATE/CAT/CAT/CATE/C	Unknown
SNOG_09081.2	ATE/CAT/CAT/CAT/C	*NPS4*	AAX09986	*C. heterostrophus*	No	33.0%	TECAT/CATE/CAT/CATE/C/T/C/T	Unknown
		PesA	CAA61605	*M. anisopliae*	Yes	58.1%	ATE/CAT/CAT/CATE/C	Unknown
SNOG_09488.2	CATE/CAT/CATE	*NPS4*	AAX09986	*C. heterostrophus*	No	26.3%	TECAT/CATE/CAT/CATE/C/T/C/T	Unknown
		PesA	CAA61605	*M. anisopliae*	No	36.6%	ATE/CAT/CAT/CATE/C	Unknown
SNOG_14098.2	ATE/CAT/CAT/CATE/C	*NPS4*	AAX09986	*C. heterostrophus*	No	32.3%	TECAT/CATE/CAT/CATE/C/T/C/T	Unknown
		Salps2	CAI38799	*Hypocrea lixii*	Yes	34.4%	TCAT/CAT/CAT	
SNOG_14368.2	AT/C/TT	*NPS6*	AAX09988	*C. heterostrophus*	Yes	73.5%	AT/C/AT	Virulence, siderophore-mediated iron metabolism, tolerance to oxidative stress
		NPS6	ABI51982	*C. miyabeanus*	Yes	73.5%	AT/C/TT	
SNOG_14834.2	TECAT/CATE/CAT/CATE/C	*NPS4*	AAX09986	*C. heterostrophus*	Yes	75.9%	TECAT/CATE/CAT/CATE/C/T/C/T	Virulence, conidial cell wall construction, spore germination/sporulation efficiency
		Ab *NPS2*[e]		*A. brassicicola*	Yes	78.2%	TECAT/CATE/CAT/CATE/C/T/C	
		Abre *Psy1*	AAP78735	*A. brassicae*	Yes	78.0%	TECAT/CATE/CAT/CATE/C/T/C/T	

[a] Domain abbreviations: A, adenylation; C, condensation; Cy, Cyclization; E, epimerization; T, thiolation; Te, Thioesterase. Potential orthologs were identified by best BLASTP hits to NRPS genes in *C. heterostrophus* (Lee et al., 2005) and by best informative hit to the nonredundant database at NCBI of e-value $<1 \times 10^{-10}$.

[b] The best hit of the identified gene in the *S. nodorum* gene set was observed reciprocally.

[c] Percentage of similarity of ortholog pairs (Needleman and Wunsch, 1970) via NEEDLE (Rice et al., 2000).

[d] Domain structure and modular organization for all sequences was determined via the online NRPS-PKS database (Ansari et al., 2004). It should be noted that domain structure predictions may vary slightly from those stated in the original studies.

[e] Protein sequence for Ab NPS2 available in Supplemental Appendix S2 online from Kim et al. (2007).

Overall, they are 54.5% similar, and apart from an acyl binding domain found only in SNOG_00308.2, the domain structures are identical. Gz *FUS1* produces fusarin C, a mycotoxin (Gaffoor et al., 2005). The best hit (with 59.4% similarity) to SNOG_00308.2 in *M. grisea* is *Ace1*, a hybrid PKS/NRPS that confers avirulence to *M. grisea* during rice (*Oryza sativa*) infection (Bohnert et al., 2004). It will be intriguing to determine if the product of SNOG_00308.2 is required for pathogenicity.

G-alpha proteins have been extensively studied in fungi (Lafon et al., 2006), and many are required for pathogenicity. Distinct roles for three g-alpha genes have been revealed in *M. grisea*, *A. nidulans*, and *N. crassa*. It was therefore a surprise to find a fourth g-alpha gene in the *S. nodorum* genome. This gene (SNOG_06158.2) was shown to be expressed in vitro and in planta. It was investigated by gene disruption, and its loss resulted in no discernable phenotype (data not shown). A fourth g-alpha protein has been recently identified in both *A. oryzae* (Lafon et al., 2006) and *Ustilago maydis* (Kamper et al., 2006), but neither shows significant sequence similarity to that of *S. nodorum.* Indeed, SNOG_06158.2 is most similar to *Gba3* among the *U. maydis* g-alphas and *GpaB* among the *A. oryzae* genes.

G-alpha proteins transduce extracellular signals leading to infection-specific development (Solomon et al., 2004b). *Pth11* and *ACI1* are two *M. grisea* genes encoding transmembrane receptors that defined a new protein domain, CFEM (Kulkarni et al., 2003), with roles in g-protein signaling. *M. grisea* has nine

Table 7. Potential Orthologs of *S. nodorum* PKS Genes

SN15 PKS	Domain/Module Structure	Best Hit	Accession Number	Organism	Reciprocal	Similarity	Inferred Clade	Domain/Module Structure	Inferred Function of Ortholog
SNOG_00477.2	KsAtKrAcp	fg12100		*F. graminearum*	No	21.6%	Fungal 6MSAS	KsAtAcp/KrAcp/Acp	6-Methylsalicylic acid synthesis (Fujii et al., 1996)
		PKS25	AAR90279	*C. heterostrophus*	Yes	59.7%		KsAtKrAcp	
		atX	BAA20102	*A. terreus*	Yes	69.0%		KsAtKrAcp	
SNOG_02561.2	KsAtErKrAcp	fg12109		*F. graminearum*	No	42.1%	Reducing PKS clade I	KsAtErKrAcp	Synthesis of diketide moiety of compactin or similar polyketide (Abe et al., 2002a, 2002b)
		PKS3	AAR90258	*C. heterostrophus*	Yes	45.8%		KsAtErKrAcp	
		mlcB	BAC20566	*P. citrinum*	Yes	47.7%		KsAtDhErKrAcp	
SNOG_04868.2	KsAtErKrAcp	fg05794		*F. graminearum*	Yes	38.9%	Reducing PKS clade I	KsAtAcp/ErKrAcp	Synthesis of squalestatin (Nicholson et al., 2001)
		PKS6	AAR90261	*C. heterostrophus*	No	46.5%		KsAtErKr	
		type I PKS	AAO62426	*Phoma* sp C2932	No	49.8%		KsAtErKrAcp	
SNOG_05791.2	KsAtErKrAcp	fg12109		*F. graminearum*	Yes	43.1%	Reducing PKS clade I	KsAtErKrAcp	Synthesis of alternapyrone (Fujii et al., 2005)
		PKS5	AAR90261	*C. heterostrophus*	Yes	73.1%		KsAtErKr	
		atl5	BAD83684	*A. solani*	Yes	79.2%		KsAtErKrAcp	
SNOG_06676.2	KsAtErKr	fg01790		*F. graminearum*	No	31.3%	Reducing PKS clade IV	KsAtErKrAcp	Synthesis of fumonisin (Proctor et al., 1999)
		PKS14	AAR92221	*G. moniliformis*	No	31.8%		KsAtErKr	
		FUM1	AAD43562	*G. moniliformis*	No	30.0%		KsAtErKrAcp	
SNOG_06682.1	KsAtAcp	fg03964		*F. graminearum*	No	36.2%	Nonreducing PKS clade III (uncharacterized)	KsAtAcp	Synthesis of citrinin (Shimizu et al., 2005)
		PKS17	AAR90253	*B. fuckeliana*	Yes	43.9%		KsAtAcp	
		pksCT	BAD44749	*M. purpureus*	Yes	44.7%		KsAtAcp	
SNOG_07866.2	KsAtKr	fg12100		*F. graminearum*	No	37.8%	Reducing PKS clade II	KsAtAcp/KrAcp/Acp	Synthesis of polyketide similar to citrinin/ lovastatin
		PKS16	AAR90270	*C. heterostrophus*	Yes	67.9%		KsAtKr	
		EqiS	AAV66106	*F. heterosporum*	Yes	39.8%		KsAtKrAcp/Acp	
SNOG_08274.2	KsAcp/AtAcp/Acp	fg12040		*F. graminearum*	No	40.3%	Nonreducing PKS clade II (e.g., melanins)	KsAtAcp	Synthesis of perithecial pigment, autofusarin, or similar polyketide (Graziani et al., 2004)
		PKS12	AAR90248	*B. fuckeliana*	No	44.5%		KsAtAcp/Acp	
		PKS	AAS48892	*N. haematococca*	No	47.8%		KsAtAcp/Acp	
SNOG_08614.2	KsAtAcp/Acp/Te	fg12125		*F. graminearum*	Yes	41.2%	Nonreducing PKS clade I	KsAtAcp/Acp	Synthesis of perithecial pigment, cercosporin, or similar polyketide; cercosporin is a reactive oxygen species–generating toxin degrading plant cell membranes (Chung et al., 2003)
		PKS3	AAR92210	*G. moniliformis*	No	49.3%		KsAtAcp/Acp	
		PKS	AAT69682	*C. nicotianae*	No	60.7%		KsAtAcp/Acp	
SNOG_09623.2	KsAtAcp/ErKrAcp	fg01790		*F. graminearum*	No	58.9%	Reducing PKS clade IV	KsAtErKrAcp	Synthesis of fumonisin (Proctor et al., 1999)
		PKS11	AAR90266	*C. heterostrophus*	Yes	56.1%		KsAtErKrAcp	
		FUM1	AAD43562	*G. moniliformis*	Yes	57.7%		KsAtErKrAcp	
SNOG_11066.2	KsKsAtErKrAcp	fg12055		*F. graminearum*	No	43.2%	Reducing PKS clade III	KsAtErKrAcp	High virulence; synthesis of T-toxin; zearalenone (Gaffoor et al., 2005; Baker et al., 2006)
		PKS8	AAR90244	*B. fuckeliana*	Yes	59.7%		KsAtErKrAcp	
		PKS2	ABB76806	*C. heterostrophus*	Yes	44.9%		KsAtErKrAcp	
SNOG_11076.2	KsAtDhErKrAcp	fg01790		*F. graminearum*	Yes	62.8%	Reducing PKS clade IV	KsAtErKrAcp	Synthesis of fumonisin (Proctor et al., 1999)
		PKS14	AAR90268	*C. heterostrophus*	Yes	82.7%		KsAtDHErKrAcp	
		FUM1	AAD43562	*G. moniliformis*	No	53.9%		KsAtErKrAcp	
SNOG_11272.2	KsKsAtErKrAcp	fg12055		*F. graminearum*	No	44.7%	Reducing PKS clade IV	KsAtErKrAcp	Synthesis of zearalenone or similar polyketide; similar to HR-type PKS (Gaffoor et al., 2005)
		PKS14	AAR92221	*G. moniliformis*	Yes	48.7%		KsAtErKr	
		PKSKA1	AAY32931	*Xylaria* sp BCC 1067	No	47.3%		KsAtErKrAcp	
SNOG_11981.2	KsAtAcp/Acp/Acp/Te	fg12040		*F. graminearum*	Yes	54.9%	Nonreducing PKS clade II	KsAtAcp	Synthesis of melanin (Moriwaki et al., 2004; Amnuaykanjanasin et al., 2005).
		PKS18	AAR90272	*C. heterostrophus*	Yes	85.7%		KsAtAcp/Acp/Te	
		PKS1	BAD22832	*B. oryzae*	Yes	88.2%		KsAtAcp/Acp/Te	
SNOG_12897.2	KsAtErKr	fg10548		*F. graminearum*	No	39.7%	Reducing PKS clade I	KsAtErKrAcp	Synthesis of alternapyrone (Fujii et al., 2005)
		PKS5	AAR92212	*G. moniliformis*	No	38.2%		KsAtAcp/ErKrAcp	
		alt5	BAD83684	*A. solani*	No	37.8%		KsAtErKrAcp	
SNOG_13032.2	KsAtAcpKr	fg01790		*F. graminearum*	No	40.4%	Reducing PKS clade IV	KsAtErKrAcp	Synthesis of fumonisin (Proctor et al., 1999)
		PKS12	AAR92219	*G. moniliformis*	No	45.1%		KsAtAcp/Acp/ErKrAcp	
		FUM1	AAD43562	*G. moniliformis*	No	39.1%		KsAtErKrAcp	
SNOG_14927.2	KsAtDhKrAcp	fg01790		*F. graminearum*	No	44.3%	Reducing PKS clade IV	KsAtErKrAcp	Synthesis of diketide moiety of compactin (Abe et al., 2002a, 2002b).
		PKS15	AAR90269	*C. heterostrophus*	No	43.5%		KsAtAcp/ErKrAcp	
		PKS	BAC20566	*P. citrinum*	No	39.1%		KsAtDhErKrAcp	

(Continued)

Table 7. (continued).

SN15 PKS	Domain/Module Structure	Best Hit	Accession Number	Organism	Reciprocal	Similarity	Inferred Clade	Domain/Module Structure	Inferred Function of Ortholog
SNOG_15829.2	KsAtAcp	fg12040		*F. graminearum*	No	42.2%	Nonreducing PKS proteins basal to clades I and II (uncharacterized)	KsAtAcp	Synthesis of melanin, autofusarin, or similar polyketide (Moriwaki et al., 2004)
		PKS1	BAD22832	*B. oryzae*	No	41.3%		KsAtAcp/Acp/Te	
		PKS14	AAR90250	*B. fuckeliana*	No	67.6%		KsAtAcp/Acp	
SNOG_15965.2	KsAtErKrAcp	fg10548		*F. graminearum*	No	43.9%	Reducing PKS clade IV	KsAtErKrAcp	Synthesis of fumonisin (Proctor et al., 1999)
		PKS10	AAR90246	*B. fuckeliana*	Yes	68.8%		KsAtErKrAcp	
		PKSKA1	AAY32931	*Xylaria* sp BCC 1067	No	54.7%		KsAtErKrAcp	

Domain abbreviations: ACP, acyl-carrier protein domain; At, acetyl transferase; Dh, dehydratase; Er, enoyl reductase; Kr, keto reductase; Ks, keto synthase; Te, thioesterase. Potential orthologs to PKSs *C. heterostrophus* and *F. graminearum* (Kroken et al., 2003; Gaffoor et al., 2005) and by best informative hit to the nonredundant database at NCBI of e-value $<1 \times 10^{-10}$. Domain structure and modular organization for all sequences was determined via the online NRPS-PKS database (Ansari et al., 2004). It should be noted that domain structure predictions may vary slightly from those stated in the original studies. Percentage of similarity of ortholog pairs was determined (Needleman and Wunsch, 1970) via NEEDLE (Rice et al., 2000).

CFEM domain proteins, and *F. graminearum* has eight. Using a combination of domain searches and BLAST searches seeded with the *M. grisea* and *F. graminearum* genes, we identified six related genes (Table 8). Four of these have at least a weak match to the CFEM domain, but only three are predicted to be transmembrane proteins. SNOG_09610.2 appears to be the ortholog of *Pth11* and the *F. graminearum* gene fg05821, suggesting these genes are ancient and conserved. *Pth11* is required for appressorial development and perception of suitable surfaces (DeZwaan et al., 1999). Neither *S. nodorum* nor *F. graminearum* form classical appressoria (Solomon et al., 2006b), suggesting that the genes have different functions in these species. SNOG_05942.2 is also closely related to *Pth11*. Knockout strains for this gene were obtained but had no obvious phenotype (data not shown). SNOG_03589.2 is also related to *Pth11* and appears to be a transmembrane protein but lacks the CFEM domain. Both SNOG_08876.2 and SNOG_15007.2 possess CFEM domains and a signal peptide but do not appear to be integral membrane proteins. Knockout strains for SNOG_08876.2 were obtained but revealed no obvious phenotype (data not shown). SNOG_15007.2 is similar to glycosylphosphatidylinositol-anchored CFEM-containing proteins from three other fungal species. It appears to be heavily expressed both in the oleate and in planta EST libraries with 17 and 27 ESTs, respectively. A clear role for this gene is as yet unknown. CFEM domain proteins are thought to be involved in surface signal perception, and three candidates for this role have been identified. The paucity of such genes in *S. nodorum* and lack of phenotype associated with deletion of two of them suggests that this function may be covered by a different class of transmembrane receptors.

Hydrophobins are small, secreted proteins with eight Cys residues in a conserved pattern that coat the fungal mycelium and spore (Wessels, 1994). Class 2 hydrophobins are restricted to ascomycetes, whereas class 1 hydrophobins are also found in other fungal divisions (Linder et al., 2005). Two class 2 and no class 1 hydrophobin genes were found in the *S. nodorum* genome (see Supplemental Figure 1 online). This is an unusual example of an ascomycete genome with only class 2 hydrophobin genes. It is interesting to note that although a range of *Neurospora* species all contained a gene orthologous to the class 1 hydrophobin *EAS* gene from *N. crassa*, they were expressed significantly only in species that produced aerial conidia (Winefield et al., 2007). *S. nodorum* produces pycnidiospores in a gelatinous cirrhus adapted for rain dispersal (Solomon et al., 2006b). It may be that the absence of aerial mitosporulation negates the need for class 1 hydrophobins.

Two rhodopsin-like genes were found in the genome, one of which is similar to the bacteriorhodopsin-like gene found in *L. maculans* that is a proton pump (Sumii et al., 2005). The likely physiological roles of these genes are currently under investigation.

The interaction between a pathogen and its host is to a large extent orchestrated by the proteins that are secreted or localized to the cell wall or cell membrane. Pathogens such as *M. grisea*, which kill and degrade host tissues, have been shown to secrete large numbers of degradative enzymes (Dean et al., 2005). More surprisingly, the biotrophic pathogen *U. maydis* was also found to secrete many proteins; many of the genes are clustered, coexpressed, and required for normal pathogenesis but are of unknown molecular functions (Kamper et al., 2006). We have therefore analyzed the putative proteome of *S. nodorum* for potentially secreted proteins and compared them with these pathogens and the saprobe *N. crassa.* A total of 1782 proteins was predicted to be extracellular based on predictions using WoLF PSORT; a further 1760 were predicted to be plasma membrane located (see Supplemental Data Set 2 online). GO annotations were assigned via Blast2GO for 551 of the putatively extracellular proteins (Table 9). They are dominated by carbohydrate and protein degradation enzymes as would be expected and which is consistent with the EST analysis (see below). The role of many fungal extracellular proteins is currently unknown. Among the *S. nodorum* predicted extracellular proteins, 1231 had no GO annotation. Of these, only 410 had significant matches to putatively extracellular proteins from *U. maydis*, *M. grisea*, or *N. crassa* (Figure 3). More of these 410 genes had homologs in *M. grisea* (13 + 91 = 104) than in *N. crassa* (3 + 39 = 42). As *M. grisea* and *N. crassa* are phylogenetically equidistant from *S. nodorum*, this suggests both that the pathogens secrete more proteins than *N. crassa* and that the genes are related. Expanding this generalization will require genome analysis of a saprobic dothideomycete. A further 251 genes (89 + 162) had homologs in both *M. grisea* and *N. crassa*. Another 118 *S. nodorum* putatively

Table 8. Potential Orthologs of *S. nodorum* CFEM Proteins Identified by Matches to PFAM Domain PF05730 and Best Hit to *M. grisea* and *F. graminearum* CFEM Proteins

Best Seed[a]	Similarity	SN15 CFEM Protein	PF05730 Match[a]	Location[b]	7tm[c]	Informative Best Hit (Nonredundant)	Reciprocal	Similarity	Inferred Function
fg08554	36.4%	SNOG_02161.2	No	Extracellular	No	None			Unknown
		SNOG_08876.2	Yes	Extracellular	No	None			Unknown
MGG06724.5	21.4%	SNOG_03589.2	No	Plasma membrane	Yes	AAD30437 *M. grisea* PTH11	No	24.8%	GPCR involved in surface perception
MGG05871.5	38.3%	SNOG_05942.2	Yes	Plasma membrane	Yes	AAD30438 *M. grisea* PTH11	No	38.3%	GPCR involved in surface perception
MGG10473.5	57.5%	SNOG_09610.2	Yes	Plasma membrane	Yes	AAD30437 *M. grisea* PTH11	Yes	33.1%	GPCR involved in surface perception
						Fg05821	Yes	59.0%	
MGG05531.5	38.0%	SNOG_15007.2	Yes	Extracellular	No	ABA33784 Pro-rich antigen-like protein (*Paracoccidioides brasiliensis*)	Yes	51.2%	Similar to Pro-rich antigens and Ag2/Pra CRoW domains more commonly identified as immunoreactive antigens of mammalian fungal pathogens (Peng et al., 2002); similar to GPI-anchored CFEM proteins
						XP_750946.1 GPI-anchored CFEM domain protein (*Aspergillus fumigatus* Af293)	Yes	56.6%	
						AAP84613.1|Ag2/Pra CRoW domain (*Coccidioides posadasii*)	Yes	25.7%	

[a] In addition to HMMER matches to PFAM accession PF05730 (using gathering cutoffs), CFEM domain–containing proteins from *M. grisea* (BROAD release 5: MGG_01149.5, MGG_01872.5, MGG_05531.5, MGG_05871.5, MGG_06724.5, MGG_06755.5, MGG_07005.5, MGG_09570.5, and MGG_10473.5) and *Fusarium graminearum* (FGDB: fg00588, fg02077, fg02155, fg02374, fg03897, fg05175, fg05821, and fg08554) were used as seeds to identify putative CFEM proteins by their best BLASTP match with an e-value $<1 \times 10^{-10}$ to *S. nodorum*.

[b,c] Cellular location was determined with WoLF PSORT (Horton et al., 2007), and the presence of transmembrane-spanning regions (7tm) was determined by consensus between TMHMM (Krogh et al., 2001), TMPRED (Hofmann and Stoffel, 1993), and Phobius (Kall et al., 2004). Best hits of *Stagonospora* CFEM proteins to the nonredundant protein database at NCBI was determined by informative BLASTP hits of e-value $<1 \times 10^{-10}$ (hits excluding seed and self hits, which are not hypothetical or unknown and preferentially yield useful functional information).

secreted genes of unknown function had significant similarity to genes encoding the *U. maydis* secretome, including 13 that uniquely hit the biotrophic basidiomycete. We found no obvious patterns of clustering of any of these secreted genes. The most striking finding was that 821 genes had no significant similarity to predicted extracellular proteins of any of these fully sequenced genomes. These large numbers of putatively secreted proteins of mostly unknown functions, whose genes appear to be rapidly evolving, points to a hitherto unsuspected complexity and subtlety in the interaction between the pathogen and its environments.

One newly recognized aspect of *S. nodorum* is the production of secreted proteinaceous toxins. SNOG_16571.2 encodes a host-specific protein toxin called ToxA that determines the interaction with the dominant wheat disease susceptibility gene *Tsn1* (Friesen et al., 2006; Liu et al., 2006). Population genetic evidence suggests that this gene was interspecifically transferred to the wheat tan spot pathogen, *P. tritici-repentis*, prior to 1941, thereby converting a minor into a major pathogen. An RGD motif that is involved in import into susceptible plant cells is required for activity (Meinhardt et al., 2002; Manning et al., 2004; Manning and Ciuffetti, 2005). Biochemical and genetic evidence suggests that *S. nodorum* isolates contain several other proteinaceous toxins (Liu et al., 2004a, 2004b). ToxA is a 13.2-kD protein, and current research indicates that the other toxins are also small proteins. Among the predicted extracellular proteins, 840 are <30 kD and 26 have an RGD motif (see Supplemental Data Set 2 online). Analysis of these toxin candidates is underway.

Gene Expression during Infection

When analyzed by biological process and molecular function, the in planta EST library was dominated by genes involved in the

Table 9. The Most Abundant Gene Ontologies Associated with the Predicted Secretome

GO Identifier	Description	Genes
Biological processes		
GO:0005975	Carbohydrate metabolism	42
GO:0008152	Metabolism	34
GO:0006118	Electron transport	30
GO:0006508	Proteolysis	26
GO:0045493	Xylan catabolism	19
GO:0044237	Cellular metabolism	17
GO:0030245	Cellulose catabolism	12
GO:0000272	Polysaccharide catabolism	10
Molecular functions		
GO:0003674	Molecular function	68
GO:0016491	Oxidoreductase activity	52
GO:0016787	Hydrolase activity	52
GO:0004553	Hydrolase activity, hydrolyzing *O*-glycosyl compounds	48
GO:0005488	Binding	34
GO:0003824	Catalytic activity	33
GO:0046872	Metal ion binding	19
GO:0030248	Cellulose binding	18
GO:0016798	Hydrolase activity, acting on glycosyl bonds	13
GO:0008233	Peptidase activity	12
GO:0008810	Cellulase activity	12
GO:0016740	Transferase activity	12
GO:0016789	Carboxylic ester hydrolase activity	10

biosynthesis of proteins and in the degradation and use of extracellular proteins and carbohydrates, specifically cellulose and arabino-xylan (Table 3; see Supplemental Table 5 online). The secretion of proteases (Bindschedler et al., 2003), xylanases, and cellulases suggests that the fungus catabolizes plant proteins and carbohydrates for energy in planta at this late stage of infection. Studies in a range of pathogens suggest that early infection is characterized by the catabolism of internal lipid stores and that mid stages are characterized by the use of external sugars and amino acids (Solomon et al., 2003a). These studies of late infection suggest that polymerized substrates are used after the more easily available substrates are exhausted and provide a novel perspective to in planta nutrition. The abundance of transcripts linked to protein synthesis indicates that this process is very active during this stage of infection and may be related to the massive secretion of degradative enzymes and to the synthesis of proteins needed for sporulation. A similar picture was observed when the genes were analyzed by cellular component (Table 3) with gene products destined for the ribosome dominating. Genes whose products are targeted to the nucleolus and cytoskeleton illustrate the importance of nuclear division and cytokinesis at this stage of infection as many new cell types are elaborated in the developing pycnidium. Thirty-six gene products targeted to the extracellular region include a protease, an α-glucuronidase, and various glucanases consistent with the picture of polymerized substrate breakdown.

Late-stage infection has rarely been examined in plant–fungal interactions, and it is therefore noteworthy that upregulation of transcripts involved in protein synthesis was also observed in late-stage infections of wheat by *M. graminicola* (Keon et al., 2005). The high rate of protein synthesis may be required for pycnidial biogenesis and may also represent an attractive target for fungicidal intervention.

The oleate EST library was dominated by genes involved in lipid and malate metabolism, the trichloroacetic acid cycle, and electron transport. To determine whether the ratio of ESTs in the in planta and oleate libraries was indicative of early- and late-stage infection, transcript levels from eight candidate genes were determined by a quantitative PCR. The genes were chosen from those found exclusively in the in planta library. Transcripts were quantified in cDNA pools made from RNA isolated from in vitro and in planta growth of SN15 at early and late infection time points. The in vitro cultures sampled at 4 DAI did not contain any pycnidia, whereas cultures sampled at 18 DAI were heavily sporulating with many pycnidia. To isolate RNA from both early and late infections of wheat, an SN15 infection latent period assay was used to provide in planta infection transcripts (Solomon et al., 2003b). Lesions were excised after 8 and 12 DAI, representing early and late infection stages. The three genes most upregulated during growth in planta (SNOG_00557.2, SNOG_03877.2, and SNOG_16499.2) all had >20 in planta ESTs, indicating that the EST frequencies were useful predictors of gene expression at these levels. SNOG_00557.2 was the gene with the highest peak expression level in planta and the largest difference between in vitro and in planta conditions. As the gene is predicted to be an arabinofuranosidase, this reaffirmed the important role of carbohydrolases during colonization of the host tissue. Overall, the four tissue types represented clear states of nonsporulation and sporulation during both in vitro and in planta growth. All the genes were found to be expressed in the 12-DAI infected sample, and this was the highest level observed in all but

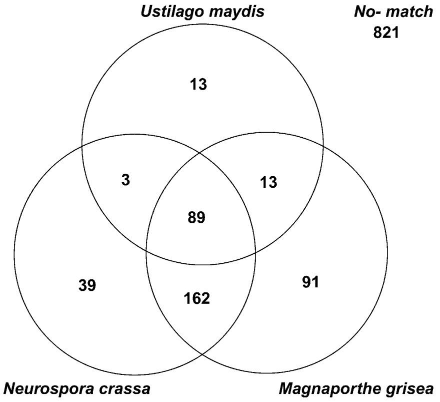


Figure 3. Homology Relationships of the *S. nodorum* Secretome.

The 1231 predicted extracellular proteins without GO annotation were compared with the latest releases of the *U. maydis*, *N. crassa*, and *M. grisea* genomes. Counts are best hits of e-value <1e-10 if predicted as extracellular by WoLF PSORT.

one case (see Supplemental Figure 2 online). On the other hand, moderate to high levels of gene expression were found in the in vitro samples from six of the eight genes. Attempts to use axenic samples to mimic infection has a long history (Coleman et al., 1997; Talbot et al., 1997; Solomon et al., 2003a; Trail et al., 2003; Thomma et al., 2006). Our data indicate that oleate feeding is not a significantly closer model for infection than starvation has proved to be.

Genome Architecture

Colinearity of gene order is a notable feature of closely related plant and animal genomes but is represented in fungi by a complex pattern of dispersed colinearity, such as observed between *Aspergillus* spp (Galagan et al., 2005). This low degree of organizational conservation can be attributed to the 200 million year separation of these species and the rarity of meiotic events that would tend to maintain chromosomal integrity. *S. nodorum* is the first dothideomycete to have its sequence publicly released; thus, the closest species for which whole-genome comparisons could be made are the Aspergilli and *N. crassa*, which are separated from *S. nodorum* by ~400 million years of evolution. Thus far, we have been unable to detect significant regions of colinearity with other sequenced genomes, but it is possible that more sensitive methods and the release of more closely related genome sequences will succeed in revealing patterns of chromosomal level evolution.

At a smaller scale, the mating type loci of *L. maculans*, *C. heterostrophus*, *A. alternata*, and *M. graminicola* have been previously analyzed (Cozijnsen and Howlett, 2003). Apart from *Mat1-1* and *orf1*, the only shared genes were a DNA ligase gene found in *L. maculans* and *M. graminicola* and *Gap1* found in *L. maculans* and *C. heterostrophus*. The colinearity between *L. maculans* and *S. nodorum* is more complete. Figure 4A shows the relationship of gene order and orientation between these sibling species. It is clear that gene order and orientation are conserved, though the intergenic regions have no discernable relationships. This confirms the close phylogenetic relationship between these species, with *S. nodorum* closest to *L. maculans* and more distantly related to *C. heterostrophus*.

There is tantalizing evidence of residual colinearity of the genes present in a contiguous 38-kb region of *L. maculans* DNA (Idnurm et al., 2003). Six of the nine *L. maculans* genes have orthologs within a 200-kb region of *S. nodorum* DNA. Gene orientation is retained, but they are interspersed with ~60 other predicted genes (Figure 4B).

A different pattern of colinearity is apparent in the quinate gene cluster. The quinate genes regulate the catabolism of quinate and have a wide distribution in filamentous fungi (Giles et al., 1991). The cluster comprises seven coregulated genes. In some cases, portions of the cluster are repeated at other genomic locations. In *S. nodorum*, the main cluster is located on scaffold 19 with two genes on scaffold 12. Figure 5 show the relationship of gene content and order in six fungal species. As the species for comparison span 400 million years of evolution, it is apparent that clustering of these genes per se is highly conserved. However, the gene order and orientation shows no conservation apart from the constant juxtaposition of homologs of Qa-1S and Qa1-F and of Qa-X and Qa-2. Even in these cases, though, some are 5′ to 5′ and others are 3′ to 3′. It appears, therefore, that there must be selection for retention of clustering of these seven genes, even if the exact orientation may not be so important. It may be that such proximity clustering is important in that it enables gene regulation by chromatin remodeling mediated by a

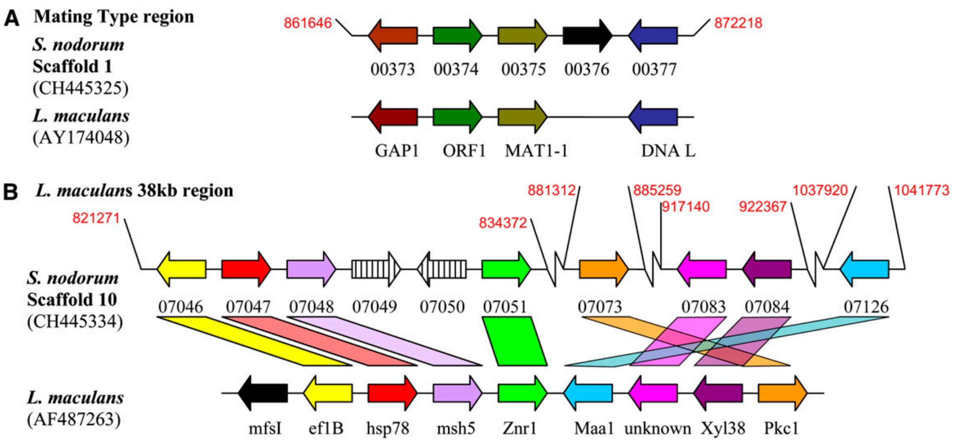


Figure 4. Colinearity between *S. nodorum* SN15 and *L. maculans* Identified by tBLASTX.

(A) Mating-type locus.
(B) A 38-kb region of *L. maculans*.
Colors indicate an orthologous relationship between genes, whereas black indicates no relationship. Numbers in red give coordinates on the *S. nodorum* scaffolds.

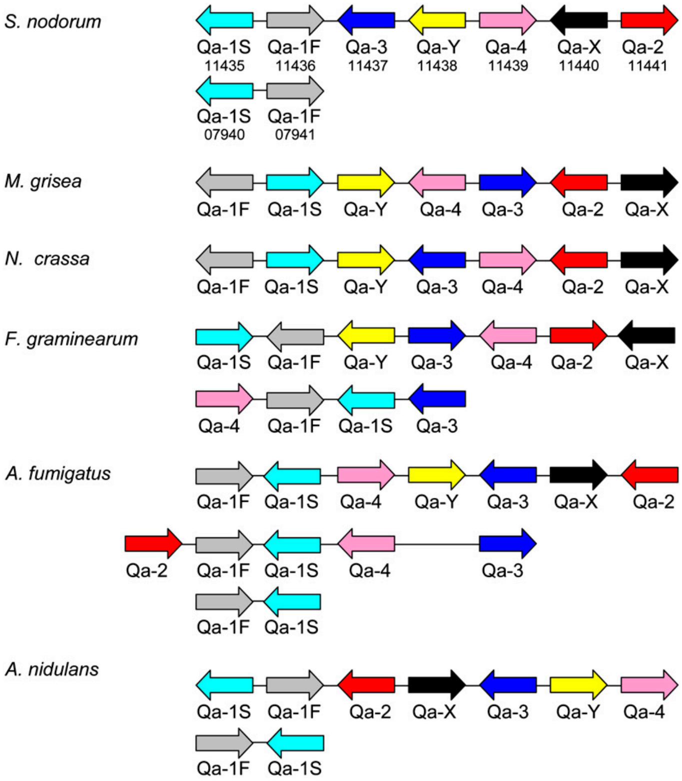

S. nodorum
Qa-1S 11435
Qa-1F 11436
Qa-3 11437
Qa-Y 11438
Qa-4 11439
Qa-X 11440
Qa-2 11441
Qa-1S 07940
Qa-1F 07941
M. grisea
Qa-1F Qa-1S Qa-Y Qa-4 Qa-3 Qa-2 Qa-X
N. crassa
Qa-1F Qa-1S Qa-Y Qa-3 Qa-4 Qa-2 Qa-X
F. graminearum
Qa-1S Qa-1F Qa-Y Qa-3 Qa-4 Qa-2 Qa-X
Qa-4 Qa-1F Qa-1S Qa-3
A. fumigatus
Qa-1F Qa-1S Qa-4 Qa-Y Qa-3 Qa-X Qa-2
Qa-2 Qa-1F Qa-1S Qa-4 Qa-3
Qa-1F Qa-1S
A. nidulans
Qa-1S Qa-1F Qa-2 Qa-X Qa-3 Qa-Y Qa-4
Qa-1F Qa-1S

Figure 5. Organization of the Quinate Cluster from Several Sequenced Fungal Genomes.

Numbers under the *S. nodorum* genes refer to the SNOG identifiers. The *S. nodorum* genes are located on scaffolds 19 and 12, respectively.

gene such as *LaeA* (Bok and Keller, 2004). SNOG_11365.2 is an apparent ortholog of *LaeA*.

Fungi have a truly ancient history, many times longer than that of animals and flowering plants (Padovan et al., 2005). The extent of genetic diversity is correspondingly large. Phylogenetic analysis of fungi has been hampered by a paucity of reliable morphological indicators. As a consequence, phylogenetic reconstruction of the fungi has been particularly unstable until the widespread introduction of multigene-based DNA sequence comparisons. This study confirms the overall monophyletic characters of the Dothideomycetes and the Pleosporales taxa. These groups contain many thousands of species and are notable for their content of major plant pathogens infecting many important plant families. The origins of this class, likely >400 million years ago, is considerably older than their plant hosts. It is particularly interesting to note that all the fungal plant pathogens that are well established as host-specific toxin producers are in this class (*Stagonospora*, *Cochliobolus*, *Pyrenophora*, *M. zeae-maydis*, and *Alternaria*). Taken as a whole, the pathogens in this class are mainly described as necrotrophic, while a few are debatably described as biotrophic (*Venturia* and *C. fulvum*) or hemibiotrophic (*L. maculans*, *M. graminicola*, and *M. fijiensis*). None of the species in this class are obligate pathogens, and none possess classical haustoria (Oliver and Ipcho, 2004); furthermore, many other species are nonpathogenic. The genome sequence of *S. nodorum* represents an important point of comparison from which to derive hypotheses about the genetic basis of pathogenicity in these organisms. Genome sequences of several of these species are in progress (Goodwin, 2004), including *A. brassicicola*, *M. fijiensis*, *M. graminicola*, *L. maculans*, and *P. tritici-repentis*, or have been completed but not released (*C. heterostrophus*; Catlett et al., 2003). Three of these species, *S. nodorum*, *M. graminicola*, and *P. tritici-repentis*, are wheat pathogens. The possibility of multiple pairwise comparisons of gene content between phylogenetically and ecologically close species promises to be a powerful method to derive workably small lists of candidate effector genes controlling pathogenicity, host specificity, and life cycle characters and thus provide ideas for the generation of novel crop protection strategies. The genome sequence provides the tools for global transcriptome analysis, thereby identifying the genes expressed during different phases of infection. This study has also highlighted secreted proteins, which appear to be markedly more numerous than in nonpathogens but which have predominately mysterious roles. Finally, the genome sequence is an essential prerequisite for the critical analysis of hypotheses of interspecific gene transfer. This has already been identified in pathogens in general (Richards et al., 2006) and *S. nodorum* in particular (Friesen et al., 2006; Stukenbrock and McDonald, 2007) and may emerge as a major feature in the evolution of these organisms.

METHODS

Fungal Strains

Stagonospora nodorum strain SN15 (Solomon et al., 2003b) was used for both genome and EST libraries and has been deposited at the Fungal Genetics Stock Center. The genome sequence was obtained as described (http://www.broad.mit.edu/annotation/genome/stagonospora_nodorum/Assembly.html). The sequences are available for download from GenBank under accession number AAGI00000000.

Phylogenetic Analysis

A combined matrix of 41 taxa was generated from DNA sequences obtained from two ribosomal (nuclear large and small subunit [nuc-LSU and nuc-SSU]) and three protein genes (elongation factor 1 α [EF-1α] and the largest and second largest subunits of RNA polymerase II [*RPB1* and *RPB2*]). Data were obtained from the Assembling the Fungal Tree of Life (AFTOL; www.aftol.org) and GenBank sequence databases, which aligned in ClustalX (Thompson et al., 1997) and manually improved where necessary. After introns and ambiguously aligned characters were excluded, 6694 bp were used in the final analyses. In some cases, genes were missing (see Supplemental Table 3 and Supplemental Data Set 3 online). The resulting data were combined and delimited into 11 partitions, including nuc-SSU, nuc-LSU, and the first, second, and third codon positions of EF-1α, RPB1, and RPB2, with unique models applied to each partition. Metropolis coupled Markov chain Monte Carlo analyses were conducted using MrBayes 3.1.2 (Huelsenbeck and Ronquist, 2001) with a six-parameter model of evolution (generalized time reversible) (Rodriguez et al., 1990) and gamma distribution approximated with four categories and a proportion of invariable sites. Trees were sampled every 100th generation for 5,000,000 generations. Three runs were completed to ensure that stationarity was reached, and 5000 trees were discarded as "burn in" for each. Posterior probabilities were determined by calculating a 50% majority-rule consensus tree of 45,000 trees from a single run. Maximum likelihood bootstrap proportions were calculated by doing 1000 replicates in RAxML-VI-HPC (Stamatakis, 2006) with the GTRCAT model approximation and 25 rate categories with the same data partitions as for the Bayesian runs.

Repetitive Elements

Repetitive elements were identified de novo using RepeatScout v1.0.0 (Price et al., 2005) with a minimum threshold of 10 matches and a minimum repeat length of 200 bp. Newly generated repeats were aligned to the genome assembly via BLASTN v2.0 (Altschul et al., 1990). Hits were discarded if sequence identity fell below 35% or alignment length was <200 bp. For each repeat, the number of filtered hits was counted, and repeats were discarded if the number of hits to the assembly was <10. Prototype repeat regions were defined by identification of sequence similarities among themselves via BLASTN and deletion of redundant repeat sequence. Regions of the genome assembly that matched to a repeat were aligned using MUSCLE v3.6 (Edgar, 2004). To identify repeat type and function, repeats were compared with the nonredundant sequence database hosted by NCBI via BLASTN/BLASTX, and subrepeat regions within repeats were also analyzed. Tandem repeats were identified using Tandem Repeat Finder (Benson, 1999), direct repeats were identified via MegaBLAST (Zhang et al., 2000), and inverted repeats were identified using both MegaBLAST and eINVERTED (Rice et al., 2000).

Putative telomeric regions were predicted by considering scaffold ends containing successive repetitive elements without interspersed predicted protein coding genes. The occurrence of repeat classes within these regions was counted and compared with occurrences throughout the genome. Where >85% of the repeats were found to be at nongenic scaffold end regions, these classes were classified as telomeric repeats. The scaffold ends were predicted to be physical telomeres if they contained three or more telomeric repeats.

To analyze and compare the prevalence of RIP mutation, we developed a program to compare RIP mutations between multiple sequences (J.K. Hane and R.P. Oliver, unpublished data). RIPCAL compares the aligned sequences against a designated model sequence (in this case the trimmed de novo repeats) and was configured to calculate RIP (Dean et al., 2005).

Gene Content

An automated genome annotation was initially created using the Calhoun annotation system. A combination of gene prediction programs, FGENESH, FGENESH+, and GENEID, and 317 manually curated transcripts were used. GENEWISE was used with non-species specific parameters to predict genes from proteins identified by BLASTX. To refine the initial genome annotation, a further 10,752 EST reads were obtained from an EST library from oleate-grown mycelium and 10,751 from *S. nodorum*–infected wheat (see below for details). These EST sequences were screened against a library of wheat ESTs, trimmed for vector sequence manually, for poly(A) tail sequence using TrimEST (Rice et al., 2000), screened for unusable sequences of poor quality, filtered for remaining sequence of ≥50 bp, and aligned to the genome assembly using Sim4 (Florea et al., 1998). ESTs with multiple genomic locations were assigned their optimum location based on percent identity, total alignment length, and best location of the EST mated pair. Gene models were manually annotated according to optimum EST alignments using Apollo (Lewis et al., 2002).

Genes that were fully supported by EST data were used to train the gene prediction program UNVEIL (Majoros et al., 2003). Second-round gene annotations were created by combining gene predictions with EST data, with EST data replacing predicted gene models, and UNVEIL predictions preferred over first-round predictions. Genes with coding regions <50 amino acids were discarded. The numerical identifiers assigned during the first-round predictions have been retained. New gene models were assigned loci from 20,000 onwards. Updated loci (UNVEIL and EST supported) have been given the numerical suffix 2 (e.g., SNOG_16571.2). The 5354 unsupported genes have retained the numerical identifiers and suffix 1. Where genic regions lack EST support, only the coding sequence is reported.

ESTs not aligned to the assembly by Sim4 were compared with the unassembled reads via BLASTN. EST with matches to unassembled reads with an e-value of <1E-10 were clustered into contigs using cap3 (Huang and Madan, 1999). The assembled ESTs were tested for single open reading frames using getORF (Rice et al., 2000), and possible genes were determined by BLASTP (Altschul et al., 1990) to protein databases at NCBI. One new gene (STAG_20208.1) was identified by this method. We have chosen not to alter genes where gene models conflicted with homologs except where colinearity evidence confirms orthology (Figure 4). Updated files of gene and protein sequences are available from R. Oliver (roliver@murdoch.edu.au).

Proteins were assigned putative functional classes by searching for relevant PFAM (Bateman et al., 1999) domains (Bateman et al., 2004) with HMMER v2.0 (Eddy, 1998) (see Supplemental Data Set 2 and Supplemental Table 2 online). Putative gene ontology was assigned via BLASTP with BLAST2GO (Conesa et al., 2005). Due to the relatively poor identification of certain PKS and NRPS domains using resources like PFAM (Bateman et al., 2004) and CD-SEARCH (Marchler-Bauer and Bryant, 2004; Marchler-Bauer et al., 2005), PKS and NRPS domains and their modular organization were further elucidated with online tools available at NRPS-PKS (Yadav et al., 2003; Ansari et al., 2004). Subcellular localization and secretion was predicted via WoLF PSORT (Horton et al., 2007).

EST Library Construction

For the construction of the in planta cDNA library, wheat (*Triticum aestivum*) cv Amery was infected with *S. nodorum* SN15 as a whole plant spray and processed as a latent period assay (Solomon et al., 2006b). Necrotic tissue was excised from the leaves at 10, 11, and 12 DAI for RNA isolation. RNA was isolated by the Trizol method (Sigma-Aldrich).

Material for the oleate library was generated as follows: minimal medium (100 mL) + sucrose (0.5% [w/v]) was inoculated with *S. nodorum* SN15 pycnidiospores (2.75 × 10^7) and incubated at 22°C with shaking (130 rpm) for 4 d. Mycelia was harvested, washed, and added to 100 mL of minimal media, including 0.05% (v/v) Tween 80 and 0.2% (w/v) oleate as the carbon source. The culture was incubated as before for 30 h before being harvested, washed, snap frozen in liquid nitrogen, and freeze-dried in a Maxi Dry Lyo (Heto Holten). mRNA was extracted using the Messagemaker mRNA purification-cloning kit (Gibco/Invitrogen) according to manufacturer's instructions. A total of 2.3 μg of mRNA was used as template for reverse transcription to cDNA. The in planta cDNA library was constructed using the SMART cDNA library construction kit (Clontech), and the oleate library was constructed in pSPORT1 (Gibco/ Invitrogen). All manipulations were performed according to the manufacturer's instructions. Phage DNA was packaged using the GigapackIII Gold phage packaging system (Stratagene) according to the manufacturer's instructions. Phages were grown, amplified, and mass-excised to bacterial clones according to the provided protocols. The final libraries were both estimated to contain ~500,000 clones.

Quantitative PCR

Total RNA (1 μg) was reverse transcribed to cDNA using iScript reverse transcriptase premix (Bio-Rad) according to the manufacturer's instructions. RNA from three biological replicates was pooled for a single cDNA synthesis. cDNA reactions were used as PCR template at 1:50 dilution for in vitro–grown SN15 samples and at 1:5 dilution for in planta–grown SN15 samples. Quantitative PCR reactions consisted of 10 μL of iQ SYBR green supermix (Bio-Rad), forward and reverse primers (each at 1.2 μM), and 5 μL of template DNA in a 20 μL reaction. Reactions were incubated in a Rotor-Gene 3000 thermocycler (Corbett Research). Cycling conditions were 3 min/95°C and then 35 cycles of 10 s/95°C, 20 s/57°C, and 72°C/20 s. Amplicon fluorescence from template of unknown concentration was compared with that from genomic DNA standards of 25, 2.5, 0.25, and 0.025 ng/reaction. All reactions were performed with two technical replicates. Data were analyzed using the Rotorgene software version 6.0 (Corbett Research). Primer sequences are listed in Supplemental Table 4 online.

Accession Number

Sequence data from this article can be found in the GenBank/EMBL data libraries under accession number AAGI00000000.

Supplemental Data

The following materials are available in the online version of this article.

Supplemental Figure 1. Aligned Kyte and Doolittle Hydrophobicity Plots.

Supplemental Figure 2. Quantitative PCR Analysis of Gene Expression from in Planta–Associated Genes.

Supplemental Table 1. Mitochondrial Genes.

Supplemental Table 2. Summary of PFAM Domains Used for Identification.

Supplemental Table 3. List of Species Used in This Study.

Supplemental Table 4. PCR Primers for Abundant in Planta–Associated Genes.

Supplemental Table 5. EST Data for Abundant in Planta–Associated Genes.

Supplemental Data Set 1. Repetitive Elements: Telomeres, Subrepeats, and Similarity Scores.

Supplemental Data Set 2. Gene Summary.

Supplemental Data Set 3. Alignments for Figure 1 as a Noninterleaved Nexus Formatted Text File.

ACKNOWLEDGMENTS

Funding was provided by the Australian Grains Research and Development Corporation (UMU14; we particularly thank James Fortune, John Sandow, and John Lovett for their support of this project), by the Swiss Federal Institute of Technology Zurich, by the Swiss National Science Foundation (Grant 3100A0-104145), by MIT, and by the National Science Foundation (DEB-0228725; Assembling the Fungal Tree of Life). We thank Soeren Rasmussen for help with bioinformatics, Robert Lee for the oleate EST library, and Tim Friesen (USDA) for comments on the manuscript. We dedicate this article to the recently retired Chris Caten who has inspired our work on *S. nodorum*.

Received May 10, 2007; revised September 11, 2007; accepted October 15, 2007; published November 16, 2007.

REFERENCES

Abe, Y., Suzuki, T., Mizuno, T., Ono, C., Iwamoto, K., Hosobuchi, M., and Yoshikawa, H. (2002b). Effect of increased dosage of the ML-236B (compactin) biosynthetic gene cluster on ML-236B production in *Penicillium citrinum.* Mol. Genet. Genomics **268:** 130–137.

Abe, Y., Suzuki, T., Ono, C., Iwamoto, K., Hosobuchi, M., and Yoshikawa, H. (2002a). Molecular cloning and characterization of an ML-236B (compactin) biosynthetic gene cluster in *Penicillium citrinum.* Mol. Genet. Genomics **267:** 636–646.

Agrios, G. (2005). Plant Pathology. (Burlington, MA: Elsevier Academic Press).

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. **215:** 403–410.

Amnuaykanjanasin, A., Punya, J., Paungmoung, P., Rungrod, A., Tachaleat, A., Pongpattanakitshote, S., Cheevadhanarak, S., and Tanticharoen, M. (2005). Diversity of type I polyketide synthase genes in the wood-decay fungus *Xylaria* sp. BCC 1067. FEMS Microbiol. Lett. **251:** 125–136.

Ansari, M.Z., Yadav, G., Gokhale, R.S., and Mohanty, D. (2004). NRPS-PKS: A knowledge-based resource for analysis of NRPS/PKS megasynthases. Nucleic Acids Res. **32:** W405–413.

Audic, S., and Claverie, J.M. (1997). The significance of digital gene expression profiles. Genome Res. **7:** 986–995.

Bailey, A.M., Kershaw, M.J., Hunt, B.A., Paterson, I.C., Charnley, A.K., Reynolds, S.E., and Clarkson, J.M. (1996). Cloning and sequence analysis of an intron-containing domain from a peptide synthetase-encoding gene of the entomopathogenic fungus *Metarhizium anisopliae.* Gene **173:** 195–197.

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

Baldwin, T.K., Winnenburg, R., Urban, M., Rawlings, C., Koehler, J., and Hammond-Kosack, K.E. (2006). The pathogen-host interactions database (PHI-base) provides insights into generic and novel themes of pathogenicity. Mol. Plant Microbe Interact. **19:** 1451–1462.

Bateman, A., Birney, E., Durbin, R., Eddy, S.R., Finn, R.D., and Sonnhammer, E.L. (1999). Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. Nucleic Acids Res. **27:** 260–262.

Bateman, A., et al. (2004). The Pfam protein families database. Nucleic Acids Res. **32:** D138–D141.

Bathgate, J.A., and Loughman, R. (2001). Ascospores are a source of inoculum of *Phaeosphaeria nodorum*, *P. avenaria f. sp. avenaria* and *Mycosphaerella graminicola* in Western Australia. Australas. Plant Pathol. **30:** 317–322.

Bearchell, S.J., Fraaije, B.A., Shaw, M.W., and Fitt, B.D.L. (2005). Wheat archive links long-term fungal pathogen population dynamics to air pollution. Proc. Natl. Acad. Sci. USA **102:** 5438–5442.

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

Benson, G. (1999). Tandem repeats finder: A program to analyze DNA sequences. Nucleic Acids Res. **27:** 573–580.

Bindschedler, L.V., Sanchez, P., Dunn, S., Mikan, J., Thangavelu, M., Clarkson, J.M., and Cooper, R.M. (2003). Deletion of the SNP1 trypsin protease from *Stagonospora nodorum* reveals another major protease expressed during infection. Fungal Genet. Biol. **38:** 43–53.

Bohnert, H.U., Fudal, I., Dioh, W., Tharreau, D., Notteghem, J.L., and Lebrun, M.H. (2004). A putative polyketide synthase/peptide synthetase from *Magnaporthe grisea* signals pathogen attack to resistant rice. Plant Cell **16:** 2499–2513.

Bok, J.W., and Keller, N.P. (2004). LaeA, a regulator of secondary metabolism in Aspergillus spp. Eukaryot. Cell **3:** 527–535.

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

Catlett, N.L., Yoder, O.C., and Turgeon, B.G. (2003). Whole-genome analysis of two-component signal transduction genes in fungal pathogens. Eukaryot. Cell **2:** 1151–1161.

Chung, K.R., Ehrenshaft, M., Wetzel, D.K., and Daub, M.E. (2003). Cercosporin-deficient mutants by plasmid tagging in the asexual fungus *Cercospora nicotianae.* Mol. Genet. Genomics **270:** 103–113.

Coleman, M., Henricot, B., Arnau, J., and Oliver, R.P. (1997). Starvation-induced genes of the tomato pathogen *Cladosporium fulvum* are also induced during growth in planta. Mol. Plant Microbe Interact. **10:** 1106–1109.

Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., Talon, M., and Robles, M. (2005). Blast2GO: A universal tool for annotation, visualization and analysis in functional genomics research. Bioinformatics **21:** 3674–3676.

Cooley, R., Shaw, R., Franklin, F., and Caten, C. (1988). Transformation of the phytopathogenic fungus *Septoria nodorum* to hygromycin B resistance. Curr. Genet. **13:** 383–386.

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

Cozijnsen, A.J., and Howlett, B.J. (2003). Characterisation of the mating-type locus of the plant pathogenic ascomycete *Leptosphaeria maculans.* Curr. Genet. **43:** 351–357.

Cummings, D., McNally, K., Domenico, J., and Matsuura, E. (1990). The complete DNA sequence of the mitochondrial genome of *Podospora anserina.* Curr. Genet. **17:** 375–402.

Dancer, J., Daniels, A., Cooley, N., and Foster, S. (1999). *Septoria tritici* and *Stagonospora nodorum* as model pathogens for fungicide discovery. In Septoria on Cereals: A Study of Pathosystems, J.A. Lucas, P. Bowyer, and H.M. Anderson, eds (New York: ABI Publishing), pp. 316–331.

Dean, R.A., et al. (2005). The genome sequence of the rice blast fungus *Magnaporthe grisea.* Nature **434:** 980–986.

Del Prado, R., Schmitt, I., Kautz, S., Palice, Z., Luecking, R., and Lumbsch, H.T. (2006). Molecular data place Trypetheliaceae in Dothideomycetes. Mycol. Res. **110:** 511–520.

DeZwaan, T.M., Carroll, A.M., Valent, B., and Sweigard, J.A. (1999). *Magnaporthe grisea* pth11p is a novel plasma membrane protein that mediates appressorium differentiation in response to inductive substrate cues. Plant Cell **11:** 2013–2030.

Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics **14:** 755–763.

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

Eisendle, M., Oberegger, H., Zadra, I., and Haas, H. (2003). The siderophore system is essential for viability of *Aspergillus nidulans*: Functional analysis of two genes encoding l-ornithine N 5-monooxygenase (sidA) and a non-ribosomal peptide synthetase (sidC). Mol. Microbiol. **49:** 359–375.

Eisendle, M., Schrettl, M., Kragl, C., Müller, D., Illmer, P., and Haas, H. (2006). The intracellular siderophore ferricrocin is involved in iron storage, oxidative-stress resistance, germination, and sexual development in *Aspergillus nidulans.* Eukaryot. Cell **5:** 1596–1603.

Elliott, C.E., and Howlett, B.J. (2006). Overexpression of a 3-ketoacyl-CoA thiolase in *Leptosphaeria maculans* causes reduced pathogenicity on *Brassica napus.* Mol. Plant Microbe Interact. **19:** 588–596.

Eriksson, O.E. (2006). Outline of Ascomycota. Myconet **12:** 1–82.

Florea, L., Hartzell, G., Zhang, Z., Rubin, G.M., and Miller, W. (1998). A computer program for aligning a cDNA sequence with a genomic DNA sequence. Genome Res. **8:** 967–974.

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

Fujii, I., Ono, Y., Tada, H., Gomi, K., Ebizuka, Y., and Sankawa, U. (1996). Cloning of the polyketide synthase gene atX from *Aspergillus terreus* and its identification as the 6-methylsalicylic acid synthase gene by heterologous expression. Mol. Gen. Genet. **253:** 1–10.

Fujii, I., Yoshida, N., Shimomaki, S., Oikawa, H., and Ebizuka, Y. (2005). An iterative type I polyketide synthase PKSN catalyzes synthesis of the decaketide alternapyrone with regio-specific octa-methylation. Chem. Biol. **12:** 1301–1309.

Gaffoor, I., Brown, D., Plattner, R., Proctor, R., Qi, W., and Trail, F. (2005). Functional analysis of the polyketide synthase genes in the filamentous fungus *Gibberella zeae* (anamorph *Fusarium graminearum).* Eukaryot. Cell **4:** 1926–1933.

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

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

Ghikas, D.V., Kouvelis, V.N., and Typas, M.A. (2006). The complete mitochondrial genome of the entomopathogenic fungus *Metarhizium* *anisopliae* var. *anisopliae*: Gene order and trn gene clusters reveal a common evolutionary course for all Sordariomycetes, while intergenic regions show variation. Arch. Microbiol. **185:** 393–401.

Giles, N.H., Geever, R.F., Asch, D.K., Avalos, J., and Case, M.E. (1991). Organization and regulation of the qa (quinic acid) genes in *Neurospora crassa* and other fungi. J. Hered. **82:** 1–7.

Goodwin, S.B. (2004). Minimum phylogenetic coverage: An additional criterion to guide the selection of microbial pathogens for initial genomic sequencing efforts. Phytopathology **94:** 800–804.

Graziani, S., Vasnier, C., and Daboussi, M.J. (2004). Novel polyketide synthase from *Nectria haematococca.* Appl. Environ. Microbiol. **70:** 2984–2988.

Guillemette, T., Sellam, A., and Simoneau, P. (2004). Analysis of a nonribosomal peptide synthetase gene from *Alternaria brassicae* and flanking genomic sequences. Curr. Genet. **45:** 214–224.

Hofmann, K., and Stoffel, W. (1993). TMbase - A database of membrane spanning proteins segments. Biol. Chem. Hoppe Seyler **374:** 166.

Horton, P., Park, K.J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C.J., and Nakai, K. (2007). WoLF PSORT: Protein localization predictor. Nucleic Acids Res. **35:** W585–587.

Huang, X., and Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res. **9:** 868–877.

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

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

Idnurm, A., Taylor, J.L., Pedras, M.S.C., and Howlett, B.J. (2003). Small scale functional genomics of the blackleg fungus, *Leptosphaeria maculans*: Analysis of a 38 kb region. Australas. Plant Pathol. **32:** 511–519.

Jaffe, D.B., Butler, J., Gnerre, S., Maucelli, E., Lindblad-Toh, K., Mesirov, J.P., Zody, M.C., and Lander, E.S. (2003). Whole genome sequence assembly for mammalian genomes: Arachne 2. Genome Res. **13:** 91–96.

James, T., et al. (2006). Reconstructing the early evolution of fungi using a six-gene phylogeny. Nature **443:** 818–822.

Jones, T., Federspiel, N.A., Chibana, H., Dungan, J., Kalman, S., Magee, B.B., Newport, G., Thorstenson, Y.R., Agabian, N., Magee, P.T., Davis, R.W., and Scherer, S. (2004). The diploid genome sequence of *Candida albicans.* Proc. Natl. Acad. Sci. USA **101:** 7329–7334.

Kall, L., Krogh, A., and Sonnhammer, E.L. (2004). A combined transmembrane topology and signal peptide prediction method. J. Mol. Biol. **338:** 1027–1036.

Kamper, J., et al. (2006). Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis.* Nature **444:** 97–101.

Keon, J., Antoniw, J., Rudd, J., Skinner, W., Hargreaves, J., and Hammond-Kosack, K. (2005). Analysis of expressed sequence tags from the wheat leaf blotch pathogen *Mycosphaerella graminicola* (anamorph *Septoria tritici*). Fungal Genet. Biol. **42:** 376–389.

Kim, K.-H., Cho, Y., La Rota, M., Cramer, R.A., Jr., and Lawrence, C.B. (2007). Functional analysis of the *Alternaria brassicicola* nonribosomal peptide synthetase gene AbNPS2 reveals a role in conidial cell wall construction. Mol. Plant Pathol. **8:** 23–29.

Kodsueb, R., Dhanasekaran, V., Aptroot, A., Lumyong, P., McKenzie, E.H.C., Hyde, K.D., and Jeewon, R. (2006). The family Pleosporaceae: Intergeneric relationships and phylogenetic perspectives based on sequence analyses of partial 28S rDNA. Mycologia **98:** 571–583.

Kouvelis, V., Ghikas, D., and Typas, M. (2004). The analysis of the complete mitochondrial genome of *Lecanicillium muscarium* (synonym *Verticillium lecanii*) suggests a minimum common gene organization in mtDNAs of Sordariomycetes: Phylogenetic implications. Fungal Genet. Biol. **41:** 930–940.

Krogh, A., Larsson, B., von Heijne, G., and Sonnhammer, E.L. (2001). Predicting transmembrane protein topology with a hidden Markov model: Application to complete genomes. J. Mol. Biol. **305:** 567–580.

Kroken, S., Glass, N.L., Taylor, J.W., Yoder, O.C., and Turgeon, B.G. (2003). Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. Proc. Natl. Acad. Sci. USA **100:** 15670–15675.

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

Lafon, A., Han, K.H., Seo, J.A., Yu, J.H., and d'Enfert, C. (2006). G-protein and cAMP-mediated signaling in aspergilli: A genomic perspective. Fungal Genet. Biol. **43:** 490–502.

Lee, B.N., Kroken, S., Chou, D.Y.T., Robbertse, B., Yoder, O.C., and Turgeon, B.G. (2005). Functional analysis of all nonribosomal peptide synthetases in *Cochliobolus heterostrophus* reveals a factor, NPS6, involved in virulence and resistance to oxidative stress. Eukaryot. Cell **4:** 545–555.

Lewis, S.E., et al. (2002). Apollo: A sequence annotation editor. Genome Biol. **3:** 82.

Linder, M.B., Szilvay, G.R., Nakari-Setala, T., and Penttila, M.E. (2005). Hydrophobins: The protein-amphiphiles of filamentous fungi. FEMS Microbiol. Rev. **29:** 877–896.

Liu, Z., Friesen, T., Ling, H., Meinhardt, S., Oliver, R., Rasmussen, J., and Faris, J. (2006). The Tsn1-ToxA interaction in the wheat-*Stagonospora nodorum* pathosystem parallels that of the wheat-tan spot system. Genome **49:** 1265–1273.

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

Liu, Z.H., Friesen, T.L., Rasmussen, J.B., Ali, S., Meinhardt, S.W., and Faris, J.D. (2004b). Quantitative trait loci analysis and mapping of seedling resistance to *Stagonospora nodorum* leaf blotch in wheat. Phytopathology **94:** 1061–1067.

Lowe, T.M., and Eddy, S.R. (1997). tRNAscan-SE: A program for improved detection of transfer RNA genes in genomic sequences. Nucleic Acids Res. **25:** 955–964.

Majoros, W.H., Pertea, M., Antonescu, C., and Salzberg, S.L. (2003). GlimmerM, Exonomy and Unveil: Three ab initio eukaryotic genefinders. Nucleic Acids Res. **31:** 3601–3604.

Manning, V.A., Andrie, R.M., Trippe, A.F., and Ciuffetti, L.M. (2004). Ptr ToxA requires multiple motifs for complete activity. Mol. Plant Microbe Interact. **17:** 491–501.

Manning, V.A., and Ciuffetti, L.M. (2005). Localization of Ptr ToxA produced by *Pyrenophora tritici-repentis* reveals protein import into wheat mesophyll cells. Plant Cell **17:** 3203–3212.

Marchler-Bauer, A., et al. (2005). CDD: A Conserved Domain Database for protein classification. Nucleic Acids Res. **33:** D192–D196.

Marchler-Bauer, A., and Bryant, S.H. (2004). CD-Search: Protein domain annotations on the fly. Nucleic Acids Res. **32:** W327–331.

Meinhardt, S.W., Cheng, W.J., Kwon, C.Y., Donohue, C.M., and Rasmussen, J.B. (2002). Role of the arginyl-glycyl-aspartic motif in the action of Ptr ToxA produced by *Pyrenophora tritici-repentis.* Plant Physiol. **130:** 1545–1551.

Moriwaki, A., Kihara, J., Kobayashi, T., Tokunaga, T., Arase, S., and Honda, Y. (2004). Insertional mutagenesis and characterization of a polyketide synthase gene (PKS1) required for melanin biosynthesis in *Bipolaris oryzae.* FEMS Microbiol. Lett. **238:** 1–8.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol. **48:** 443–453.

Nicholson, T.P., Rudd, B.A., Dawson, M., Lazarus, C.M., Simpson, T.J., and Cox, R.J. (2001). Design and utility of oligonucleotide gene probes for fungal polyketide synthases. Chem. Biol. **8:** 157–178.

Oide, S., Moeder, W., Krasnoff, S., Gibson, D., Haas, H., Yoshioka, K., and Turgeon, B.G. (2006). NPS6, encoding a nonribosomal peptide synthetase involved in siderophore-mediated iron metabolism, is a conserved virulence determinant of plant pathogenic ascomycetes. Plant Cell **18:** 2836–2853.

Oliver, R.P., and Ipcho, S.V.S. (2004). Arabidopsis pathology breathes new life into the necrotrophs vs. biotrophs classification of fungal pathogens. Mol. Plant Pathol. **5:** 347–352.

Orbach, M. (1989). Electrophoretic characterisation of the *Magnaporthe grisea* genome. Fungal Genet. Newsl. **14:** 14.

Orbach, M., Vollrath, D., Davis, R., and Yanofsky, C. (1988). An electrophoretic karyotype for *Neurospora crassa.* Mol. Cell. Biol. **8:** 1469–1473.

Padovan, A.C., Sanson, G.F., Brunstein, A., and Briones, M.R. (2005). Fungi evolution revisited: Application of the penalized likelihood method to a bayesian fungal phylogeny provides a new perspective on phylogenetic relationships and divergence dates of ascomycota groups. J. Mol. Evol. **60:** 726–735.

Peng, T., Shubitz, L., Simons, J., Perrill, R., Orsborn, K.I., and Galgiani, J.N. (2002). Localization within a proline-rich antigen (Ag2/PRA) of protective antigenicity against infection with *Coccidioides immitis* in mice. Infect. Immun. **70:** 3330–3335.

Price, A.L., Jones, N.C., and Pevzner, P.A. (2005). De novo identification of repeat families in large genomes. Bioinformatics **21**(Suppl 1): i351–i358.

Proctor, R.H., Desjardins, A.E., Plattner, R.D., and Hohn, T.M. (1999). A polyketide synthase gene required for biosynthesis of fumonisin mycotoxins in *Gibberella fujikuroi* mating population A. Fungal Genet. Biol. **27:** 100–112.

Rawson, J.M. (2000). Transposable Elements in the Phytopathogenic Fungus *Stagonospora nodorum.* PhD dissertation (Birmingham, UK: University of Birmingham).

Rice, P., Longden, I., and Bleasby, A. (2000). EMBOSS: The European Molecular Biology Open Software Suite. Trends Genet. **16:** 276–277.

Richards, T.A., Dacks, J.B., Jenkinson, J.M., Thornton, C.R., and Talbot, N.J. (2006). Evolution of filamentous plant pathogens: Gene exchange across eukaryotic kingdoms. Curr. Biol. **16:** 1857–1864.

Robbertse, B., Reeves, J.B., Schoch, C.L., and Spatafora, J.W. (2006). A phylogenomic analysis of the Ascomycota. Fungal Genet. Biol. **43:** 715–725.

Rodriguez, F., Oliver, J.F., Martin, A., and Medina, J.R. (1990). The general stochastic model of nucleotide substitution. J. Theor. Biol. **142:** 485–501.

Schoch, C.L., Shoemaker, R.A., Seifert, K.A., Hambleton, S., Spatafora, J.W., and Crous, P.W. (2006). A multigene phylogeny of the Dothideomycetes using four nuclear loci. Mycologia **98:** 1041–1052.

Shimizu, T., Kinoshita, H., Ishihara, S., Sakai, K., Nagai, S., and Nihira, T. (2005). Polyketide synthase gene responsible for citrinin biosynthesis in *Monascus purpureus.* Appl. Environ. Microbiol. **71:** 3453–3457.

Solomon, P.S., Tan, K.-C., and Oliver, R.P. (2003a). The nutrient supply of pathogenic fungi; a fertile field for study. Mol. Plant Pathol. **4:** 203–210.

Solomon, P.S., Thomas, S.W., Spanu, P., and Oliver, R.P. (2003b). The utilisation of di/tripeptides by *Stagonospora nodorum* is dispensable for wheat infection. Physiol. Mol. Plant Pathol. **63:** 191–199.

Solomon, P.S., Lee, R.C., Wilson, T.J.G., and Oliver, R.P. (2004a). Pathogenicity of *Stagonospora nodorum* requires malate synthase. Mol. Microbiol. **53:** 1065–1073.

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

Solomon, P.S., Tan, K.C., Sanchez, P., Cooper, R.M., and Oliver, R.P. (2004b). The disruption of a G alpha subunit sheds new light on the pathogenicity of *Stagonospora nodorum* on wheat. Mol. Plant Microbe Interact. **17:** 456–466.

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

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

Spatafora, J.W., et al. (2006). A five-gene phylogenetic analysis of the Pezizomycotina. Mycologia **98:** 1018–1028.

Stamatakis, A. (2006). RAxML-VI-HPC: Maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics **22:** 2688–2690.

Stukenbrock, E.H., Banke, S., and McDonald, B.A. (2006). Global migration patterns in the fungal wheat pathogen *Phaeosphaeria nodorum.* Mol. Ecol. **15:** 2895–2904.

Stukenbrock, E.H., and McDonald, B.A. (2007). Geographic variation and positive diversifying selection in the host specific toxin *SnToxA.* Mol. Plant Pathol. **8:** 321–322.

Sumii, M., Furutani, Y., Waschuk, S.A., Brown, L.S., and Kandori, H. (2005). Strongly hydrogen-bonded water molecule present near the retinal chromophore of *Leptosphaeria rhodopsin*, the bacteriorhodopsin-like proton pump from a eukaryote. Biochemistry **44:** 15159–15166.

Talbot, N., Salch, Y., Ma, M., and Hamer, J. (1993). Karyotypic variation within clonal lineages of the rice blast fungus, *Magnaporthe grisea.* Appl. Environ. Microbiol. **59:** 585–593.

Talbot, N.J., McCafferty, H.R.K., Ma, M., Moore, K., and Hamer, J.E. (1997). Nitrogen starvation of the rice blast fungus *Magnaporthe grisea* may act as an environmental cue for disease symptom expression. Physiol. Mol. Plant Pathol. **50:** 179–195.

Tambor, J., Guedes, R., Nobrega, M., and Nobrega, F. (2006). The complete DNA sequence of the mitochondrial genome of the dermatophyte fungus *Epidermophyton floccosum.* Curr. Genet. **49:** 302–308.

Thomma, B., Bolton, M.D., Clergeot, P.H., and De Wit, P. (2006). Nitrogen controls in planta expression of *Cladosporium fulvum* Avr9 but no other effector genes. Mol. Plant Pathol. **7:** 125–130.

Thompson, J.D., Gibson, T.J., Plewniak, F., Jeanmougin, F., and Higgins, D.G. (1997). The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res. **25:** 4876–4882.

Trail, F., Xu, J.R., San Miguel, P., Halgren, R.G., and Kistler, H.C. (2003). Analysis of expressed sequence tags from *Gibberella zeae* (anamorph *Fusarium graminearum*). Fungal Genet. Biol. **38:** 187–197.

Vizcaino, J.A., Cardoza, R.E., Dubost, L., Bodo, B., Gutierrez, S., and Monte, E. (2006). Detection of peptaibols and partial cloning of a putative peptaibol synthetase gene from *T. harzianum* CECT 2413. Folia Microbiol. (Praha) **51:** 114–120.

Wessels, J.G.H. (1994). Developmental regulation of fungal cell wall formation. Annu. Rev. Phytopathol. **32:** 413–437.

Winefield, R.D., Hilario, E., Beever, R.E., Haverkamp, R.G., and Templeton, M.D. (2007). Hydrophobin genes and their expression in conidial and aconidial Neurospora species. Fungal Genet. Biol. **44:** 250–257.

Winka, K., and Eriksson, O.E. (1997). Supraordinal taxa of Ascomycota. Mol Phylogenet Evol. **1:** 1–16.

Wolpert, T.J., Dunkle, L.D., and Ciuffetti, L.M. (2002). Host-selective toxins and avirulence determinants: What's in a name? Annu. Rev. Phytopathol. **40:** 251–285.

Woo, P., et al. (2003). The mitochondrial genome of the thermal dimorphic fungus *Penicillium marneffei* is more closely related to those of molds than yeasts. FEBS Lett. **555:** 469–477.

Yadav, G., Gokhale, R.S., and Mohanty, D. (2003). SEARCHPKS: A program for detection and analysis of polyketide synthase domains. Nucleic Acids Res. **31:** 3654–3658.

Yoder, O.C., and Turgeon, B.G. (2001). Fungal genomics and pathogenicity. Curr. Opin. Plant Biol. **4:** 315–321.

Yun, S.H., Turgeon, B.G., and Yoder, O.C. (1998). REMI-induced mutants of *Mycosphaerella zeae-maydis* lacking the polyketide PM-toxin are deficient in pathogenesis to corn. Physiol. Mol. Plant Pathol. **52:** 53–66.

Zhang, Z., Schwartz, S., Wagner, L., and Miller, W. (2000). A greedy algorithm for aligning DNA sequences. J. Comput. Biol. **7:** 203–214.

Zolan, M.E. (1995). Chromosome-length polymorphism in fungi. Microbiol. Rev. **59:** 686–698.

Chapter 3: Attribution Statement

Title:	**RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences.**
Authors:	**James K. Hane** and Richard P. Oliver
Citation:	*BMC Bioinformatics* 9:478 (2008)

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 (JKH) made the following contributions to this chapter:

- Performed bioinformatics analyses described in this chapter.

The following contributions were made by co-authors:

- JKH and RPO wrote the manuscript.

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

--   --

James Hane (Ph. D. candidate)   Date

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

--   --

Richard Oliver (Principal supervisor)   Date

BMC Bioinformatics

BioMed Central

Methodology article

RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences

James K Hane and Richard P Oliver*

Address: Australian Centre for Necrotrophic Fungal Pathogens, Faculty of Health Sciences, Murdoch University, South Street, Murdoch, 6150, Australia

Email: James K Hane - j.hane@murdoch.edu.au; Richard P Oliver* - roliver@murdoch.edu.au

* Corresponding author

Published: 12 November 2008

BMC Bioinformatics 2008, **9**:478 doi:10.1186/1471-2105-9-478

Received: 28 April 2008
Accepted: 12 November 2008

This article is available from: http://www.biomedcentral.com/1471-2105/9/478

Abstract

Background: Repeat-induced point mutation (RIP) is a fungal-specific genome defence mechanism that alters the sequences of repetitive DNA, thereby inactivating coding genes. Repeated DNA sequences align between mating and meiosis and both sequences undergo C:G to T:A transitions. In most fungi these transitions preferentially affect CpA di-nucleotides thus altering the frequency of certain di-nucleotides in the affected sequences. The majority of previously published *in silico* analyses were limited to the comparison of ratios of pre- and post-RIP di-nucleotides in putatively RIP-affected sequences – so-called RIP indices. The analysis of RIP is significantly more informative when comparing sequence alignments of repeated sequences. There is, however, a dearth of bioinformatics tools available to the fungal research community for alignment-based RIP analysis of repeat families.

Results: We present RIPCAL http://www.sourceforge.net/projects/ripcal, a software tool for the automated analysis of RIP in fungal genomic DNA repeats, which performs both RIP index and alignment-based analyses. We demonstrate the ability of RIPCAL to detect RIP within known RIP-affected sequences of *Neurospora crassa* and other fungi. We also predict and delineate the presence of RIP in the genome of *Stagonospora nodorum* – a Dothideomycete pathogen of wheat. We show that RIP has affected different members of the *S. nodorum* rDNA tandem repeat to different extents depending on their genomic contexts.

Conclusion: The RIPCAL alignment-based method has considerable advantages over RIP indices for the analysis of whole genomes. We demonstrate its application to the recently published genome assembly of *S. nodorum*.

Background

Over 100 fungal genome sequences have been obtained or are in the pipeline [1] and next-generation sequencing technologies will further accelerate the accumulation of data over the next decade. This rapidly growing array of sequence information presents many new challenges for analysis. There is an urgent need to develop and implement efficient tools to describe features of new genomes. Repeat-induced point mutation (RIP) is one such area of fungal biology requiring efficient analytical tools. RIP is an irreversible genome defence mechanism first detected in *Neurospora crassa* [2,3] and subsequently in *Magnaporthe grisea* [4,5], *Podospora anserina* [6] and *Leptosphaeria maculans* [7]. RIP is believed to be a defence against transposons, rendering them inactive and protecting sexual progeny from the expression of transposon genes.

Direct experimental observation of RIP requires both that the fungal species can be crossed under laboratory conditions and that the strain can be transformed with multiple copies of a transgene. Very few fungal species are amenable to such analysis and these procedures are slow in all cases. RIP-like processes can also be detected by *in-silico* analysis of repeated elements in whole or partial genomic sequences. Prior examples include *Aspergillus fumigatus* [8], *Fusarium oxysporum* [9-11], *Aspergillus nidulans* [12], *Microbotryum violaceum* [13], *Magnaporthe oryzae* [14], *Aspergillus niger* [15] and *Penicillium chysogenum* [15]. We now have the opportunity to detect and measure RIP *in silico* from genomic sequences of diverse species.

RIP involves transitions from C:G to T:A nucleotides in pairs of duplicated sequences during the dikaryotic phase between mating and meiosis [2,3]. RIP changes are scattered throughout both sequences where pairs share more than ~80% identity [16] and are over 400 bp in length [17]. C:G transitions are not random within affected sequences. Particular CpN dinucleotides are preferentially altered over others (Table 1). In *N. crassa*, CpA di-nucleotides were preferentially altered [18]. Thus a strong bias towards CpA to TpA changes (or TpG to TpA in the complementary strand) was observed. This resulted in a relative decrease in CpA and TpG and a corresponding increase in TpA di-nucleotides within RIP-affected sequences. These changes in di-nucleotide frequencies can be used to identify RIP-affected repeats by measuring the ratios of pre-RIP and post-RIP di-nucleotides within a set of repeated sequences. This generates a single statistic called a "RIP index" (plural: RIP indices). High frequencies of post-RIP and low frequencies of pre-RIP di-nucleotides are straightforward to detect by this method and useful for identifying RIP-affected sequences. The RIP indices TpA/Apt and (CpA+TpG)/(ApC+GpT), originally developed by Margolin *et al* [19], are commonly used to detect RIP *in silico* [8,12,19,20]. TpA/ApT is the simplest index and measures the frequency of TpA RIP products with correction for false positives due to A:T rich regions. Higher values of TpA/ApT indicate a stronger RIP response. The index (CpA+TpG)/(ApC+GpT) is similar in principle to TpA/ApT but measures the depletion of the RIP targets CpA and TpG. In this case lower values of (CpA+TpG)/(ApC+GpT) are indicative of stronger RIP.

Table 1: The four possible CpN→TpN di-nucleotide RIP mutations and their reverse complements which form the basis for comparisons to determine the dominant form of RIP mutation in both alignment-based and statistical analyses.

RIP Mutation	Reverse Complement*
CpA→TpA	TpG→TpA
CpC→TpC	GpG→GpA
CpG→TpG	CpG→CpA
CpT→TpT	ApG→ApA

(*The forward orientation is often referred to in the text, however should be understood to include the reverse complement unless specified otherwise.)

RIP-indices are simple to calculate and do not require complete knowledge of the genome sequence or repeat families. They are also applicable to heavily mutated repeat families for which an alignment is not possible or questionable. However, RIP indices are insensitive tools which obscure many interesting features of RIP. These include the direction of RIP changes (i.e. which sequence is closer to the ancestral precursor of the RIP-affected sequence), the degree of RIP along the length of repeat alignments and differences in RIP profiles between members of the repeat class.

As RIP operates on aligned sequences, these questions are better answered using an alignment-based approach. Alignment-based analysis of RIP involves the multiple alignment of a repeat family and counting RIP mutations along the alignment for all sequences. This method has been previously used to identify RIP within the Ty1 transposon family of *Microbotryum violaceum* using the software tool Sequencher. Such manual calculation of RIP as was used by Hood *et al* [13] does not lend itself to whole genome RIP analysis. To enable a thorough, facile and automated analysis of RIP in the plethora of new fungal genomes, we have developed the free software tool RIPCAL (available at http://www.sourceforge.net/projects/ripcal. RIPCAL incorporates both RIP index and alignment-based methods. Its capabilities are demonstrated with examples taken from *de novo*-defined repeat families of the recently published *Stagonospora nodorum* genome, a major fungal pathogen of wheat [21,22].

Results

Validation of RIP detection by the alignment-based method

The RIPCAL alignment-based method was applied to both the 5S rDNA repeat family of *Neurospora crassa*, which is reportedly free from RIP mutation due to its short sequence length [17], and to the Tad1 transposons of *N. crassa*, which are reported to be heavily prone to CpA→TpA RIP mutation [23]. The 5S rDNA and Tad1 repeat families served as negative and positive controls for RIP respectively. Analysis showed low levels of RIP mutation among 5S rDNAs, whereas high levels of RIP mutation were detected amongst Tad1 transposons as expected (Additional file 1). Interestingly, while CpA↔TpA changes were highly increased in the Tad1 family, these were overshadowed by a major increase in CpT↔TpT mutation, which has not been previously detected [23]. This may be due to the fact that the former study compared Tad1 sequences between different strains of *Neurospora crassa*, whereas this comparison was restricted to all repeats within a single strain.

Identification of the dominant CpN to TpN di-nucleotide mutation in RIP-affected sequences

De novo RIP analysis of a fungal repeat unit first requires the identification of the most affected CpN di-nucleotides. The MATE transposon repeat family of *Aspergillus nidulans* and the Ty1 Copia-like transposon family of the Basidiomycete *Microbotryum violaceum* were analysed by RIPCAL. *A. nidulans* MATE repeats are reported to exhibit a dual preference for CpG→TpG and CpA→TpA RIP mutation in descending order of magnitude [24]. The Ty1 repeats of *M. violaceum* were reported to exhibit a strong preference for CpG→TpG di-nucleotide RIP mutation [13]. High levels of CpG→TpG and CpA→TpA RIP mutation were detected in the MATE transposons (Additional file 2). RIPCAL also detected the CpG→TpG bias in the Ty1 repeats of *M. violaceum* (Additional file 1). Hood *et al* have reported preferential mutation of the tri-nucleotide TpCpG to TpTpG in Ty1 [13], however RIPCAL is not currently designed to detect a tri-nucleotide RIP bias.

Di-nucleotide frequency and index analysis of RIP mutation in Stagonospora nodorum

RIPCAL di-nucleotide frequency analyses of the previously identified de novo repeat families Molly, Pixie, Elsa, Y1 (rDNA repeat), R8, R9, R10, R22, R25, R31, R37, R38, R39, R51, X0, X3, X11, X12, X15, X23, X26, X28, X35, X36, X48 and X96 [21] of the *S. nodorum* genome were performed and indicated depletion of the CpA, CpC, CpG, GpG and TpG di-nucleotide targets of RIP-mutation (Figure 1, Additional file 2). Of the RIP di-nucleotide products, only TpA showed a corresponding increase. This suggests that CpA to TpA is the dominant form of CpN→TpN di-nucleotide mutation in repeats of *S. nodorum*, as observed in *N. crassa* and *P. anserina* [6,20]. This is corroborated by RIP index analysis. RIP indices for TpA/ApT were well in excess of *S. nodorum* non-repetitive control sequences indicating high frequencies of the TpA RIP product in the repeat families. The (CpA+TpG)/(ApC+GpT) index was below control levels indicating depletion of the CpA and TpG RIP targets in the repeats. Both dinucleotide frequency and RIP index analyses strongly indicated that the mutation of CpA to TpA was the dominant form of di-nucleotide RIP mutation in the repeat families of *S. nodorum* (Table 2, Additional file 2).

Alignment-based analysis of RIP mutation in Stagonospora nodorum

Repeat families of *S. nodorum* were aligned and scanned for RIP-like di-nucleotide changes using RIPCAL. RIP mutation statistics for all repeat families of *S. nodorum* are summarised in Additional file 2. Alignment-based analysis indicated that the dominant form of CpN-targeted RIP mutation in *S. nodorum* repeats was CpA to TpA as observed by index analysis. High levels of CpT to TpT mutation were also observed in some repeat classes (Additional file 2).

In this analysis we introduce a statistic called 'RIP dominance'. RIP dominance is the ratio of a particular CpN↔TpN RIP mutation over the sum of the other 3 alternative CpN↔TpN mutations within a multiple alignment (or sub-alignment). This was used to determine the relative strength of CpA to TpA type RIP mutations in *S. nodorum* (Table 2).

RIPCAL analysis of the XO repeat family of predicted non-LTR transposons is shown in Figure 2. The alignment (Figure 2A) displays the range of repeat sizes, sequence coverages and locations of RIP mutation for individual repeats. The repeat with the highest total G:C content was chosen as the least RIP-mutated model for comparison to all aligned sequences. CpN↔TpN di-nucleotide changes are colour-coded and show that CpA to TpA changes far outweighed all other CpN to TpN di-nucleotide mutations. Figure 2B shows the same data summarised as a rolling frequency graph. The RIP dominance for CpA↔TpA mutation in XO was 2.13, meaning that the CpA↔TpA mutation was more than twice as frequent as the sum of CpC, CpG and CpT-targeted RIP mutations. Each repeat element in this family showed a relatively equal degree of RIP. A slight tendency towards higher RIP incidence was found towards the ends of the alignment. XO appears to be a simple repeat unit which is highly and evenly RIP-affected.

The *S. nodorum* rDNA repeat family provided a more complex example. *S. nodorum* Y1/rDNA repeats are located within a large tandem array on scaffold 5 and as non-tandem remnants scattered elsewhere throughout the genome. The non-tandem remnants were sub-divided into those longer or shorter than 1 kb. rDNA sub-classes differed markedly from the non-repetitive control by changes in di-nucleotide frequency (Figure 3). Tandem rDNA repeats appeared to be the least RIP affected in terms of Cp(A/C/G) depletion and increases in TpA, followed by the non-tandem and short repeats. RIP index analysis showed a similar trend (Table 2). Tandem, non-tandem and short rDNA repeats had TpA/ApT index scores of 2.08, 2.68 and 3.55 respectively. These values were among the highest TpA/ApT scores of all repeat classes suggesting extreme RIP mutation. The (CpA + TpG)/(ApC + GpT) index gave a similar result. Tandem, non-tandem and short rDNA repeats scored 0.94, 0.69 and 0.25 respectively. These values were among the lowest for all repeat classes, again suggesting extreme RIP muta-

Table 2: Analysis of *Stagonospora nodorum* repeat families for evidence of RIP ranked by CpA↔TpA dominance.

Repeat Family	$\frac{TpA}{ApT}$	$\frac{CpA+TpG}{ApC+GpT}$	CpA↔TpA Dominance	Alignment Length	Description/Homology
R8	1.70 ± 0.03	0.74 ± 0.02	2.96	9548	Ubiquitin conjugating enzyme
X0	1.75 ± 0.02	0.49 ± 0.02	2.13	4103	Non LTR transposon
R10	1.76 ± 0.06	0.56 ± 0.05	1.91	1360	Unknown
R9	1.99 ± 0.03	0.48 ± 0.02	1.88	4483	Non LTR transposon
X48	1.35 ± 0.11	1.26 ± 0.13	1.82	275	Sub-telomeric repeat
rDNA Non-tandem	2.68 ± 0.18	0.69 ± 0.04	1.50	9938	Non-array rDNA repeats ≥ 1 kb
X35	1.76 ± 0.07	0.58 ± 0.07	1.50	1185	Sub-telomeric repeat
MOLLY	1.90 ± 0.06	0.40 ± 0.04	1.21	1946	Mariner-like transposon
R22	1.73 ± 0.08	0.27 ± 0.04	1.20	710	Sub-telomeric repeat
X26	1.70 ± 0.03	0.52 ± 0.02	1.16	5034	Sub-telomeric repeat/Transposon remnant
R31	1.65 ± 0.04	0.44 ± 0.04	0.99	3119	Unknown
X36	1.97 ± 0.18	0.44 ± 0.10	0.89	516	Unknown
X96	1.87 ± 0.19	0.56 ± 0.18	0.87	319	Unknown
ELSA	1.67 ± 0.04	0.46 ± 0.05	0.86	5273	Copia-like transposon
X11	2.04 ± 0.03	0.38 ± 0.02	0.83	7570	Gypsy-like transposon
X28	2.22 ± 0.13	0.39 ± 0.03	0.83	1975	Unknown
PIXIE	1.84 ± 0.07	0.36 ± 0.03	0.77	1918	Mariner-like transposon
X12	2.06 ± 0.07	0.33 ± 0.04	0.67	2059	Gypsy-like transposon
X3	1.91 ± 0.03	0.74 ± 0.01	0.63	10673	Helicase
X15	1.87 ± 0.04	0.33 ± 0.02	0.61	6437	Sub-telomeric repeat/Gypsy-like transposon
R39	1.92 ± 0.08	0.30 ± 0.03	0.59	2102	Unknown
rDNA Tandem	2.08 ± 0.09	0.94 ± 0.02	0.53	9938	rDNA repeats in tandem array
R37	1.85 ± 0.03	0.28 ± 0.02	0.49	2264	Mariner-like transposon
R51	1.93 ± 0.07	0.27 ± 0.03	0.47	870	Unknown
X23	1.85 ± 0.09	0.31 ± 0.03	0.45	613	Unknown
rDNA Short	3.55 ± 0.39	0.25 ± 0.03	0.26	280*	Non-array rDNA repeats < 1 kb
R25	2.16 ± 0.15	0.31 ± 0.03	0.25	3407	Transposon remnant
R38	2.10 ± 0.18	0.24 ± 0.05	0.20	391	Unknown
Non-Repetitive Control	0.83 ± 0.01	1.25 ± 0.00	N/A	N/A	Genomic regions not corresponding to repeat matches

Two RIP index scores are given within a 95% confidence interval. The alignment-based comparison of CpA↔TpA RIP-mutation is used to give dominance score. CpA↔TpA dominance is a numerical indicator of frequency of that mutation over other CpN↔TpN mutations as described in the methods. Values for the rDNA repeat are sub-classified according to physical location and length. The length of the alignment is also given. (*rDNA short alignments are a subset of the full-length rDNA alignment.)

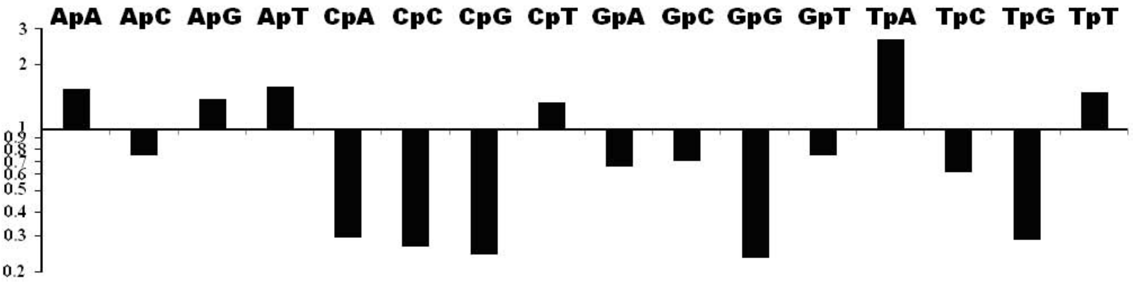


Figure 1

Fold changes in di-nucleotide abundances for all repeat families of *Stagonospora nodorum* compared to non-repetitive control sequence on a Log_{10} scale. This conforms to the expected pattern associated with classical CpA→TpA type RIP mutation: high TpA and low CpA and TpG abundances.

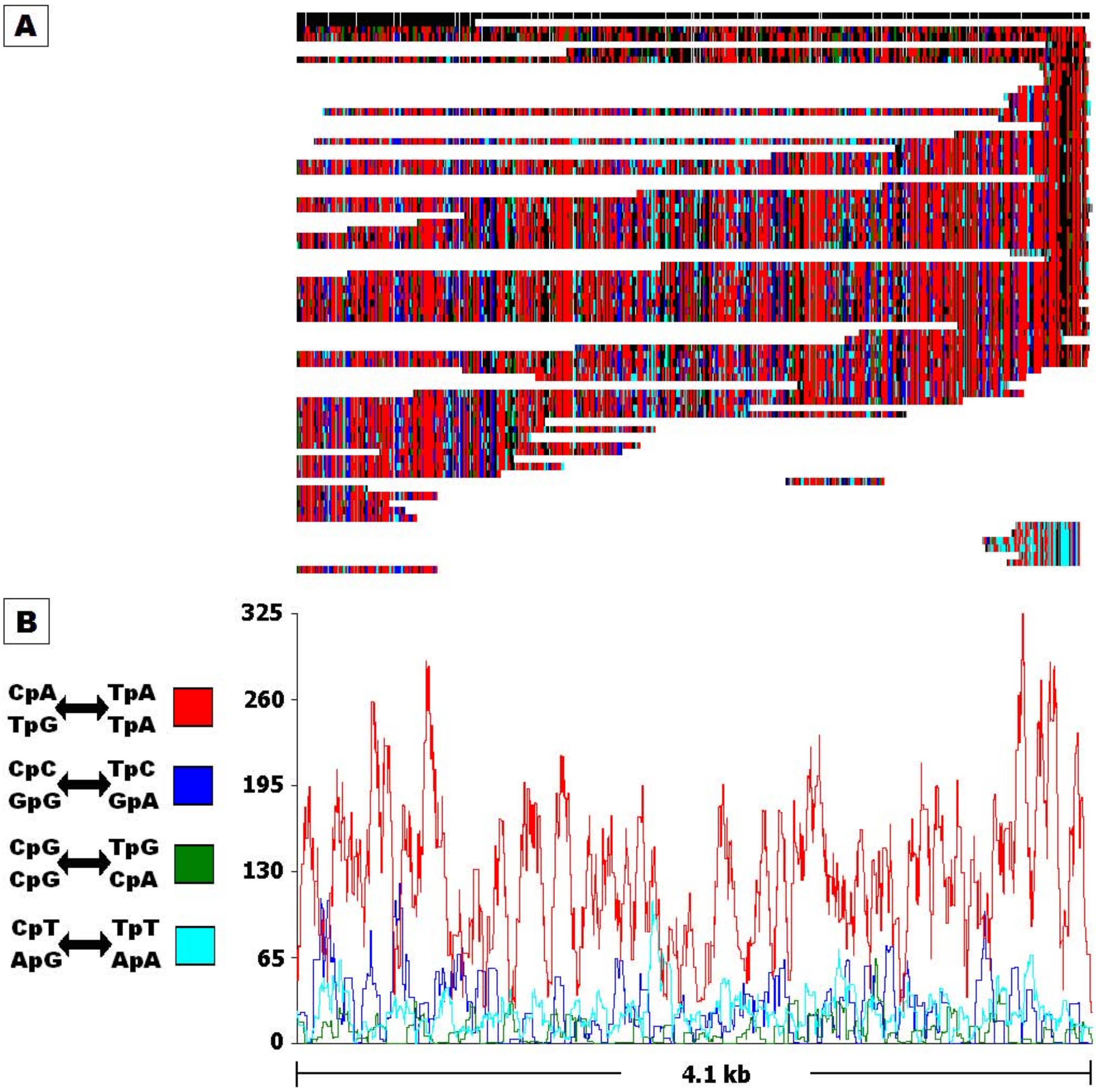


Figure 2
RIPCAL analysis of the X0 repeat family of *Stagonospora nodorum*, representative of a repeat family exhibiting strongly dominant classical CpA→TpA type RIP mutation. A) multiple alignment of the putative transposon repeat family X0 compared to highest G:C content model. Incomplete repeated regions are typical for repeat family alignments illustrated by the blocks in white in panel A. Black = match; grey = mismatch; white = gap. Mismatches corresponding to selected di-nucleotide changes are coloured as indicated. B) Overall RIP mutation frequency graph over a 50 bp scanning window, corresponding to the alignment above, demonstrating the overall dominance of the CpA↔TpA mutation over other CpN↔TpN mutations for the X0 repeat family.

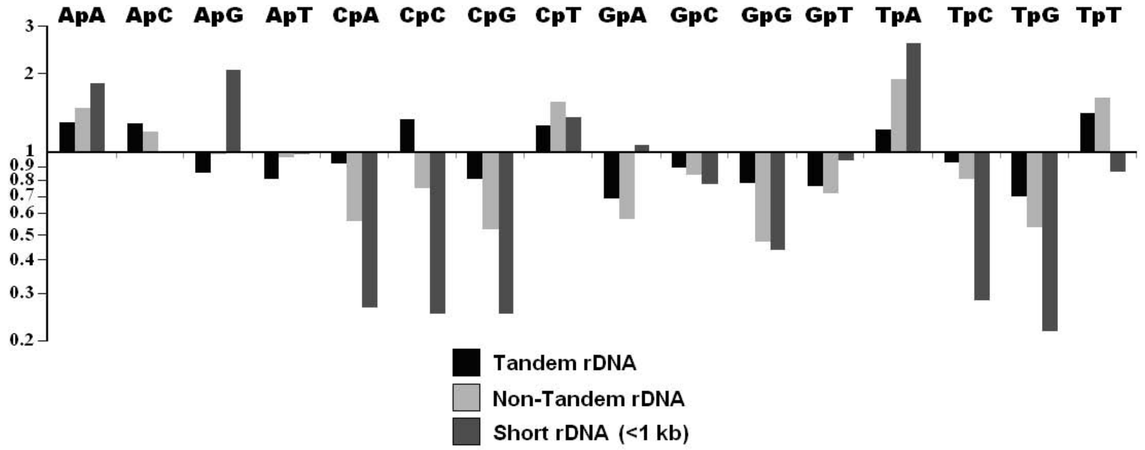


Figure 3

Fold changes in di-nucleotide abundances between *Stagonospora nodorum* rDNA repeat sub-categories. Tandem (black), non-tandem (light-grey) and short < 1 kb (dark grey) on a Log_{10} scale. Tandem rDNA repeats exhibit lesser variations in TpA, CpA and TpG counts, therefore are less RIP-affected than non-tandem and short < 1 kb rDNA repeats.

tion in the rDNA repeat sub-classes. In all cases, the short rDNA repeats had particularly extreme scores, suggesting that these were the most RIP-affected.

When analysed by alignment (Figure 4), a more comprehensive picture emerged. The frequencies of CpN to TpN mutations (Figure 4B) indicated that CpA to TpA mutation was the dominant form of RIP mutation for the rDNA repeat family. However the distribution of RIP mutation within the alignment (Figure 4A) shows distinct differences in RIP profiles between the three rDNA sub-classes. The tandem rDNA repeats were generally unaffected by CpN-targeted mutation. Interestingly, a single tandem repeat was identified that had undergone extensive CpA to TpA changes. This proved to be the 5' terminal repeat within the rDNA array. The long non-tandem repeats were heavily affected by CpA to TpA RIP mutation, especially in the central regions. The short repeats showed no evidence of CpA to TpA RIP but did exhibit a high level of CpT to TpT RIP mutation. The CpA↔TpA RIP dominance score for non-tandem rDNA repeats was 1.5, whereas the tandem and short sub-classes had low scores of 0.53 and 0.26 (Table 2). This indicated heavy RIP mutation in non-tandem repeats and absence or low levels of RIP in tandem and short rDNA repeats.

Discussion

The alignment-based method employed by RIPCAL is an efficient, accurate and reliable method of RIP detection and characterisation. RIPCAL successfully detected the presence and absence of RIP in the positive and negative *N. crassa* control sequences. RIPCAL also accurately determined the preferential CpN mutation bias in RIP-affected sequences. The CpG bias in Ty1 repeats of *M. violaceum* and the dual CpG and CpA bias in MATE repeats of *A. nidulans* were also identified consistent with previously published results [13,24].

Di-nucleotide frequency, RIP index and alignment-based analyses all indicated that CpA to TpA mutation was the dominant CpN-targeted mutation in the repeat families of *S. nodorum*. This preference is common to most known RIP-affected fungi. The high incidence of CpT to TpT mutation detected by alignment is less common, but has been observed in *Magnaporthe grisea* accompanying CpA-targeted mutation in RIP-affected sequences [4,5]. However high levels of CpT to TpT mutation within *S. nodorum* short rDNA repeats, which are presumably unaffected by RIP, suggest that CpT-targeted mutation may not related to RIP in *S. nodorum*. Further experimental evidence is required to confirm to relevance of CpT to TpT mutation to RIP in *S. nodorum* and other Fungi.

RIPCAL alignment-based analysis displays the physical distribution of RIP along an alignment as shown in for the X0 repeat family in Figure 2 and the Y1/rDNA repeat family in Figure 4. This allows detection of individual repeats with anomalous changes, such as the single RIP-affected tandem rDNA repeat (Figure 4A). The lack of CpA to TpA mutation within the tandem rDNA repeats adds further supporting evidence for RIP-resistance within the rDNA nucleolus organiser region (NOR) [2,25]. However, the RIP-affected tandem repeat, located at the terminus of the rDNA array suggests that protection from RIP within the NOR has a finite range.

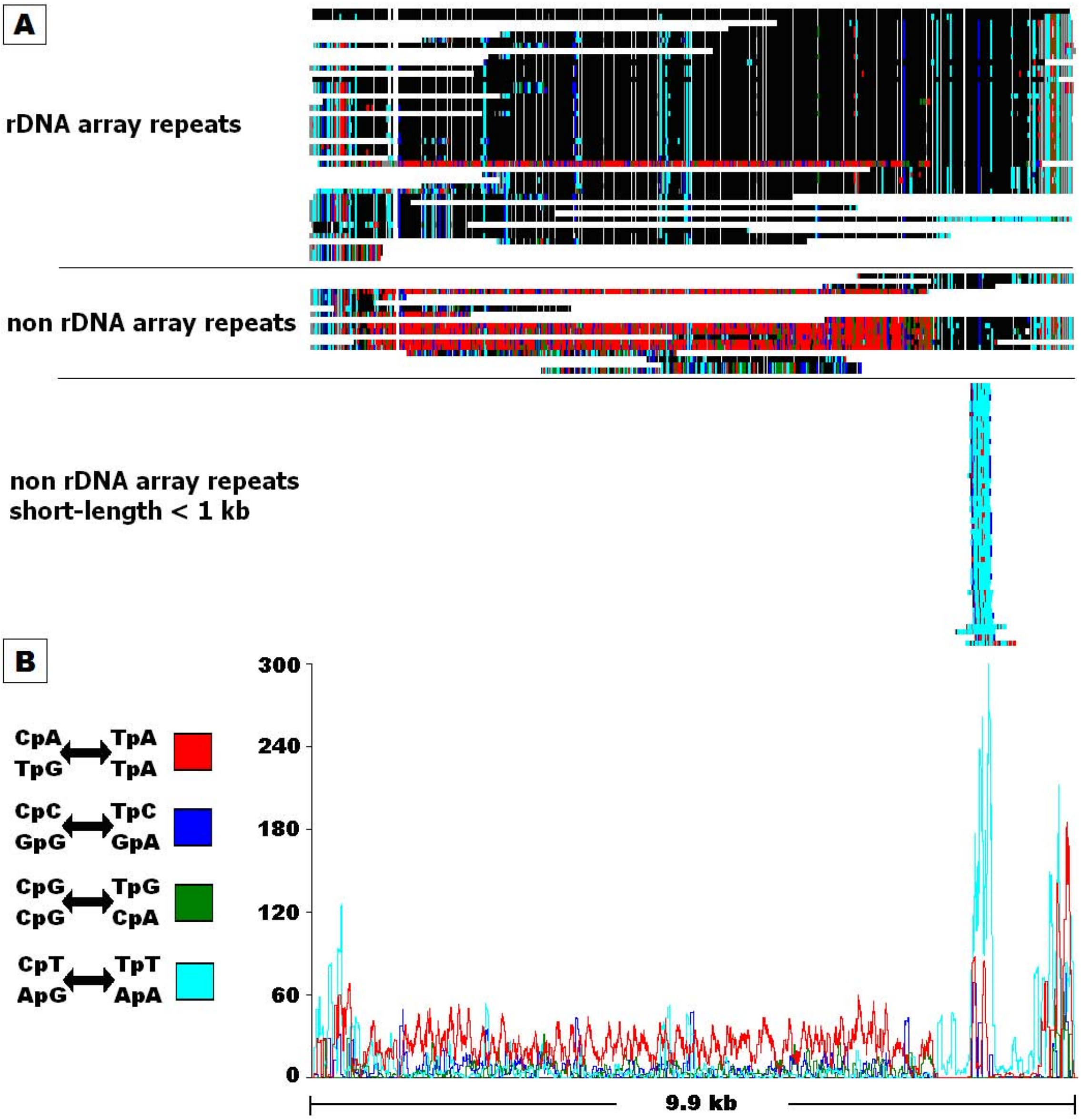


Figure 4

RIPCAL analysis of the rDNA tandem repeat of *Stagonospora nodorum*. A) multiple alignment of the rDNA repeat family compared to highest G:C content model. Annotation is as for figure 3. Classical CpA↔TpA type RIP mutations are generally limited to full length rDNA-like repeats not located within the rDNA tandem array. One copy within the rDNA array exhibits RIP-like alterations. B) Overall RIP mutation frequency graph over a 50 bp scanning window, corresponding to the alignment above, demonstrating even dominance of CpA↔TpA changes in the non-array full-length repeats except near each end of the alignment.

The close examination of the *S. nodorum* rDNA repeat sub-classes by alignment highlighted the poor performance of the RIP index based analyses. Differences in the extent of RIP mutation between DNA sub-classes by both TpA/ApT and (CpA + TpG)/(ApC + GpT) RIP indices were not as expected. This was particularly true for the short rDNA repeats which were predicted to exhibit the highest levels of RIP. Furthermore, both RIP indices predicted extreme RIP mutation in all sub-classes, which was only expected for the non-tandem rDNA repeats. Repeat order ranked by CpA↔TpA dominance is clearly different from that produced by either RIP index method (Table 2). The relationship between RIP index and CpA↔TpA dominance is shown in Figure 5A. There is no correlation ($R^2 = 0.135$) between the TpA/ApT RIP index and the CpA↔TpA dominance of *S. nodorum* repeats. Furthermore there was no significant correlation ($R^2 = 0.090$) between the two RIP indices (Figure 5B). We conclude that simple RIP indices are not reliable indicators of RIP mutation.

The length of a *S. nodorum* repeat class and the degree of RIP mutation did not appear to be related (Table 2). This was highlighted by X48, a short sub-telomeric repeat, which had a high CpA↔TpA dominance score of 1.82. Its length of 275 bp was well below the 400 bp length considered the minimum for RIP in *N. crassa* [17] and the 280 bp length of the S. nodorum short rDNA repeats (which do not display CpA to TpA changes). Alignment-based analysis predicted that sub-telomeric repeats were among the most RIP-susceptible. This may explain the high CpA↔TpA dominance of X48 as chromosome ends may be physically more accessible to the molecular RIP machinery. Alternatively, the X48 repeat may be recognised in conjunction with adjacent repeats as a single unit. Unlike the NOR, fungal telomeres do not appear to be immune to RIP. RIP-like changes have also been reported in the sub-telomeric gene *TLH* of *Magnaporthe oryzae* [14].

Conclusion

We present RIPCAL as a versatile and efficient tool for the analysis of RIP which simplifies existing index-based analyses and adds alignment-based RIP analysis as a feasible alternative for whole genome analysis. These analyses highlight significant deficiencies in index-based methods of RIP detection. The alignment-based approach is biologically relevant and reveals novel features and predictions that can be tested experimentally in appropriate organisms. Sifting through the expected flood of fungal genome sequences for RIP-like phenomena may provide insights on fungal lifestyle, genomics and evolution.

Methods

RIPCAL has multiple modes of operation involving different combinations of RIP index and alignment-based methods. RIPCAL can be run in either command-line or graphical modes and is Perl-based. It is also compiled as a Windows executable. Dependent on the analysis method, RIPCAL accepts sequence input in Fasta format, pre-aligned sequence input in Fasta or ClustalW format and repeat coordinate input in either version 2 or 3 GFF format. If pre-aligned input is not provided, RIPCAL can interface with a local installation of ClustalW [26]. Refer to Additional file 3 for more detailed information.

RIP index analysis

Index analysis can proceed from either direct Fasta input, or from both Fasta and GFF coordinate inputs. RIP index analyses count frequencies of single nucleotides and the 16 possible di-nucleotide combinations, which are used to calculate RIP indices. Sequences were divided into sub-sequences of ≤ 100 bp length and di-nucleotide counts were normalised for N content by:

$$\frac{Count \times (Length - Ncount)}{Length} \quad (1)$$

Where *Count* = di-nucleotide count, *Length* = length of sub-sequence and *Ncount* = count of unknown 'N' bases in sequence. Di-nucleotide counts were ignored where (*Length* - *Ncount*) < 10. The following indices have been published previously [19,27]:

$$\frac{TpA}{ApT} \quad (2)$$

$$\frac{CpA+TpG}{ApC+GpT} \quad (3)$$

Additional RIP indices that can be defined are of the form (CpN+NpG)/(TpN+NpA), which represents a ratio of conversion of pre-RIP di-nucleotides to post-RIP di-nucleotides, for the characteristic di-nucleotide mutation CpN→TpN and its reverse complement NpG→NpA (Table 1):

$$\frac{CpA+TpG}{TpA} \quad (4)$$

$$\frac{CpC+GpG}{TpC+GpA} \quad (5)$$

$$\frac{CpG}{TpG+CpA} \quad (6)$$

$$\frac{CpT+ApG}{TpT+ApA} \quad (7)$$

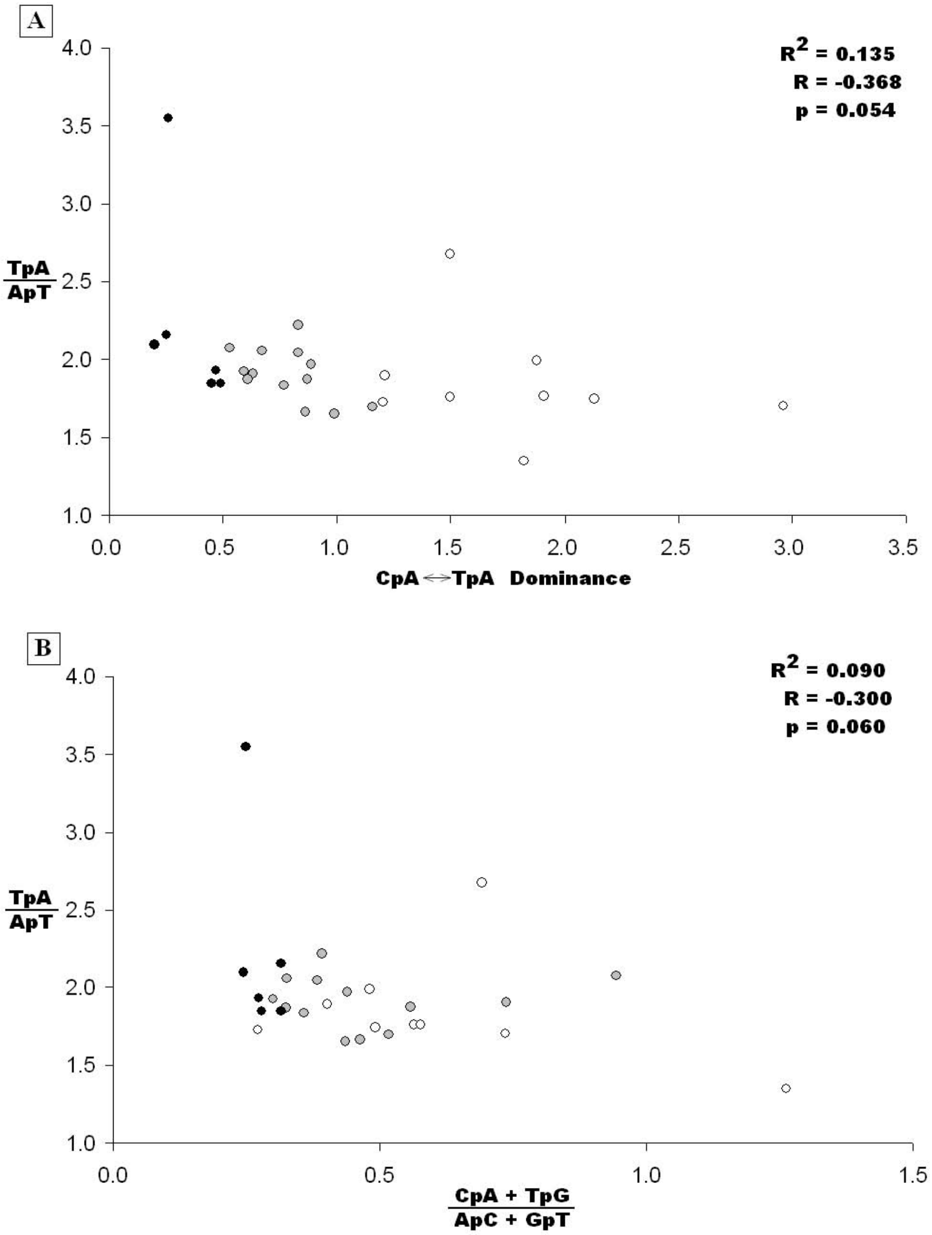


Figure 5

Comparison of RIP indices with alignment-based RIPCAL comparisons for repeat families of *Stagonospora nodorum*. A) Comparison of TpA/ApT RIP index with the alignment-based CpA↔TpA dominance. A positive correlation was expected however was not observed. B) Comparison of the TpA/ApT and (CpA+TpG)/(ApC+GpT) RIP indices. A negative correlation would be expected. Repeat families exhibiting low levels of RIP by alignment based analysis are represented by black dots (CpA↔TpA dominance < 0.5); medium families are grey (0.5 ≤ CpA↔TpA dominance ≥ 1.2); and high are white (CpA↔TpA dominance > 1.2).

When using GFF input, RIP index data for repeat features was compared to a non-repetitive control family. If repeat family information is contained within the GFF input (via the target attribute) then this process was also separated by family. Fold changes between repeat families and the control were determined by ΔNpN = (repeat NpN count)/(control NpN count), where NpN represents any di-nucleotide combination.

RIP index sequence scan

RIP indices are calculated over a user-defined window (default 200 bp). Using index thresholds as criteria for RIP, RIP-affected sub-regions were predicted and the output is given in GFF format. The default criteria for RIP within a sequence window were based on previously published data [19,27].

$$\frac{TpA}{ApT} \geq 0.89$$

$$\frac{CpT+ApT}{ApC+GpT} \leq 1.03$$

Where two windows meeting the above criteria overlap, the predicted sub-region was extended (Additional file 3). Sub-regions were subject to a minimum size threshold (default 300 bp) reflecting the existence of an experimentally observed size threshold for RIP [17]. Non-published indices were excluded by default, but can be employed as additional/replacement criteria using thresholds based on results obtained in this paper (Additional file 2). This method can be used to predict *de novo* ancient/non-repeated RIP-affected sequences. However, caution should be used with this method as the above threshold values are calibrated for RIP in *N. crassa*.

Alignment-based analysis

RIPCAL's alignment-based analysis indicates the presence, type and location of a putatively RIP-generated mutation within each copy of a repeat family. The input is accepted as Fasta or as both Fasta and GFF inputs. "Repeat_region" features in the GFF input were aligned by family via ClustalW (Additional file 4, Additional file 5). The prevalence of internal direct repeats within repeat families can result in poor alignment. Therefore the ClustalW default parameters have been adjusted for fast alignment, pairwise window length = 50 and k-tuple word-size = 2 to improve repeat family alignment. In some cases custom alignment parameters or manual alignment curation was used and is recommended. Sequence-only inputs are also accepted as pre-aligned Fasta files. It is assumed for sequence-only inputs that all sequences belong to the same family.

Aligned sequences are compared to a model sequence which can be either a sequence with highest total G:C content in the alignment, the alignment consensus or a user-defined sequence. The default model selection method is highest total G:C content. As RIP mutations deplete the G:C content, this default is assumed to select the least RIP-affected sequence as the model. RIPCAL also provides alternative methods of model selection, one of which is to define a majority consensus of the aligned sequences. The degenerate nucleotide code is used if two or more nucleotides are present in equal frequency (Additional file 3). The third option is for the model to be user-defined. This would be appropriate if the non-RIP-affected sequence was known, as in the case of experimentally transformed strains.

Following alignment and choice of model, the mutation frequencies are compared along the alignment for each sequence. Where the consensus sequence is degenerate, the probability of mutation at that location is added to the total count. The final output is a repeat family alignment and corresponding RIP frequency graph in GIF format. A summary of RIP mutation type versus total sequence divergence per sequence is also generated based on the alignment.

Validation of alignment-based RIP analysis

The alignment-based method was tested using the Tad1 transposon and 5S rDNA repeats from *Neurospora crassa* as positive and negative controls for detection of RIP mutation. These sequences [GenBank:L25662, GenBank:AF181821] were mapped to the *N. crassa* genome (release 7) [20] via RepeatMasker [28]. The genomic matches were compared via RIPCAL for RIP mutation. *Aspergillus nidulans* MATE transposon sequences [24] [GenBank:.BK001592, GenBank:.BK001593, GenBank:.BK0015924, GenBank:.BK001595, GenBank:.BK001596, GenBank:X78051] were compared via RIPCAL using MATE-9 [GenBank:.BK001592] as the model for comparison to test for detection of non-classical (non Cpa→TpA) RIP mutation. RIP mutation of Ty1 Copia-like transposons of *Mycrobotryum violaceum* [PopSet:55418573] was also analysed using the degenerate consensus model to observe RIP detection in sequences with a known tri-nucleotide mutation bias [13].

RIP Analysis of S. nodorum de novo *repeat families*

Results herein use data from a recent survey of the genome of *S. nodorum* [21] (Additional file 4, Additional file 5). Repeat family genomic coordinates can be found in the supplementary data (Additional file 4). Repetitive sequences were identified *de novo* via RepeatScout [29], and filtered for ≥ 200 bp length; ≥ 10 × genomic match coverage and ≥ 75% identity. *De novo* repeats were mapped to the *S. nodorum* genome via RepeatMasker [28]. A total of 26 repeat families were identified, corresponding to roughly 4.5% of the assembled genomic sequence. The repeat families were aligned via ClustalW (Additional file 5). Some repeat families were predicted to be telomeric, where ≥ 85% of genomic matches resided on scaffold termini relative to overall localisation. The tandem rDNA repeats were defined by location within the rDNA tandem array on scaffold 5 [GenBank:CH445329] from base pair position 1310974 to 1594765. rDNA repeats at other locations were divided into non-tandem (≥ 1 kb) and short-length (< 1 kb) sub-families. The predicted repeat type was assigned based on BLAST versus NCBI and REPBASE [30]. RIP mutation 'dominance' represents the preponderance of a particular type of RIP di-nucleotide mutation relative to all other alternative forms of RIP mutation. CpA↔TpA dominance as referred to in Table 2 was determined by:

$$\overline{\left(\frac{(CpA\leftrightarrow TpA)}{(CpC\leftrightarrow TpC)+(CpG\leftrightarrow TpG)+(CpT\leftrightarrow TpT)}\right)} \quad (8)$$

Other CpN↔TpN dominance equations (Additional file 2) were of a similar format to the one above (8).

Time of Operation

All data was generated on a 2.99 GHz Dual-core ×64 Intel PC with 2 GB RAM. The combined run-time of the di-nucleotide and alignment-based analyses for the *S. nodorum* whole genome assembly was approximately 4 hours. Pre-aligned inputs with few sequences (i.e. < 20) can be expected to complete under a minute.

Authors' contributions

JKH developed the RIPCAL software. JKH and RPO wrote the manuscript.

Additional material

Additional file 1

Control Data. *Compressed (.zip) file containing data relevant to control tests with* Neurospora crassa *Tad1 and 5S rDNA repeats, containing RIPCAL graphical (.png), tabular text (.txt) and fasta (.fas) alignments of repeat family matches. Also contains files (.png, .txt and .fas) for the MATE repeats from* Aspergillus nidulans *and Ty1 transposons from* Microbotryum violaceum.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-478-S1.zip]

Additional file 2

Supplementary Data Tables. *Excel (.xls) file with tabular data relating to RIPCAL analyses of* Stagonospora nodorum *repeats.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-478-S2.xls]

Additional file 3

Supplementary Methods. *Word (.doc) file explaining some of the methods in more detail.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-478-S3.doc]

Additional file 4

Snodorum Repeats. *GFF3 (.gff) file containing NCBI genomic accessions and coordinates of* Stagonospora nodorum *repeats used in this publication.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-478-S4.gff]

Additional file 5

Snodorum Alignments. *Compressed (.zip) file containing alignments for all* Stagonospora nodorum *repeat families in Fasta (.fas) format.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-9-478-S5.zip]

Acknowledgements

We thank Barbara Howlett and Thierry Rouxel for their expert advice and the Grains Research & Development Corporation (Barton, ACT, 2600, Australia) for providing the PhD scholarship to JKH.

References

1. Soanes DM, Richards TA, Talbot NJ: **Insights from sequencing fungal and oomycete genomes: what can we learn about plant disease and the evolution of pathogenicity?** *Plant Cell* 2007, **19(11):**3318-3326.
2. Selker EU: **Premeiotic instability of repeated sequences in *Neurospora crassa*.** *Annu Rev Genet* 1990, **24:**579-613.
3. Selker E, Cambareri E, Jensen B, Haack K: **Rearrangement of duplicated DNA in specialised cells of *Neurospora*.** *Cell* 1987, **51:**741-752.
4. Ikeda K, Nakayashiki H, Kataoka T, Tamba H, Hashimoto Y, Tosa Y, Mayama S: **Repeat-induced point mutation (RIP) in *Magnaporthe grisea*: implications for its sexual cycle in the natural field context.** *Mol Microbiol* 2002, **45(5):**1355-1364.
5. Nakayashiki H, Nishimoto N, Ikeda K, Tosa Y, Mayama S: **Degenerate MAGGY elements in a subgroup of *Pyricularia grisea*: a possible example of successful capture of a genetic invader by a fungal genome.** *Mol Gen Genet* 1999, **261(6):**958-966.
6. Graia F, Lespinet O, Rimbault B, Dequard-Chablat M, Coppin E, Picard M: **Genome quality control: RIP (repeat-induced point mutation) comes to *Podospora*.** *Mol Microbiol* 2001, **40(3):**586-595.
7. Idnurm A, Howlett BJ: **Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete *Leptosphaeria maculans*.** *Fungal Genet Biol* 2003, **39(1):**31-37.
8. Neuveglise C, Sarfati J, Latge JP, Paris S: **Afut1, a retrotransposon-like element from *Aspergillus fumigatus*.** *Nucleic Acids Res* 1996, **24(8):**1428-1434.
9. Hua-Van A, Hericourt F, Capy P, Daboussi MJ, Langin T: **Three highly divergent subfamilies of the impala transposable element coexist in the genome of the fungus *Fusarium oxysporum*.** *Mol Gen Genet* 1998, **259(4):**354-362.
10. Hua-Van A, Langin T, Daboussi MJ: **Evolutionary history of the impala transposon in *Fusarium oxysporum*.** *Mol Biol Evol* 2001, **18(10):**1959-1969.
11. Julien J, Poirier-Hamon S, Brygoo Y: **Foret1, a reverse transcriptase-like sequence in the filamentous fungus *Fusarium oxysporum*.** *Nucleic Acids Res* 1992, **20(15):**3933-3937.
12. Nielsen ML, Hermansen TD, Aleksenko A: **A family of DNA repeats in *Aspergillus nidulans* has assimilated degenerated retrotransposons.** *Mol Genet Genomics* 2001, **265(5):**883-887.
13. Hood ME, Katawczik M, Giraud T: **Repeat-induced point mutation and the population structure of transposable elements in *Microbotryum violaceum*.** *Genetics* 2005, **170(3):**1081-1089.
14. Farman ML: **Telomeres in the rice blast fungus *Magnaporthe oryzae*: the world of the end as we know it.** *FEMS microbiology letters* 2007, **273(2):**125-132.

15. Braumann I, Berg M van den, Kempken F: **Repeat induced point mutation in two asexual fungi, *Aspergillus niger* and *Penicillium chrysogenum*.** *Curr Genet* 2008.
16. Cambareri E, Singer M, Selker E: **Recurrence of repeat-induced point mutation (RIP) in *Neurospora crassa*.** *Genetics* 1991, **1:**.
17. Watters MK, Randall TA, Margolin BS, Selker EU, Stadler DR: **Action of repeat-induced point mutation on both strands of a duplex and on tandem duplications of various sizes in *Neurospora*.** *Genetics* 1999, **153(2):**705-714.
18. Cambareri E, Jensen B, Schabtach E, Selker E: **Repeat-induced G-C to A-T mutations in *Neurospora*.** *Science* 1989, **244:**1571-1575.
19. Margolin BS, Garrett-Engele PW, Stevens JN, Fritz DY, Garrett-Engele C, Metzenberg RL, Selker EU: **A methylated *Neurospora* 5S rRNA pseudogene contains a transposable element inactivated by repeat-induced point mutation.** *Genetics* 1998, **149(4):**1787-1797.
20. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, *et al.*: **The genome sequence of the filamentous fungus *Neurospora crassa*.** *Nature* 2003, **422(6934):**859-868.
21. Hane JK, Lowe RGT, Solomon PS, Tan KC, Schoch CL, Spatafora JW, Crous PW, Kodira C, Birren BW, Galagan JE, *et al.*: **Dothideomycete-plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*.** *Plant Cell* 2007, **19(11):**3347-3368.
22. Solomon PS, Lowe RGT, Tan KC, Waters ODC, Oliver RP: ***Stagonospora nodorum*: cause of *Stagonospora nodorum* blotch of wheat.** *Mol Plant Pathol* 2006, **7(3):**147-156.
23. Cambareri EB, Helber J, Kinsey JA: **Tad1-1, an active LINE-like element of *Neurospora crassa*.** *Mol Gen Genet* 1994, **242(6):**658-665.
24. Clutterbuck AJ: **MATE transposable elements in *Aspergillus nidulans*: evidence of repeat-induced point mutation.** *Fungal Genet Biol* 2004, **41(3):**308-316.
25. Perkins DD, Metzenberg RL, Raju NB, Selker EU, Barry EG: **Reversal of a *Neurospora* translocation by crossing over involving displaced rDNA, and methylation of the rDNA segments that result from recombination.** *Genetics* 1986, **114(3):**791-817.
26. Larkin MA, Blackshields G, Brown NP, Chenna R, McGettigan PA, McWilliam H, Valentin F, Wallace IM, Wilm A, Lopez R, *et al.*: **Clustal W and Clustal X version 2.0.** *Bioinformatics* 2007, **23(21):**2947-2948.
27. Selker EU, Tountas NA, Cross SH, Margolin BS, Murphy JG, Bird AP, Freitag M: **The methylated component of the *Neurospora crassa* genome.** *Nature* 2003, **422(6934):**893-897.
28. **RepeatMasker Open-3.0** [http://www.repeatmasker.org]
29. Price AL, Jones NC, Pevzner PA: ***De novo* identification of repeat families in large genomes.** *Bioinformatics* 2005, **21(Suppl 1):**i351-358.
30. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenetic and genome research* 2005, **110(1–4):**462-467.

Appendix 3A: RIPCAL/deRIP Manual

Version 1.0 Last updated: 1st June 2011

1.1 FEEDBACK

Spotted a bug? Report it. If you have comments or questions regarding RIPCAL:

- E-mail jameshane@users.sourceforge.net
- or leave a post on the RIPCAL forums at http://sourceforge.net/projects/ripcal/support

1.2 WHAT IS RIPCAL?

RIPCAL is a software tool for the bioinformatic analysis of repeat-induced point mutation (RIP) in fungal genome sequences. For comprehensive reviews of RIP and RIP-related analyses in fungi refer to:

- Clutterbuck AJ (2011) Genomic evidence of repeat-induced point mutation (RIP) in filamentous ascomycetes. Fungal Genetics & Biology 48(3):306-26.
- Galagan JE & Selker EU (2004) RIP: the evolutionary cost of genome defense. Trends in Genetics 20(9):41723.
- Hane JK, Williams AH & Oliver RP (2011) Genomic and Comparative Analysis of the Class Dothideomycetes. The Mycota vol. 14, chapter 9.

RIPCAL performs a range of RIP-based calculations in a simple and user friendly manner, but also has additional capabilities for advanced users. RIPCAL can be applied to a series of repeat families within a whole genome, a single repeat family or a single sequence.

RIPCAL can answer one or more of the following questions:

- Is my genome/sequence likely to be RIP mutated?
- Which repeat classes are more/less RIP mutated in my genome?
- Which form of CpN->TpN mutation is the dominant RIP-like di-nucleotide mutation?
- Is there a bias for the RIP-mutation of certain repeat elements in my genome/sequences?
- Is there a locational bias for RIP mutation within individual repeats or within a repeat family as a whole?
- Is my genome/sequence more RIP mutated than another genome/sequence?

1.3 WHAT IS deRIP?

The RIPCAL suite is packaged together with its sister tool called deRIP. DeRIP detects RIP within a repeat family multiple alignment in a similar way to RIPCAL and then predicts the most probable consensus sequence for what the repeat family would have looked like prior to RIP mutation. As RIP inactivates sequence function by mutating them to the point where they are often unrecognisable by sequence similarity searches, deRIP can provide insight into the nature and origin of some repeats.

1.4 HOW TO CITE RIPCAL/deRIP

To cite RIPCAL and deRIP and for examples of how these tools can be applied to your research, see the following publications:

- Hane & Oliver (2008) RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. BMC Bioinformatics 9:478
- Hane & 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

1.5 CITATIONS

To date, RIPCAL has been applied in the following studies:

- Rouxel et al. (2011) Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. Nature Communications 2:202
- Di Guistini et al. (2011) Genome and transcriptome analyses of the mountain pine beetle-fungal symbiont *Grosmannia clavigera*, a lodgepole pine pathogen. PNAS 108(6):2504-2509
- Gao et al. (2011) Genome Sequencing and Comparative Transcriptomics of the Model Entomopathogenic Fungi *Metarhizium anisopliae* and *M. acridum*. PLoS Genetics 7(1): e1001264.
- Van de Wouw 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 Pathogens 6(11): e1001180.
    - **Note:** this study employed a variation on the standard RIPCAL analyis which is currently unavailable but may be included in a future update.

1.6 VERSION HISTORY

1.6.1.1 Version 1.0.5→2.0

- added deripcal perl script, which predicts ancient pre-RIP alignment consensus sequence from fasta or clustalw format alignment input

- added ripcal_summarise perl script, which give basic RIP summary statistics from an input derived from a RIPCAL table output

1.6.1.2 Version 1.0.4→1.0.5

- fixed errors causing user defined, fasta/clustal only inputs to fail to run
- added numbering of sequence id's in alignment based graphical outputs
- fixed bug introduced in v1.0.4 where alignment order either from pre-aligned or local clustalw generated inputs is lost in final output

1.7 INSTALLATION

1.7.1 Requirements

RIPCAL can run in various form on both Windows and Linux. Running the compiled RIPCAL executable in windows should have no requirements. Linux users can run the RIPCAL perl script. Running deRIP and other miscellaneous RIPCAL scripts also requires perl. This should already be installed if you are a Linux user. If you are a windows user you can install ActivePerl here (http://www.activestate.com/activeperl). Running alignments locally (during RIPCAL analysis on your PC as opposed to prior to RIPCAL on an online server) also requires the installation of ClustalW (http://www.ebi.ac.uk/Tools/clustalw2/index.html).

1.7.2 Instructions

- extract zip contents to new folder
- if running a WINDOWS OS, double click the exe.
- if running LINUX/MAC OS, set ripcal_x_x_x.pl to executable and run.
- if running LINUX/MAC OS, depending on your shell commands you may have to modify the lines of perl code where 'system()' is used. You should change this to whatever is appropriate to run clustalw on your shell. If your executable is called 'clustalw2' instead of 'clustalw' you will either have to edit 'clustalw' to 'clustalw2' in the perl script or create a symbolic link 'clustalw' which points to the location of 'clustalw2'.

2 RIPCAL

2.1 Analysis modes

There are 3 modes available:

- Alignment-based
- Di-nucleotide frequency analysis
- RIP-index scan

2.1.1 Alignment-based

Calculates RIP mutation frequencies based on a multiple alignment of input sequences. Input may be prealigned (fasta/clustalw format) or not aligned (fasta or fasta + gff formats, this requires local installation of clustalw).

2.1.1.1 RIP model sequences

Alignment-based analyses quantify RIP-like mutations in a multiple alignment by comparing each aligned sequence to a „model‟ sequence. There are 3 methods used for the selection of model (comparison) sequence for alignment based analyses. The choice of model

2.1.1.1.1 Highest G and C content

The sequence with the highest total G+C content is selected as the model on the basis that RIP mutation depletes G+C content, therefore highest G+C content should indicate the least RIP affected sequence.

Note: This may not be the appropriate method of model selection if the aligned sequences are of variable lengths as a longer sequence is more likely to be chosen than shorter one.

2.1.1.1.2 Majority consensus

This method determines the most common base at each position of the alignment (where sequence number > 2). Degenerate base letters are used where 2 or more base counts are equal. The degenerate consensus method assigns degenerate bases W, S, M K, R, Y, B, D, H, V or N:

Degenerate base letter	Corresponds to
W	A/T
S	G/C
M	A/C
K	G/T
R	G/A
Y	T/C
B	G/T/C OR not A
D	A/G/T OR not C
H	A/C/T OR not G
V	G/C/A OR not T
N	A/C/G/T

“N” is used in the degenerate consensus to refer to any base pair combination but is not assigned a probability of RIP mutation when calculating RIP mutation from a degenerate consensus.

Because each sequence in the alignment is now compared to an ambiguous consensus, in this mode RIPCAL converts absolute mutation counts to „probabilities of mutation‟. The table below outlines the probability of nucleotide identity for each degenerate base letter:

1/1	**1/2**	**1/3**
A	M/R/W	D/H/V
C	M/S/Y	B/H/V
G	K/R/S	B/D/V
T	K/W/Y	B/D/H

RIP probability at a particular position along an alignment is determined by the table above. i.e. for consensus dinucleotide MpD mutating to TpA in aligned sequences, there is a (1/2*1/3=1/6) chance that this is a CpA→TpA mutation.

In some cases this may be the most appropriate method to use, as it can detect RIP mutation among a repeat family of diverse sequence and RIP mutation profiles, as opposed to the highest G+C method, which chooses the sequence in the family most likely to be the least RIP affected. If the highest G+C sequence appears to be an anomaly compared to the majority of sequences, this method is a good choice.

2.1.1.1.3 User-defined

The choice of model sequence is left to the user. See sections 2.2.1.1 and 2.2.1.2 for details on how to define RIP models. If no model is defined for a repeat family RIPCAL defaults to GC mode.

2.1.1.2 RIP dominance

This is a measure of the pre-dominance of a particular type of RIP-like CpN→TpN mutation over another. RIP dominance can be calculated individually for each sequence (section 2.2.2.2) or for all sequences in a repeat family (section 2.3.2.3) or in one or two directions (section 2.1.1.3).

CpA↔TpA dominance is calculated by

$$\left(\frac{(CpA \leftrightarrow TpA)}{(CpC \leftrightarrow TpC)+(CpG \leftrightarrow TpG)+(CpT \leftrightarrow TpT)}\right)$$

CpC↔TpC dominance is calculated by

$$\left(\frac{(CpC \leftrightarrow TpC)}{(CpA \leftrightarrow TpA)+(CpG \leftrightarrow TpG)+(CpT \leftrightarrow TpT)}\right)$$

CpG↔TpG dominance is calculated by

$$\left(\frac{(CpG \leftrightarrow TpG)}{(CpA \leftrightarrow TpA)+(CpC \leftrightarrow TpC)+(CpT \leftrightarrow TpT)}\right)$$

CpT↔TpT dominance is calculated by

$$\left(\frac{(CpT \leftrightarrow TpT)}{(CpA \leftrightarrow TpA)+(CpC \leftrightarrow TpC)+(CpG \leftrightarrow TpG)}\right)$$

2.1.1.3 Direction of mutation

- **Why is there a bidirectional arrow between mutated dinucleotides in the RIP dominance statistics show above?** When comparing repeat families it is not known which sequences are truly less/more RIP affected. In fact, all repeats within a genome are likely to interact with each other and become affected by RIP. Therefore CpN mutations in both directions are counted. Counting RIP mutations in a single direction would be appropriate if the repeat alignment contained a known precursor repeat (e.g. from a different organism) as the comparison model.

Notes:

- Bi-directional RIP counts are used in the graphical outputs.
- Both bi-directional and directional RIP mutation counts of alignments are provided in the RIPCAL tabular output.

2.1.2 Di-nucleotide frequency analysis

Calculates the relative frequencies of dinucleotides (pairs of adjacent nucleotides) for input sequences (or subsequences if fasta+gff selected).

2.1.3 RIP-index scan

- Predicts RIP-mutated regions within input sequence based on high scoring RIP-indices. The RIPCAL GUI uses default index thresholds only. Thresholds can be adjusted by running on the command line. Currently outputs GFF coordinates sequence regions with high or low scoring RIP indices.

- RIPCAL breaks down long sequences into smaller chunks (default chunk size 200bp). If these small sub-regions are above the threshold for RIP for the selected RIP indices, these chunks are stored in memory. Overlapping RIP-affected chunks are merged into longer regions, which are subject to a minimum size threshold of 300bp (default).

- Default index thresholds for TpA/ApT and (CpA+TpG)/(ApC+GpT) (see section 2.3.4.2) are based on values determined experimentally in *Neurospora crassa* by Margolin et al. (A methylated *Neurospora* 5S rRNA pseudogene contains a transposable element inactivated by repeat-induced point mutation. Genetics 149, 1787-1797 (1998)).

Additional (un-published) indices are also suggested as additions, but not replacements to the previously published RIP indices TpA/ApT and (CpA+TpG)/(ApC+GpT). The rationale behind these new RIP indices is that the ratio of pre-RIP to post-RIP dinucleotides should be low in RIP mutated sequences. By comparison of these four indices it is also possible to detect non-conventional (ie not CpA→TpA) dinucleotide bias (which the previous indices are designed to target).

2.2 File formats

2.2.1 Input

There are 2 types of input options which are context sensitive depending on analysis type (**FASTA + GFF** or **FASTA only**). These options are used only in "alignment-based" and "di-nucleotide frequency" modes (sections 2.1.1 and 2.1.2)

2.2.1.1 FASTA

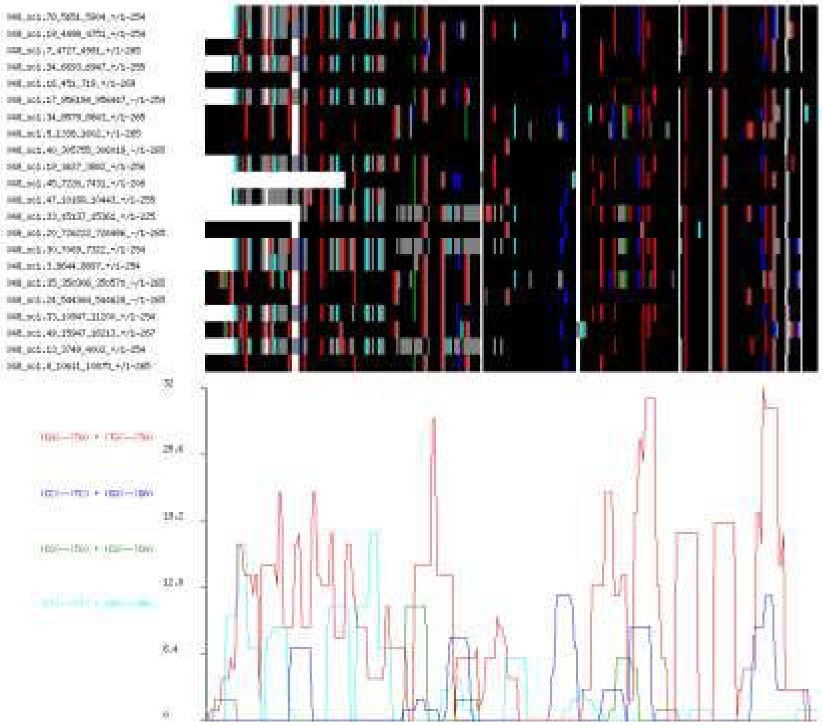
- RIPCAL accepts standard FASTA/Pearson sequence format (http://www.ncbi.nlm.nih.gov/BLAST/fasta.shtml) or FASTA multiple alignment format (http://www.bioperl.org/wiki/FASTA_multiple_alignment_format). There can be multiple sequence records contained in a single fasta file.
- Fasta files can be of any extension as long as they are in the correct format, however GUI mode browses for *.fa, *.fas and *.fasta by default.
- For alignment based and dinucleotide frequency analyses, if a GFF input is not provided, all sequences from the FASTA file are grouped together as members of the same repeat family.
- In alignment-based mode with no accompanying GFF input, the first sequence in a multi-fasta file is used as the model for RIP comparison if the RIP model is set as "user-defined" (refer to section 2.1.1.1.3). RIPCAL will also guess whether fasta inputs are pre-aligned by searching for gap '-' characters in the sequences. If your input is pre-aligned but happens to contain NO GAPS, add a '-' to the end of one of your sequences. This will not affect the RIP calculation.

2.2.1.2 GFF

- The GFF format is described here (GFF3: http://www.sequenceontology.org/gff3.shtml). GFF input is used only when more than one repeat family is being analysed. This option is not available for index scan analysis.
- Only 'repeat_region' type features are considered by RIPCAL. Features are grouped according to repeat family, which is indicated by the "target" attribute i.e. "Target=repeatfamilyID X Y;", although coordinates X and Y in the target attribute are ignored.
- For alignment-based mode, the inclusion of the attribute "note=model" defines a sequence as the model for RIP comparison (refer to section 2.1.1.1.3). It is advisable to check that there is only one defined model per repeat family.

2.2.2 Output

There are several types of output which are dependent on analysis type:



2.2.2.1 GIF

This output is created if alignment-based analysis is selected (section 2.1.1).

The alignment diagram is a visual representation of the alignment file received from ClustalW/pre-aligned input. Sequences appear in identical order to that of the alignment input. Usually this means that sequences are grouped according to similarity (default ClustalW alignment ordering) and this overall sequence similarity usually corresponds to similar RIP profiles.

If reference sequence selection methods are by G:C content or user-defined, one of these sequences will be represented in black and white only – i.e. indicating no sequence variation from itself (the reference). This will not be the case if the method chosen is degenerate consensus as the model sequence does not exist within the set of aligned sequences.

The y-axis of the plot at the bottom of the output represents the overall frequency of RIP mutations (type indicated by colour) along a scanning window at each position of the alignment. This means that at alignment position „x", the total RIP mutation counts are determined from all sequences in the alignment from position (x-24) to position (x+25). Window-size defaults to 1/50th of the alignment length, to a minimum of 10 bp, but can be adjusted in command-line mode. This can show the localised effects of RIP changes in discrete sequence regions.

2.2.2.2 *_RIPALIGN.TXT

This output is created if alignment-based analysis is selected (section 2.1.1). This is a tabular summary of the data presented in the *.GIF graphical alignment-based output. The tabular data provides more in depth information than can be presented in graphical form, including the polarity of RIP mutation.

2.2.2.3 *_DINUC.TXT

This output is created if dinucleotide frequency analysis is selected (section 2.1.2). This creates a tabular dinucleotide frequency table for individual sequences (fasta only) or repeat families (fasta+gff).

2.2.2.4 *_SCAN.TXT

This output is created in RIP-index scan mode (section 2.1.3). The contents of this file will be GFF format. This groups high and low scoring (by RIP index) regions as GFF features.

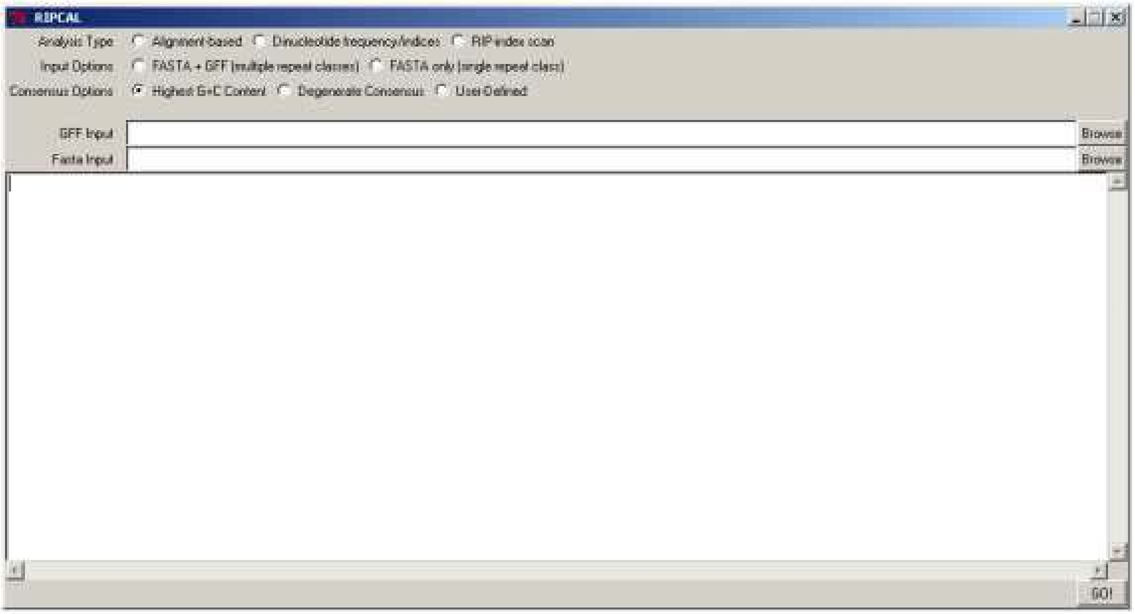
2.3 USAGE

RIPCAL may be run either as a graphical user interface (GUI) or from the command line. To access the GUI run ripcal as an executable or from the command-line with no arguments. Use the '-h' or '--help' arguments to see the full list of available command-line arguments.

Note: the command-line and GUI versions of RIPCAL are almost identitical in their capabilities, however RIP index scan thresholds can only be adjusted in command-line mode.

2.3.1 General

2.3.1.1 GUI



Analysis Type ○ Alignment-based ○ Dinucleotide frequency/indices ○ RIP index scan
Input Options ○ FASTA + GFF (multiple repeat classes) ○ FASTA only (single repeat class)
Consensus Options ● Highest G+C Content ○ Degenerate Consensus ○ User-Defined

GFF Input [Browse]
Fasta Input [Browse]

- run the exe (windows) or "perl ripcal" through the command-line (*nix/mac)
- select one of the 3 available analysis types (alignment-based, dinucleotide or index scan)
- select FASTA+GFF input or FASTA only (GFF input field will become disabled)
- select consensus calculation option (highest G and C content, degenerate consensus or user-defined)

Argument	Description	Default
--help OR -h	RIPCAL options help (lists these command-line arguments)	
--command OR -c	use command line interface	
--type OR -t	RIP analysis type [align OR index OR scan]	align

2.3.1.2 Command-line options (perl ripcal --command)

- If -fasta and -gff selected, the input is assumed to contain multiple repeat families
- If -fasta only the input is assumed to contain a single repeat family.
- If -m=user models are interpreted as the first sequence in the alignment for single family inputs
- single repeat family, aligned RIP analysis also accepts prealigned input in FASTA OR CLUSTALW format

2.3.2 Alignment-based mode

2.3.2.1 GUI

- Select "Alignment-based"
- Select "FASTA+GFF" for whole-genome analysis or "FASTA only" for single repeat family analysis
- Select Consensus options

- Browse for FASTA and/or GFF input files
- Click "GO!"

2.3.2.2 Command-line options (perl ripcal --command --type align)

Argument	Description	Default
--seq OR -s	input sequence file [fasta or clustalw format] (**REQUIRED**)	
--gff OR -g	input gff file [gff3 format: http://www.sequenceontology.org/gff3.shtml]	
--model OR -m	Alignment model [gc OR consensus OR user]	gc
--windowsize OR -z	RIP frequency graph window (align-mode only)	alignment length/50, minimum of 10 bp

2.3.2.3 Summarising the results

This is a simple script for summarising the results of the *_RIPALIGN.TXT files (section 2.2.2.2) which contain detailed comparisons of RIP-like mutations between ALL SEQUENCES and the RIP model, into a simplified tabular format which summarises the WHOLE repeat family.

Run through the command line (on first use): perl ripcal_summarise inputfile > outputfile or use the following for subsequent uses on different input files (to append summary to single output file) perl ripcal_summarise inputfile >> outputfile

2.3.3 Di-nucleotide mode

2.3.3.1 GUI

- Select "Dinucleotide frequency/indices"
- Select "FASTA+GFF" for whole-genome analysis or "FASTA only" for single repeat family analysis
- Browse for FASTA and/or GFF input files
- Click "GO!"

2.3.3.2 Command-line options (perl ripcal --command --type index)

Argument	Description	Default
--seq OR -s	input sequence file [fasta or clustalw format] (**REQUIRED**)	
--gff OR -g	input gff file [gff3 format: http://www.sequenceontology.org/gff3.shtml]	

2.3.4 Index scan mode

2.3.4.1 GUI

- Select "RIP-index scan"
- Browse for FASTA input file
- Click "GO!"

2.3.4.2 Command-line options (perl ripcal --command --type scan)

In RIP index scanning mode, RIPCAL calculates RIP indices within a defined window-size. RIPCAL then repeats this calculation for neighbouring windows, moving along the sequence by a distance set according to the increment parameter.

Argument	Description	Default
--seq OR -s	input sequence file [fasta or clustalw format] (**REQUIRED**)	
-l	Length of scanning subsequence (window size) (bp)	300
-i	Scanning subsequence increment (bp)	150
-q	TpA/ApT threshold (>=)	0.89
-w	CpA+TpG/ApC+GpT threshold (<=)	1.03
-e	CpA+TpG/TpA threshold (<=)	-
-r	CpC+GpG/TpC+GpA threshold (<=)	-
-y	CpG/TpG+CpA threshold (<=)	-
-u	CpT+ApG/TpT+ApA threshold (<=)	-

modifying RIP-index scanning parameters only available in command-line mode
disable index thresholds by setting to 0

3 deRIP

Building upon the RIPCAL procedure, deRIP is a technique to reverse the effects of RIP mutation *in silico*. The deRIP process involves scanning a multiple alignment of a repeat family for RIP-like polymorphism and reverting the alignment consensus to the putative pre-RIP-mutated sequence. The resultant "deRIPped" sequence is a prediction of what a RIP-mutated repeat DNA may have looked like prior to RIP mutation.

3.1 USAGE

3.1.1 Running deRIP

deRIP can only be run through the command-line:

```
perl deripcal <format> <inputfile> <outputfile>
```

Argument	Description	Default
<format>	"fasta" or "clustalw"	-
<inputfile>	Multiple-alignment file in fasta (i.e. with "-" characters) or clustalw format	-
<outputfile>	The multiple alignment in aligned fasta format with the addition of a new first sequence which is the deRIP consensus. Output file name also used as a prefix for the following 2 outputs: <outputfile>.consensus <outputfile>.deripcons Which are the sequences of the majority and derip consensus respectively. **Note:** These will still contain gap characters to allow cutting/pasting into alignments for comparison if desired. Run a simple text replace for gap characters in your text-editor of choice to obtain	-

deripcal will insert a gap in the consensus sequences if there is only 1 X sequence coverage at a particular part of the alignment. This is sensible for most repeat consensus sequences as there are sometimes rare insertions in some copies. Depending on the repeat family this value could even be raised. RIP cannot be calculated by alignment for a single copy sequence so 2 was set as the minimum coverage threshold.
To change this behaviour edit line 149 of the deripcal script (change "2" to appropriate threshold value):

```
if (($consensusgapcount > $maxcount)&&($maxcount < 2)) {
```

3.2 CAVEATS

Bioinformatic analysis of repetitive sequences can be problematic and current tools are not yet perfect. Below are a few general recommendations for analysis and processing of repeat data prior to RIPCAL analysis:

RIP Indices are ratios of dinucleotide frequencies in a sequence. These are rough indicators of RIP mutation, however dinucleotide frequencies are affected by factors other than RIP and their reliability should be treated with caution (Hane & Oliver 2008). Nevertheless RIP-indices are still useful when there is little repetitive sequence available for analysis
(i.e. a single copy sequence of a repeat), or for scanning a genome for potential regions of "ancient" RIP -which due to RIP, other mutations, deletions or lateral transfer may no longer be similar enough to its former sister repeats to be recognised as coherent gene family.

When using *de novo* repeat-finding tools (such as RepeatScout, RECON, REPET, etc.), check your consensus repeat families for redundancy. Sometimes there are repeats which are a slightly different version of a sub-region of a larger repeat. These types of repeat families should be combined. When analysing RIP, it is good practice to consider all repeats that could potentially interact with one another.

When using RepeatMasker (or a similar repeat-mapping method) don't rely on its ability to join HSPs (high-scoring pairs -basically short regions of aligned sequence that usually don't represent complete sequences) together into whole repeat annotations. Inspect the processed and unprocessed outputs carefully and assess whether you can trust the processed result. RIPCAL analysis will be more effective if you are able to join up the disparate repeat HSPs in the unprocessed output, but less so if the joining is inaccurate.

When aligning repeat sequences using multiple alignment tools always manually inspect the alignments. Jalview (http://www.jalview.org/) is useful for this.

You should also keep the following considerations in mind when setting
alignment parameters:

- Some repeats can have large insertions/deletions
  - o Lower the gap extension penalty accordingly.
- Some repeats have short internal repeats and low complexity sequences.
  - o Set window size (or equivalent parameters) as large as possible to control for mis-alignment.
- RIP involves C→T mutations (or G→A if reverse complemented).
  - o If your alignment tool allows the use of custom matrices you should allow for a slightly stronger match score for matches between C and T and between G and A.

Some large/complex repeats refuse to align properly. In these cases, choose the best automatically-generated alignment and manually curate the alignment with CINEMA or similar alignment-editing tool (http://utopia.cs.man.ac.uk/utopia/cinema).

Appendix 3B: Response to Thesis Examination Comments

The candidate suggests using two types of alignments as input to the RIPCAL procedure: a consensus repeat, or the most GC-rich copy of a repeat that is available in the genome under study. There are problems with both of these approaches:

1) ***a majority consensus sequence based on alignment of RIP'd repeats may end up containing RIP'd bases in several positions***
2) ***unless most GC rich copy of repeat has escaped RIP entirely, it will not be able to provide information about nucleotide positions where it, itself, is RIP'd***

thus the efficacy of this approach is highly dependent on having elements that escape RIP.

The RIP calculations performed by RIPCAL are dependent upon prior identification and alignment between individual repeats belonging to a repeat family. The alignments used by RIPCAL are multiple-alignments of the whole repeat family. As the examiner points out, RIP can only be detected by RIPCAL if at least one of the sequences in the multiple alignment has escaped RIP. However, this is a limitation of any *in silico* alignment-based RIP calculation method, including the examiner's suggested alternative (see below). Calculation of RIP without the presence of RIP-free regions is only possible using RIP indices, which have been demonstrated to be inaccurate in Chapter 3 due to background noise caused by G:C content variation.

RIPCAL has three 'model' selection options, the sequence of highest G+C content, the consensus of the multiple alignment or manually selected sequence. The 'model' sequence is a point of reference to which all other aligned sequences are compared, but the model does not actually have to be the least RIP-affected sequence in order to measure RIP in that repeat family. At any given position along the repeat-family multiple alignment, as long as there is a RIP-like polymorphism in one of the sequences it will be detected as RIP. RIPCAL can calculate RIP both directionally, i.e. CpN → TpN (from model to other repeat) or TpN to CpN (from other repeat to model, or bi-directionally i.e. CpN ↔ TpN. By default, RIP dominance and other metrics are reported in bi-directional terms. RIPCAL is only intended to be used directionally in cases where a sequence (typically a user-defined model) is known to be unaffected by RIP. By calculating RIP bi-directionally over a whole alignment, rather than in a single direction relative to a model sequence, RIPCAL is able to determine all sites of RIP-polymorphism across a whole repeat class.

A better approach is to blast the genome against itself, examine every pairwise alignment and keep a tally of RIP (and nonRIP) mutations at each nucleotide position in the query sequence. Assigning RIP sites to specific repeat residues is easily accomplished by mapping the genome coordinates of the RIP'd sites back to individual repeats.

The examiner suggests that by aligning each repeat copy to all other copies, one is able to catch every instance of RIP, and implies that RIPCAL is not able to do this. This criticism may stem from a misconception about RIPCAL relying on directional detection of RIP-like polymorphism relative to a 'model' sequence (see above). RIPCAL calculates RIP bi-directionally within a whole alignment by default, not just in a single direction from the selected model sequence, thus detecting all sites of RIP-polymorphism present within the multiple alignment. To accurately detect RIP, RIPCAL does however depend upon accurate *de novo* prediction of repeats, a process separate to and preceding RIPCAL analysis.

The examiner may have suggested this alternative method out of concern that RIPCAL would not detect RIP sites within multiple alignments of large repeat families if they contained internal sub-repeats (e.g. terminal inverted repeats). The examiner's suggested local-alignment method should be able to detect all pair-combinations between sub-repeats, whereas global alignment based methods such as RIPCAL would only detect RIP between a single pair-combination. Nevertheless, local-alignment analysis of RIP would appear to be unnecessary. Longer, i.e. more complete copies of repeats, reduce or prevent the effects of RIP upon smaller sub-repeats (Bhat & Kasbekar 2001, Fehmer et al. 2001, Vyas et al. 2006, Singh & Kasbekar 2008). Therefore sub-repeats need only be considered in their longest pair-combination, which is most likely to be represented by a global alignment.

Whether or not an alignment-based approach provides accurate insights into genome-wide patterns of RIP depends heavily on how the analyses are conducted. Unfortunately there was insufficient information provided in the RIPCAL paper, in the supplemental information, or in the relevant thesis chapter, for me to be able to evaluate the validity of the findings. I was not able to identify relevant subroutines within the script but this could be due to the fact that it wasn't annotated very well.

The RIPCAL publication and supplementary data did not omit any information pertaining to how the analyses were performed. Regrettably the format of a peer-reviewed journal, in which the primary goal is to present high-impact research data, did not lend itself to describing the finer points and pitfalls of the algorithm. This was addressed in the manual (see Appendix 3A), but its release was delayed by work on other publications during the PhD candidature. As such the examiner may have missed this.

I wholeheartedly agree with the examiner on the need for high quality prediction and alignment of repeat families prior to running RIPCAL. A significant portion of the "caveats" section of the RIPCAL manual (Appendix 3A) was devoted to discussion of this issue. Regarding the identification of repeat families, the RIPCAL publication cites the preceding publication (Chapter 2) in which the repetitive regions of the *P. nodorum* genome assembly were identified. RIPCAL can either be used to launch clustalw to align repeats, or can calculate RIP on a pre-aligned input file. The RIPCAL publication states "RIPCAL can interface with a local installation of ClustalW. Refer to Additional File 3 for more detailed information". Within Additional File 3 the ClustalW parameters used for repeat-alignment are described.

As for the program itself, the code within the script "ripcal" (contained within the download package for RIPCAL version 2.0) contains annotation "########## ALIGN ##########" on line 703 indicating that this is the section pertaining to alignment-based methods. Within this section (delimited by section annotations from lines 703 to 1839) contains a sub-routine called aligncount on line 940). Although not heavily annotated, this section contains perl functions using basic arithmetic with variable names such as "$ca2ta" to indicate counts of mutations from CpA to TpA and so on. On the other hand, going by the majority of open-source bioinformatics code I have come across so far, a certain level of bewilderment is only to be expected when encountering code that you haven't written yourself.

Specific deficiencies include the following: In calculating the relative frequencies of CpN to TpN di-nucleotide mutations, there is no mention of how adjacent RIP sites are treated. For example, with a CC<-->TT mutations, one cannot determine the context of the 5' mutation because the order in which the mutations occured is unknown. The same is true for CG <-> TA (i.e. RIP when has occured on both strands). Does RIPCAL take these situations into account and ignore di-nucleotides where the 3' position is unknown? If not, then the dominance of CpA (or CpT in some cases) could be an artifact of the data analysis procedures.

The examiner makes a series of criticisms which I believe can be broken down into three separate issues: 1) how RIPCAL treats the directionality of RIP-mutation in general, 2) how RIPCAL treats mutation of nucleotides of ambiguous sequence and 3) whether RIPCAL corrects for certain RIP-mutations which are ambiguous (i.e. the target of one type of RIP-mutation is the product of another).

1) I am in agreement with the examiner that it is impossible without prior knowledge (e.g. the sequence of a known active transposon) to determine the context (i.e. directionality) of RIP-like polymorphism. For this reason, as mentioned above, RIPCAL identifies RIP bi-directionally within a whole multiple alignment by default. That is, at any given position along a multiple alignment, the direction of a RIP mutation relative to the model sequence is irrelevant. What is relevant is that a RIP-like polymorphism was detected at that alignment position.
2) Where the 3' or 5' nucleotide in any di-nucleotide sequence is ambiguous, i.e. a C to T mutation adjacent in one or both to an unknown nucleotide ("N") or one using the IUPAC degenerate nucleotide code, RIPCAL converts the count to a probability. For example: CpA aligned with TpA will add 1 to the CpA to TpA running total, CpA aligned with YpA (Y = C or T) will add 0.5 to the CpA to TpA running total. The behaviour for all degenerate nucleotides is described in detail in the RIPCAL manual (Appendix 3X).
   Note: RIPCAL does not recognise gap ("-") characters, which are the product of gaps in the multiple alignment, as ambiguous bases for RIP calculation purposes.
3) The examiner refers to CpG to TpA which would be a product of CpG to Tpg followed by TpG to TpA. CpC to TpT would occur if CpCpN was mutated to TpCpN followed by mutation to TpCpN mutation to TpTpN. RIPCAL does currently not take these situations into account. As these types of CpN mutations are generally of lower frequency in repeat famillies that have been experimentally confirmed to be RIP-mutated in *P. nodorum*, this would not be an issue. Indeed RIP-related CpG and CpC mutations are of relatively low abundance in most species analysed thus far (Clutterbuck 2011), the overwhelming bias being towards CpA mutation.

There is no real value to the RIP dominance scores.

The RIP dominance scores have more relevance to detectable RIP mutations than the previously used RIP indices. Rather than being a ratio of nucleotide abundances, the calculation of RIP dominance scores involves counts of RIP-like polymorphisms that have been detected via alignment. The RIP dominance score measures the ratio (X divided by Y) of the frequency of a single CpN mutation (X = CpA mutation) versus the other three alternative CpN mutations (Y = CpC, CpG and CpT mutations) (X + Y = total pool of CpN mutation). Dominance functions were given for single CpN mutations because that it has been overwhelming observed in most fungal species that the preponderance of a single type of CpN mutation is far more frequent than the other three alternatives. Nevertheless, RIPCAL does more than just report RIP-dominances. It is at its heart a mutation counting tool, and its tabular outputs contain a high level of detail regarding polymorphism frequencies. It is a simple matter for a researcher to import this output into a spreadsheet and develop custom metrics, should the need arise.

Once a CpN bias has been postulated as associated with RIP, either via dominance values or other means, the dominance value specific to that bias can indicate the strength of RIP in a repeat family. The complete set of four dominance values could also be used to indicate atypical CpN mutation biases that may be specific to some repeat families but not others.

The choice of a low, medium and high dominance threshold seems arbitrary. To address this concern, it would have been relatively straightforward to assign statistical significances to the values based on comparison with random samplings of "non-RIP" mutations.

The assignment of low, medium or high dominance descriptors was somewhat arbitrary, but based on pre-existing knowledge. Table 2 shows CpA ↔ TpA dominance values for the 26 repeat families (the 3 rDNA sub-types are contained within a single repeat family) identified in *P. nodorum* (Chapters 2, 3 and 4). The low, medium and high dominance thresholds (<0.5, >0.5 and <1.2, and >1.2 respectively) were assigned based on based on incorporating these observations with details from the literature on RIP (REF). We had identified the nucleolus organiser region (NOR) containing tandem rDNA repeats, which is purported to be unaffected by RIP (REF Camamberi). Therefore the CpA ↔ TpA domainance value for tandem rDNA (to 1 decimal place = 0.5) was used as a benchmark for the "low" threshold. Out of the 3 transposable elements (MOLLY, ELSA and PIXIE (REF)), MOLLY had the highest RIP dominance value (1.2) but there were other de novo predicted repeat families with higher dominance scores. Therefore MOLLY was used as a benchmark for the "high" threshold.

The dominance score is prone to misinterpretation. James concluded that the sequences with a high RIP dominance have been RIP'd while those with low dominance have not. Specifically, he

stated that the non-tandem and short rDNAs "are presumably unaffected by RIP" and that the extensive "CpT-targeted mutations may not be related to RIP in S. nodorum".

As mentioned above, the short and **tandem** (NOR-localised) rDNA repeats were postulated to be non-RIP-affected, whereas **non-tandem** rDNA repeats were affected by RIP. In addition to the assumptions based on literature, we can clearly observe a lack of RIP mutation in tandem repeats and high frequency of CpA mutation in non-tandem repeats simply by visual inspection (Chapter 3: Figure 4).

Prior to speculating that CpT mutation may not be involved to RIP in *P. nodorum*, an analysis of dinucleotide abundance was performed to identify the dominant form of CpN mutation across all repeats relative to non-repetitive sequences (Chapter 3: Figure 1). CpA, CpC and CpG dinucleotides were less abundant in repeats, whereas CpT dinucleotides were more abundant in repeats. While some of the CpT dinucleotides contributing to this over-abundance may have arisen from mutation of the second cytosine in a CpC dinucleotide, this would not appear to be at a high enough level to significantly skew this result. All NpT dinucleotides are over-abundant in repeats except for GpT, however their fold increases are approximately 1.5 times or less. In contrast the fold depletion of CpA, CpC and CpG is approximately 3 to 4 times. CpT mutations were also observed at relatively high frequency within rDNA repeat fragments of < 200 bp (Chapter 3: Figure 4). These repeat fragments likely escape RIP due to their short length (as has been observed for short repeats in *Neurospora crassa (*REF) and have a distinctive RIP-mutation pattern compared to RIP-affected and NOR-localised (non-RIP-affected) rDNA repeats (Chapter 3: Figure 4). Based on those two pieces of information, the likelihood of CpT targeted mutation being a dominant form of RIP in *P. nodorum* was greatly reduced.

This chapter seems to imply that RIP is absolutely defined by a superabundance of CpA mutations and anything that does not conform to this pattern is not caused by RIP. This ignores what is known about RIP in other fungi. For example, a hygromycin resistance gene that was RIP'd in a cross of M. oryzae exhibited a distinct mutational bias toward CpT residues and would have yielded a dominance score of 0.4. These mutations would have been classified as non-RIP according to the dominance score and, yet, by all other criteria the hygromycin resistance gene almost certainly was RIP'd. In relation to the above discussion, it is important to remember that RIP stands for Repeat-induced Point mutations, is believed to occur prior to meiosis and is characterised by a

superabundance of transition mutations over transversions. There is nothing that says it has to exhibit CpA bias. Indeed, RIP's di-nucleotide preference is almost certainly determined by the specificity of the cytosine methylase that is believed to initiate the process. Therefore, the most likely explanation for differences in mutational spectrum is the methylase's recognition specificity.

Defining RIP mutation as the irreversible mutation of cytosine to thymine mediated by recognition by a cytosine methylase is accurate, however neglects other observed characteristics of RIP. CpA bias is one of these. While not observed across all fungi it has been clearly established within the Pezizomycotina (syn. Filamentous Ascomycota) (Clutterbuck 2011). As the Pezizomycotina make up a large number of species thus far analysed for RIP, the literature does often and justifiably present CpA mutation as the primary indicator of RIP. Given the focus on Pezizomycotina species during my PhD candidature this dissertation also focuses on CpA-biased RIP mutations. Chapters 3 to 6 use CpA mutation as an indicator of RIP because the organisms of interest, *Phaeosphaeria nodorum* and *Leptosphaeria maculans,* are phylogenetically placed within the Pezizomycotina.

RIPCAL does not however arbitrarily assume a CpA bias. The publication includes examples of repeat families which have been experimentally confirmed to be RIP-affected, some of which have a non-CpA mutation bias. The Ty1 Copia-like transposon repeat family of *Microbotryum violaceum* and the MATE transposon family of *Aspergillus nidulans* have been observed to exhibit a bias for CpG mutation, which RIPCAL accurately detected (Chapter3: Additional File 1).

Correct use of the RIPCAL tools would involve an initial analysis identifying the dominant form of CpN mutation, or if there is any preference at all. If a single CpN mutation is identified to be more prevalent than its counterparts, this bias can be used as an indicator of RIP mutation. If, as in the examiner's example of the *M. oryzae* hygromycin resistance gene, one identifies a CpN mutation bias that is not CpA, one would simply calculate CpT dominance values (Chapter 3: methods) rather than a CpA dominance values.

The examiner appears to be vigorously opposed to the role of a CpA bias in RIP mutation. I understand this may stem from caution, as little is known about RIP and focussing too narrowly onto phenomena such as CpA mutation bias may prevent researchers from observing all aspects of RIP. Nevertheless, there is no denying that CpA bias is a commonly reported feature of RIP. In appendix 4A, in response to a more direct criticism of CpA bias, I also outline a common-sense explanation as to why there may be a selective pressure to retain a CpA bias.

My main issue with the RIPCAL procedure is that it appears to have a fundamental flaw. The issue centres around the very fact that most of the sequences under study are repeated and, therefore, arose through duplication. Consequently, many (if not most) of the RIP mutations surveyed by RIPCAL will be duplications of already mutated sites, as opposed to being independent RIP mutations. It follows that analysis of multiple alignments will lead to repeated counting of a single mutational event.

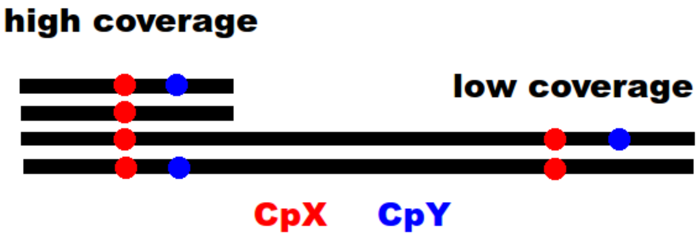
This fundamental flaw does not exist. It true that 1) graphical outputs will display a frequency plot of various CpN mutations along the alignment, 2) that these plots will be higher in frequency overall in regions of an alignment which have a higher coverage of repeat sequences, and 3) regions with higher repeat coverage will lead to repeated counting of single RIP mutation events. The main metric used to assess and quantify RIP by RIPCAL is however a ratio. As such variations for positional variation in repeat number are 'flattened out' (see next comment) in the final assessments of RIP presence and strength.

Looking at figures 2A and 4A, it is apparent that most of the X0 repeats share a large number of RIP'd sites, as do the rDNA types. Even taking into account RIP bias, it is highly unlikely that they would share so many mutant sites through independently RIP-ing. Instead, the data suggest that most of the mutations were present in a RIP'd progenitor that was subsequently duplicated. A related issue is that figures 2B and 4B do not appear to show frequency graphs as stated in the respective legends. As far as I can tell (the y-axis is not labelled), it is total counts in the sliding window. This distinction is critically important because total counts will be affected by the number of alignments for each region of a given repeat. I can see that James noted in Figure 2B that "a tendency toward higher RIp incidence was found at the ends of the alignment." However, a careful look at figure 2A shows that there are more alignments in these regions. Consequently, one would expect more RIP mutations to have been counted. It follows that the sliding window analysis does not properly serve one of its intended functions - to show how RIP incidence varies across a repeat. Instead, it provides an indication on alignment density in different regions. This will be especially problematical when aligning LTR retro-elements due to the presence of solo LTRs, which will cause very prominent spikes at the ends of RIP counts graphs. This problem is easily addressed by weighting RIP counts according to how often the nucleotide in question was present in an

alignment. This will produce a true frequency graph. Incidentally, it will also help to correct for the over-counting of RIP mutations due to replication of already RIP'd repeats.

The examiner makes a pertinent point and highlights a need to adjust the presentation of this data in the frequency plot. While in all other outputs RIPCAL deals with ratios, the frequency plots are presented as absolute counts which indicate alignment density more clearly than the relative prevalence of CpN mutations. This could potentially lead to mis-interpretation of the RIP-status of a repeat, therefore this will be amended in an upcoming release of RIPCAL (available online at www.sourceforge.net/projects/ripcal).

This criticism does not however accurately describe the behaviour of the RIPCAL algorithm. As mentioned above, the metrics RIPCAL uses to measure RIP-like polymorphism within the alignments are ratio-based. The effects of using ratios to measure repeatedly detected polymorphisms is that the coverage biases are 'flattened out' (refer to diagram below).



In the diagram above there is a region of 2X coverage (low) and a region of 4X coverage (high). For the sake of simplicity there are only two CpN mutations shown here, CpX and CpY. The Dominance value for CpX will be determined by CpX/CpY. For the high coverage region this is (4 / 2 =) 2. For the low coverage region this is (2 / 1 =) 2. Assuming the rates of RIP for each CpN mutation are constant across all regions of a multiple alignment, the dominance ratios will be 'normalised' for coverage.

In summary i dont think that the RIPCAL program has much value because it is unable to detect RIP in organisms, or at loci, that don't exhibit a significant mutational bias toward certain di-nucleotides. The most important feature of RIP - and the most reliable diagnositically is that it

produces exclusively C to T, G to A transitions. It follows that the best metric to identify RIP'd repeats is to compare the frequences of C<->T and G<->A mutations relative to all others.

The statement that RIPCAL is unable to detect RIP in organisms or loci without significant mutational bias is innacurate. RIPCAL is capable of detecting any RIP-like CpN to TpN polymorphism that exists between an aligned family of repeats. RIPCAL also provides several metrics to assist in 1) determining whether a mutational bias exists, and 2) using that identified bias as an indicator of the presence and extent of RIP. If one considers the most important feature of RIP to be C to T (or G to A) transitions, these values can be extracted from the RIPCAL tabular output along with many other values. It is largely up to the researcher to decide which metrics are the most appropriate to use under the circumstances.

The alignment-based algorithms used in RIPCAL were not revolutionary but an extension of alignment-based techniques employed in several previous studies (such as Hood et al. 2005). The RIPCAL program was created to automate these techniques on either a repeat family or whole-genome scale. A true test of its value is whether it has been used for this purpose. Within 3 years since its publication, RIPCAL has been applied to whole-genome fungal repeat analysis in three fungal genome studies that were not associated with this thesis dissertation (Di Guistini et al. 2011, Gao et al. 2011, Klosterman et al. 2011).

References:

Bhat A, Kasbekar DP (2001) Escape from repeat-induced point mutation of a gene-sized duplication in *Neurospora crassa* crosses that are heterozygous for a larger chromosome segment duplication. Genetics 157(4):1581–1590.

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

Di Guistini et al. (2011) Genome and transcriptome analyses of the mountain pine beetle-fungal symbiont *Grosmannia clavigera*, a lodgepole pine pathogen. PNAS 108(6):2504-2509

Fehmer M, Bhat A, Noubissi FK, Kasbekar DP (2001) Wild-isolated *Neurospora crassa* strains that increase fertility of crosses with segmental aneuploids used to establish that a large duplication suppresses RIP in a smaller duplication. Fungal Genet Newslett 48:13–14

Gao et al. (2011) Genome Sequencing and Comparative Transcriptomics of the Model Entomopathogenic Fungi *Metarhizium anisopliae* and *M. acridum*. PLoS Genetics 7(1): e1001264.

Hood ME, Katawczik M, Giraud T (2005) Repeat-induced point mutation and the population structure of transposable elements in *Microbotryum violaceum*. Genetics 170(3):1081-1089

Klosterman SJ, Subbarao KV, Kang S, Veronese P, Gold SE, Thomma BP, Chen Z, Henrissat B, Lee YH, Park J, Garcia-Pedrajas MD, Barbara DJ, Anchieta A, de Jonge R, Santhanam P, Maruthachalam K, Atallah Z, Amyotte SG, Paz Z, Inderbitzin P, Hayes RJ, Heiman DI, Young S, Zeng Q, Engels R, Galagan J, Cuomo CA, Dobinson KF, Ma LJ. (2011) Comparative genomics yields insights into niche adaptation of plant vascular wilt pathogens. PLoS Pathogens 7(7):e1002137

Singh PK, Kasbekar DP (2008) Titration of repeat-induced point mutation (RIP) by chromosome segment duplications in *Neurospora crassa*. Genetica 134(3):267-75.

Van de Wouw et al. (2010) Evolution of Linked Avirulence Effectors in *Leptosphaeria maculans i*s Affected by Genomic Environment and Exposure to Resistance Genes in Host Plants. PLoS Pathogens 6(11): e1001180.

Vyas M, Ravindran C, Kasbekar DP (2006) Chromosome segment duplications in *Neurospora crassa* and their effects on repeat-induced point mutation (RIP) and meiotic silencing by unpaired DNA. Genetics 172:1511–1519

Rouxel et al. (2011) Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. Nature Communications 2:202

Chapter 4: Attribution Statement

Title:	***In silico* reversal of repeat-induced point mutation (RIP) identifies the origins of repeat families and uncovers obscured duplicated genes.**
Authors:	**James K. Hane** and Richard P. Oliver
Citation:	*BMC Genomics* 11:655 (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 (JKH) made the following contributions to this chapter:

- Designed the deRIP algorithm and performed the bioinformatics analyses.

The following contributions were made by co-authors:

- JKH and RPO co-wrote the manuscript.

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

-- --

James Hane (Ph. D. candidate) Date

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

-- --

Richard Oliver (Principal supervisor) Date

BMC Genomics

RESEARCH ARTICLE Open Access

In silico reversal of repeat-induced point mutation (RIP) identifies the origins of repeat families and uncovers obscured duplicated genes

James K Hane[1,3], Richard P Oliver[2*]

Abstract

Background: Repeat-induced point mutation (RIP) is a fungal genome defence mechanism guarding against transposon invasion. RIP mutates the sequence of repeated DNA and over time renders the affected regions unrecognisable by similarity search tools such as BLAST.

Results: DeRIP is a new software tool developed to predict the original sequence of a RIP-mutated region prior to the occurrence of RIP. In this study, we apply deRIP to the genome of the wheat pathogen *Stagonospora nodorum* SN15 and predict the origin of several previously uncharacterised classes of repetitive DNA.

Conclusions: Five new classes of transposon repeats and four classes of endogenous gene repeats were identified after deRIP. The deRIP process is a new tool for fungal genomics that facilitates the identification and understanding of the role and origin of fungal repetitive DNA. DeRIP is open-source and is available as part of the RIPCAL suite at http://www.sourceforge.net/projects/ripcal.

Background

Repeat-induced point mutation (RIP) is a genome defence mechanism found within filamentous ascomycete fungi that is purported to combat transposon invasion. RIP mutates duplicated DNA sequences during sexual reproduction, thereby inactivating genes encoded in both copies. First discovered in *Neurospora crassa* [1,2], RIP was later demonstrated in the Ascomycetes *Magnaporthe oryzae* [3,4], *Podospora anserina* [5], *Leptosphaeria maculans* [6] and *Fusarium graminearum* [7]. Putative RIP events have also been detected bioinformatically in *Aspergillus fumigatus* [8], *Fusarium oxysporum* [9-11], *Aspergillus nidulans* [12], *Neurospora tetrasperma* [13], *Microbotryum violaceum* [14], *Aspergillus oryzae* [15], *Magnaporthe oryzae* [16], *Colletotrichum cereal* [17], *Aspergillus niger* [18], *Penicillium chysogenum* [18] and most recently in *Stagonospora nodorum* [19,20]. Given this broad distribution, it is reasonable to assume that RIP is widespread across, but so far restricted to, filamentous ascomycota and basidiomycota.

The mechanism by which RIP operates is yet to be fully understood, but the following observations have been made. RIP involves transition mutations from C:G to T:A nucleotide base pairs in duplicated DNA; this affects both copies of the repeat and occurs prior to meiosis [1,2]. In the majority of cases studied so far, there is a strong bias for the mutation of C:G nucleotide base pairs followed by A:T nucleotide base pairs [18,21,22]. Thus CpA di-nucleotides are more frequently affected than any of the other 15 di-nucleotides. CpA nucleotides are converted to TpA. Coincidentally, the complementary TpG di-nucleotide on the opposite strand is also converted to TpA (Table 1). In *N. crassa*, RIP requires ≥ 80% identity of duplicated DNA over a length of ≥ 400 bp [23,24].

The consequences of RIP are that repeated DNA segments, such as would result from the transposition of a retrotransposon, or the duplication of a gene, are mutated and inactivated. RIP would be expected to operate in successive sexual cycles until the sequence identity between duplicated sequences is reduced below the minimum homology threshold required by the RIP

* Correspondence: Richard.Oliver@curtin.edu.au
[2]Department of Environment and Agriculture, Curtin University, Perth, Western Australia, 6102, Australia
Full list of author information is available at the end of the article

Table 1 The four potential di-nucleotide RIP mutations detected by RIPCAL

RIP mutation				Counted di-nucleotides	
Forward		Reverse complement		Forward	Reverse complement
pre-RIP	post-RIP	pre-RIP	post-RIP		
CpA	TpA	TpG	TpA	CpA, TpA	TpG, TpA
CpC	TpC	GpG	GpA	CpC, TpC	GpG, GpA
CpG	TpG	CpG	CpA	CpG, TpG	CpG, CpA
CpT	TpT	ApG	ApA	CpT, TpT	ApG, ApA

The deRIP process counts the occurrence of the contributing di-nucleotides incrementally across a multiple alignment of repeats and alters the consensus sequence at each position to the appropriate pre-RIP di-nucleotide sequence.

machinery. The genome would then contain a repeat family consisting of relics of the duplication event degraded to varying degrees.

The rapid increase in the number of fungal genome assemblies has created a demand for methods to detect and quantify RIP. Two approaches have been used; RIP indices and alignment methods. RIP increases the frequency of particular di-nucleotides (TpA in most cases studied to date) in affected regions of DNA. Thus RIP can be identified by comparing ratios of di-nucleotide frequencies in pre-RIP to post-RIP sequences; these ratios are referred to as "RIP indices" [8,12,25,26]. However, in reality RIP depends upon the alignment of two similar regions of double-stranded DNA [23] and therefore it is more appropriate to use alignments of repeat families to identify and quantify RIP. We have previously introduced a rapid, automated alignment-based procedure for the whole-genome analysis of RIP mutation called RIPCAL [20]. Using this procedure, we readily identified and quantified the degree of RIP in all repeated DNA families within the genome of the necrotrophic fungal wheat pathogen *S. nodorum*.

Stagonospora (syn. *Septoria*) *nodorum* [teleomorph: *Phaeosphaeria* (syn. *Leptosphaeria*) *nodorum* (Müll) Hedjar.] is a major pathogen of wheat and is a model for the fungal class Dothideomycetes, a taxon that includes many important pathogens of crops [27]. *S. nodorum* infects wheat crops in most wheat-growing areas of the world [28]. Infection is predominantly determined by the presence of various effectors (host-specific toxins) harboured by different strains of the fungus [29]. The fungus is heterothallic (out-crossing) and the mating types are evenly distributed [30]. The fungus over-summers as ascospores on stubble [28] and multiplies via asexual reproduction during the growing season. The pathogen displays high levels of variability as determined by genomic analyses [31,32] and this has been exploited to determine the biogeographic history of the pathogen [33]. The pattern of micro-satellite markers is consistent with a pattern whereby the pathogen originated in the "Golden Triangle" region and spread as wheat cultivation was adopted in Eurasia and North Africa several thousand years ago and into North and South America, South Africa and Australia since European colonisation.

An initial survey of the nuclear genome sequence of a West Australian isolate (strain SN15) [19] identified 26 repeat families which comprised 6.2% of assembly. The role and origin of several repeat families could not be inferred by homology. We ascribed this to RIP mutation, after which all copies were unrecognisable. RIPCAL analysis showed that the repetitive DNA of SN15 was subject to RIP-like changes [20]. The rDNA repeat (Y1) exhibits selective susceptibility to RIP mutation (Figure 1). RIP does not affect copies located within the tandem rDNA array (also referred to as the nucleolus organiser region, or NOR, Figure 1: regions 3 & 4) [1,34]. One exception was found in a repeat at the array terminus, which showed evidence of RIP at similar levels to those of non-rDNA array repeats. rDNA repeats were also found scattered throughout the genome. Within the non-rDNA array repeats, short repeats (defined as < 1 kb, however the majority were < 300 bp) did not show evidence of RIP whilst the long repeats (> 1 kb) were RIP-affected [20]. Due to the presence of both RIP-affected and non-RIP-affected copies, the rDNA repeat was perfectly suited to be used as a test case for the validity of bioinformatic predictions of RIP.

The presence and activity of transposons in *S. nodorum* had previously been studied using a transposon trap procedure [35]. Several strains of *S. nodorum* from the United Kingdom (UK) were plated on chlorate [36] to select for mutations in the nitrate reductase (*Nia1*) structural gene. Using the cloned *Nia1* gene as a probe [37] several insertional mutants were identified. Three insertions were cloned and sequenced [35]. These insertion sequences, named Molly, Pixie and Elsa, represented intact copies of active transposons (Table 2). Southern blots probed with these transposons revealed large variations in copy number, band size and band intensity between strains. When the sequences of the intact copies of these transposons were compared to the genome sequence of the SN15 strain [19], related repetitive regions were identified. However, no active (non-RIP-affected) copies of these transposons were found in the SN15 assembly. The lack of active transposons within SN15 was intriguing and raises the question of the relationship of the repeat families to the active transposons in the UK isolates. This relationship is addressed here.

Building upon the RIPCAL procedure, we describe here a new technique to reverse the effects of RIP mutation *in silico*: "deRIP". The deRIP process involves

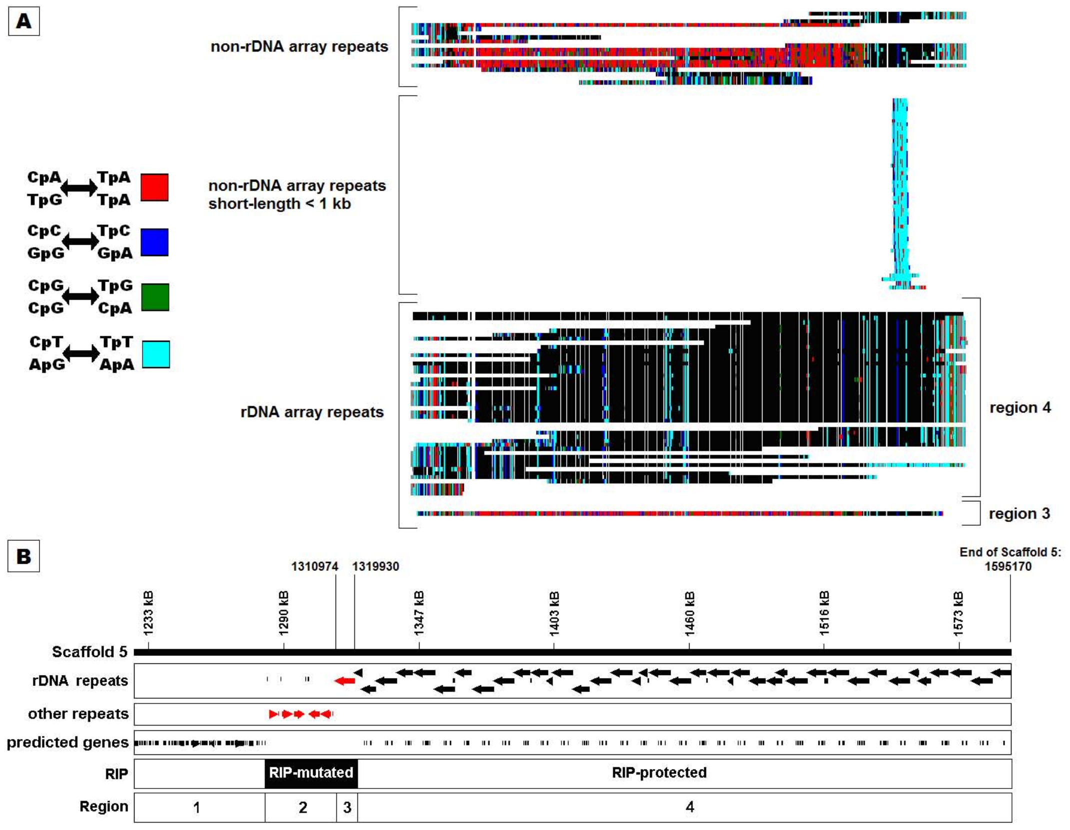


Figure 1 The distribution of the Y1 family of rDNA repeats and their susceptibility to RIP mutation. (A) A multiple alignment of Y1 rDNA repeats found in the genome of *S. nodorum* strain SN15. Each repeat was compared for mutation with the alignment majority consensus (black = match, grey = mismatch, white = gap). The mutation of CpN di-nucleotides is color-coded according to the legend (left). In *S. nodorum* RIP is characterised by the mutation of CpA di-nucleotides (red). Y1 rDNA repeats are grouped according to their genomic location and length. Full length rDNA repeats scattered randomly throughout the genome are prone to RIP whereas short, incomplete copies (defined as <1 kb but generally <300 bp) are not affected. rDNA repeats located in a tandem array at the 3' end of scaffold 5 [NCBI: CH445329] are protected from RIP, excepting a single repeat. (B) The *S. nodorum* tandem rDNA array, also known as the nucleolus organiser region (NOR), and flanking regions. Region 1 contains gene encoding regions, region 2 contains non-rDNA repeats and regions 3 and 4 comprise the tandem rDNA array. RIP mutates repetitive DNA, hence genes in region 1 are not RIP-mutated but repeats in region 2 are RIP-mutated (indicated in red). The tandem rDNA array repeats are protected from RIP (region 4), except for a single repeat at the array terminus (region 3).

scanning a multiple alignment of a repeat family for RIP-like polymorphism and reverting the alignment consensus to the putative pre-RIP-mutated sequence. The resultant "deRIPped" sequence is a prediction of what a RIP-mutated repeat DNA may have looked like prior to RIP mutation. We have applied the deRIP process to the repetitive DNA of *S. nodorum* SN15 which has increased the number of recognisable repeat families from 65% (17/26) to 92% (23/25).

Results

Validating the deRIP process using known non-RIP-affected repeats

The repetitive DNA content of the *S. nodorum* SN15 nuclear genome was previously estimated to contain 26 families comprising 6.2% of the assembly [19]. A repeat family was defined if there were 10 or more copies, of greater than 200 bp and sharing greater than 65% sequence identity. Each family had been analysed by

Table 2 Validation of the deRIP technique comparing homology of majority- and deRIP-consensus sequences with non-RIP-affected sequences

		Blastn homology					Needleman-Wunsch Global Alignment		
		Majority consensus		deRIP consensus		deRIP improvement factor	Majority consensus	deRIP consensus	deRIP improvement to percent identity
Repeat class	Hit Accession	e-value	bitscore	e-value	bitscore		Percent identity		
(A) Comparisons to active transposon sequences									
Elsa	AJ277966	1.00E-51	216	1.00E-121	381	1.8 X	69.2%	73.1%	3.9%
Molly	AJ488502	7.00E-07	66	3.00E-86	329	5.0 X	72.3%	77.5%	5.2%
Pixie	AJ488503	5.00E-07	66	2.00E-28	137	2.1 X	72.5%	75%	2.5%
(B) Comparisons to RIP-protected rDNA array consensus (Figure 1: region 4)									
Long, non-rDNA array repeats > 1 kb		0	12800	0	17220	1.3 X	89.5%	94.0%	4.5%
Short, non-rDNA array repeats < 1 kb		3.00E-10	58	1.00E-27	122	2.1 X	46.2% [a]	45.6% [a]	-0.6%
RIP-mutated terminal rDNA array repeat (Figure 1: region 3)		0	8258	–	–	–	85.8%	–	–

[a] Needleman-Wunsch global alignment was performed using a sub-region of long rDNA repeats corresponding to the short rDNA repeat consensus

Blastn hits and pairwise global percent identities to non-RIP-affected sequences were compared between the majority consensus and deRIP consensus versions. (A) The transposons Elsa, Molly and Pixie of *S. nodorum* SN15 were compared to active copies of an alternate strain. In all 3 cases the deRIP sequences match best to the active transposons. This is indicated by the 'deRIP improvement' factor and the differences in percent identities for global alignments. DeRIP improvement is a measure of how much better the deRIP consensus matched the hit compared to the majority consensus. DeRIP improvement > 1 indicates that the repeat family was derived from the hit or a related homolog, but was subsequently mutated by RIP. (B) RIP-protected copies of the *S. nodorum* rDNA repeat are located within a tandem array (Figure 1). RIP-susceptible copies were grouped by size into long (> 1 kB) and short (< 1 kB) categories and compared to the RIP-protected copies. Homology between RIP-protected repeats in rDNA array and long RIP-susceptible non-rDNA array repeats were improved by deRIP. The rDNA array also contains one RIP-affected repeat at its terminus which shows similar levels of homology to the rDNA array as the majority consensus of the long non-rDNA array repeats.

RIPCAL and the extent of RIP measured using the CpA↔TpA dominance statistic [20]. RIP dominance varied from 0.2 to 2.96 (by comparison to the highest G: C content sequence). Blast comparisons predicted the origin of 17 out of the 26 repeat families.

Functional and authentic transposon homologues of the repeat families Molly, Pixie and Elsa had been previously characterized. Elsa was identified as a LTR retrotransposon; Molly and Pixie as Tc-1 Mariner elements [38]. Characterized sequences were derived from UK isolates of *S. nodorum* [35]. The maximum sequence identity between the proteins encoded by the active copies and matches within the SN15 genome assembly was approximately 66% by blastx (Additional file 1).

To determine whether the SN15 repeats were derived from the active copies via RIP mutation, the deRIP procedure was applied to the alignment of the Molly, Elsa and Pixie-like sequences. The example shown in Figure 2 illustrates the deRIP process applied to the transposon repeat Molly.

Molly-like repeat sequences of SN15 were aligned and analysed for RIP mutation via RIPCAL [20] (Figure 2A). The alignment includes 18 full length copies and 22 incomplete copies. Mismatches between individual repeats and the majority consensus are colour-coded; vertical red bars represent the CpA/TpG to TpA di-nucleotide substitution previously shown to be the predominant RIP-induced change in *S. nodorum* [20]. The predominance of red changes indicates that the repeat family has been affected by RIP.

The Molly alignment was processed using the new deRIP algorithm. The process is illustrated in Figure 2B in a 51 bp subsection of the alignment from position 1900 to 1950. At position 1900-1901 of the alignment there is a TpA di-nucleotide in 23 out of 24 copies and TpG in one copy. This set of di-nucleotides corresponds to the TpG→ TpA mutation, which is characteristic of RIP (the reverse complement of CpA→ TpA, Table 1). It was assumed that the TpA copies were derived from an ancestral TpG via RIP. Therefore while the majority consensus (alignment consensus by base majority) was TpA at this position, the deRIP process changed this to the most probable pre-RIP sequence - TpG. This process was extended across the length of the repeat alignment, producing a new sequence called the 'deRIP consensus'. This deRIP consensus sequence was compared to the majority consensus as well as the sequence of the active copy of Molly [NCBI: AJ488502.1] (Figure 2B). In this example, deRIP changes were labelled as "correct" where alterations in the deRIP consensus agreed with the sequence of the active copy. Nine such cases occurred in the highlighted section.

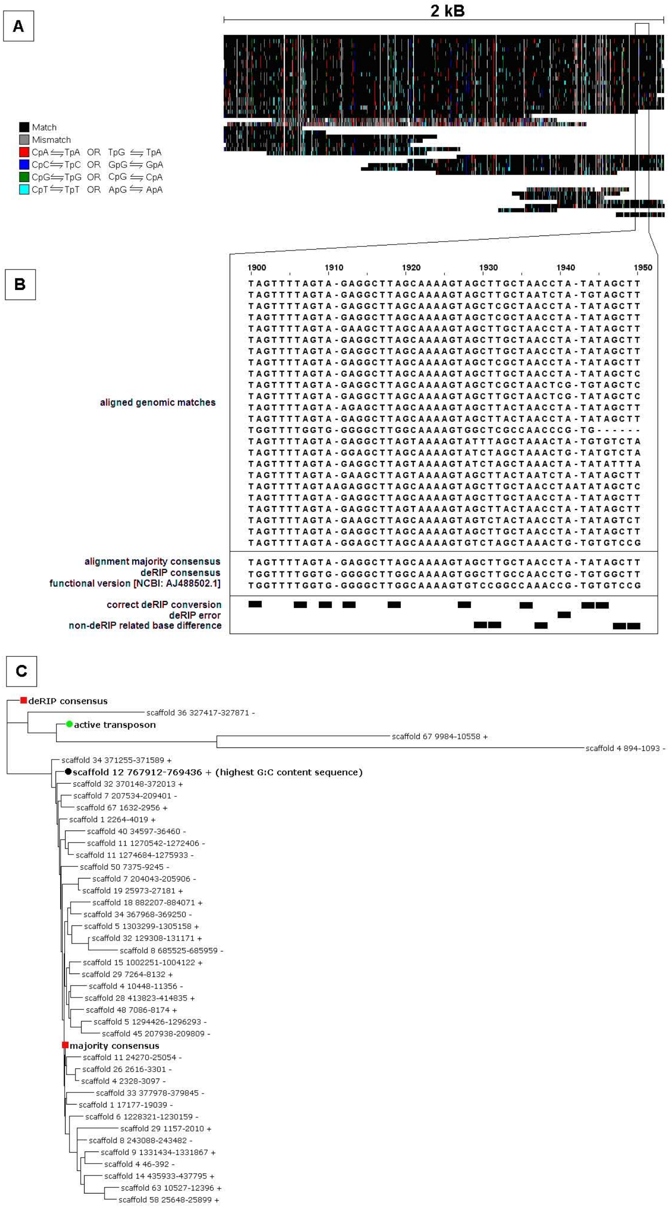

A
2 kB
Match
Mismatch
CpA ⇋ TpA OR TpG ⇋ TpA
CpC ⇋ TpC OR GpG ⇋ GpA
CpG ⇋ TpG OR CpG ⇋ CpA
CpT ⇋ TpT OR ApG ⇋ ApA
B
1900
1910
1920
1930
1940
1950
aligned genomic matches
TAGTTTTAGTA-GAGGCTTAGCAAAAGTAGCTTGCTAACCTA-TATAGCTT
TAGTTTTAGTA-GAGGCTTAGCAAAAGTAGCTTGCTAATCTA-TGTAGCTT
TAGTTTTAGTA-GAGGCTTAGCAAAAGTAGCTCGCTAACCTA-TATAGCTT
TAGTTTTAGTA-GAGGCTTAGCAAAAGTAGCTCGCTAACCTA-TATAGCTT
TAGTTTTAGTA-GAAGCTTAGCAAAAGTAGCTTGCTAACCTA-TATAGCTT
TAGTTTTAGTA-GAGGCTTAGCAAAAGTAGCTTGCTAACCTA-TATAGCTT
TAGTTTTAGTA-GAGGCTTAGCAAAAGTAGCTTGCTAACCTA-TATAGCTT
TAGTTTTAGTA-GAGGCTTAGCAAAAGTAGCTCGCTAACCTA-TATAGCTT
TAGTTTTAGTA-GAGGCTTAGCAAAAGTAGCTTGCTAACCTA-TATAGCTC
TAGTTTTAGTA-GAGGCTTAGCAAAAGTAGCTCGCTAACTCG-TGTAGCTC
TAGTTTTAGTA-GAGGCTTAGCAAAAGTAGCTTGCTAACTCG-TATAGCTC
TAGTTTTAGTA-AGAGCTTAGCAAAAGTAGCTTACTAACCTA-TATAGCTT
TAGTTTTAGTA-GAGGCTTAGCAAAAGTAGCTTACTAACCTA-TATAGCTT
TGGTTTTGGTG-GGGGCTTGGCAAAAGTGGCTCGCCAACCCG-TG------
TAGTTTTAGTA-GAGGCTTAGTAAAAGTATTTAGCTAAACTA-TGTGTCTA
TAGTTTTAGTA-GGAGCTTAGCAAAAGTATCTAGCTAAACTG-TATGTCTA
TAGTTTTAGTA-GAGGCTTAGTAAAAGTATCTAGCTAAACTA-TATATTTA
TAGTTTTAGTA-GAAGCTTAGTAAAAGTAGCTTACTAATCTA-TATAGCTT
TAGTTTTAGTAAGAGGCTTAGCAAAAGTAGCTTGCTAACCTAATATAGCTC
TAGTTTTAGTA-GAGGCTTAGCAAAAGTAGCTTGCTAACCTA-TATAGCTT
TAGTTTTAGTA-GAGGCTTAGTAAAAGTAGCTTACTAACCTA-TATAGCTT
TAGTTTTAGTA-GAAGCTTAGCAAAAGTAGTCTACTAACCTA-TATAGTCT
TAGTTTTAGTA-GAGGCTTAGCAAAAGTAGCTTGCTAACCTA-TATAGCTT
TAGTTTTAGTA-GGAGCTTAGCAAAAGTGTCTAGCTAAACTG-TGTGTCCG
alignment majority consensus
deRIP consensus
functional version [NCBI: AJ488502.1]
TAGTTTTAGTA-GAGGCTTAGCAAAAGTAGCTTGCTAACCTA-TATAGCTT
TGGTTTTGGTG-GGGGCTTGGCAAAAGTGGCTTGCCAACCTG-TGTGGCTT
TGGTTTTGGTG-GGGGCTTGGCAAAAGTGTCCGGCCAAACCG-TGTGTCCG
correct deRIP conversion
deRIP error
non-deRIP related base difference
C
deRIP consensus
scaffold 36 327417-327871 -
active transposon
scaffold 67 9984-10558 +
scaffold 4 894-1093 -
scaffold 34 371255-371589 +
scaffold 12 767912-769436 + (highest G:C content sequence)
scaffold 32 370148-372013 +
scaffold 7 207534-209401 -
scaffold 67 1632-2956 +
scaffold 1 2264-4019 +
scaffold 40 34597-36460 -
scaffold 11 1270542-1272406 -
scaffold 11 1274684-1275933 -
scaffold 50 7375-9245 -
scaffold 7 204043-205906 -
scaffold 19 25973-27181 +
scaffold 18 882207-884071 +
scaffold 34 367968-369250 -
scaffold 5 1303299-1305158 +
scaffold 32 129308-131171 +
scaffold 8 685525-685959 -
scaffold 15 1002251-1004122 +
scaffold 29 7264-8132 +
scaffold 4 10448-11356 -
scaffold 28 413823-414835 +
scaffold 48 7086-8174 +
scaffold 5 1294426-1296293 -
scaffold 45 207938-209809 -
majority consensus
scaffold 11 24270-25054 -
scaffold 26 2616-3301 -
scaffold 4 2328-3097 -
scaffold 33 377978-379845 -
scaffold 1 17177-19039 -
scaffold 6 1228321-1230159 -
scaffold 29 1157-2010 +
scaffold 8 243088-243482 -
scaffold 9 1331434-1331867 +
scaffold 4 46-392 -
scaffold 14 435933-437795 +
scaffold 63 10527-12396 +
scaffold 58 25648-25899 +

Figure 2 Application of the deRIP process to the Molly transposon repeat family of *Stagonospora nodorum* SN15. Molly is one of three *S. nodorum* repeats with known functionally transposable sequence available [NCBI AJ488502.1]. (A) Genomic matches to the Molly repeat family were aligned and compared for RIP-like polymorphism against a model sequence (in this case the majority consensus). RIP mutation of the form CpN ←→ TpN was color-coded as indicated in the legend. (B) The deRIP process was applied to a 51 bp sub-region of the alignment. A 'majority' consensus of the alignment represented the most abundant nucleotide at each alignment position. The deRIP consensus was derived from the majority consensus, however where di-nucleotides were detected exhibiting RIP-like polymorphism (Table 1) they were reverted back to their pre-RIP state. Changes in sequence between majority and deRIP consensus sequences was compared to the sequence of the active transposon. (C) Phylogram showing relationships between all genomic regions, majority consensus, deRIP consensus and active copy of the Molly repeat family. The deRIP consensus resembled the functional transposon more closely than the majority consensus, highest G:C content sequence and the majority of matching genomic regions.

DeRIP changes were labelled as "errors" where the deRIP changes and the active copy sequence did not agree. There was one deRIP error in the sub-alignment at alignment position 1940-1. Non-deRIP related base differences, common between the majority and deRIP consensus sequences but different in the active copy sequence, occurred five times in the sub-alignment.

All the Molly-like repeats, the active copy, the alignment 'majority' consensus sequence and the new deRIP consensus sequence were compared via RAxML (using the gamma model and maximum-likelihood phylogeny) [39] (Figure 2C). The deRIP-predicted sequence was a closer match to the authentic, active copy than the majority consensus. The relative levels of sequence similarity between the active transposon and the majority and deRIP consensus sequences were also tested via Needleman-Wunsch global alignment [40]. The sequence identity between the active copy and majority consensus was 72.3% whereas the identity between the active copy and the deRIP consensus was 77.5% (Table 2).

The deRIP process was applied to the other transposon repeat families with pre-existing characterized active copy sequences, Pixie and Elsa. Table 2A summarises the results for all three previously identified active *S. nodorum* transposons. In the case of Molly, the majority consensus by blastn had an e-value to the active copy of 7e-07 (bitscore = 66) whereas the deRIP consensus by blastn had ane-value of 3e-86 (bitscore = 329). These results can be summarised as a "deRIP improvement" of 329/66 = 5.0. DeRIP improvement factors were 1.8 and 2.1 and global percent identities to the active transposons were improved 3.9% and 2.5% for Elsa and Pixie respectively (Table 2). The overall improvement in maximum bit scores to active transposons indicates that the deRIP versions were significantly better matches to the functional transposons that were the presumed ancestors of the sequences in the Australian SN15 strain.

The rDNA repeat family Y1 had been previously demonstrated to show differential susceptibility to RIP between its various copies (Figure 1) [20]. rDNA repeats within a tandem rDNA array were not RIP-affected except for one repeat at the array terminus. Non-rDNA array repeats greater than 1 kb (which we call "long") showed evidence of RIP, however non-rDNA array repeats less than 1 kb ("short") did not. After deRIP was applied to the consensus of long, non-rDNA array repeats, the percent identity to the non-RIP-affected rDNA-array consensus was improved by 4.5% - from 89.5% to 94% and the deRIP improvement factor was 1.3 (Table 2). Conversely, the percent identity between the non-rDNA array short repeat consensus and the rDNA array consensus was not improved by deRIP (Table 2). The RIP-affected terminal rDNA repeat and the majority consensus of the RIP-affected long non-rDNA array repeats both had similar levels of homology to the rDNA array (Table 2).

Determining the role and origin of RIP-degraded repeats in *S. nodorum*

The deRIP process was extended to all repeat families of *S. nodorum* SN15 (Additional files 2, 3, 4). Table 3 summarises the copy number and size of repeat families as estimated previously [19,20]. The extent to which repeat families were affected by RIP is indicated by the RIP dominance scores. RIP dominance [20] was calculated using a variety of comparative models including: sequence of highest G:C content; alignment majority consensus and; consensus sequence predicted by deRIP (Additional file 5). A RIP dominance of greater than 0.6 by comparison to the repeat with highest G:C content was considered a reliable threshold for RIP [20].

DeRIP consensus-generated RIP dominances correlated with the highest G:C content RIP dominance scores better (correlation coefficient = 0.88) than those of the majority consensus (0.85). This supports the reliability of the deRIP consensus as an accurate prediction of the pre-RIP-mutated progenitor sequence.

Table 3 also lists Blast hits to the NCBI NR and GIRI Repbase. The number of hits of the majority consensus is compared to those of the deRIP consensus sequence. Similarly, the number of hits of either sequence to Repbase is also reported. The deRIP improvement factor used the ratio of highest bit scores of the deRIP and majority sequences to either NR by blastx or Repbase by tblastx respectively. An improvement factor can only be calculated if both consensus sequences have hits above the thresholds (see methods).

In the great majority of cases the number of hits of the deRIP consensus matched or exceeded the number achieved by the majority consensus sequences (Table 3). In two cases, (X35, X0 to NR) the deRIP sequence found a hit where none had been found before (Table 4). In other cases, very substantial increases in hit number were observed (R8, R9 to NR; R9 to Repbase). In a few cases the number of hits was reduced (X12, R37 and R51 by NR; R10, Pixie, R31 and R37 by Repbase).

The deRIP improvement factor was greater than one in all cases for NR and in all but two case for Repbase indicating a general increase in the confidence and significance of a hit and hence a functional assignment. The factor ranged up to 3.78 for NR and up to 2.18 for Repbase hits (Table 3). In two cases (R25 and X23) the factor with Repbase was less than 1. This can occur if the hit present in the reference database had been submitted in its non-functional, RIP-affected form.

Blast information was used to determine the origin of several RIP-degraded repeat families of *S. nodorum*

Table 3 Summary of RIP mutation in the repeat families of S. nodorum strain SN15

			RIP Dominance Scores			NCBI NR Protein Blastx			GIRI Repbase Tblastx		
Repeat Family	**Copy Number**	**Full Length (bp)**	**RIP dominance by highest G:C content [20]**	**RIP dominance by majority consensus**	**RIP dominance by deRIP consensus**	**Hits to deRIP consensus**	**Hits to majority consensus**	**deRIP Improvement factor (Maximum value)**	**Hits to deRIP consensus**	**Hits to majority consensus**	**deRIP Improvement factor (Maximum value)**
R8	48	9143	2.96	1.95	2.91	126	77	1.96			
R10	59	1241	1.91	0.96	2.07	3	2	2.13	0	3	
X0	76	3862	2.13	0.97	2.05	1	0		3	1	1.34
R9	72	4108	1.88	0.92	1.77	250	25	2.75	124	4	1.28
Molly	40	1862	1.21	0.64	1.73	250	161	3.78	34	15	1.92
X3	213	9364	0.63	0.81	1.62	11	10	2.8			
X35	19	1157	1.5	1.34	1.43	1	0				
X96	14	308	0.87	0.89	1.39						
X48	22	265	1.82	1.16	1.33						
R22	23	678	1.2	0.84	1.28				2	2	1.06
X26	38	4628	1.16	1.08	1.19	57	57	1.38			
Pixie	28	1845	0.77	0.57	1.06	250	190	1.79	17	18	1.25
R37	98	1603	0.49	0.25	0.95	0	55		4	18	1.16
R31	23	3031	0.99	0.83	0.9	16	15	1.44	3	7	1.14
X23	29	685	0.45	0.4	0.9				3	3	0.82
X36	10	512	0.89	0.78	0.87	2	1	1.43			
Elsa	17	5240	0.86	0.78	0.82	250	231	2.06	65	30	1.44
R51	39	833	0.47	0.31	0.8	0	3		0	3	
X11	36	8555	0.83	0.71	0.78	250	250	1.35	250	228	2.18
X12	29	2263	0.67	0.43	0.76	0	1		10	10	1.44
R39	29	2050	0.59	0.28	0.74	173	149	1.54	34	31	1.88
X28	30	1784	0.83	0.59	0.73						
R25	23	3320	0.25	0.6	0.65	4	4	1.19	3	1	0.86
X15	37	6231	0.61	0.46	0.61	250	250	1.45	243	217	1.66
R38	25	358	0.2	0.14	0.5						

RIP dominance, a measure of the strength of RIP mutation, is reported for all 3 different RIPCAL comparison methods: versus the highest G:C content sequence; versus the alignment 'majority' consensus and; versus the deRIP consensus. Measures of how much the predicted deRIP consensus of a repeat family resembles its original version, hit discovery scores and deRIP improvement factors, are also summarised for comparisons against NCBI NR Proteins via blastx and the GIRI Repbase database of repetitive elements via tblastx.

SN15. Previously the probable origin of 17 out of 26 repeat families had been identified. However after deRIP had been applied to each repeat family, 23 out of 25 have now been categorised. In six cases (R10, R31, R39, R51, X23 and X36) no previous homology information had been available. Repeat families R31, R39, R51, X23 and X36 were re-classified as transposons after deRIP analysis (Table 4). The repeat family R10, also previously unknown, was identified as corresponding to *S. nodorum* genes SNOG_15997, SNOG_11270 and SNOG_16585 [NCBI: EAT76576.1, EAT81769.1, EAT76052.1].

The previous classification of X15 as a Gypsy class transposon remnant was confirmed after deRIP. The deRIP improvement factors for Gypsy sequences were 1.45 and 1.66 for NR Proteins and Repbase sequences respectively (Table 4). X26, previously predicted to be a transposon remnant, was found after deRIP to contain regions corresponding to a telomere-associated RecQ helicase (Table 4). R25 was previously classified as a putative transposon remnant. After deRIP, some weak homology to DNA transposons was detected versus Repbase but a region of homology to histone H3 proteins was also detected (Table 4). Repeat family R25 was thus re-classified as originating from a (presumably) endogenous gene-encoding region.

R8 and X3 were previously predicted to contain the remnants of an ubiquitin conjugating enzyme and helicase genes respectively [19,20]. DeRIP analysis was used to predict the ancestral sequence and identified matches to nine copies of a cluster of endogenous *S. nodorum* genes (Additional file 3). Analysis of repeats X3 and R8 indicated that several copies of these repeat families

Table 4 Classification of repeat family origin in S. nodorum SN15

Repeat family	Predicted origin [19,20]	Predicted origin after deRIP	comparison type	informative hits	Majority Consensus e-value	deRIP consensus e-value	deRIP improvement factor (maximum)
X26	Sub-telomeric, transposon remnant	Telomere-associated RecQ helicase	blastx vs NR	EAL89306.1 telomere-associated RecQ helicase, putative *Aspergillus fumigatus* Af293	1.00E-07	2.00E-12	1.25
R25	Transposon remnant	Histone H3	blastx vs NR	EDU47581.1 histone H3 *Pyrenophora tritici-repentis* Pt-1C-BFP	0.032	2.00E-04	1.16
			tblastx vs Repbase	TDD4 DNA transposon *Dictyostelium_discoideum*		6.00E-04	
R10	Unknown	Uncharacterized endogenous gene region and DNA transposon	blastx vs NR	EAT76576.1 hypothetical protein SNOG_15997 *Phaeosphaeria nodorum* SN15	2.2	2.00E-13	2.13
			blastx vs NR	EAT81769.1 hypothetical protein SNOG_11270 *Phaeosphaeria nodorum* SN15		5.00E-11	
			blastx vs NR	EAT76052.1 hypothetical protein SNOG_16585 *Phaeosphaeria nodorum* SN15	0.006	2.00E-08	1.40
			tblastx vs Repbase	CR1-3_HM CR1 *Hydra magnipapillata*	9.00E-06		
R31	Unknown	DNA Transposon	blastx vs NR	CAP79587.1 Pc23g00930 *Penicillium chrysogenum* Wisconsin 54-1255	0.013	1.00E-06	1.28
			tblastx vs Repbase	hAT-1_AN hAT DNA transposon *Emericella nidulans*	1.00E-05	1.00E-06	1.07
R39	Unknown	Mariner/Tc1-like DNA transposon	blastx vs NR	EAT91063.1 hypothetical protein SNOG_01414 *Phaeosphaeria nodorum* SN15	2.00E-62	3.00E-73	1.15
			blastx vs NR	EED11513.1 pogo transposable element, putative *Talaromyces stipitatus* ATCC 10500	2.00E-28	1.00E-36	1.20
			tblastx vs Repbase	Mariner-9_AN Mariner/Tc1 *Emericella_nidulans*	8.00E-37	1.00E-25	1.01
R51	Unknown	Mariner/Tc1-like DNA transposon	tblastx vs Repbase	P-29_HM P *Hydra magnipapillata*	1.00E-05		
			tblastx vs Repbase	Mariner-31_HM Mariner/Tc1 *Hydra magnipapillata*	3.00E-05		
X23	Unknown	LTR Retrotransposon	tblastx vs Repbase	ATCOPIA80_I Copia *Arabidopsis thaliana*		1.00E-04	
			tblastx vs Repbase	CR1-3_HM CR1 *Hydra magnipapillata*	9.00E-05	3.00E-04	0.82
X36	Unknown	Retrotransposon	blastx vs NR	EAS29858.1 hypothetical protein CIMG_08604 *Coccidioides immitis* RS	4.9	2.00E-04	1.43
			blastx vs NR	gag-pol polyprotein *Podospora anserina*		4.00E-03	

Table 4 Classification of repeat family origin in S. nodorum SN15 *(Continued)*

X3X3R8	X3: Helicase	Endogenous gene cluster containing tandem duplicated Rad5/SNF2-like helicase, Rad6/ubiquitin conjugating enzyme and uncharacterised ORFs	blastx vs NR	EAT83378.1 hypothetical protein SNOG_09186 EAT91019.1 hypothetical protein SNOG_01370 *Phaeosphaeria nodorum* SN15	1.00E-165	0	1.48
			blastx vs NR	EAT90556.1 hypothetical protein SNOG_02344 *Phaeosphaeria nodorum* SN15	6.00E-75	1.00E-122	1.54
			blastx vs NR	EAT83381.1 hypothetical protein SNOG_09189 EAT91023.1 hypothetical protein SNOG_01374 *Phaeosphaeria nodorum* SN15	8.00E-93	1.00E-117	1.51
			blastx vs NR	EAT90553.1 hypothetical protein SNOG_02341 *Phaeosphaeria nodorum* SN15	7.00E-61	1.00E-100	1.51
			blastx vs NR	EAT92620.1 hypothetical protein SNOG_16597 *Phaeosphaeria nodorum* SN15	9.00E-39	2.00E-49	1.43
			blastx vs NR	EAT91018.1 hypothetical protein SNOG_01369 *Phaeosphaeria nodorum* SN15	3.00E-30	1.00E-48	1.44
			blastx vs NR	EAT90555.1 hypothetical protein SNOG_02343 *Phaeosphaeria nodorum* SN15	7.00E-36	5.00E-33	1.31
			blastx vs NR	EAT90554.1 hypothetical protein SNOG_02342 *Phaeosphaeria nodorum* SN15	1.00E-36	2.00E-26	1.34
			blastx vs NR	EAT83379.1 hypothetical protein SNOG_09187 *Phaeosphaeria nodorum* SN15	3.00E-14	1.00E-21	1.28
			blastx vs NR	EAT91020.2 hypothetical protein SNOG_01371 *Phaeosphaeria nodorum* SN15	3.00E-15	2.00E-20	1.19
			blastx vs NR	EAT83294.1 hypothetical protein SNOG_09102 *Phaeosphaeria nodorum* SN15	2.00E-13	4.00E-20	1.25
			blastx vs NR	EAT83377.2 hypothetical protein SNOG_09185 *Phaeosphaeria nodorum* SN15	1.00E-13	8.00E-20	1.23
			blastx vs NR	EAT92618.1 hypothetical protein SNOG_16595 *Phaeosphaeria nodorum* SN15	3.00E-07	4.00E-19	1.60
			blastx vs NR	EAT83380.1 hypothetical protein SNOG_09188 EAT91022.1 hypothetical protein SNOG_01373 EAT92619.1 hypothetical protein SNOG_16596 *Phaeosphaeria nodorum* SN15	4.00E-05	7.00E-15	1.57

Table 4 Classification of repeat family origin in S. nodorum SN15 *(Continued)*

	blastx vs NR	EAT91021.1 hypothetical protein SNOG_01372 *Phaeosphaeria nodorum* SN15	2.00E-06	8.00E-14	1.40
	blastx vs NR	EDU40406.1 ubiquitin-conjugating enzyme E2-21 kDa *Pyrenophora tritici-repentis* Pt-1C-BFP	2.00E-10	2.00E-16	1.27
	blastx vs NR	EAW17873.1 ubiquitin conjugating enzyme (*UbcC*), putative *Neosartorya fischeri* NRRL 181	1.00E-07	2.00E-13	1.29
R8: Ubiquitin conjugating enzyme	blastx vs NR	EAT91013.2 hypothetical protein SNOG_01364 *Phaeosphaeria nodorum* SN15	0	0	0.91
	blastx vs NR	EAT92627.2 hypothetical protein SNOG_16589 *Phaeosphaeria nodorum* SN15	0	0	0.91
	blastx vs NR	EAT83373.2 hypothetical protein SNOG_09181 *Phaeosphaeria nodorum* SN15	1.00E-177	0	0.95
	blastx vs NR	EAT90557.2 hypothetical protein SNOG_02345 *Phaeosphaeria nodorum* SN15	2.00E-65	1.00E-106	2.80
	blastx vs NR	EAT90559.2 hypothetical protein SNOG_02347 *Phaeosphaeria nodorum* SN15	1.00E-62	5.00E-91	1.38
	blastx vs NR	EAT91015.1 hypothetical protein SNOG_01366 *Phaeosphaeria nodorum* SN15	3.00E-40	3.00E-62	1.35
	blastx vs NR	EAT85951.1 hypothetical protein SNOG_06120 *Phaeosphaeria nodorum* SN15	2.00E-18	3.00E-24	1.19
	blastx vs NR	EAT91016.1 hypothetical protein SNOG_01367 *Phaeosphaeria nodorum* SN15	1.00E-15	2.00E-23	1.28
	blastx vs NR	EAT83374.2 hypothetical protein SNOG_09182 *Phaeosphaeria nodorum* SN15	5.00E-05	4.00E-08	1.23
	blastx vs NR	EAT91014.2 hypothetical protein SNOG_01365 *Phaeosphaeria nodorum* SN15	1.00E-04	2.00E-04	1.15

After deRIP analysis the predicted origin of 8 repeat families has been altered from that described in Hane & Oliver (2008) [20]. Details of the blast hits which were most informative in re-classifying a repeat family are listed below. E-values are shown for matches to both the majority and deRIP consensus sequences. DeRIP improvement is a measure of how much better the deRIP consensus matched the hit compared to the majority consensus. DeRIP improvement > 1 indicates that the repeat family was derived from the hit or a related homolog, but was subsequently mutated by RIP.

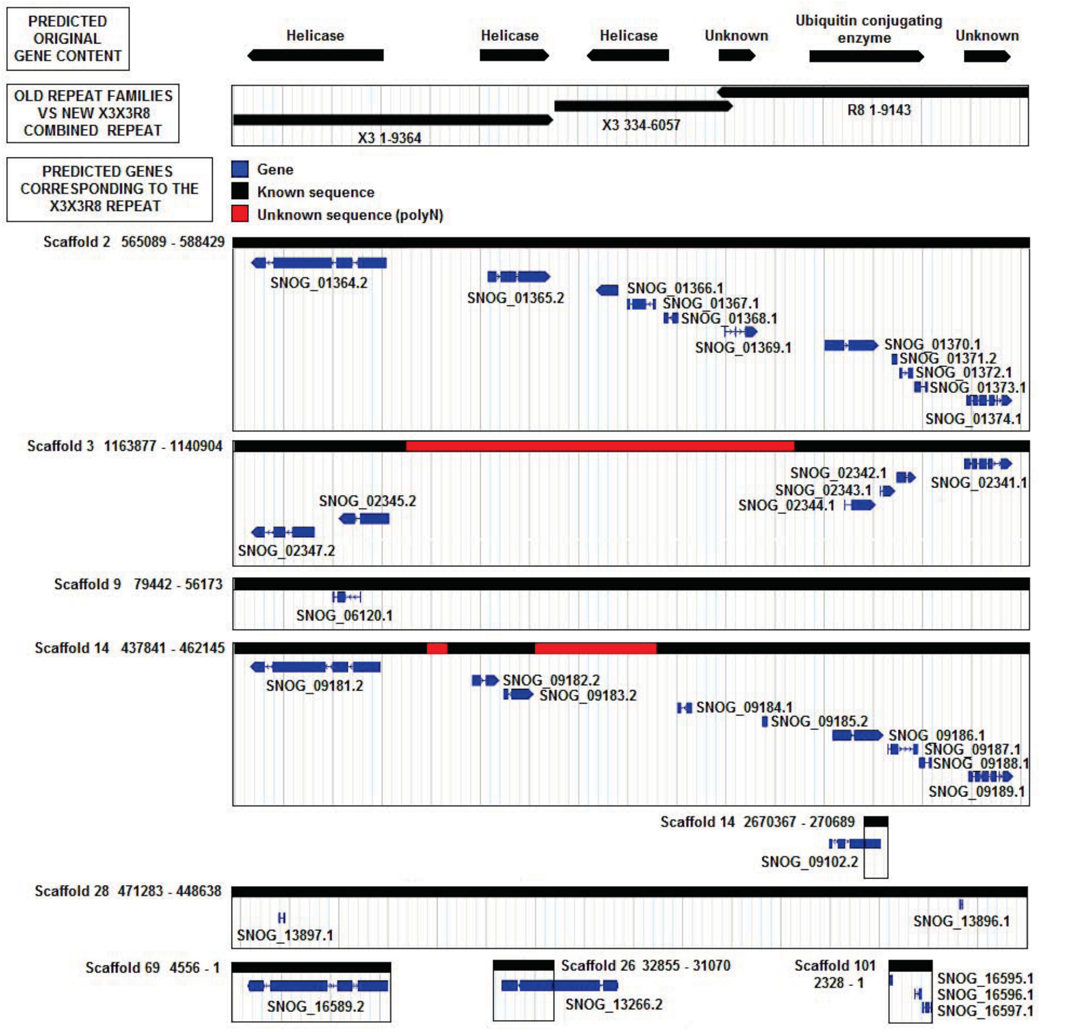


Figure 3 Nine copies of the repeat X3X3R8 contain predicted gene annotations in the *S. nodorum* genome. Blast analysis of the deRIP consensus sequence led to the hypothesis that 3 helicase genes, an ubiquitin conjugating enzyme and 2 unknown genes originally occupied this region. The effects of RIP mutation have led to the disruption of open-reading frames in several of these genes resulting in multiple, short-length gene predictions which are highly likely to be pseudogenes.

were physically adjacent (Additional file 3). Studying the location of X3 and R8 repeats revealed that these repeats were frequently arranged in a distinctive pattern roughly corresponding to two tandem X3 repeats followed by a reversed R8 repeat (Figure 3, Figure S# 4). These two repeat classes were combined and renamed X3X3R8.

In addition to the cluster of endogenous genes, the X3X3R8 deRIP consensus also hit known ubiquitin conjugating enzymes with greater homology than the majority consensus (Table 4). Homology relationships to DNA excision/repair helicase regions were inferred from hits to the endogenous *S. nodorum* genes residing within X3X3R8 (Additional file 2).

Since the deRIP consensus is a prediction of the sequence prior to RIP-mutation it can be presumed that the X3X3R8 repeats with functional genes are closely related to the deRIP consensus. The majority of repeats with predicted gene annotations were found to be highly similar to the deRIP consensus (Figure 4, green circles). Current evidence supports the functionality of some or all of the genes contained in six out of these nine X3X3R8 repeats.

Discussion

The deRIP algorithm was designed to reverse the affect of RIP upon repeat families in-silico and thereby help determine the evolutionary history of repeated elements. The process automates the selective alteration of bases within a consensus according to a set of rules that can be determined by considering the RIP machinery operating in the organism concerned. Despite its validation with known non-RIP-affected sequences, deRIP has some limitations and will not necessarily perfectly predict the ancestral sequence. DeRIP can only choose within options provided by the aligned set of repeats. In the example given in Figure 2 the TpA sequence at position 1900-1901 was converted to TpG. To do this, at least one of the copies of the repeat must have the presumably ancestral di-nucleotide TpG at this site. If all extant copies had been mutated, the reversion would have no support. The deRIP process is therefore critically dependent on the degree of RIP within a repeat. The success of deRIP is also dependent upon the accuracy of the alignment. A noteworthy aside is that default alignment parameters often fail to align fungal repeats correctly due to complex internal repeat structures. Finally, the diagnostic metrics of deRIP success, deRIP improvement and hit discovery, are only possible to calculate if appropriate matching sequences exist in the queried databases. If a repeat sequence is truly novel, its "homology" cannot be improved until homologs are found.

RIPCAL uses a model sequence to compare to aligned repeats for RIP-like polymorphism. Selecting the repeat with the highest total count of G and C nucleotides assumes that high G:C content is representative of the least RIP-affected repeat. The majority consensus model on the other hand could be representative of the least or most RIP-affected repeat depending on the level of RIP mutation within the repeat family. Previously, we had selected the sequence with the highest G:C content as the RIPCAL model [20]. While in most cases this rationale is sound, the G:C model has several shortcomings. If a sequence with the highest total G:C content does not span the full length of its alignment, RIP data from the un-covered regions would be lost. Alternatively, a repeat longer than the least RIP-affected repeat (e.g. resulting from a large sequence insertion into a RIP-affected repeat) may have higher total G:C content merely due to its greater length. The G:C model is also sensitive to variations in G:C content not related to RIP. Furthermore, RIP occurs between multiple combinations of repeats over time. The G:C model sequence therefore comprises of an amalgam of pre-RIP and post-RIP di-nucleotides relative to the alignment as a whole.

RIP mutations have directionality (Table 1), so the combination of pre- and post-RIP sites makes it necessary to consider RIP mutation both towards and away from the G:C model sequence. In contrast, a deRIP consensus model, being a prediction of the pre-RIP-mutated sequence, has the advantage of polarity. As such, deRIP mutation calculations can be restricted to one direction: proceeding from the deRIP consensus to the RIP-affected repeat.

The prior isolation of active copies of three transposons, Molly Pixie and Elsa, as well as the differential effect of RIP on the rDNA repeats, allowed a thorough test of the power of deRIP to reconstruct the ancestral sequence. In all four examples the predicted deRIP consensus of the RIP-affected sequences was the best match to the active copy indicating that deRIP was able to accurately revert the RIP-degraded repeats close to their original states. These analyses helped define the concepts of hit discovery number and deRIP improvements as applied more broadly in Table 3.

The deRIP process serves to highlight the effectiveness of RIP as a transposon-silencing mechanism. In most observed cases, the resemblance between RIP-degraded repeats and their non-RIP-affected, functional counterparts is minimal (Additional file 1). In the case of the Molly, Elsa and Pixie transposons, functional sequences of transposon proteins were available for comparison. No viable open-reading frames could be found in any of their respective genomic matches in *S. nodorum* SN15 (Additional file 1). Some repeat families could not even be classified by homology prior to deRIP (Table 4). The deRIP process is therefore an essential tool which facilitates the identification and understanding of the role and origin of fungal repetitive DNA. The effectiveness of deRIP was such that functional assignments were improved quantitatively or qualitatively in nearly all cases. This was most clearly the case when repeat families were most clearly affected by RIP (Table 3, Table 4).

Conversely when the repeat family was not RIP-affected, the deRIP process was not able to improve the homology assignment. An example is the transposon repeat family R37 which had low RIP dominance (by majority consensus) of 0.25, indicating that R37 is not greatly affected by RIP mutation.

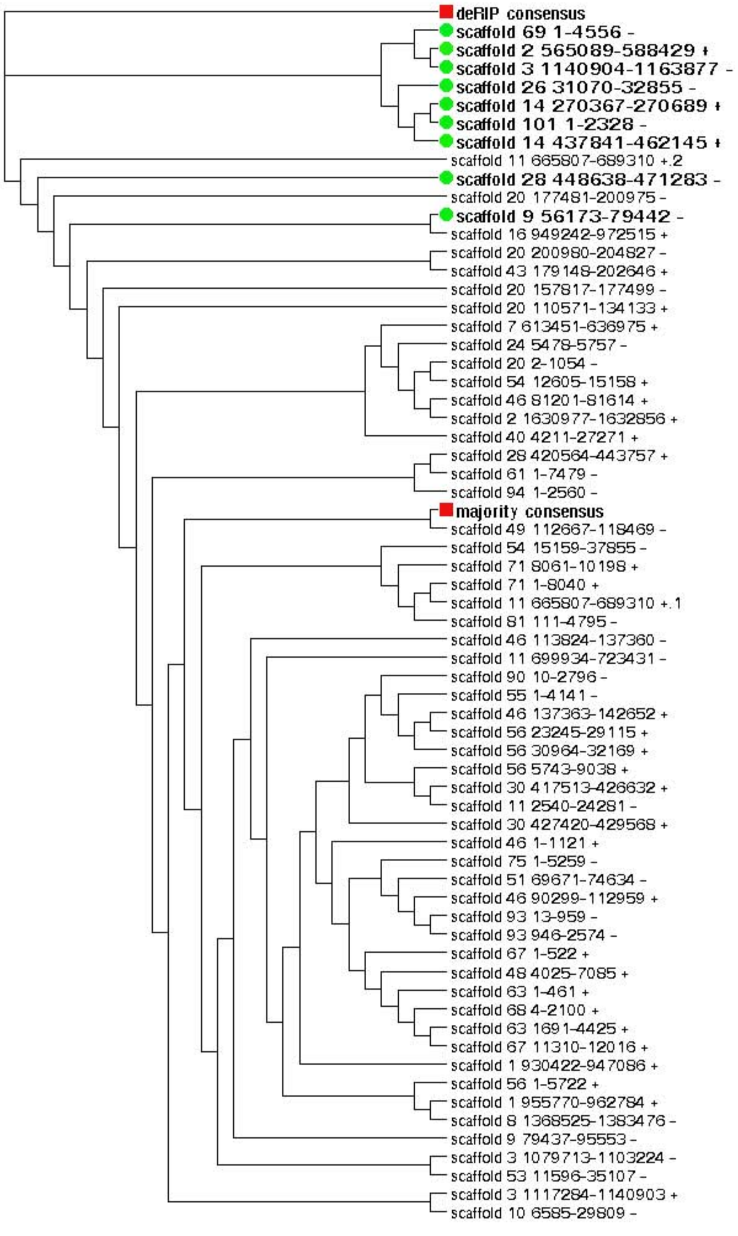

deRIP consensus
scaffold 69 1-4556 -
scaffold 2 565089-588429 +
scaffold 3 1140904-1163877 -
scaffold 26 31070-32855 -
scaffold 14 270367-270689 +
scaffold 101 1-2328 -
scaffold 14 437841-462145 +
scaffold 11 665807-689310 +.2
scaffold 28 448638-471283 -
scaffold 20 177481-200975 -
scaffold 9 56173-79442 -
scaffold 16 949242-972515 +
scaffold 20 200980-204827 -
scaffold 43 179148-202646 +
scaffold 20 157817-177499 -
scaffold 20 110571-134133 +
scaffold 7 613451-636975 +
scaffold 24 5478-5757 -
scaffold 20 2-1054 -
scaffold 54 12605-15158 +
scaffold 46 81201-81614 +
scaffold 2 1630977-1632856 +
scaffold 40 4211-27271 +
scaffold 28 420564-443757 +
scaffold 61 1-7479 -
scaffold 94 1-2560 -
majority consensus
scaffold 49 112667-118469 -
scaffold 54 15159-37855 -
scaffold 71 8061-10198 +
scaffold 71 1-8040 +
scaffold 11 665807-689310 +.1
scaffold 81 111-4795 -
scaffold 46 113824-137360 -
scaffold 11 699934-723431 -
scaffold 90 10-2796 -
scaffold 55 1-4141 -
scaffold 46 137363-142652 +
scaffold 56 23245-29115 +
scaffold 56 30964-32169 +
scaffold 56 5743-9038 +
scaffold 30 417513-426632 +
scaffold 11 2540-24281 -
scaffold 30 427420-429568 +
scaffold 46 1-1121 +
scaffold 75 1-5259 -
scaffold 51 69671-74634 -
scaffold 46 90299-112959 +
scaffold 93 13-959 -
scaffold 93 946-2574 -
scaffold 67 1-522 +
scaffold 48 4025-7085 +
scaffold 63 1-461 +
scaffold 68 4-2100 +
scaffold 63 1691-4425 +
scaffold 67 11310-12016 +
scaffold 1 930422-947086 +
scaffold 56 1-5722 +
scaffold 1 955770-962784 +
scaffold 8 1368525-1383476 -
scaffold 9 79437-95553 -
scaffold 3 1079713-1103224 -
scaffold 53 11596-35107 -
scaffold 3 1117284-1140903 +
scaffold 10 6585-29809 -

Figure 4 Comparison of the repeats of the X3X3R8 repeat family with respect to predicted gene content. The deRIP consensus (red square, top) is a prediction of the original repeat sequence prior to RIP-degradation. Nine X3X3R8 repeats contained predicted gene annotations (green circles, refer to Figure 3). All gene-annotation containing repeats were more closely related to the deRIP consensus than to the majority consensus (red square, middle).

The R10 repeat contained regions corresponding to three *S. nodorum* genes (SNOG_15997, SNOG_11270 and SNOG_16585 [NCBI: EAT76576.1, EAT81769.1, EAT76052.1]) (Table 4) which are located in separate regions of the genome assembly. The sub-telomeric repeat X26, which contained telomere-associated RecQ helicase sequence and was subject to relatively high levels of RIP mutation (Table 3). RecQ helicase plays a critical role in genome maintenance and is essential for DNA replication in eukaryotes [41]. Twenty six putative functional copies (i.e annotated gene models) of RecQ are present within the *S. nodorum* genome (Additional file 6). It is currently unclear by what mechanism function is preserved in certain copies of this highly repeated gene family, but not in others.

The repeat family X3X3R8, which replaced the previously defined repeat families X3 and R8, matched to a cluster of endogenous *S. nodorum* SN15 genes (Table 4) - some of these coding for a DNA repair helicase and ubiquitin conjugating enzyme (Figure 3). The helicase and ubiquitin conjugating enzyme genes within X3X3R8 were homologous to *Saccharomyces cerevisiae* proteins Rad5 [SGD: YLR032W] and Rad6 [SGD: YGL058W] respectively (Additional file 2). These proteins are involved in the post-replication repair of UV-damaged DNA, the epigenetic silencing of telomeres and sporulation in yeast [42-44].

The absence of active copies of the Elsa, Molly and Pixie families in the Australian SN15 isolate contrasts with the situation in the UK. Rawson screened several isolates for active transposons and used a trapping process to isolate active copies [35]. Using these transposons as probes showed great variation in copy number and band intensity amongst a collection of UK isolates. We have only looked at one Australian isolate, but it appears to be devoid of active transposons (Additional file 1). Consideration of the properties of RIP, the need for sexual reproduction by the fungus in Mediterranean climates and the biogeography of *S. nodorum*, could explain the absence of transposons in the Australian isolate. Repeated elements over a threshold size and above a threshold identity would be subject to RIP. This would inactivate all copies of a transposon during a meiotic event that appears to be necessary for survival over the hot summer [30]. It would not be conceivable for an active copy to be reconstituted in an asexual population derived from such an event. The survival of a transposon in a population of a sexually reproducing fungus would require mating with an isolate with an active (and presumably single) copy of the transposon. The invasion of *S. nodorum* into Australia most likely occurred via the propagation of a small founder population consistent with the reduced polymorphism of populations found here [33]. We speculate that no active transposons have survived within any Australian isolate capable of RIP. Screening of a larger population of Australian and Eurasian isolates to determine differences in frequency and distribution of active copies between the founder and derived populations would be required to confirm this.

Conclusions

In summary, we present a facile and rapid method to assist the annotation of repetitive elements of ascomycete genomes. The deRIP process can predict ancestral functional sequence from degraded repeat elements. Analysis of the repeat families of the fungal phytopathogen *Stagonospora nodorum* (strain SN15) using deRIP-converted sequences increased the number of recognisable repeat families from 65% (17/26) to 92% (23/25). This has enabled the characterization of many repeat families and has advanced our progress towards the goal of understanding and accounting for the evolutionary history of all regions of a genome.

Methods

Analysis of RIP-mutation of *S. nodorum* repetitive DNA

The 26 distinct repeat families of *S. nodorum* SN15 [19] were analysed for RIP mutation using RIPCAL [20]. RIPCAL requires an appropriate model sequence, which is a template to which all other aligned sequences of the repeat family are compared for RIP-like polymorphism. Previously we used the sequence of highest total G:C content as the model sequence. As RIP irreversibly converts G:C nucleotide pairs to A:T, it was assumed that the sequence with the highest G:C content was the least RIP-affected repeat in the family. In this study, we have performed RIPCAL analyses using 3 different models: highest G:C content, alignment majority consensus and predicted sequence of the repeat family prior to RIP-mutation.

Predicting the original repeat sequence prior to RIP-degradation

The deRIP process predicts the sequence of the pre-RIP-mutated version of the repeat alignment. Firstly, the majority consensus was generated by counting the nucleotide frequency at each position of the repeat alignment. The majority consensus sequence was determined by the highest frequency nucleotide. Secondly, at each position of the multiple alignment, counts of di-nucleotides exhibiting RIP-like polymorphism (CpN → TpN) were calculated (Table 1). A RIP mutation with the highest corresponding di-nucleotide count was presumed to be dominant and therefore the majority consensus was converted to the appropriate pre-RIP di-nucleotide sequence. This predicted sequence is henceforth referred to as the 'deRIP consensus'.

Validating the deRIP technique

The sequences of the active copies (which are presumably non-RIP-degraded) of the *S. nodorum* transposon repeats Molly, Pixie and Elsa [NCBI; AJ277966, AJ488502, AJ488503] [35] were compared to their respective majority and deRIP consensus sequences from strain SN15 via blastn [45]. Majority and deRIP consensus sequences of these repeat families were also globally aligned against their respective active copy via needle [40]. The SN15 rDNA repeat (Y1) was previously shown to be differentially susceptible to RIP [20]. The majority consensus of the non-RIP-affected copies of Y1 were compared to the majority and deRIP consensus sequences of the RIP-susceptible copies as above. The relative difference in alignment bit scores between majority and deRIP consensus sequences with their respective active copies was used to measure the degree of 'improvement' of the deRIP consensus over the majority consensus:

$$\frac{\text{Bit score of best HSP (deRIP consensus)}}{\text{Bit score of best HSP (majority consensus)}}$$

A 'deRIP improvement factor' greater than 1 indicated that the deRIP process had modified the RIP-affected sequence to resemble the sequence of the active copy.

Predicting the origin of RIP-degraded repeats

Majority and deRIP consensus sequences were compared to the NCBI NR protein database via blastx [45] and to the GIRI Repbase database of repetitive DNA [46] via tblastx. The results of these comparisons were used to infer repeat family origin and function. In this analysis, NCBI and Repbase sequences were both assumed to represent active transposons. Stronger deRIP matches to either database indicated that the deRIP algorithm was able to convert a RIP-inactivated sequence back into that of an active transposon. A maximum e-value threshold of 10 was imposed on hits against both the majority or deRIP consensus, with one of these also required to be less than 1e-3. DeRIP improvement factors were calculated for each hit as above. However for the purpose of summarising this data in Table 3, the maximum value was reported for each respective repeat family. 'Hit discovery scores' are the number of hits that the deRIP or majority consensus sequences have to the NR or GIRI databases. The scores illustrate the extent to which the deRIP process was able to discover new homology relationships that were previously lost due to RIP.

Additional material

Additional file 1: Test for viable copies of the transposons Molly, Pixie and Elsa in the *S. nodorum* SN15 genome.

Additional file 2: Summary of deRIP improvement and hit discovery scores. Contains summaries of the RIPCAL analyses for highest G:C content, majority consensus and deRIP consensus comparisons. Also contains details of majority and deRIP consensus hits by blastx to the NCBI NR Protein database and by tblastx to the GIRI Repbase database.

Additional file 3: Merging of the previously identified repeat families X3 and R8 to form the new repeat family X3X3R8.

Additional file 4: Merging of the previously identified repeat families X3 and R8 to form the new repeat family X3X3R8. Supplementary Figure, PNG format. The previously predicted X3 and R8 repeat families (HANE and OLIVER 2008) were found to correspond to genomic regions in a distinctive repeated pattern which spanned 26 kB. This region was classified as a new repeat family, X3X3R8, which supersedes the old repeat families R8 and X3. The MUMMER dot-plot above illustrates how the nucleotide majority consensus sequences of R8 and X3 relate to X3X3R8. The first third of the X3X3R8 majority consensus corresponds to a full length copy of X3. The second third of X3X3R8 is comprised of a second, incomplete copy of X3 which in matching regions is 10-20% divergent from the X3 consensus. The final third corresponds to a complete copy of the R8 repeat, in the reverse orientation with respect to its previously defined sequence.

Additional file 5: deRIP RIPCAL analysis of the repetitive DNA of *S. nodorum* SN15. RIPCAL outputs for highest G:C, consensus and deRIP models versus S. nodorum repeat families, tab-delimited txt and gif formats.

Additional file 6: List of predicted functional RecQ helicases in the *S. nodorum* genome.

Acknowledgements
This work was supported by the Grains Research and Development Corporation [grants: UMU00022 and GRS128]. Barbara Howlett and Thierry Rouxel were consulted during the writing of this manuscript.

Author details
[1]Faculty of Health Sciences, Murdoch University, Perth, Western Australia, 6150, Australia. [2]Department of Environment and Agriculture, Curtin University, Perth, Western Australia, 6102, Australia. [3]Current address: CSIRO Plant Industry, CELS Floreat, Perth, Western Australia, 6014, Australia.

Authors' contributions
JKH designed the deRIP algorithm and performed the bioinformatics analyses. JKH and RPO wrote the manuscript. All authors read and approved the final manuscript.

Received: 12 August 2010 Accepted: 24 November 2010
Published: 24 November 2010

References

1. Selker EU: **Premeiotic instability of repeated sequences in *Neurospora crassa*.** *Annual review of genetics* 1990, **24**:579-613.
2. Selker EU, Cambareri EB, Jensen BC, Haack KR: **Rearrangement of duplicated DNA in specialized cells of *Neurospora*.** *Cell* 1987, **51(5)**:741-752.
3. Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, Thon M, Kulkarni R, Xu JR, Pan H, *et al*: **The genome sequence of the rice blast fungus *Magnaporthe grisea*.** *Nature* 2005, **434(7036)**:980-986.
4. Ikeda K, Nakayashiki H, Kataoka T, Tamba H, Hashimoto Y, Tosa Y, Mayama S: **Repeat-induced point mutation (RIP) in *Magnaporthe grisea*:**

implications for its sexual cycle in the natural field context. *Molecular microbiology* 2002, **45(5)**:1355-1364.

5. Graia F, Lespinet O, Rimbault B, Dequard-Chablat M, Coppin E, Picard M: **Genome quality control: RIP (repeat-induced point mutation) comes to *Podospora*.** *Molecular microbiology* 2001, **40(3)**:586-595.
6. Idnurm A, Howlett BJ: **Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete *Leptosphaeria maculans*.** *Fungal Genet Biol* 2003, **39**:31-37.
7. Cuomo CA, Guldener U, Xu JR, Trail F, Turgeon BG, Di Pietro A, Walton JD, Ma LJ, Baker SE, Rep M, *et al*: **The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization.** *Science (New York, NY)* 2007, **317(5843)**:1400-1402.
8. Neuveglise C, Sarfati J, Latge JP, Paris S: **Afut1, a retrotransposon-like element from *Aspergillus fumigatus*.** *Nucleic acids research* 1996, **24(8)**:1428-1434.
9. Hua-Van A, Hericourt F, Capy P, Daboussi MJ, Langin T: **Three highly divergent subfamilies of the impala transposable element coexist in the genome of the fungus *Fusarium oxysporum*.** *Mol Gen Genet* 1998, **259(4)**:354-362.
10. Hua-Van A, Langin T, Daboussi MJ: **Evolutionary history of the impala transposon in *Fusarium oxysporum*.** *Molecular biology and evolution* 2001, **18(10)**:1959-1969.
11. Julien J, Poirier-Hamon S, Brygoo Y: **Foret1, a reverse transcriptase-like sequence in the filamentous fungus *Fusarium oxysporum*.** *Nucleic acids research* 1992, **20(15)**:3933-3937.
12. Nielsen ML, Hermansen TD, Aleksenko A: **A family of DNA repeats in *Aspergillus nidulans* has assimilated degenerated retrotransposons.** *Mol Genet Genomics* 2001, **265(5)**:883-887.
13. Bhat A, Tamuli R, Kasbekar DP: **Genetic transformation of *Neurospora tetrasperma*, demonstration of repeat-induced point mutation (RIP) in self-crosses and a screen for recessive RIP-defective mutants.** *Genetics* 2004, **167(3)**:1155-1164.
14. Hood ME, Katawczik M, Giraud T: **Repeat-induced point mutation and the population structure of transposable elements in *Microbotryum violaceum*.** *Genetics* 2005, **170(3)**:1081-1089.
15. Montiel MD, Lee HA, Archer DB: **Evidence of RIP (repeat-induced point mutation) in transposase sequences of *Aspergillus oryzae*.** *Fungal Genet Biol* 2006, **43(6)**:439-445.
16. Farman ML: **Telomeres in the rice blast fungus *Magnaporthe oryzae*: the world of the end as we know it.** *FEMS microbiology letters* 2007, **273(2)**:125-132.
17. Crouch JA, Glasheen BM, Giunta MA, Clarke BB, Hillman BI: **The evolution of transposon repeat-induced point mutation in the genome of *Colletotrichum cereale*: reconciling sex, recombination and homoplasy in an "asexual" pathogen.** *Fungal Genet Biol* 2008, **45(3)**:190-206.
18. Braumann I, van den Berg M, Kempken F: **Repeat induced point mutation in two asexual fungi, *Aspergillus niger* and *Penicillium chrysogenum*.** *Current genetics* 2008, **53(5)**:287-297.
19. Hane JK, Lowe RG, Solomon PS, Tan KC, Schoch CL, Spatafora JW, Crous PW, Kodira C, Birren BW, Galagan JE, *et al*: **Dothideomycete plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*.** *The Plant cell* 2007, **19(11)**:3347-3368.
20. Hane JK, Oliver RP: **RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences.** *BMC bioinformatics* 2008, **9**:478.
21. Cambareri EB, Jensen BC, Schabtach E, Selker EU: **Repeat-induced G-C to A-T mutations in *Neurospora*.** *Science (New York, NY)* 1989, **244(4912)**:1571-1575.
22. Galagan JE, Selker EU: **RIP: the evolutionary cost of genome defense.** *Trends Genet* 2004, **20(9)**:417-423.
23. Cambareri EB, Singer MJ, Selker EU: **Recurrence of repeat-induced point mutation (RIP) in *Neurospora crassa*.** *Genetics* 1991, **127(4)**:699-710.
24. Watters MK, Randall TA, Margolin BS, Selker EU, Stadler DR: **Action of repeat-induced point mutation on both strands of a duplex and on tandem duplications of various sizes in *Neurospora*.** *Genetics* 1999, **153(2)**:705-714.
25. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, *et al*: **The genome sequence of the filamentous fungus *Neurospora crassa*.** *Nature* 2003, **422(6934)**:859-868.
26. Margolin BS, Garrett-Engele PW, Stevens JN, Fritz DY, Garrett-Engele C, Metzenberg RL, Selker EU: **A methylated *Neurospora* 5 S rRNA pseudogene contains a transposable element inactivated by repeat-induced point mutation.** *Genetics* 1998, **149(4)**:1787-1797.
27. Schoch CL, Shoemaker RA, Seifert KA, Hambleton S, Spatafora JW, Crous PW: **A multigene phylogeny of the Dothideomycetes using four nuclear loci.** *Mycologia* 2006, **98(6)**:1041-1052.
28. Solomon PS, Lowe RG, Tan KC, Waters OD, Oliver RP: ***Stagonospora nodorum*: cause of stagonospora nodorum blotch of wheat.** *Molecular plant pathology* 2006, **7(3)**:147-156.
29. Friesen TL, Faris JD, Solomon PS, Oliver RP: **Host-specific toxins: effectors of necrotrophic pathogenicity.** *Cellular microbiology* 2008, **10(7)**:1421-1428.
30. Solomon PS, Parker K, Loughman R, Oliver RP: **Both mating types of *Phaeosphaeria* (anamorph *Stagonospora*) *nodorum* are present in Western Australia.** *Eur J Plant Pathol* 2004, **110**:763-766.
31. Cooley RN, Caten CE: **Variation in electrophoretic karyotype between strains of *Septoria nodorum*.** *Mol Gen Genet* 1991, **228**:17-23.
32. Keller SM, McDermott JM, Pettway RE, Wolfe MS, McDonald BA: **Gene flow and sexual reproduction in the wheat glume blotch pathogen *Phaeosphaeria nodorum* (anamorph *Stagonospora nodorum*).** *Phytopathology* 1997, **87(3)**:353-358.
33. Stukenbrock EH, Banke S, McDonald BA: **Global migration patterns in the fungal wheat pathogen *Phaeosphaeria nodorum*.** *Molecular ecology* 2006, **15(10)**:2895-2904.
34. Perkins DD, Metzenberg RL, Raju NB, Selker EU, Barry EG: **Reversal of a *Neurospora* translocation by crossing over involving displaced rDNA, and methylation of the rDNA segments that result from recombination.** *Genetics* 1986, **114(3)**:791-817.
35. Rawson JM: **PhD Thesis: Transposable elements in the phytopathogenic fungus *Stagonospora nodorum*.** Birmingham: University of Birmingham; 2000.
36. Cove DJ: **Chlorate toxicity in *Aspergillus nidulans*: the selection and characterisation of chlorate resistant mutants.** *Heredity* 1976, **36(2)**:191-203.
37. Cutler SB, Cooley RN, Caten CE: **Cloning of the nitrate reductase gene of *Stagonospora* (*Septoria*) *nodorum* and its use as a selectable marker for targeted transformation.** *Current genetics* 1998, **34(2)**:128-137.
38. Kempken F, Kuck U: **Transposons in filamentous fungi–facts and perspectives.** *Bioessays* 1998, **20(8)**:652-659.
39. Stamatakis A, Hoover P, Rougemont J: **A Rapid Bootstrap Algorithm for the RAxML Web-Servers.** *Systematic Biology* 2008, **75(5)**:758-771.
40. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite.** *Trends Genet* 2000, **16(6)**:276-277.
41. Killoran MP, Keck JL: **Sit down, relax and unwind: structural insights into RecQ helicase mechanisms.** *Nucleic acids research* 2006, **34(15)**:4098-4105.
42. Gangavarapu V, Haracska L, Unk I, Johnson RE, Prakash S, Prakash L: **Mms2-Ubc13-dependent and -independent roles of Rad5 ubiquitin ligase in postreplication repair and translesion DNA synthesis in *Saccharomyces cerevisiae*.** *Molecular and cellular biology* 2006, **26(20)**:7783-7790.
43. Jentsch S, McGrath JP, Varshavsky A: **The yeast DNA repair gene *RAD6* encodes a ubiquitin-conjugating enzyme.** *Nature* 1987, **329(6135)**:131-134.
44. Torres-Ramos CA, Prakash S, Prakash L: **Requirement of *RAD5* and *MMS2* for postreplication repair of UV-damaged DNA in *Saccharomyces cerevisiae*.** *Molecular and cellular biology* 2002, **22(7)**:2419-2426.
45. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool.** *Journal of molecular biology* 1990, **215(3)**:403-410.
46. Jurka J, Kapitonov VV, Pavlicek A, Klonowski P, Kohany O, Walichiewicz J: **Repbase Update, a database of eukaryotic repetitive elements.** *Cytogenetic and genome research* 2005, **110(1-4)**:462-467.

doi:10.1186/1471-2164-11-655
Cite this article as: Hane and Oliver: **In silico reversal of repeat-induced point mutation (RIP) identifies the origins of repeat families and uncovers obscured duplicated genes.** *BMC Genomics* 2010 **11**:655.

Appendix 4A: Response to Thesis Examination Comments

The Derip algorithm is not explained very clearly

The deRIP is explained in the deRIP publication to the level that the journal reviewers and editors deemed appropriate for the target audience. Details omitted from this publication can be found in the RIPCAL user manual (Appendix 3A) or in this section (below).

It appears that this process is too aggressive because rare, non-RIP mutations will also be reversed by this process. Moreover, should a sequence that has been mutated entirely through other processes make its way into the alignment, this could elicit illegitimate reversals. Is there some kind of kind of "filter" to remove such sequences from the alignment? It would have been useful had James discussed the predicted frequency of illegitimate reversals using his procedure. For example, he could have considered how many non-RIP mutations were present, on average, in a RIP'd alignment. Using my own RP analysis scripts to investigate genome-wide RIP in Stagonospora, I note that in any given alignment, approximately 10 non-RIP mutations accompany every 400 RIp-type mutations. Thus, we can expect that approximately 10 of the RIP-type mutations are also not actually caused by RIP but by other mutational mechanisms. Under the current implementation of the DeRIP script, some of these would be "reversed", as long as they occurred in a CpA context. With possibly tens of alignments for any given repeat, it seems to me that there is a great danger of over-DeRIP'ping. Would it not have ben more judicious to require that a certain number of repeat copies possess the ancestral non-RIP'd base at a given position? The number of instances of ancestral bases required before enacting deRIP could easily have been included as an option in the script.

The examiner seems to imply that any mutation in an alignment would be incorporated into the 'deRIPped consensus'. This is not the case at all, instead what deRIP actually does is quite similar to the examiner's suggestion of requiring “a certain number of repeat copies posses the ancestral non-RIP'd base at any given position”. At each site of an alignment, deRIP tallies both RIP target and product dinucleotides for each type of CpN mutation. The RIP-target dincleotide for the CpN mutation (NOT the individual dinucleotides) with the highest frequency is then substituted into the consensus sequence. This tally must also be above a threshold count for the substitution to occur (details of modifying this threshold value are contained in the RIPCAL manual, Appendix 3A).

While it is still true that non-RIP mutations are occasionally incorporated, these will only occur if the following conditions are met:

1) the falsely introduced non-RIP-like polymorphism is the predicted RIP-target of a CpN mutation detected at high abundance, or
2) the falsely introduced non-RIP-like polymorphism is not recognised as a CpN mutation but is the most abundant nucleotide at its position in the alignment i.e. the majority consensus sequence

This does not introduce a high rate of false mutations as is indicated by the relatively strong alignment statistics of deRIPped sequences presented in chapter 4: Table 2 and Additional File 2. Discussion of false-positive rates are in my opinion inappropriate for a tool such as deRIP, a method which is highly dependent upon the sequence diversity of a repeat family, actual distribution and extent of RIP, for its success. For certain repeat families that have been too highly RIP-affected it is not likely to work well at all. This is something that investigators should be aware of when considering the appropriateness of the tool for their analyses. Having not seen the examiner's own script code for RIP calculation makes it impossible to ascertain the validity of what he defines as an illegitimate reversal. However if we accept his method, 10 out of 410 predicted reversals were incorrect in the published analyses of *P. nodorum*. This comes to a 2.5% error rate, which to my mind indicates that RIPCAL has performed well.

In the background statement for the "in silico reversal" paper,, James states that RIP is a genome defense mechanism that guards against transposon invasion. This is not true. If the purpose of RIP is to suppress transposon activity, then it has been spectacularly unsuccessful. Most of the fungi that show evidence of RIP have extremely large families of repeated sequences. Thus, transposons have been massively amplified within fungal genomes in spite of RIP.

I propose a common-sense explanation for a strong CpA specificity being under strong selective pressure, which to my knowledge has not been adequately stressed in the literature. During the course of protein translation, actively replicating transposons draw upon a limited pool of intracellular resources. At the same time fungi rely upon this resource pool to survive and compete with other organisms. On face value, RIP does indeed appear to do a poor job of protecting the fungal genome from tranposons. High proportions of repetitive contents have been reported in a number of sequenced fungal genomes, so we can infer that repeat invasion in and of itself is not a major impediment to survival. In chapter 8 we observe that genomes are highly rearranged, so presumably sequence rearrangements that arise due to repetitive DNA have little or no deleterious effects upon genome viability. Nevertheless, while a fungal genome may be able to tolerate (and even benefit from the presence of repetitive elements (Van de Wouw et al. 2010), inactivating these transposons would prevent them from draining proteomic resources which could otherwise be reserved for the translation of endogenous genes. An effective way of doing this would be to introduce stop codons into transposase gene-coding regions. The analysis of stop-codon densities of transposon repeats of *P. nodorum* strain SN15 (Chapter 3, Additional File 1) showed that all copies of the highly RIP-affected Molly, Pixie and Elsa transposons had high stop codon densities. This suggests the RIP-mediated interruption of transposase open-reading frames (ORFs). On a related note, RIP-mediated ORF-silencing explains the observed prevalence of a CpA mutation bias in most fungal species analysed thus far. The four possible CpN dinucleotide mutations are listed in the table below, next to the stop codons that they can potentially introduce. From this table, assuming RIP sites are chosen at random cytosines, CpA mutation would have the highest probability of introducing a stop-codon and thus would more efficiently prevent the replication of transposons. In conclusion, RIP is an effective means of protection against the deleterious effects of transposon invasion.

Fwd mutation	Rev mutation	Potential Fwd Stops	Potential Rev Stops
CpA → TpA	**TpG → TpA**	**TAA, TAG**	**TAA, TAG**
CpC → TpC	GpG → GpA		TGA
CpG → TpG	CpG → CpA	TGA	
CpT → TpT	ApG → ApA		TAA

I don't understand why DeRIP was necessary to classify repeat family X26 as a telomere-linked helicase. We identified the telomere-linked helicase in Stagonospora several years ago, simply through blastx searches - no DeRIP'ing required! It is possible that the appropriate query hadn't been tried before DeRIP was employed, or perhaps different parameters were tried?

Sequences of the X26 repeat family were compared to the NCBI NR protein database via blastx in the publication Hane et al 2007 (Chapter 2: Supplementary dataset 1). As the examiner did not elaborate upon his blastx parameters, this comment assumes that we have both used the defaults. The published blastx comparison for X26 resulted in hits to a helicase of *Aspergillus fumigatus* Af293 [PROTEIN ACCESSION: EAL89306.1] with e-values from 3.4 e -5 to 5.7 e-8. However these hits were not unambiguous, as there were also numerous hits of similar significance to Frigida-like proteins of the legume *Medicago truncatula*, P-553 of *Borrelia hermsii,* as well as other transposase, chemotaxis transducer and unidentified proteins. In hindsight after the application of deRIP we can see that the transposase protein was the correct hit, but it would have been incorrrect to annotate the X26 repeat family prior to deRIP based on the evidence available at that time.

Another way to benchmark the accuracy of DeRIP'ping is to DeRIP the whole genome and look at LTR retrotransposon copies whole LTR sequences had diverged due to RIP. Accuracy can be assessed based on how often the LTR sequences have been restored to identity.

This is essentially the control experiments using the Molly, Pixie and Elsa transposons presented in the publication. For each of these transposons, an active copy has been sequenced. It is to these active sequences that the deRIP consensus sequences have been compared. To a lesser extent, the rDNA repeat also serves as a control experiment as given what is currently known in the literature certain copies of this repeat (arranged in a tandem array and localised to the nucleolus organiser region (NOR)) are unlikely to have been affected by RIP.

Page 80 paragraph 2, states that the maximum sequence identity between the active transposon copies in the UK isolates versus the SN15 genome was 66% by blastx. should this not have been blastn? If blastx were required to align these sequences then i doubt that deRIP would even be possible. Moreover, Figure 2 shows a blastn alignment.

Sequence identity for amino acid alignments will always look unimpressive compared to nucleotide identity values, as for most amino acids there are a handful of alternatives with similar properties which can be substituted without major effect on overall protein function. An identity of 66% by blastx is in my opinion actually quite reasonable.

The use of blastx rather than blastn may seem counter-intuitive but concerns expediency of the analysis and the current availability of repetitive DNA resources. Ideally, we would have access to a complete curated database of repeat sequences, pristine and untouched by RIP. We could use this to determine whether a deRIPping a repeat had improved the match to the pre-RIP version by blastn. Unfortunately we do not have such a resource for the majority of identified repeat famillies. Although there are some online repositories of fungal repetitive DNA such as those in REPbase and within the NCBI nucleotide database, not enough work has been done (if any) to annotate their RIP status. Because of this, we are for the most part unaware of which repeats to use as a benchmark. The NOR-localised rDNA, Molly, Pixie, Elsa and other repeats used as validation controls in the RIPCAL paper are exceptions. On the other hand, presence of a sequence in a protein database implies that the underlying nucleotide region is non-RIP-affected since it contains un-interrupted open-reading frames (see above discussion of stop-codon insertion by RIP). Hence without a great deal of time to devote to curation of repeat-databases, we can safely use protein sequences as a benchmark for the improvement of deRIPped repeat matches via blastx.

Figure 2(B) shows neither a blastx or blastn alignment, which are pairwise local alignment algorithms, but a multiple global alignment alignment generated by clustalw. The examiner correctly points out however that Figure 2 presented nucleotide-based alignments, not amino-acid based ones such as blastx. To clarify, blastx was not used in the generation of the deRIP consensus, but was used to evaluate the improvement in sequence matches to protein databases by the deRIP consensus compared to the majority consensus of a repeat family. In addition to the reasons for using blastx outlined above, blastx was used as t

1) There was no guarantee that the active sequences and the sequences of copies that invaded strain SN15 prior to RIP were identical as there is no lineage data for the various strains involved. If deRIP had perfectly reverted an SN15 transposon but this did not match the active sequence, this would also introduce errors into the analysis.
2) Of primary concern is whether deRIP is able to restore open-reading frames to the point where protein hits can be detected in a sequence similarity search (i.e. via blastx).
3) Nucleotide mutations, either stemming from RIP or other mechanisms, may result in amino-acid translations with identical or functionally equivalent properties to the active copy.

The section of the text immediately to which the examiner refers references additional file 1, not figure 2. This shows the results of an experiment in which the "viability" of transposons is estimated based on amino-acid homology to active transposon proteins. The reason for this, in addition to those immediately above, is that the transposase protein is the major determinant of the successful excision/duplication of the mobile element. Additionally, this experiment tested for the presence of stop codons within the blastx alignments, and completeness of coverage of the full length active proteins by blastx alignments. This indicated in all cases that genomic copies of Molly, Pixie and Elsa in *P. nodorum* SN15 encoded truncated protein products. Thus in all likelihood, the replication of these mobile elements has been inactivated.

Chapter 5: Attribution Statement

Title:	**Evolution of linked avirulence effectors in *Leptosphaeria maculans* is affected by genomic environment and exposure to resistance genes in host plants.**
Authors:	Van de Wouw A. P., Cozijnsen A. J., **Hane J. K.**, Brunner P. C., McDonald B. A., Oliver R. P., Howlett B. J.
Citation:	*PLOS Pathogens* 6(11):e1001180 (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 (JKH) made the following contributions to this chapter:

- Carried out and interpreted bioinformatic analyses of repeat-induced point mutations.

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

-- --

James Hane (Ph. D. candidate) Date

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

-- --

Richard Oliver (Principal supervisor) Date

Evolution of Linked Avirulence Effectors in *Leptosphaeria maculans* Is Affected by Genomic Environment and Exposure to Resistance Genes in Host Plants

Angela P. Van de Wouw[1], Anton J. Cozijnsen[1], James K. Hane[2], Patrick C. Brunner[3], Bruce A. McDonald[3], Richard P. Oliver[2], Barbara J. Howlett[1]*

1 School of Botany, the University of Melbourne, Victoria, Australia, **2** Australian Centre for Necrotrophic Fungal Pathogens, Curtin University, Bentley, Western Australia, Australia, **3** Plant Pathology Group, Institute of Integrative Biology, ETH Zurich, Zurich, Switzerland

Abstract

Brassica napus (canola) cultivars and isolates of the blackleg fungus, *Leptosphaeria maculans* interact in a 'gene for gene' manner whereby plant resistance (*R*) genes are complementary to pathogen avirulence (*Avr*) genes. Avirulence genes encode proteins that belong to a class of pathogen molecules known as effectors, which includes small secreted proteins that play a role in disease. In Australia in 2003 canola cultivars with the *Rlm1* resistance gene suffered a breakdown of disease resistance, resulting in severe yield losses. This was associated with a large increase in the frequency of virulence alleles of the complementary avirulence gene, *AvrLm1*, in fungal populations. Surprisingly, the frequency of virulence alleles of *AvrLm6* (complementary to *Rlm6*) also increased dramatically, even though the cultivars did not contain *Rlm6*. In the *L. maculans* genome, *AvrLm1* and *AvrLm6* are linked along with five other genes in a region interspersed with transposable elements that have been degenerated by Repeat-Induced Point (RIP) mutations. Analyses of 295 Australian isolates showed deletions, RIP mutations and/or non-RIP derived amino acid substitutions in the predicted proteins encoded by these seven genes. The degree of RIP mutations within single copy sequences in this region was proportional to their proximity to the degenerated transposable elements. The RIP alleles were monophyletic and were present only in isolates collected after resistance conferred by *Rlm1* broke down, whereas deletion alleles belonged to several polyphyletic lineages and were present before and after the resistance breakdown. Thus, genomic environment and exposure to resistance genes in *B. napus* has affected the evolution of these linked avirulence genes in *L. maculans*.

Citation: Van de Wouw AP, Cozijnsen AJ, Hane JK, Brunner PC, McDonald BA, 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(11): e1001180. doi:10.1371/journal.ppat.1001180

Editor: Anita Sil, University of California San Francisco, United States of America

Received March 23, 2010; **Accepted** October 6, 2010; **Published** November 4, 2010

Funding: This research was supported by the Grains Research and Development Corporation, Australia. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: bhowlett@unimelb.edu.au

Introduction

The fungus *Leptosphaeria maculans* causes blackleg (phoma stem canker) and is the major disease of *Brassica napus* (canola) worldwide [1]. The major source of inoculum is wind-borne ascospores, which are released from sexual fruiting bodies on infected stubble (crop residue) of previous crops and can be transmitted several kilometres. This fungus has a 'gene for gene' interaction with its host (canola) such that pathogen avirulence alleles render the pathogen unable to attack host genotypes with the corresponding resistance genes [2].

Twelve genes conferring resistance to *L. maculans* (*Rlm1-9, LepR1-3*) have been identified from *Brassica* species [3,4]. Nine of these genes have been mapped but none have yet been cloned [5,6]. Of the corresponding avirulence genes in *L. maculans*, seven have been mapped to two gene clusters, *AvrLm1-2-6* and *AvrLm3-4-7-9*, located on separate *L. maculans* chromosomes [7]. Three genes, *AvrLm1*, *AvrLm6* and *AvrLm4-7*, have been cloned and characterised [8,9,10]. Avirulence proteins belong to a class of molecules called effectors. Effectors are small molecules or proteins produced by the pathogen that alter host-cell structure and function, facilitate infection (for example, toxins) and/or induce defence responses (for example, avirulence proteins), and are generally essential for disease progression [11]. Many effectors are small, secreted proteins (SSPs), which are cysteine-rich, share no sequence similarity with genes from other species, are often highly polymorphic between isolates of a single species and are expressed highly *in planta* [12]. The AvrLm1, AvrLm4-7 and AvrLm6 proteins of *L. maculans* are effector molecules with an avirulence function. *AvrLm6* and *AvrLm4-7* encode SSPs with six and eight cysteine residues, respectively, whilst the *AvrLm1* protein, which is also a SSP, has only one cysteine [8,9,10].

Leptosphaeria maculans undergoes sexual recombination prolifically and populations can rapidly adapt to selection pressures imposed by the host such as exposure to resistance conferred by single or major genes. This situation increases the frequency of virulent isolates and can cause resistance to break down, often resulting in severe yield losses [13]. Three examples of this

Author Summary

The fungus *Leptosphaeria maculans* causes blackleg, the major disease of canola worldwide. Populations of this fungus rapidly adapt to selection pressures such as the extensive sowing of canola with particular disease resistance genes. This can lead to a breakdown of resistance and severe economic losses. We describe mutations in key fungal genes involved in the interaction with canola, and we report the first large scale study of evolutionary processes affecting such genes in any fungal plant pathogen. We relate these changes to the genomic environment of these genes and to the breakdown of disease resistance in canola.

breakdown are discussed below, the most dramatic one occurring in Australia in 2003.

Prior to 2000 the Australian canola industry relied on 'polygenic' cultivars with multiple resistance genes. Yield losses were generally low. In 2000, cultivars with major gene resistance, termed 'sylvestris' resistance, were released commercially and grown extensively in some areas of Australia. These cultivars were derived from a synthetic *B. napus* line produced by crossing the two progenitor species, *B. oleracea* subsp. *alboglabra* and an accession of *B. rapa* subsp. *sylvestris* that had a high level of resistance to *L. maculans* [14]. For several years these cultivars showed little or no disease but in 2003, resistance failed, resulting in up to 90% yield losses in the Eyre Peninsula, South Australia, costing the industry between $5–10 million AUD [15,16]. These cultivars contained resistance genes *Rlm1* and *RlmS*, suggesting that both sources of resistance failed simultaneously [17]. After 2004, these cultivars were withdrawn from sale although they were still grown in yield trial sites around Australia. A similar but less dramatic situation occurred in France when resistance conferred by *Rlm1* was rendered ineffective within five years of commercial release of *Rlm1*-containing cultivars [13,18]. Another breakdown of resistance was observed in field trial experiments in France that had been designed to assess the durability of a resistance gene *Rlm6*. In these experiments *B. napus* lines containing *Rlm6* were sown into *L. maculans*-infected stubble of a *Rlm6*–containing line over a four year period [19]. After three years of this contrived selection, the frequency of virulence in fungal populations towards *Rlm6* was so high that this resistance was rendered ineffective and the lines suffered extremely high levels of disease.

The three avirulence genes, *AvrLm1*, *AvrLm6* and *AvrLm4-7*, cloned from *L. maculans* are located within AT-rich, gene-poor regions that are riddled with degenerated copies of transposable elements [8,9,10]. These transposable elements appear to have been inactivated via repeat-induced point (RIP) mutations [20]. RIP is an ascomycete-specific process that alters the sequence of multicopy DNA. Nucleotide changes CpA to TpA and TpG to TpA are conferred during meiosis, often generating stop codons, thereby inactivating genes [21]. Additionally, RIP mutations have been inferred bioinformatically in various transposable elements throughout the *L. maculans* genome and in *AvrLm6* [22]. *AvrLm1* and *AvrLm6* are genetically linked and different types of mutations leading to virulence have been reported. The entire *AvrLm1* locus was deleted in 285 of 290 (98%) isolates that were virulent towards *Rlm1* [20]. Fudal et al. characterised the *AvrLm6* locus in a different set of 105 isolates, most of which were cultured from *Rlm6*–containing lines during the field trial in France described above [22,23]. Deletions and RIP were responsible for virulence in 45 (66%) and 17 (24%) isolates, respectively, that were virulent towards the *Rlm6* gene [22].

In this paper we relate changes in the types and frequencies of mutations in genes including *AvrLm1* and *AvrLm6* to the selection pressure imposed by extensive regional sowing of *B. napus* cultivars with sylvestris resistance.

Results

Virulence of Australian isolates of *L. maculans* and analysis of mutations at *AvrLm1* and *AvrLm6*

In preliminary experiments to see if the breakdown of 'sylvestris resistance' in *B. napus* seen in the field [15] correlated with changes in frequency of virulence towards *Rlm1* in individual *L. maculans* isolates, 11 isolates collected prior to the breakdown (before 2004) and 12 isolates collected after the breakdown (2004 onwards) were screened for virulence on *B. napus* cultivars Q2 (with *Rlm3*) and Columbus (*Rlm1, Rlm3*). Cotyledons of 14 day old plants were inoculated with individual isolates and symptoms were scored 17 days later. All isolates were virulent on the susceptible control, cv. Q2. Ten of the 11 isolates collected prior to 2004 were avirulent on the *Rlm1*-containing cultivar, Columbus. However, seven of ten isolates collected from 2004 onwards were virulent towards *Rlm1* (Table 1). A subset of these isolates was inoculated onto the *B. napus* cultivar Aurea (*Rlm6*). Of seven of the 11 isolates collected prior to 2004, only one was virulent towards *Rlm6*. However, of 12 isolates collected from 2004 onwards, eight were virulent towards *Rlm6* (Table 1). These results suggested that there was a significant change in frequencies of virulent alleles of *AvrLm1* and *AvrLm6* associated with breakdown of 'sylvestris resistance'. These isolates were then genotyped at the *AvrLm1*, *AvrLm6* and mating type loci using PCR-based screening [20,22,24] and as expected, the virulence phenotypes corresponded with *avrLm1* and /or *avrLm6* genotypes (Table 1).

Changes in allele frequencies were then examined in a total of 295 Australian isolates. Of these 137 were collected between 1987 and 2003, prior to the breakdown of sylvestris resistance, whilst the remaining 158 isolates were collected between 2004 and 2008, after the resistance breakdown (Table S1). These isolates were collected from stubble of a range of canola cultivars with different resistance genes. One third of the isolates had a deletion of *AvrLm1* (Table 2), whilst 63% had the allele of isolate v23.1.3, whose genome has been sequenced. Alleles of this isolate hereafter are referred to as wild type alleles (e.g. *AvrLm1-0*). As expected, isolates with *AvrLm1-0* were avirulent towards *Rlm1* (Tables 1 and 2). The remaining eight isolates comprised four alleles with coding sequence changes conferring non-synonymous substitutions ($I^{125}K$ and/or $H^{155}Y$). Isolates harbouring these alleles were avirulent on the *Rlm1*- containing *B. napus* line (Table 1). Thirteen alleles of *AvrLm6* were detected (Table 2). *AvrLm6* was deleted in 20% of isolates, thus conferring virulence towards *Rlm6*, whilst 24% had the allele of the sequenced isolate and conferred avirulence towards *Rlm6* (Tables 1 and 2). Other isolates virulent towards the *Rlm6*-containing cultivar had alleles with stop codons conferred by base changes reminiscent of RIP mutation. Accordingly, allele sequences were analysed by RIPCAL, a software tool that visualises the physical distribution of RIP mutation and reports a RIP dominance score indicating the degree of RIP in each sequence. Sequences with RIP dominance scores >1 are highly RIP-affected having a high proportion of CpA to TpA, or TpG to TpC RIP mutations [25]. RIPCAL analysis did not detect RIP in any *AvrLm1* alleles, but detected RIP in seven *AvrLm6* alleles. These RIP mutations resulted in numerous non-synonymous changes as well as premature stop codons (between 4 and 6), which were within all RIP-affected alleles (Table 2). Southern hybridisation results suggest that there is only a single

Table 1. Pathogenicity of Australian isolates of *Leptosphaeria maculans* on cotyledons of *Brassica napus* cultivars Q2 and Columbus and *B. juncea* cv. Aurea.

Isolate	Cultivar						Genotype[a]	
	Q2 (*Rlm3*)		Columbus (*Rlm1, Rlm3*)		Aurea (*Rlm5, Rlm6*)			
	Path. score	Phenotype	Path. score	Phenotype	Path. score	Phenotype	*AvrLm1* allele	*AvrLm6* allele
Pre sylvestris breakdown (before 2004)								
LM535	6.5	Vir	2.6	Avir	Not tested		0	0
1317	6.4	Vir	1.9	Avir	Not tested		0	0
LM752	4.5	Vir	1.2	Avir	1.5	Avir	0	1
LM526	6.5	Vir	1.4	Avir	2.3	Avir	0	1
LM641	6.6	Vir	1.4	Avir	Not tested		0	1
LM749	6.5	Vir	1.8	Avir	1.0	Avir	0	1
GA2	7.0	Vir	1.5	Avir	1.6	Avir	0	3
IBCN18	7.0	Vir	2.3	Avir	6.7	Vir	1	del
V4	7.1	Vir	1.3	Avir	Not tested		1	del
LM691	7.0	Vir	1.1	Avir	1.2	Avir	3	0
IBCN17	6.7	Vir	6.6	Vir	1.6	Avir	del	0
Post sylvestris breakdown (2004 onwards)								
04P017	6.2	Vir	1.3	Avir	1.5	Avir	0	1
04P042	5.7	Vir	1.2	Avir	1.3	Avir	0	1
04S005	7.0	Vir	7.8	Vir	1.8	Avir	del	0
04S014	7.0	Vir	7.2	Vir	1.6	Avir	del	1
05P032	7.0	Vir	6.5	Vir	6.4	Vir	del	2
05P033	6.5	Vir	6.7	Vir	6.0	Vir	del	2
05P034	7.0	Vir	6.7	Vir	5.6	Vir	del	2
06S013	4.8	Vir	7.1	Vir	4.1	Vir	del	6
06S039	5.8	Vir	6.9	Vir	4.6	Vir	del	8
06P042	7.0	Vir	1.6	Avir	3.6	Vir	0	9
06P039	7.0	Vir	1.1	Avir	4.1	Vir	0	9
06P040	7.0	Vir	1.3	Avir	4.8	Vir	1	*del*

Cotyledons of 14 day old seedlings were infected with spores of individual isolates. Mean pathogenicity scores (Path. score) were determined by assessing 40 inoculation sites at 17 dpi. Isolates with Path. scores $\leq$3.9 are classified as avirulent whilst those with Path. scores $\geq$4.0 are classified as virulent.

[a]Alleles as described in Table 2.

doi:10.1371/journal.ppat.1001180.t001

copy of the *AvrLm6* locus within isolates harbouring the RIP alleles (Figure S1). The remaining alleles of *AvrLm6* (*AvrLm6-1, -2,-3* and *-4*) harboured single or few nucleotide changes leading to synonymous or amino acid substitutions (G^{123}C, K^{127}E, F^{54}L) compared to isolate v23.1.3 (*AvrLm6-0*). These amino acid substitutions were generated via non-RIP like mutations (henceforth referred to as non-RIP amino acid substitutions). Isolates harbouring *AvrLm6-1, AvrLm6-3* or *AvrLm6-4* alleles were avirulent towards cv. Aurea, whilst the four isolates harbouring the *AvrLm6-2* allele (G^{123}C) were virulent (Table 1).

No RIP-affected alleles were present in isolates collected prior to the breakdown of sylvestris resistance. However, the 158 isolates collected after the breakdown had seven *AvrLm6* RIP alleles (frequency of 8.9%) and there was a very large increase in the frequency of deletion alleles of *AvrLm1* (22.6 to 41.8%) and *AvrLm6* (4.4 to 38.7). This was a nine-fold increase in frequency of *avrLm6* (Table S2). Thirty four (11.5%) isolates had deletions of both *AvrLm1* and *AvrLm6* and only one of these isolates was collected prior to the breakdown. All 295 isolates were grouped into four genotypic classes (*AvrLm1, AvrLm6*; *AvrLm1, avrLm6*; *avrLm1, Avrlm6*; *avrLm1, avrlm6*). The frequency of *AvrLm1, AvrLm6* isolates was 73.7% prior to the breakdown, but decreased to 37.3% afterwards (Table 3). Conversely, the frequency of *avrLm1, avrlm6* isolates was only 0.8% prior to the breakdown, but increased to 28.5% afterwards.

Nine of the 14 isolates harbouring RIP alleles (64%) had a deletion allele at the *AvrLm1* locus and all these isolates were cultured from stubble of cultivars with sylvestris resistance. Additionally, all four isolates harbouring the virulent *AvrLm6*-2 allele (G^{123}C) had a deletion allele at *AvrLm1* (Table S3). When the isolates cultured from 2004 onwards were categorised in terms of the stubble from which they were derived, all those cultured from 'sylvestris stubble' had the *avrLm1* allele (Table 3) as expected, due to the presence of *Rlm1* in these cultivars [17]. Conversely, only 15% of isolates cultured from stubble of polygenic cultivars, which do not have *Rlm1* nor *Rlm6*, had the *avrLm1* allele. The frequency of *avrLm6* was 38.7% in isolates cultured from 'polygenic' stubble, compared to 68.7% of isolates cultured from 'sylvestris' stubble (Table 3). For all comparisons of isolates, there were no significant differences in allele frequencies at the mating type locus. Since the mating-type locus is not under selection pressure, a 1:1 ratio of each allele suggests that sampling of isolates has been random (data not shown).

Table 2. Alleles of *AvrLm1*, *AvrLm6, LmCys1* and *LmCys2* in 295 Australian isolates of *Leptosphaeria maculans*.

Gene[a, b]	Allele	Isolates (frequency %)	Nucleotide changes[c]				Coding sequence changes[d]	RIP dominance score[e]
			No.	Type				
				CpA to TpA	TpG to TpA	Other		
AvrLm1	0	185 (62.8)	0	0	0	0		0
	1	9 (3.0)	1	0	0	1 (T to A)	$I^{125}K$	0
	2	1 (0.3)	2	0	0	2 (T to A, G to T)	$I^{125}K$	0
	3	2 (0.7)	2	1	0	1 (T to A)	$I^{125}K$, $H^{155}Y$	0
	4	1 (0.3)	3	1	0	2 (T to A, G to T)	$I^{125}K$, $H^{155}Y$	0
	del	97 (32.9)	deletion					*NC*
AvrLm6	0	70 (23.7)	0	0	0	0		0
	1	134 (45.4)	1	0	1	0	SYN	0
	2	4 (1.4)	2	0	1	1 (G to T)	$G^{123}C$	0
	3	4 (1.4)	2	0	1	1 (A to G)	$K^{127}E$	0
	4	2 (0.7)	3	0	1	2 (T to C, A to G)	$F^{54}L$, $K^{127}E$	0
	5	1 (0.3)	38	10	15	13 (all G to A or C to T)	16 N-S, 5 SC	3.6
	6	1 (0.3)	41	11	13	17 (all G to A or C to T)	22 N-S, 4 SC	3.0
	7	3 (1.0)	42	11	14	17 (all G to A or C to T)	21 N-S, 4 SC	2.8
	8	3 (1.0)	41	14	15	12 (all G to A or C to T)	21 N-S, 6 SC	9.7
	9	3 (1.0)	46	18	12	16 (all G to A or C to T)	18 N-S, 6 SC	7.5
	10	1 (0.3)	42	13	13	16 (all G to A or C to T)	20 N-S, 5 SC	3.3
	11	2 (0.7)	46	13	14	19 (all G to A or C to T)	25 N-S, 5 SC	3.9
	del	67 (22.8)	deletion					*NC*
LmCys1	0	52 (17.6)	0	0	0	0		0
	1	238 (80.7)	1	0	0	1 (A to C)	$N^{121}H$	0
	2	2 (0.7)	3	0	0	3 (A to G, A to G, C to G)	SYN, $N^{121}G$, $Q^{134}E$	0
	3	2 (0.7)	4	0	0	4 (A to G, C to G, A to G, A to G)	SYN, $T^{53}A$, $N^{121}G$	0
	4	1 (0.3)	63	32	12	19 (all G to A or C to T)	26 N-S, 18 SC	7.3
LmCys2	0	293 (99.3)	0	0	0	0		0
	del	2 (0.7)	deletion					*NC*

[a]The reference sequences are AM084345 (designated as *AvrLm1-0*), AM2539336 (*AvrLm6-0*) and GU332625 (*LmCys1-0*) and GU332629 (*LmCys2-0*) in isolate v23.1.3 [8,9].
[b]Sizes of amplified products were 676 bp for *AvrLm1-0*, 751 bp for *AvrLm6-0*, 667 bp for *LmCys1-0* and 1050 bp for *LmCys2-0*.
[c]Deletions were confirmed by Southern analysis of selected isolates (Figure S1).
[d]SYN, synonymous amino acid substitutions; N-S, non-synonymous substitutions; SC, premature stop codons; *NC* = Not calculated.
[e]Allele sequences were analysed by RIPCAL for the presence of RIP mutations [25]. All sequences were compared to the wild-type allele (designated *-0*). RIP dominance scores of >1 are highly RIP-affected, whilst scores of 0 reflect the absence of RIP.

doi:10.1371/journal.ppat.1001180.t002

Table 3. Changes in the frequency of the profile of alleles of *AvrLm1* and *Avrlm6* of *Leptosphaeria maculans* isolates before and after the breakdown of 'sylvestris resistance' and in relation to stubble source of isolates.

Genotype	Number of isolates (frequency %)		Number of isolates (frequency %)[a]	
	before 2004	2004 onwards	Polygenic[b]	Sylvestris[b]
AvrLm1, AvrLm6	101 (73.7)	59 (37.3)	43 (53.8)	0 (0)
AvrLm1, avrLm6	5 (3.6)	33 (20.9)	25 (31.2)	0 (0)
avrLm1, AvrLm6	30 (21.9)	21 (13.3)	6 (7.5)	15 (31.3)
avrLm1, avrLm6	1 (0.8)	45 (28.5)	6 (7.5)	33 (68.7)

[a]Only isolates collected between 2004 and 2008 were analysed.
[b]Polygenic and sylvestris refers to the resistance of cultivars from which isolates were cultured.

doi:10.1371/journal.ppat.1001180.t003

Characteristics of the genomic environment of *AvrLm1* and *AvrLm6*

Because of the marked difference in the *AvrLm1* and *AvrLm6* allele frequencies before and after the breakdown of sylvestris resistance, the region flanking these genes was characterised to identify any features that might have influenced allele frequency. A 520 kb AT-rich genomic region bordered by *AvrLm1* and *LmCys2* (Figure 1A) was examined. Part of this region has been described previously in isolate v23.1.3 [8] and includes two additional genes encoding SSPs, *LmCys1* and *LmCys2*. Three other genes, *LmTrans*, *LmGT* and *LmMFS* were present; *LmCys1, LmTrans, LmGT, LmMFS* and *LmCys2* have been reported previously [8,9] but sequence data are only in the form of BAC clones. The features of the proteins are listed below and in Table S4.

The predicted LmCys1 protein (220 aa) contained eight cysteine residues and its single match was to a 'secreted in xylem' Six1 effector (also known as Avr3) of *Fusarium oxysporum* (26% identity, 40% similarity, accession number CAE55870.1) [26]. The predicted LmCys2 protein (247 aa), also containing eight cysteine residues, had no matches within the NCBI database. Both LmCys1 and LmCys2 were predicted to be secreted with signal peptides of 19 and 18 aa, respectively. To determine whether *LmCys1* and *LmCys2* were expressed highly *in planta*, which would be expected of genes encoding effector-like proteins, quantitative reverse transcriptase PCR (RT-PCR) analyses were performed. Transcript levels of *LmCys1* and *LmCys2 in planta* were six times higher than those of actin at seven days after inoculation of cotyledons of a susceptible *B. napus* cultivar (Figure 2). *LmCys1* and *LmCys2* were expressed at 0.1 and 0.01 times, respectively, that of actin in seven day *in vitro* cultures. Similar results were obtained with a second isolate (data not shown). This extremely high level of expression *in planta* compared to *in vitro* is similar to that seen for *AvrLm1* and *AvrLm6* (Figure 2) [8]. These characteristics (small cysteine-rich secreted proteins with no or few matches in databases and high *in planta* expression) strongly suggested that *LmCys1* and *LmCys2*, like *AvrLm1* and *AvrLm6*, encode effector-like proteins.

The LmTrans protein (373 aa) contained a DDE superfamily endonuclease domain predicted to be involved in efficient DNA transposition, and its best match was a putative transposase from *Stagonospora nodorum* (61% identity, 68% similarity, accession number CAD32687.1). The predicted LmGT protein (414 aa) was a putative glycosyltransferase with best matches to hypothetical protein PTRG_04076 of *Pyrenophora tritici-repentis* (89% identity, 95% similarity, EDU46914.1). The *LmMFS* protein (591 aa) belonged to the Major Facilitator Superfamily (MFS). This protein had best matches to putative protein SNOG_04897 from *S. nodorum* (74% identity, 82% similarity, EAT87288.2).

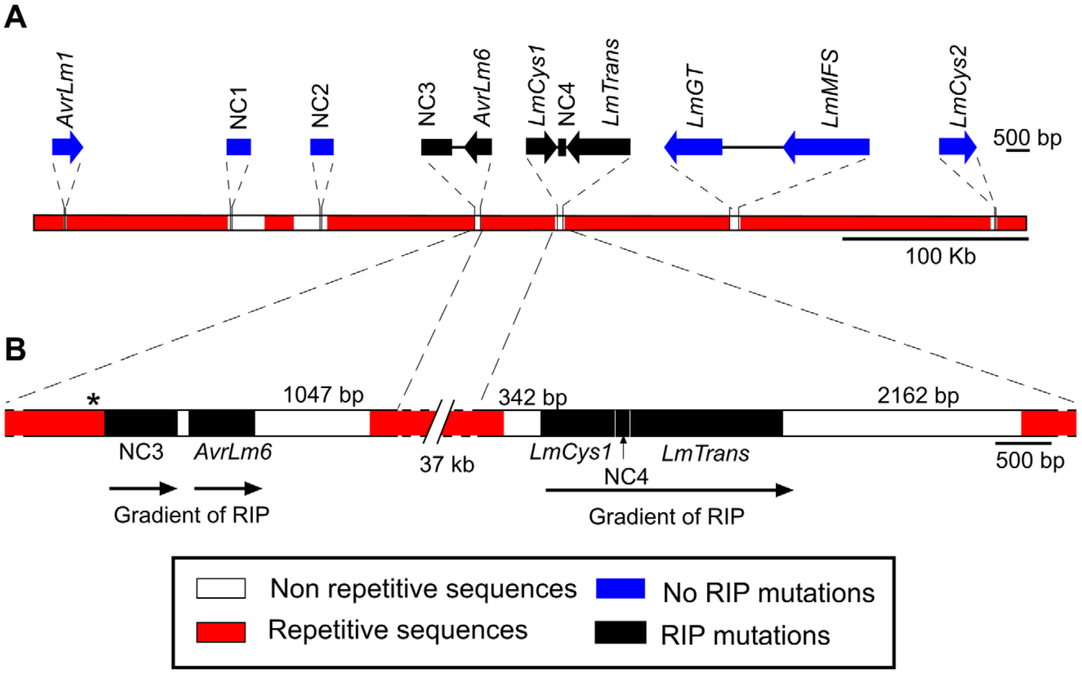


Figure 1. Location of the genes and non-coding, non-repetitive regions analysed from a 520 kb region of the *Leptosphaeria maculans* genome located on a 2.6 Mb chromosome in isolate v23.1.3. (A) Schematic representation of the *AvrLm1-LmCys2* genomic region whereby highly-repetitive sequences with low GC content (red) flank single copy sections with high GC content (white). Four genes encoding small-secreted proteins, *AvrLm1*, *AvrLm6*, *LmCys1* and *LmCys2*, were sequenced from 295 Australian isolates, whilst the remaining three genes and four non-coding, non-repetitive regions (NC1-4) were sequenced in 84 of the 295 isolates. Repeat-induced point (RIP) mutations were detected in *AvrLm6*, *LmCys1*, *LmTrans* and two non-coding, non-repetitive regions (NC3 and NC4) (black). (B) Location of single-copy sequences relative to the flanking repetitive regions. Within each of the two single copy regions(black) analysed, the frequency of RIP alleles and distribution of RIP mutations decreased in a 5′ to 3′gradient (arrows). The single copy region (149 bp) between NC3 and *AvrLm6* was not analysed and so the gradient arrow is discontinuous. * denotes a repeat region (620 bp) directly upstream of NC3 that is highly RIP-affected.
doi:10.1371/journal.ppat.1001180.g001

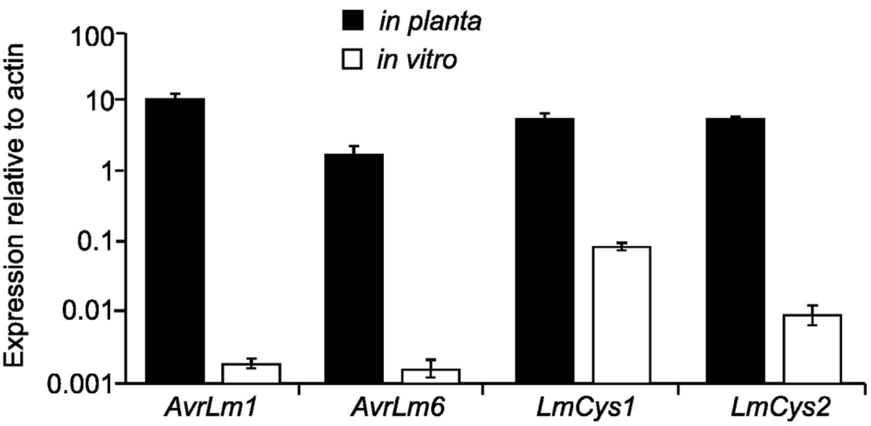


Figure 2. Quantitative reverse transcriptase (RT)-PCR analyses of *AvrLm6, AvrLm1, LmCys1* and *LmCys2*. At least 100 fold changes in expression were observed for each gene *in planta* compared to *in vitro* growth. RNA was prepared from seven day old cultures of isolates with the wild type alleles *AvrLm6-0, AvrLm1-0, LmCys1-0* and *LmCys2-0*, which were growing in 10% V8 juice. Additionally RNA was prepared from cotyledons of *B. napus* cv. Beacon 7 dpi with the same isolates. Two isolates were used. Transcript levels of each gene were compared to those of *L. maculans* actin within each sample. Data points are the average expression of each gene relative to actin determined from the two isolates, and three technical replicates for each sample. Expression relative to actin is presented as a log scale (Y-axis). Standard errors are represented.
doi:10.1371/journal.ppat.1001180.g002

Since *LmCys1* and *LmCys2* had effector-like properties and thus putative roles in the plant-fungal interaction, these genes were sequenced in the 295 isolates. Five and two alleles were detected for *LmCys1* and *LmCys2*, respectively (Table 1), including the wild type alleles obtained from published BAC sequences. RIPCAL analysis showed that only one isolate, which was collected after the breakdown of sylvestris resistance had a RIP allele of *LmCys1*, and there were no deletion alleles, whereas there were no RIP alleles of *LmCys2*, but two isolates collected before breakdown of sylvestris resistance had a deletion of *LmCys2* (Table 1). Overall, the frequencies of individual alleles ranged from 99% for *LmCys2-0* to 0.3% for *AvrLm1-4*. The 295 isolates comprised 34 haplotypes based on alleles of *AvrLm1*, *AvrLm6*, *LmCys1* and *LmCys2* (Table S3).

Four non-coding, non-repetitive regions (Figure 1A) ranging in size between 247 and 657 bp were analysed, to see whether single copy non-coding regions, like the single copy genes, were affected by RIP mutation. These regions and the three other genes *LmTrans, LmGT* and *LmMFS* in this region (Figure 1A) were sequenced in a subset of 84 of the 295 isolates, which included isolates representing all 34 haplotypes described above (Table S3). Three, four, 14 and two alleles were detected for NC1, NC2, NC3 and NC4, respectively, whilst two, one, and three alleles were detected for *LmTrans*, *LmGT* and *LmMFS*, respectively (Table 4 and Table S5). The mutations giving rise to the alleles of NC1, NC2 and *LmMFS* were single non-RIP-like nucleotide substitutions (both synonymous and non-synonymous), and single base pair deletions. No polymorphisms were detected in *LmGT* and no RIP alleles were detected for NC1, NC2, *LmMFS* or *LmGT*. A single isolate had RIP alleles for NC4 and *LmTrans* in addition to NC3, *AvrLm6* and *LmCys1*. The NC3 region had an extremely high frequency of RIP mutation and the ten RIP alleles were all associated with *AvrLm6* RIP alleles. Based on all seven genes and four non-coding regions, 51 haplotypes were identified among the 84 isolates (Table S6).

Distribution and extent of RIP mutations

The distribution and degree of RIP mutation across each allele of each gene was determined. This was represented as a ratio of the number of mutated CpA and TpG sites relative to number of potential RIP sites in isolate v23.1.3 in a 100 bp rolling window. The seven RIP alleles of *AvrLm6* had the highest frequency of RIP towards the 3′ end (Figure 3A). The frequency of RIP was higher within the 3′ UTR and the exons, than in the introns and 5′ UTR (Figure 3B). The ten RIP alleles of NC3 and single RIP allele of *LmCys1* showed the highest frequency of RIP at their 5′ ends, whilst RIP mutations were evenly distributed throughout the single RIP allele of *LmTrans* (Figure S2), the latter gene being the furthest 3′ characterised single copy sequence affected by RIP mutation within the *AvrLm1-LmCys2* genomic region (Figure 3C). The NC3-*Avrlm6* and *LmCys1*-NC4-*LmTrans* single copy regions in which RIP mutations were detected, were separated by 37 kb of repetitive elements (Figure 1B). Both these single copy regions were closer to upstream repetitive elements (<342 bp) than to downstream repetitive elements (>1 kb) (Figure 1B). The degree of RIP within these single copy regions was proportional to their proximity to repetitive elements, and thus a gradient of RIP mutations was apparent in a 5′ to 3′ direction (Figures 1B and 3B).

Since amongst these single copy regions the degree of RIP mutations was highest in NC3, a repeat region (620 bp) directly upstream was analysed to see if it was severely RIP-affected and thus might act as a point of 'leakage' of RIP into NC3 and *AvrLm6* (Figure 1B). BLAST analysis of the genomic sequence of isolate v23.1.3 revealed 293 copies of this repeat. RIPCAL analysis showed that the repeat directly upstream of NC3 was amongst the most highly RIP-affected copy of that repeat in the genome.

Transcription of *AvrLm1*, *AvrLm6*, *LmCys1* and *LmCys2* alleles

The mutations of *AvrLm1*, *AvrLm6*, *LmCys1* and *LmCys2* (deletions or RIP mutations including stop codons) would be expected to lead to lack of transcription of these alleles. This hypothesis was assessed in a range of isolates seven days after inoculation of cotyledons of *B. napus* cv. Beacon, which all the isolates could attack. Primers designed to amplify 500–700 bp products within the coding region of these genes were used in end-point RT-PCR (Table 5). Isolates with either *AvrLm1-0* or *AvrLm1-1* allele had an *AvrLm1* transcript of the appropriate size. As expected, isolates with the deletion allele had no transcript. The

Table 4. Nucleotide changes in non-coding, non-repetitive regions within the *AvrLm1-LmCys2* genomic region in Australian isolates of *Leptosphaeria maculans*.

Region[a, b]	Allele	Number of isolates (frequency %)	Nucleotide changes				RIP dominance score
			No.	Type			
				CpA to TpA	TpG to TpA	Other	
NC1	0	64 (76.2)	0	0	0	0	0
	1	1 (1.2)	1	0	0	1 (C to G)	0
	2	19 (22.6)	1	0	0	1 (T to C)	0
NC2	0	53 (63.1)	0	0	0	0	0
	1	27 (32.1)	1	0	0	1 (T to A)	0
	2	1 (1.2)	1	0	0	1 (A to T)	0
	3	3 (3.6)	3	0	0	1 (T to A, del G, del T)	0
NC3	0	67 (79.6)	0	0	0	0	0
	1	1 (1.2)	1	0	0	1 (T to C)	0
	2	1 (1.2)	1	0	0	1 (A to G)	0
	3	1 (1.2)	1	0	0	1 (C to G)	0
	4	2 (2.4)	28	9	7	12 (all G to A or C to T)	1.5
	5	1 (1.2)	29	9	7	13 (all G to A or C to T)	1.5
	6	1 (1.2)	29	9	7	13 (all G to A or C to T)	1.3
	7	1 (1.2)	30	9	7	14 (all G to A or C to T)	1.5
	8	1 (1.2)	38	13	12	13 (all G to A or C to T)	3.6
	9	1 (1.2)	34	9	11	14 (all G to A or C to T)	2.0
	10	1 (1.2)	38	13	11	14 (all G to A or C to T)	3.0
	11	2 (2.4)	39	9	12	18 (all G to A or C to T)	2.1
	12	3 (3.6)	40	9	12	19 (all G to A or C to T)	2.1
	13	1 (1.2)	29	8	8	12 (all G to A or C to T)	1.6
NC4	0	83 (98.8)	0	0	0	0	
	1	1 (1.2)	9	4	0	5 (all G to A or C to T)	4.0

These isolates represented all 34 haplotypes based on *AvrLm1*, *AvrLm6*, *LmCys1* and *LmCys2* alleles.

[a] The reference sequences are CT485790 (designated as NC1-0), CT485667 (NC2-0), CT485649 (NC3-0) and CT485669 (NC4-0) identified from isolate v23.1.3 [8,9].

[b] The sizes of the amplified products were 517 bp for NC1-0, 496 bp for NC2-0, 657 bp for NC3-0 and 139 bp for NC4-0.

doi:10.1371/journal.ppat.1001180.t004

AvrLm6-0, *AvrLm6-1* or *AvrLm6-2* alleles were expressed, whilst the RIP alleles, *AvrLm6-6*, *AvrLm6-7*, *AvrLm6-8* or *AvrLm6-9* were not. As expected, no expression of *AvrLm6* or of *LmCys2* was detected in isolates with the deletion alleles of these genes. Isolates with *LmCys1-0*, *LmCys1-1* or *LmCys1-2* alleles had an *LmCys1* transcript, whilst the isolate with the RIP allele, *LmCys1-4*, did not. Actin transcripts were detected in all isolates.

Rate of mutations of genes

To determine the rates of mutations, genes were analysed by phylogeny-based likelihood ratio tests (LRT) implemented in the program HyPhy [27]. These tests suggested that nucleotides of *AvrLm1* and *LmMFS* mutated at a constant rate, i.e. mutations did not deviate significantly ($p = 0.544$) from a clock-like rate of evolution. In contrast, an accelerated mutation rate was detected for *AvrLm6* and *LmCys1* loci ($P<0.001$) compared to expectations under constant (i.e. clock-like) evolution. However, when RIP alleles were excluded, mutations evolved at a clock-like rate (Table 6). Genetic divergence between haplotypes was calculated using Molecular Evolutionary Genetics Analysis (MEGA) software [28]. Relative rates of sequence evolution of non-RIP alleles were much lower than those of RIP alleles (Table 6). To determine whether the seven proteins were undergoing positive selection, the rates of non-synonymous and synonymous substitutions were compared. All RIP alleles were excluded from this analysis since they contained multiple stop codons. Evolution of codon changes within AvrLm1 and LmCys1 was best explained by a model of positive selection, as shown by a likelihood ratio test implemented using two complementary approaches, the sitewise likelihood-ratio (SLR) and phylogenetic analysis by maximum likelihood (PAML) methods (Table 7). This interpretation was supported by the finding of positive selection at a single codon site in AvrLm1 and at three sites in LmCys1 (Table 7). Analysis of the AvrLm6 protein using the SLR approach suggested a single amino acid (codon 54) may be undergoing positive selection; however, this finding was not supported by the PAML analysis. Similar analyses could not be performed on LmCys2, LmGT or LmTrans as these genes had fewer than three alleles.

Evolution and phylogenetic relationships of RIP and deletion alleles

Two hypotheses for the evolution of RIP alleles, monophyly (having a single origin or having evolved only once) and polyphyly (multiple origins or having evolved several times independently) were tested by comparing tree topologies using the Kishino–Hasegawa (KH) test [29]. For both the Maximum Parsimony (MP) and

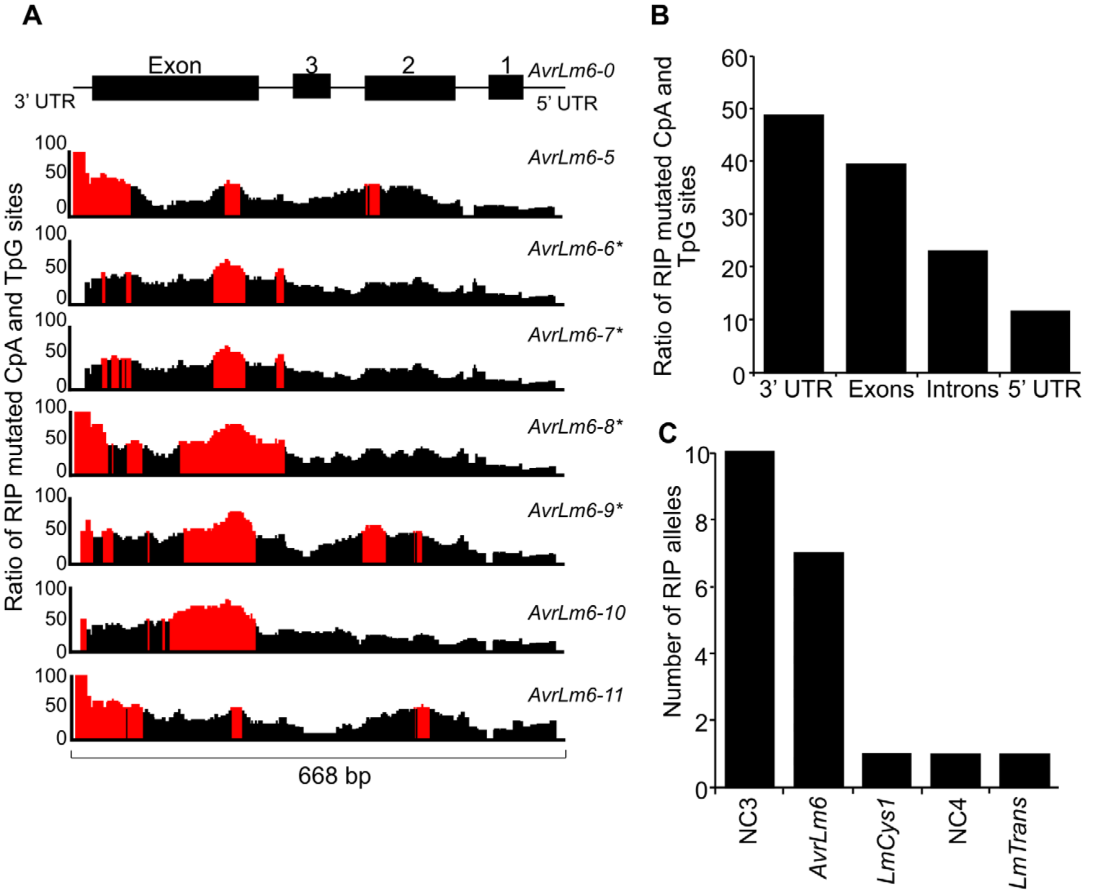


Figure 3. Distribution of RIP mutations across the *AvrLm1-LmCys2* genomic region in *Leptosphaeria maculans.* (A) The ratio of mutated CpA or TpG sites within the seven *AvrLm6* RIP alleles compared to the number of available CpA and TpG nucleotides present in the wild type *AvrLm6-0* allele over a 100 bp window. Regions where greater than 50% of potential RIP sites are mutated are highlighted in red. There is a higher proportion of RIP mutations towards the 5′ end of the region (3′ end of the *AvrLm6* gene). *AvrLm6-0, -1, -2, -3* and *-4* alleles do not display RIP mutations. (B) The average ratio of RIP mutations within the untranslated regions (UTRs), exons and introns for the seven RIP alleles of *AvrLm6* relative to the number of potential RIP sites in the wild type sequence. The 3′ UTR and exons are undergoing the highest frequency of RIP. (C) The number of RIP alleles for each of the genes and non-coding, non-repetitive (NC) regions analysed across the *AvrLm1-LmCys2* genomic region in 84 isolates (see also Figure S2). The number of RIP alleles is highest for NC3 and decreases in the genes and non coding regions downstream, consistent with RIP occurring in a directional manner.* represent RIP alleles of *AvrLm6* that are not transcribed and that confer a virulence phenotype on an *Rlm6*-containing cultivar. The remaining RIP alleles of *AvrLm6* were not tested for virulence towards *Rlm6*.
doi:10.1371/journal.ppat.1001180.g003

Maximum Likelihood (ML) approach, trees based on the assumption of a monophyletic origin of RIP alleles performed significantly better than those based on polyphyletic origin (Table 8). The phylogenetic relationship of all detected haplotypes is depicted in Figure 4. The RIP and non-RIP associated alleles form two distinct clusters, supporting the hypothesis of a single origin of RIP alleles. In contrast, haplotypes associated with gene deletions are associated with multiple clades of the tree and have probably arisen several times.

Discussion

Selection pressure on frequencies of *AvrLm1* and *AvrLm6* alleles is imposed by extensive sowing of cultivars with *Rlm1* resistance

Breakdown of disease resistance has been observed in other plant fungal pathogen systems where fungal populations evolve rapidly (for review see [30]). In the canola- *L. maculans* interaction described here, strong selection pressure was exerted on the *AvrLm1* locus, due to extensive sowing of sylvestris cultivars with *Rlm1*, which was consistent with the finding of a rapid increase in the frequency of isolates with virulent (*avrLm1*) alleles after the breakdown of resistance. Surprisingly the frequency of isolates virulent (with the *avrLm6* allele) towards *Rlm6* increased, although cultivars with polygenic or with sylvestris resistance have not been shown to contain *Rlm6* [17,31]. The linkage and genomic location of *AvrLm1* and *AvrLm6* may have led to a selective sweep whereby selection at *AvrLm1* affected the frequency of *avrLm6* alleles through ‘hitchhiking’. Thus strong selection imposed by widespread deployment of a plant resistance gene that favors a complementary effector allele in a pathogen could affect evolution of closely-linked effector genes. It is intriguing that in France recurrent sowing of *Rlm6*-containing cultivars in localised field

Table 5. Expression analysis of *AvrLm1*, *AvrLm6*, *LmCys1* and *LmCys2* alleles in *Leptosphaeria maculans* isolates *in planta*.

Isolate	Genotype (allele)				Gene expression (7 days post inoculation)				
	AvrLm1	*AvrLm6*	*LmCys1*	*LmCys2*	*AvrLm1*	*AvrLm6*	*LmCys1*	*LmCys2*	actin
05P031	*0*	*1*	*0*	*0*	+	+	+	+	+
06P042	*0*	*9*[a]	*1*	*0*	+	−	+	+	+
04P012	*1*	*0*	*1*	*0*	+	+	+	+	+
06P040	*1*	*del*	*1*	*0*	+	−	+	+	+
IBCN18	*1*	*del*	*2*	*del*	+	−	+	−	+
05P033	*del*	*2*	*1*	*0*	−	+	+	+	+
05P034	*del*	*2*	*1*	*0*	−	+	+	+	+
06S013	*del*	*6*[a]	*1*	*0*	−	−	+	+	+
06S014	*del*	*7*[a]	*1*	*0*	−	−	+	+	+
06S039	*del*	*8*[a]	*4*[a]	*0*	−	−	−	+	+

Cotyledons of 14 day old seedlings of *Brassica napus* cv. Beacon were infected with conidia of individual *L. maculans* isolates. Seven days post inoculation, tissue was harvested and RNA extracted from 20 inoculation sites per isolate. Expression of genes was determined using end point RT-PCR. '+' band of expected size. '−' no band.

[a] alleles associated with RIP.

doi:10.1371/journal.ppat.1001180.t005

trials led to an increase in *avrLm6* isolates and a corresponding increase in isolates avirulent at the *AvrLm1* locus [19]. This difference may be due to the different targets of selection pressure - *Rlm6* in France and *Rlm1* in our study.

This situation whereby selection pressure on one gene affected allele frequencies of another may be partly due to the presence of these two effector genes in a repeat-rich region, where there is a low recombination frequency. Effectors in other fungi are present in repeat-rich regions. In the rice blast fungus, *Magnaporthe oryzae*, avirulence gene *Avr-Pita* is located 48 bp from telomere repeats whilst *Avr1-CO39* is associated with transposable elements [32,33]. An avirulence gene (SIX1) of *Fusarium oxysporum* f.sp. *lycopersici* is flanked by transposable elements [34] and other effectors are localised on a single, transposon-rich chromosome [35]. Toxin-encoding genes, *Tox3 and ToxA*, are located next to repetitive elements in *Stagonospora nodorum* [36,37,38] and effectors are in repeat-rich regions of the genome of the oomycete, *Phytophthora infestans* [39]. In the sequenced isolate of *L. maculans*, *AvrLm1* and *AvrLm6* are located amongst multiple copies of long terminal repeat retrotransposons, namely *Pholy*, *Olly*, *Polly* and *Rolly*, which are generally incomplete. Furthermore the distribution and number of these elements within this genomic location varies considerably between isolates [20,22]. The presence of effector genes within such regions is suggested to promote their adaptation and diversification when exposed to strong selection pressure [40]. Rep and Kistler have speculated that the presence of highly repetitive regions containing transposons, may promote mutation of resident effector genes [41].

Table 6. Analysis for the presence of a molecular clock, and relative rates of sequence evolution for genes within the *AvrLm1-LmCys2* genomic region of *Leptosphaeria maculans*.

Gene	Analysis of molecular clocks (HyPhy) [a]				Relative rates of substitution (MEGA) [b]	
	LogLh-no clock	LogLh-clock	p-value	Interpretation	Nucleotide (K2P)	Interpretation
AvrLm1	−1073.67	−1077.13	0.544	Clock-like	0.002 (0.001)	
AvrLm6	−1578.58	−1633.62	<0.001	Accelerated	0.051 (0.006)	
excl. RIPs	−956.75	−963.27	0.111	Clock-like	0.003 (0.001)	Genetic distances <10 fold lower with RIP alleles excluded
LmCys1	−1175.04	−1240.44	<0.001	Accelerated	0.046 (0.006)	
excl. RIPs	−964.78	−968.33	0.311	Clock-like	0.005 (0.002)	Genetic distances <10 fold lower with RIP alleles excluded
LmTrans	*NC*	*NC*	*NC*	*NC*	0.086 (0.008)	
excl. RIPs	*NC*	*NC*	*NC*	*NC*	*NC*	
LmGT	*NC*	*NC*	*NC*	*NC*	*NC*	
LmMFS	−2726.22	−2729.52	0.159	Clock-like	0.001 (0.001)	
LmCys2	*NC*	*NC*	*NC*	*NC*	*NC*	

Presence of a molecular clock was analysed using a phylogeny-based likelihood ratio test.

MEGA was used to infer relative rates of sequence evolution by calculating means of genetic distances (Kimura-2-Parameter (K2P) between haplotypes.

[a] When RIP alleles were excluded from HyPhy analysis (excl. RIPs), p values for all genes became non-significant (clock-like).

[b] Since both HyPhy and MEGA approaches are phylogeny-based, a minimum of three distinct alleles are required for analysis. Genes with less than three alleles could not be analysed (*NC*).

doi:10.1371/journal.ppat.1001180.t006

Table 7. Analysis of positive selection on amino acids of proteins encoded within the *AvrLm1-LmCys2* genomic region of *Leptosphaeria maculans*.

Protein	SLR approach		PAML approach (M7 vs. M8)			
	Positively selected codons	***P* value**	**LRT statistics [a]**	***P* value**	**Positively selected codons**	***P* value**
AvrLm1	125	0.002	8.126	0.017	125	0.001
AvrLm6	54	0.003	0.214	0.900	54	NS
LmCys1	53	<0.001	24.018	<0.001	53	0.002
	121	<0.001			121	<0.001
	134	0.002			134	<0.001
LmTrans	*NC*	–	*NC*	–	*NC*	–
LmGT	*NC*	–	*NC*	–	*NC*	–
LmMFS	none	–	0.00	1.00	none	–
LmCys2	*NC*	–	*NC*	–	*NC*	–

Evidence for non-neutral selection was assessed by comparing the rate of non-synonymous and synonymous substitutions using two approaches, SLR and PAML [52,53]. These approaches are phylogeny-based and require a minimum of three distinct alleles for analysis, so values for LmTrans, LmGT and LmCys2 were not calculated (*NC*). For the PAML approach, the comparison included the likelihood estimates of the neutral null model (M7) and the alternative model of positive selection (M8). RIP alleles were excluded from these analyses since such alleles encode sequences with stop codons.
NS = not significant.
[a]LRT statistics are compared against a χ^2 distribution with two degrees of freedom.
doi:10.1371/journal.ppat.1001180.t007

Repeat-induced point mutations may be 'leaking' from adjacent inactivated transposable elements into single copy regions

The genes and non-coding regions undergoing RIP within the *AvrLm1- LmCys2* gene region are single copy and therefore are not expected to be targeted by RIP. Two explanations for the presence of RIP mutations in these genes are as follows. Firstly, if these genes are the result of an ancestral duplication, RIP mutation may be acting directly on them. However, Fudal et al showed that no closely related paralogs of *AvrLm6* exist in the genome of isolate v23.1.3 suggesting that RIP would not be targeting these sequences due to a duplication event [22]. Alternatively, the *AvrLm6* locus may be completely or partially duplicated in isolates where RIP was detected. However, southern hybridisations suggest only a single copy of the *Avrlm6* locus in isolates where RIP was detected, which does not support the possibility of RIP being targeted to this locus. A more likely explanation is that RIP mutations 'leak' from adjacent repetitive sequences. As RIP mutation is traditionally observed to be restricted to repetitive regions and not single copy regions, leakage of RIP mutation might occur within a relatively short distance of a RIP-affected repeat, as suggested by Fudal et al [22]. Indeed, this has been reported in *N. crassa* whereby leakage of RIP was detected in single copy sequences at least 930 bp from the boundary of neighbouring duplicated sequences [42]. This is consistent with our finding of a high frequency of RIP mutations in single copy regions of *L. maculans* with the degree of RIP mutation being proportional to the proximity of flanking repetitive elements. The potential 'leakage' of RIP mutations into closely linked effector genes highlights the power of this process to lead to major evolutionary changes to

Table 8. Hypothesis-testing of multiple (polyphyletic) vs. single (monophyletic) origin of RIP mutation-associated haplotypes of *Leptosphaeria maculans*.

Tree	Tree score [a]	Score difference[b]	*P*
Maximum parsimony (MP) (gaps treated as missing)			
Monophyletic origin of RIP	15488	10	<0.01
Polyphyletic origin of RIP	15478	best	
Maximum parsimony (MP) (gaps treated as fifth state)			
Monophyletic origin of RIP	15550	32	<0.01
Polyphyletic origin of RIP	15518	best	
Maximum likelihood (ML)			
Monophyletic origin of RIP	15345	58	<0.01
Polyphyletic origin of RIP	15287	best	

Kishino–Hasegawa tests were used to assess different hypotheses of RIP evolution by comparing tree topologies as implemented in PAUP*. The probabilities (*P*) of obtaining better trees were assessed using two-tailed tests, the full optimization criterion and 1000 bootstrap replicates.
[a]Tree scores refer to branch lengths for the tree topologies. For the maximum likelihood analysis, tree scores are given as –ln of branch length.
[b]'Best' refers to the 'shortest' tree, which is likely to be the most accurate.
doi:10.1371/journal.ppat.1001180.t008

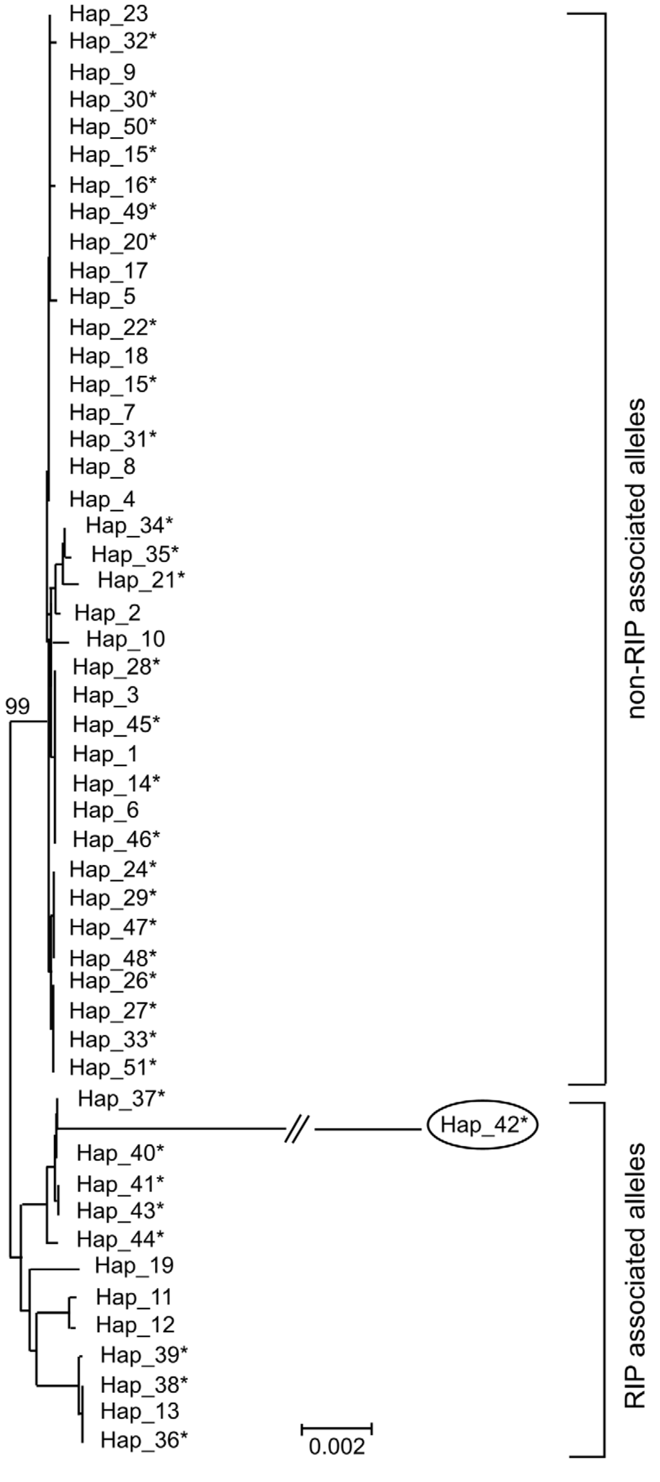

Hap_23
Hap_32*
Hap_9
Hap_30*
Hap_50*
Hap_15*
Hap_16*
Hap_49*
Hap_20*
Hap_17
Hap_5
Hap_22*
Hap_18
Hap_15*
Hap_7
Hap_31*
Hap_8
Hap_4
Hap_34*
Hap_35*
Hap_21*
Hap_2
Hap_10
Hap_28*
Hap_3
Hap_45*
Hap_1
Hap_14*
Hap_6
Hap_46*
Hap_24*
Hap_29*
Hap_47*
Hap_48*
Hap_26*
Hap_27*
Hap_33*
Hap_51*
Hap_37*
Hap_42*
Hap_40*
Hap_41*
Hap_43*
Hap_44*
Hap_19
Hap_11
Hap_12
Hap_39*
Hap_38*
Hap_13
Hap_36*
99
0.002
non-RIP associated alleles
RIP associated alleles

Figure 4. Genetic diversity and phylogenetic relationships between haplotypes of *Leptosphaeria maculans*. Un-rooted strict consensus of 1000 ML trees (-ln L = 16381.75) constructed from the concatenated DNA sequences of the seven genes and four non-coding, non-repetitive regions of the *AvrLm1-LmCys2* genomic region. All CpA to TpA and TpG to TpA nucleotide changes were removed from the data set prior to analysis. Haplotypes (see Table S6) associated with RIP alleles cluster in a distinct clade with high bootstrap support (99%), suggesting a single origin. In contrast, haplotypes related to deletion alleles (*) are associated with multiple clades suggesting multiple origins. Note that RIP associated alleles have much longer branches (i.e. larger genetic distance) due to an accelerated evolution compared to non-RIP alleles (see Table 6). Haplotype 42 (circled) has RIP alleles at five of the loci examined (NC3, *AvrLm6*, *LmCys1*, NC4 and *LmTrans*), resulting in an exceptionally long branch.
doi:10.1371/journal.ppat.1001180.g004

genes such as effectors that play an important role in the lifestyle of an organism.

Deletion alleles of *AvrLm1* and *AvrLm6* had multiple origins whilst RIP alleles had a single origin

All haplotypes associated with RIP alleles of *AvrLm6* and *LmCys1* appeared to have a single origin indicating that the event that led to the RIP occurred only once. Furthermore, haplotypes associated with the virulent *AvrLm6-2* allele (cysteine substitution) arose as a single clade. These single origins suggest that events leading to both the RIP mutations and non-RIP amino acid substitutions are much rarer than those leading to the deletion mutations. The RIP event occurred after the breakdown of sylvestris resistance, as isolates with RIP mutations were not detected before this time. This rapid appearance of RIP mutations is not unprecedented; such mutations have been shown to occur after one generation in *L. maculans*. When transformants with several tandemly linked copies of a hygromycin resistance gene were crossed with a wild type strain, none of the progeny were resistant to hygromycin due to the presence of multiple stop codons generated by RIP mutations in the hygromycin resistance gene [43].

Mutations associated with resistance to azole fungicides in *Mycosphaerella graminicola* also are derived from a single origin. Resistance emerged only once following strong selection due to widespread use of azole fungicides [44]. In both the *M. graminicola* and *L. maculans* examples, a similar phylogeny was found despite differences in origin and type of evolutionary pressure. In *L.maculans* haplotypes associated with deletion alleles conferring virulence towards *AvrLm1* and *AvrLm6* appeared to have a polyphyletic origin. Isolates with these deletions were detected prior to 2004 when the resistance breakdown occurred, albeit at a much lower frequency than afterwards. Haplotypes associated with deletion of both *AvrLm1* and *AvrLm6* might have been derived directly from the ancestral wild type rather than via the deletion of one gene followed by that of the second.

Deletions, non-synonymous and RIP mutations can confer virulence

Deletions, RIP mutations and non-RIP amino acid substitutions conferred virulence at the *AvrLm6* locus, whilst only deletions were responsible for virulence at the *AvrLm1* locus. Similar types of mutations were detected in French populations of *L. maculans* isolates [22]. The finding of a virulence allele of *AvrLm6* arising from a non-synonymous, non-RIP like mutation, of a glycine to cysteine substitution, was intriguing. In other avirulence proteins the loss rather than the gain of cysteine through non-synonymous substitutions confers virulence. For example, in the AVR4 protein of *Cladosporum fulvum*, loss of cysteine residues renders the isolate virulent towards tomato cultivars with the corresponding *Cf-4* resistance gene [45]. The *AvrLm6-2* allele of *L. maculans* gives rise to a protein with seven rather than six cysteine residues in the protein encoded by *AvrLm6-0*. This latter protein is proposed to have two disulphide bridges between C^{109} and C^{130}, and C^{103} and C^{122} [8] on the basis of the SCRATCH disulphide bond prediction program [46]. The program predicts that the presence of the additional cysteine (C^{123}) in the *AvrLm6-2* allele would result in a third disulphide bridge, between C^{26} and C^{123}.

As well as the mechanisms leading to inactivation of alleles described above, some of the proteins were undergoing non-RIP amino acid substitutions which did not lead to a change in phenotype. Some of these mutations, in AvrLm1 and LmCys1, were the results of positive selection, which favours new mutations that confer a fitness advantage and thus lead to an increase in gene diversity [12,47]. Positive selection has been detected in pathogen effector genes including the avirulence gene that encodes NIP in *Rhynchosporium secalis* [47] and in genes encoding host-specific toxins such as *S. nodorum* ToxA [36,48]. In contrast to *AvrLm1, AvrLm6* and *LmCys1*, the remaining genes in the 520 kb AT- rich region, including *LmCys2* showed very little variation. Despite positive selection driving amino acid substitutions within some of the effector-like proteins, deletion and RIP mutations are by far the major mechanisms leading to virulence at the *AvrLm1* and *AvrLm6* loci.

Materials and Methods

Brassica cultivars and *Leptosphaeria maculans* isolates

Stubble of cultivars of *B. napus* and *B. juncea* infested with *L. maculans* was collected each year from 1997 to 2008 from 25 locations across Australia (Table S1). For instance, stubble from a crop sown in 2003 was collected from the field in 2004 and isolates were then cultured from it. Cultivars with 'polygenic' resistance (Beacon, Dunkeld, Emblem, Grace, Hyden, Jade, Pinnacle, Skipton, Pinnacle and Tornado TT) had one or more *Rlm* genes, but none had *Rlm1* nor *Rlm6*. The identity of resistance genes in some of these cultivars has been reported [44]. The category of 'sylvestris resistance' refers to cultivars (Surpass 400, Surpass 501TT, Surpass 603CL, 45Y77 and 46Y78) with resistance derived from *B. rapa* spp. *sylvestris* [14] have *Rlm1* and *RlmS* [17]. The category of 'juncea' resistance refers to cultivars and lines of *B. juncea* (cv. Dune and lines JC05002, JC05006 and JC05007). Stubble of the latter two categories was not collected prior to 2004. From 2004 onwards although cultivars with sylvestris resistance were withdrawn from sale, these lines were grown in yield trials across Australia, and stubble was collected from them. Isolates (287) were cultured from individual ascospores discharged from stubble collected the previous year as described previously [15]. In addition, eight Australian isolates collected in 1987 and 1988 were analysed. All isolates were maintained on 10% Campbell's V8 juice agar.

Virulence testing

Virulence of a subset of isolates was tested on three *B. napus* and one *B. juncea* cultivars. The *B. napus* cv. Beacon and cv. Q2 are susceptible controls that all isolates could attack. Cultivar Columbus contains *Rlm1* and *Rlm3* and *B. juncea* cv. Aurea contains *Rlm5* and *Rlm6* [3]. No *Rlm1*-only or *Rlm6*-only cultivars were available. Cotyledons of 14-day old seedlings were wounded and inoculated with conidia of individual isolates representing different haplotypes for *AvrLm1* and *AvrLm6*. Symptoms were assessed at 10, 14 and 17 days post-inoculation (dpi) and pathogenicity scores determined at 17 dpi by scoring lesions on a scale from 0 (no darkening around wounds) to 9 (large grey-green lesions with profuse sporulation). Mean pathogenicity scores (determined from 40 inoculation sites) ≤3.9 were assigned as an avirulent phenotype whilst scores ≥4.0 were assigned as a virulent phenotype [17].

Gene identification and primer design

Non-coding, non-repetitive regions in the *AvrLm1-LmCys2* genomic region and genes 3′ of *AvrLm6* (Figure 1) were identified using published information [22] and by BLAST searches. Primers were designed upstream and downstream of start and stop codons to allow analysis of the sequences of entire open reading frames. For transcriptional analyses, primers were designed to amplify a 500–700 bp region of the coding sequence, flanking an intron where possible. All primers were designed using the program Primer3 [49] (Table S7). Primers to amplify the mating type locus and actin have been described previously [8,24].

The sequence information for all genes has been deposited in GenBank with the following accession numbers; *AvrLm1*, AM084345 [1], *AvrLm6*, AM259336 [2], *LmCys1*, GU332625, *LmTrans*, GU332626, *LmGT*, GU332627, *LmMFS*, GU332628 and *LmCys2*, GU332629.

Allele genotyping

Genomic DNA was isolated from mycelia. Conditions for all PCR experiments were 95°C for 3 min; 35 cycles of 95°C for 30 sec, 59°C for 30 sec and 72°C for 1 min; 72°C for 6 min. PCR products were purified using QIAquick PCR purification kit (Qiagen) and sequenced using BigDyeTM terminator cycling conditions. Sequences were analysed using Sequencher v 4.0.5. Deletion genotypes were assigned if no band was produced following amplification with the *AvrLm1*, *AvrLm6* or *LmCys2*-specific primers. Amplification of the mating type locus was a positive control for DNA quality. In a subset of eight isolates, deletion alleles were confirmed by Southern analysis of genomic DNA that had been digested with *Hin*dIII and hybridised with the appropriate probe (Figure S1). Additionally, PCR screens were performed on genomic DNA from two (for *LmCys2*) to 25 (for *AvrLm6*) isolates using multiple primer sets that amplify specific regions of the *AvrLm1*, *AvrLm6* and *LmCys2* gene regions. These amplifications confirmed that the entire locus was deleted in all isolates tested (data not shown).

Allele sequences were analysed by RIPCAL for the presence of RIP mutations [25]. RIPCAL generates a RIP dominance score, which is the frequency of the dominant dinucleotide RIP mutation (in this case CpA→TpA) relative to the sum of the alternative mutations (CpC→TpC, CpG→TpG and CpT→TpT). Sequences with RIP dominance scores >1 are considered to be highly RIP-affected. The 'model' sequence used for all RIPCAL analyses was the 'wild type' allele (designated with a *-0* suffix) from the isolate (v23.1.3) whose genome has been sequenced. The spatial distribution of RIP was assessed for each gene and four non-coding regions by comparing the ratio of mutated CpA or TpG sites detected by RIPCAL, relative to the number of available CpA and TpG nucleotides present within the wild type allele over a 100 bp rolling window.

Expression analysis

Ten infected cotyledons of *B. napus* cv. Beacon were harvested at 7 dpi. Necrotic tissue surrounding the inoculation wounds of each cotyledon was harvested using a hole punch (diameter 8 mm). Total RNA was purified from this tissue using the RNeasy Plant Mini Kit (Qiagen) and was treated with DNaseI (Invitrogen) before cDNA was synthesized using a first strand cDNA synthesis kit and an oligo-dT primer. End point RT-PCR was used to assess expression of *AvrLm1*, *AvrLm6*, *LmCys1*, *LmCys2* and actin.

Quantitative RT-PCR was used to determine levels of expression of *AvrLm1*, *AvrLm6, LmCys1* and *LmCys2 in planta* and *in vitro* culture. RNA was prepared from seven day old cultures of isolates with the wild type alleles of these genes, which were growing in 10% V8 juice. Additionally RNA was prepared from cotyledons of *B. napus* cv. Beacon 7 dpi infected with isolates with the wild type alleles of these genes. Total RNA and cDNA synthesis was performed as described above. Controls lacking reverse transcriptase were included. Quantitative RT-PCR was performed using Rotor-Gene 3000 equipment (Corbett Research, Australia) and QuantiTect SYBR Green PCR kit (QIAgen). A standard curve of amplification efficiency of each gene was generated from purified RT-PCR products [50]. Diluted RT product (1 μl) was added to 19 μl of PCR mix and subjected to 40 cycles of PCR (30 s at 94°C, then 60°C and then 72°C). All samples were analysed in triplicate. The amplified product was detected every cycle at the end of the 72°C step. Melt curve analysis after the cycling confirmed the absence of non-specific products in the reaction. The fluorescence threshold (Ct) values were determined for standards and samples using the Rotor-Gene 5 software. Ct values were exported to Microsoft Excel and analysed [51]. Actin was used as a reference gene.

Phylogenetic analyses

Deviation from a constant rate of molecular evolution within the data sets (i.e. a "molecular clock") was assessed using the phylogeny-based likelihood ratio test (LRT) implemented in the program HyPhy [27]. To estimate the contribution of the RIP alleles, likelihoods were calculated both for the total data sets and for data sets excluding RIP alleles. MEGA was also used to infer relative rates of sequence evolution by calculating means of genetic distances (Kimura-2-Parameter) between haplotypes.

Evidence for non-neutral evolution was assessed using two complementary approaches by comparing the rate of non-synonymous substitutions with the rate of synonymous substitutions ($dN/dS = \omega$). Firstly, the analysis was based on the "sitewise likelihood-ratio" method as implemented in the SLR software package [52]. The test consists of performing a likelihood-ratio test on a site-wise basis, testing the null model (neutrality, $\omega = 1$) against an alternative model $\omega \neq 1$ (i.e. purifying selection $\omega < 1$; positive selection $\omega > 1$). Secondly, $dN/dS = \omega$ was tested using a phylogenetic analysis based on maximum likelihood as implemented in the PAML software package [53]. Two codon substitution models were compared via likelihood ratio tests (LRT). The comparison included the likelihood estimates of the neutral null model (M7) and the alternative model of positive selection (M8). RIP alleles were excluded from these analyses since such alleles encode sequences with stop codons.

To test different hypotheses of emergence of haplotypes associated with RIP alleles (Table S6), tree topologies using concatenated DNA sequences of all the genes (*AvrLm1, AvrLm6, LmCys1, LmTrans, LmGT, LmMFS* and *LmCys2*) and non-coding, non-repetitive regions (NC1-4) were generated and compared using the Kishino–Hasegawa (KH) test [29] as implemented in PAUP* 4.0b 10. Since the RIP mechanism produces the same mutations at specific sites, it is likely that formerly unrelated nucleotide sequences converge, leading to the false impression of similarity due to common descent. To avoid this bias, all CpA to TpA and TpG to TpA nucleotide changes were removed from the data set prior to inferring the phylogenetic relationships of haplotypes. Two alternative hypotheses were then compared; (i) haplotypes containing RIP alleles were monophyletic, i.e. they emerged only once. Trees representing this hypothesis were "constrained" by restricting RIP alleles to cluster only amongst each other (ii) haplotypes containing RIP alleles were polyphyletic, i.e. they emerged several times independently. Trees representing this alternative hypothesis were "unconstrained", i.e. the pairing of particular alleles in the topology was not restricted. One thousand trees were generated representing each hypothesis, and the probabilities (*P*) of obtaining better trees were assessed using two-tailed tests, the full optimization criterion and 1000 bootstrap replicates. The KH test was conducted for trees constructed under the maximum likelihood and the maximum parsimony criterion.

The phylogenetic relationship among isolates based on the concatenated DNA sequences of all genes and non-coding, non-repetitive regions was assessed by PAUP* using maximum likelihood. Tree searches were conducted with the "fast-stepwise-addition" option and 1000 bootstrap replicates to assess statistical significance of nodes. The GTR-model with estimated substitution-rate matrix was used to evaluate molecular rate constancy.

Supporting Information

Figure S1 Confirmation of deletion alleles for *AvrLm1, AvrLm6* and *LmCys2* by Southern analysis of genomic DNA. (A) Hybridisation of an *AvrLm1* probe to genomic DNA of two isolates that produced an amplicon after PCR using primers specific for *AvrLm1* (lanes 1 and 7) and five isolates that did not (lanes 2–6). (B) Hybridisation of an *AvrLm6* probe to genomic DNA of five isolates that produced an amplicon after PCR using primers specific for *AvrLm6* (lane 1–3 and 7–8) and three isolates that did not (lanes 4–6). (C) Hybridisation of a *LmCys2* probe to genomic DNA of a single isolate that produced an amplicon after PCR using primers specific for *LmCys2* (lane 1) and one isolate that did not (lane 2). All genomic DNA was digested with *Hind*III. For three isolates, a 300 bp size difference was observed for genomic DNA fragments hybridising to the *AvrLm6* probe. Since no *Hind*III sites are present in the coding sequence of *AvrLm6* the size difference must be due to polymorphisms in *Hind*III sites outside the region analysed.
Found at: doi:10.1371/journal.ppat.1001180.s001 (1.55 MB EPS)

Figure S2 Distribution of RIP mutations across the NC3 region, *LmCys1* and *LmTrans.* The ratio of mutated CpA or TpG sites was compared to the number of CpA and TpG nucleotides present in the wild type allele over a 100 bp window for the NC3 region (A) and *LmCys1* (B) and *LmTrans* (C) genes. Regions where greater than 50% of potential RIP sites are mutated are highlighted in grey. The gene structure of *LmCys1* and *LmTrans* are represented above the respective graphs.
Found at: doi:10.1371/journal.ppat.1001180.s002 (1.19 MB EPS)

Table S1 *Leptosphaeria maculans* isolates used in this study.
Found at: doi:10.1371/journal.ppat.1001180.s003 (0.04 MB DOC)

Table S2 Primers used in this study.
Found at: doi:10.1371/journal.ppat.1001180.s004 (0.06 MB DOC)

Table S3 Predicted gene structure of open reading frames within the *AvrLm1-LmCys2* genomic region of *Leptosphaeria maculans* isolate v23.1.3.
Found at: doi:10.1371/journal.ppat.1001180.s005 (0.03 MB DOC)

Table S4 Haplotype characterisation of 295 Australian isolates of *Leptosphaeria maculans* based on alleles of *AvrLm1, AvrLm6, LmCys1* and *LmCys2*.
Found at: doi:10.1371/journal.ppat.1001180.s006 (0.05 MB DOC)

Table S5 Alleles of *LmTrans, LmGT* and *LmMFS* in 84 Australian isolates of *Leptosphaeria maculans.*
Found at: doi:10.1371/journal.ppat.1001180.s007 (0.04 MB DOC)

Table S6 Haplotype characterisation of 84 Australian isolates of *Leptosphaeria maculans* based on alleles of seven genes and four non-coding, non-repetitive regions.
Found at: doi:10.1371/journal.ppat.1001180.s008 (0.11 MB DOC)

Table S7 Changes in allele frequencies of *AvrLm1, AvrLm6, LmCys1* and *LmCys2.*
Found at: doi:10.1371/journal.ppat.1001180.s009 (0.04 MB DOC)

Acknowledgments

We thank Steve Marcroft, Horsham for providing some of the *L. maculans* isolates and Thierry Rouxel and Marie-Hélène Balesdent, INRA-Bioger, Patrick Wincker, Genoscope, and Joelle Anselem, URGI, France for access to the genome sequence of *L. maculans* isolate v23.1.3.

Author Contributions

Conceived and designed the experiments: APVdW JKH BAM RPO BJH. Performed the experiments: APVdW AJC JKH PCB. Analyzed the data: APVdW AJC JKH PCB BAM RPO BJH. Contributed reagents/materials/analysis tools: APVdW AJC BJH. Wrote the paper: APVdW JKH PCB BAM RPO BJH.

References

1. Fitt BDL, Brun H, Barbetti MJ, Rimmer SR (2006) World-wide importance of phoma stem canker (*Leptosphaeria maculans* and *L. biglobosa*) on oilseed rape (*Brassica napus*). Eur J Plant Pathol 114: 3–15.
2. Flor HH (1955) Host-parasite interactions in flax rust - its genetic and other implications. Phytopathology 45: 680–685.
3. Balesdent MH, Attard A, Kuhn ML, Rouxel T (2002) New avirulence genes in the phytopathogenic fungus *Leptosphaeria maculans*. Phytopathology 92: 1122–1133.
4. Yu F, Lydiate DJ, Rimmer SR (2005) Identification of two novel genes for blackleg resistance in *Brassica napus*. Theor Appl Genet 110: 969–979.
5. Delourme R, Chevre AM, Brun H, Rouxel T, Balesdent MH, et al. (2006) Major gene and polygenic resistance to *Leptosphaeria maculans* in oilseed rape (*Brassica napus*). Eur J Plant Pathol 114: 41–52.
6. Yu F, Lydiate DJ, Rimmer SR (2008) Identification and mapping of a third blackleg resistance locus in *Brassica napus* derived from *B. rapa* subsp. sylvestris. Genome 51: 64–72.
7. Rouxel T, Balesdent MH (2005) The stem canker (blackleg) fungus, *Leptosphaeria maculans*, enters the genomic era. Mol Plant Pathol 6: 225–241.
8. Fudal I, Ross S, Gout L, Blaise F, Kuhn ML, et al. (2007) Heterochromatin-like regions as ecological niches for avirulence genes in the *Leptosphaeria maculans* genome: map-based cloning of *AvrLm6*. Mol Plant Microbe Interact 20: 459–470.
9. Gout L, Fudal I, Kuhn ML, Blaise F, Eckert M, et al. (2006) Lost in the middle of nowhere: the *AvrLm1* avirulence gene of the Dothideomycete *Leptosphaeria maculans*. Mol Microbiol 60: 67–80.
10. Parlange F, Daverdin G, Fudal I, Kuhn ML, Balesdent MH, et al. (2009) *Leptosphaeria maculans* avirulence gene *AvrLm4-7* confers a dual recognition specificity by the *Rlm4* and *Rlm7* resistance genes of oilseed rape, and circumvents *Rlm4*-mediated recognition through a single amino acid change. Mol Microbiol 71: 851–863.
11. Hogenhout SA, Van der Hoorn RA, Terauchi R, Kamoun S (2009) Emerging concepts in effector biology of plant-associated organisms. Mol Plant Microbe Interact 22: 115–122.

12. Stukenbrock EH, McDonald BA (2009) Population genetics of fungal and oomycete effectors involved in gene-for-gene interactions. Mol Plant Microbe Interact 22: 371–380.
13. Sprague SJ, Balesdent MH, Brun H, Hayden HL, Marcroft SJ, et al. (2006) Major gene resistance in *Brassica napus* (oilseed rape) is overcome by changes in virulence of populations of *Leptosphaeria maculans* in France and Australia. Eur J Plant Pathol 114: 33–44.
14. Crouch JH, Lewis BG, Mithen RF (1994) The effect of A genome substitution on the resistance of *Brassica napus* to infection by *Leptosphaeria maculans*. Plant Breeding 112: 265–278.
15. Sprague SJ, Marcroft SJ, Hayden HL, Howlett BJ (2006) Major gene resistance to blackleg in *Brassica napus* overcome within three years of commercial production in southeastern Australia. Plant Disease 90: 190–198.
16. Van de Wouw AP, Stonard JF, Howlett BJ, West JS, Fitt BD, et al. (2010) Determining frequencies of avirulent alleles in airborne *Leptosphaeria maculans* inoculum using quantitative PCR. Plant Pathology 59: 809–818.
17. Van de Wouw AP, Marcroft SJ, Barbetti MJ, Hua L, Salisbury PA, et al. (2009) Dual control of avirulence in *Leptosphaeria maculans* towards a *Brassica napus* cultivar with 'sylvestris-derived' resistance suggests involvement of two resistance genes. Plant Pathology 58: 305–313.
18. Rouxel T, Penaud A, Pinochet X, Brun H, Gout L, et al. (2003) A 10-year survey of populations of *Leptosphaeria maculans* in France indicates a rapid adaptation towards the *Rlm1* resistance gene of oilseed rape. Eur J Plant Pathol 109: 871–881.
19. Brun H, Chevre AM, Fitt BD, Powers S, Besnard AL, et al. (2010) Quantitative resistance increases the durability of qualitative resistance to *Leptosphaeria maculans* in *Brassica napus*. New Phytol 185: 285–299.
20. Gout L, Kuhn ML, Vincenot L, Bernard-Samain S, Cattolico L, et al. (2007) Genome structure impacts molecular evolution at the *AvrLm1* avirulence locus of the plant pathogen *Leptosphaeria maculans*. Environ Microbiol 9: 2978–2992.
21. Selker EU, Garrett PW (1988) DNA sequence duplications trigger gene inactivation in *Neurospora crassa*. Proc Natl Acad Sci U S A 85: 6870–6874.
22. Fudal I, Ross S, Brun H, Besnard AL, Ermel M, et al. (2009) Repeat-induced point mutation (RIP) as an alternative mechanism of evolution toward virulence in *Leptosphaeria maculans*. Mol Plant Microbe Interact 22: 932–941.
23. Brun H, Levivier S, Somda I, Ruer D, Renard M, et al. (2000) A field method for evaluating the potential durability of new resistance sources: application to the *Leptosphaeria maculans-Brassica napus* pathosystem. Phytopathology 90: 961–966.
24. Cozijnsen AJ, Howlett BJ (2003) Characterisation of the mating-type locus of the plant pathogenic ascomycete *Leptosphaeria maculans*. Curr Genet 43: 351–357.
25. Hane JK, Oliver RP (2008) RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. BMC Bioinformatics 9: 478.
26. Rep M, van der Does HC, Meijer M, van Wijk R, Houterman PM, et al. (2004) A small, cysteine-rich protein secreted by *Fusarium oxysporum* during colonization of xylem vessels is required for I-3-mediated resistance in tomato. Mol Microbiol 53: 1373–1383.
27. Pond SL, Frost SD, Muse SV (2005) HyPhy: hypothesis testing using phylogenies. Bioinformatics 21: 676–679.
28. Tamura K, Dudley J, Nei M, Kumar S (2007) MEGA4: Molecular Evolutionary Genetics Analysis (MEGA) software version 4.0. Mol Biol Evol 24: 1596–1599.
29. Kishino H, Hasegawa M (1989) Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from DNA sequence data, and the branching order in hominoidea. J Mol Evol 29: 170–179.
30. McDonald BA, Linde C (2002) Pathogen population genetics, evolutionary potential, and durable resistance. Annu Rev Phytopathol 40: 349–379.
31. Rouxel T, Willner E, Coudard L, Balesdent MH (2003) Screening and identification of resistance to *Leptosphaeria maculans* (stem canker) in *Brassica napus* accessions. Euphytica 133: 219–231.
32. Couch BC, Fudal I, Lebrun MH, Tharreau D, Valent B, et al. (2005) Origins of host-specific populations of the blast pathogen *Magnaporthe oryzae* in crop domestication with subsequent expansion of pandemic clones on rice and weeds of rice. Genetics 170: 613–630.
33. Farman ML, Eto Y, Nakao T, Tosa Y, Nakayashiki H, et al. (2002) Analysis of the structure of the *AVR1-CO39* avirulence locus in virulent rice-infecting isolates of *Magnaporthe grisea*. Mol Plant Microbe Interact 15: 6–16.
34. Rep M (2005) Small proteins of plant-pathogenic fungi secreted during host colonization. FEMS Microbiol Lett 253: 19–27.
35. Ma LJ, van der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, et al. (2010) Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature 464: 367–373.
36. Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H, et al. (2006) Emergence of a new disease as a result of interspecific virulence gene transfer. Nat Genet 38: 953–956.
37. Liu Z, Faris JD, Oliver RP, Tan KC, Solomon PS, et al. (2009) SnTox3 acts in effector triggered susceptibility to induce disease on wheat carrying the *Snn3* gene. PLoS Pathog 5: e1000581.
38. Hane JK, Lowe RG, Solomon PS, Tan KC, Schoch CL, et al. (2007) Dothideomycete plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*. Plant Cell 19: 3347–3368.
39. Haas BJ, Kamoun S, Zody MC, Jiang RH, Handsaker RE, et al. (2009) Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. Nature 461: 393–398.
40. Farman ML (2007) Telomeres in the rice blast fungus *Magnaporthe oryzae*: the world of the end as we know it. FEMS Microbiol Lett 273: 125–132.
41. Rep M, Kistler HC (2010) The genomic organization of plant pathogenicity in *Fusarium* species. Curr Opin Plant Biol 13: 1–7.
42. Irelan JT, Hagemann AT, Selker EU (1994) High frequency repeat-induced point mutation (RIP) is not associated with efficient recombination in *Neurospora*. Genetics 138: 1093–1103.
43. Idnurm A, Howlett BJ (2003) Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete *Leptosphaeria maculans*. Fungal Genet Biol 39: 31–37.
44. Brunner PC, Stefanato FL, McDonald BA (2008) Evolution of the *CYP51* gene in *Mycosphaerella graminicola*: evidence for intragenic recombination and selective replacement. Mol Plant Pathol 9: 305–316.
45. van den Burg HA, Westerink N, Francoijs KJ, Roth R, Woestenenk E, et al. (2003) Natural disulfide bond-disrupted mutants of AVR4 of the tomato pathogen *Cladosporium fulvum* are sensitive to proteolysis, circumvent *Cf-4*-mediated resistance, but retain their chitin binding ability. J Biol Chem 278: 27340–27346.
46. Cheng J, Randall AZ, Sweredoski MJ, Baldi P (2005) SCRATCH: a protein structure and structural feature prediction server. Nucleic Acids Res 33: W72–76.
47. Schurch S, Linde CC, Knogge W, Jackson LF, McDonald BA (2004) Molecular population genetic analysis differentiates two virulence mechanisms of the fungal avirulence gene *NIP1*. Mol Plant Microbe Interact 17: 1114–1125.
48. Stukenbrock EH, McDonald BA (2007) Geographical variation and positive diversifying selection in the host-specific toxin SnToxA. Mol Plant Pathol 8: 321–332.
49. Rozen S, Skaletsky H (2000) Primer3 on the WWW for general users and for biologist programmers. Methods Mol Biol 132: 365–386.
50. Gardiner DM, Cozijnsen AJ, Wilson LM, Pedras MS, Howlett BJ (2004) The sirodesmin biosynthetic gene cluster of the plant pathogenic fungus *Leptosphaeria maculans*. Mol Microbiol 53: 1307–1318.
51. Muller PY, Janovjak H, Miserez AR, Dobbie Z (2002) Processing of gene expression data generated by quantitative real-time RT-PCR. Biotechniques 32: 1372–1374, 1376, 1378–1379.
52. Massingham T, Goldman N (2005) Detecting amino acid sites under positive selection and purifying selection. Genetics 169: 1753–1762.
53. Yang Z (2007) PAML 4: phylogenetic analysis by maximum likelihood. Mol Biol Evol 24: 1586–1591.

Appendix 5A: Response to Thesis Examination Comments

I have one question about the calculation of RIP dominance in Table 2. I don't see how the values were reached, given the frequencies of mutations listed in the table. For example, AvrLm6 allele 5 (listed with a dominance score of 3.6) exhibited 38 nt changes, 10 of which were CpA to TpA, 15 were TpG to TpA, while the remaining changes were G to A or C to T changes in other contexts. Wouldn't this give a dominance score of (10 + 15) / 13 = 1.9?

The sum of all nucleotide changes (38) for allele 5, is not the same as the sum of all CpN (or reverse complement NpG) mutations, which is a subset of this total. That is why using the (38 – (10+15) =) 13 as the denominator in the examiners example calculation produces a different result ((10+15)/13=1.9) than presented in table 2 ((10+15)/7=3.6).

For reference purposes, the RIP dominance formulas are presented in the methods section of Chapter 3.

Chapter 6: Attribution Statement

Title: **Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations**

Authors: Thierry Rouxel, Jonathan Grandaubert, **James K. Hane**, Claire Hoede, Angela P. van de Wouw, Arnaud Couloux, Victoria Dominguez, Véronique Anthouard, Pascal Bally, Salim Bourras, Anton J. Cozijnsen, Lynda M. Ciuffetti, Alexandre Degrave, Azita Dilmaghani, Laurent Duret, Isabelle Fudal, Stephen B. Goodwin, Lilian Gout, Nicolas Glaser, Juliette Linglin, Gert H. J. Kema, Nicolas Lapalu, Christopher B. Lawrence, Kim May, Michel Meyer, Bénedicte Ollivier, Julie Poulain, Conrad L. Schoch, Adeline Simon, Joseph W. Spatafora, Anna Stachowiak, B. Gillian Turgeon, Brett M. Tyler, Delphine Vincent, Jean Weissenbach, Joëlle Amselem, Hadi Quesneville, Richard P. Oliver, Patrick Wincker, Marie-Hélène Balesdent, Barbara J. Howlett

Citation: *Nature Communications* 2:202 (2011)

This thesis chapter is submitted in the form of a collaboratively-written draft manuscript which at the time of writing was accepted pending revision by *Nature Communications*. As such, not all work contained in this chapter can be attributed to the Ph. D. candidate.

The PhD candidate (JKH) made the following contributions to this chapter:

- Comparison of synteny between *Leptosphaeria maculans* and *Stagonospora nodorum* and synteny-based prediction of genome assembly finishing in *Leptosphaeria maculans*.
- Analysis of inter-genic distances in *Leptosphaeria maculans* in A:T-rich and G:C-equilibrated sequence regions.
- Characterisation of repetitive DNA families, analysis of repeat-induced point mutation (RIP) in the repetitive DNA of *Leptosphaeria maculans* and prediction of original sequences prior to RIP.
- Manual curation of ribosomal DNA repeat.

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

-- --

James Hane (Ph. D. candidate) Date

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

-- --

Richard Oliver (Principal supervisor) Date

nature COMMUNICATIONS

OPEN

ARTICLE

Received 20 Apr 2010 | Accepted 11 Jan 2011 | Published 15 Feb 2011

DOI: 10.1038/ncomms1189

Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations

Thierry Rouxel[1,*], Jonathan Grandaubert[1,*], James K. Hane[2], Claire Hoede[3], Angela P. van de Wouw[4], Arnaud Couloux[5], Victoria Dominguez[3], Véronique Anthouard[5], Pascal Bally[1], Salim Bourras[1], Anton J. Cozijnsen[4], Lynda M. Ciuffetti[6], Alexandre Degrave[1], Azita Dilmaghani[1], Laurent Duret[7], Isabelle Fudal[1], Stephen B. Goodwin[8], Lilian Gout[1], Nicolas Glaser[1], Juliette Linglin[1], Gert H. J. Kema[9], Nicolas Lapalu[3], Christopher B. Lawrence[10], Kim May[4], Michel Meyer[1], Bénédicte Ollivier[1], Julie Poulain[5], Conrad L. Schoch[11], Adeline Simon[1], Joseph W. Spatafora[6], Anna Stachowiak[12], B. Gillian Turgeon[13], Brett M. Tyler[10], Delphine Vincent[14], Jean Weissenbach[5], Joëlle Amselem[3], Hadi Quesneville[3], Richard P. Oliver[15], Patrick Wincker[5], Marie-Hélène Balesdent[1] & Barbara J. Howlett[4]

Fungi are of primary ecological, biotechnological and economic importance. Many fundamental biological processes that are shared by animals and fungi are studied in fungi due to their experimental tractability. Many fungi are pathogens or mutualists and are model systems to analyse effector genes and their mechanisms of diversification. In this study, we report the genome sequence of the phytopathogenic ascomycete *Leptosphaeria maculans* and characterize its repertoire of protein effectors. The *L. maculans* genome has an unusual bipartite structure with alternating distinct guanine and cytosine-equilibrated and adenine and thymine (AT)-rich blocks of homogenous nucleotide composition. The AT-rich blocks comprise one-third of the genome and contain effector genes and families of transposable elements, both of which are affected by repeat-induced point mutation, a fungal-specific genome defence mechanism. This genomic environment for effectors promotes rapid sequence diversification and underpins the evolutionary potential of the fungus to adapt rapidly to novel host-derived constraints.

[1] INRA-Bioger, UR1290, Avenue Lucien Brétignières, BP 01, Thiverval-Grignon F-78850, France. [2] Murdoch University, South Street, Murdoch, Western Australia 6150, Australia. [3] INRA-URGI, Route de Saint Cyr, Versailles Cedex F-78026, France. [4] School of Botany, University of Melbourne, Victoria 3010, Australia. [5] GENOSCOPE, Centre National de Séquençage, Institut de Génomique CEA/DSV, 2, rue Gaston Crémieux, CP 5706, Evry Cedex F-91057, France. [6] Department of Botany and Plant Pathology, Cordley Hall 2082, Oregon State University, Corvallis, Oregon 97331-2902, USA. [7] Laboratoire Biométrie et Biologie Evolutive, UMR CNRS 5558, Université Lyon 1, 43 Bld du 11 Novembre 1918, Villeurbanne cedex F-69622, France. [8] USDA-ARS, Crop Production and Pest Control Research Unit, Purdue University, 915 West State Street, West Lafayette, Indiana 47907-2054, USA. [9] Wageningen UR, Plant Research International, Department of Biointeractions and Plant Health, P.O. Box 69, Wageningen 6700 AB, The Netherlands. [10] Virginia Bioinformatics Institute, Virginia Polytechnic Institute and State University, Blacksburg, Virginia 24061-0477, USA. [11] NIH/NLM/NCBI, 45 Center Drive, MSC 6510, Bethesda, Maryland 20892-6510, USA. [12] Institute of Plant Genetics, Polish Academy of Sciences, Strzeszynska 34, Poznan PL-60479, Poland. [13] Deparment of Plant Pathology & Plant-Microbe Biology, Cornell University, Ithaca, New York 14853, USA. [14] INRA, UMR1202 BIOGECO, 69 Route d'Arcachon, Cestas F-33612, France. [15] Australian Centre for Necrotrophic Fungal Pathogens, Curtin University, Perth, Western Australia 6845, Australia. *These authors contributed equally to this work. Correspondence and requests for materials should be addressed to T.R. (email: rouxel@versailles.inra.fr).

Fungi are the most important pathogens of cultivated plants, causing about 20% yield losses worldwide. Such diseases are a major cause of malnutrition worldwide[1]. Their phenotypic diversity and genotypic plasticity enable fungi to adapt to new host species and farming systems and to overcome new resistance genes or chemical treatments deployed in attempts to limit losses to crop yields[2]. Along with such genotypic plasticity, natural or anthropogenic long-distance dispersal of fungi allows the emergence of novel, better-adapted phytopathogens and more damaging diseases. These processes of adaptation are exemplified by *Leptosphaeria maculans* 'brassicae' (Phyllum Ascomycota, class Dothideomycetes), which causes stem canker (blackleg) of oilseed rape (*Brassica napus*) and other crucifers. This fungus has been recorded on crucifers (mainly cabbages) since 1791, but only began to cause substantial damage to broad acre Brassica species and spread around the world in the last four decades[3]. Other phytopathogens often rapidly cause lesions on plants to ensure asexual reproduction. In contrast, *L. maculans* shows an unusually complex parasitic cycle with alternating saprotrophy associated with sexual reproduction on stem debris, necrotrophy and asexual sporulation on leaf lesions, endophytic and symptomless systemic growth, and a final necrotrophic stage at the stem base[3].

Some features of filamentous fungal genomes are remarkably constant; for instance, size &0.+4. Mb* typically about 34 Mb), gene number &/.*...+/1*...' gene content, intron size and number, and the low content of repeated sequences[4]. Comparative genomic approaches have shown that most of the candidate 'pathogenicity genes' (for example, those encoding hydrolytic enzymes that can degrade plant cell walls, or involved in formation of infection structures) analysed in the last decade in a gene-by-gene approach are shared by saprobes and pathogens[4]. These genes were probably recruited as pathogenicity factors when phytopathogens evolved from saprobes, but they do not account for host range or host specificity of phytopathogens. Such roles are played by 'effector' proteins, which modulate host innate immunity, enable parasitic infection and are generally genus, species, or even isolate-specific[5,6]. Such effector genes include those with a primary function as avirulence genes or encoding toxins or suppressors of plant defense. While bacteria produce few effectors (typically <30), which mostly seem to suppress plant innate immunity[7], hundreds of candidate effectors have been identified in oomycetes[8–10]. In fungi, in contrast, such a catalogue of effectors has only been established to-date in the hemibasidiomycete pathogen of maize, *Ustilago maydis*, in which many of the effector genes are organized as gene clusters[11].

In *L. maculans*, the only characterized effectors include a toxic secondary metabolite, sirodesmin PL[12] and the products of three avirulence genes, *AvrLm1*, *AvrLm6* and *AvrLm4-7*, of which at least one, *AvrLm4-7*, is implicated in fungal fitness[13–15]. These three avirulence genes show typical features of effector genes; that is, they are predominantly expressed early in infection, encode small proteins predicted to be secreted (SSPs) into the plant apoplast and have no or few matches in databases. Intriguingly, all three are located within large AT-rich, heterochromatin-like regions that are mostly devoid of other coding sequences[13,15].

In this paper, we describe the genome of *L. maculans*. We speculate how the genome, characterized by a distinct division into guanidine–cytosine (GC)-equilibrated and AT-rich blocks of homogenous nucleotide composition, has been reshaped following massive invasion by and subsequent degeneration of transposable elements (TEs). We also predict the repertoire of pathogenicity effectors for the first time in an ascomycete genome and we propose how the unusual genome structure may have led to the diversification and evolution of effectors.

Results

General features of the *L. maculans* genome. The haploid genome of strain v23.1.3 of *L. maculans* 'brassicae' was sequenced using a whole-genome shotgun strategy. This fungus is closely related to *Phaeosphaeria (Stagonospora) nodorum*, *Pyrenophora tritici-repentis* and *Cochliobolus heterostrophus,* as seen in the phylogeny based on sequence analysis of a range of genes (Supplementary Table S1; Fig. 1). The genome assembly had a total size of 45.12 Mb, scaffolded into 76 SuperContigs (SCs; 30 large SCs >143 kb; Tables 1 and 2; Supplementary Table S2). The correspondence of SCs to chromosomes was inferred by a combination of approaches (Fig. 2; Supplementary Figs S1 and S2). Conglomerated data are consistent with the presence of 17 or 18 chromosomes, ten of which correspond to single SCs (Supplementary Fig. S1; Supplementary Table S2).

Gene models were identified using the EuGene prediction pipeline (Supplementary Tables S3 and S4), and the resulting total of 12,469 genes is consistent with that in other Dothideomycetes (Table 2). Expression of 84.4% of predicted genes was detected using NimbleGen custom-oligoarrays in free-living mycelium or during early stages of oilseed rape infection (Table 3). About 10% of the genes were significantly overexpressed during infection (Table 3). Taking into account expressed-sequence-tag (EST), transcriptomic, and proteomic support, 84.8% of the gene models were biologically validated (Table 3). The genes are shorter than those in the other Dothideomycetes whose genomes have been sequenced (Table 2). Intergenic distances are shorter than those of *P. nodorum*, the closest relative to have been sequenced, and bi-directional promoters are common (Supplementary Table S5).

Automated finding and annotation of repeated elements in the genome using the REPET pipe-line (http://urgi.versailles.inra.fr/index.php/urgi/Tools/REPET) showed that they comprise one-third of the genome compared to 7% in *P. nodorum* (Table 2). Although most of the repeat elements are truncated and occur as mosaics of multiple families, their origin as TEs is clear (Supplementary Data S1 and S2). Class I elements (see ref. 16 and Table 4 for classification of TEs) dominate with nine families comprising 80% of the repeated elements (Table 4, Supplementary Data S1). Of these, just four families comprise 11.37 Mb, which is 25% of the genome assembly. Very few, if any, of the TEs are transcribed, as shown by EST inspection and transcriptomic analysis. TEs are clustered in blocks distributed across SCs, and the number of TE copies per SC correlates with size of the SC ($R^2=0.86$; Supplementary Fig. S3).

The TEs are RIP affected. Alignment and comparison of repeat families also showed a pattern of nucleotide substitution consisting mainly of C-to-T and G-to-A changes, suggesting the presence of repeat-induced point mutation (RIP). RIP is a premeiotic repeat-inactivation mechanism specific to fungi and has been previously experimentally identified in *L. maculans*[17]. The *L. maculans* genome possesses orthologues of all the *Neurospora crassa* genes currently postulated to be necessary for RIP[18] (Supplementary Table S6). Analysis using RIPCAL, a quantitative alignment-based method[19], indicated that C bases within CpA dinucleotides were mutated to T, more frequently than the sum of CpC, CpG and CpT dinucleotides, confirming the action of RIP on all of the TEs (Supplementary Figs S4 and S5; Supplementary Data S2).

The compartmentalized genome of *L. maculans*. The *L. maculans* genome is larger and has a lower overall GC content (44.1% GC) than those of the related Dothideomycetes *P. nodorum*, *Alternaria brassicicola*, *C. heterostrophus*, *P. tritici-repentis* or the more divergent species *Mycosphaerella graminicola* (Table 2). As previously reported for a broader range of fungi[20], the larger size is consistent with the genome having been extensively invaded by TEs. The GC content of ESTs and other known coding sequences is 50.5%, and the low genome GC content is due to the compartmentalized structure of the genome into GC-equilibrated regions (51.0% GC content, sizes between 1 and 500 kb, average 70.4 kb; henceforth denoted as GC-blocks) alternating with AT-rich regions (henceforth denoted as AT-blocks; averaging 33.9% GC content; with sizes between

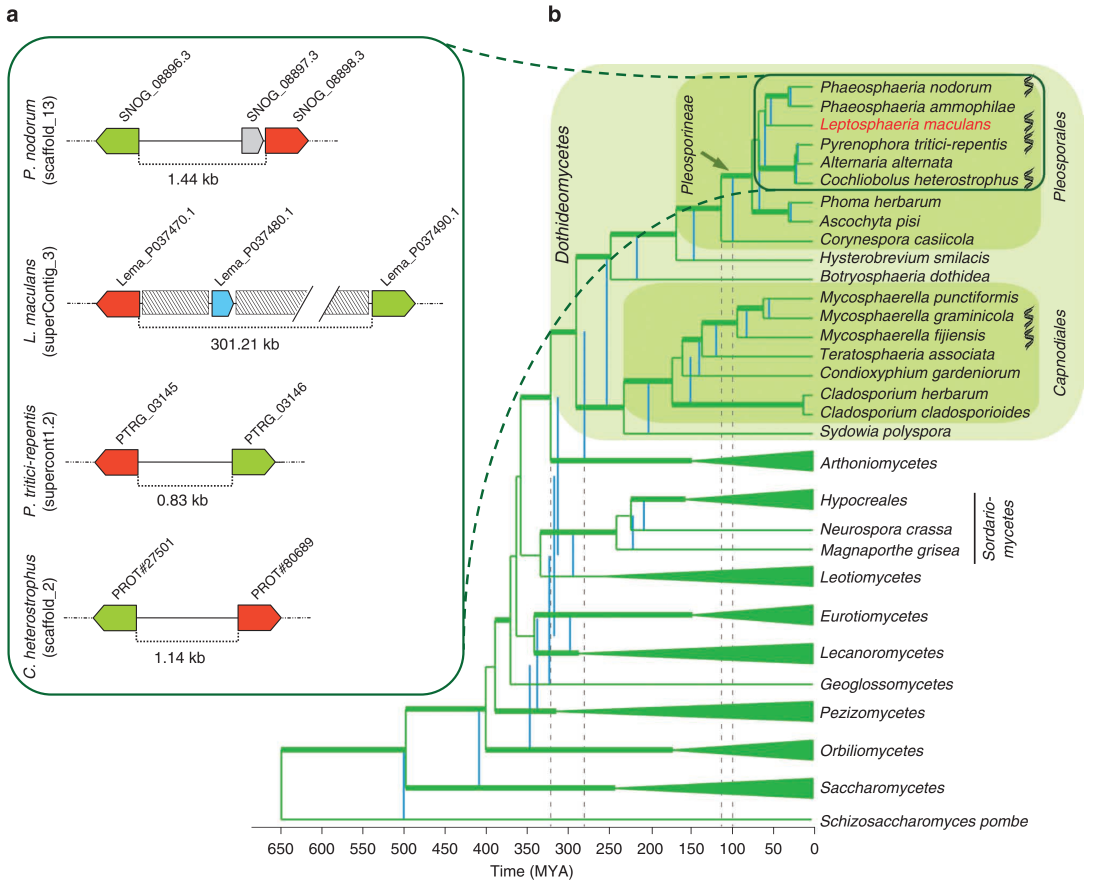


Figure 1 | Phylogenetic relationships between Dothideomycetes and an example of microsynteny between related species. (**a**) An example of microsynteny between *L. maculans* and closely related Dothideomycetes, *P. nodorum*, *C. heterostrophus* and *P. tritici-repentis*, showing the integration of an AT-rich genomic region (grey boxes) between two orthologous genes encoding for fungal transcription factors (red and green arrows) of the three other species, along with generation of one novel small-secreted protein-encoding gene (blue arrow) in *L. maculans* only. Grey arrow, *P. nodorum* predicted gene. The ID of each gene in the corresponding genome sequence is indicated. The intergenic distance (expressed in kb) is shown. (**b**) A phylogenetic tree and estimated time divergences of major lineages in Ascomycota with a selection of plant pathogenic lineages in Dothideomycetes. The phylogenetic analysis was performed using RaxML[44] and the chronogram, calibrated using recent data from the literature and fossil dates, produced using r8s (ref. 45). Classes outside of the Dothideomycetes were collapsed in TreeDyn, except for Sordariomycetes where the order *Hypocreales* represented an important calibration point. The blue vertical lines correlate with divergence times when the root of the tree was fixed at 500 MYA, whereas the green lines of the tree represent a fixed root of 650 MYA. The range of dates for the emergence of Dothideomycetes and *Pleosporineae* are highlighted with stippled lines. Thickened branches on the tree represents nodes that had more than 70% bootstrap values in a RAxML run. Species with genome data are marked with a DNA logo.

Table 1 | Assembly statistics for the *L. maculans* genome.

	SuperContigs	Contigs
Number	76	1,743
Size (Mb)	45.12	43.76
N50 (kb)	1,770	61
Min/max size (kb)	0.49/4,258.57	0.22/395.37
Mean size (kb)	594	26
Median size (kb)	29	11

1 and 320 kb, average of 38.6 kb). Whole-genome analysis identified 413 AT-blocks and 399 GC-blocks (Supplementary Table S7). The AT-blocks cover 36% of the genome and are distributed within the large SCs, comprising between 23.1 and 49.2% (Fig. 2c, Supplementary Table S7; Supplementary Fig. S6). SC22, corresponding to a minichromosome,[21] contains nine AT-blocks amounting to 92.5% of the SC (Supplementary Table S7).

As well as differences in GC content, the two types of genomic regions are dissimilar in terms of recombination frequency and gene content. The number of crossovers (CO) along a chromosome ranges between 1.16 and 3.31, depending on size of the chromosome, with one CO every 820 kb on average. The recombination frequency is significantly higher between marker pairs located within GC-blocks than those located on each side of one AT-block (*F*-Fisher = 5.873, *P* = 0.019; Fig. 2d, Supplementary Fig. S7).

GC-blocks contain 95% of the predicted genes of the genome, at a higher density (4.2 per 10 kb) than in other Dothideomycetes (Table 2) and are mostly devoid of TEs. In contrast, AT-blocks are gene-poor, comprising only 5.0% of the predicted coding sequences,

Table 2 | Features of genomes of *L. maculans* and other related Dothideomycetes.

	*L. maculans**	*P. (Stagonospora) nodorum**	*P. tritici-repentis**	*C. heterostrophus**	*A. brassicicola**	*M. graminicola**
No. of chromosomes	17–18	19	11	15–16	9–11	21
Genome size (Mb)	45.1	36.6	37.8	34.9	30.3	39.7
No of contigs	1,743	496	703	400	4,039	21
No of SuperContigs (SCs)	76	107	47	89	838	21
SC N50 (Mb)	1.8	1.1	1.9	1.3	2.4	NA†
Gaps (%)	2.5	0.4	1.7	1.1	5.4	0.01
No. of predicted genes	12,469	10,762	12,141	9,633	10,688	10,952
Average gene length (bp)	1,323	1,326	1,618	1,836	1,523	1,600
GC content (%)	44.1	50.3	50.4	52–54	50.5	55.0
Repeat content (%)	34.2	7.1	16.0	7.0	9.0	18.0
'Core' genome size (Mb)‡	29.7	34.5	31.7	32.5	27.6	32.6
Gene density/core genome (no. of gene per 10 kb)	4.2	3.1	3.8	3.0	3.9	3.4

*References for the genomes as follows: *L. maculans*[54], *P. nodorum*[55], *P. tritici-repentis*[56], *C. heterostrophus*[57], *A. brassicicola*[58], *M. graminicola*[59]; unpublished reannotation of *P. nodorum* genome was provided by J. K. Hane and R. P. Oliver.
†Not applicable, as the *M. graminicola* genome is finished; that is, each SC corresponds to a chromosome.
‡'Core' genome excluding the repeated elements, but including the gaps in the genome sequence.

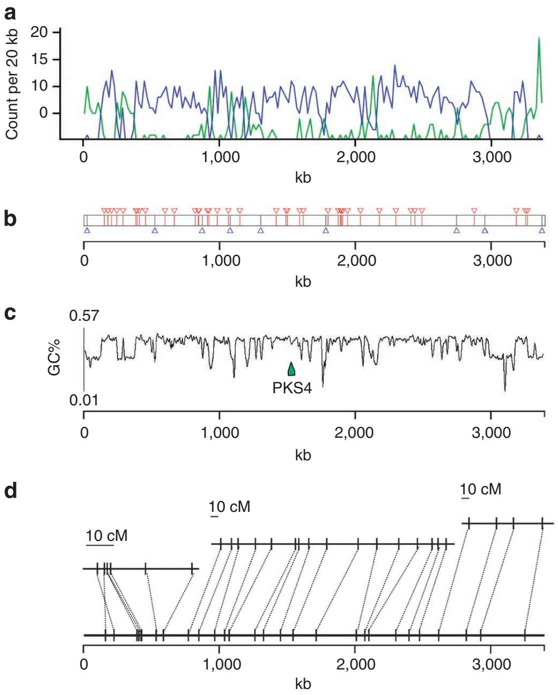


Figure 2 | Main features of the *L. maculans* genome as exemplified by chromosome 5 SuperContig 1. (**a**) Transposable elements (TEs) distribution and gene density along the supercontig. TE density is drawn in green and gene density is in blue. (**b**) Location of SSPs (small-secreted protein encoding genes). Blue arrowheads, SSP in AT-blocks, corresponding to TE-rich regions in a; red arrowheads, SSP in GC-blocks, corresponding to gene-rich genome regions in a. (**c**) GC content along the SC showing alternating GC-equilibrated and AT-rich regions, with location of a polyketide synthase-encoding gene, PKS4. (**d**) Genetic (upper part, expressed in centiMorgan—cM) to physical (expressed in kb) distance relationship as a function of the isochore-like structure. Lower part: physical location of genetic markers. Upper part of the panel: genetic map using MapMaker/Exp 3.0 with parameters set at likelihood ratio value >3.0 and minimum distance = 20 cM. Only markers drawn from the sequence data are represented.

and mainly contain mosaics of TEs mutated by RIP, thus resulting in a low-GC content of TEs. There are three categories of AT-blocks: telomeres, which include a *Penelope* retroelement[22] (Supplementary Fig. S8); large AT-blocks (216 sized 13–325 kb); and mid-sized AT-blocks (197 sized 1–13 kb; Supplementary Fig. S9), mostly corresponding to single integrations of only two families of DNA transposons (Supplementary Table S8).

In almost half of the cases where pairs of orthologues are on the same SC, the genes flanking AT-blocks in *L. maculans* have orthologues in *P. nodorum* that are either two consecutive genes or genes separated by only a few others (Fig. 1a; Supplementary Data S3). A similar pattern was observed for *C. heterostrophus* and *P. tritici-repentis*, suggesting that the TEs invaded the genome after the separation of *Leptosphaeria* from other species of suborder *Pleosporineae* 50–57 million years ago (MYA; Fig. 1b).

The ribosomal DNA repeat is extensively affected by RIP. In eukaryotes, the ribosomal DNA (rDNA) comprises a multigene family organized as large arrays of tandem repeats. The core unit is a single transcription unit that includes the 18S or Small Subunit, 5.8S, and 28S or Large Subunit separated by internal transcribed spacers (ITS1 and ITS2). Each transcription unit is separated by the Intergenic Spacer (Fig. 3a). Although essential duplicated regions would be expected to be protected from RIP mutations, the rDNA repeats in *L. maculans* are in part affected by RIP (Fig. 3b,c, Supplementary Fig. S10). The number of rDNA repeats ranges between 56 and 225 in different *L. maculans* isolates[23]. The assembly of strain v23.1.3 has >150 repeats, only two of which are highly similar (99.6% identity) and are not affected by RIP. Fifty complete rDNA units and 107 incomplete units are present, and most of them are on extreme ends of SC2 and SC19, which are not complete chromosomes. Many of these repeats are severely affected by RIP (Fig. 3, Supplementary Fig. S10). Selker[24] has suggested that rDNA repeats in the nucleolus organizer region are protected from RIP. Our data indicate that this is not the case in *L. maculans*, at least for a part of the array of tandem repeats.

AT-blocks as niches for effectors. As described above, AT-blocks have few genes. Furthermore, 76% of these genes are located close to the borders with GC-blocks; only 24% (148 genes) are located within AT-blocks (Table 3; Supplementary Data S4 and S5). Protein comparisons and Gene Ontology (GO) analysis indicate that AT-blocks are enriched in genes likely to have a role in pathogenicity (Supplementary Fig. S11). These include orphan genes such as those

Table 3 | Comparative features of SSP-encoding genes occurring in diverse genome environments.

	All predicted genes	SSPs in GC-equilibrated regions	Non-SSPs in borders*	Non-SSPs within AT-rich regions	SSPs in borders*	SSPs within AT-rich regions
No.	12,469	529 (4.2%)	407 (3.3%)	91 (0.7%)	65 (0.5%)	57 (0.5%)
BLAST hits (%)†	71.3	48.4	60.2	34.1	15.4	8.8
GC content (%)	54.1	54.6	52.9	48.9	51.1	48.2
TpA/ApT	1.04	1.20	1.12	1.49	1.19	1.44
TpA/ApT >1.5 (%)‡	6.9	16.4	11.5	36.3	20.0	38.6
EST, transcriptomic or proteomic support (%)	84.8	77.1	73.7	54.9	56.9	60.0
No. of genes present on the NimbleGen array, and with transcriptomic support	10,524	396	298	47	35	33
Genes overexpressed *in planta* 7 dpi (%)§	9.9	19.1	11.1	36.2	13.9	72.7
Genes overexpressed *in planta* 14 dpi (%)§	11.0	15.4	11.8	8.5	22.2	24.2
Average protein size (amino acid)	418.4	167.7	396.1	192.4	111.6	98.6
% Cysteines in the predicted protein	1.7	2.9	1.9	2.1	3.8	4.5

TpA/ApT, frequency of occurrence of dinucleotide TA over dinucleotide AT.

* 'Borders' refer to 859±385 bp transition regions between AT-rich and GC-equilibrated genomic regions.

†BLAST to nr cutoff=1×e^{-10}.

‡Percentage of genes showing a TpA/ApT RIP index above 1.5. This cutoff corresponds to that observed in the majority of RIP-inactivated *AvrLm6* alleles[39].

§Genes with more than 1.5-fold change in transcript level and an associated *P*-value <0.05 were considered as significantly differentially expressed during infection (7 or 14 dpi) compared with growth *in vitro*; expressed as a percent of genes with transcriptomic support. The same genes may be overexpressed at 7 and 14 dpi.

encoding SSPs, genes involved in response to chemical or biotic stimuli (Supplementary Fig. S11), as well as non-ribosomal peptide synthetases and polyketide synthases, which encode enzymes involved in biosynthesis of secondary metabolites (Supplementary Tables S9 and S10; Supplementary Fig. S12).

One hundred and twenty-two (~20%) of the genes located in AT-blocks encode putative SSPs (Table 3; Supplementary Data S4). Only 4.2% of the genes in the GC-blocks encode SSPs (529 genes), and these lack many features of known effectors of *L. maculans* (Table 3). In contrast, the SSPs encoded in AT-blocks have features indicative of effectors such as low EST support in *in vitro* grown cultures, low abundance in *in vitro* secretome samples, increased expression upon plant infection, lack of recognizable domains or homologues in other fungi, and high cysteine content (Table 3; Supplementary Data S4). Three TEs, the retrotransposon, RLx_*Ayoly*, and two DNA transposons, DTF_*Elwe* and DTx_*Gimli*, are significantly over-represented in the immediate vicinity of SSPs (Supplementary Fig. S13). Although SSPs are never embedded within a single TE, four SSPs are inserted between two tandemly repeated copies of the DNA transposon DTM_*Sahana*.

As well as the avirulence genes, two SSPs, LmCys1 and LmCys2, have been functionally analysed. LmCys1 contributes to fungal growth *in planta*, whereas LmCys2 contributes to suppression of plant defence responses, reflecting their roles as effectors (I. Fudal, unpublished data). Expression of 70.2% of the SSP-encoding genes was detected (Table 3). Of these, 72.7% of the SSP-encoding genes located within AT-blocks (compared with 19.1–22.2% in GC-blocks) were over-expressed at early stages of infection of cotyledons compared with *in vitro* mycelium growth (Table 3; Supplementary Fig. S14). Accordingly, these are postulated to be effectors. In addition, 45% of the predicted SSPs in AT-blocks show a presence/absence polymorphism in field populations, as is the case for avirulence genes in *L. maculans* and other fungi[25]. The SSPs in GC-blocks include 110 (20.8%) with best BLAST hits to hypothetical proteins from *P. nodorum*. In contrast, very few SSPs in AT-blocks have identifiable orthologues; only two (1.8%) had a best match to a predicted protein of *P. nodorum* (Supplementary Data S4). In addition to their lack of orthologues, SSPs in AT-blocks also lack paralogues; only seven genes belong to gene families comprising one to four paralogues. Biases in codon usage occur: in GC-blocks, the preferred codon for each of the 20 amino acids ends with a C or a G and the preferred stop codon is TGA, whereas in SSP genes located in AT-blocks, the preferred codon ends with an A or T for 13 amino acids and the preferred stop codon is TAA (Supplementary Table S11). This, however, only has a limited impact on amino acid favoured usage by SSPs (Supplementary Table S12).

Motifs resembling the RxLR translocation motifs of oomycetes were sought[26] following the validation that one such motif, RYWT, present in the N-terminal part of AvrLm6 allows translocation into plant and animal cells[26]. Searches for ⟨[RKH] X [LMIFYW] X⟩ or ⟨[RKH] [LMIFYW] X [RKH]⟩ showed that up to 60% of SSPs in AT-blocks and up to 73% of SSP in GC-blocks have putative 'RxLR-like' motifs, implicating these SSPs as candidate effectors that enter plant cells (Supplementary Data S4).

History of genome invasion by TEs. A range of 278–320 MYA is estimated for the origin of the *Dothideomycetes* with the crown radiation of the class during the Permian (251–289 MYA; Fig. 1b). The origins of the plant pathogenic *Pleosporineae* is determined at 97–112 MYA, placing it in the Cretaceous at a time when flowering plants were beginning to become widespread and eudicots were emerging, during the late Cretaceous and Paleocene. *Leptosphaeria* likely diverged from the other species analysed between 50 and 57 MYA (Fig. 1b). Phylogenetic analyses suggest three main features of genome invasion by TEs: transposition bursts mostly after separation of *L. maculans* from other species of suborder *Pleosporineae* as indicated by a 'recent' divergence of the TE families, estimated to 4–20 MYA (Fig. 4a); a single or few wave(s) of massive transposition(s) followed by a 'rapid' decay, with some cases like DTM_*Sahana* where divergence between copies is extremely low; and no on-going waves of genome invasion by TEs (Fig. 4b). Like other organisms with a high density of TEs, the *L. maculans* genome exhibits 'nesting', where repeats occur within previously inserted TEs. In this fungus, TEs are commonly invaded by other TEs generating a complex 'nesting network'. Eighty-five % of these cases correspond to TEs invading one other TE (primary nesting relationship). Most of the retrotransposon families investigated can invade or be invaded to similar extents (Supplementary Table S8). They also can invade TEs from the same family (self-nests), but usually at a very low frequency compared with invasion of retrotransposons from other families. In contrast, the DNA transposons are more commonly invaded (23.3% of the cases) than acting as invaders (3.5% of the cases; Supplementary Table S8).

Table 4 | Main families and characteristics of transposable elements and other repeats in the *L. maculans* genome.

	Genome coverage	Size (bp)	LTR/TIR size (bp)	Number of copies	Number of complete copies	Complete/ incomplete copies	Superfamily
*Class I (retrotransposons)**							
LTR: 9 families.	12.30 Mb (27.26%)						
RLG_*Olly*	3.06 Mb	7,246	250	1,085	187	0.172	*Ty3/Gypsy*
RLG_*Polly*	2.97 Mb	6,928	179	1,014	164	0.162	*Ty3/Gypsy*
RLG_*Rolly*	2.24 Mb	11,875	235	594	46	0.077	*Ty3/Gypsy*
RLC_*Pholy*	3.10 Mb	6,981	281	1,020	83	0.081	*Ty1/Copia*
RLG_*Dolly*	0.30 Mb	6,620	228	85	30	0.353	*Ty3/Gypsy*
RLC_*Zolly-1 and -2*	0.16 Mb	5,306	177	97	14	0.144	*Ty1/Copia*
RLx_*Jolly*	0.02 Mb	803	259	57	5	0.088	Unknown
RLx_*Ayoly*	0.40 Mb	10,397	217	164	8	0.049	Unknown
RLG_*Brawly*	0.05 Mb	7,289	None	22	3	0.136	*Ty3/Gypsy*
Class II (DNA transposons)							
TIR: 10 families	1.19 Mb (2.64%)						
DTF_*Elwe*	199.3 kb	2,173	57	158	54	0.342	*Fot1-Pogo*
DTM_*Lenwe*	25.6 kb	3,489	49	36	3	0.083	*Mutator*
DTx_*Olwe*	5.3 kb	866	49	15	4	0.267	Unknown
DTx_*Valwe*	11.5 kb	1,793	37	73	2	0.027	Unknown
DTT_*Finwe-1*	2.7 kb	529	29	7	4	0.571	*Tc1-Mariner*
DTT_*Finwe-2*	10.8 kb	523	29	31	11	0.355	*Tc1-Mariner*
DTT_*Finwe-3*	7.4 kb	806	29	15	4	0.267	*Tc1-Mariner*
DTM_*Sahana*	782.8 kb	5,992	None	873	49	0.056	*Mutator*
DTx_*Gimli*	112.9 kb	606	None	279	51	0.183	Unknown
DTM_*Ingwe*	33.4 kb	3,582	37	48	1	0.021	*Mutator*
Uncharacterized repeats (11 families)	159.9 kb (0.35%)						
rDNA repeats†	767 kb (1.70%)	7,800‡		>100	50		
Telomeric repeats§	935.0 kb (2.07%)						

* Classification of TEs according to Wicker *et al.*[16]: the three-letter code refers to class (R, retrotransposon; D, DNA transposon), order (L, Long terminal repeat—LTR; T, terminal inverted repeat—TIR; P, *Penelope*-like element—PLE) and superfamily (G, *Gypsy*; C, *Copia*; P, *Penelope*; F, *Fot1-Pogo*; T, *Tc1-Mariner*; M, *Mutator*, x, unknown superfamily) followed by the family (or subfamily) name italicized.
†Including a rDNA-specific LINE element.
‡Excluding variable length short-tandem repeats flanking almost every rDNA repeat.
§Including telomere-associated *Penelope*-like retroelement RPP-*Circe* and RecQ telomere-linked helicase.

In accordance with overlapping divergence time estimates (Figs 1b and 4), these data indicate periods of overlapping transpositional activity for the long terminal repeats retrotransposons that form the major part of AT-blocks. In such a scenario, the later insertions would be preferentially tolerated in existing decayed transposons. These TEs, having undergone RIP in their turn, would initiate a positive reinforcement loop that would create large AT-rich and gene-poor blocks of homogeneous nucleotide composition.

Discussion

The peculiar genomic structure of the *L. maculans* genome is reminiscent of that discovered in mammals and some other vertebrates: the base composition (GC-content) varies widely along chromosomes, but locally, base composition is relatively homogenous. Such structural features have been termed 'isochores'[27]. In *L. maculans*, AT-blocks are gene-poor, rich in TEs and deficient in recombination compared with GC-blocks, as in mammals[27]. However, despite these similarities, these genomic landscapes seem to result from different mechanisms. In mammals, the evolution of GC-rich isochores is most likely driven by recombination: genomic regions sized between 100 kb to several Mb with a high recombination rate tend to increase in GC content relative to the rest of the genome. This pattern is not due to a mutational effect of recombination, but most probably due to biased gene conversion[28]. In *L. maculans*, variations in base composition occur at a much finer scale (the isochore-like blocks are about 10–20 times smaller than in mammals), and it is unknown whether biased gene conversion contributes to increase the GC content of GC-blocks. Conversely, *L. maculans* isochores can be attributed to the AT-biased mutational pattern induced by RIP mutation of TEs and their flanking regions, thus leading to the evolution of AT-rich isochores.

Although the evolutionary forces we postulate shaped the *L. maculans* genome are common to many species, no fungal genome characterized so far has a similar isochore-like structure. This structure reflects extensive genome invasion by TEs that are nonetheless tolerated by the pathogen and the existence of an active RIP machinery (Supplementary Table S6) that has so far been restricted to the Pezizomycotina subphylum of the Ascomycota and maintenance of sexual reproduction (necessary for RIP). Whereas many species seem to have maintained an active RIP machinery, most of the sequenced fungal genomes are poor in TEs, indicating that run-away genome expansion is normally deleterious. Also, many fungal species have lost the ability to cross in nature (for example, *Fusarium oxysporum*, *Magnaporthe oryzae*) and no case of large-scale sculpting of repeat-rich regions is found in these species, only some ancient signatures of RIP are found[29].

On the basis of the characteristics of avirulence genes in *L. maculans*, we have described a comprehensive repertoire of putative effectors, which has not previously been done for an Ascomycete. In *L. maculans*, AT-rich blocks are enriched in effector-like sequences. Location of effector genes has been investigated in only some eukaryotic genomes. A few of the effectors of *M. oryzae* are

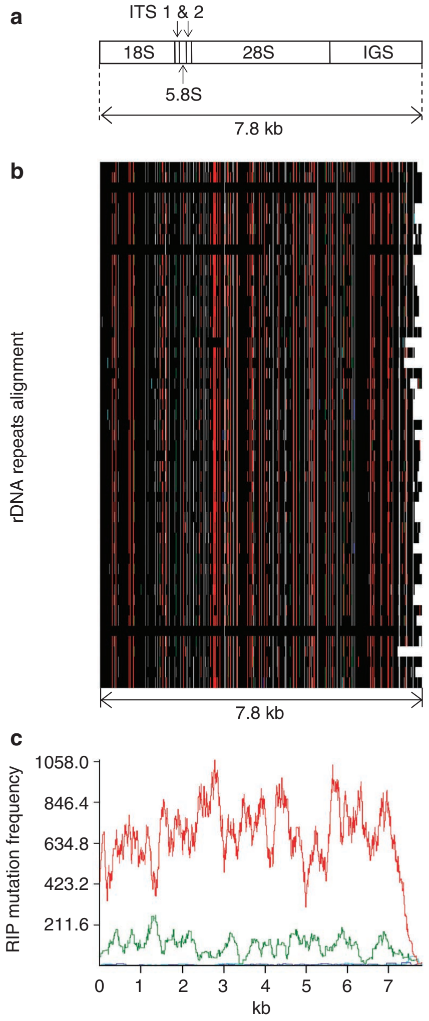


Figure 3 | Repeat-induced point (RIP) mutation in ribosomal DNA of *L. maculans* shown as RIPCAL output. (**a**) Schematic representation of the rDNA unit in *L. maculans* (ITS, internal transcribed spacers; IGS, intergenic spacer); (**b**) a schematic multiple alignment of the 7.8 kb 'complete' ribosomal DNA (rDNA) units occurring in SuperContigs 2 and 19. Polymorphic nucleotides are coloured as a function of the type of RIP mutation observed, with black, invariant nucleotide; red, CpA ←→ TpA or TpG ←→ TpA mutations; dark blue, CpC ←→ TpC or GpG ←→ GpA mutations; pale blue, CpT ←→ TpT or ApG ←→ ApA mutations; green, CpG ←→ TpG or CpG ←→ CpA mutations; (**c**) RIP mutation frequency plot over a rolling sequence window, corresponding to the multiple alignment directly above. Nucleotide polymorphisms (against the alignment consensus, which is also the highest GC-content sequence) mostly correspond to CpA ←→ TpA or TpG ←→ TpA (red curve) and CpG ←→ TpG or CpG ←→ CpA (green curve).

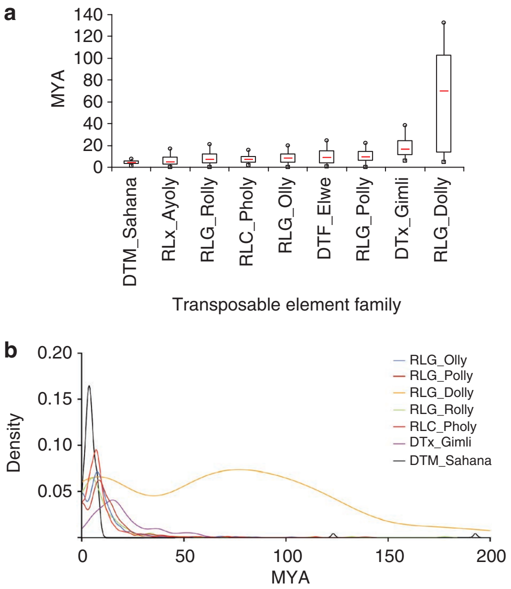


Figure 4 | Dynamics of transposable elements in the *L. maculans* genome. A phylogenetic analysis was used to retrace the evolutionary history of each transposable element (TE) family after elimination of mutations due to repeat-induced point mutations. Terminal fork branch lengths were assumed to correspond to an evolutionary distance used to estimate the age of the last transposition activity. The divergence values were converted to estimated divergence time using a substitution rate of 1.05×10^{-9} substitution per location per year[52,53] (expressed as 'million years ago' MYA). (**a**) Box plot graph of divergence times. The red line represents the median value; the boxes include values between the first and the third quartile of the distribution; squares and circles, first and ninth decile, respectively. (**b**) Kernel density of divergence plots. A R-script was written to plot a histogram of the terminal fork branch length with kernel density estimate for each family.

subtelomeric[30], as are those in protozoan parasites of animals, such as *Plasmodium* and *Trypanosoma*[31]. Genomes of many *Fusarium* species contain supernumerary 'B' chromosomes enriched in strain-specific effectors and accounting for the host range of each '*forma specialis*'[32,33]. The genome of the oomycete *Phytophthora infestans* has a plethora of effector candidates embedded in repetitive DNA, and diversification of these effectors is postulated to occur via segmental duplication and variation in intraspecific copy number, resulting in rapidly diverging multigene families[8]. The association between one family of effectors and a LINE in *Blumeria graminis*, the barley powdery mildew fungus, is proposed to provide a mechanism for amplifying and diversifying effectors[34]. Diversification of effectors in the species mentioned above is postulated to be associated with TE-driven gene duplication and generation of multigene families. In the *L. maculans* genome, SSP-encoding genes are associated with only a few TE families, which may indicate the ability of TEs to 'pickup and move' effectors. In contrast to the above examples, duplicated effector genes are not present in *L. maculans*, a finding consistent with the steady inactivation of TEs by RIP and with ancient transposition activity before underdoing RIP.

The origins of some effector genes might be at least partially ascribed to lateral gene transfer, a specialty of species within the Pezizomycotina[35–37]. Regardless of the origin of the effector genes, our data suggest that RIP is an important mechanism for generating diversity for genes occurring within AT-blocks of the genome of *L. maculans*, in a manner not previously documented in any other species. RIP has previously been reported to be restricted to duplicated DNA, but most SSPs or other genes in AT-blocks are present in single copies. How, then, can RIP act on SSP-encoding genes?

Studies in *N. crassa* indicated that the RIP machinery can occasionally overrun the repeated region into adjacent single-copy genes[38]. The embedding of SSPs within RIP-degenerated TEs would then favour such RIP leakage (Supplementary Fig. S4c), while selection pressure to maintain functional effectors would prevent them from becoming extinct due to an excessive degree of RIP. This would result in extensive mutation of the affected gene and could account for the mutation rate required for diversifying selection. In contrast, effector genes that became detrimental to pathogen fitness, such as avirulence genes subjected to resistance gene selection, would be lost rapidly as alleles that have undergone extensive RIP are selected for[39]. Evidence for this scenario is provided by examination of RIP indices and in alignment-based studies of alleles of SSPs[39]. The genes (including SSPs) within AT-blocks had higher TpA/ApT indices than those in GC-blocks (Table 3; Supplementary Fig. S15), consistent with former genes having been RIP affected. RIP indices for the effectors located within AT-blocks thus would be a compromise between values leading to complete degeneration of the sequence and values enabling sequence diversification while retaining functionality. In plant-pathogen systems, diversifying selection operates on effector genes whose products interact with host proteins[25]. This has been demonstrated for both resistance and avirulence genes, but mechanisms for the diversifying selection of effectors have not been proposed. RIP is shown here to be a potential factor to create the genetic (hyper)variation needed for selection to occur in *L. maculans*, and this process may also act on effectors in other fungi[40].

These findings allow speculation about an evolutionary scenario for birth of isochore-like structures in the *L. maculans* genome and its incidence on effector diversification. First, the genome was invaded by a few families of TEs over a (relatively) short time period, mostly after the separation of *L. maculans* from other related fungi. This TE invasion is unlikely to have been targeted to pre-existing effector-rich genome regions as seen in microsynteny analyses (Fig. 1a). Accordingly, the most recent invader, DTM_*Sahana*, is not specifically associated with SSPs. Second, waves of overlapping transposition occurred with probable transduction, translation or duplication of genes, resulting in the large amplification of a few families. Such transpositions were primarily targeted to other TEs as shown by the nesting of retrotransposons within other TEs. In parallel, duplicated copies of TEs and genes (either duplicated or not) hosted within TE-rich regions underwent RIP either to extinction for TEs or to generate gene diversity in cases where a strong selection pressure to retain genes was exerted. This eventually resulted in complete inactivation of transposition events, and the sculpting of the genome in an isochore-like structure. Effector genes were maintained in AT-blocks to favour rapid response to selection pressure[39,41] and probable epigenetic concerted regulation of their expression (Supplementary Fig. S14b). *L. maculans* shows intriguing evolutionary convergence with both higher eukaryotes in terms of an isochore-like genome structure, and with oomycetes in terms of hosting effectors in highly dynamic 'plastic' regions of the genome[8]. It differs in exploiting a RIP-based mechanism for diversification and inactivation of effector genes.

The sequencing of genomes of several species or subspecies of the recent and more ancient outgroups that derived from a common ancestor with *L. maculans* will provide more information on origin of effectors, genome invasion by TEs and the subsequent effect on generation/diversification of effectors, and thus test the validity of the proposed evolutionary scenario.

Methods

Phylogenetic analysis. A taxon set containing representatives of most classes in Ascomycota was selected from the data matrices produced in two previous papers[42,43]. Sequences were concatenated from the Small Subunit and Large Subunit of the nuclear ribosomal RNA genes and three protein coding genes, namely the translation elongation factor-1α and the largest and second largest subunits of RNA polymerase II (Supplementary Table S1). A phylogenetic analysis was performed using RAxML v. 7.0.4 (ref. 44) applying unique model parameters for each gene and codon. A combined bootstrap and maximum likelihood (ML) tree search was performed in RAxML with 500 pseudo replicates. The best scoring ML tree was analysed in the program R8sv1.7 (ref. 45) to produce a chronogram (Fig. 1b).

Sequencing and assembly. *L. maculans* 'brassicae' isolate v23.1.3. was sequenced because it harbours numerous avirulence genes, three of which have been cloned by a map-based strategy involving large-scale sequencing of surrounding genomic regions[13,15,41]. Isolate v23.1.3. results from a series of *in vitro* crosses between European field isolates[46] and is representative of the populations of the pathogen prevalent in Europe in the mid-1990s.

DNA was provided as agarose plugs containing partly digested conidia[21]. Whole-genome shotgun sequencing of three types of libraries (high-copy-number plasmids with 3.3 kb inserts; low-copy-number plasmids with 10 kb inserts and fosmids with inserts 35 or 40 kb; Supplementary Table S13) was performed, and also six cDNA libraries, including ones derived from infected plants, were sequenced (Supplementary Table S14). Sequencing reads were assembled using Arachne[47] (Table 1) and the correspondence of SCs to chromosome was inferred by aligning the genetic map to the genome sequence, hybridization of single-copy markers to chromosomal DNA separated by pulsed-field gel electrophoresis, identification of telomere-specific repeats, and by mesosynteny analyses (conserved gene content) with genomes of other Dothideomycetes (Supplementary Table S2).

***L. maculans* genome annotation**. Automated structural annotation of the genome was performed using the URGI genomic annotation platform, including pipelines, databases and interfaces, developed or locally set up for fungi. The EuGene prediction pipeline v. 3.5a (ref. 48), which integrates *ab initio* (Eugene_IMM, SpliceMachine and Fgenesh 2.6 (www.softberry.com)) and similarity methods (BLASTn, GenomeThreader, BLASTx), was used to predict gene models. The functional annotation pipeline was run using InterProScan[49]. Genome assembly and annotations are available at INRA (http://urgi.versailles.inra.fr/index.php/urgi/Species/Leptosphaeria).

Genome assemblies together with predicted gene models and annotations were deposited at DNA European Molecular Biology Laboratory/GenBank under the accession numbers FP929064 to FP929139 (SC assembly and annotations). ESTs were submitted to dbEST under accession nos. FQ032836 to FQ073829.

Full description and associated references for sequencing, assembly and gene annotation are provided as Supplementary Methods.

Annotation and analysis of repeated elements. TEs[16] were identified and annotated using the 'REPET' pipeline (http://urgi.versailles.inra.fr/index.php/urgi/Tools/REPET), optimized to better annotate nested and fragmented TEs. Repeats were searched with BLASTER for an all-by-all BLASTn genome comparison, clustered with GROUPER, RECON and PILER, and consensuses built with the MAP multiple sequence alignment program. Consensuses were classified with BLASTER matches, using tBLASTx and BLASTx against the Repbase Update databank[50] and by identification of structural features such as long terminal repeats, terminal inverted repeats[16] and so on. Additional steps of clustering and manual curation of data were performed, resulting in a series of consensuses used as an input for the REPET annotation pipeline part, comprising the TE detection software BLASTER, RepeatMasker and Censor, and the satellite detection softwares RepeatMasker, TRF and Mreps.

Analysis of the dynamics of genome invasion by TEs was first based on phylogenetic analysis of each family of repeats, retracing the evolutionary history, regardless of truncation, insertion in other TEs and deletion events[51]. After elimination of all RIP targets, the tree topology was used to retrace the dynamics and demography of TE invasion in the genome. Terminal fork branch lengths from the trees were used to calculate the age of the last transposition events of the copies in the genome. The divergence values were converted in estimated divergence time using a substitution rate of 1.05×10^{-9} nucleotide per site per year as applied to fungi[52,53].

Dynamics of TE aggregation over time was also analysed by a visual analysis of nesting relationships between TEs. Following the long join annotation, mosaics of TEs were visualised using Artemis v. 12.0 (http://www.sanger.ac.uk/Software/Artemis/) in SC0-22 and a data matrix recording the frequency with which a given TE family was inserted into another one (invader) and the frequency with which one given TE was recipient of an insertion from one or multiple other TEs (invaded TE) was generated (Supplementary Table S8). The statistical identification and significance of the favoured invasion of other TE families as compared with random association was evaluated with a χ^2-test for given probabilities with simulated *P*-values, based on 20,000 replicates, as implemented in R.

RIP and DeRIP analyses. Automated analysis of RIP in *L. maculans* genomic DNA repeats was performed using RIPCAL (http://www.sourceforge.net/projects/ripcal), a software tool that performs both RIP index and alignment-based analyses[19]. In addition, RIP indices such as TpA/ApT and (CpA + TpG)/(ApC + GpT) were used to evaluate the effect of RIP on genes or genome regions for which multiple alignments could not be generated. DeRIP analyses, which predict putative ancient pre-RIP sequences, were performed using an updated version of RIPCAL, including the Perl script 'deripcal' and ripcal_summarise.

Analysis of AT-blocks. AT- and GC-blocks were manually discriminated from each scaffold using Artemis (http://www.sanger.ac.uk/Software/Artemis/), and a Python script was used to extract sequences and features of AT-blocks. TE content of AT- and GC-blocks was analysed using the REPET pipeline. Size distribution of AT-blocks, occurrence of AT-blocks on chromosomes and relationship between AT-blocks, TE content and chromosome length were calculated.

To evaluate meiotic recombination differences between AT- and GC-blocks, micro- and minisatellites located either in GC-blocks or located on both sides of a single AT-block were mapped in a reference cross, and the number of CO between two consecutive markers was calculated. The recombination frequency between two successive markers was calculated, plotted against the physical distance between the two markers and subjected to analysis of variance and a non-parametric test (Mann–Whitney test) using XLStat, to compare recombination frequencies between and within GC-blocks.

Intergenic distances were compared between AT- and GC-blocks in *L. maculans*, and also compared with those of the closely related Dothideomycete, *P. nodorum* (Supplementary Table S5).

GO annotations were compared between genes occurring in AT- and GC-blocks using the blast2GO program.

Identification and features of SSPs. Non-repeated regions within AT-blocks were identified following masking of TEs with RepeatMasker. The EMBOSS:GETORF program was used on these genomic regions to refine the identification of genes encoding SSPs with a size limit set at 600 amino acids (lower limit: 60 amino acids). A dedicated script combined the outputs of GETORF, FgeneSH and EuGene and a pipeline written in Python screened the predicted proteins according to their size and the presence of signal peptide and transmembrane domains (SignalP 3.0, TargetP and TMHMM). Base composition of the genes encoding SSPs (percent of each base in the sequence, GC content and GC3 content) and amino-acid count of the SSPs (as % of each amino acid in the protein) were calculated by custom Python scripts. Statistical bias in amino acid occurrence was evaluated by an *F*-test to determine if the variances were equal in both sets, followed by Student's *t*-test (95% confidence level) to compare the mean use of each amino acid in each set of predicted proteins. Biases in codon usage were evaluated using EMBOSS: CHIPS. A χ^2-test for given probabilities with simulated values (20,000 replicates) as implemented in R was performed to test random association of SSP-encoding genes in AT-blocks with specific TEs. Motifs similar to the RxLR motif necessary for oomycete effectors to be translocated within plant cells were sought in predicted SSPs, using a Python script aiming at identification of motifs (⟨[RKH] X [LMIFYW] X⟩ or ⟨[RKH] [LMIFYW] X [RKH]⟩).

Analysis of expression patterns of SSP-encoding genes were compared between *in vitro* (mycelium grown in axenic medium) and *in planta* (3, 7 and 14 days after inoculation of oilseed rape cotyledons), either using the *L. maculans* whole-genome expression array (manufactured by NimbleGen Systems) or by quantitative reverse transcription-PCR on a selected subset of SSP-encoding genes.

References

1. Skamniotia, P. & Gurr, S. J. Against the grain: safeguarding rice from rice blast disease. *Trends Biotechnol.* **27,** 141–150 (2009).
2. Oliver, R. P. & Solomon, P. S. Recent fungal diseases of crop plants: is lateral gene transfer a common theme? *Mol. Plant-Microbe Interact.* **21,** 287–293 (2008).
3. Rouxel, T. & Balesdent, M. H. The stem canker (blackleg) fungus, *Leptosphaeria maculans*, enters the genomic era. *Mol. Plant Pathol.* **6,** 225–241 (2005).
4. Soanes, D. N. *et al.* Comparative genome analysis of filamentous fungi reveals gene family expansions associated with fungal pathogenesis. *Plos One* **3,** 1–15 (2008).
5. Stergiopoulos, I. & de Wit, P. J. G. M. Fungal effector proteins. *Annu. Rev. Phytopathol.* **47,** 233–263 (2009).
6. Rouxel, T. & Balesdent, M. H. Avirulence genes. in: *Encyclopedia of Life Sciences (ELS)* (John Wiley & Sons, 2010) (doi: 10.1002/9780470015902.a00212672010).
7. Alfano, J. R. Roadmap for future research on plant pathogen effectors. *Mol. Plant Pathol.* **10,** 805–813 (2009).
8. Haas, B. J. *et al.* Genome sequence and analysis of the Irish potato famine pathogen *Phytophthora infestans*. *Nature* **461,** 393–398 (2009).
9. Jiang, R. H. Y., Tripathy, S., Govers, F. & Tyler, B. M. RXLR effector reservoir in two *Phytophthora* species is dominated by a single rapidly evolving superfamily with more than 700 members. *Proc. Natl Acad. Sci. USA* **105,** 4874–4879 (2008).
10. Tyler, B. M. *et al. Phytophthora* genome sequences uncover evolutionary origins and mechanisms of pathogenesis. *Science* **313,** 1261–1266 (2006).
11. Kämper, J. *et al.* Insights from the genome of the biotrophic fungal plant pathogen *Ustilago maydis*. *Nature* **444,** 97–101 (2006).
12. Elliott, C. E., Gardiner, D. M., Thomas, G., Cozijnsen, A. J., van de Wouw, A. & Howlett, B. J. Production of the toxin sirodesmin PL by *Leptosphaeria maculans* during infection of *Brassica napus*. *Mol. Plant Pathol.* **8,** 791–802 (2007).
13. Fudal, I. *et al.* Heterochromatin-like regions as ecological niches for avirulence genes in the *Leptosphaeria maculans* genome: map-based cloning of *AvrLm6*. *Mol. Plant-Microbe Interact.* **20,** 459–470 (2007).
14. Huang, Y. J., Li, Z. Q., Evans, N., Rouxel, T., Fitt, B. D. L. & Balesdent, M. H. Fitness cost associated with loss of the *AvrLm4* function in *Leptosphaeria maculans* (Phoma stem canker of oilseed rape). *Eur. J. Plant Pathol.* **114,** 77–89 (2006).
15. Parlange, F. *et al. Leptosphaeria maculans* avirulence gene *AvrLm4-7* confers a dual recognition specificity by *Rlm4* and *Rlm7* resistance genes of oilseed rape, and circumvents *Rlm4*-mediated recognition through a single amino acid change. *Mol. Microbiol.* **71,** 851–863 (2009).
16. Wicker, T. *et al.* A unified classification system for eukaryotic transposable elements. *Nat. Rev. Genet.* **8,** 973–982 (2007).
17. Idnurm, A. & Howlett, B. J. Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete *Leptosphaeria maculans*. *Fungal Genet. Biol.* **39,** 31–37 (2003).
18. Espagne, E. *et al.* The genome sequence of the model ascomycete fungus *Podospora anserina*. *Genome Biol.* **9,** R77 (2008).
19. Hane, J. K. & Oliver, R. P. RIPCAL: a tool for alignment-based analysis of repeat-induced point mutations in fungal genomic sequences. *BMC Bioinformatics* **9,** 478 (2008).
20. Feschotte, C., Keswani, U., Ranganathan, N., Guibotsy, M. L. & Levine, D. Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in Eukaryotic genomes. *Genome Biol. Evol.* **1,** 205–220 (2009).
21. Leclair, S., Ansan-Melayah, D., Rouxel, T. & Balesdent, M. H. Meiotic behaviour of the minichromosome in the phytopathogenic ascomycete *Leptosphaeria maculans*. *Curr. Genet.* **30,** 541–548 (1996).
22. Gladyshev, E. A. & Arkhipova, I. R. Telomere-associated endonuclease-deficient *Penelope*-like retroelements in diverse eukaryotes. *Proc. Natl Acad. Sci USA* **104,** 9352–9357 (2007).
23. Howlett, B. J., Cozijnsen, A. J. & Rolls, B. D. Organisation of ribosomal DNA in the ascomycete *Leptosphaeria maculans*. *Microbiol. Res.* **152,** 1–7 (1997).
24. Selker, E. U. Premeiotic instability of repeated sequences in *Neurospora crassa*. *Annu. Rev. Genet.* **24,** 579–613 &/990).
25. Stukenbrock, E. H. & McDonald, B. A. Population genetics of fungal and oomycete effectors involved in gene-for-gene interactions. *Mol. Plant-Microbe Interact.* **22,** 371–380 (2009).
26. Kale, S. D. *et al.* External lipid PI-3-P mediates entry of eukaryotic pathogen effectors into plant and animal host cells. *Cell* **142,** 284–295 (2010).
27. Eyre-Walker, A. & Hurst, L. D. The evolution of isochores. *Nat. Rev. Genet.* **2,** 549 (2001).
28. Duret, L. & Galtier, N. Biased gene conversion and the evolution of mammalian genomic landscapes. *Annu. Rev. Genomics Hum. Genet.* **10,** 285–311 (2009).
29. Ikeda, K. *et al.* Repeat-induced point mutation (RIP) in *Magnaporthe grisea*: implications for its sexual cycle in the natural field context. *Mol. Microbiol.* **45,** 1355–1364 (2002).
30. Farman, M. L. Telomeres in the rice blast fungus *Magnaporthe oryzae*: the world of the end as we know it. *FEMS Microbiol. Lett.* **273,** 125–132 (2007).
31. Pain, A. *et al.* The genome of the simian and human malaria parasite *Plasmodium knowlesi*. *Nature* **455,** 799–803 (2008).
32. Ma, L. J. *et al.* Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium oxysporum*. *Nature* **464,** 367–373 (2010).
33. Coleman, J. J. *et al.* The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion. *PloS Genet.* **5,** e1000618 (2009).
34. Sacristan, S. *et al.* Coevolution between a family of parasite virulence effectors and a class of LINE-1 retrotransposons. *PloS One* **4,** e7463 (2009).
35. Friesen, T. L. *et al.* Emergence of a new disease as a result of interspecific virulence gene transfer. *Nat. Genet.* **38,** 953–956 (2006).
36. Marcet-Houben, M. & Gabaldón, T. Acquisition of prokaryotic genes by fungal genomes. *Trends Genet.* **26,** 5–8 (2010).
37. Khaldi, N. & Wolfe, K. H. Elusive origins of the extra genes in *Aspergillus oryzae*. *Plos One* **3,** e3036 (2008).
38. Irelan, J. T., Hagemann, A. T. & Selker, E. U. High frequency repeat-induced point mutation (RIP) is not associated with efficient recombination in *Neurospora*. *Genetics* **138,** 1093–1103 (1994).
39. Fudal, I. *et al.* Repeat-induced point mutation (RIP) as an alternative mechanism of evolution towards virulence in *Leptosphaeria maculans*. *Mol. Plant-Microbe Interact.* **22,** 932–941 (2009).
40. Stergiopoulos, I., De Kock, M. J. D., Lindhout, P. & de Wit, P.J.G.M. Allelic variation in the effector genes of the tomato pathogen *Cladosporium fulvum* reveals different modes of adaptive evolution. *Mol. Plant-Microbe Interact.* **20,** 1271–1283 (2007).
41. Gout, L. *et al.* Genome structure impacts molecular evolution at the AvrLm1 avirulence locus of the plant pathogen *Leptosphaeria maculans*. *Environ. Microbiol.* **9,** 2978–2992 (2007).

42. Schoch, C. L. *et al.* A class-wide phylogenetic assessment of *Dothideomycetes*. *Stud. Mycol.* **64,** 1–15S10 (2009).
43. Schoch, C. L. *et al.* The *Ascomycota* Tree of Life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. *Syst. Biol.* **58,** 224–239 (2009).
44. Stamatakis, A. RAxML-VI-HPC: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics* **22,** 2688–2690 (2006).
45. Sanderson, M. J. r8s: inferring absolute rates of molecular evolution and divergence times in the absence of a molecular clock. *Bioinformatics* **19,** 301–302 (2003).
46. Balesdent, M. H., Attard, A., Ansan-Melayah, D., Delourme, R., Renard, M. & Rouxel, T. Genetic control and host range of avirulence toward *Brassica napus* cultivars Quinta and Jet Neuf in *Leptosphaeria maculans*. *Phytopathology* **91,** 70–76 (2001).
47. Jaffe, D. B. *et al.* Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res.* **13,** 91–96 (2003).
48. Foissac, S. *et al.* Genome annotation in plants and fungi: EuGene as a model platform. *Curr Bioinformatics* **3,** 87–97 (2008).
49. Quevillon, E. *et al.* InterProScan: protein domains identifier. *Nucleic Acids Res.* **33** (Suppl. 2), W116–W120 (2005).
50. Jurka, J., Kapitonov, V. V., Pavlicek, A., Klonowski, P., Kohany, O. & Walichiewicz, J. Repbase Update, a database of eukaryotic repetitive elements. *Cytogenet. Genome Res.* **110,** 462–467 (2005).
51. Fiston-Lavier, A. S. Etude de la dynamique des répétitions dans les génomes eucaryotes: de leur formation à leur élimination. PhD Thesis, University Pierre et Marie Curie, Paris, France (2008).
52. Berbee, M. L. & Taylor, J. W. Dating the molecular clock in fungi—how close are we? *Fungal Biol. Rev.* **24,** 1–16 (2010).
53. Kasuga, T., White, T. J. & Taylor, J. W. Estimation of nucleotide substitution rates in Eurotiomicete fungi. *Mol. Biol. Evol.* **19,** 2318–2324 (2002).
54. Rouxel, T., Balesdent, M.H., Amselem, J. & Howlett, B. J. GnpGenome: a Genome Browser for *Leptosphaeria maculans* structural annotation (2010)http://urgi.versailles.inra.fr/cgi-bin/gbrowse/lmaculans_pub/.
55. Hane, J. K. *et al.* Dothideomycete plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*. *Plant Cell* **19,** 3347–3368 (2007).
56. Cuiffetti, L. M. *Pyrenophora tritici-repentis* database (2008) http://www.broadinstitute.org/annotation/genome/pyrenophora_tritici_repentis/Home.html.
57. Turgeon, B. G. *Cochliobolus heterostrophus* C5 whole genome project (2008) http://genome.jgi-psf.org/CocheC5_1/CocheC5_1.home.html.
58. Lawrence, C. B. *Alternaria brassicicola* whole genome project (2006) http://genome.jgi-psf.org/Altbr1/Altbr1.home.html.
59. Goodwin, S. B. & Kema, G. H. J. *Mycosphaerella graminicola* whole genome project (2008) http://genome.jgi-psf.org/Mycgr3/Mycgr3.home.html.

Acknowledgments

We acknowledge Marc-Henri Lebrun (INRA-Bioger) and Francis Martin (INRA, Interactions arbres/micro-organismes, Champenoux, France) for support and advice. The genome sequencing of *L. maculans* was funded by the Genoscope, Institut de Génomique, CEA, France. The establishment of databases and interfaces was funded by Agence Nationale de la Recherche (GnpAnnot project; ANR-07-GPLA-051G). Whole-genome effector analysis was funded by ANR (FungEffector project; ANR-06-BLAN-0399). Recombination analysis and P.B. were funded by ANR (AvirLep project; ANR-07-GPLA-015). B.J.H and R.P.O. thank the Australian Grains Research and Development Corporation for funding. C.L.S was supported in part by the Intramural Research Program of the NIH, National Library of Medicine. J.W.S acknowledges support from the U.S. National Science Foundation, grant number DEB-0717476. B.M.T. was supported in part by the U.S. National Science Foundation, grant number IOS-0924861. Thanks are due to INRA-SPE department, the '*Leptosphaeria maculans*' scientific and applied community, and the 'Dothideomycete' community for support of the *L. maculans* genome initiative.

Author contributions

J.K.H., C.H., A.P.vdW, A.C. and V.D. contributed equally to this work as second authors. J.A., H.Q., R.P.O., P.W., M.H.B. and B.J.H. coordinated genome sequencing, annotation and data analyses, and made equivalent contributions as senior authors. Individual contributions were as follows: T.R., M.H.B. and B.J.H. initiated the sequencing project; M.H.B was responsible for DNA production; P.W. and J.W. coordinated the sequencing; J.P., A.C. and V.A. performed the sequencing and assembled the genome; J.A. was responsible for annotation pipelines, databases and interfaces; V.D., C.H., and J.A. did the *ab initio* annotation of gene models; M.M., A.J.C. and B.J.H. provided EST/cDNA information; and V.D., C.H. and J.A. did the cDNA clustering, defined the training set for *ab initio* gene finder steps and inserted annotation data in the database. Genome statistics were performed by J.A., H.Q., and J.G. J.K.H., R.P.O. and J.G. carried out the genome synteny analyses; M.H.B., P.B., L.G., A.J.C., A.P.vdW., A.St., and J.G. identified and designed mini- and microsatellite markers; M.H.B and A.P.vdW. built the genetic maps; J.K.H. and R.P.O. performed mesosynteny and RIPCAL analyses; A.J.C. and K.M. hybridized electrokaryotypes and annotated NRPSs and PKSs; H.Q., V.D., L.G. and T.R. analysed TEs; V.D. and J.G. estimated time for TE transposition events; J.G., T.R. and M.H.B. analysed TE nesting; N.L. performed automated functional analysis; N.L. and S.B. performed GO analyses; B.M.T. and J.G. carried out RXLR analysis of effector candidates; I.F., A.Si., J.G., B.O. and J.L. designed microarrays and analysed microarray data; A.De., B.O., N.G. and I.F. performed expression analysis of effectors by reverse transcription-PCR and quantitative reverse transcription-PCR; A.Di. analysed polymorphism of effectors in field populations. D.V. carried out proteomic and secretomic analyses. L.M.C., S.B.G., C.B.L., G.H.J.K. and B.G.T. contributed to comparative genomics approaches. L.D. analysed isochores-like blocks and contributed to comparative analysis with those of mammals. C.L.S. and J.W.S. performed phylogenetic analyses and estimated divergence time. T.R. organized co-ordination between groups. T.R. wrote and edited the paper with major input from B.J.H. and R.P.O. Final editing of the text, Tables and Figures was done by S.B., J.G., M.H.B., B.J.H. and T.R.

Additional information

Acccession codes: Genome assemblies together with predicted gene models and annotations have been deposited in the DNA European Molecular Biology Laboratory/GenBank nucleotide core database under the accession numbers FP929064 to FP929139 (SC assembly and annotations). ESTs were submitted to dbEST under accession numbers FQ032836 to FQ073829. Oligo array data has been deposited in the gene Expression Omnibus under accession code GSE27152.

Supplementary Information accompanies this paper at http://www.nature.com/naturecommunications

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Rouxel, T. *et al.* Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by repeat-induced point mutations. *Nat. Commun.* 2:202 doi: 10.1038/ncomms1189 (2011).

Chapter 7: Attribution Statement

Title:	**Deep proteogenomics; high throughput gene validation by multidimensional liquid chromatography and mass spectrometry of proteins from the fungal wheat pathogen *Stagonospora nodorum*.**
Authors:	Scott Bringans, **James K. Hane**, Tammy Casey, Kar-Chun Tan, Richard Lipscombe, Peter S. Solomon and Richard P. Oliver
Citation:	*BMC Bioinformatics* 10:301 (2009)

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 (JKH) made the following contributions to this chapter:

- Performed all bioinformatics analyses described in this chapter.
- Co-wrote the manuscript.

The following contributions were made by co-authors:

- KCT performed fungal culture, protein extraction and sample preparation.
- SB, TC and RL performed mass-spectrometry analysis.
- PSS and RPO devised the experimental plan and provided intellectual input.
- JKH and RPO co-wrote the manuscript.
- All authors read and approved the manuscript.

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

-- --

James Hane (Ph. D. candidate) Date

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

-- --

Richard Oliver (Principal supervisor) Date

BMC Bioinformatics

Research article

Open Access

Deep proteogenomics; high throughput gene validation by multidimensional liquid chromatography and mass spectrometry of proteins from the fungal wheat pathogen *Stagonospora nodorum*

Scott Bringans[1], James K Hane[2], Tammy Casey[1,3], Kar-Chun Tan[2], Richard Lipscombe[1,3], Peter S Solomon*[4] and Richard P Oliver[2]

Address: [1]Proteomics International, Perth, WA, Australia, [2]Australian Centre for Necrotrophic Fungal Pathogens, Murdoch University, Perth, WA, Australia, [3]Centre for Food and Genomic Medicine, Western Australian Institute for Medical Research, Perth, WA, Australia and [4]Plant Cell Biology, Research School of Biology, The Australian National University, Canberra 0200, ACT, Australia

Email: Scott Bringans - scott@proteomics.com.au; James K Hane - J.Hane@murdoch.edu.au; Tammy Casey - tammy@proteomics.com.au; Kar-Chun Tan - Kar-Chun.Tan@murdoch.edu.au; Richard Lipscombe - richard@proteomics.com.au;
Peter S Solomon* - peter.solomon@anu.edu.au; Richard P Oliver - R.Oliver@murdoch.edu.au

* Corresponding author

Published: 22 September 2009

BMC Bioinformatics 2009, **10**:301 doi:10.1186/1471-2105-10-301

Received: 3 July 2009
Accepted: 22 September 2009

This article is available from: http://www.biomedcentral.com/1471-2105/10/301

Abstract

Background: *Stagonospora nodorum*, a fungal ascomycete in the class dothideomycetes, is a damaging pathogen of wheat. It is a model for necrotrophic fungi that cause necrotic symptoms via the interaction of multiple effector proteins with cultivar-specific receptors. A draft genome sequence and annotation was published in 2007. A second-pass gene prediction using a training set of 795 fully EST-supported genes predicted a total of 10762 version 2 nuclear-encoded genes, with an additional 5354 less reliable version 1 genes also retained.

Results: In this study, we subjected soluble mycelial proteins to proteolysis followed by 2D LC MALDI-MS/MS. Comparison of the detected peptides with the gene models validated 2134 genes. 62% of these genes (1324) were not supported by prior EST evidence. Of the 2134 validated genes, all but 188 were version 2 annotations. Statistical analysis of the validated gene models revealed a preponderance of cytoplasmic and nuclear localised proteins, and proteins with intracellular-associated GO terms. These statistical associations are consistent with the source of the peptides used in the study. Comparison with a 6-frame translation of the *S. nodorum* genome assembly confirmed 905 existing gene annotations (including 119 not previously confirmed) and provided evidence supporting 144 genes with coding exon frameshift modifications, 604 genes with extensions of coding exons into annotated introns or untranslated regions (UTRs), 3 new gene annotations which were supported by tblastn to NR, and 44 potential new genes residing within un-assembled regions of the genome.

Conclusion: We conclude that 2D LC MALDI-MS/MS is a powerful, rapid and economical tool to aid in the annotation of fungal genomic assemblies.

Background

The primary goal of most, if not all, genome sequence projects is to elucidate the gene, and hence protein, content of the organism. The gene set is the key tool to elucidate the interesting biological aspects of the organism. The prediction of genes from assembled genomic data has traditionally relied on two types of data; sequenced transcripts and homology to gene sequences in related organisms. Based on these data, various *in silico* methods to predict gene models can be applied. Experience in intensively studied model organisms suggests that such methods still struggle to provide a reliable gene set. As more and more genome sequences of distantly related non-model species become available, the need for efficient, rapid and accurate methods of gene prediction becomes more and more pronounced.

Just as gene sequences can predict protein sequences, protein sequences can predict gene sequences [1]. Until recently, all methods to analyse peptide sequences in complex mixtures were too slow and too expensive to be considered a viable method of whole genome gene-model validation. Transcriptomic methods have been core to gene prediction as they can efficiently identify transcribed regions and define intron-exon boundaries. Proteomic methods focus on the translated regions of genes. They have been used to identify processed N and C termini of proteins and provide information about post-translational modifications [2-4]. Proteomics has also been used to measure the quantity of each protein because this is only poorly predicted by the quantity of transcript [3,5]. Proteomic analyses have hitherto been used to provide specific and complementary information about cellular function, but were not used as a primary method of gene annotation. Recent developments in proteomic techniques, using liquid chromatography, have begun to challenge the speed, cost and efficiency of gene validation by transcriptomics. A number of recent studies have reported the use of LC-based high-throughput proteomics to assist in refining genome annotation [6-9].

Stagonospora (syn. Septoria or Phaeosphaeria) nodorum is a major fungal pathogen of wheat in many parts of the world, causing Stagonospora nodorum blotch (SNB). In Western Australia it currently causes greater than $100 m losses per annum corresponding to 9% of the yield [10,11]. It is a member of the class dothideomycetes, a taxon that includes many important crop pathogen genera such as *Leptosphaeria, Mycosphaerella* and *Pyrenophora* [12]. A draft 37.1 Mbp genome assembly of a West Australian isolate (called SN15) was obtained in 2005. A total of 380,000 Sanger reads were obtained, corresponding to about 10× coverage. The reads were assembled into 496 contigs, 107 scaffolds and the mitochondrial genome. The total amount of gaps was estimated at 154 kb. A total of 15,455 reads were not included in the assembly [12].

The assembly was searched for genes using the Broad annotation pipeline. The pipeline used the then available EST data (just 317 manually curated transcripts) and conserved homology. This predicted 16,957 genes. This number was significantly higher than expected and so the annotation was checked by the acquisition of 21,503 EST sequences, principally from two libraries; an axenic library with oleate as carbon source and an *in planta* library made from heavily infected leaves with sporulating colonies. A total of 795 genes were manually annotated from comprehensive EST evidence. A second gene prediction run using Unveil predicted 10,762 nuclear and 14 mitochondrial version 2 (v2) genes [12,13]. One gene was present on the un-assembled reads. Of the nuclear genes, 2696 were supported by EST data. A further 5354 genes not supported by the second round of gene prediction were tentatively retained as version 1 (v1) gene models.

This uncertainty in gene content was hampering research efforts particularly because, in this case, homology-based gene prediction methods were unreliable. *S. nodorum* was the first dothideomycete to be sequenced and the nearest sequenced relative organisms are separated by 400 Mya [12]. A number of approaches to improving the confidence in gene models could be envisaged. In this study, we analysed soluble mycelial proteins using 2D LC-MS/MS to generate a library of mass spectra. The data were used to verify our current gene prediction models. In addition, we generated six-frame translations of the assembled and un-assembled genome sequences to facilitate the discovery of unidentified genes and correction of current coding exon boundaries and frame assignment. To our knowledge, this study is the first that describes the extensive use of high-throughput proteomics in assisting gene annotation of a phytopathogenic fungus. The data supports 2253 genes, a number comparable to that supported by an extensive EST project. It also highlighted many potential gene model problems and identified new gene candidates.

Methods

Growth and Maintenance of Stagonospora nodorum

S. nodorum SN15 and *gna1* strains were maintained on CZV8CS agar as previously described [14]. These two strains were chosen as part of a relative quantitation analysis coupled to this proteome mapping experiment. For proteomic analysis, 100 mg of fungal mycelia were inoculated into minimal medium broth supplemented with 25 mM glucose as the sole carbon source. The fungi were grown to a vegetative state by incubation at 22°C with shaking at 150 rpm for three days. Vegetative mycelia were harvested via cheesecloth filtration and freeze-dried overnight.

Protein Extraction

Soluble intracellular proteins were extracted from freeze-dried mycelia as previously described [2]. Briefly, freeze-dried mycelia were mechanically broken with a cooled mortar and pestle and proteins were solubilised with 10 mM Tris-Cl (pH 7.5). The crude homogenate was collected and centrifuged at 20,000 *g* for 15 min at 4°C. The resulting supernatant was retained and treated with nucleases to remove nucleic acids. All protein samples were checked via SDS-PAGE to ensure that proteolysis was minimal during sample preparation (data not shown).

Sample Preparation

Proteins from SN15 and *gna1-35* strains were precipitated individually by adding five volumes of acetone, incubating for 1 hour at -20°C and pulse centrifuging for 5-10 seconds. The protein pellets were resuspended in 0.5 M triethylammonium bicarbonate (TEAB) (pH 8.5) before reduction and alkylation according to the iTRAQ protocol (Applied Biosystems, Foster City, CA, USA). Samples were centrifuged at 13,000 *g* for 10 min at room temperature before the supernatant was removed and assayed for protein concentration (Bio-Rad protein assay kit, Hercules, CA, USA). A total of 55 μg of each sample was digested overnight with 5.5 μg trypsin at 37°C. Each digest was desalted on a Strata-X 33 μm polymeric reverse phase column (Phenomenex, Torrance, CA, USA) and dried. The entire experiment was performed in triplicate (including the generation of ground mycelia).

Strong Cation Exchange Chromatography

Dried peptides were dissolved in 70 μl of 2% acetonitrile and 0.05% trifluoroacetic acid (TFA) and separated by strong cation exchange chromatography on an Agilent 1100 HPLC system (Agilent Technologies, Palo Alto, CA, USA) using a PolySulfoethyl column (4.6 × 100 mm, 5 μm, 300 Å, Nest Group, Southborough, MA, USA). Peptides were eluted with a linear gradient of Buffer B (1 M KCl, 10% acetonitrile and 10 mM KH_2PO_4, pH 3). A total of 37 fractions were collected, pooled into 8 fractions, desalted, dried and resuspended in 20 μl of 2% acetonitrile and 0.05% TFA.

Reverse Phase Nano LC MALDI-MS/MS

Peptides were separated on a C18 PepMap100, 3 μm column (LC Packings, Sunnyvale, CA, USA) with a gradient of acetonitrile in 0.1% formic acid using the Ultimate 3000 nano HPLC system (LC Packings-Dionex, Sunnyvale, CA, USA). The eluent was mixed with matrix solution (5 mg/ml α-cyano-4-hydroxycinnamic acid) and spotted onto a 384 well Opti-TOF plate (Applied Biosystems, Framingham, MA, USA) using a Probot Micro Fraction Collector (LC Packings, San Francisco, CA, USA).

Peptides were analysed on a 4800 MALDI-TOF/TOF mass spectrometer (Applied Biosystems, Framingham, MA, USA) operated in reflector positive mode. MS data were acquired over a mass range of 800-4000 *m/z* and for each spectrum a total of 400 shots were accumulated. A job-wide interpretation method selected the 20 most intense precursor ions above a signal/noise ratio of 20 from each spectrum for MS/MS acquisition but only in the spot where their intensity was at its peak. MS/MS spectra were acquired with 4000 laser shots per selected ion with a mass range of 60 to the precursor ion -20.

Data Analysis

Mass spectral data from all three biological replicates were combined and analysed using the Mascot sequence matching software (Matrix Science, Boston, USA) with the support of the facilities at the Australian Proteomics Computational Facility (Victoria, Australia). Search parameters were: Enzyme, Trypsin; Max missed cleavages, 1; Fixed modifications, iTRAQ4plex (K), iTRAQ4plex(N-term), Methylthio(C); Variable modifications, Oxidation(M); Peptide tol, 0.6 Da; MS/MS tol, 0.6 Da. The MOWSE algorithm (MudPIT scoring) of Mascot was used to score the significance of peptide/protein matches with $p < 0.05$ for each protein identification. Four protein datasets were constructed for proteogenomic screening: the combination of version 1 and 2 proteins as defined from annotation of the SN15 genome sequence [12]; a between-stop codon 6-frame translation of the *S. nodorum* genome assembly; 6-frame translated, CAP3-generated [15] contigs of un-assembled reads of the *S. nodorum* assembly, and; 6-frame translated singleton un-assembled reads which did not assemble into contigs via CAP3. All 6-frame open reading frames (ORFs) were subject to a 10 amino acid minimum length threshold.

For the purpose of false discovery rate (FDR) calculation, randomised sequences from the version 1 and 2 proteins and the 6-frame translated assembly protein datasets were generated as Mascot decoy databases [16] (as detailed at http://www.matrixscience.com/help/decoy_help.html).

Characterisation of peptide-supported genes

Peptide supported genes were analysed for abundance of assigned gene ontology (GO) terms [12]. Gene counts for GO terms were compared between peptide supported and unsupported genes via Fisher's exact test. A p-value threshold of 0.05 was imposed to determine significance. Gene counts for SignalP [17] and WolfPsort [18] cellular location predictions and relative molecular mass predictions were also compared by this method.

De novo proteogenomics

MudPIT-filtered peptide matches to the 6-frame translated assembly were mapped back to their genomic location. Peptides mapping in the same orientation with either overlapping genomic coordinates or within the proximity of 200 bp were combined as peptide clusters (referred to herein as peptide clusters). The purpose of peptide cluster formation was merely to reduce the redundancy in the peptide data to aid in the interpretation of subsequent comparisons with annotated gene features, therefore clusters with a single peptide were retained. Individual peptides and peptide clusters were compared for overlap and proximity within 200 bp to *S. nodorum* version 1 and 2 genes.

Potential homologs to *S. nodorum* SN15 genes were detected by tblastn comparison of the genome assembly with the proteomes of the dothideomycete fungi *Leptosphaeria maculans, Pyrenophora tritici-repentis, Cochliobolus heterostrophus, Alternaria brassicicola, Mycosphaerella graminicola* and *Mycosphaerella fijiensis*. Tblastn high-scoring pairs (HSPs) were grouped according to hit, but also subject to additional criteria: best HSP e-value < 1e-10; individual HSP e-values < 1e-5; HSPs mapped on *S. nodorum* genome no further than 2 kb apart or split into subgroups each subject to the previous criteria. Grouped HSP sequence coordinates were compared to both peptide cluster and annotated gene coordinates on the *S. nodorum* genome assembly for overlap (Figure 1). By this method we detected peptide clusters which could be linked back to a nearby gene model through a shared homolog or peptide clusters representing potential new gene annotations with homology support.

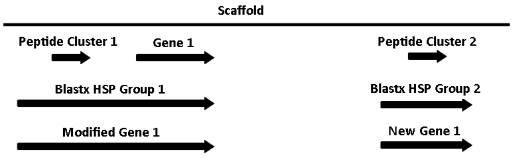


Figure 1
Peptide clusters which did not confirm an existing gene model or conflict with an existing gene model in the opposing orientation were compared to grouped blastx HSPs from related dothideomycete proteins. Some of these peptide clusters were linked back to existing gene models as illustrated by Peptide Cluster 1 and Gene 1, which share a homology relationship with Blastx HSP group 1. This provides strong evidence for the reannotation of Gene 1 to become Modified Gene 1. Other peptide clusters could not be linked to existing gene annotations such as Peptide Cluster 2, which provides evidence for the creation of New Gene 2.

Results and Discussion

Gene models confirmed by Mascot analysis of the existing gene model database

Comparison of the detected peptides against a database built from both v1 and v2 predictions of SN15 genes matched a total of 2134 gene sequences with high confidence (Tables 1, 2 and 3). Of these, all but 188 were v2 genes (Figure 2; new and modified genes are listed in Additional Files 1 and 2). The results indicate the greater reliability of the v2 prediction. The proteomic analysis matched 1324 genes that were not directly supported by any of the 21,503 EST sequences (Figure 2). This clearly shows that proteomics targets a complementary set of gene products to transcriptomic approaches. The false discovery rate (FDR) for the mass spectra matched against *S. nodorum* proteins was determined to be 13% by the Mascot decoy method. While this is relatively high, the purpose of this study was the discovery of genes not detectable by conventional techniques. Hence we favoured sensitivity over accuracy.

The genes validated by the 2D-LC-MS/MS procedure were searched for features that were overrepresented compared to the total set of predicted genes. Peptide-supported genes predicted by WolfPsort to be localised in the cytoplasm were over-represented whereas cytoskeletal, extra-

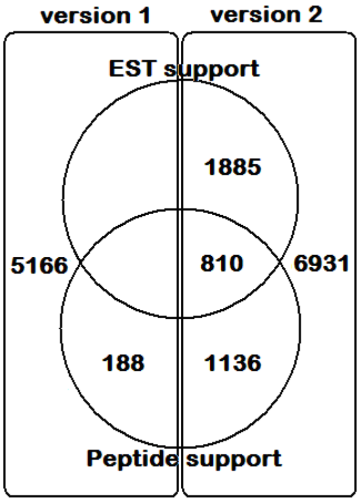


Figure 2
Comparison of *S. nodorum* annotated gene versions and confirmation by either MASCOT peptide matching or EST alignment. Version 2 genes are derived from EST alignments and a second round of EST-trained gene predictions. Version 2 genes and are considered to be more reliable than the remaining tentative version 1 gene annotations.

Table 1: Sub-cellular localisation predictions using WolfPsort of proteins over-represented within the Mascot peptide supported gene annotations of *S. nodorum*.[1]

Localisation	Peptide supported genes (all)	Peptide supported genes (v1)	Peptide supported genes (v2)	Unsupported genes	Expected genes	Total Genes
cytoplasm	605	24	581	2118	361	2723
cytoplasm/ nucleus	87	7	80	491	77	578
other	1442	157	1285	11346	1094	12788
unassigned	0	0	0	27	4	27
TOTAL	**2134**	**188**	**1946**	**13982**	**2136**	**16116**

[1] Numbers of peptide supported genes in each category were compared to expected counts from a random sampling of the whole genome via Fisher's exact test at a significance threshold of <0.05.

Table 2: Sub-cellular localisation predictions using SignalP of proteins over-represented within the Mascot peptide supported gene annotations of *S. nodorum*.[1]

	Peptide supported genes (all)	Peptide supported genes (v1)	Peptide supported genes (v2)	Unsupported genes	Expected genes	Total Genes
non-secretory protein	1847	163	1684	11259	1773	13106
signal anchor OR signal peptide	287	25	262	2723	436	3010
TOTAL	**2134**	**188**	**1946**	**13982**	**2209**	**16116**

[1] Numbers of peptide supported genes in each category were compared to expected counts from a random sampling of the whole genome via Fisher's exact test at a significance threshold of <0.05.

Table 3: Relative molecular masses of predicted proteins significantly over-represented within the Mascot peptide supported gene annotations of *S. nodorum*.[1]

	Peptide supported genes (all)	Peptide supported genes (v1)	Peptide supported genes (v2)	Unsupported genes	Expected genes	Total Genes
0 to 20 kDa	284	60	224	4495	633	4779
20 to 100 kDa	1536	121	1415	8816	1371	10352
100 to 500 kDa	310	7	303	669	130	979
>500 kDa	4	0	4	2	1	6
TOTAL	**2134**	**188**	**1946**	**13982**	**2135**	**16116**

[1] Numbers of peptide supported genes in each category were compared to expected counts from a random sampling of the whole genome via Fisher's exact test at a significance threshold of <0.05.

Table 4: Summary of 12947 6-frame translated genome-mapped peptides and 1840 peptide clusters corresponding to annotated gene features categorised by either direct overlap or close proximity (within 200 bp).

Match Type	Overlap	Within 200 bp	No match	Total
Peptide-CDS[1]	11635	323	989	**12947**
Peptide cluster-CDS	1520	54	266	**1840**

[1] CDS features are coding exons (gene sequence from translation start to stop, excluding introns). The number of proximal peptide and peptide cluster matches is greatly reduced in whole gene comparisons relative to CDS matches, indicating significant mis-annotations in EST-supported UTR regions and/or intron regions.

cellular, mitochondrial, peroxisomal and plasma membrane samples were under-represented. Peptides with predicted masses greater than 20 kDa were over-represented. Detailed analyses of GO-terms over- and under-represented are given in Additional Files 3 and 4. These analyses are consistent with the source of the peptides used in this analysis in which only soluble, intracellular proteins were isolated and analysed. The data indicates that using extracellular, membrane and cell-wall material would significantly increase the number of proteins detected. If a similar proportion of protein detection was achieved for these proteins as was for the soluble, cytoplasmic proteins, the number of extra genes detected would be about 1400.

Gene models confirmed by Mascot analysis of the 6-frame translated genome database

The proteomic data can also be used to search for gene models with errors such as inaccurate exon-intron boundaries and missed or superfluous exons as well as entirely unsuspected genes. We approached this by mapping peptides onto the 6-frame translation of the nuclear genome assembly. The coordinates of overlapping or nearby mapped peptides were amalgamated into peptide clusters Peptide clusters were defined so that internal gaps between constituent peptides of up to 200 bp were permitted, which is about four times the average intron size in *S. nodorum* [12].

Initially the mapped-peptides were compared to existing annotated coding exons (Table 4). Because of the uncertainty in defining the ends of exons, hits within 200 bp were also scrutinised. A total of 12947 spectra mapped to the genome assembly. Of these 11635 mapped within coding exons, 323 mapped within 200 bp of a coding exon and 989 mapped distantly from known genes.

A similar analysis was also carried out by comparing peptide clusters to the genome assembly (Figure 1). A total of 1840 peptide clusters were analysed. Of these, 1520 mapped within coding exons, 54 mapped within 200 bp and 266 mapped distantly from known genes. The 1520 confirmed peptide clusters correspond to 905 genes (Table 5). Of these, 119 were not identified by the previous Mascot analysis of the gene models, bringing the total of confirmed genes to 2253 (Additional File 2). The genes identified by 6-frame analysis corresponded overwhelmingly to v2 genes. Only 41 were v1 genes and 30 of these exhibited potential conflicts with the current gene model. All genes with EST support had previously been designated v2. In these cases, it was possible to define un-translated (UTR) regions, within both introns and terminal exons. The v2 confirmed genes included 300 without conflict and 564 with potential exon conflicts, divided between genes with (355) and without (209) EST support.

These analyses so far have merely mapped 6-frame translated genome-mapped peptide or peptide cluster coordinates to exon or gene coordinates. Next we considered the predicted open reading frames of the exons and the individual mapped peptides overlapping annotated coding exons (Table 6). Reassuringly, 11224 peptides matched exactly to the predicted frame whereas 482 entirely matched to a different frame. When analysed by gene, 715 were fully confirmed; of these 13 were v1 genes and 702 were v2. In a total of 144 genes, there were frame mis-

Table 5: Counts of *S. nodorum* version 1 and 2 gene annotations matching 6-frame translated genome-mapped peptide clusters.[1]

Annotation version	Confirmed no conflict	UTR/intron conflict (EST support)	UTR/intron conflict (no EST support)	No-match	Total
1	11	0	30	5313	5354
2	300	355	209	9898	10762

[1]Un-translated region (UTR)/intron conflicts were determined based on peptide matches outside of coding exon regions but within either 200 bp or within the boundaries of known UTRs. UTRs were known for genes for which EST alignments were available, where UTR regions were defined as EST-aligned regions not corresponding to coding exons or introns. 41% of peptide-supported version 2 (reliable) genes with UTR regions confirmed by EST support have suspected UTR/intron conflicts (355 out of 864).

Table 6: Summary of frame conflicts within coding-exon (CDS) annotations confirmed by overlapping 6-frame translated genome-mapped peptides.

TOTAL peptide-CDS matches in frame		11224
TOTAL peptide-CDS matches out of frame		482
Genes with all peptide matches to CDS in frame		715
	Version 1	13
	Version 2	702
Genes with all peptide matches to CDS out of frame		86
	Version 1	10
	Version 2	76
Genes with peptide matches to CDS both in and out of frame		58
	Version 1	1
	Version 2	57

The majority of peptide matches agree with current coding-exon frames, however there are 144 (86+58) gene annotations requiring frame reassessment.

matches detected by all (86) or some (58) of the supporting peptides.

Gene models identified by Mascot analysis of the 6 frame translated un-assembled read database

Finally, the mass spectra were compared to the collection of un-assembled DNA reads. The 15,455 reads were re-assessed for overlapping sequence missed during the first genome assembly by clustering into contigs using CAP3 [15]. 4616 reads were clustered into 939 contigs, with 10,839 singleton reads remaining. 423 contigs and 651 singleton DNA reads matched peptides that were detected by MS.

Sifting new and modified gene models by homology criteria

The probabilistic nature of the matching of the peptides to the genome is reflected in the false discovery rate, estimated at 13% for the matches against existing gene models. Thus not all the gene confirmations, the conflicts with existing models or the potential new genes can be expected to survive scrutiny.

Overall, these data suggest the confirmation of 2254 previously defined v1 and v2 genes (Figure 3). Of these, about 594 (355 + 30 + 209, Table 5) are identified as having questionable exon boundaries and 144 (86 + 58) as having questionable frame assignments (Table 6). Confirmation or rejection of the new gene models will require gene-by-gene sifting of the evidence; however this analysis remains an effective method of highlighting gene annotations with potential problems.

Analysis of the assembly identified a potential 266 peptide clusters that mapped distant from known genes and represent candidate novel genes. Assessment of the

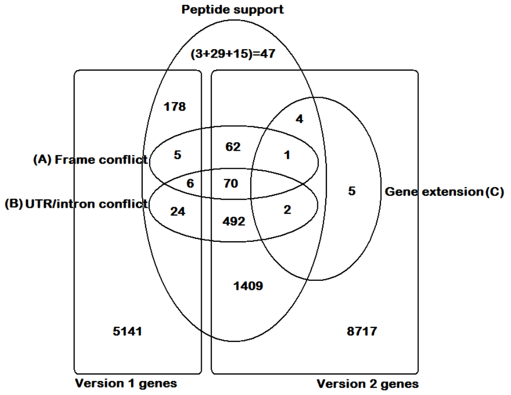


Figure 3
Summary of version 1 and 2 *S. nodorum* genes confirmed and modified by peptide support (either by conventional protein database or by 6-frame genome translation-derived peptide matches). Genes were identified as candidates for re-annotation if the 6-frame translated genome-matched peptides indicated: conflicts in annotated coding-exons open reading frames (A); peptide-genome matches residing within annotated introns or untranslated regions (UTRs) (B) or; peptide-genome matches matched to the genome which could be linked back to a gene model via tblastn homology between the genome sequence and selected dothideomycete genomes (C). 47 new gene candidates were identified by a multiple methods: 3 peptide clusters which could not be linked to an existing gene annotation via a tblastn homolog; 29 unassembled read-contigs matching dothideomycete proteins via blastx but not matching *S. nodorum* proteins and; 15 unassembled read-singletons matching dothideomycete proteins via blastx but not matching *S. nodorum* proteins.

Table 7: Summary of the 266 6-frame translated genome-mapped peptide clusters[1] not confirming existing *S. nodorum* CDS annotations by either overlap or proximity within 200 bp.

In conflicting orientation with existing gene annotation	113
No conflict, no supporting evidence	135
Overlaps genomic tblastn hit	18
Genomic tblastn hit links to an existing gene	15
Genomic tblastn hit	3

[1] Peptide clusters were assessed for orientation conflicts with genes in the opposing strand and for overlap with grouped regions of tblastn homology to related dothideomycete genomes (*L. maculans*, *P. tritici-repentis*, *C. heterostrophus*, *A. brassisicola*, *M. graminicola* and *M. fijiensis*).

unassembled reads identified a further 1074 (423+ 651) candidate genes (Tables 7 and 8).

To sift through this large number of new gene candidates, we applied two tests. First we discarded gene candidates that mapped to known genes in the opposing orientation. This removed 113 of the 266 peptide clusters matching the assembly. As the genomes of several dothideomycetes have been released in the last three years, we were able to compare the predicted new genes to the predicted proteomes of these related organisms. Overall 68% (10899/16116) of *S. nodorum* genes (and 86.4% of v2 (9299/10762) genes) have a homolog among these related species. Only 18 peptide clusters (12%) corresponded to significant tblastn hits between dothideomycete proteins and the genome assembly. Of these, 15 matched known *S. nodorum* genes (Additional File 2A) and six of these corresponded to a single gene (SNOG_01477). The three other peptide clusters correspond to potentially novel genes (Additional File 2B). In the case of the 1074 candidate genes on the un-assembled reads, 707 (270 + 437) significantly matched via blastx to dothideomycete proteins of which 663 (241 + 422) hit existing *S. nodorum* genes. The remaining 44 (29 + 15) genes are dominated by transposon-related genes (as would be expected for the repeat-dominated un-assembled reads) but also include several metabolically and structurally critical gene functions (Table 8 and Additional File 2C).

Conclusion

The flood of genome sequences that are resulting from the wave of "next-generation" sequencing technologies demands the development of time and cost-efficient methods of genome annotation. Annotation pipelines utilising transcriptomic data will remain the first choice option in many cases but the results presented here show that proteomics based on LC and tandem TOF approaches can efficiently complement transcriptomic-based annotation. The number of genes confirmed by EST analysis (2696) and proteomics (2253) are comparable both in terms of experimental time and equipment and consumables costs. The confirmation of existing gene models by proteomics is computationally straightforward as is the matching of spectra to 6-frame translated genome databases. Merging and resolution of multiple datasets of differing evidence levels is more complicated. In this paper, we have developed an annotation protocol based on defining peptide clusters and comparing first their coordinates to existing genes and then their sequences to genes from related organisms. Building upon this, we have created a pipeline that highlights potential problems with existing genes as well as new genes. This approach does not replace the need for manual annotation but reduces the scale of the task whilst providing an additional layer of evidence for gene annotation refinement.

Table 8: Summary of 6-frame translated unassembled reads supported by MASCOT peptides.

Unassembled Read Contigs	**939**
with peptide support	423
with blastx hit to dothideomycetes	270
Hits *S. nodorum* gene	241
does not hit *S. nodorum* gene	29
Without blastx hit to dothideomycetes	153
without peptide support	516
Unassembled Read Singletons	**10839**
with peptide support	651
with blastx hit to dothideomycetes	437
hits *S. nodorum* gene	422
does not hit *S. nodorum* gene	15
Without blastx hit to dothideomycetes	214
without peptide support	10188

15455 reads that were not included in the main genome assembly of *S. nodorum* were re-assessed for overlap. 4616 reads were clustered into 939 contigs, with 10839 singleton reads remaining.

Authors' contributions

KCT performed fungal culture, protein extraction and sample preparation techniques. PSS and RPO devised the experimental plan and provided intellectual input. SB, TC and RL performed mass-spectrometry analyses. JKH performed the bioinformatic analyses. JKH and RPO wrote the manuscript and RPO, PSS, KCT, JKH and RL revised the manuscript. All authors have read and approved the final manuscript.

Additional material

Additional file 1

Summary of the potentially new and modified genes identified by 6-frame proteogenomics. 18 6-frame translated genome-matching Peptide clusters not supporting an existing gene annotation with tblastn homology evidence supported the modification (by extension and/or merging) of 12 existing gene annotations (A) or the creation of a new gene annotation (B). 6 of the 15 Peptide clusters linked to existing gene annotations corresponded a single gene, SNOG_01477. 29 contigs of unassembled reads had a blastx hit to a dothideomycete genome but no similarity with S. nodorum *annotated genes (C). These represent potential new genes that were excluded from the* S. nodorum *genome due to assembly errors. A further 15 unassembled reads which did not form contigs also had a blastx hit to a dothideomycete genome but no similarity with* S. nodorum *annotated genes (D). These represent a less reliable set of potential new genes excluded from the main genome assembly. % Hit aligned is the percentage of the length of the best blastp hit subsequently globally aligned via the Needleman-Wunsch algorithm that aligns to the corresponding* S. nodorum *protein. % Identity is the percentage of identical amino acids contained within this alignment, whereas % Similarity is the percentage of amino acids with similar properties.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-301-S1.PDF]

Additional file 2

Summary tables of supporting evidence for peptide-supported genes and coordinate data for 6-frame translated genome-mapped peptides and peptide clusters.
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-301-S2.XLS]

Additional file 3

Summary of gene ontology (GO) terms over and under represented in peptide supported S. nodorum *genes relative to a random sampling of the whole genome of* S. nodorum. *Significance of representation was determined via Fisher's exact test, subject to a p-value threshold of 0.05.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-301-S3.PDF]

Additional file 4

Summary of gene ontology (GO) terms over and under represented in peptide supported S. nodorum *genes relative to genes supported by EST alignments. Significance of representation was determined via Fisher's exact test, subject to a p-value threshold of 0.05.*
Click here for file
[http://www.biomedcentral.com/content/supplementary/1471-2105-10-301-S4.PDF]

Acknowledgements

The proteomics analyses were performed in facilities funded by the (WA) Lotterywest State Biomedical Facility - Proteomics Node, Western Australian Institute for Medical Research, Perth, Australia. The proteomics data analyses were performed with the support of the facilities at the Australian Proteomics Computational Facility (housed at the Ludwig Institute for Cancer Research, Victoria, Australia, and funded by the National Health and Medical Research Council of Australia (NHMRC grant number 381413). The research was supported by the Australian Grains and Research Corporation (UMU00022 and a studentship to JKH).

References

1. Yanofsky C: **Using studies on tryptophan metabolism to answer basic biological questions.** *Journal of Biological Chemistry* 2003, **278(13):**10859-10878.
2. Tan KC, Heazlewood JL, Millar AH, Thomson G, Oliver RP, Solomon PS: **A signaling-regulated, short-chain dehydrogenase of *Stagonospora nodorum* regulates asexual development.** *Eukaryotic Cell* 2008, **7(11):**1916-1929.
3. Tan K-C, Heazlewood JL, Millar AH, Oliver RP, Solomon PS: **Proteomic identification of extracellular proteins regulated by the *Gna1* Gα subunit in *Stagonospora nodorum*.** *Mycological Research* 2009, **113(5):**523-531.
4. Carapito C, Klemm C, Aebersold R, Domon B: **Systematic LC-MS analysis of labile post-translational modifications in complex mixtures.** *Journal of Proteome Research* 2009, **8(5):**2608-2614.
5. Fessler MB, Malcolm KC, Duncan MW, Worthen GS: **A genomic and proteomic analysis of activation of the human neutrophil by lipopolysaccharide and its mediation by p38 mitogen-activated protein kinase.** *Journal of Biological Chemistry* 2002, **277(35):**31291-31302.
6. Kalume DE, Peri S, Reddy R, Zhong J, Okulate M, Kumar N, Pandey A: **Genome annotation of *Anopheles gambiae* using mass spectrometry-derived data.** *BMC Genomics* 2005, **6:**128.
7. Vaughan A, Chiu SY, Ramasamy G, Li L, Gardner MJ, Tarun AS, Kappe SH, Peng X: **Assessment and improvement of the *Plasmodium yoelii yoelii* genome annotation through comparative analysis.** *Bioinformatics* 2008, **24(13):**i383-389.
8. Wright JC, Sugden D, Francis-McIntyre S, Riba-Garcia I, Gaskell SJ, Grigoriev IV, Baker SE, Beynon RJ, Hubbard SJ: **Exploiting proteomic data for genome annotation and gene model validation in *Aspergillus niger*.** *BMC Genomics* 2009, **10:**.
9. Jaffe JD, Berg HC, Church GM: **Proteogenomic mapping as a complementary method to perform genome annotation.** *Proteomics* 2004, **4(1):**59-77.
10. Solomon PS, Lowe RGT, Tan KC, Waters ODC, Oliver RP: ***Stagonospora nodorum*: cause of stagonospora nodorum blotch of wheat.** *Molecular Plant Pathology* 2006, **7(3):**147-156.
11. Solomon PS, Wilson TJG, Rybak K, Parker K, Lowe RGT, Oliver RP: **Structural characterisation of the interaction between *Triticum aestivum* and the dothideomycete pathogen *Stagonospora nodorum*.** *European Journal of Plant Pathology* 2006, **114(3):**275-282.
12. Hane JK, Lowe RGT, Solomon PS, Tan KC, Schoch CL, Spatafora JW, Crous PW, Kodira C, Birren BW, Galagan JE, *et al.*: **Dothideomycete-plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*.** *Plant Cell* 2007, **19(11):**3347-3368.
13. Majoros WH, Pertea M, Antonescu C, Salzberg SL: **GlimmerM, Exonomy and Unveil: three ab initio eukaryotic genefinders.** *Nucleic Acids Res* 2003, **31(13):**3601-3604.
14. Solomon PS, Rybak K, Trengove RD, Oliver RP: **Investigating the role of calcium/calmodulin-dependent protein signalling in *Stagonospora nodorum*.** *Molecular Microbiology* 2006, **62:**367-381.
15. Huang X, Madan A: **CAP3: A DNA sequence assembly program.** *Genome Research* 1999, **9(9):**868-877.
16. Elias JE, Haas W, Faherty BK, Gygi SP: **Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations.** *Nature Methods* 2005, **2(9):**667-675.
17. Bendtsen JD, Nielsen H, von Heijne G, Brunak S: **Improved prediction of signal peptides: SignalP 3.0.** *Journal of Molecular Biology* 2004, **340(4):**783-795.
18. Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K: **WoLF PSORT: protein localization predictor.** *Nucleic Acids Research* 2007:W585-587.

Chapter 8: Attribution Statement

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

This thesis chapter is submitted in the form of a collaboratively-written draft manuscript which at the time of writing was in submission to *Science*. As such, not all work contained in this chapter can be attributed to the Ph. D. candidate.

The PhD candidate (JKH) made the following contributions to this chapter:

- The PhD candidate (JKH) performed all bioinformatics analyses described in this chapter.

The following contributions were made by co-authors:

- JKH and RPO wrote the manuscript.

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

-- --

James Hane (Ph. D. candidate) Date

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

-- --

Richard Oliver (Principal supervisor) Date

Hane *et al. Genome Biology* 2011, **12**:R45
http://genomebiology.com/2011/12/5/R45

Genome **Biology**

RESEARCH **Open Access**

A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi

James K Hane[1,2], Thierry Rouxel[3], Barbara J Howlett[4], Gert HJ Kema[5], Stephen B Goodwin[6] and Richard P Oliver[7*]

Abstract

Background: Gene loss, inversions, translocations, and other chromosomal rearrangements vary among species, resulting in different rates of structural genome evolution. Major chromosomal rearrangements are rare in most eukaryotes, giving large regions with the same genes in the same order and orientation across species. These regions of macrosynteny have been very useful for locating homologous genes in different species and to guide the assembly of genome sequences. Previous analyses in the fungi have indicated that macrosynteny is rare; instead, comparisons across species show no synteny or only microsyntenic regions encompassing usually five or fewer genes. To test the hypothesis that chromosomal evolution is different in the fungi compared to other eukaryotes, synteny was compared between species of the major fungal taxa.

Results: These analyses identified a novel form of evolution in which genes are conserved within homologous chromosomes, but with randomized orders and orientations. This mode of evolution is designated mesosynteny, to differentiate it from micro- and macrosynteny seen in other organisms. Mesosynteny is an alternative evolutionary pathway very different from macrosyntenic conservation. Surprisingly, mesosynteny was not found in all fungal groups. Instead, mesosynteny appears to be restricted to filamentous Ascomycetes and was most striking between species in the Dothideomycetes.

Conclusions: The existence of mesosynteny between relatively distantly related Ascomycetes could be explained by a high frequency of chromosomal inversions, but translocations must be extremely rare. The mechanism for this phenomenon is not known, but presumably involves generation of frequent inversions during meiosis.

Background

The evolutionary history of organisms, as revealed by comparisons of genome sequences, is of the greatest biological significance and interest. The current explosion in the number of genome assemblies of species within the same class, order and genus is allowing the whole-genome interrelationships between organisms to be examined in ever greater detail. There is a long history of comparisons of individual orthologous gene sequences and these have revolutionized our understanding of phylogenetic relationships [1]. A more complete understanding of both the mechanism and results of evolution can be obtained by comparing entire genomes [2]. These comparisons have refined the concept of synteny. This term is used loosely by many authors. Originally it was used in cytogenetics to describe two or more loci that are located on the same chromosome. As DNA sequencing and comparative genomics became commonplace, the term synteny acquired the additional property of co-linearity; i.e. the conservation of gene order and orientation. In this study we refer to synteny in the original cytogenetic sense and describe co-linearity as a sub-category of synteny. If orthologs of multiple genes that are co-located in the genome of one organism are co-located in another species, the chromosomes on which the genes reside are said to be syntenic. Synteny can also be quantitative; chromosomes that contain all of the same genes are 100% syntenic.

The process of speciation occurs when two independent populations diverge into reproductively isolated species. Initially the daughter species would have had chromosomes that shared both gene content (synteny) and order (co-linearity). Over evolutionary time, the

* Correspondence: richard.oliver@curtin.edu.au
[7]Australian Centre for Necrotrophic Fungal Pathogens, Curtin University, Perth, 6845, Australia
Full list of author information is available at the end of the article

degree of synteny and co-linearity would be degraded through various processes, including chromosomal duplications, gene losses/gains and chromosomal rearrangements (Additional file 1), until orthologous genes in one species occur randomly in the genome of the other.

The related concepts of synteny and co-linearity have been refined mostly in plants, animals and bacteria. Synteny has been differentiated qualitatively based on the length and completeness of co-linear regions. Macrosynteny describes co-linearity observable at a whole-chromosome scale, involving hundreds or thousands of genes of which a backbone are co-linear. Microsynteny describes co-linearity spanning a small number (for example, two to ten) of successive genes. Comparisons of vertebrate and flowering plant species within taxonomic families often have shown extensive macrosynteny [3-8]. Macrosynteny has been exploited to assist genetic mapping and gene cloning; examples include the use of the *Arabidopsis* genome to find genes in canola [9], and rice/*Brachypodium* synteny to locate genes in wheat and barley [10].

Filamentous fungi form an ancient, large and diverse group of organisms. Until the last decade, the phylogenetics of fungi was problematic but the application of techniques based on gene sequence variation has created a stable taxonomy. The ascomycete filamentous fungi are mostly within the sub-phylum Pezizomycotina (Figure 1) [11]. This sub-phylum contains four major classes: Dothideomycetes, Eurotiomycetes, Sordariomycetes and Leotiomycetes. The Dothideomycetes contains more than 20,000 species amongst which are many of the most important plant pathogens worldwide, including those in the genera *Phaeosphaeria*, *Leptosphaeria* and *Mycosphaerella*.

Evolutionary diversity within the filamentous ascomycete fungi is much higher than in flowering plants or vertebrate animals [12]. A number of reasons have been proposed to account for this. Filamentous fungi reflect approximately 400 million years of evolutionary history, comparable to that of the vertebrates but approximately four times longer than that of the flowering plants [13] (Figure 1; Table 1). The generation times of fungi are typically measured in hours or days, whereas plants and

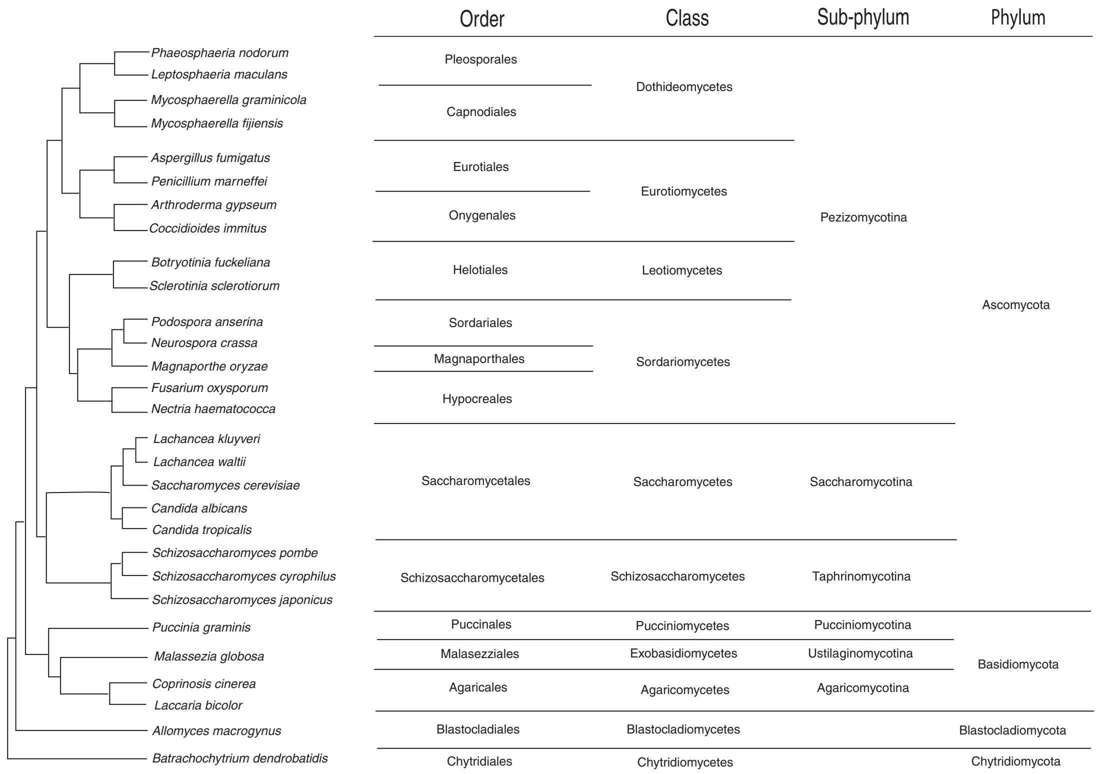


Figure 1 Cladogram of species used in whole-genome comparisons in this study. Detailed information on each species is provided in Additional file 3.

Table 1 Summary of whole-genome synteny relationships across selected fungal orders

	Sub-phylum	**Pezizomycotina**								**Saccharomycotina**	**Taphrinomycotina**
	Class	**Dothideomycetes**		**Eurotiomycetes**		**Sordariomycetes**			**Leotiomycetes**	**Saccharomycetes**	**Schizosaccharomycetes**
Class	**Order**	**Capnodiales**	**Pleosporales**	**Eurotiales**	**Onygenales**	**Hypocreales**	**Magnaporthales**	**Sordariales**	**Helotiales**	**Saccharomycetales**	**Schizosaccharomycetales**
Dothideomycetes	Capnodiales	Meso/150	Meso	Demeso	Demeso	Demeso	None	Demeso	Demeso	None	None
	Pleosporales	300	Meso/120	Demeso	Demeso	Demeso	None	Demeso	Demeso	None	None
Eurotiomycetes	Eurotiales	370	370	Demacro/< 160	Demacro	None	None	None	Demeso	None	None
	Onygenales	370	370	150	Demacro/< 160	None	None	None	Demeso	None	None
Sordariomycetes	Hypocreales	370	370	370	370	Macro/170	Demeso	Demeso	Demeso	None	None
	Magnaporthales	370	370	370	370	240	NA	Demeso	Demeso	None	None
	Sordariales	370	370	370	370	225	240	Demeso	Demeso	None	None
Leotiomycetes	Helotiales	370	370	370	370	340	340	340	Macro/250	None	None
Saccharomycetes	Saccharomycetales	500	500	500	500	500	500	500	500	Demacro/none/250	None
	Schizosaccharomycetales	650	650	650	650	650	650	650	650	650	Demacro/none/240

Whole-genome synteny, indicated above the diagonal, was classified as either macrosynteny (macro), degraded macrosynteny (demacro), mesosynteny (meso), degraded mesosynteny (demeso) or no synteny (none). 'NA' indicates lack of sufficient data to perform whole-genome comparisons. Time since divergence between orders, previously predicted within the Ascomycetes [38], is indicated below the diagonal in millions of years.

animals have generation times of many weeks, years or even decades. Meiosis is a powerful force stabilising chromosomal structure and may occur less commonly in some fungi compared to plants and animals; whilst nearly all filamentous fungi undergo germline asexual reproduction, only a subset have known sexual phases. Furthermore, many filamentous fungi can acquire genetic material by lateral gene transfer, which can increase their rate of evolution [14-16]. All of these factors would tend to reduce or eliminate the extent of synteny between species. It was not surprising, therefore, when initial comparisons between fungal genome sequences failed to find extensive evidence of interspecific macro- or microsynteny [17-22] and, with the exception of the aspergilli, even between species from the same genus [23-25].

The number of sequenced fungal genomes has increased dramatically since 2008. There is now a sufficient number of sequenced species within each fungal class to begin to assess whole-genome patterns of evolution. In this paper, we have applied a simple dot-plot approach to fungal genome comparisons and observed a striking pattern of chromosome-level evolutionary conservation. This pattern is characterized by the conservation of gene content in chromosomes, without conservation of gene order or orientation; that is, synteny without co-linearity. We propose to call this sequence conservation 'mesosynteny' to distinguish it from micro- and macrosynteny. Mesosynteny appears to be peculiar to the filamentous Ascomycetes (syn. Pezizomycotina), particularly in the class Dothideomycetes. This phenomenon has interesting implications for the study of genome evolution and may have applications in the sequencing and assembly of fungal genomes.

Results

Dot plots are a well-established method of representing sequence comparisons [26]. Comparison of co-linear genomes (Supplementary Figure S1a in Additional file 1) gives a series of dots that lie on the diagonal (Supplementary Figure S1b in Additional file 1). Random gene loss from either chromosome without major rearrangements (Supplementary Figure S1c, d in Additional file 1) progressively destroys microsynteny but retains macrosynteny. Inversions are visualised on dot plots by diagonal lines with the opposite slope, while translocations are indicated when the genes on a chromosome of one species share syntenic blocks with two or more chromosomes. Conservation of short, contiguous runs of genes, whether on the same or different chromosomes, retains microsynteny but not macrosynteny.

The fungus *Phaeosphaeria* (syn. *Stagonospora*, *Septoria*) *nodorum* is a major pathogen of wheat [27]. It is a member of the class Dothideomycetes (Figure 1), a taxon that includes more than 20,000 species amongst which are many dominant crop pathogens [28]. Its genome, which is believed to comprise 14 to 19 chromosomes [29], was assembled as 107 nuclear scaffolds [21]. Expressed sequence tag and proteomic data have refined the annotations to a set of 12,194 genes [30-32]. Pathogenicity in *P. nodorum* has been linked to the expression of a suite of necrotrophic effectors [33-36] (formerly called host-specific toxins), some of which appear to have been acquired by lateral gene transfer [14,16].

The genome sequences of other Dothideomycetes species have become available recently, allowing whole-genome comparisons with relatively closely related taxa. We used the software tool MUMmer [37] to generate dot plots that compare the scaffolds of the *P. nodorum* assembly with the 21 finished chromosomes of *Mycosphaerella graminicola* [38,39] (Figure 2a). These species are classified respectively in the Pleosporales and the Capnodiales, order-level taxa within the Dothideomycetes (Figure 2), with an estimated divergence time of (very approximately) 300 million years ago (Mya). The figure is arranged with the chromosomes or scaffolds of each species arranged in size order along the axes. Dots correspond to regions of sequence similarity and are color-coded to indicate their degree of identity.

Our expectation was that we would see either dispersed diagonal lines or a completely random distribution of very short matches ('dots'). Instead, the dot plot shows a highly non-random distribution whereby dots from individual chromosomes of *M. graminicola* appear to be strongly associated with one or a few scaffolds of *P. nodorum*, indicated by 'boxes' within columns and rows. For example, dots corresponding to *P. nodorum* scaffold 7 were almost exclusively found within the box corresponding to *M. graminicola* chromosome 12. Reciprocally, dots corresponding to *M. graminicola* chromosome 12 appeared predominantly within the box corresponding to *P. nodorum* scaffold 7. The dots within this box did not fall on any obvious diagonal lines and were instead arranged quasi-randomly. When these two sequences were aligned (Figure 2b), lines joining regions of significant similarity were distributed quasi-randomly. The orientation of the genes (color coded as red for parallel and blue for inverted) also appeared to be random. The dot plots used six-frame back translations of the genomes. Similar results were obtained with raw nucleotide sequences or when validated genes were used (Additional file 2). This indicated that the majority of the dots corresponded to genes.

We call this pattern of dots-within-boxes 'mesosynteny'. The non-random distribution implies conservation of the gene content of scaffolds (and by implication, chromosomes) during evolution; hence, this is a form of

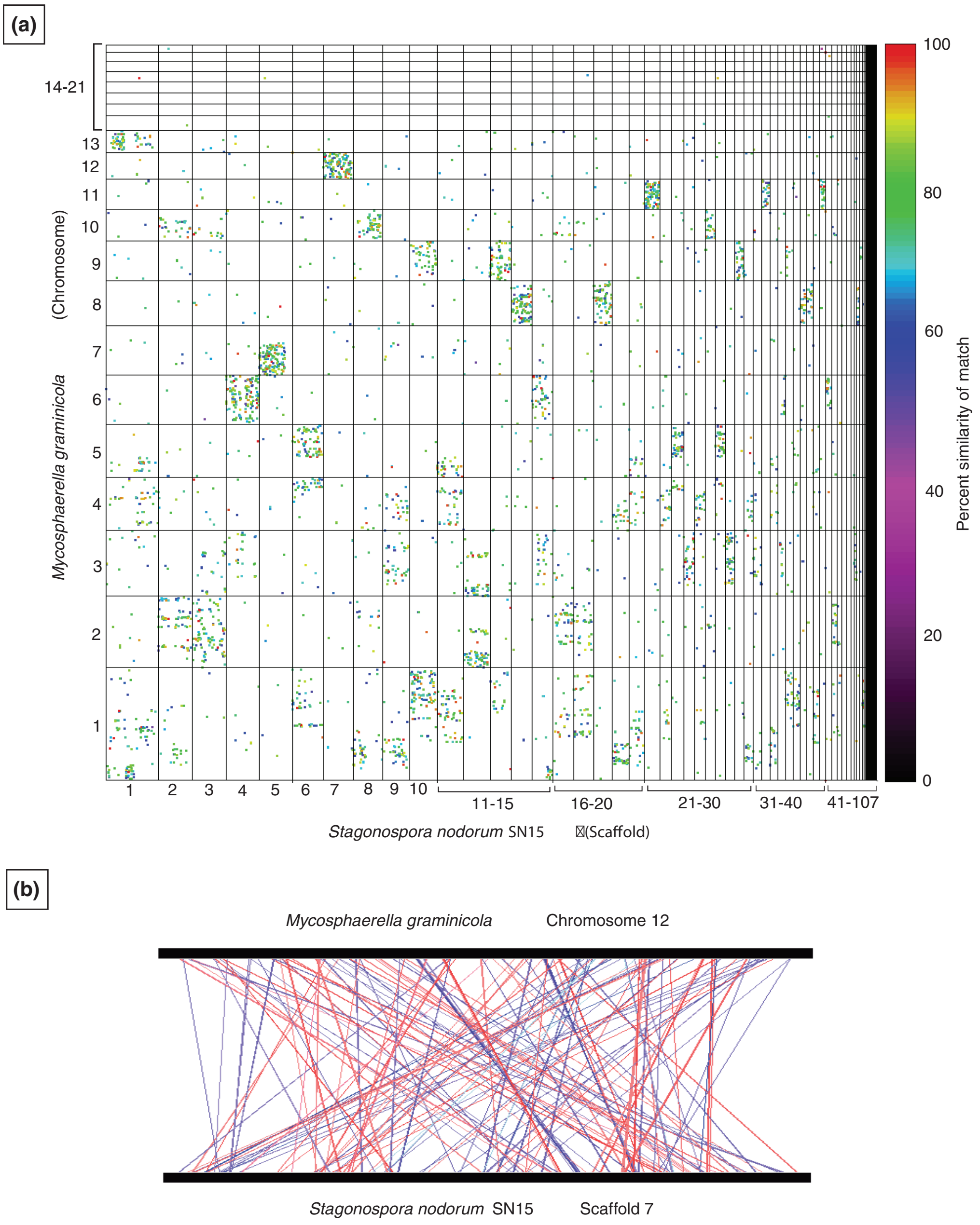


Figure 2 Whole-genome dot-plot comparison between the Dothideomycetes species *Mycosphaerella graminicola* and *Phaeosphaeria nodorum*. **(a)** The six-frame translations of both genomes were compared via MUMmer 3.0. Homologous regions are plotted as dots, which are color coded for percent similarity as per the color bar (right). Chromosomes 1 to 21 from *M. graminicola* are displayed by decreasing size along the y axis and scaffolds 1 to 107 from *P. nodorum* are displayed along the x axis. Dots represent matching regions between translated scaffold sequences. Mesosyntenic regions appear as dots within boxes without any obvious diagonal lines. **(b)** Homology relationships between chromosome 12 of *M. graminicola* and scaffold 17 of *P. nodorum*. Red lines link parallel homologous pairs and blue lines link anti-parallel pairs.

synteny [40] that does not involve the retention of colinearity as found in both macro- and microsynteny.

Taxonomic distribution of mesosynteny across the fungal kingdom

To test the extent and generality of mesosynteny within the fungi, the analysis was extended to other species within the Dothideomycetes, other classes within the Pezizomycotina and other fungal phyla. These comparisons were tested for chromosomal-scale genome conservation and were classified as macrosyntenic, mesosyntenic, or non-syntenic (Figure 3; Additional file 3).

Visual inspection of dot plots distinguished the comparisons neatly into three classes: no synteny,

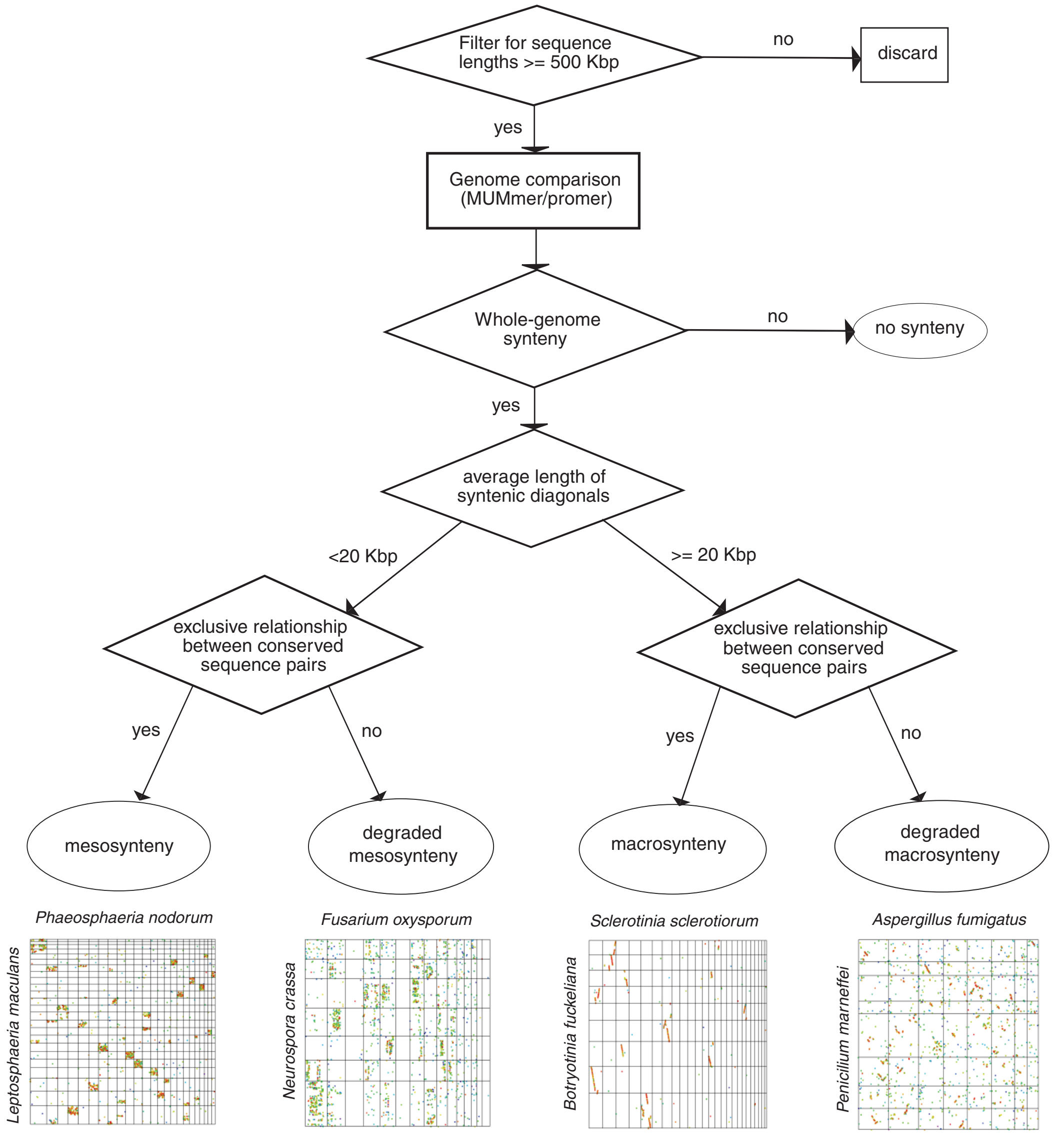


Figure 3 The whole-genome synteny classification process. Genome sequences of fewer than 500 kb were discarded before analysis. The results of MUMmer comparisons between a pair of genomes were tested for synteny (see Materials and methods). If a genome pair was determined to be syntenic at a whole-genome level, synteny was classified as 'mesosyntenic' or 'macrosyntenic' based on the average length of co-linear diagonals.

macrosynteny or mesosynteny. Quantifying the degree of synteny between species required the development of new statistical tests. Significant sequence conservation was tested between pairs of scaffolds by a one-tailed cumulative binomial test, requiring a probability of ≥0.99. The whole-genome conservation was defined as significant when ≥25% of the expected number of scaffold pairs (assuming perfect whole-genome synteny) were conserved. Species pairs showing synteny were classified as macro- or mesosynteny based on the average length of co-linear runs of sequence matches between both genomes; an average co-linear diagonal length of ≥20 kb was considered macrosyteny and < 20 kb was classified as mesosynteny (Additional file 3). Mesosynteny and macrosynteny were further categorized into 'degraded' or 'non-degraded' (Figure 3). Synteny was classified as degraded when significant clusters of 'dots' or 'lines' were found outside of the primary box (that is, for any given 'box', < 75% of the total length of conserved sequences within its corresponding rows and columns resided within the dominant box). Scaffolds shorter than 500 kb were excluded from these analyses.

Dot-plot comparisons between the Dothideomycetes species *P. nodorum*, *M. graminicola*, *Mycosphaerella fijiensis* and *Leptosphaeria maculans* showed significant mesosynteny (Figure 4). The comparison between *P. nodorum* and *L. maculans* (both in the order Pleosporales) was especially striking (Additional file 2). The dot plot was dominated by matches of 80 to 100% similarity, compared to 60 to 80% in the case of *P. nodorum* versus the species in the Capnodiales, *M. graminicola* or *M. fijiensis*. The dots in the comparison between *P. nodorum* and *L. maculans* were almost exclusively restricted to single boxes within both rows and columns. As before, there was no indication of the diagonal lines characteristic of macrosynteny. This pattern of nearly exclusive dots within single boxes was also observed when comparisons were made between these genomes and the other released but so far unpublished Dothideomycetes genomes available via the JGI and Broad Institute web sites ([39,41,42] and data not shown).

Dothideomycetes species also showed a discernable level of mesosynteny-like conservation with species representing the classes Eurotiomycetes (*Aspergillus fumigatus*), and the Leotiomycetes (*Sclerotinia sclerotiorum*; *S. sclerotiorum* sequencing project [43]), but not with the Sordariomycetes (*Magnaporthe oryzae*) or the Saccharomycetes (*Saccharomyces cerevisiae*) (Figure 4; Additional file 1). Comparisons of *P. nodorum* and *M. graminicola* with *A. fumigatus* and *S. sclerotiorum* had a statistically significant non-random distribution of dots within boxes. In contrast to intra-Dothideomycetes comparisons, dots appeared in multiple boxes within a row and column. This is an example of degraded mesosynteny. Comparisons between Dothideomycetes and *M. oryzae* (Sordariomycetes) and the yeast *S. cerevisiae* (Saccharomycetes) failed to find a statistically significant degree of synteny, reflected in the apparently random distribution of dots. These comparisons had an average of 1 and 0 sequences with binomial probabilities of significant sequence conservation >0.99. No statistically significant syntenic relationships were found when either *M. oryzae* or any yeast was compared with other filamentous fungal genomes.

A similar series of dot-plot comparisons between the class Eurotiomycetes and Leotiomycetes and species from classes of the Ascomycota is shown in Figure 5. The test species are *A. fumigatus* and *S. sclerotiorum*. The comparisons between *S. sclerotiorum* and *Botryotinia fuckeliana* exhibited a highly conserved pattern with many obvious diagonal lines made up of red and yellow dots representing highly similar (90 to 100%) sequence pairs. The average length of co-linear regions was much greater than 20 kb. This is a classical macrosyntenic pattern reflecting very recent divergence between these closely related genera. A weaker macrosyntenic pattern was observed between *A. fumigatus* and *Penicillium marneffei*, two species in the Eurotiales. Less than 25% of matches in columns and rows resided within a single box, characteristic of degraded macrosynteny. Comparisons between *A. fumigatus* and *S. sclerotiorum* and the Dothideomycetes, represented by *L. maculans*, revealed degraded mesosynteny. This was also observed between *S. sclerotiorum* and the two members of the Eurotiales, *A. fumigatus* and *P. marneffei*.

The Sordariomycetes *Fusarium oxysporum* exhibited mixed patterns of synteny in comparisons between species from the related orders Sordariales and Hypocreales and from other classes in the Pezizomycotina (Figure 6; Additional file 4). Striking macrosynteny was observed between chromosomes 1, 2, 4, 5 and 7 to 10 of *F. oxysporum* and chromosomes 1 to 6 and 7 to 10 of *Nectria haematococca*. Parts of chromosomes 3, 6 and 11 to 14 of *F. oxysporum* exhibited a mesosyntenic pattern with chromosomes 7 and 11 to 14 of *N. haematococca*. Mesosynteny was strongest between *N. haematococca* chromosome 14 and parts of *F. oxysporum* chromosomes 3, 6, 14 and 15. Degraded mesosynteny was observed between *F. oxysporum* and *Neurospora crassa*, *S. sclerotiorum*, *A. fumigatus* and with *P. nodorum*. However, in all comparisons (excluding *N. haematococca*), dots were conspicuously absent from rows corresponding to *F. oxysporum* chromosomes 3, 6, 14 and 15 (Figure 6).

The comparison between *N. crassa* and *Podospora anserina* (order Sordariales) showed a dominant pattern of mesosynteny, with some macrosyntenic regions particularly between the largest chromosome of both species

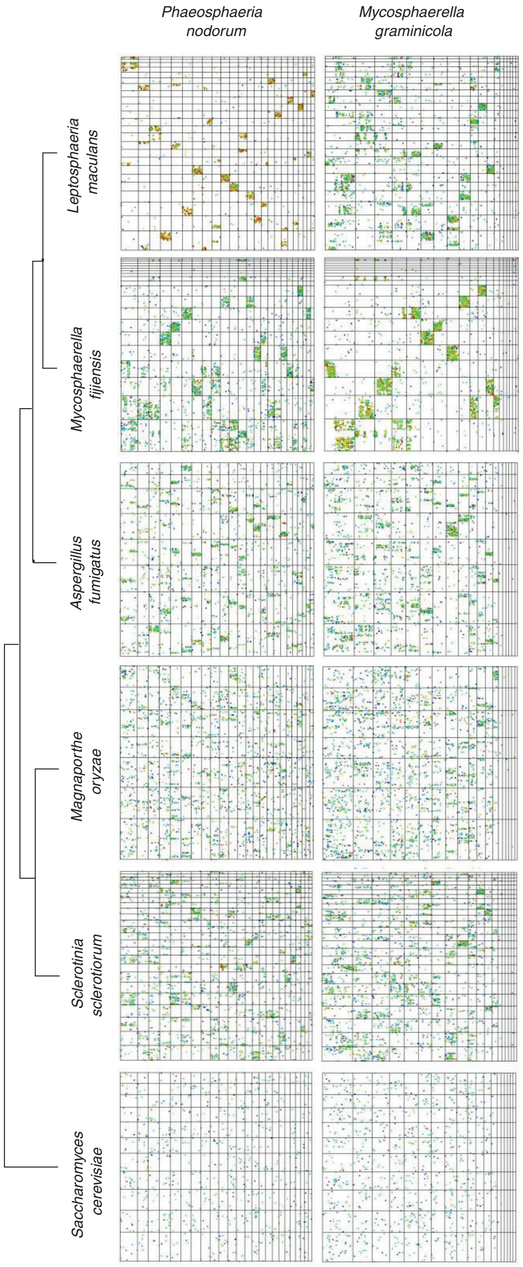


Figure 4 Dot-plot comparisons between the class Dothideomycetes and related classes. Scaffolds greater than 500 kb in length are ordered in ascending alpha-numeric order upwards along the y-axis and left-to-right along the x-axis. The orders Pleosporales (represented by *P. nodorum* and *L. maculans*) and Capnodiales (*Mycosphaerella* spp.) of the Dothideomycetes exhibit a tightly clustered pattern of mesosynteny between and within each order. This degrades into a mesosynteny-like pattern in comparisons between Dothideomycetes and the classes Eurotiomycetes (*A. fumigatus*), Sordariomycetes (*M. oryzae*) and Leotiomycetes (*S. sclerotiorum*). Clustered blocks can still be observed in these comparisons. The dot-plots comparing Dothideomycetes and Saccharomycetes (*S. cerevisiae*) appear to be random - that is, there was no synteny.

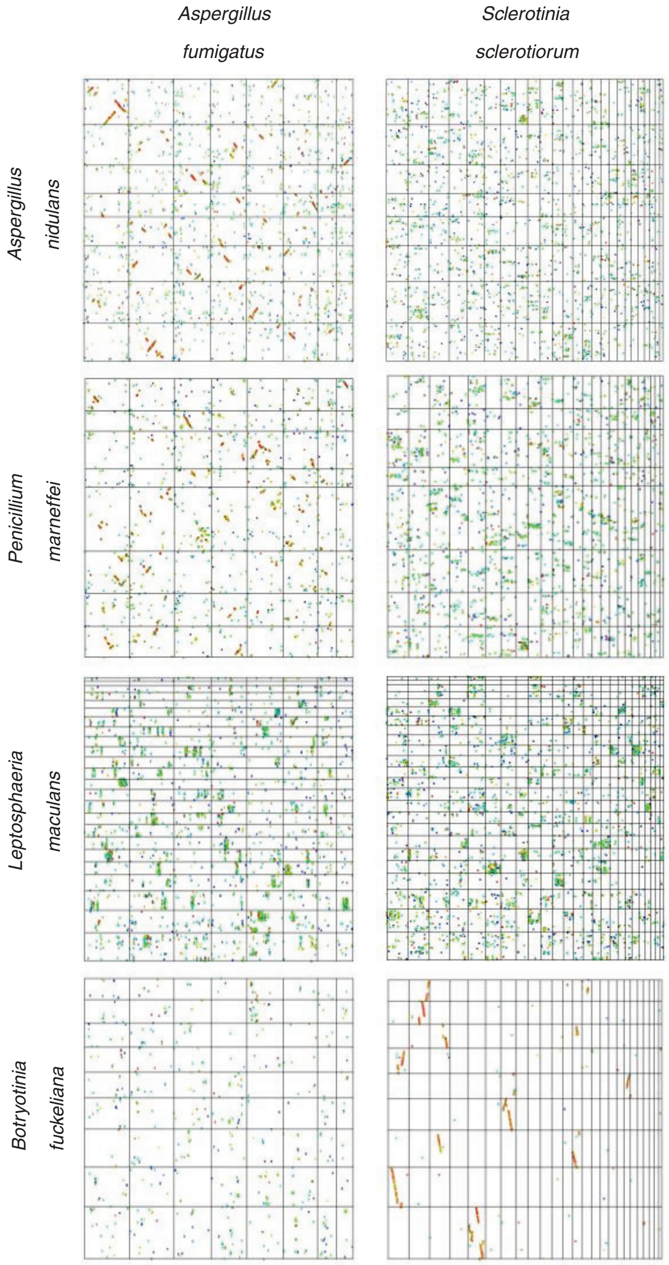


Figure 5 Dot-plot comparisons between representatives of the classes Eurotiomycetes and Leotiomycetes and related classes. The orders Eurotiales (*Aspergillus* spp.) and Onygenales (*P. marneffei*) of the Eurotiomycetes exhibit degraded macrosynteny between and within each order. The Leotiomycetes exhibit macrosynteny between species of the order Helotiales. A mesosynteny-like pattern is observed in comparisons between the Eurotiomycetes, Leotiomycetes and the more distantly related class Dothideomycetes (*L. maculans*).

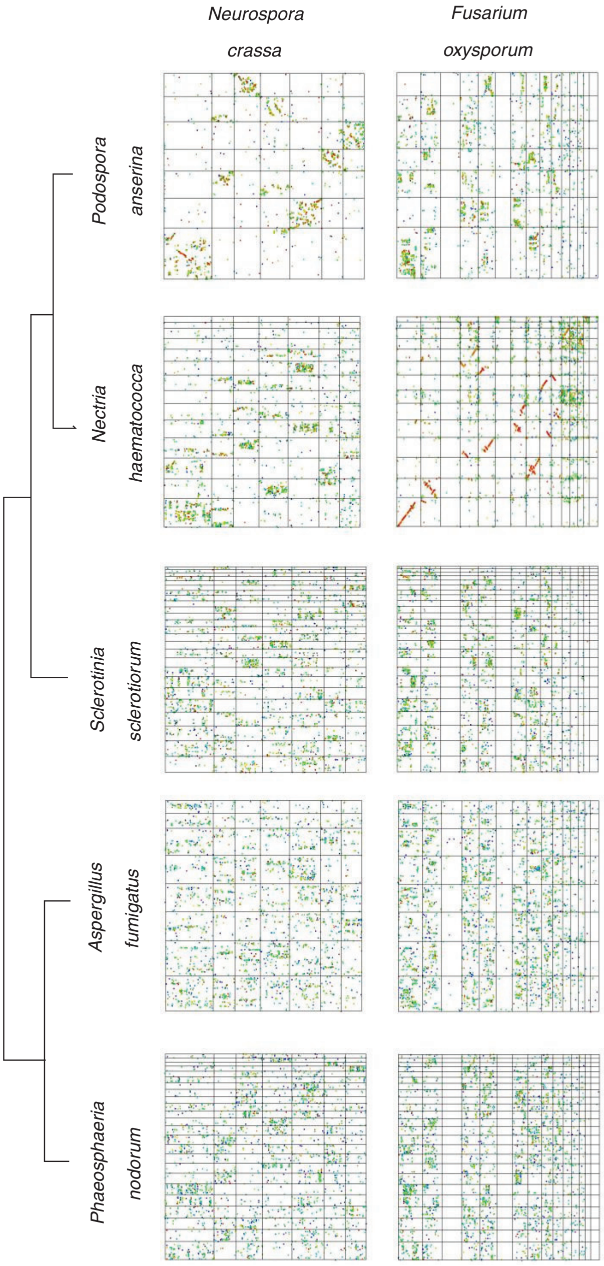


Figure 6 Dot-plot comparisons between two members of the class Sordariomycetes and related classes. The orders Sordariales (*N.* crassa and *P. anserina*) and Hypocreales (*F. oxysporum* and *N. haematococca*) of the Sordariomycetes generally exhibit mesosynteny-like conservation. The closely related pair of Hypocreales species *Fusarium oxysporum* and *Nectria haematococca* exhibit mostly macrosynteny with mesosynteny between a few chromosomes. Comparisons between the Sordariales and Hypocreales with the classes Leotiomycetes, Eurotiomycetes and Dothideomycetes exhibit degraded mesosynteny. Lack of synteny between the conditionally dispensable chromosomes 3, 6, 14 and 15 of *F. oxysporum* (excluding *N. haematococca*) with chromosomes of any other species is evident by an absence of dots in those comparisons.

[20]. *N. crassa* exhibited degraded mesosynteny when compared to *S. sclerotiorum*, *A. fumigatus* and *P. nodorum* (Figure 6).

We expanded these comparisons beyond the classes presented above to include additional species of the subphyla Saccharomycotina and Taphrinomycotina within the phylum Ascomycota and species from the phyla Basidiomycota, Blastocladiomycota and Chytridiomycota (Table 1; Figure 1). Non-filamentous Ascomycetes (classes Saccharomycetes and Schizosaccharomycetes) exhibited either macrosynteny or no synteny within their respective classes and no synteny with other fungal classes (Additional file 5). The class Agaricomycetes exhibited degraded macrosynteny between species within the class and no synteny with other fungal classes (Additional file 6). No non-Pezizomycotina taxa showed any level of synteny when compared to species of a different class (Table 1; Additional file 1).

Discussion

A novel and unexpected mode of chromosome-level sequence conservation, which we have called mesosynteny, has been detected between species of filamentous Ascomycetes, and in particular the Dothideomycetes. Mesosynteny implies the conservation of gene content within chromosomes but without conservation of gene order or orientation. It contrasts markedly with the macrosynteny observed commonly in plants and animals and the absence of synteny seen in other eukaryotes such as distantly related yeast species. The cause of mesosyntenic chromosomal evolution is not known. However, a mesosyntenic pattern would be expected to occur if intra-chromosomal recombination (including inversions) occurred significantly more frequently than inter-chromosomal recombinational events such as translocations.

Mesosynteny is distinct from macrosynteny. Macrosynteny would be expected to arise when the predominant modes of chromosomal evolution are inter-chromosomal recombination and gene loss. These considerations suggest that different patterns of mutagenic events can lead either to mesosynteny or to macrosynteny as chromosomes evolve following a speciation event.

Mesosynteny also is distinct from microsynteny, which is characterized by co-linearity between clusters of two to about ten genes with both order and orientation conserved. Earlier comparisons of synteny in related filamentous fungi frequently found clusters of genes with related functions but without retention of gene order or orientation. An example is the quinate cluster, which is conserved across species of the Ascomycetes [21,44]. This pattern of shuffled cluster retention is akin to what we observe at the whole-chromosome level.

Our results suggest that mesosyntenic chromosomal conservation is restricted to the Pezizomycotina and is most pronounced in the Dothideomycetes [45]. The Dothideomycetes are the only group to exhibit non-degraded mesosynteny between species of the different genera (estimated to have diverged approximately 120 to 150 Mya) and orders (approximately 300 Mya). A recognizable yet degraded form of mesosynteny was found between many species of Pezizomycotina outside the Dothideomycetes. The estimated time of divergence within Dothideomycetes orders are comparable to other orders within the Pezizomycotina that exhibited either degraded mesosynteny or no detectable synteny (Table 1) [38]. No mesosynteny was observed in any of the fungal groups outside of the Pezizomycotina that were surveyed: yeasts, Basidiomycetes, Blastocladiomycetes and Chytridiomycetes. The evolutionary separation between these groups and the Dothideomycetes (500 to 650 Mya) [38] may be so great that both mesosynteny and macrosynteny have decayed below the limit of detection. To our knowledge, mesosynteny has not been observed in non-fungal eukaryotes. Superficially similar dot-plots have been occasionally observed in comparisons of chordate genomes [45] but appear to be due to the amplification of paralogous copies of genes within chromosomes. Overall, either macrosynteny or no synteny has been found outside the Pezizomycotina.

Chromosomal conservation akin to mesosynteny had been observed previously in a number of inter-species comparisons within the Pezizomycotina, but its full extent was not analyzed. These include comparisons between the Pezizomycetes *Tuber melanosporum* and the Eurotiomycetes *Coccidioides immitus* [46] and the Sordariomycetes *P. anserina* and *N. crassa* [20]. As *N. crassa* and *P. anserina* are heterothallic, the authors suggested that the observed conservation may be specific to out-crossing (heterothallic) fungi. However, evidence from this study suggests otherwise as mesosynteny was observed between both heterothallic and homothallic species (Table 1; Additional files 3 and 5). For example, two homothallic Sordariomycetes species, which diverged approximately 225 Mya [38], exhibited degraded mesosynteny (*Fusarium graminearum* and *Chaetomium globosum*; Additional file 7).

Mesosynteny was observed in species both with and without (*F. oxysporum* [47] and *Penicillium marneffei* [48]) a known sexual stage. It may be that sexual crossing has been lost relatively recently in these species. Nonetheless, this finding suggests that mesosyntenic relationships were not quickly lost in the absence of meiosis. Amongst the mesosyntenic Pezizomycotina, mesosynteny was weakest in comparisons against the *M. oryzae* genome (Figures 5 and 6; Table 1). *M. oryzae* is believed to exist in nature in purely asexual lineages. The sequenced isolate of *M. oryzae* was a fertile derivative of two asexual lineages [19]. We speculate that a history of asexual reproduction and/or the process of laboratory domestication may have destroyed the remnants of mesosynteny in this isolate. This hypothesis could be tested by comparisons with genome sequences from additional isolates of *M. oryzae* or related species in the Magnaporthales.

In some species there was an uneven distribution of syntenic relationships between different chromosomes. The genome of *M. graminicola* has been finished [39] and comprises 21 chromosomes, the eight smallest of which have been shown to be dispensable [49]. These dispensable chromosomes displayed little sequence conservation with genes from any other species, and therefore no detectable synteny of any type (Figure 4). In contrast, the *M. graminicola* core chromosomes exhibited a typical mesosyntenic pattern in most comparisons with other Pezizomycotina species. Similarly, the conditionally dispensable chromosomes (CDCs) of *F. oxysporum* (3, 6, 14 and 15) showed no synteny with almost all species tested. All of these supernumerary chromosomes are thought to have originated by lateral transfer from unknown donor species [39,50]. Whether the lack of synteny of most supernumerary chromosomes is because they come from distantly related species or because they evolve more rapidly than core chromosomes is not known.

Surprisingly, CDC 14 of *N. haematococca* was mesosyntenic to *F. oxysporum* CDCs (chromosome 14 and the terminal end of chromosomes 3 and 6; Figure 6; Additional file 4). This is in contrast to the core chromosomes of each species, which exhibited macrosynteny and to previous comparisons that had indicated that these CDCs were non-syntenic. A comparison of the *F. oxysporum* genome with the closely related *Fusarium verticillioides* indicated that the *F. oxysporum* CDCs were not syntenic [50]. A possible explanation for this phenomenon is that mutations and rearrangements in supernumerary chromosomes accumulate more rapidly because these chromosomes are rarely required for survival. Faster accumulation of mutations potentially coupled with origins in distantly related donor species may allow the sequences of supernumerary chromosomes to diverge to the point where no sequence similarity remains (as in *M. graminicola*). The occurrence of mesosyntenic rearrangement in *F. oxysporum* and *N. haematococca* may also be related to the origin of their CDCs. These may have arisen in their common ancestor from a single chromosome, which subsequently mutated and broke into smaller chromosomes. Alternatively, they may have been recently transferred laterally from a common (or closely related) donor.

Whole-genome shotgun sequencing involves the generation of many short DNA reads that are assembled into longer segments. Macrosyntenic relationships are commonly used to assist the assembly and finishing of fragmented genome sequences, particularly in prokaryotic genomes. Sequences that are macrosyntenic to a long sequence of a closely related genome can be confidently hypothesized to be joined physically. Mesosynteny between a new genome assembly with a reference genome also may be used to suggest which scaffolds are juxtaposed. This could significantly reduce the cost and complexity of assembling and finishing genomes. To test whether mesosynteny could be used to predict scaffold joins in genomic sequences, early and late assemblies of the *M. graminicola* genome were analyzed to determine whether the joining of contigs or scaffolds in the finished genome could have been predicted by mesosyntenic relationships of the draft genome to *P. nodorum* [39]. Mesosynteny was remarkably successful in predicting separate scaffolds that should be joined and for identifying mis-joins in the initial assembly. This approach has the potential to assist with assembly and finishing of other genomes within the Pezizomycotina.

Conclusions

We have unearthed a novel mode of evolution in which chromosomes retain their content but shuffle the order and orientation of genes. We propose to call this phenomenon mesosynteny. What is the origin and mechanism of mesosynteny? The phenomenon is observed only in the Pezizomycotina and especially in the Dothideomycetes. The Dothideomycetes sequenced to date have several (ten or more) relatively small chromosomes, hinting at the ubiquity of supernumerary chromosomes within this taxon. The Pezizomycotina exhibit repeat-induced point mutation and higher frequencies of lateral gene transfer compared to other fungi [15,51]. Are these phenomena causally related?

The mechanism for mesosynteny may occur through a high frequency of inversions during meiosis. Whether the Dothideomycetes have a higher propensity for inversions is not known but should be the subject of further investigation. Alternatively, lateral gene transfer may be the driving force behind mesosynteny. The mechanism of lateral gene transfer is not well understood, but recent evidence suggests that the sequence transferred can be very large, even up to the size of entire chromosomes [49,52]. Fungi are capable of fusing with other fungal species through either conidial or hyphal anastomosis tubes [53]. Fusion can lead to exchange of nuclei and the transient formation of heterkaryotic strains. If the transferred DNA carried a gene that was beneficial to the recipient species, the chromosome (or a large section) carrying this gene may be retained whilst other donated chromosomes would be lost. As mesosynteny tends to retain genes on the same chromosomes, a recipient species may be able to accept a substitute chromosome from a reasonably closely related species without major disruption of phenotype. Recombination between the new and old chromosome would shuffle the order and orientation of genes, with remnant duplicated genes being removed in further cycles of repeat-induced point mutation. Recombinants with a complete core gene content plus any advantageous laterally transferred genes would then be selected, resulting in the mesosyntenic pattern of chromosomal conservation we see today. Mesosynteny may, therefore, be an adaptive mechanism that both allowed and resulted from lateral acquisition of large chromosomal sections.

Materials and methods

Whole-genome comparisons

The synteny classification method is outlined in Figure 3. Genome sequence assemblies of the species listed in Figure 1 were obtained from the sources described in Additional file 1. Phylogenetic data (Figure 1; Table 1) were inferred from previous publications [1,28,54,55]. Individual sequences (contigs, scaffolds or chromosomes) less than 500 kbp in length were discarded from the analysis. Whole-genome comparisons were performed using promer (MUMmer 3.0, [37]) with the '–mum' parameter. Promer outputs were filtered for repetitive matches using the program 'delta-filter' (MUMmer 3.0) with the '-g' parameter. Genome dot plots were generated using 'mummerplot' (MUMmer 3.0) and coordinates of promer matches were derived from filtered promer outputs using the 'show-coords' program (MUMmer 3.0).

Determination of significant sequence conservation

For the purposes of this study, only synteny observable at a whole-genome level was considered. For a given pair of genomes (genome A and genome B), all combinations of their sequence (contigs, scaffolds or chromosomes) pairs (one sequence from genome A (sequence A) and one from genome B (sequence B)) were tested for significant conservation. Lengths of conserved regions in sequences A and B were derived from MUMmer outputs. The probability of synteny (P_{syn}) for sequence pairs was calculated via a one-tailed cumulative binomial test:

$$p_{syn} = F(x, p, n) = \sum_{i=0}^{x} \binom{n}{i} (p)^{i} (1-p)^{n-1}$$

where x = (Length conserved in sequence A × Length conserved in sequence B)/(Length of sequence A × Length of sequence B); n = 100; p = (Total length conserved in Genome A × Total length conserved in Genome B)/(Total length genome A × Total length genome B). P_{syn} was required to be ≥ 0.99 to indicate significant amounts of sequence conservation between a sequence pair.

Analysis of syntenic regions between conserved sequences

The lengths of syntenic regions were analyzed for significantly conserved sequence pairs. Extended co-linearity of sequence matches visible as uninterrupted diagonal lines on a dot plot was used as an indicator of macrosynteny (Additional file 1). Dot plots between sequence pairs were considered as individual scatter plots. Promer matches between a pair of sequences were converted into a series of points on the scatter plot, with a point added every 1 kb along each match. R^2 values were calculated along the axis of sequence A in 20-kb windows (incrementing along by 2 kb). A window was considered to be co-linear if it contained a minimum of 15 data points with an $R^2 \geq 0.9$ The end coordinates of co-linear windows were subsequently modified to exclude the coordinate range of overlapping non-co-linear windows. The data points of co-linear windows within 50 kb of one another were combined (including intermediate data points if not overlapping) and were merged into larger co-linear windows if (Slope of window 1/Slope of window 2) > 0.8 and < 1.2. The start and end points of co-linear windows with a length of ≥ 5 kb were used as the coordinates of 'syntenic regions'. The same process was repeated along the axis of sequence B.

Classification of synteny type

Whole-genome synteny was identified by the 'significant pair ratio' statistic, which is an indicator of the proportion of conserved sequences relative to the expected number of conserved sequences. The significant pair ratio was determined by:

$$\text{significant pair ratio} = \frac{N_{scp}}{\sqrt{S_a \times S_b}}$$

where N_{scp} is the number of significantly conserved pairs between genomes A and B; S_a is the number of sequences in genome A ≥ 500 kb; and S_b is the number of sequences in genome B ≥ 500 kb.

Whole-genome synteny was identified when the significant pair ratio was ≥ 0.25. Genome pairs failing this criterion were classified as 'non-syntenic'. Genome pairs passing the test for whole-genome synteny were subcategorized as either macrosyntenic or mesosyntenic, defined by an average length of syntenic regions (combined between both compared genomes) of greater than or less than 20 kb, respectively. Synteny type was further categorized into 'degraded' or 'non-degraded' based on the statistic 'pair exclusivity'. For a given sequence pair, consisting of sequence A of genome A and sequence B of genome B, the 'pair exclusivity' was calculated by:

$$\text{Pair exclusivity} = \frac{C_{ab}}{C_{Ab} + C_{aB} - C_{ab}}$$

where C_{ab} is the total length of conserved matches between sequences A and B; C_{Ab} is the total length of conserved matches for sequence A and all sequences of genome B; and C_{aB} is the total length of conserved matches for sequence B and all sequences of genome A.

Synteny was classified as 'degraded' if the maximum value of all pair exclusivities was less than 0.75.

Additional material

Additional file 1: Supplementary Figure S1. The origins of the different types of syntenic relationships. Immediately after a speciation event, equivalent chromosomes in two daughter species retain the gene content, order and orientation of the parent species. **(a)** Diagrammatic representation of a chromosome with sequential elements A to Z. **(b)** A dot plot comparing the chromosomes in (a), with letters substituted for dots. The unbroken series of letters on the diagonal indicates macrosynteny. **(c, d)** Loss of sequences from each chromosome (c) will degrade the diagonal co-linearity when visualized as a dot plot (d).

Additional file 2: Supplementary Figure S2. **(a, b)** Correspondence between promer-derived dot plots (a) and blastp-derived protein comparisons of annotated genes (b) between *Phaeosphaeria nodorum* and *Leptosphaeria maculans*. Sequence pairs ('boxes') in (a) containing non-random distributions of 'dots' correspond to those in (b), indicating that the back-translated genome matches in (a) correspond to regions of conserved gene content.

Additional file 3: Supplementary File 1. Predictions of synteny between all species involved in this study in an Excel file.

Additional file 4: Supplementary Figure S3. Presence of both macrosyntenic and mesosyntenic conservation patterns between the genomes of *Fusarium oxysporum* and *Nectria haematococca*. Core chromosomes (indicated by black bars along the axes) are macrosyntenic between the two species. Dispensable chromosomes (red bars along the axes) are either non-syntenic (*N. haematococca* chromosomes 15 to 17) or mesosyntenic (*N. haematococca* chromosomes 7 and 11 to 14, *F. oxysporum* chromosome 14). The majority of chromosomes 3 and 6 of *F. oxysporum* had no similarity to the chromosomes of *N. haematococca* except for regions near their telomeres.

Additional file 5: Supplementary Figure S4. Degradation of whole-genome synteny in the classes Saccharomycetes and Schizosaccharomyces. Whole-genome dot plots have been limited to scaffolds or chromosomes greater than 500 kb. Species of the *Saccharomyces* and *Schizosaccharomyces* do not exhibit whole-genome conservation with each other. Certain species within each class exhibit macrosynteny whereas others exhibit no synteny.

Additional file 6: Supplementary Figure S5. Degradation of whole-genome synteny between a member of the class Agaricales and related orders. Whole-genome dot plots have been limited to scaffolds or chromosomes greater than 500 kb. Species in the Agaricales exhibited macrosynteny with each other. However, the Agaricales exhibited no synteny with the closest related classes represented in this study, the Exobasidiomycetes and Pucciniomycetes.

Additional file 7: Supplementary Figure S6. Evidence of degraded mesosynteny between the genomes of two homothallic Sordariomycete species, *Fusarium graminearum* (order Hypocreales) and *Chaetomium globosum* (order Sordariales). These two species are estimated to have diverged approximately 225 Mya. Sequence matches (dots) are arranged in blocked clusters typical of mesosynteny. Chromosomes and scaffolds do not share a one-to-one relationship, with multiple mesosyntenic clusters appearing in the same row or column.

Abbreviations

CDC: conditionally dispensable chromosome; Mya: million years ago.

Author details

[1]CSIRO Plant Industry, Centre for Environment and Life Sciences, Private Bag 5, Perth, 6193, Australia. [2]Faculty of Health Sciences, Murdoch University, Perth, 6150, Australia. [3]INRA-Bioger, Avenue Lucien Brétignières, BP 01, Thiverval-Grignon, 78850, France. [4]School of Botany, The University of Melbourne, Melbourne, 3010, Australia. [5]Wageningen UR, Plant Research International, Department of Biointeractions and Plant Health, PO Box 69, Wageningen, 6700 AB, The Netherlands. [6]USDA-ARS, Crop Production and Pest Control Research Unit, Purdue University, 915 West State Street, West Lafayette, IN 47907-2054, USA. [7]Australian Centre for Necrotrophic Fungal Pathogens, Curtin University, Perth, 6845, Australia.

Authors' contributions

RPO and JKH conceived and designed the study. JKH developed algorithms and performed mesosynteny analyses. JKH and RPO wrote the manuscript. TR, BJH, GHJK and SBG contributed data. JKH, RPO, TR, BJH, GHJK and SBG edited the manuscript. All authors read and approved the final manuscript.

Received: 20 February 2011 Revised: 27 April 2011
Accepted: 24 May 2011 Published: 24 May 2011

References

1. Schoch CL, Sung GH, Lopez-Giraldez F, Townsend JP, Miadlikowska J, Hofstetter V, Robbertse B, Matheny PB, Kauff F, Wang Z, Gueidan C, Andrie RM, Trippe K, Ciufetti LM, Wynns A, Fraker E, Hodkinson BP, Bonito G, Groenewald JZ, Arzanlou M, de Hoog GS, Crous PW, Hewitt D, Pfister DH, Peterson K, Gryzenhout M, Wingfield MJ, Aptroot A, Suh SO, Blackwell M, *et al*: **The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits.** *Syst Biol* 2009, **58**:224-239.
2. Sims GE, Jun SR, Wu GA, Kim SH: **Alignment-free genome comparison with feature frequency profiles (FFP) and optimal resolutions.** *Proc Natl Acad Sci USA* 2009, **106**:2677-2682.
3. McLysaght A, Enright AJ, Skrabanek L, Wolfe KH: **Estimation of synteny conservation and genome compaction between pufferfish (*Fugu*) and human.** *Yeast* 2000, **17**:22-36.
4. Pennacchio LA: **Insights from human/mouse genome comparisons.** *Mamm Genome* 2003, **14**:429-436.
5. Kohn M, Kehrer-Sawatzki H, Vogel W, Graves JA, Hameister H: **Wide genome comparisons reveal the origins of the human X chromosome.** *Trends Genet* 2004, **20**:598-603.
6. Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, Gouzy J, Wang X, Mudge J, Vasdewani J, Schiex T, Spannagl M, Monaghan E, Nicholson C, Humphray SJ, Schoof H, Mayer KF, Rogers J, Quetier F, Oldroyd GE, Debelle F, Cook DR, Retzel EF, Roe BA, Town CD, Tabata S, Van de Peer Y, Young ND: **Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes.** *Proc Natl Acad Sci USA* 2006, **103**:14959-14964.
7. Phan HT, Ellwood SR, Hane JK, Ford R, Materne M, Oliver RP: **Extensive macrosynteny between Medicago truncatula and Lens culinaris ssp. culinaris.** *Theor Appl Genet* 2007, **114**:549-558.
8. Shultz JL, Ray JD, Lightfoot DA: **A sequence based synteny map between soybean and *Arabidopsis thaliana*.** *BMC Genomics* 2007, **8**:8.
9. Parkin AP, Lydiate DJ, Trick M: **Assessing the level of collinearity between *Arabidopsis thaliana* and *Brassica napus* for *A. thaliana* chromosome 5.** *Genome* 2002, **45**:356-366.
10. Ma B: **Synteny between *Brachypodium distachyon* and *Hordeum vulgare* as revealed by FISH.** *Chromosome Res* 2010, **18**:841-850.
11. Zhang Y, Schoch CL, Fournier J, Crous PW, de Gruyter J, Woudenberg JH, Hirayama K, Tanaka K, Pointing SB, Spatafora JW, Hyde KD: **Multi-locus**

phylogeny of Pleosporales: a taxonomic, ecological and evolutionary re-evaluation. *Stud Mycol* 2009, **64**:85-102S5.

12. Marande W, López-García P, Moreira D: **Eukaryotic diversity and phylogeny using small- and large-subunit ribosomal RNA genes from environmental samples.** *Environ Microbiol* 2009, **11**:3179-3188.
13. Wang H, Guo S, Huang M, Thorsten LH, Wei J: **Ascomycota has a faster evolutionary rate and higher species diversity than Basidiomycota.** *Sci China* 2010, **53**:1163-1169.
14. Oliver RP, Solomon PS: **Recent fungal diseases of crop plants: is lateral gene transfer a common theme?** *Mol Plant Microbe Interact* 2008, **21**:287-293.
15. Marcet-Houben M, Gabaldon T: **Acquisition of prokaryotic genes by fungal genomes.** *Trends Genet* 2010, **26**:5-8.
16. 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.
17. Chibana H, Oka N, Nakayama H, Aoyama T, Magee BB, Magee PT, Mikami Y: **Sequence finishing and gene mapping for *Candida albicans* chromosome 7 and syntenic analysis against the *Saccharomyces cerevisiae* genome.** *Genetics* 2005, **170**:1525-1537.
18. Borkovich KA, Alex LA, Yarden O, Freitag M, Turner GE, Read ND, Seiler S, Bell-Pedersen D, Paietta J, Plesofsky N, Plamann M, Goodrich-Tanrikulu M, Schulte U, Mannhaupt G, Nargang FE, Radford A, Selitrennikoff C, Galagan JE, Dunlap JC, Loros JJ, Catcheside D, Inoue H, Aramayo R, Polymenis M, Selker EU, Sachs MS, Marzluf GA, Paulsen I, Davis R, Ebbole DJ, *et al*: **Lessons from the genome sequence of *Neurospora crassa*: Tracing the path from genomic blueprint to multicellular organism.** *Microbiol Mol Biol Rev* 2004, **68**:1-108.
19. Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, Orbach MJ, Thon M, Kulkarni R, Xu JR, Pan H, Read ND, Lee YH, 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 MH, Bohnert H, Coughlan S, Butler J, Calvo S, Ma LJ, *et al*: **The genome sequence of the rice blast fungus *Magnaporthe grisea*.** *Nature* 2005, **434**:980-986.
20. Espagne E, Lespinet O, Malagnac F, Da Silva C, Jaillon O, Porcel BM, Couloux A, Aury JM, Segurens B, Poulain J, Anthouard V, Grossetete S, Khalili H, Coppin E, Dequard-Chablat M, Picard M, Contamine V, Arnaise S, Bourdais A, Berteaux-Lecellier V, Gautheret D, de Vries RP, Battaglia E, Coutinho PM, Danchin EG, Henrissat B, Khoury RE, Sainsard-Chanet A, Boivin A, Pinan-Lucarre B, *et al*: **The genome sequence of the model ascomycete fungus *Podospora anserina*.** *Genome Biol* 2008, **9**:R77.
21. Hane JK, Lowe RG, Solomon PS, Tan KC, Schoch CL, Spatafora JW, Crous PW, Kodira C, Birren BW, Galagan JE, Torriani SF, 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.
22. Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, Arroyo J, Berriman M, Abe K, Archer DB, Bermejo C, Bennett J, Bowyer P, Chen D, Collins M, Coulsen R, Davies R, Dyer PS, Farman M, Fedorova N, Fedorova N, Feldblyum TV, Fischer R, Fosker N, Fraser A, Garcia JL, Garcia MJ, Goble A, Goldman GH, Gomi K, Griffith-Jones S, *et al*: **Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*.** *Nature* 2005, **438**:1151-1156.
23. Machida M, Terabayashi Y, Sano M, Yamane N, Tamano K, Payne GA, Yu J, Cleveland TE, Nierman WC: **Genomics of industrial Aspergilli and comparison with toxigenic relatives.** *Food Addit Contam Part A Chem Anal Control Expo Risk Assess* 2008, **25**:1147-1151.
24. Machida M, Asai K, Sano M, Tanaka T, Kumagai T, Terai G, Kusumoto K, Arima T, Akita O, Kashiwagi Y, Abe K, Gomi K, Horiuchi H, Kitamoto K, Kobayashi T, Takeuchi M, Denning DW, Galagan JE, Nierman WC, Yu J, Archer DB, Bennett JW, Bhatnagar D, Cleveland TE, Fedorova ND, Gotoh O, Horikawa H, Hosoyama A, Ichinomiya M, Igarashi R, *et al*: **Genome sequencing and analysis of *Aspergillus oryzae*.** *Nature* 2005, **438**:1157-1161.
25. Pel HJ, de Winde JH, Archer DB, Dyer PS, Hofmann G, Schaap PJ, Turner G, de Vries RP, Albang R, Albermann K, Andersen MR, Bendtsen JD, Benen JA, van den Berg M, Breestraat S, Caddick MX, Contreras R, Cornell M, Coutinho PM, Danchin EG, Debets AJ, Dekker P, van Dijck PW, van Dijk A, Dijkhuizen L, Driessen AJ, d'Enfert C, Geysens S, Goosen C, Groot GS, *et al*: **Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88.** *Nat Biotechnol* 2007, **25**:221-231.
26. Maizel JV Jr, Lenk RP: **Enhanced graphic matrix analysis of nucleic acid and protein sequences.** *Proc Natl Acad Sci USA* 1981, **78**:7665-7669.
27. Solomon PS, Lowe RGT, Tan KC, Waters ODC, Oliver RP: ***Stagonospora nodorum*: cause of stagonospora nodorum blotch of wheat.** *Mol Plant Pathol* 2006, **7**:147-156.
28. Schoch CL, Crous PW, Groenewald JZ, Boehm EW, Burgess TI, de Gruyter J, de Hoog GS, Dixon LJ, Grube M, Gueidan C, Harada Y, Hatakeyama S, Hirayama K, Hosoya T, Huhndorf SM, Hyde KD, Jones EB, Kohlmeyer J, Kruys A, Li YM, Lucking R, Lumbsch HT, Marvanova L, Mbatchou JS, McVay AH, Miller AN, Mugambi GK, Muggia L, Nelsen MP, Nelson P, *et al*: **A class-wide phylogenetic assessment of Dothideomycetes.** *Stud Mycol* 2009, **64**:1-15S10.
29. Cooley RN, Caten CE: **Variation in electrophoretic karyotype between strains of *Septoria nodorum*.** *Mol Gen Genet* 1991, **228**:17-23.
30. Bringans S, Hane JK, Casey T, Tan KC, Lipscombe R, Solomon PS, Oliver RP: **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.
31. Casey T, Solomon PS, Bringans S, Tan KC, Oliver RP, Lipscombe R: **Quantitative proteomic analysis of G-protein signalling in *Stagonospora nodorum* using isobaric tags for relative and absolute quantification.** *Proteomics* 2010, **10**:38-47.
32. Tan K-C, Heazlewood JL, Millar AH, Oliver RP, Solomon PS: **Proteomic identification of extracellular proteins regulated by the *Gna1* Gα subunit in *Stagonospora nodorum*.** *Mycol Res* 2009, **113**:523-531.
33. Oliver RP, Solomon PS: **New developments in pathogenicity and virulence of necrotrophs.** *Curr Opin Plant Biol* 2010, **13**:415-419.
34. Liu Z, Faris JD, Oliver RP, Tan KC, 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 Path* 2009, **5**:e1000581.
35. Friesen TL, Faris JD, Solomon PS, Oliver RP: **Host-specific toxins: Effectors of necrotrophic pathogenicity.** *Cell Microbiol* 2008, **10**:1421-1428.
36. Friesen TL, Zhang Z, Solomon PS, Oliver RP, Faris JD: **Characterization of the interaction of a novel *Stagonospora nodorum* host-selective toxin with a wheat susceptibility gene.** *Plant Physiol* 2008, **146**:682-693.
37. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, Antonescu C, Salzberg SL: **Versatile and open software for comparing large genomes.** *Genome Biol* 2004, **5**:R12.
38. Rouxel T, Grandaubert J, Hane J, Hoede C, van de Wouw A, Couloux A, Dominguez V, Anthouard V, Bally P, Bourras S, Cozijnsen A, Ciuffetti L, Dimaghani A, Duret L, Fudal I, Goodwin S, Gout L, Glaser N, Kema G, Lapalu N, Lawrence C, May K, Meyer M, Ollivier B, Poulain J, Turgeon G, Tyler BM, Vincent D, Weissenbach J, Amselem J, *et al*: **The compartmentalized genome of Leptosphaeria maculans: diversification of effectors within genomic regions affected by Repeat Induced Point mutations.** *Nat Commun* 2011, **2**:art.202.
39. Goodwin SB, Ben M'Barek S, Dhillon B, Wittenberg A, Crane CF, Van der Lee TAJ, Grimwood J, Aerts A, Antoniw J, Bailey A, Bluhm B, Bowler J, Bristow J, Brokstein P, Canto-Canche B, Churchill A, Conde-Ferràez L, Cools H, Coutinho PM, Csukai M, Dehal P, Donzelli B, Foster AJ, Hammond-Kosack K, Hane J, Henrissat B, Kilian A, Koopmann E, Kourmpetis Y, Kuo A, *et al*: **Finished genome of *Mycosphaerella graminicola* reveals stealth pathogenesis and dispensome structure.** *PLoS Genetics -D-10-00112R2* 2011.
40. Passarge E, Horsthemke B, Farber RA: **Incorrect use of the term synteny.** *Nat Genet* 1999, **23**:387.
41. **Broad Institute.** [http://www.broadinstitute.org/].
42. **DOE Joint Genome Institute.** [http://www.jgi.doe.gov/].
43. ***S. sclerotiorum* Sequencing Project.** [http://www.broadinstitute.org/annotation/genome/sclerotinia_sclerotiorum/Info.html].
44. Giles NH, Geever RF, Asch DK, Avalos J, Case ME: **Organization and regulation of the qa (quinic acid) genes in *Neurospora crassa* and other fungi.** *J Hered* 1991, **82**:1-7.
45. Putnam NH, Butts T, Ferrier DEK, Furlong RF, Hellsten U, Kawashima T, Robinson-Rechavi M, Shoguchi E, Terry A, Yu K Jr, Benito-Gutiérrez È, Dubchak I, Garcia-Fernàndez J, Gibson-Brown JJ, Grigoriev IV, Horton AC, De

Jong PJ, Jurka J, Kapitonov VV, Kohara Y, Kuroki Y, Lindquist E, Lucas S, Osoegawa K, Pennacchio LA, Salamov AA, Satou Y, Sauka-Spengler T, Schmutz J, Shin-I T, *et al*: **The amphioxus genome and the evolution of the chordate karyotype.** *Nature* 2008, **453**:1064-1071.

46. Martin F, Kohler A, Murat C, Balestrini R, Coutinho PM, Jaillon O, Montanini B, Morin E, Noel B, Percudani R, Porcel B, Rubini A, Amicucci A, Amselem J, Anthouard V, Arcioni S, Artiguenave F, Aury JM, Ballario P, Bolchi A, Brenna A, Brun A, Buee M, Cantarel B, Chevalier G, Couloux A, Da Silva C, Denoeud F, Duplessis S, Ghignone S, *et al*: **Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis.** *Nature* 2010, **464**:1033-1038.
47. Yun SH, Arie T, Kaneko I, Yoder OC, Turgeon BG: **Molecular organization of mating type loci in heterothallic, homothallic, and asexual Gibberella/Fusarium species.** *Fungal Genet Biol* 2000, **31**:7-20.
48. Fisher MC, Hanage WP, De Hoog S, Johnson E, Smith MD, White NJ, Vanittanakom N: **Low effective dispersal of asexual genotypes in heterogeneous landscapes by the endemic pathogen penicillium marneffei.** *PLoS Path* 2005, **1**:0159-0165.
49. Wittenberg AH, van der Lee TA, Ben M'barek S, Ware SB, Goodwin SB, Kilian A, Visser RG, Kema GH, Schouten HJ: **Meiosis drives extraordinary genome plasticity in the haploid fungal plant pathogen *Mycosphaerella graminicola*.** *PLoS One* 2009, **4**:e5863.
50. Ma LJ, van der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, Di Pietro A, Dufresne M, Freitag M, Grabherr M, Henrissat B, Houterman PM, Kang S, Shim WB, Woloshuk C, Xie X, Xu JR, Antoniw J, Baker SE, Bluhm BH, Breakspear A, Brown DW, Butchko RA, Chapman S, Coulson R, Coutinho PM, Danchin EG, Diener A, Gale LR, Gardiner DM, Goff S, *et al*: **Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*.** *Nature* 2010, **464**:367-373.
51. Clutterbuck JA: **Genomic evidence of repeat-induced point mutation (RIP) in filamentous ascomycetes.** *Fungal Genet Biol* 2011, **48**:306-326.
52. Akagi Y, Akamatsu H, Otani H, Kodama M: **Horizontal chromosome transfer, a mechanism for the evolution and differentiation of a plant-pathogenic fungus.** *Eukaryot Cell* 2009, **8**:1732-1738.
53. Roca M, Read N, Wheals A: **Conidial anastomosis tubes in filamentous fungi.** *FEMS Micro Letts* 2005, **249**:191-198.
54. Liu Y, Leigh JW, Brinkmann H, Cushion MT, Rodriguez-Ezpeleta N, Philippe H, Lang BF: **Phylogenomic analyses support the monophyly of Taphrinomycotina, including *Schizosaccharomyces* fission yeasts.** *Mol Biol Evol* 2009, **26**:27-34.
55. Shertz CA, Bastidas RJ, Li W, Heitman J, Cardenas ME: **Conservation, duplication, and loss of the Tor signaling pathway in the fungal kingdom.** *BMC Genomics* 2010, **11**:510.

doi:10.1186/gb-2011-12-5-r45
Cite this article as: Hane *et al.*: A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. *Genome Biology* 2011 **12**:R45.

Appendix 8A: Response to Thesis Examination Comments

In the introduction James defines microsynteny as the retention of linear runs of 2-10 genes. However in the results section he defines macrosynteny based on an average co-linear diagonal length of 20 kb or longer. With an average gene density of only 1 gene per 3 kb, 20 kb encompasses only 6 genes on average. This is still well witihn the microsynteny range. Would it not be better to define macrosynteny based on some majority of co-linear diagonals being greater than 30kb in length (and therefore clearly outside the macrosynteny range)?

The examiner highlights a problem with the explanation of microsynteny (co-linear runs of "2-10" genes), in that it can be confused with the definition of macrosynteny which includes co-linear runs of 20 kb. According to data presented in chapter 10 (table 2), this should on average, result in a co-linear run of six genes in *P. nodorum.* It should be noted that this number will vary between different fungal species, as they have different gene densities, but will usually still lie within the range of a run of 5-10 genes.

To clarify for readers interested in the algorithms used to determine synteny type, the 20 kb co-linearity threshold was one of several criteria used in combination. Refer to Chapter 8: methods for more details. The 20 kb threshold was an empirically determined value, based on visual observations of macro-, and mesosynteny, which in combination with other criteria served the purpose of distinguishing between these different synteny types. The description of co-linear runs of up to 10 genes in the introduction to chapter 8 was used "anecdotally" to illustrate the concept of microsynteny . Greater care should have been taken to ensure that this did not conflict with the data presented. Preliminary synteny analysis performed during my PhD candidature (unpublished) appear to indicate that co-linear runs rarely contain more than 5 genes, which does not conflict with a threshold of 6 genes.

In the discussion section, James says that mesosyntenic conservation had been observed in a number of previous studies. If it has already been reported, can it really be described as a novel form of fungal genome evolution?

Prior to this publication, mesosynteny was a conservation pattern that had generally been ignored as an overall lack of synteny. This publication represents the first detailed analysis of the phenomenon of mesosynteny. It goes beyond the few previous descriptions of mesosynteny which were limited to two species and defines the phylogenetic boundaries of this mode of evolution.

Furthermore, publications which have mentioned the phrase 'mesosynteny' had done so through having contact with the Australian Centre for Necrotrophic Fungal Pathogens, as elements of this research had been presented internationally almost two years prior to final publication.

There are three lineages in the background of the sequenced strain of M. oryzae. The Dean et al. reference is not the appropriate citation for this statement. Se Chao and Ellingboe, 1991, Phytopathology), the argument that laboratory domestication and subsequence history of asexual reproduction is nonsensical. First, it does not consider what is involved in the domestication process. The first step is crossing a wild isolate with lab strains. Approximately 50% of the 70-15 genome is only one meiotic division away from a wild strain, another ~25% is only 2 divisions away. Earlier in the paragraph, James implies that meiosis ought to maintain mesosynteny. Therefore, loss of mesosynteny would not be expected to happen during the first domestication step. Second, the "subsequence history of asexual reproduction" would amount to only two or three subcultures, which would pale in comparison to the numerous years of asexual reproduction in the field prior to isolate collection. Simply put, there has not been enough time (nuclear divisions) for the gene order to have been scrambled in 70-15 (Besides, we know from Nick Talbot's sequencing of Guy11 - a recurrent parent of 70-15 - that gene order has been maintained through the domestication process).

Surprisingly, *M. oryzae* was observed to have a more advanced degradation of the mesosyntenic pattern compared to species of equivalent phylogenetic distance. The comments regarding *M. oryzae* 70-15 speculated that this may be related to its history of lab domestication. While either position has yet to be proven or disproven my personal opinion is in agreement with the examiner, however I had conceded to the more informed opinions of my co-authors on matters concerning the subculture of fungal lab strains.

I don't see the relevance of the discussion of RIP. What bearing does this have on mesosynteny? I also don't see the relevance of the discussion of lateral gene transfer. It was not discussed in the context of mesosynteny.

One of the major discoveries outlined in the publication was the phylogenetic boundaries of mesosynteny. That is, mesosynteny is restricted to the sub-family taxon Pezizomycotina (the filamentous Ascomycota). This coincides with the phylogenetic boundaries of observed RIP-like polymorphism with a CpA bias (Clutterbuck 2010). Furthermore, while the exact mechanism of mesosynteny is yet to be determined, repetitive DNA is a strong candidate as a means of generating rearrangements, either through transposase activity or homologous recombination. The publication draws no unfounded conclusions of a direct link between the two phenomena, but based on the current literature the discussion of RIP is highly relevant in the context of mesosynteny.

James implies that mesosyntenic relationships would be maintained by meiosis. At first I didn't understand this logic. I wonder if he really means that chromosomal structure is maintained by meiosis and this would lead to retention of mesosyntenic relationships with other genomes (as long as their structures are also maintained)?

This is speculative, however the common phylogenetic link between RIP and mesosynteny strengthens the link between mesosynteny and meiosis, as RIP does not occur without meiosis. Based on what we currently know about mesosynteny, which is very little, when searching for a potential cause of mesosyntenic rearrangements we would regardless have considered meiotic “crossing-over” as a strong candidate. If this does prove to be the case, In this sense meioisis could be seen as maintaining the mesosyntenic pattern. The meioitic requirement for the pairing of sister-chromosomes could potentially also play a role in maintaining the high frequency of intra-chromosomal rearrangement relative to inter-chromosomal rearrangement.

References:

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

Chapter 9: Attribution Statement

Title: **Finished Genome of *Mycosphaerella graminicola* Reveals Stealth Pathogenesis and Extreme Plasticity.**

Authors: Stephen B. Goodwin, Sarrah Ben M'Barek, Braham Dhillon, Alexander H. J. Wittenberg, Charles F. Crane, Theo A. J. Van der Lee, Jane Grimwood, Andrea Aerts, John Antoniw, Andy Bailey, Burt Bluhm, Judith Bowler, Jim Bristow, Blondy Canto-Canche, Alice Churchill, Laura Conde-Ferràez, Hans Cools, Pedro M. Coutinho, Michael Csukai, Paramvir Dehal, Pierre De Wit, Bruno Donzelli, Andrew J. Foster, Kim Hammond-Kosack, **James Hane**, Bernard Henrissat, Andrzej Kilian, Edda Koopmann, Yiannis Kourmpetis, Arnold Kuzniar, Erika Lindquist, Vincent Lombard, Chris Maliepaard, Natalia Martins, Rahim Mehrabi, Richard Oliver, Alisa Ponomarenko, Jason Rudd, Asaf Salamov, Jeremy Schmutz, Henk J. Schouten, Harris Shapiro, Ioannis Stergiopoulos, Stefano F. F. Torriani, Hank Tu, Ronald P. de Vries, Ad Wiebenga, Lute-Harm Zwiers, Igor V. Grigoriev, Gert H. J. Kema

Citation: *PLOS Genetics* 7(6): e1002070 (2011)

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 (JKH) made the following contributions to this chapter:

- Assisted finishing of genome sequence via analysis of synteny between *Mycosphaerella graminicola* and *Phaeosphaeria nodorum*.

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

--	--
James Hane (Ph. D. candidate)	Date

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

--	--
Richard Oliver (Principal supervisor)	Date

OPEN ACCESS Freely available online

PLoS GENETICS

Finished Genome of the Fungal Wheat Pathogen *Mycosphaerella graminicola* Reveals Dispensome Structure, Chromosome Plasticity, and Stealth Pathogenesis

Stephen B. Goodwin[1☯*], Sarrah Ben M'Barek[2], Braham Dhillon[3], Alexander H. J. Wittenberg[2], Charles F. Crane[1], James K. Hane[4], Andrew J. Foster[5], Theo A. J. Van der Lee[2], Jane Grimwood[6,7], Andrea Aerts[7], John Antoniw[8], Andy Bailey[9], Burt Bluhm[10], Judith Bowler[11], Jim Bristow[7], Ate van der Burgt[2], Blondy Canto-Canché[12], Alice C. L. Churchill[13], Laura Conde-Ferràez[12], Hans J. Cools[8], Pedro M. Coutinho[14], Michael Csukai[11], Paramvir Dehal[7], Pierre De Wit[15], Bruno Donzelli[16], Henri C. van de Geest[2], Roeland C. H. J. van Ham[2], Kim E. Hammond-Kosack[8], Bernard Henrissat[14], Andrzej Kilian[17], Adilson K. Kobayashi[18], Edda Koopmann[19], Yiannis Kourmpetis[15], Arnold Kuzniar[15], Erika Lindquist[7], Vincent Lombard[14], Chris Maliepaard[15], Natalia Martins[20], Rahim Mehrabi[21], Jan P. H. Nap[2], Alisa Ponomarenko[3], Jason J. Rudd[8], Asaf Salamov[7], Jeremy Schmutz[6,7], Henk J. Schouten[2], Harris Shapiro[7], Ioannis Stergiopoulos[15], Stefano F. F. Torriani[22], Hank Tu[7], Ronald P. de Vries[23], Cees Waalwijk[2], Sarah B. Ware[2], Ad Wiebenga[23], Lute-Harm Zwiers[23], Richard P. Oliver[24], Igor V. Grigoriev[7☯*], Gert H. J. Kema[2☯*]

1 USDA–Agricultural Research Service, Purdue University, West Lafayette, Indiana, United States of America, **2** Plant Research International B.V., Wageningen, The Netherlands, **3** Department of Botany and Plant Pathology, Purdue University, West Lafayette, Indiana, United States of America, **4** School of Veterinary and Biomedical Sciences, Murdoch University, Perth, Australia, **5** IBWF e.V., Institute for Biotechnology and Drug Research, Kaiserslautern, Germany, **6** HudsonAlpha Institute of Biotechnology, Huntsville, Alabama, United States of America, **7** DOE Joint Genome Institute, Walnut Creek, California, United States of America, **8** Rothamsted Research, Department of Plant Pathology and Microbiology, Harpenden, United Kingdom, **9** School of Biological Sciences, University of Bristol, Bristol, United Kingdom, **10** University of Arkansas, Fayetteville, Arkansas, United States of America, **11** Syngenta, Jealott's Hill Research Centre, Bracknell, United Kingdom, **12** Unidad de Biotecnología, Centro de Investigación Científica de Yucatán, A.C., (CICY), Mérida, México, **13** Department of Plant Pathology and Plant-Microbe Biology, Cornell University, Ithaca, New York, United States of America, **14** Architecture et Fonction des Macromolecules Biologiques, CNRS, Marseille, France, **15** Wageningen University and Research Centre, Wageningen, The Netherlands, **16** USDA–Agricultural Research Service, Ithaca, New York, United States of America, **17** Diversity Arrays Technology Pty Ltd, Yarralumla, Australia, **18** Embrapa Meio-Norte, Teresina, Piauí, Brazil, **19** Bayer CropScience AG, Monheim, Germany, **20** Embrapa-Cenargen, Brasilia, Brazil, **21** Department of Genetics, Seed and Plant Improvement Institute, Karaj, Iran, **22** Plant Pathology, Institute of Integrative Biology, Swiss Federal Institute of Technology (ETH), Zürich, Switzerland, **23** CBS–KNAW Fungal Biodiversity Centre, Utrecht, The Netherlands, **24** Environment and Agriculture, Curtin University, Bentley, Australia

Abstract

The plant-pathogenic fungus *Mycosphaerella graminicola* (asexual stage: *Septoria tritici*) causes septoria tritici blotch, a disease that greatly reduces the yield and quality of wheat. This disease is economically important in most wheat-growing areas worldwide and threatens global food production. Control of the disease has been hampered by a limited understanding of the genetic and biochemical bases of pathogenicity, including mechanisms of infection and of resistance in the host. Unlike most other plant pathogens, *M. graminicola* has a long latent period during which it evades host defenses. Although this type of stealth pathogenicity occurs commonly in Mycosphaerella and other Dothideomycetes, the largest class of plant-pathogenic fungi, its genetic basis is not known. To address this problem, the genome of *M. graminicola* was sequenced completely. The finished genome contains 21 chromosomes, eight of which could be lost with no visible effect on the fungus and thus are dispensable. This eight-chromosome dispensome is dynamic in field and progeny isolates, is different from the core genome in gene and repeat content, and appears to have originated by ancient horizontal transfer from an unknown donor. Synteny plots of the *M. graminicola* chromosomes versus those of the only other sequenced Dothideomycete, *Stagonospora nodorum*, revealed conservation of gene content but not order or orientation, suggesting a high rate of intra-chromosomal rearrangement in one or both species. This observed "mesosynteny" is very different from synteny seen between other organisms. A surprising feature of the *M. graminicola* genome compared to other sequenced plant pathogens was that it contained very few genes for enzymes that break down plant cell walls, which was more similar to endophytes than to pathogens. The stealth pathogenesis of *M. graminicola* probably involves degradation of proteins rather than carbohydrates to evade host defenses during the biotrophic stage of infection and may have evolved from endophytic ancestors.

Citation: Goodwin SB, Ben M'Barek S, Dhillon B, Wittenberg AHJ, Crane CF, et al. (2011) Finished Genome of the Fungal Wheat Pathogen *Mycosphaerella graminicola* Reveals Dispensome Structure, Chromosome Plasticity, and Stealth Pathogenesis. PLoS Genet 7(6): e1002070. doi:10.1371/journal.pgen.1002070

Editor: Harmit S. Malik, Fred Hutchinson Cancer Research Center, United States of America

Received October 17, 2010; **Accepted** March 24, 2011; **Published** June 9, 2011

Funding: The work conducted by the U.S. Department of Energy Joint Genome Institute was supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. Supported in part by USDA CRIS projects 3602-22000-013-00D and 3602-22000-015-00D. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: Steve.Goodwin@ARS.USDA.gov, sgoodwin@purdue.edu (SBG); gert.kema@wur.nl (GHJK); IVGrigoriev@lbl.gov (IVG)

◑ These authors contributed equally to this work.

Introduction

The ascomycete fungus *Mycosphaerella graminicola* (Figure S1) causes septoria tritici blotch (STB), a foliar disease of wheat that poses a significant threat to global food production. Losses to STB can reduce yields of wheat by 30 to 50% with a huge economic impact [1]; global expenditures for fungicides to manage STB total hundreds of millions of dollars each year [2–3]. This fungus is difficult to control because populations contain extremely high levels of genetic variability [4] and it has very unusual biology for a pathogen. Unlike most other plant pathogens [5–7], *M. graminicola* infects through stomata rather than by direct penetration and there is a long latent period of up to two weeks following infection before symptoms develop. The fungus evades host defenses [8] during the latent phase, followed by a rapid switch to necrotrophy immediately prior to symptom expression 12–20 days after penetration [5,9–10]. Such a switch from biotrophic to necrotrophic growth at the end of a long latent period is an unusual characteristic shared by most fungi in the genus Mycosphaerella. Very little is known about the cause or mechanism of this lifestyle switch [9–10] even though Mycosphaerella is one of the largest and most economically important genera of plant-pathogenic fungi.

A striking aspect of *M. graminicola* genetics is the presence of many dispensable chromosomes [11]. These can be lost readily in sexual progeny with no apparent effect on fitness. However, the structure and function of dispensable chromosomes are not known. Here we report the first genome of a filamentous fungus to be finished according to current standards [12]. The 21-chromosome, 39.7-Mb genome of *M. graminicola* revealed an apparently novel origin for dispensable chromosomes by horizontal transfer followed by extensive recombination, a possible mechanism of stealth pathogenicity and exciting new aspects of genome structure. The genome provides a finished reference for the Dothideomycetes, the largest class of ascomycete fungi, which also includes the apple scab pathogen *Venturia inaequalis*, the southern corn leaf blight pathogen *Cochliobolus heterostrophus*, the black Sigatoka pathogen of banana, *M. fijiensis*, and numerous other pathogens of almost every crop.

Results

Features of the finished genome

The finished genome of *M. graminicola* isolate IPO323 consists of 21 complete chromosomes, telomere to telomere (Figure S2), with the exceptions of one telomere of chromosome 21 and two internal gaps of unclonable DNA that are missing from chromosome 18 (Table 1). Alignments between the 21 chromosomes and two genetic linkage maps yielded an excellent correspondence (Figure 1 and Figure S3), representing the most complete and the first finished sequence of a filamentous fungus. The next most complete genome of a filamentous fungus is that of *Aspergillus fumigatus*, which did not include centromere sequences and contained 11 gaps in total [13]. The complete 43,960-bp mitochondrial genome also was obtained and has been described elsewhere [14].

Sexual activation of chromosome plasticity and repeat-induced point mutation

Comparative genome hybridizations using a whole-genome tiling array made from the genome sequence of IPO323 demonstrated striking sexually activated chromosomal plasticity in progeny isolates (Figure 2) and chromosome number polymorphisms in field isolates. For example, isolate IPO94269, a field strain from bread wheat in the Netherlands, was missing two chromosomes that were present in IPO323 (Figure 2A).

Sexual-driven genome plasticity was particularly evident among progeny isolates in the two mapping populations, including losses of chromosomes that were present in both parents and disomy for others [11]. For example, progeny isolate #51 of the cross between IPO323 and IPO94269 lost chromosomes 14 and 21 (Figure 2B) even though they were present in both parents. This isolate also was missing chromosome 20, which was polymorphic for presence between the parents of the cross. More surprisingly, this isolate was disomic for chromosomes 4 and 18 (Figure 2B), indicating that chromosomes can be both gained and lost during meiosis. For chromosome 18, both copies must have originated from IPO323 because no homolog was present in IPO94269. Molecular markers for chromosome 4 appeared to be heterozygous indicating that both parents contributed a copy to progeny isolate #51 (data not shown). Progeny isolate #2133 of the cross between isolates IPO323 and IPO95052 showed loss of three dispensable chromosomes (15, 18 and 21) that were present in both parents (Figure 2C), most likely due to non-disjunction during meiosis. Thus, extreme genome plasticity was manifested as chromosome number and size polymorphisms [11] that were generated during meiosis and extended to core as well as dispensable chromosomes.

The whole-genome hybridizations also indicated that the core and dispensable chromosomes can be remarkably uniform for gene content, given the high capacity of the latter for change. Comparative genome hybridizations between IPO323 and IPO95052, an isolate from a field of durum wheat in Algeria, showed that they had the same complement of core and dispensable chromosomes (Figure 2D). This was surprising, because populations of the pathogen from durum wheat (a tetraploid) usually are adapted to that host and not to hexaploid bread wheat, yet the chromosomal complements of isolates from these hosts on different continents were the same.

Evidence for repeat-induced point mutation (RIP), a mechanism in fungi that inactivates transposons by introducing C to T

Author Summary

The plant-pathogenic fungus *Mycosphaerella graminicola* causes septoria tritici blotch, one of the most economically important diseases of wheat worldwide and a potential threat to global food production. Unlike most other plant pathogens, *M. graminicola* has a long latent period during which it seems able to evade host defenses, and its genome appears to be unstable with many chromosomes that can change size or be lost during sexual reproduction. To understand its unusual mechanism of pathogenicity and high genomic plasticity, the genome of *M. graminicola* was sequenced more completely than that of any other filamentous fungus. The finished sequence contains 21 chromosomes, eight of which were different from those in the core genome and appear to have originated by ancient horizontal transfer from an unknown donor. The dispensable chromosomes collectively comprise the dispensome and showed extreme plasticity during sexual reproduction. A surprising feature of the *M. graminicola* genome was a low number of genes for enzymes that break down plant cell walls; this may represent an evolutionary response to evade detection by plant defense mechanisms. The stealth pathogenicity of *M. graminicola* may involve degradation of proteins rather than carbohydrates and could have evolved from an endophytic ancestor.

transitions in repeated sequences [15–16], was seen in genome-wide analyses of transition:transversion ratios in long terminal repeat (LTR) pairs from 20 retrotransposon insertions which had 255 transitions and 6 transversions for a ratio of 42.5:1. Similarly high transition:transversion ratios were found in all repetitive sequences analyzed and extended to the coding regions in addition to the LTRs [17]. The reverse transcriptase coding regions from transposon families RT11 and RT15 had transition:transversion ratios of 27.8:1 and 25.3:1, respectively, instead of the 1:1 ratio expected among 6,939 mutations analyzed. This high incidence of transitions most likely reflects changes caused by RIP. The coding regions of all transposons with more than 10 copies included stop codons that prevent proper translation, indicating that they were inactivated.

Core and dispensable chromosomes are highly divergent

There were significant differences in structure and gene content between the 13 core and eight dispensable chromosomes (Table 1 and Table 2); the latter are referred to collectively as the dispensome. The dispensome constituted about 12% of the genomic DNA but contained only 6% of the genes. In contrast, the 13 core chromosomes had twice as many genes per Mb of DNA, about half as much repetitive DNA, a significantly higher G+C content, and much higher numbers of unique genes (Table 1 and Table 2). Genes in the dispensome were significantly shorter, usually were truncated relative to those on the core chromosomes (Table 2) and had dramatic differences in codon usage (Figure S4).

About 59% of the genes on core chromosomes could be annotated compared to only 10% of those on the dispensome (Table 2). Some unique genes in the dispensome with intact, presumably functional reading frames, had possible paralogs on the core chromosomes (Figure S5) that appeared to be inactivated by mutations (Figure S6). A majority of the annotated dispensome genes coded for putative transcription factors or otherwise may function in gene regulation or signal transduction (Table S1). Most of the redundant genes on the dispensome were copies of genes present on core chromosomes, yet no syntenic relationships could be identified. Instead, each dispensable chromosome contained genes and repetitive sequences from all or most of the core chromosomes (Figure 3 and Figure S7) with additional unique genes of unknown origin. Sharing of genetic material applied to core chromosomes as well as the dispensome, consistent with a high level of recombination (Figure S8). Whether the primary direction of transfer is from core to dispensable chromosomes or vice versa is not known.

The dispensome contained fewer genes encoding secreted proteins such as effectors and other possible pathogenicity factors compared to the core set. Signal peptides showed no enrichment on the dispensome (Table S1) except for a few clusters overlapping with transposon-related repeats. Although mature microRNAs have not been demonstrated in fungi, they may be important regulatory molecules. In the *M. graminicola* genome, 418 non-overlapping loci potentially encoding pre-microRNA-like small RNA (pre-milRNA) were predicted computationally based on the RFAM database [18]. This number was similar to the 434 loci predicted in the 41-Mb genome of *Neurospora crassa* using the same approach. Of the 418 putative pre-milRNA loci predicted in the genome of *M. graminicola*, 88 (21%) are located on the 11% of the genome present as dispensome. This is about twice as much as is expected on the basis of a random distribution. Therefore, the dispensome is enriched for pre-milRNA loci.

The 418 pre-milRNA loci code for 385 non-redundant pre-milRNA sequences that can give rise to distinguishable mature milRNAs. The occurrence of mature milRNAs derived from the predicted set was analyzed in a small-RNA data consisting of almost 6 million reads (Illumina platform) generated from germinated spores of *M. graminicola* isolate IPO323 (Table S2).

Many of the non-redundant predicted milRNA sequences were represented in the RNA reads, at widely different amounts per sequence. In total, 65 of the 385 non-redundant sequences were observed 10 times or more. Two predicted sequences occurred more than a thousand times each, experimentally confirming the presence of putatively mature milRNAs derived from computationally predicted pre-milRNA sequences. In *N. crassa*, computationally predicted putative milRNA sequences also were confirmed experimentally [19], supporting the likelihood of their existence in *M. graminicola*.

The origin of the dispensome of *M. graminicola* is not clear. The two most likely origins would be degeneration of copies of the core chromosomes or by horizontal transfer. Disomy for core chromosomes, as seen in one of the progeny isolates, could provide the origin for a dispensable chromosome. If one of the two chromosome copies became preferentially subject to RIP followed by breakage or interstitial deletions this could result in a degenerated copy of that core chromosome. However, in that case we would expect the dispensome to share large regions of synteny with specific core chromosomes, and this was not observed, which renders this explanation less likely.

The large differences in codon usage between core and dispensable chromosomes could be explained by horizontal transfer or possibly by RIP. To discriminate between these hypotheses, RIP was simulated on the genes of the core chromosomes. Principal components analysis (PCA) of the simulated data set did not reduce the differences in codon bias (Figure S9A); if anything, it made them farther apart. This result was consistent whether it included only putative functional, truncated copies or entire pseudogenes after RIPping (data not shown). DeRIPping of genes on the dispensable chromosomes also did not affect the results (Figure S9B), so RIP could not explain the observed differences in codon usage between core and dispensable

Table 1. Sizes and gene contents of the 21 chromosomes of *Mycosphaerella graminicola* isolate IPO323.

Chromosome		All genes		Unique genes[a]		Signal peptides	Average gene size (bp)	Genes/Mb DNA	Percent G+C	Percent repetitive	milRNAs/Mb DNA[b]
Number	Size	Number	Annotated	Number	Annotated						
1	6,088,797	1,980	1,258	1,067	497	208	1338.6	325	53.1	9.5	9.7
2	3,860,111	1,136	650	607	238	108	1402.7	294	52.4	15.7	9.6
3	3,505,381	1,071	630	583	246	122	1337.1	306	52.6	14.2	6.3
4	2,880,011	821	498	421	182	81	1388.6	285	52.2	16.1	13.2
5	2,861,803	778	489	389	180	91	1352.6	272	52.0	19.1	18.9
6	2,674,951	692	427	328	152	66	1353.0	259	51.4	22.2	12.3
7	2,665,280	766	357	462	131	96	1202.7	287	52.6	14.0	16.1
8	2,443,572	689	397	384	159	62	1311.2	282	51.7	17.6	13.5
9	2,142,475	604	353	305	134	69	1345.1	282	51.5	20.8	18.7
10	1,682,575	516	298	266	110	46	1418.7	307	52.5	14.1	9.5
11	1,624,292	488	279	270	115	65	1352.5	300	52.8	10.5	5.5
12	1,462,624	408	227	232	96	59	1254.3	279	52.3	14.5	10.9
13	1,185,774	330	183	165	68	47	1195.7	278	52.0	17.8	17.7
14	773,098	114	25	48	5	3	920.1	147	48.5	36.7	23.3
15	639,501	86	6	44	1	2	773.7	134	51.0	34.4	25.0
16	607,044	88	5	40	1	5	898.5	145	51.5	25.6	31.3
17	584,099	78	6	36	1	1	777.9	134	52.0	26.4	18.8
18[c]	573,698	64	7	28	4	0	965.1	112	48.6	40.3	33.1
19	549,847	87	8	53	3	4	658.3	158	51.3	25.1	23.6
20	472,105	79	4	41	2	4	863.1	167	51.5	21.1	25.4
21[d]	409,213	58	4	21	1	2	921.6	142	51.9	30.1	14.7
Total	39,686,251	10,933	6,111	5,790	2,326	1,141					13.5

[a] At a BLAST cutoff value of $1 \times e^{-20}$.

[b] Predicted numbers of loci for pre-microRNA-like small RNAs.

[c] This chromosome contains two internal gaps of unclonable DNA marked by gaps of 1.4 and 4.5 kb; all other chromosomes are complete.

[d] The sequence of one telomere is missing from this chromosome; all other telomeres are complete.

doi:10.1371/journal.pgen.1002070.t001

chromosomes. PCA of a sample of genes shared between core and dispensable chromosomes showed few differences in codon bias (Figure S9C) or amino acid composition (Figure S9D), consistent with an origin by duplication and exchange among chromosomes. This conclusion was supported when the analysis was expanded to include all genes with putative homologs on core and dispensable chromosomes (Figure S9E) even though these genes had a very different codon usage compared to the entire sets of genes on the core chromosomes (Figure S9F).

To test the horizontal transfer hypothesis, additional PCAs were performed on simulated horizontal transfer data sets made by combining the genome of *M. graminicola* with those of two other fungi. Best non-self BLAST hits for genes on the *M. graminicola* dispensome most often were to fungi in the Pleosporales or Eurotiales (Table S3), so published genomes from species representing those orders were chosen for analysis. PCA of the combined genomes of *M. graminicola* and *Stagonospora nodorum* (representing the Pleosporales) gave separate, tight clusters for the core chromosomes of *M. graminicola* versus most of those from *S. nodorum* (Figure S10A). Dispensable chromosomes of *M. graminicola* formed a looser, distinct cluster, and a fourth cluster was comprised of *M. graminicola* chromosome 14 plus scaffolds 44 and 45 of *S. nodorum* (Figure S10A); this may indicate the existence of dispensable chromosomes in the latter species. Analysis of the combined genomes of *M. graminicola* plus *Aspergillus fumigatus* (Eurotiales) gave a similar result (Figure S10B). The separate clustering by PCA of the *M. graminicola* dispensome and core chromosomes is consistent with an origin by horizontal transfer, but not from either of the two species tested. PCA on the frequencies of repetitive elements also indicated a separation between core and dispensable chromosomes (Figure S11), consistent with the horizontal transfer hypothesis.

A more refined test of the RIP hypothesis was performed by using the observed rates of all mutations in families of transposons with 10 or more elements to simulate mutational changes on replicated samples drawn from the core chromosomes. Observed mutation rates were calculated from aligned sequences; multicopy transposons were chosen for this analysis because they are the most likely to have been processed through the RIP machinery so will reflect the actual biases that occur in *M. graminicola*. Codon bias and other parameters in the mutated samples were then compared to those in the dispensome and in the original, non-mutated samples. Application of the mutational changes moved the samples drawn from the core chromosomes closer to the value observed for the dispensome, but the dispensome remained distinct except for a few of the analyses that are least likely to be affected by selection (Figure 4). This confirmed that the dispensome has been subject to RIP but that this alone was not sufficient to explain the observed pattern of codon usage.

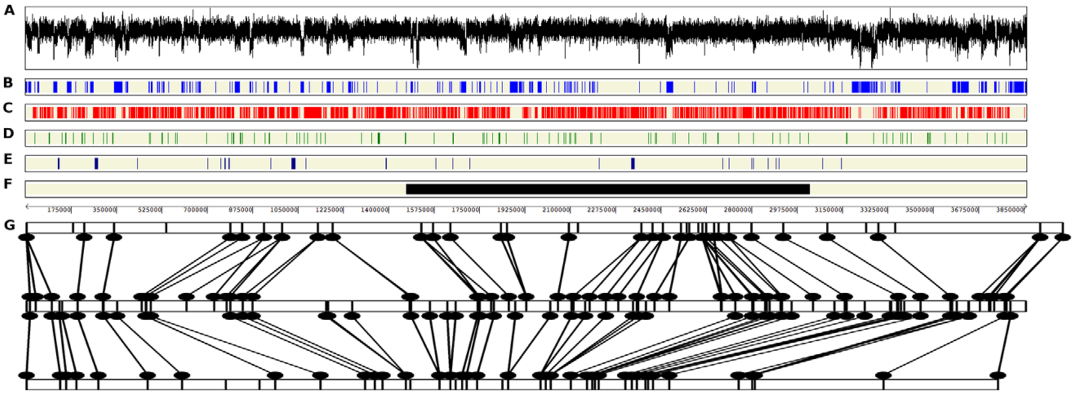


Figure 1. Features of chromosome 2 of *Mycosphaerella graminicola* and alignment to genetic linkage maps. A, Plot of GC content. Areas of low GC usually correspond to regions of repetitive DNA. B, Repetitive regions of the *M. graminicola* genome. C, Single-copy (red) regions of the *M. graminicola* genome. D, Locations of genes for proteins containing signal peptides. E, Locations of homologs involved in pathogenicity or virulence that have been experimentally verified in species pathogenic to plant, animal or human hosts. F, Approximate locations of quantitative trait loci (QTL) for pathogenicity to wheat. G, Alignments between the genomic sequence and two genetic linkage maps of crosses involving isolate IPO323. Top half, Genetic linkage map of the cross between IPO323 and the Algerian durum wheat isolate IPO95052. Bottom half, Genetic linkage map of the cross between bread wheat isolates IPO323 and IPO94269. The physical map represented by the genomic sequence is in the center. Lines connect mapped genetic markers in each linkage map to their corresponding locations on the physical map based on the sequences of the marker loci. Exceptions to the almost perfect alignment between the three maps are indicated by crossed lines, most likely due to occasional incorrect scorings of the marker alleles. Chromosome 2 was used for this illustration because no QTL mapped to chromosome 1.
doi:10.1371/journal.pgen.1002070.g001

A new type of synteny

Pairwise sequence comparisons between the chromosomes of *M. graminicola* and scaffolds of *Stagonospora nodorum*, another wheat pathogen in the Dothideomycetes but in a different order from Mycosphaerella, revealed multiple regions with approximately 70–90% similarity (Figure 5). However, the similarity did not extend to the dispensome, which generally was different from all of the *S. nodorum* scaffolds. Detailed examination showed that each region of similarity generally represents only one or a few genes in both organisms. Comparisons between the initial draft genome (version 1.0) of *M. graminicola* (Figure 5A) and the finished sequence (Figure 5B) revealed some misassemblies and also indicated scaffolds that ultimately were joined in the final assembly.

A surprising result was that the dot-plot patterns were very different from those that characterize the macro- or microsynteny seen in other organisms when viewed at a whole-scaffold/chromosome scale. Instead of the expected diagonal lines indicating chromosomal regions with content in the same order and orientation, the dots are scattered quasi-randomly within 'blocks' defined by scaffold/chromosome boundaries (Figure 5). For many *S. nodorum* scaffolds the vast majority of dots related are shared exclusively with one or a small number of *M. graminicola* chromosomes. For example, there are predominant one-to-one relationships between *M. graminicola* version 3 chromosomes 11 and 12 with *S. nodorum* scaffolds 21 and 7 (Figure 5B, circle V), respectively. Similarly, *M. graminicola* chromosomes 5–10 each had strong relationships with 2 to 4 scaffolds of *S. nodorum*. We refer to this conservation of gene content but not order or orientation among chromosomes as 'mesosynteny'. Analyses of additional genomes has shown that mesosynteny as defined here occurs among all Dothideomycetes tested and may be unique to that class of fungi (data not presented).

Mesosynteny as a tool to assist genome assembly

Macrosyntenic relationships are used commonly to assist the assembly and finishing of fragmented genome sequences [20–23], particularly in prokaryotes. Sequences that are macrosyntenic to a long segment of a closely related genome are highly likely to be joined physically. If mesosynteny between a new genome assembly and a reference genome also may be used to suggest scaffolds that should be juxtaposed it could significantly reduce the cost and complexity of assembling and finishing genomes. To test whether mesosynteny could be used to predict scaffold or contig joins in a genomic sequence, versions 1 and 2 of the *M. graminicola* genome assembly were analyzed to determine whether any of the improvements in the finished genome could have been predicted bioinformatically by mesosynteny (Dataset S1).

The first version of the *M. graminicola* genome consisted of 129 scaffolds (http://genome.jgi-psf.org/Mycgr1/Mycgr1.home.html). Comparison of *M. graminicola* version 1 scaffolds with those of the *P. nodorum* genome predicted all scaffold joins made in version 2 (Figure 5, Dataset S1). Version 1 scaffolds 10 and 14 (Figure 5: group I), 7 and 17 (groups II, VII and IX), and 12 and 22 (groups III and VIII) were joined into chromosomes 7, 5 and 10, respectively. Mesosynteny also indicated both instances where version 1 scaffolds were assembled incorrectly and subsequently were split in version 2. Compared to the scaffolds of *P. nodorum*, *M. graminicola* version 1 scaffold 4 exhibited regions of mesosynteny adjacent to regions of no synteny. Corrections to the assembly made in version 2 separated these two distinct regions into separate chromosomes. Version 1 scaffolds 4 and 9 (Figure 5: groups IV/VI and V) were corrected to version 2 chromosomes 6 and 16 (Figure 5: group IV/VI) and chromosomes 12 and 21 (Figure 5: group V) respectively. Mesosynteny was remarkably successful and has great potential to assist the assembly and finishing of fungal genomes.

A mechanism of stealth pathogenesis

Generally, gene families involved in cell wall degradation are expanded in fungal plant pathogens [24–25]. However, in *M. graminicola*, gene families characterized by the Carbohydrate-

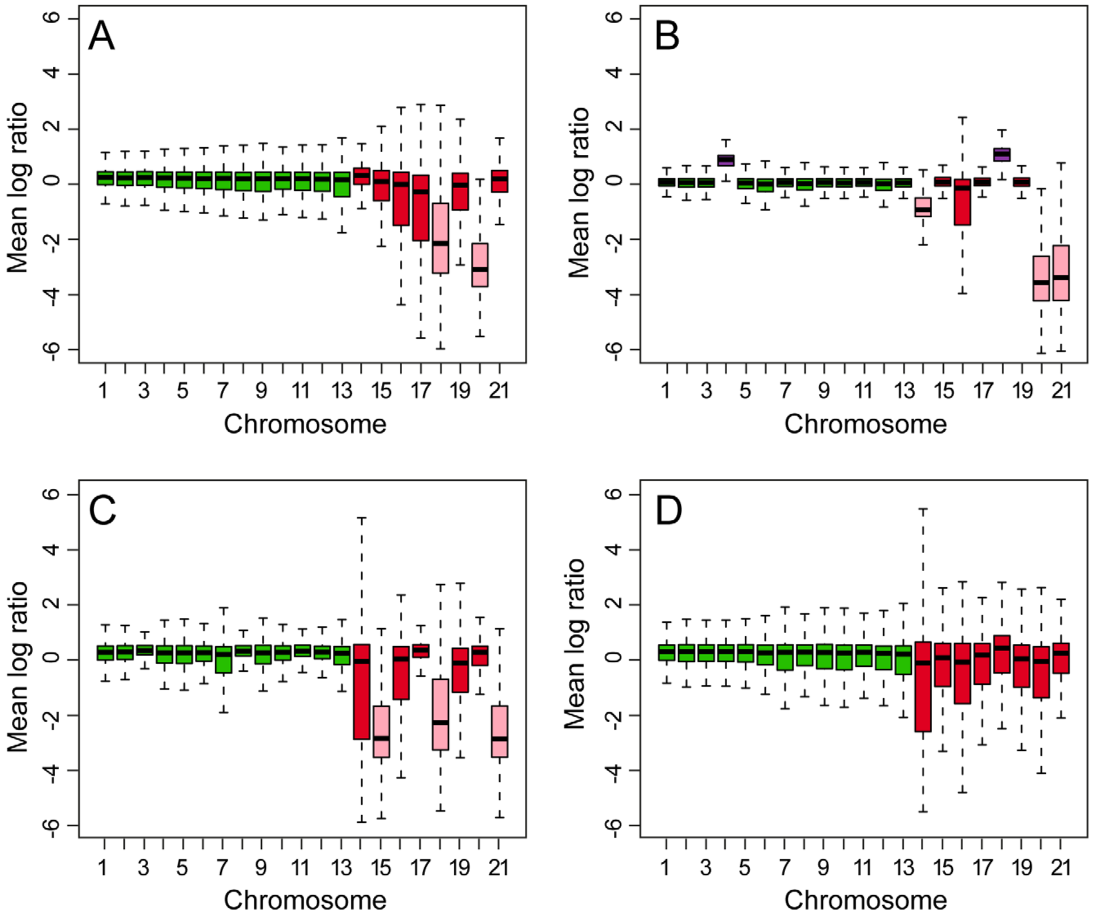


Figure 2. Box plots of comparative genome hybridizations (CGH) of DNA from five isolates of *Mycosphaerella graminicola* to a whole-genome tiling array made from the finished sequence of isolate IPO323. A, CGH between IPO323 and the Dutch field isolate IPO94269. B, CGH between IPO323 and progeny isolate #51 from the cross between IPO323 and IPO94269. C, CGH between IPO323 and progeny isolate #2133 of the cross between IPO323 and IPO95052. D, CGH between IPO323 and Algerian field isolate IPO95052, which was isolated from and is adapted to durum (tetraploid) wheat. The genomic difference between the strains for each CGH is shown by 21 box plots, one for each chromosome of *M. graminicola*. The horizontal line in each box is the median log ratio of hybridization signals of the two strains; the upper and lower ends of a box represent the 25% and 75% quartiles. The whiskers extending from each box indicate 1.5 times the interquartile range, the distance between the 25% and 75% quartiles. The larger the deviation from 0, the greater the difference between the strains for a particular chromosome. Pink boxes that are significantly less than the zero line indicate missing chromosomes. The purple boxes in panel B (4 and 18) that are significantly higher than the zero line indicate chromosomes that are disomic.
doi:10.1371/journal.pgen.1002070.g002

Active Enzyme database (CAZy) [26] as plant cell wall polysaccharidases were severely reduced in size (Figure 6). According to the CAZy analysis, the genome of *M. graminicola* contains fewer genes for cellulose degradation than those of six other fungi with sequenced genomes including both grass pathogens and saprophytes (Table 3), and only about one-third as many genes for cell wall degradation in total compared to the other plant pathogens (Table S4). This reduction in CAZymes in *M. graminicola* was very visible when the putative genes were divided based on polysaccharide substrate (Table S4). In addition, genes involved in appressorium formation, which are required for pathogenesis of many plant pathogens including *Magnaporthe oryzae* [27], were absent or reduced in the *Mycosphaerella graminicola* genome, reflecting its alternative host-penetration strategy.

To further analyze the mechanism of stealth pathogenesis, we profiled the growth on polysaccharides of *M. graminicola* compared to *Stagonospora nodorum* and *Magnaporthe oryzae*, two pathogens of the cereals wheat and rice, respectively, with sequenced genomes (Figure S12). Growth of *M. graminicola* corresponded with the CAZy annotation for a strongly reduced number of genes encoding putative xylan-degrading enzymes. Furthermore, the CAZy annotation demonstrated that *M. graminicola* contains a much smaller set of glycoside hydrolases, carbohydrate esterases, and carbohydrate binding modules (CBMs) compared to the other two cereal pathogens (Table S5). The strong reduction of CBMs in *M. graminicola* suggests a different strategy in the degradation of plant cell walls compared to the other two species. The *M. graminicola* genome is particularly depauperate for enzymes

Table 2. Differences between essential and dispensable chromosomes in the genome of *Mycosphaerella graminicola* isolate IPO323.

	Chromosomes		
Statistic	**Core (1–13)**	**Dispensable (14–21)**	**Combined (1–21)**
Size in bp			
Total	35,077,646	4,608,605	39,686,251
Mean	2,698,280	576,076***	1,889,821
Percent	88.4	11.6	100.0
All genes			
Total	10,279	654	10,933
Mean	790.7	81.8***	521
Percent of total	94.0	6.0	100.0
Unique genes[a]			
Total	5,479	311	5,790
Mean	421.5	38.9***	276
Percent of all	53.3	47.6	53.0
Annotated genes			
Total	6,046	65	6,111
Mean	465.1	8.1***	291.0
Percent of all	58.8	9.9	55.9
Unique total	2,308	18	2,326
Unique mean	177.5	2.3***	110.8
Transcript size, mean in bp	1327.1	847.3***	1144.3
Gene density, Mb^{-1}	288.9	142.4***	233.1
Repetitive DNA, mean	15.9%	30.0%***	21.2%
G+C, mean	52.3%	50.9%**	51.7%

[a] At a BLAST cutoff value of $1\times e^{-20}$.

***The mean for the dispensable chromosomes is significantly different from that for the essential chromosomes at $P<0.001$ by one-tailed t test.

**The mean for the dispensable chromosomes is significantly different from that for the essential chromosomes at $P=0.012$ by one-tailed t test.

doi:10.1371/journal.pgen.1002070.t002

degrading cellulose, xylan and xyloglucan compared to the other two species, so is very atypical for a cereal pathogen.

A possible mechanism of stealth pathogenesis was indicated by gene families that were expanded in the genome of *M. graminicola*. In comparative analyses of gene families and PFAM domains with several other fungi, the most striking expansions were observed for peptidases (M3, S28, pro-kuma, M24, metalloendopeptidase, metalloproteinase) and alpha amylases (glycoside hydrolase family 13) (Tables S6 and S7). This suggests that alternative nutrition sources during the biotrophic phase of infection may be proteins which are available in the apoplast, or possibly starch from chloroplasts that are released early in the infection process [5]. Overall, these analyses revealed that the genome of *M. graminicola* differs significantly from those of other cereal pathogens with respect to genes involved in plant penetration as well as polysaccharide and protein degradation (Figure 6, Table 3), which most likely reflects its stealthy mode of pathogenesis.

Differences in gene expression during the different stages of infection were evident from an analysis of EST sequences [9] from wheat leaves 5, 10 and 16 days after inoculation (DAI) with *M. graminicola*. Most genes were present at only one sampling time with little overlap, particularly between the library from the biotrophic stage of infection (5 DAI) compared to the other two (Figure S13A). Lack of overlap extended to a library from minimal medium minus nitrogen to simulate the nitrogen starvation thought to occur during infection (Figure S13B). Expression of genes for cell wall-degrading enzymes also was reduced during the biotrophic stage of infection [9], consistent with the stealth-pathogenicity hypothesis.

Discussion

The dispensome as defined here includes all parts of the genome that can be missing in field or progeny isolates with no obvious effects on fitness in axenic culture, on a susceptible host or during mating. For *M. graminicola*, this includes the eight known dispensable chromosomes in isolate IPO323 plus any others that may be discovered in the future. The core genome consists of all chromosomes that are always present in field and progeny isolates, presumably because they contain genes that are vital for survival so cannot be lost. Both core and dispensable chromosomes may be present in two or possibly more copies, but core chromosomes are never absent.

The dispensome of *M. graminicola* is very different from the supernumerary or B chromosomes in plants and some animals. The B chromosomes of plants contain few if any genes and are composed mostly of repetitive elements assembled from the A chromosomes. They may have a negative effect on fitness [28] and appear to be maintained primarily by meiotic drive [29]. In contrast, the dispensome of *M. graminicola* contains many unique and redundant genes and is not maintained by meiotic drive, as individual chromosomes are lost readily during meiosis [11].

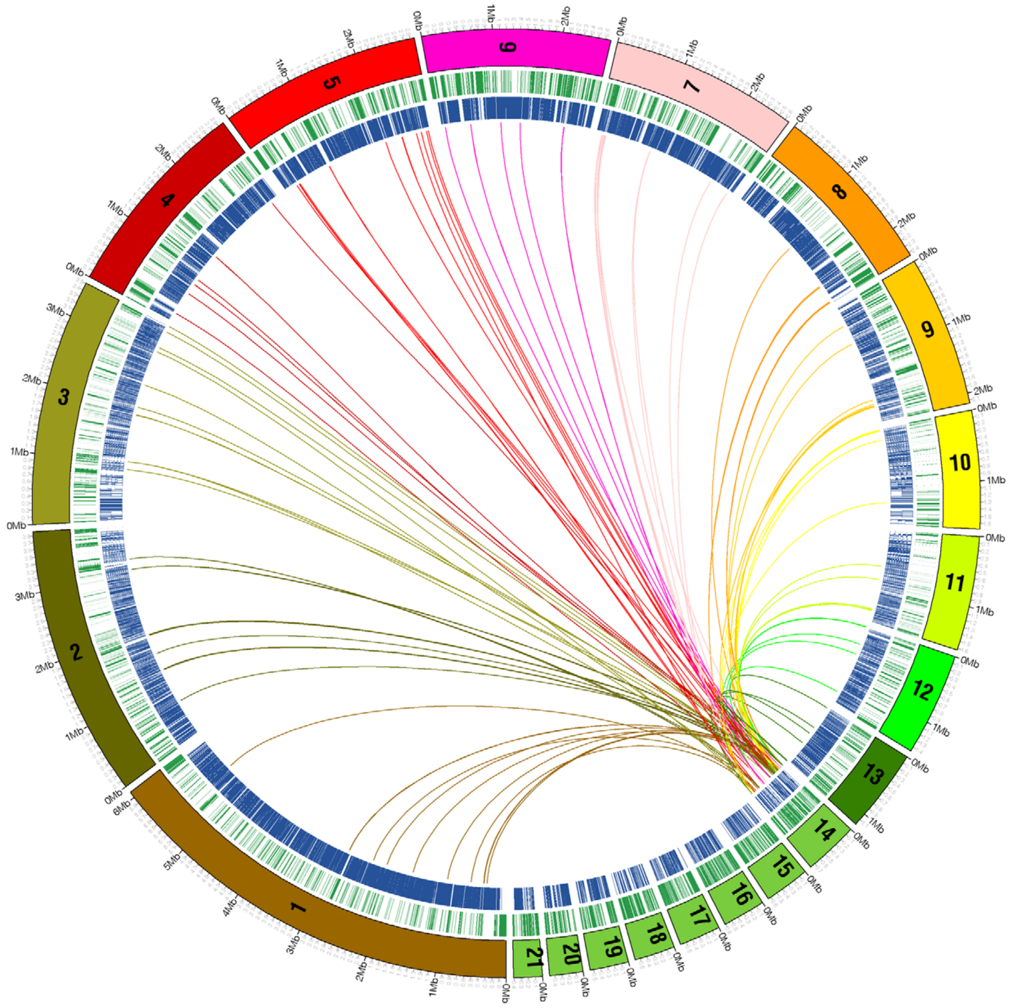


Figure 3. Analysis of genes that are shared between dispensable chromosome 14 and the 13 core chromosomes of *Mycosphaerella graminicola* isolate IPO323. Each chromosome is drawn to scale as a numbered bar around the outer edge of the circle, and the sequence was masked for repetitive DNA prior to analysis. Lines connect regions of 100 bp or larger that are similar between each core chromosome and the corresponding region on chromosome 14 at 1×e^{-5} or lower. Chromosome 14 is an amalgamation of genes from all of the core chromosomes but they are mixed together with no synteny. Genes on the other dispensable chromosomes were not included in this analysis.
doi:10.1371/journal.pgen.1002070.g003

Dispensable chromosomes have been reported in other fungi but they are significantly fewer and larger (from 0.7 to 4.9 Mb with an average of about 1.5 to 2.0 Mb) than those in *M. graminicola* (from 0.42 to 0.77 Mb) and mostly are composed of repetitive DNA with few known genes [30]. Unlike the dispensome of *M. graminicola*, the few genes on dispensable chromosomes in other fungi often are pathogenicity factors [31–33] and whole chromosomes may be transferred asexually [34]. Dispensable chromosomes in other fungi are different from the dispensome of *M. graminicola* except for the conditionally dispensable or lineage-specific chromosomes reported recently in *Nectria haematococca* (asexual stage: *Fusarium solani*) and other species of Fusarium [35–36], which also were different from core chromosomes in structure and gene content and contained numerous unique genes. However, unlike those in *M. graminicola*, dispensable chromosomes of Fusarium species had clear functions in ecological adaptation, were transferred more or less intact among closely related species [35] and did not show extensive recombination with core chromosomes.

The high instability of the *M. graminicola* dispensome during meiosis and mitosis would cause it to be eliminated unless it provided a selective advantage to the pathogen at least under some conditions. The unique genes with annotations indicated possible functions in transcription or signal transduction. There also was an enrichment for predicted pre-milRNAs, which may indicate that parts of the dispensome are involved in gene regulation. Based on dispensable chromosomes in other plant pathogens, genes on the dispensome were expected to be involved with host adaptation or

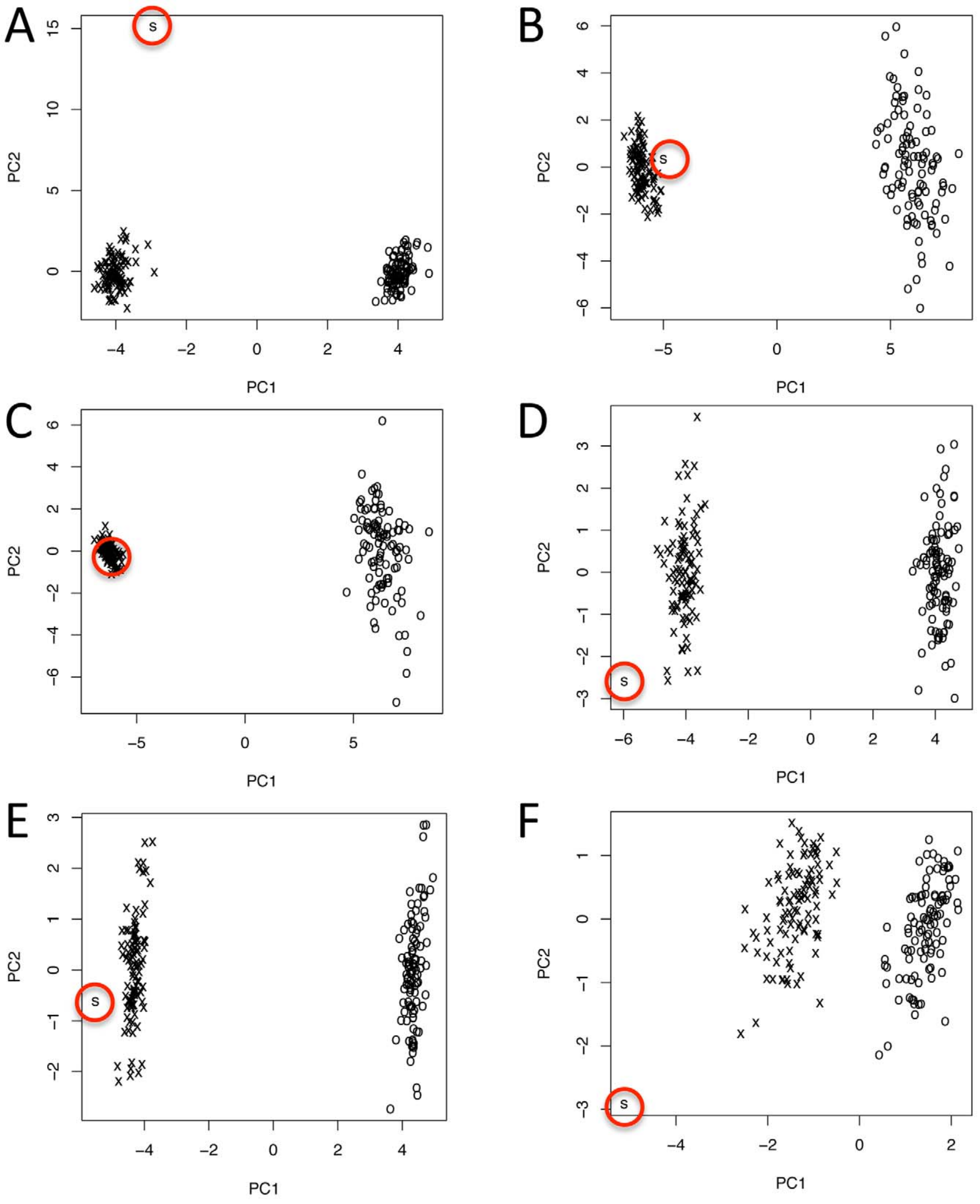


Figure 4. Principal Component Analysis of: S, observed genes on the dispensome; O, observed samples of genes on the core chromosomes before mutation; and x, samples of genes from the core chromosomes after mutation. Mutation was simulated using observed frequencies of all mutations in families of transposable elements with ten or more copies, and included mutations from RIP and other processes. Mutating the samples of genes from the core chromosomes always made them more similar to the observed value for the dispensome but only rarely included the dispensome value (see panel C). This occurred primarily with codon preference and GC content by amino acid, which are the quantities that are least subject to natural selection for protein function. A, amino acid frequency using the values for the aligned sequence with the highest GC content to build the table of mutation frequencies; B, codon preference using the consensus of the aligned sequences to make the table of mutation frequencies covering only the 5′ portion of each gene; C, codon preference using the values for the aligned sequence with the highest

GC content to build the table of mutation frequencies covering only the 5′ portion of each gene; D, codon usage using the values for the aligned sequence with the highest GC content to build the table of mutation frequencies but with all mutation frequencies cut in half; E, codon usage using the values for the aligned sequence with the highest GC content to build the table of mutation frequencies; and F, GC skew using the consensus of the aligned sequences to make the table of mutation frequencies. The first principal component always separated out the pre- and post-mutated chromosome samples. The locations of the observed values for the dispensome (S) are circled.
doi:10.1371/journal.pgen.1002070.g004

pathogenicity, yet so far no genes for pathogenicity or fitness of *M. graminicola* have been mapped to the dispensome [37]. A more interesting possibility is that the dispensome facilitates high recombination among chromosomes and could provide a repository of genes that may be advantageous under certain environmental conditions. This hypothesis should be tested by additional experimentation.

A recent comparison of the *M. graminicola* genome with that of its closest known relative, the unnamed species S1 from wild grasses in Iran, identified probable homologs for all of the dispensome chromosomes in the sibling species except for chromosome 18 [38]. These putative homologs presumably are dispensable also in species S1, but this has not been proven and only one isolate has been sequenced. Species S1 and *M. graminicola* are thought to have diverged approximately 10,500 years ago [39], concomitant with the domestication of wheat as a crop. Therefore, unlike dispensable chromosomes in other fungi, the dispensome of *M. graminicola* appears to be relatively ancient and has survived at least one speciation event. Analyses of two recently sequenced Dothideomycetes with Mycosphaerella sexual stages, *M. pini* (asexual stage: *Dothistroma septosporum*) and *M. populorum* (asexual stage: *Septoria musiva*), showed that they contained clear homologs of all of the core chromosomes of *M. graminicola*, but none of their chromosomes corresponded to the dispensome (B. Dhillon and S. B. Goodwin, unpublished). Taken together, these observations indicate that the dispensome of *M. graminicola* most likely was acquired prior to its divergence from a common ancestor with species S1 more than 10,000 years ago, but after the split of the *M.graminicola*-S1 lineage from that which gave rise to *M. pini* and *M. populorum*. The mechanism for the longevity of this dispensome with no obvious effects on fitness is not known.

More than half of the genes on the dispensome and almost all of the transposons also were present on core chromosomes. Moreover, there was no increase in gene numbers so a simple transfer of chromosomes from another species does not explain all of the observations. Instead, we propose a new model for the origin of dispensable chromosomes in *M. graminicola* by horizontal transfer followed by degeneration and extensive recombination with core chromosomes. The tight clustering of the dispensable chromosomes in the PCAs, with the possible exception of chromosome 14, indicates that they probably came from the same donor species. However, it is difficult to explain why they are so numerous. The most likely mechanism of horizontal transfer is via a sexual or somatic fusion with another species that had eight or more chromosomes, in which only a few genes were maintained on each donated chromosome. Chromosome segments that were redundant with the core set could be eliminated, leaving only those that are unique or that could confer some sort of selective advantage to the individual or to the dispensome. The fitness advantage could be transitory or occur only under certain conditions to allow those chromosomes to be dispensable, at least on an individual or population basis. Another possibility is that the numerous dispensable chromosomes are fragments from one or two larger chromosomes that were broken, acquired additional telomeres and lost content to result in their current, reduced complements of genes. High recombination within chromosomes and transfer of content between the donor and host chromosomes must have occurred to explain the observed pattern of shared genes.

The recombination hypothesis is supported by degenerated copies of some unique genes that were found on core chromosomes. These most likely represent genes that were copied from core to dispensable chromosomes, after which the copy on the core chromosome became inactivated, probably by RIP. Duplication, diversification and differential gene loss were proposed recently as the origin of lineage-specific gene islands in *Aspergillus fumigatus* [40], but that process seems to be very different from what occurred in *M. graminicola*. In *A. fumigatus*, large blocks of genes with synteny to other chromosomes were found, the opposite of what was seen for *M. graminicola*. The origin and evolution of the dispensome in *M. graminicola* seems to be very different from those reported for dispensable chromosomes in other fungi [35]. Unlike other fungi in which single chromosomes seem to have been transferred recently, the dispensome of *M. graminicola* most likely originated by ancient horizontal transfer of many chromosomes thousands of years ago. So far it is not known to be conditionally dispensable, unlike dispensable chromosomes in other fungi, which have clear roles in ecological adaptation.

The mesosyntenic analyses provided a new approach that complements the use of genetic linkage maps to support whole-genome assembly. Gene content was highly conserved on syntenic chromosomes in the two distantly related species, but there was little or no conservation of gene order or orientation. The comparison of the version 1 assembly of *M. graminicola* with the related *S. nodorum* genome sequence indicated scaffolds that should be merged and others that were erroneously assembled into one scaffold. Hence, mesosynteny validated the high-density genetic analyses and may provide an additional tool for whole-genome assembly for fungi where linkage maps do not exist or cannot be generated. Groups of genes in *S. nodorum* that corresponded to more than one group in *M. graminicola* may indicate scaffolds that should be joined in *S. nodorum* or, more likely, may reflect chromosomal rearrangements that have occurred since the divergence of *S. nodorum* and *M. graminicola* from an ancient common ancestor.

Considering their early divergence [41] relative to species within the same genus, the degree of mesosyntenic conservation between *M. graminicola* and *S. nodorum* is striking. However, it is very surprising that the synteny only applied to gene content but not order or orientation. In comparisons between other organisms, synteny plots usually yield diagonal lines even between unrelated species such as humans and cats [42]. The lack of diagonal lines in the comparisons of *S. nodorum* with *M. graminicola* indicate a high rate of shuffling of genes on chromosomal blocks that have remained constant over long periods of evolutionary time. The mechanism by which these small chromosomal rearrangements occur is not known.

The greatly reduced number of cell wall-degrading enzymes (CWDEs) in the genome of *M. graminicola* compared with other sequenced fungal genomes might be an evolutionary adaptation to avoid detection by the host during its extended, biotrophic latent phase and thus evade plant defenses long enough to cause disease. Similar loss of CWDEs in the ectomycorrhizal fungus *Laccaria bicolor* was thought to represent an adaptation to a symbiotic

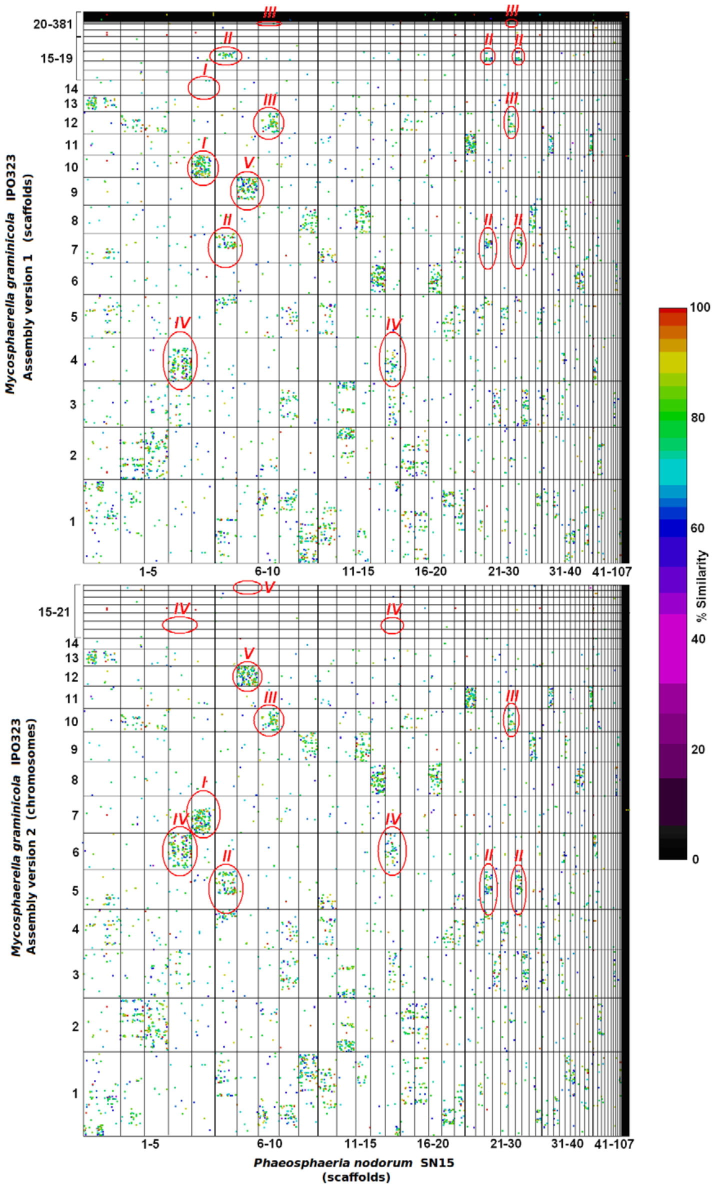


Figure 5. Comparisons of *Mycosphaerella graminicola* genome assembly versions 1 and 2 against that of *Stagonospora nodorum* isolate SN15. Scaffolds/chromosomes are ordered along their respective axes according to both decreasing length and increasing number. The 6-frame translations of both genomes were compared via MUMMER 3.0 [53]. Homologous regions are plotted as dots, which are color coded for percent similarity as per the bar on the right. Amendments made in the version 2 assembly and their corresponding regions in assembly version 1 are circled in red. Version 2 chromosomes 5 (B, circle II), 7 (B, circle I) and 10 (B, circle III) were derived from joined version 1 scaffolds 7 and 17 (A, circle II), 10 and 14 (A, circle I) and 12 and 22 (A, circle III), respectively, validating the method. Observation of the mesosyntenic pattern also could be used to identify inappropriately joined scaffolds. For example, *M. graminicola* v2 chromosomes 6 and 16 (B, circle IV) and 12 and 21 (B, circle V) were derived from split version 1 scaffolds 4 (A, circle IV) and 9 (A, circle V), respectively. These scaffolds are characterized by an abrupt termination of the mesosyntenic block at the split point as indicated by red lines (A, circles IV and V). A total of 21 predictions was made and 14 were validated.
doi:10.1371/journal.pgen.1002070.g005

lifestyle [43]. Based on these results we propose a novel, biphasic mechanism of stealth pathogenesis. During penetration and early colonization, *M. graminicola* produces a reduced set of proteins that facilitate pathogenicity and function as effectors in other fungi. Instead of the usual carbohydrate metabolism, nutrition during the extended biotrophic phase may be by degradation of proteins

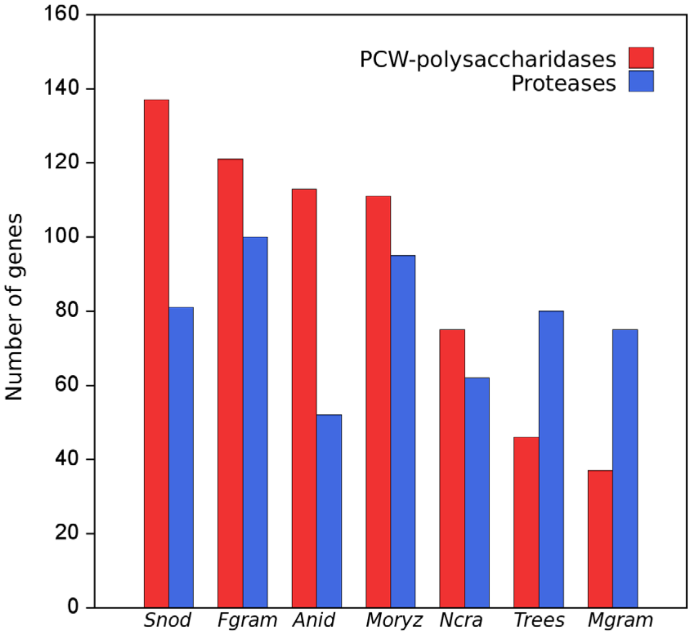


Figure 6. Numbers of genes for proteases and plant cell wall (PCW) degrading polysaccharidases in the genomes of seven fungi with sequenced genomes. Genes for PCW-polysaccharidases were severely reduced in the genome of *Mycosphaerella graminicola* but proteases were about the same. The overall profile of the enzymes in *M. graminicola* was most similar to that of *T. reesei* than to any of the other plant pathogens. Species analyzed included the saprophytes *Aspergillus nidulans* (Anid), *Neurospora crassa* (Ncra), and *Trichoderma reesei* (Trees), and the plant pathogens *Fusarium graminearum* (Fgram), *Mycosphaerella graminicola* (Mgram), *Magnaporthe oryzae* (Moryz), and *Stagonospora nodorum* (Snod).
doi:10.1371/journal.pgen.1002070.g006

rather than carbohydrates in the apoplastic fluid and intercellular spaces. The large number of proteases expressed during the early stages of the infection process supports this hypothesis. The biotrophic phase terminates by a switch to necrotrophic growth, production of specific cell wall-degrading enzymes and possibly by triggering programmed cell death [5,9–10].

Table 3. Numbers of predicted enzymes degrading cellulose across seven ascomycete species with sequenced genomes.

	Saprophytes[a]			Pathogens[a]			
CAZy family[b]	Anid	Ncra	Trees	Fgram	Mgram	Moryz	Snod
GH5 cellulases[c]	3	1	2	2	0	2	3
GH6	2	3	1	1	0	3	4
GH7	3	5	2	2	1	6	5
GH12	1	1	2	4	1	3	4
GH45	1	1	1	1	1	1	3
GH61	9	14	3	15	2	17	30
GH74	2	1	1	1	0	1	0
CBM1	8	19	15	12	0	22	13
Total cellulases	29	45	27	38	5	55	62

[a]Species analyzed included the saprophytes *Aspergillus nidulans* (Anid), *Neurospora crassa* (Ncra), and *Trichoderma reesii* (Trees), and the plant pathogens *Fusarium graminearum* (Fgram), *Mycosphaerella graminicola* (Mgram), *Magnaporthe oryzae* (Moryz), and *Stagonospora nodorum* (Snod).
[b]Families defined in the Carbohydrate-active enzymes database (www.cazy.org).
[c]GH5 is a family containing many different enzyme activities; only those targeting cellulose are included.
doi:10.1371/journal.pgen.1002070.t003

Stealth biotrophy raises the intriguing possibility that *M. graminicola* and possibly other Dothideomycetes may have evolved originally as endophytes or could be evolving towards an endophytic lifestyle. The finished genome of *M. graminicola* provides a gold standard [12] for this class of fungi, which is the largest and most ecologically diverse group of Ascomycetes with approximately 20,000 species, classified in 11 orders and 90 families, and provides a huge advantage for comparative genomics to identify the genetic basis of highly divergent lifestyles.

Materials and Methods

Biological material

Mycosphaerella graminicola isolates IPO323 and IPO94269 are Dutch field strains that were isolated in 1984 and 1994 from the wheat cultivar Arminda and an unknown cultivar, respectively. Isolate IPO95052 was isolated from a durum (tetraploid) wheat sample from Algeria. All isolates are maintained at the CBS-KNAW Fungal Biodiversity Centre of the Royal Netherlands Academy of Arts and Sciences (Utrecht, the Netherlands) under accession numbers CBS 115943 (IPO323), CBS 115941 (IPO94269) and CBS 115942 (IPO95052). Mycelia of each isolate were used to inoculate 200 mL of YG broth (10 g of yeast extract and 30 g of glucose per L) and were cultured until cloudy by shaking at 120 rpm at 18°C, after which the spores were lyophilized, 50 mg of lyophilised spores were placed in a 2-mL tube and ground with a Hybaid Ribolyser (model n° FP120HY-

230) for 10 s at 2500 rpm with a tungsten carbide bead. DNA was extracted using the Promega Wizard Magnetic DNA Purification system for food according to instructions provided by the manufacturer.

Initial sequencing and assembly

Whole-genome shotgun (WGS) sequencing of the genome of *M. graminicola* used three libraries with insert sizes of 2–3, 6–8, and 35–40 kb. The sequenced reads were screened for vector using cross_match, trimmed for vector and quality, and filtered to remove reads shorter than 100 bases. WGS assembly was done using Jazz, a tool developed at the JGI [44]. After excluding redundant and short scaffolds, the assembly v1.0 contained 41.2 Mb of sequence in 129 scaffolds, of which 4.0 Mb (7.5%) was in gaps (Table S8). The sequence depth derived from the assembly was 8.88±0.04.

Gap closure and finishing

To perform finishing, the *M. graminicola* WGS assembly was broken down into scaffold-size pieces and each piece was reassembled with phrap. These scaffold pieces were then finished using a Phred/Phrap/Consed pipeline. Initially, all low-quality regions and gaps were targeted with computationally selected sequencing reactions completed with 4:1 BigDye terminator: dGTP chemistry (Applied Biosystems). These automated rounds included resequencing plasmid subclones and walking on plasmid subclones or fosmids using custom primers.

Following completion of the automated rounds, a trained finisher manually inspected each assembly. Further reactions were then manually selected to complete the genome. These included additional resequencing reactions and custom primer walks on plasmid subclones or fosmids as described above guided by a genetic map of more than 2,031 sequenced markers plus paired-end reads from a library of Bacterial Artificial Chromosome clones. Smaller repeats in the sequence were resolved by transposon-hopping 8-kb plasmid clones. Fosmid and BAC clones were shotgun sequenced and finished to fill large gaps, resolve larger repeats and to extend into the telomere regions.

Each assembly was then validated by an independent quality assessment. This included a visual examination of subclone paired ends using Orchid (http://www-hagsc.org), and visual inspection of high-quality discrepancies and all remaining low-quality areas. All available EST resources were also placed on the assembly to ensure completeness. The finished genome consists of 39,686,251 base pairs of finished sequence with an estimated error rate of less than 1 in 100,000 base pairs. Genome contiguity is very high with a total of 21 chromosomes represented, 19 of which are complete and 20 of which are sequenced from telomere to telomere.

Genome annotation

Both draft (v1.0) and finished (v2.0) assemblies of *M. graminicola* were processed using the JGI annotation pipeline, which combines several gene predictors:1) putative full-length genes from EST cluster consensus sequences; 2) homology-based gene models were predicted using FGENESH+ [45] and Genewise [46] seeded by Blastx alignments against sequences from the NCBI non-redundant protein set; 3) *ab initio* gene predictor FGENESH [45] was trained on the set of putative full-length genes and reliable homology-based models. Genewise models were completed using scaffold data to find start and stop codons. ESTs were used to extend, verify, and complete the predicted gene models. Because multiple gene models per locus were often generated, a single representative gene model for each locus was chosen based on homology and EST support and used for further analysis. Those comprised a filtered set of gene models supported by different lines of evidence. These were further curated manually during community annotation and used for analysis.

All predicted gene models were annotated using InterProScan [47] and hardware-accelerated double-affine Smith-Waterman alignments (www.timelogic.com) against the SwissProt (www.expasy.org/sprot) and other specialized databases such as KEGG [48]. Finally, KEGG hits were used to map EC numbers (http://www.expasy.org/enzyme/), and Interpro hits were used to map GO terms [49]. Predicted proteins also were annotated according to KOG [50,51] classification.

Following the machine annotation, manual validation and correction of selected gene sets was performed by more than 30 annotators through a jamboree held at the JGI facilities in Walnut Creek, California, USA. Annotators were trained by JGI staff and continue to make modifications as necessary.

Potential microRNA-like small RNA (milRNAs) loci were annotated using the INFERNAL software tool and based on 454 microRNA families (covarion models) from the RFAM database version 9.1 [52]. milRNAs were predicted if their scores were higher than thresholds, defined individually for each family, in the same way as PFAM domains are predicted.

Experimental validation of the predicted milRNAs was done by sequencing of an RNA library Total RNA was isolated from spores germinated on water agar of *M. graminicola* isolate IPO323. A small RNA library was prepared according to the protocol for Illumina sequencing; small RNAs from 16–~50 nt were isolated from gels, sequenced with an Illumina/Solexa single read DNA 50 cycles Genome Analyzer II, and compared by BLAST search against the list of 535 predicted pre-milRNAs from the genome sequence.

Assembly and annotations of the *M. graminicola* finished genome are available from the JGI Genome Portal at http://www.jgi.doe.gov/Mgraminicola and were deposited at DDBJ/EMBL/GenBank under the project accession ACPE00000000.

Microarray analyses

Whole-genome tiling microarrays were designed by choosing one 50-mer primer every 100 bases spanning the entire finished genome. The arrays were manufactured and hybridized by the Nimblegen Corporation with total DNA extracted from each field isolate.

Principal component analyses of core and dispensable chromosomes

The CodonW package (http://codonw.sourceforge.net/) was used for correspondence analysis of codon usage, which mathematically is identical to principal component analysis. CodonW requires as an input a set of coding sequences, usually of individual genes. For chromosome-level analyses coding sequences from the frozen gene catalog models for each chromosomes were concatenated, forming 21 'superORFs', one for each chromosome. Because partial models may introduce some potential frameshifts with internal stop codons they were removed from the analysis; this did not affect the results as their total number is low. CodonW has no graphical outputs, so they were used as inputs for scatter plots in R (http://www.r-project.org/).

For *M. graminicola* only a similar analysis was done for repeats. RepeatScout was run on the genome to produce a set of *ab initio*-identified repeat sequences. From that set 81 distinct repeat sequences, each with an occurrence exceeding 20 times in the genome, were extracted. For each chromosome a vector of length 81 was calculated with the relative frequency of each repeat. A PC analysis was run on the resulting vectors using the standard principal component function pcomp in R. Separation at the repeat level means that these chromosomes have distinct evolutionary profiles not only on the protein-coding level, but also on other parts of the chromosomes, suggesting that entire chromosomes may be transferred horizontally.

Mesosynteny

Dot plots were generated via MUMMER 3.0 [53] with data derived from default PROmer comparisons between the *M. graminicola* genome assembly versions 1 and 2 (http://genome.jgi-psf.org/Mycgr3/Mycgr3.home.html) and *S. nodorum* SN15 assembly version 2 [54], available under GenBank accessions CH445325–CH445384, CH445386–CH445394 and CH959328–CH959365, or AAGI00000000. Additional comparisons and statistical analyses were made with custom-designed perl scripts.

Data from the *M. graminicola* version 2 comparison with *S. nodorum* were used to test the efficacy of mesosyntenic comparisons to assist the completion of fungal genomes. The mesosynteny-based prediction of scaffold joining involved 3 stages: determining the percent coverage of scaffolds/chromosomes for each scaffold/chromosome pair (i.e., a function of the number of 'dots' per 'block'); determining which scaffold/chromosome pairs were significantly related and forming groups of joined scaffolds; and filtering out background levels of similarity due to sequence redundancy and incomplete genome assemblies.

Coordinates of homologous regions were obtained from PROmer coordinate outputs (MUMMER 3.0) and used to determine the percent of sequence covered by matches to a sequence from the alternate genome for each sequence pair. Where match coordinates overlapped on the sequence of interest, those matches were merged into a single feature to avoid redundancy. A perl script for conversion of PROmer coordinate outputs to a table of percent coverage is available on request.

Coverage values for each *M. graminicola*-*S. nodorum* sequence pair were subject to a binomial test for significance. The threshold for significance (*Psig*)≥0.95 was:

$$P_{sig} = F(x,p,n) = \sum_{i=0}^{x} \binom{n}{i} (p)^{i} (1-p)^{n-1}$$

where x is the percent coverage, n equals 100, and p is the probability of chromosome homology.

The probability of chromosome homology (p) was equal to 1/(21×19), which was derived from the number of *M. graminicola* chromosomes (21) and the approximate PFGE estimate of *S. nodorum* chromosomes (19) [55]. This is the likelihood that a given sequence pair represents related chromosomes. This model assumes that no whole-genome/chromosome duplication events have occurred previously between either fungal genome since divergence from their last common ancestor.

The significance of percent coverage (*Psig*) was tested bidirectionally for each sequence pair (i.e., for sequence pair A–B, both coverage of sequence A by B and coverage of sequence B by A were tested). Sequence pairs were significantly related if a test in either direction was successful. A minimum length threshold of 1 kb was also imposed for both sequences. Where multiple scaffolds of *M. graminicola* were significantly related to the same *S. nodorum* scaffold, those *M. graminicola* scaffolds formed a 'joined group' of candidates for representation of the same chromosome.

All possible paired combinations of *M. graminicola* scaffolds present within predicted joined groups were subject to filtering for high levels of background similarity as follows:

$$retention = \frac{\#\text{joined groups (scaffolds joined)}}{\#\text{joined groups (either scaffold present)}}$$

The retention score is a measure of the reliability of scaffold join relationships. Joins between *M. graminicola* scaffold pairs with retention scores <0.25 were discarded.

CAZy annotation and growth profiling

Annotation of carbohydrate-related enzymes was performed using the Carbohydrate-Active Enzyme database (CAZy) annotation pipeline [26]. BLAST was used to compare the predicted proteins of *M. graminicola* to a collection of catalytic and carbohydrate-binding modules derived from CAZy. Significant hits were compared individually by BLAST to assign them to one or more CAZy families. Ambiguous family attributions were processed manually along with all identified models that presented defects (deletions, insertions, splicing issues, etc.).

Growth profiling of *S. nodorum* and *M. graminicola* was on *Aspergillus niger* minimal medium [56]. Cultures were grown at 25 degrees for seven days after which pictures were taken for growth comparison. Carbon sources used were: glucose (Sigma); soluble starch (Difco); alpha-cellulose (Sigma); Guar Gum (Sigma, galactomannan); Oat spelt xylan (Sigma); and Apple Pectin (Sigma).

Genome structure analyses

Comparisons of sequence content between core and dispensable chromosomes was with Circos [57]. This tool draws ribbons connecting sequences that align in different data sets.

Supporting Information

Dataset S1 The method and calculations for using mesosynteny to predict scaffold joins from version 1 to version 2 of the *Mycosphaerella graminicola* genomic sequence.
(XLS)

Figure S1 Aspects of the *in vitro* and *in vivo* lifestyle of *Mycosphaerella graminicola*. 1. Typical colony appearance of *M. graminicola* isolates grown under light (upper two rows) and dark (lower low) conditions. Light stimulates yeast-like growth whereas darkness induces filamentous growth. 2. Close-up of yeast-like growth on V8 agar. 3. *In vitro* production of asexual fructifications (pycnidia; arrow) on wheat leaf extract agar. 4. Penetration of a wheat leaf stoma (arrow) by a pycnidiospore germ tube. 5. Simultaneous penetration of a wheat leaf stoma by three germ tubes of sexual airborne ascospores (arrows) that are transported over vast distances. 6. Colonization of the mesophyll tissue by an intercellular hypha (arrows) during the symptomless biotrophic phase of pathogenesis. 7. Initiation (arrow head) of a pycnidium in the substomatal cavity of a wheat leaf. 8. Ripe pycnidia in a primary leaf of a susceptible wheat seedling. High humidity stimulates the extrusion of cyrrhi, tendril-like mucilages containing asexual pycnidiospores that are rain-splash dispersed over short distances. 9. Typical infection of the primary leaf of a resistant cultivar. Note the low fungal density in the apoplast (arrow) and the response of the mesophyll cells (arrow head), particularly the chloroplasts, to the presence of intercellular hyphae. 10. Typical symptoms on a primary seedling leaf of a highly susceptible wheat cultivar. 11. Typical response on a primary leaf of a highly resistant wheat cultivar. 12. Adult-plant evaluation plots are inoculated at the adult plant stage with individual isolates using air-driven equipment. 13. Symptoms on an adult plant flag leaf

after field inoculations. 14. Symptoms on a naturally infected adult plant flag leaf.
(TIF)

Figure S2 The 21 chromosomes of the *Mycosphaerella graminicola* genome drawn to scale. Red indicates regions of single-copy sequence; repetitive sequences are in shown blue. Chromosome 1 is almost twice as long as any of the others. The core chromosomes 1–13 are the largest. Dispensable chromosomes 14–21 are smaller than the core chromosomes and have a higher proportion of repetitive DNA as indicated by the blue bands.
(TIF)

Figure S3 Features of chromosome 14, the largest dispensable chromosome of *Mycosphaerella graminicola*, and alignment to genetic linkage maps. A, Plot of GC content. Areas of low GC usually correspond to regions of repetitive DNA. B, Repetitive regions of the *M. graminicola* genome. C, Single-copy (red) regions of the *M. graminicola* genome. D, Locations of genes for proteins containing signal peptides. E, Locations of homologs of pathogenicity or virulence genes that have been experimentally verified in species pathogenic to plant, animal or human hosts. F, Approximate locations of quantitative trait loci (QTL) for pathogenicity to wheat. G, Alignments between the genomic sequence and two genetic linkage maps of crosses involving isolate IPO323. Top half, Genetic linkage map of the cross between IPO323 and the Algerian durum wheat isolate IPO95052. Bottom half, Genetic linkage map of the cross between bread wheat isolates IPO323 and IPO94269. The physical map represented by the genomic sequence is in the center. Lines connect mapped genetic markers in each linkage map to their corresponding locations on the physical map based on the sequences of the marker loci. Very few secreted proteins (track D) or pathogenicity-related genes (E) and no pathogenicity QTL mapped to the dispensome.
(TIF)

Figure S4 Principal Component Analysis of codon usage in 21 chromosomes of the *Mycosphaerella graminicola* finished genome. Factor 1 gave good discrimination between core (blue circles) and dispensable (red) chromosomes.
(TIF)

Figure S5 Examples of unique genes on dispensable chromosomes with an inactivated copy on a core chromosome. A unique gene on chromosome 14 and two on chromosome 18 showed excellent alignments to footprints of genes on chromosome 1. The copies on chromosome 1 matched those on the dispensable chromosomes with an expected value of 1×10^{-5} or better, but contained numerous stop codons indicating that they were pseudogenes and possibly could have been the progenitor copies for the intact, unique genes on dispensable chromosomes 14 and 18. The graphs above chromosome 14 and below chromosome 18 indicate GC content.
(TIF)

Figure S6 Examples of amino acid alignments between protein sequences of unique genes on dispensable chromosomes to their inactivated putative homologs on core chromosomes. A, A unique gene on dispensable chromosome 14 aligned to a footprint of its homologous pseudogene on core chromosome 1. B and C, Alignments between two genes on dispensable chromosome 18 to homologous pseudogenes on core chromosome 1. Identical amino acids are shaded blue. Stop codons in pseudogenes are indicated by X and are shaded red. Details are provided beneath each alignment. Each unique gene is at least 26% identical and 46% similar to its putative homolog.
(TIF)

Figure S7 Analysis of genes and repetitive DNAs that are shared between dispensable chromosome 14 and the 13 core chromosomes of *Mycosphaerella graminicola*. Each chromosome is drawn to scale as a numbered bar around the outer edge of the circle. Lines connect regions of 100 bp or larger that are similar between each core chromosome and the corresponding region on chromosome 14 at $1\times e^{-5}$ or lower. Chromosome 14 contains parts of all of the core chromosomes that are mixed in together with no synteny. Genes on the other dispensable chromosomes were not included in this analysis.
(TIF)

Figure S8 Analysis of genes that are shared between each of the nine largest core chromosomes (1–9) and all other chromosomes of the *Mycosphaerella graminicola* genome. Each chromosome is drawn to scale as a numbered bar around the outer edge of the circle. Lines connect regions of 100 bp or larger that are similar between the indicated core chromosome and each of the remaining 20 chromosomes at $1\times e^{-5}$ or lower. Each chromosome contains parts of all of the other chromosomes mixed in together with no synteny. Genes on the 12 smallest chromosomes were similar but are not shown.
(TIF)

Figure S9 Principal Component Analysis of codon usage. A, in 21 chromosomes of the *Mycosphaerella graminicola* finished genome after simulated RIPping. B, in 21 chromosomes of the *M. graminicola* finished genome after simulated deRIPping. C, of about 150 genes with shared putative homologs between the core and dispensable chromosomes of *M. graminicola*. D, of amino acid composition of about 150 genes with shared putative homologs between the core and dispensable chromosomes of *M. graminicola*. E, of all genes with shared putative homologs between the core and dispensable chromosomes of *M. graminicola*. F, of all genes on dispensable chromosomes with shared putative homologs on core chromosomes against all genes on the core chromosomes of *M. graminicola*.
(TIF)

Figure S10 Principal Component Analysis of codon usage. A,between the genomes of *M. graminicola* and *Stagonospora nodorum*. B, between the genomes of *M. graminicola* and *Aspergillus fumigatus*. Values for the chromosomes of *M. graminicola* are indicated by red circles, those for *S. nodorum* and *A. fumigatus* by green triangles.
(TIF)

Figure S11 Principal Component Analysis of repeats in 21 chromosomes of the *Mycosphaerella graminicola* finished genome. Core chromosomes (black circles) were clearly separated from the dispensome (red).
(TIF)

Figure S12 Growth of *Mycosphaerella graminicola*, *Stagonospora nodorum* and *Magnaporthe oryzae* (*M. grisea*) on glucose and several plant polysaccharides. Growth of *M. graminicola* was decreased on xylan, consistent with the CAZy annotation for fewer genes involved in degradation of that substrate.
(TIF)

Figure S13 Venn diagrams showing the expression of *Mycosphaerella graminicola* genes at different times during the infection process and with a sample grown *in vitro*. A, Libraries MgEST_08, MgEST_09, and MgEST_10 contain EST sequences from wheat leaf tissue collected at 5, 10 and 16 days after inoculation, respectively. B, four-way diagram with the same three in vitro-produced libraries plus *in vitro* library MgEST_05, grown on minimal medium minus nitrogen to mimic the early stages of the infection process.
(TIF)

Table S1 List of functional domains or other annotations for 65 genes on dispensable chromosomes 14–21 of the genome of *Mycosphaerella graminicola*.
(DOCX)

Table S2 Analysis of small RNA sequences (generated on the Illumina platform) for the presence of computationally predicted pre-microRNA-like (milRNA) sequences in germinated spores of *Mycosphaerella graminicola* isolate IPO323.
(DOCX)

Table S3 Best non-self BLAST hits for 654 called genes on dispensable chromosomes of *Mycosphaerella graminicola* queried with *tblastn* against a combined database containing the GenBank nt and EST datasets plus *M. graminicola* version 2.0 and *M. fijiensis* v1.0 from the Joint Genome Institute.
(DOCX)

Table S4 Numbers of predicted enzymes degrading hemicellulose, pectin and cutin across seven ascomycete species with sequenced genomes.
(DOCX)

Table S5 Total numbers of predicted CAZymes in *Mycosphaerella graminicola* and selected ascomycetes.
(DOCX)

Table S6 PFAM domains that are expanded in the genome of *Mycosphaerella graminicola* relative to those of five other Ascomycetes[a] and two plant-pathogenic Stramenopiles[b].
(DOCX)

Table S7 PFAM domains that are expanded in the genome of *Mycosphaerella graminicola* relative to those of five other Ascomycetes[a] but not the two plant-pathogenic Stramenopiles[b].
(DOCX)

Table S8 Assembly statistics for the *Mycosphaerella graminicola* version 1 (8.9× draft) and version 2 (finished) sequences compared to the 10× draft sequence of *Stagonospora nodorum*.
(DOCX)

Acknowledgments

We thank Els C. P. Verstappen, Jessica R. Cavaletto, Ineke Ph. De Vries, and Odette Mendes for excellent technical support.

Author Contributions

Conceived and designed the experiments: SB Goodwin, IV Grigoriev, GHJ Gema. Performed the experiments: S Ben M'Barek, AHJ Wittenberg, TAJ Van der Lee, PM Coutinho, B Henrissat, V Lombard, SB Ware, C Waalwijk. Analyzed the data: SB Goodwin, S Ben M'Barek, B Dhillon, AHJ Wittenberg, CF Crane, JK Hane, AJ Foster, J Grimwood, J Antoniw, A Bailey, B Bluhm, J Bowler, A Burgt, B Canto-Canché, ACL Churchill, L Conde-Ferràez, HJ Cools, M Csukai, P Dehal, P De Wit, B Donzelli, HC van de Geest, KE Hammond-Kosack, RCHJ van Ham, B Henrissat, A Kilian, AK Kobayashi, E Koopmann, Y Kourmpetis, A Kuzniar, E Lindquist, C Maliepaard, N Martins, R Mehrabi, JPH Nap, A Ponomarenko, JJ Rudd, A Salamov, J Schmutz, HJ Schouten, I Stergiopoulos, SFF Torriani, RP de Vries, A Wiebenga, L-H Zwiers, RP Oliver, IV Grigoriev, GHJ Gema. Contributed reagents/materials/analysis tools: CF Crane, JK Hane, A Aerts, E Lindquist, H Shapiro, H Tu. Wrote the paper: SB Goodwin, GHJ Gema. Managed the Community Sequencing Program: J Bristow

References

1. Eyal Z, Schare AL, Prescott JM, van Ginkel M (1987) The Septoria Diseases of Wheat: Concepts and Methods of Disease Management. Mexico, DF: CIMMYT.
2. Hardwick NV, Jones DR, Slough JE (2001) Factors affecting diseases of winter wheat in England and Wales, 1989–98. Plant Pathol 50: 453–462.
3. McDougall P (2006) Phillips McDougall Agriservice Report. Scotland, UK: Pathhead, Midlothian.
4. Linde CC, Zhan J, McDonald BA (2002) Population structure of *Mycosphaerella graminicola*: From lesions to continents. Phytopathology 92: 946–955.
5. Kema GHJ, Yu D, Rijkenberg FHJ, Shaw MW, Baayen RP (1996) Histology of the pathogenesis of *Mycosphaerella graminicola* in wheat. Phytopathology 86: 777–786.
6. Duncan KE, Howard RJ (2000) Cytological analysis of wheat infection by the leaf blotch pathogen *Mycosphaerella graminicola*. Mycol Res 104: 1074–1082.
7. Jing H-C, Lovell D, Gutteridge R, Jenk D, Kornyukhin D, et al. (2008) Phenotypic and genetic analysis of the *Triticum monococcum–Mycosphaerella graminicola* interaction. New Phytol 179: 1121–1132.
8. Adhikari TB, Balaji B, Breeden J, Goodwin SB (2007) Resistance of wheat to *Mycosphaerella graminicola* involves early and late peaks of gene expression. Physiol Mol Plant Pathol 71: 55–68.
9. Kema GHJ, van der Lee TAJ, Mendes O, Verstappen ECP, Lankhorst RK, et al. (2008) Large-scale gene discovery in the septoria tritici blotch fungus *Mycosphaerella graminicola* with a focus on in planta expression. Mol Plant-Microbe Interact 21: 1249–1260.
10. Keon J, Antoniw J, Carzaniga R, Deller S, Ward JL, et al. (2007) Transcriptional adaptation of *Mycosphaerella graminicola* to programmed cell death (PCD) of its susceptible wheat host. Mol Plant-Microbe Interact 20: 178–193.
11. Wittenberg AHJ, van der Lee TAJ, Ben M'Barek S, Ware SB, Goodwin SB, et al. (2009) Meiosis drives extraordinary genome plasticity in the haploid fungal plant pathogen *Mycosphaerella graminicola*. PLoS ONE 4: e5863. doi:10.1371/journal.pone.0005863.
12. Chain PS, Grafham DV, Fulton RS, FitzGerald MG, Hostetler J, et al. (2009) Genome project standards in a new era of sequencing. Science 326: 236–237.
13. Nierman WC, Pain A, Anderson MJ, Wortman JR, Kim HS, et al. (2005) Genomic sequence of the pathogenic and allergenic filamentous fungus *Aspergillus fumigatus*. Nature 438: 1151–1156.
14. Torriani SFF, Goodwin SB, Kema GHJ, Pangilinan JL, McDonald BA (2008) Intraspecific comparison and annotation of two complete mitochondrial genome sequences from the plant pathogenic fungus *Mycosphaerella graminicola*. Fungal Genet Biol 45: 628–637.
15. Selker EU (2002) Repeat-induced gene silencing in fungi. Adv Genet 46: 439–450.
16. Cambareri EB, Jensen BC, Schabtach E, Selker EU (1989) Repeat-induced G-C to A-T mutations in *Neurospora*. Science 244: 1571–1575.
17. Dhillon B, Cavaletto JR, Wood KV, Goodwin SB (2010) Accidental amplification and inactivation of a methyltransferase gene eliminates cytosine methylation in *Mycosphaerella graminicola*. Genetics 186: 67–77.
18. Gardner PP, Daub J, Tate JG, Nawrocki EP, Kolbe DL, et al. (2008) Rfam: updates to the RNA families database. Nucleic Acids Res 37: D136–D140.
19. Lee H-C, Li L, Gu W, Xue Z, Crosthwaite SK, et al. (2010) Diverse pathways generate microRNA-like RNAs and dicer-independent small interfering RNAs in fungi. Molecular Cell 38: 803–814.
20. Fedorova ND, Khaldi N, Joardar VS, Maiti R, Amedeo P, et al. (2008) Genomic islands in the pathogenic filamentous fungus *Aspergillus fumigatus*. PLoS Genet 4: e1000046. doi:10.1371/journal.pgen.1000046.
21. Ma L-J, van der Does HC, Borkovich KA, Coleman JJ, Daboussi M-J, et al. (2010) Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature 464: 367–373.
22. Kay S, Hahn S, Marois E, Hause G, Bonas U (2007) A bacterial effector acts as a plant transcription factor and induces a cell size regulator. Science 318: 648–651.
23. Giles NH, Geever RF, Asch DK, Avalos J, Case ME (1991) Organization and regulation of the qa (quinic acid) genes in *Neurospora crassa* and other fungi. J Hered 82: 1–7.
24. Martin F, Kohler A, Murat C, Balestrini R, Coutinho PM, et al. (2010) Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis. Nature 464: 1033–1038.
25. Yun SH, Arie T, Kaneko I, Yoder OC, Turgeon BG (2000) Molecular organization of mating type loci in heterothallic, homothallic, and asexual Gibberella/Fusarium species. Fungal Genet Biol 31: 7–20.
26. Dean RA, Talbot NJ, Ebbole DJ, Farman ML, Mitchell TK, et al. (2005) The genome sequence of the rice blast fungus *Magnaporthe grisea*. Nature 434: 980–986.
27. Cuomo CA, Güldener U, Xu J-R, Trail F, Turgeon BG, et al. (2007) The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. Science 317: 1400–1402.
28. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, et al. (2009) The Carbohydrate-Active enZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res 37: D233–D238.
29. Caracuel-Rios Z, Talbot NJ (2007) Cellular differentiation and host invasion by the rice blast fungus *Magnaporthe grisea*. Curr Opin Microbiol 10: 339–345.

30. Jones RN, Viegas W, Houben A (2008) A century of B chromosomes in plants: so what?. Ann Bot 101: 767–775.
31. Jones N, Houben A (2003) B chromosomes in plants: escapees from the A chromosome genome? Trends Plant Sci 8: 417–423.
32. Covert SF (1998) Supernumerary chromosomes in filamentous fungi. Curr Genet 33: 311–319.
33. Miao VP, Covert SF, VanEtten HD (1991) A fungal gene for antibiotic resistance on a dispensable ("B") chromosome. Science 254: 1773–1776.
34. Hatta R, Ito K, Hosaki Y, Tanaka T, Tanaka A, et al. (2002) A conditionally dispensable chromosome controls host-specific pathogenicity in the fungal plant pathogen *Alternaria alternata*. Genetics 161: 59–70.
35. Wang C, Skrobek A, Butt TM (2003) Concurrence of losing a chromosome and the ability to produce destruxins in a mutant of *Metarhizium anisopliae*. FEMS Microbiol Lett 226: 373–378.
36. Masel AM, He CZ, Poplawski AM, Irwin JAG, Manners JM (1996) Molecular evidence for chromosome transfer between biotypes of *Colletotrichum gloeosporioides*. Mol Plant-Microbe Interact 9: 339–348.
37. Coleman JJ, Rounsley SD, Rodriguez-Carres M, Kuo A, Wasmann CC, et al. (2009) The genome of *Nectria haematococca*: Contribution of supernumerary chromosomes to gene expansion. PLoS Genet 5: e1000618. doi:10.1371/journal.pgen.1000618.
38. Ware SB (2006) Aspects of sexual reproduction in Mycosphaerella species on wheat and barley: genetic studies on specificity, mapping, and fungicide resistance. Wageningen University, The Netherlands: Ph.D. thesis.
39. Stukenbrock EH, Jørgensen FG, Zala M, Hansen TT, McDonald BA, et al. (2010) Whole-genome and chromosome evolution associated with host adaptation and speciation of the wheat pathogen *Mycosphaerella graminicola*. PLoS Genet 6: e1001189. doi:10.1371/journal.pgen.1001189.
40. Stukenbrock EH, Banke S, Javan-Nikkhah M, McDonald BA (2007) Origin and domestication of the fungal wheat pathogen *Mycosphaerella graminicola* via sympatric speciation. Mol Biol Evol 24: 398–411.
41. James T, Kauff F, Schoch CL, Matheny PB, Hofstetter V, et al. (2006) Reconstructing the early evolution of fungi using a six-gene phylogeny. Nature 443: 818–822.
42. Housworth EA, Postlethwait J (2002) Measures of synteny conservation between species pairs. Genetics 162: 441–448.
43. Martin F, Aerts A, Ahrén D, Brun A, Danchin EGJ, et al. (2008) The genome of *Laccaria bicolor* provides insights into mycorrhizal symbiosis. Nature 452: 88–92.
44. Aparicio S, Chapman J, Stupka E, Putnam N, Chia J-M, et al. (2002) Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. Science 297: 1301–1310.
45. Salamov A, Solovyev VV (2000) Ab initio gene finding in *Drosophila* genomic DNA. Genome Res 10: 516–522.
46. Birney E, Durbin R (2000) Using GeneWise in the *Drosophila* annotation experiment. Genome Res 10: 547–548.
47. Zdobnov EM, Apweiler R (2001) InterProScan – an integration platform for the signature-recognition methods in InterPro. Bioinformatics 17: 847–848.
48. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32: D277–D280.
49. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, et al. (2000) Gene Ontology: tool for the unification of biology. Nat Genet 25: 25–29.
50. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, et al. (2003) The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4: 41.
51. Koonin EV, Fedorova ND, Jackson JD, Jacobs AR, Krylov DM, et al. (2004) A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. Genome Biol 5: R7.
52. Griffiths-Jones S, Moxon S, Marshall M, Khanna A, Eddy SR, et al. (2005) Rfam: annotating non-coding RNAs in complete genomes. Nucleic Acids Res 33: D121–D124.
53. Kurtz S, Phillippy A, Delcher AL, Smoot M, Shumway M, et al. (2004) Versatile and open software for comparing large genomes. Genome Biology 5: R12.
54. Hane JK, Lowe RGT, Solomon PS, Tan K-C, Schoch CL, et al. (2007) Dothideomycete–plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*. The Plant Cell 19: 3347–3368.
55. Cooley RN, Caten CE (1991) Variation in electrophoretic karyotype between strains of *Septoria nodorum*. Mol Gen Genet 228: 17–23.
56. de Vries RP, Frisvad JC, van de Vondervoort PJI, Burgers K, Kuijpers AFA, et al. (2005) *Aspergillus vadensis*, a new species of the group of black Aspergilli. Antonie Van Leeuwenhoek 87: 195–203.
57. Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. (2009) Circos: An information aesthetic for comparative genomics. Genome Res 19: 1639–1645.

Appendix 9A: Response to Thesis Examination Comments

If I understand the mesosynteny-based genome finishing approach correctly it is based on the principle that scaffolds present on a single chromosome in one species will tend to be retained on the "homologous" chromosome in another species. This allows one to narrow down the list of possible neighbouring scaffolds, in turn making PCR based gap-finishing feasible. However, looking at the MUMMER plots, it would appear that only about 50% of genes, on average, are retained on the homologous chromosome. Does this mean that mesosynteny-based finishing only works about 50% of the time. It would help to have some clarification on this issue in one of the relevant chapters.

For most whole-genome comparisons between two species, it is reasonable to expect that for a given pair of "homologous" chromosomes they will contain regions of repetitive sequences that are also conserved with "non-homologous" chromosomes. It follows that to accurately identify synteny between two species, only regions that match uniquely between the two species should be considered.

MUMMER plots were generated using the "maximum unique match" parameter in the promer algorithm, followed by the "filter" parameter in mummerplot program (which draws the plots) (http://mummer.sourceforge.net/manual/). The effect of these two parameters is that "non-unique" sequence alignments have been excluded from the dot-plots and subsequent mesosynteny calculations. It is for this reason that a considerable portion of two genomes that may otherwise be homologous, will not be reported as matching.

Chapter 10: Attribution Statement

Title: **Genomic and comparative analysis of the class Dothideomycetes.**

Authors: **James K. Hane**, Angela H. Williams and Richard P. Oliver

Citation: *The Mycota* XIV, S. Poggeler & J. Wostemeyer, Springer-Verlag, Berlin Heidelberg, Chapter 9, 205-229 (2011)

This thesis chapter is submitted in the form of a collaboratively-written book chapter which has been accepted for publication in *The Mycota* volume XIV. As such, not all work contained in this chapter can be attributed to the Ph. D. candidate.

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

- Performed research of scientific literature and wrote the manuscript.

The following contributions were made by co-authors:

- AHW performed analysis of Dothideomycete genome assembly characteristics.
- AHW and RPO edited the manuscript.
- All authors read and approved the manuscript.

Note: Articles writted for *The Mycota* contain 'petite' sections of text. These are indicated by text of a smaller font and contain background information elaborating on the subject matter, but are not critical towards the understanding of the article as a whole.

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

--   --

James Hane (Ph. D. candidate)   Date

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

--   --

Richard Oliver (Principal supervisor)   Date

9 Genomic and Comparative Analysis of the Class Dothideomycetes

JAMES K. HANE[1,2], ANGELA H. WILLIAMS[1], RICHARD P. OLIVER[3]

CONTENTS

I. Introduction

The class Dothideomycetes (Schoch et al. 2009) is a recently defined taxon within the phylum Ascomycota. It is one of the largest classes within the Ascomycota, with approximately 20000 member species. Its members span a wide spectrum of host interactions and lifestyles, which include pathogens of plants, animals and other fungi, epiphytes, saprobes and lichens. Of these, the group which have the greatest human impact and which have received the vast majority of scientific attention are phytopathogenic species. These cause major losses to the agriculture and forestry industries.

The class Dothideomycetes has replaced the class Loculascomycetes (Luttrell 1955), in which species were grouped based on the morphology of their sexual fruiting bodies. Ascus morphology in the Dothideomycetes is distinct from other fungal taxons in that the asci are double-walled (bitunicate). The fruiting bodies (pseudothecia) also form distinctive cavities or 'locules' in which sexual spores (ascospores) are subsequently formed (ascolocular [syn. ascostromatic] development).

Since the creation of this class, the taxonomic placements of Dothideomycete genera and species have been refined several times based on molecular phylogenetic data (Schoch et al. 2006). Recent molecular phylogenies predict two major sub-classes within the Dothideomycetes (Schoch et al. 2009; Fig. 9.1). The Pleosporomycetidae is the larger of the two sub-classes and contains species possessing pseudoparaphyses in contrast to Dothideomycetidae species which do not. The Pleosporomycetidae and Dothideomycetidae are each divided again into four orders. The sub-class Pleosporomycetidae contains the orders Pleosporales, Hysteriales, Mytilinidiales and Jahnulales (Boehm et al. 2009). The Pleosporales is the largest order within the Pleosporomycetidae (Zhang et al. 2009). The sub-class Dothideomycetidae contains the orders Dothideales, Capnodiales, Myriangiales and Trypetheliales (Aptroot et al. 2008). Two remaining orders, the Patellariales and Botryosphaeriales, have not been assigned to either sub-class.

II. Significant Phytopathogenic Species

A. Order Pleosporales

Phaeosphaeria nodorum (anamorph *Stagonospora nodorum*, syn. *Leptosphaeria nodorum*, syn.

[1] Faculty of Health Sciences, Murdoch University, Perth, WA 6150, Australia
[2] CSIRO Plant Industry, CELS Floreat, Perth, WA 6014, Australia
[3] Department of Environment and Agriculture, Curtin University, Perth, WA 6102, Australia; e-mail: richard.oliver@curtin.edu.au

Evolution of Fungi and Fungal-Like Organisms, The Mycota XIV
S. Pöggeler and J. Wöstemeyer (Eds.)
© Springer-Verlag Berlin Heidelberg 2011

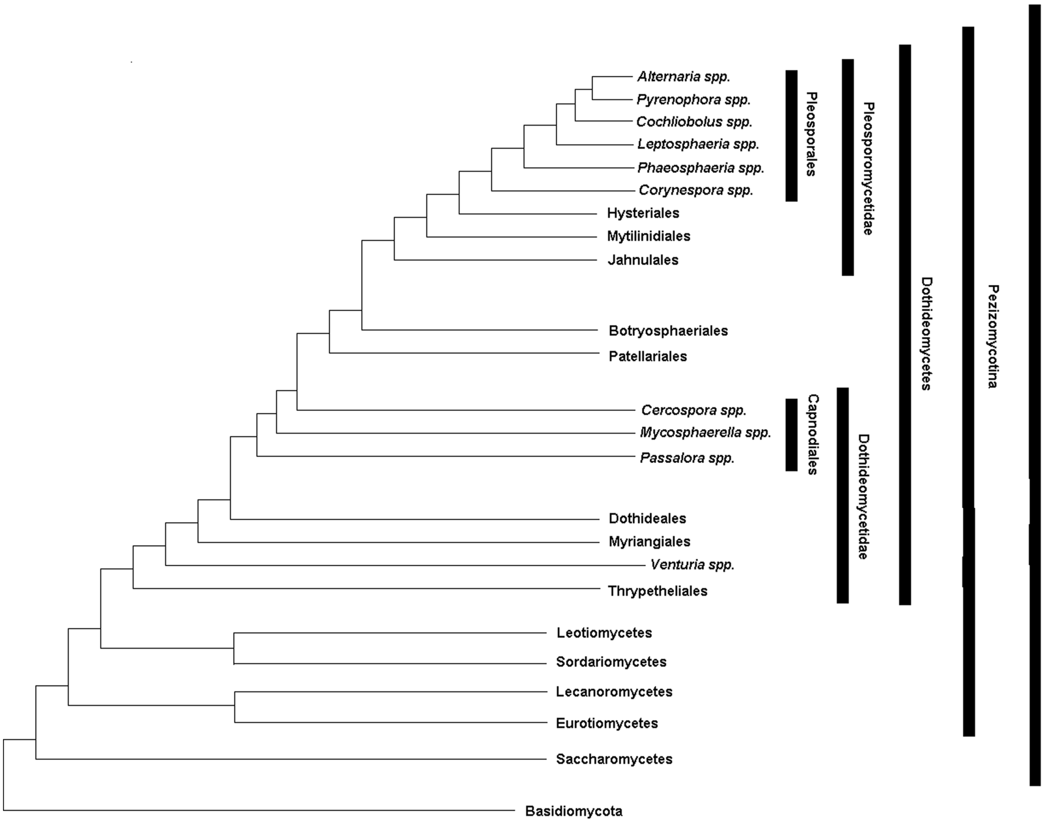


Fig. 9.1. Cladogram of significant phytopathogens within the class Dothideomycetes and their phylogenetic relationships with various classes of the Ascomycota (adapted from Schoch et al. 2009)

Septoria nodorum) is the causal agent of *Stagonospora nodorum* blotch (SNB) and glume blotch in wheat. *P. nodorum* is widely distributed (Solomon et al. 2006) and can cause up to 31% yield losses (Bhathal et al. 2003). Within Western Australia it is the primary wheat pathogen and is estimated to cause AUD 108 million in annual yield losses (Murray and Brennan 2009).

Pyrenophora tritici-repentis (anamorph *Drechslera tritici-repentis*, syn. *Helminthosporium tritici-repentis*) is the causal agent of tan spot (syn. yellow spot) and is a globally spread wheat pathogen. *P. tritici-repentis* emerged relatively recently as a major pathogen of wheat and has since increased rapidly in severity, incidence and distribution. The sudden increase in virulence of *P. tritici-repentis* is believed to have occurred after a lateral gene transfer event with *S. nodorum* (Friesen et al. 2006).

Leptosphaeria maculans (anamorph *Phoma lingam*) is the causal agent of blackleg (syn. stem canker) of oilseed rape (syn. canola) and other crucifers. *L. maculans* is rapidly adaptable to fungicide treatment and has recently overcome *Rlm1* resistance (using 'Surpass 400') in Australia (Li and Cowling 2003; Rouxel et al. 2003) and also *Rlm6* resistance under experimental conditions in France (Brun et al. 2000). Like *P. tritici-repentis*, the incidence and severity of *L. maculans* infection has increased in recent years, particularly in Eastern Europe, replacing the endemic but less virulent *L. biglobosa* (Fitt et al. 2006).

Cochliobolus heterostrophus (anamorph *Bipolaris maydis*), the causal agent of southern corn leaf blight (SCLB) of maize, is widespread in warm and humid areas (Smith and White 1988). *C. heterostrophus* has three distinct races (Hooker et al. 1970): the endemic race O, race C which is currently restricted to China and infects C male-sterile cytoplasm cultivars (Wei et al. 1988), and the highly virulent race T which infects Texas male-sterile cytoplasm cultivars (Levings and Siedow 1992). An epidemic of race T during the 1970s in North America caused a 15% reduction in yield (Ullstrup 1972). The sister species *C. victoriae*, an oat pathogen, was also responsible for an epidemic of Victoria oat blight during the 1940s in the United States (Curtis and Wolpert 2004).

The genus *Alternaria* contains several important phytopathogenic species (Andrew et al. 2009) which also impact upon human health due to mycotoxin contamination of food supplies and the allergenic properties of their airborne spores (Thomma 2003). *A. alternata* causes leaf spot on a wide range of plant species and produces a vast array of mycotoxins and phytotoxins (Thomma 2003). *A. brassicicola* and *A. brassicae* infect most *Brassicae* species (cabbages syn. mustards), causing brassica dark leaf spot and grey leaf spot respectively.

Corynespora cassiicoli is the causal agent of target spot (syn. leaf fall disease). It is a cosmopolitan pathogen with a wide host range including tomato, cucumber, cotton, soybean, tobacco, cocoa and cowpea (Silva et al. 2000). *C. cassiicoli* most significantly impacts upon rubber tree plantations in the south-east Asian tropics (Hashim 1998).

B. Order Capnodiales

Mycosphaerella graminicola (anamorph, *Septoria tritici*) is the most economically significant pathogen of wheat in Western Europe (Palmer and Skinner 2002), causing *Septoria tritici* blotch (STB, syn. *Septoria* leaf blotch, syn. speckled blotch, syn. leaf spot). In the last century, STB has replaced SNB as the major pathogen of wheat in this region due to changes in agricultural practices, air pollution, pathogen control strategies and climactic factors (Hardwick et al. 2001; Bearchell et al. 2005). The genus *Mycosphaerella* also contains *M. fijiensis* (anamorph *Pseudocercospora fijiensis*), which causes black leaf streak disease (syn. black sigatoka) in banana and is widespread in all banana growing regions of the world (Ploetz 2001).

The genus *Cercospora* contains several pathogenic species infecting a wide variety of fruits, vegetables and ornamentals. *Cercospora beticola*, the causal agent of *Cercospora* leaf spot, is the most economically significant pathogen of sugar beet worldwide (Groenewald et al. 2005).

Passalora fulva (syn. *Cladosporium fulvum*) causes tomato leaf mould. *P. fulva* is a highly refined model for gene for gene host–pathogen interactions in which several pathogenicity effectors have been characterised (Oliver 1992; Wulff et al. 2009).

C. Dothideomycetidae of Uncertain Phylogenetic Placement

Venturia inaequalis is the causal agent of apple scab and is geographically widespread in apple-growing regions (Gladieux et al. 2008) and is rapidly adaptable to resistant cultivars and fungicide treatments.

III. Genome Sequencing in the Dothideomycetes

The class Dothideomycetes contains many agriculturally and economically significant species, yet whole genome sequencing within this class was undertaken relatively late compared with other classes of fungi (Fig. 9.2). At the time of writing, only three Dothideomycete genomes have been sequenced and have had a genome analysis published. The first publically released Dothideomycete genome was that of *P. nodorum* in 2005 (Table 9.1), which was followed by the publication of its genome analysis in 2007 (Hane et al. 2007). The genomes of *M. graminicola* and *L. maculans* were released into the public domain in 2007 and 2010 respectively and genome analyses of both were published in 2011 (Table 9.1; Rouxel et al. 2011; Goodwin et al. 2011). The genomes of *C. heterostrophus*, *A. brassicicola* and *M. fijiensis* are also available although publication of their respective genome analyses is still pending (Table 9.1).

A. *Phaeosphaeria nodorum*

The *P. nodorum* genome (Tables 9.1, 9.2) was thefirst species in the class Dothideomycete to be publicly released, analysed and published (Hane et al. 2007). It was sequenced and assembled in 2005 by the Broad Institute (www.broad.mit.edu) in conjunction with the Australian Centre for Necrotrophic Fungal Pathogens (ACNFP), producing 109 scaffold sequences.

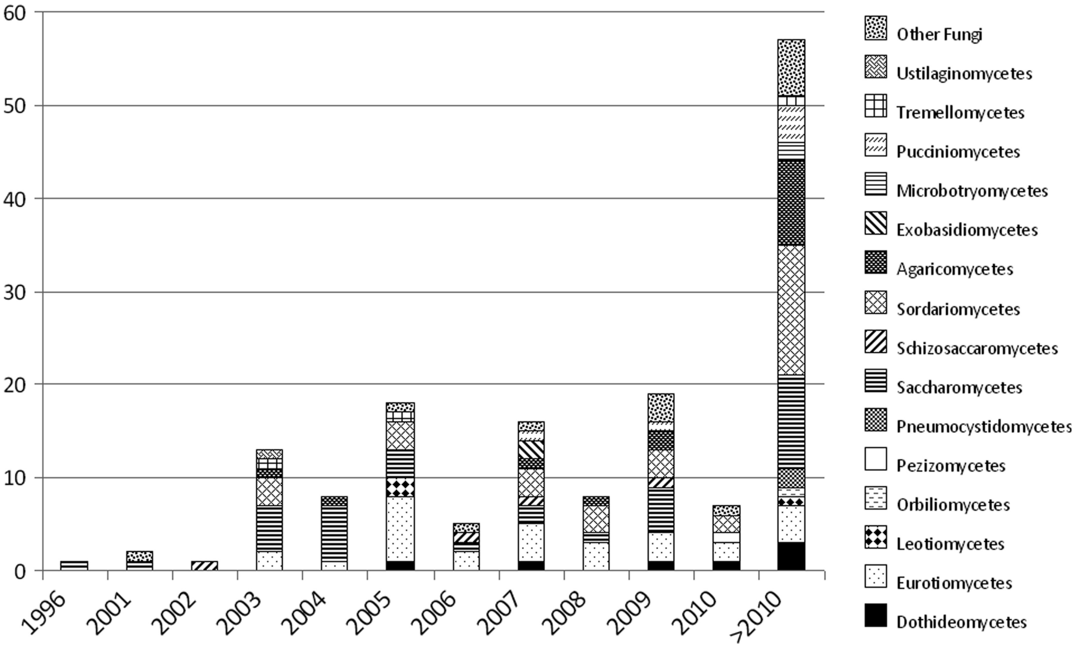


Fig. 9.2. The number of fungal species with publically available whole genome sequences, grouped by class, submitted to the NCBI Genome database over time (http://www.ncbi.nlm.nih.gov/genomes). The column >2010 indicates species for which genome projects are scheduled for upcoming release or proposed for sequencing. The class Dothideomycetes (*black*), although containing several significant species, is relatively under-represented by whole-genome sequencing compared to most classes of Fungi

The Broad also performed an *in silico* gene prediction which produced 16597 hypothetical gene models (Table 9.3; version 1). This gene dataset has been substantially revised and is now obsolete (see below).

The *P. nodorum* genome assembly and gene annotations have since been considerably improved by the ACNFP, combining multiple layers of supporting evidence (Table 9.3). The analysis in Hane et al. (2007) defined a revised gene list (version 2) supported by expressed sequence tags (ESTs), manual curation and EST-trained gene prediction. Expressed sequence tags from oleate-grown mycelium (10752 ESTs) and infected wheat libraries (10751 ESTs) were aligned to the genome assembly, supporting 2695 genes. We manually curated genes with EST support and used 795 fully EST-supported genes to train a second round of gene prediction. This produced a reliable set of 10762 predicted genes (Table 9.3; version 2). A total of 5354 gene models from version 1 were unsupported in version 2 and were retained, but did not form part of the version 2 dataset. The version 2 assembly has 107 scaffolds instead of the 109 in version 1. Two scaffolds were removed from the nuclear assembly and combined to form a circular mitochondrial genome of 49.8 kb (Hane et al. 2007). Version 2 also included minor assembly revisions based on the discovery of sequencing and assembly errors after the re-sequencing of various genomic regions.

Fourteen to 19 chromosomes had been predicted in *P. nodorum* by pulsed-field gel electrophoresis (PFGE; Cooley and Caten 1991). Between 19 and 38 sub-telomeric regions were predicted in the version 2 assembly, which was consistent with the PFGE chromosome estimate (Hane et al. 2007); and 98% of ESTs aligned to the genome assembly, indicating that the gene-rich regions of these chromosomes were well represented by these 107 scaffolds (Hane et al. 2007).

In a subsequent publication, the gene annotations from version 2 were curated using supporting evidence from both proteomic and proteogenomic experiments (Bringans et al. 2009).

Table 9.1. Online bioinformatic resources for currently available dothideomycete species and future dothideomycete genome projects

Species	Strain/isolate	Organisation[a]	Project/data type	Method	Project status	Availability
Alternaria brassicicola	ATCC 96836	WUGC	Draft whole genome assembly	Sanger sequencing	Completed 2006	JGI, NCBI [WGS: ACIW00000000]
		WUGC	Whole genome non-assembled reads	Sanger sequencing	Completed 2006	NCBI [Trace Archive]
		COGEME	EST sequencing	Sanger sequencing		COGEME
		JGI	Gene annotation		Completed 2007	JGI
		WUGC	Genome finishing	BAC physical mapping	Completed 2008	WUGC
			Miscellaneous sequences		Ongoing	NCBI [NUC/PRO]
Cochliobolus heterostrophus	C5	JGI	Draft whole genome assembly	Sanger sequencing	Completed 2007	JGI, NCBI [WGS: ACIW00000000]
		JGI	Whole genome non-assembled reads	Sanger sequencing		NCBI [Trace Archive]
		JGI	Transcriptome sequencing	Roche 454 GS-FLX		NCBI [SRA:SRX014521, SRX014522]
		JGI	EST sequencing	Sanger sequencing		NCBI [EST]
		JGI	Gene annotation		Completed 2008	JGI
			Miscellaneous sequences		Ongoing	NCBI [NUC/PRO]
Leptosphaeria maculans	v23.1.3	Genoscope/INRA	Draft whole genome assembly	Sanger sequencing	Completed 2007	N/A
		Genoscope/INRA	Finished whole genome assembly	Sanger sequencing	Completed 2010	N/A
		Genoscope/INRA	Whole genome non-assembled reads	Sanger sequencing	Completed 2007	NCBI [Trace Archive]
		Genoscope/INRA	Genome finishing	Genetic mapping	Completed 2010	N/A
		INRA	Genome analysis		Published 2011	Rouxel et al. (2011) NCBI [Genome Project: 63129]
		COGEME	EST sequencing			COGEME
			Miscellaneous sequences		Ongoing	NCBI [NUC/PRO]
Mycosphaerella fijiensis	CIRAD86	JGI	Draft whole genome assembly	Sanger sequencing	Completed 2007	JGI, NCBI
		JGI	Finished whole genome assembly	Genetic mapping	Completed 2010	JGI
		JGI	Whole genome non-assembled reads	Sanger sequencing	Completed 2007	NCBI [Trace Archive]
		JGI	EST sequencing	Sanger sequencing	Completed 2007	NCBI [EST]
			Miscellaneous sequences		Ongoing	NCBI [NUC,PRO]
M. graminicola	IPO323	JGI	Draft whole-genome assembly	Sanger sequencing	Complete	JGI

continued

Table 9.1. continued

Species	Strain/isolate	Organisation[a]	Project/data type	Method	Project status	Availability
		JGI	Finished whole-genome assembly	Genetic mapping, re-sequencing, mesosynteny	Completed 2007	
		JGI	Mitochondrial genome assembly	Sanger sequencing	Completed 2007	NCBI [NUC:EU090238]
		JGI	Whole genome non-assembled reads	Sanger sequencing	Completed 2007	NCBI [Trace Archive]
		JGI	Gene annotation			JGI NCBI [Genome Project: 19047]
		JGI/USDA	Genome analysis		Published 2011	Goodwin et al. (2011)
		JGI	EST sequencing	Sanger sequencing	Completed 2007	JGI
			Miscellaneous sequences			NCBI [NUC,PRO]
Phaeosphaeria nodorum	SN15	Broad	Draft whole-genome assembly	Sanger sequencing	Completed 2005	JGI, NCBI [WGS: AAGI00000000]
			Mitochondrial genome assembly	Sanger sequencing	Completed 2005	NCBI [NUC:EU053989]
			Whole-genome non-assembled reads	Sanger sequencing	Completed 2005	NCBI [Trace Archive]
		ACNFP	Gene annotation		Completed 2007	JGI, NCBI [Genome Project: 19047]
		ACNFP	Genome analysis		Completed 2007	Hane et al. (2007)
		ACNFP/Broad	EST sequencing	Sanger sequencing	Completed 2006	NCBI [EST]
		ACNFP	Transciptome analysis	Microarray	Completed 2010	Ip-cho et al. (2010)
			Miscellaneous sequences		Ongoing	NCBI [NUC/PRO]
Pyrenophora tritici-repentis	Pt-1C-BFP	Broad	Draft whole-genome assembly	Sanger sequencing	Completed 2007	Broad JGI, NCBI [WGS: AAXI00000000]
		Broad	Finished whole-genome assembly	Optical mapping	Completed 2007	Broad, JGI
			Whole-genome non-assembled reads	Sanger sequencing	Completed 2007	NCBI [Trace Archive]
			Gene annotation		Completed 2007	JGI, BROAD
			EST sequencing	Sanger sequencing	Completed 2007	NCBI [EST]
			Miscellaneous sequences			NCBI [NUC/PRO/GSS]
			Genome finishing	Optical mapping	Completed 2007	Broad
Aigialus grandis		JGI	Draft genome	454 Titanium and Illumina GAII	Awaiting material	
Baudoinia compinacensis		JGI	Draft genome	454 Titanium and Illumina GAII	Awaiting material	
Cercospora zeae-maydis		JGI	Transcriptome	454 Titanium and Illumina GAII	Complete	

Cenococcum geophilum		JGI	Draft genome	454 Titanium and Illumina GAII	Awaiting material
Cercospora beticola	303B	JGI	Transcriptome	454 Titanium and Illumina GAII	Complete
C. zeae-maydis	SCOH1-5	JGI	Draft genome	454 Titanium and Illumina GAII	Library construction
Cryomyces antarticus	CBS116301	JGI	Draft genome	454 Titanium and Illumina GAII	Awaiting material
Dothistroma septosporum	NZE10	JGI	Draft genome	454 Titanium and Illumina GAII	Library construction
Mycosphaerella graminicola	STIR04 3.11.1	BGI-Shenzhen	Resequencing	Illumina HiSeq	Awaiting material
Passalora (syn. *Cladosporium*) *herbarum*	CBS121621	JGI	Draft genome	454 Titanium and Illumina GAII	Awaiting material
Pyrenophora tritici-repentis	race 4	JGI	Resequencing	454 Titanium and Illumina GAII	Complete
P. tritici-repentis	DW7 race 5	JGI	Draft genome	454 Titanium and Illumina GAII	Complete
Septoria musiva	SO2202	JGI	Draft genome	454 Titanium and Illumina GAII	Library construction
Setosphaeria turcica	Et28A	JGI	Draft genome	454 Titanium and Illumina GAII	Library construction
Trypethelium spp.		JGI	Draft genome	454 Titanium and Illumina GAII	Awaiting material
Zasmidium cellare		JGI	Draft genome	454 Titanium and Illumina GAII	Awaiting material

[a] ACFNP = www.envbio.curtin.edu.au; BGI-Shenzhen = www.genomics.cn; Broad = www.broad.mit.edu; COGEME = cogeme.ex.ac.uk; Genoscope = www.cns.fr/spip; INRA = www.international.inra.fr; JGI = www.jgi.doe.gov; NCBI = www.ncbi.nlm.nih.gov/genomes; USDA = www.usda.gov; WUGC = genome.wustl.edu

Table 9.2. Comparison of sequenced dothideomycete genome characteristics[a] (adapted from Rouxel et al. 2011). Genome statistics were correct at the time of printing

Species	*P. nodorum*	*L. maculans*	*M. graminicola*	*M. fijiensis*	*P. tritici-repentis*	*C. heterostrophus*	*A. brassisicola*
Strain/isolate	SN15	v23.1.3	IPO323	CIRAD 86	Pt-1C (sub-culture BFP, race 1)	C5	ATCC 96866
Assembly version	2	2	2	2	1	1	1
Nuclear genome size (Mbp)	37.21	45.1	39.7	74.1	37.8	34.9	32.0
Assembly coverage	10×	9×	8.9×	7.1×	6.9×	9.9×	6.4×
Sequencing institution	Broad	Genoscope	JGI	JGI	Broad	JGI	WUGC
Assembler algorithm	Arachne	Arachne	Jazz	Jazz	Arachne	Arachne	PCAP
Number of scaffolds[b]	107	72	21	56	47	89	838
Number of chromosomes or linkage groups[c]	14–19 (PFGE)	17–18 (WGF)	21 (WGF)	10–23 (GM)	11 (OM)	15–16 (GM)	9–11 (PFGE) 172 (PM)
Number of scaffolds not contained in chromosomes/ linkage groups	107	45	0	Unknown	22	11	Unknown
Assembly N_{50}	13	10	6	5	4[d]	11	6
Assembly L_{50} (Mbp)	1.0	1.8	2.7	5.9	3.1[d]	1.3	2.5
Assembly $L_{\max}$ (Mbp)	2.5	4.3	6.1	8.6	9.5[d]	2.7	4.2
Assembly $L_{\min}$ (Kbp)	2	0.5	409.2	1.01	6.0[d]	1.3	0.1
Number of annotated genes	12 194	12 469	10 952	13 107	12 169	9633	10 688
Annotation version	3	2	2	2	1	1	1
Gene density (genes per 10 kb)	3.1	4.2[a]	3.4	1.4	3.8	3	3.9
Average gene length (bp)	1326	1323[a]	1600	1599	1618	1836	1523
% repetitive DNA	6.2	34.2	18.0	Unknown	16.0	7.0	9.0

[a] Excludes genes within A:T-rich isochores (see Section II.C).

[b] Excludes scaffolds identified to be mitochondrial in origin.

[c] OM = optical mapping, PFGE = pulse field gel electrophoresis, PM = physical mapping, WGF = whole genome finishing (Cooley and Caten 1991; Tzeng et al. 1992; Malkus et al. 2009; Rouxel et al. 2011).

[d] Calculated using artificial chromosome sequences consisting of scaffolds joined by mapping and unmapped scaffolds.

Table 9.3. Summary of scaffolds, genes and evidence supporting the three versions of the *P. nodorum* genome assembly

Version	1	2	3
Status	Static	Static	Revision in progress
Availability	Broad	NCBI/JGI	Richard.Oliver@curtin.edu.au
Number of scaffolds	109	107 (+1 mitochondrial)	107 (+1 mitochondrial)
Number of genes	16 597	10 762	12 194
Supporting evidence			
Re-sequencing	No	Yes	Yes
Transcriptomics: EST	No	Yes	Yes
Transcriptomics: microarray	No	No	Yes
Proteomics	No	No	Yes
Orthology	No	No	Yes

Proteomics involves matching the masses and fractionation patterns of semi-purified proteins which have been trypsin-digested and analysed via mass-spectrometry (MS) against a peptide database generated in silico from a set of known or predicted protein sequences. Proteogenomics uses complex mixtures of proteins, which are then trypsin-digested, separated by two-dimensional liquid chromatography and analysed by tandem mass spectrometry

(MS/MS). The MS data are then compared to both the predicted proteome and a larger database generated by the wholesale translation of open reading frames from both strands of the entire genome assembly. Both proteogenomic and conventional proteomics techniques can improve gene/protein identification. As draft genome assemblies rely heavily on gene prediction algorithms to determine the proteome, there is a high risk of incorrectly identifying proteins or not detecting them at all if a protein dataset is incorrect. As proteogenomics maps peptides directly to the genome sequence it can bypass any errors introduced by inefficient gene prediction algorithms. This 'direct to genome' peptide mapping also makes proteogenomics a powerful tool for genome curation. Proteogenomics can identify errors in start and stop codons and exon–intron boundaries in hypothetical gene models. It can also identify incorrect gene translations caused by frame shifts that have been introduced by either misannotation, genome sequencing errors or assembly errors. Additionally, proteogenomics is capable of detecting protein coding regions of a genome not previously identified by gene prediction software, thus providing data for the annotation of new genes.

Proteomic and proteogenomic analysis of intracellular proteins of *P. nodorum* supported 1946 version 2 gene models and 188 version 1 gene models (1946 + 188 = 2134; Bringans et al. 2009), which was comparable to that supported by EST sequencing (2695). Of the genes detected by proteogenomics, 62% had not previously been detected by EST sequencing. Several gene annotations were revised after consideration of proteogenomic data, with frame-shift errors detected in 144 genes and exon boundary errors detected in 604 genes. Three new genes were also predicted.

The curation of the latest version of the *P. nodorum* genome, version 3, is still in progress and is currently comprised of 12194 gene models. Version 3 improves upon version 2, incorporating the proteogenomic data summarised in Bringans et al. (2009) as well as additional proteomic, microarray (IpCho 2010) and orthology data (Hane 2011). It is available on request from Professor Richard Oliver (Richard.Oliver@curtin.edu.au) and its use over previous versions is strongly encouraged.

B. *Mycosphaerella graminicola*

The genome of *M. graminicola* (Tables 9.1, 9.2) was publically released in 2006 (version 1) and revised in 2008 (version 2). The version 2 genome sequence has been curated to a high degree of accuracy and completeness, containing nearly the entire genome from one telomere to the other on all chromosomes (Goodwin et al. 2011).

The *M. graminicola* genome contains 10952 annotated genes on a total of 21 chromosomes. Thirteen of these chromosomes are essential or 'core' chromosomes and 8 dispensable chromosomes are non-essential and can be absent in some isolates. These dispensable chromosomes make up 12% of the total genome size, and are distinct from the core chromosomes in being far smaller in size (0.42–0.77 Mb) and having higher repeat contents, lower gene densities and abnormal structure. Over half (54%) of core genes were assigned a functional prediction by Goodwin et al. (2011), compared to 2% of genes in dispensable chromosomes. The majority of dispensable chromosome genes were divergent paralogs of core genes and were highly represented by transcription factors and genes relating to regulation of expression and signal transduction. Dispensable chromosomes were also enriched for microRNAs, containing 21% of predicted microRNAs within the whole genome. Some regions of dispensable chromosomes corresponded to fragments of core chromosome sequences. Goodwin et al. (2011) proposed that the dispensable chromosomes are byproducts of meiotic recombination. However, no significant synteny (see Section III.A) between dispensable chromosomes and the core chromosomes or chromosomes of other species was detected.

C. *Leptosphaeria maculans*

The genome assembly of *Leptospaeria maculans* v23.1.3 (Tables 9.1, 9.2) was completed in 2007 (Rouxel et al. 2011). The genome comprises 45.1 Mb in 77 scaffolds. Genetic mapping, CHEF hybridisation and bioinformatic predictions (refer to section III.A.1) determined that these scaffolds are contained on 17–18 chromosomes. A total of 42222 ESTs across 15 different libraries were sequenced in three isolates of *L. maculans* (IBCN18, PL86, v23.1.3). Rouxel et al. (2011) predicted 12 469 genes with coding sequences greater than 100 bp based on the genomic alignment of ESTs and in silico predictions. Genes with coding sequences greater than 300 bp and with supporting evidence from EST alignment or association with conserved domains, a total of 11561 genes, were considered to be reliable gene models.

The remaining 908 less reliable gene models were manually revised.

The genome is divided into two types of isochores (regions of distinct G:C content), G:C-equilibrated (51% G:C content) and A:T rich isochores which have low (34%) G:C content. The gene density in G:C-equilibrated isochores is high, compared to the heterochromatin-like A:T-rich isochores which have low gene densities and high repetitive DNA content. Putative pathogenicity effectors make up a large proportion (20%) of predicted genes within the A:T-rich isochores. Many of these putative effectors are cysteine-rich, small secreted proteins (SSPs) with plant translocation signals. The whole-genome microarray was designed with probes specific to 12396 Eugene-predicted gene models as well as 63 non-predicted, putative SSP genes and 1316 consensus sequences of EST clusters that did not align to the genome. Gene expression was analysed in 1-week-old culture grown on media and infected oilseed rape sampled at 7 and 14 days post-infection (dpi). Most of the EuGene predictions (84.4%) and unmapped EST clusters (90.8%) were expressed in at least one of these conditions, whereas this applied to only half (51.0%) of the unpredicted SSP genes. Microarray analysis found the expression of SSP genes residing within AT-rich isochores to be up-regulated during infection (at 7 dpi).

Gene predictions were also validated by proteomic analysis of mycelial and culture filtrate samples. The majority (83%) of detected proteins in these samples contained a predicted signal peptide sequence. However, only 39 SSPs were detected, of which none resided within AT-rich isochores.

IV. Comparative Genomics

A. Synteny

Comparing whole genomes can reveal the extent to which species have diverged over time and also highlights regions of synteny. In the context of genomics, synteny is a term used to describe the preservation of the physical order of orthologous genes between two species.

Synteny between two genomes can be comprehensibly visualised in the form of a 'dot-plot'. A dot-plot is a two-dimensional graph representing matching sequences between two genomes. The sequences belonging to either species are represented by the lengths of the *x*- and *y*-axes. Sequence matches are represented by lines drawn between the *x*- and *y*-coordinates corresponding to their genomic location in either species. At a whole genome resolution, these lines are usually visualised as dots. Syntenic regions appear as many dots in a linear arrangement hence synteny is also commonly referred to as 'co-linearity'.

Synteny can be qualitatively distinguished based on the length of the matching region. 'Macrosynteny' refers to synteny across large regions which are observable at a whole chromosome scale (Fig. 9.3A). 'Microsynteny' refers to synteny at the scale of a handful of genes (Fig. 9.3B). Several cases of macrosynteny have been observed between the genomes of animals (McLysaght et al. 2000; Pennacchio 2003; Kohn et al. 2004) and between those of plants (Cannon et al. 2006; Phan et al. 2007; Shultz et al. 2007). However within Fungi, macrosynteny has only been reported between species within the genus *Aspergillus* (Galagan et al. 2005; Machida et al. 2005; Pel et al. 2007).

Synteny between two species implies either a relatively short length of time since speciation or a selective pressure to retain the physical arrangement of genes. Within the Dothideomycetes, most reported examples of microsyntenic gene clusters are involved in the biosynthesis of secondary metabolites. *Leptospaeria maculans* has a gene cluster for the synthesis of the polyketide sirodesmin which is syntenic with *Aspergillus fumigatus*, *Chaetomium globosum*, *Magnaporthe grisea* and *Fusarium graminearum* (Gardiner et al. 2004). *L. maculans* also has a gliotoxin biosynthesis gene cluster which is syntenic with *A. fumigatus* (Gardiner and Howlett 2005). *Phaeosphaeria nodorum* has a putative polyketide biosynthesis cluster that is partially syntenic with *Sordaria macrospora* (Nowrousian et al. 2010) and a quinate biosynthesis cluster that is well conserved between a wide range of fungi including *A. nidulans*, *M. grisea*, *Neurospora crassa*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe* (Hane et al. 2007). Nonetheless, in all of the cases above, co-linearity is never precisely conserved and a certain number of genes are rearranged in order or orientation (see Chapter 10 in this volume).

Whole genome comparisons of *L. maculans* (Rouxel et al. 2011) and *M. graminicola* (Goodwin et al. 2011) with *P. nodorum* (Hane et al. 2007)

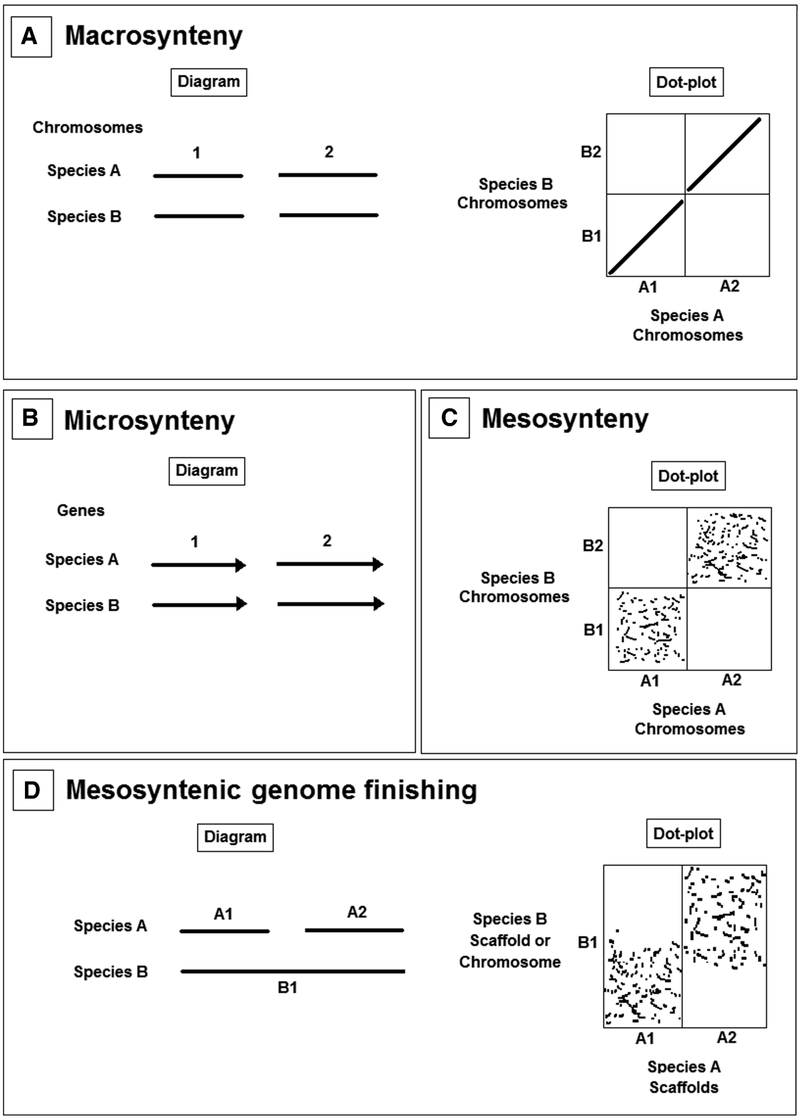


Fig. 9.3. Different types of synteny relationships. (**A**) Macrosynteny is characterised by co-linear conservation of sequence that can be observed at a chromosomal scale. In the diagram provided, chromosomes A1 and A2 of species A are respectively equivalent to chromosome B1 and B2 of species B. This can be represented in a dot-plot, where the axes represent the length of the sequence and lines within the boxes represent matching sequences. (**B**) Microsynteny is characterised by co-linear conservation of sequence and order observable at gene level. In the diagram provided, genes A1 and A2 of species A correspond respectively to genes B1 and B2 of species B. This type of relationship is usually too small to be visualised in a dot-plot at chromosomal level. (**C**) Mesosynteny is characterised by conservation of sequence between equivalent chromosomes without the conservation of sequence order. In the dot-plot provided, chromosomes A1 and A2 of species A correspond respectively to chromosomes B1 and B2 of species B. The mesosyntenic pattern appears as a scattered 'block' of 'dots', which represent multiple regions of microsynteny rearranged in order. (**D**) Mesosynteny can be used to complete genomes which are comprised of partial chromosome sequences, or scaffolds. In the diagram, scaffolds A1 and A2 are incomplete scaffolds which correspond to the complete B1 sequence. By observing the mesosyntenic conservation with B1, shown on the dot-plot, A1 and A2 can be hypothesised to be physically co-located to the same chromosome in species A

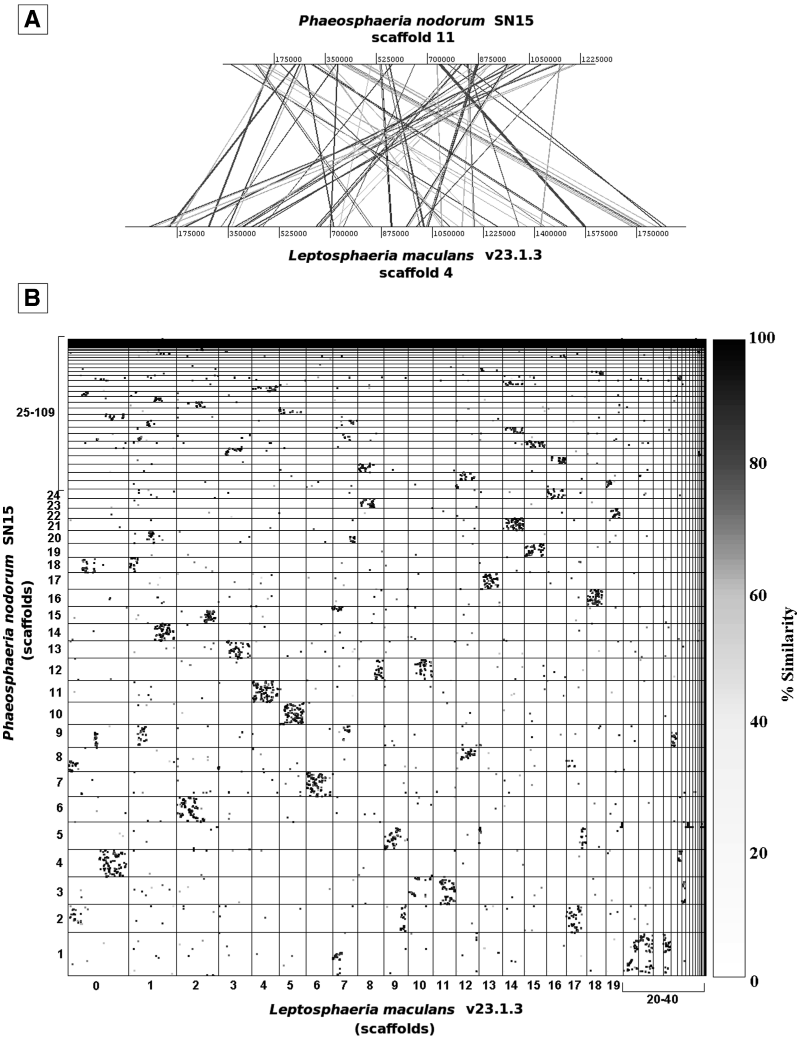


Fig. 9.4. Mesosyntenic genome finishing in *L. maculans* (adapted from Rouxel et al. 2011). (**A**) Mesosyntenic conservation between two scaffolds. Scaffold 4 of *L. maculans* and scaffold 11 of *P. nodorum* share multiple regions of homology. *Lines* drawn between the sequence pair represent regions matched by tblastx with >75% identity over >500 bp length. (**B**) The six-frame translations of both genomes were compared via MUMMER 3.0 (Kurtz et al. 2004). Scaffolds 0–40 from *L. maculans* are aligned along the *x*-axis and scaffolds 1–109 from *P. nodorum* are aligned along the *y*-axis. *Dots* represent matching regions, similar to the *lines* in **A**, between translated scaffold sequences. Presented as a dot-plot, mesosyntenic regions appear as rectangular 'blocks' comprised of many dots. Mesosyntenic conservation between scaffold 4 of *L. maculans* and scaffold 11 of *P. nodorum* **A** is retained between this sequence pair almost exclusively, suggesting that these scaffolds correspond to equivalent chromosomes

both exhibit a novel pattern of sequence conservation (Figs. 9.3, 9.4). This conservation pattern is characterised by a distinct lack of co-linearity and extensive rearrangement of short sections of a sequence relative to the compared species (Fig. 9.4A). This pattern is represented on a dot-plot as 'blocks' of many 'dots' whereby regions homologous to one scaffold are predominantly found in one or a few scaffolds of the other fungus (Fig. 9.4B). We also note that the dots are not found in diagonal lines, characteristic of macrosynteny. As the dots are mainly genes, the pattern means that genes are conserved on homologous chromosomes but their order and orientation is not conserved. We have coined the term 'mesosynteny' to refer to this novel conservation pattern (Rouxel et al. 2011; Goodwin et al. 2011; Hane et al 2011).

This pattern degrades with increasing evolutionary distance as indicated by the level of sequence similarity between *P. nodorum* and *L. maculans* (Fig. 9.4, Pleosporales vs Pleosporales, 80–100%) compared to that between *P. nodorum* and *M. graminicola* (Fig. 9.5; Pleosporales vs. Capnodiales, 60–80%). Similarly, the density of dots within blocks was also observed to become sparser with increasing evolutionary distance. This type of conservation is observed beyond the previously mentioned species and is common between the genomes of Dothideomycete fungi.

A whole-genome comparison between the Sordariomycetes *Podospora anserina* and *Neurospora crassa* performed by Espange et al. (2008) also reported a mesosyntenic-like pattern. The order of microsyntenic regions was extensively rearranged with respect to the alternate genome (as in Fig. 9.3C). As the genomes of both species had been assembled into whole chromosomes (Galagan et al. 2003; Espagne et al. 2008), Espange et al. (2008) observed that these microsyntenic regions were almost exclusively co-located on 'equivalent' chromosomes (i.e., chromosomes derived from the same chromosome of their last common ancestor) and intra-chromosomal rearrangements occurred at a much higher rate than that of translocations between non-sister chromosomes.

1. Completion of Draft Genome Assemblies Using Mesosyntenic Predictions

As chromosomal content is retained (albeit rearranged in order and orientation) on equivalent chromosomes between species exhibiting mesosynteny, it can be applied to the process of genome 'finishing' (Rouxel et al. 2011; Goodwin et al. 2011). Whole genome sequencing involves the sequencing of many short DNA fragments which are bioinformatically assembled into longer sequences. Unfortunately, repetitive sequences often cannot be assembled unambiguously. Therefore, draft assemblies of fungal genomes are not composed of complete chromosome sequences but of contigs or scaffold sequences which represent partially sequenced regions of chromosomes. Conventional finishing methods include genetic mapping (Rouxel et al. 2011), optical mapping (Schwartz et al. 1993), HAPPY mapping (Dear and Cook 1993) and re-sequencing. Using mesosyntenic predictions to aid genome finishing has certain advantages over these methods. It is applicable to fungal species in which a sexual phase is absent or cannot be induced under laboratory conditions, making genetic mapping impossible. It also has the potential to achieve the same end result as these traditional techniques at lower expense and throughput. 'Mesosyntenic finishing' (Fig. 9.3C) of a species of interest (species A) relies on the availability of a whole or draft genome of a related species (species B). All sequence combinations between two genomes are tested for significant proportions of matching sequence. A bioinformatic method which automates this process has been developed by the authors of this article (unpublished data). If two scaffold sequences (A1 and A2) from species A both exhibit mesosyntenic conservation with a single sequence (B1) of species B, then scaffolds A1 and A2 can be hypothesised to be co-located on the same chromosome. The hypothesis can be tested in a number of ways. Scaffold-specific primers can be designed leading outwards from the scaffold termini. PCR amplification can only occur between such primers if the two scaffolds are physically joined thus the presence of an amplicon validates the hypothesis. In addition, sequencing the amplicon can improve the genome assembly by filling in gaps in the sequence. Alternatively, hybridisation techniques such as Southern blotting, binding scaffold-specific probes to fragmented genomic DNA can be employed on a much smaller scale than would be feasible by a 'blind' approach. Chromosomal co-location, order and orientation of scaffolds can be determined by hybridization of multiple probes to the same DNA fragment.

These techniques contributed to finalising the draft genomes of *L. maculans* and *M. graminicola*. Mesosyntenic analysis predicted joins between *L. maculans* scaffolds, based on comparisons with *P. nodorum* (Fig. 9.4): 8 and 10, also 20, 21 and 23. It also predicted joins based on comparisons with other Dothideomycetes (data not shown): 2 and 19, 3 and 31, also 6, 11 and 29, also 12, 15 and 32. Mesosyntenic predictions for

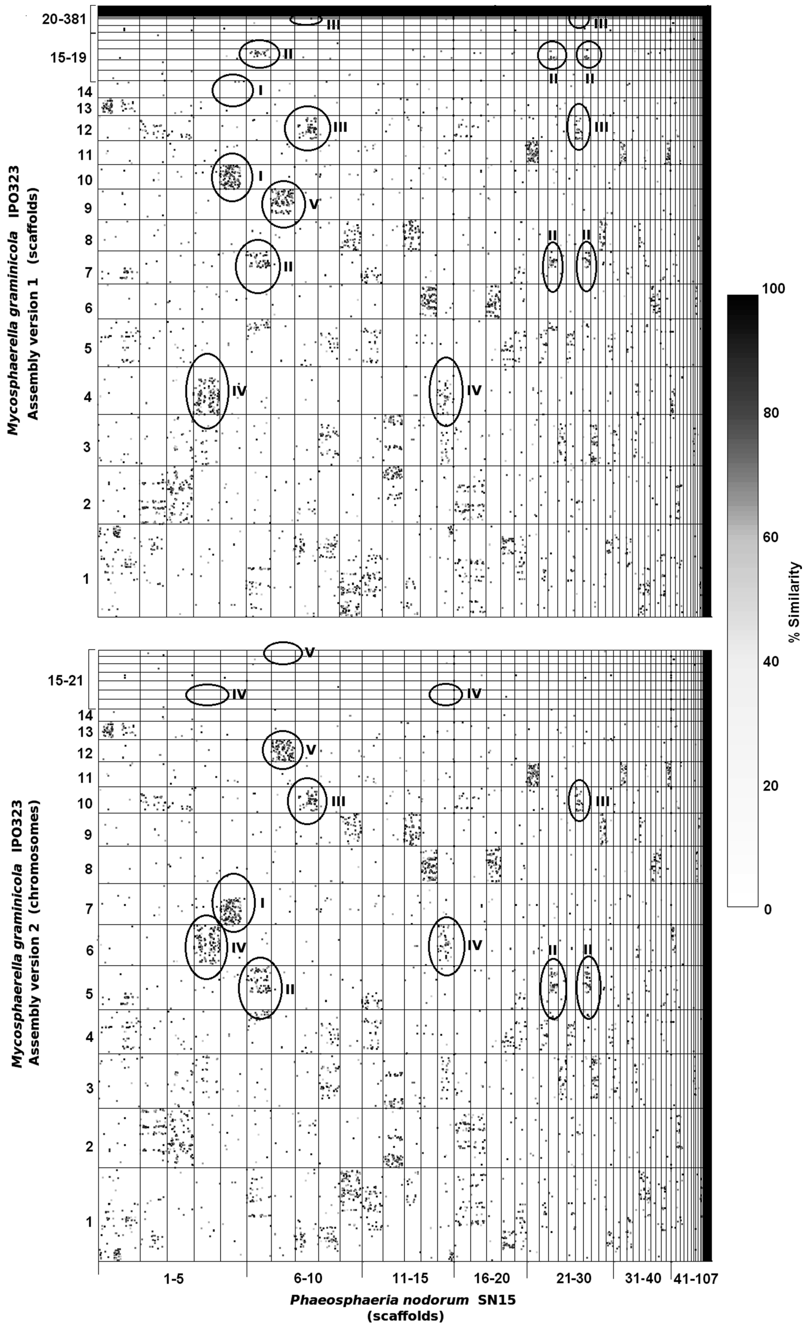


Fig. 9.5. Mesosyntenic genome finishing in *M. graminicola* (Goodwin et al. 2011; Hane 2011). Comparisons of *M. graminicola* genome assembly (*y*-axis) versions 1 (*top*) and 2 (*bottom*) against the genome assembly of *P. nodorum* (*x*-axis). Scaffolds/chromosomes are ordered along their respective axes according to both decreasing length and increasing number. The six-frame translations of both genomes were compared via MUMMER 3.0 (Kurtz et al. 2004). Homologous regions are plotted as dots, which are shaded for percent similarity (*right*). Mesosyntenic predictions were used to finalise the assembly of scaffolds (Version 1) into whole chromosomes (Version 2) as indicated by the numbered *circles*. Groups I–III demonstrate instances where *M. graminicola* scaffolds were joined to form a chromosome and those in group IV and V illustrate where a scaffold was cleaved to form two separate chromosomes

Table 9.4. Features of dothideomycete mitochondrial genome sequences (adapted from Hane et al. 2007; Torriani et al. 2008; Rouxel et al. 2011)

	P. nodorum	*M. graminicola*	*L. maculans*
Reference	Hane et al. (2007)	Torriani et al. (2008)	Rouxel et al. (2011)
Strain/isolate	SN15	IPO323, STBB1	v23.1.3
Size (bp)	49761	43960	154863
G:C content	29.4	32	30
Genetic code[a]	4	4	–
tRNA	27	27	–
rRNAs	2	2	–
Large ribosomal RNA subunit (*rnl*)	Yes	Yes	–
Small ribosomal RNA subunit (*rns*)	Yes	Yes	–
Mitochondrial protein-encoding genes	12	14	–
ATP synthase subunits	1	3	–
atp6	Yes	Yes	–
atp8	No	Yes	–
atp9	No	Yes	–
Cytochrome oxidase subunits	3	3	–
cox1	Yes	Yes	–
cox2	Yes	Yes	–
cox3	Yes	Yes	–
Cytochrome b (*cytb*)	Yes	Yes	–
Nicotinamide adenine dinucleotide ubiquinone oxidoreductase subunits	7	7	–
nad1	Yes	Yes	–
nad2	Yes	Yes	–
nad3	Yes	Yes	–
nad4	Yes	Yes	–
nad4L	Yes	Yes	–
nad5	Yes	Yes	–
nad6	Yes	Yes	–
5S ribosomal protein (*rps5*)	Yes	No	–
Unknown ORFs	3	8	–
Intronic endonucleases	4	0	–

[a] 4 indicates the mold, protozoan, and coelenterate mitochondrial code and the Mycoplasma/Spiroplasma code, http://www.ncbi.nlm.nih.gov/taxonomy/utils/wprintgc.cgi

M. graminicola in comparison with *P. nodorum* predicted all scaffold joins between version 1 scaffolds (scaffolds: 10 and 14, 7 and 17, 12 and 22) and scaffold breaks (scaffolds 4 and 9) which were updated in version 2 (Fig. 9.5). These predictions were verified by genetic mapping, hybridisation and re-sequencing in *L.maculans* and genetic mapping and re-sequencing in *M. graminicola* (Rouxel et al. 2011; Goodwin et al. 2011).

B. Mitochondrial Genomes

Mitochondrial genome sequences (mtDNAs) are currently available for all three published Dothideomycete genomes (*P. nodorum*, *M. graminicola*, *L. maculans*) however only those of *P. nodorum* and *M. graminicola* have been annotated and studied in detail (Hane et al. 2007; Torriani et al. 2008; Rouxel et al. 2011). All three mtDNAs have similarly low G:C contents, ranging from 29.4 to 32.0% (Table 9.4) which are typical of the mitochondrial genomes of filamentous fungi (Torriani et al. 2008). The mitochondrial genomes of *P. nodorum* and *M. graminicola* are similar in size but the mtDNA of *L. maculans* is approximately three times larger (Table 9.4). The *L. maculans* assembly also has four mitochondrial linear plasmids ranging in size from 4.4 to 7.6 kb which contain polymerase genes (Rouxel et al. 2011).

Both *P. nodorum* and *M. graminicola* mtDNA encode the large (rnl) and small (rns) subunits of the mitochondrial ribosomal RNA complex and 27 transfer RNAs (tRNAs) which can attach to all 20 amino acids (Hane et al. 2007; Torriani et al. 2008). These tRNAs are contained within five separate clusters in both mtDNAs (Fig. 9.6, Table 9.5). The two largest of these clusters in both species are those flanking the rnl gene. The 5' flanking tRNA cluster is well conserved, however the 3' cluster is variable between Dothideomycetes species and also across the Ascomycetes.

Some tRNAs in *P. nodorum* and *M. graminicola* were atypical. Both had mitochondrial tRNAs for threonine and phenylalanine with nine nucleotides instead of the typical seven in the their anti-codon loops and tRNA-Arg2 of *P. nodorum* had 11 nucleotides in its anticodon loop.

Fungal mtDNAs typically encode for 12 mitochondrial proteins, which are hydrophobic subunits of respiratory chain complexes (Table 9.4). *P. nodorum* lacks the *atp8* and *atp9* genes whereas *M. graminicola* lacks the *rps5* gene. *M. graminicola* mitochondrial genes differ from those of most fungi in that they lack introns

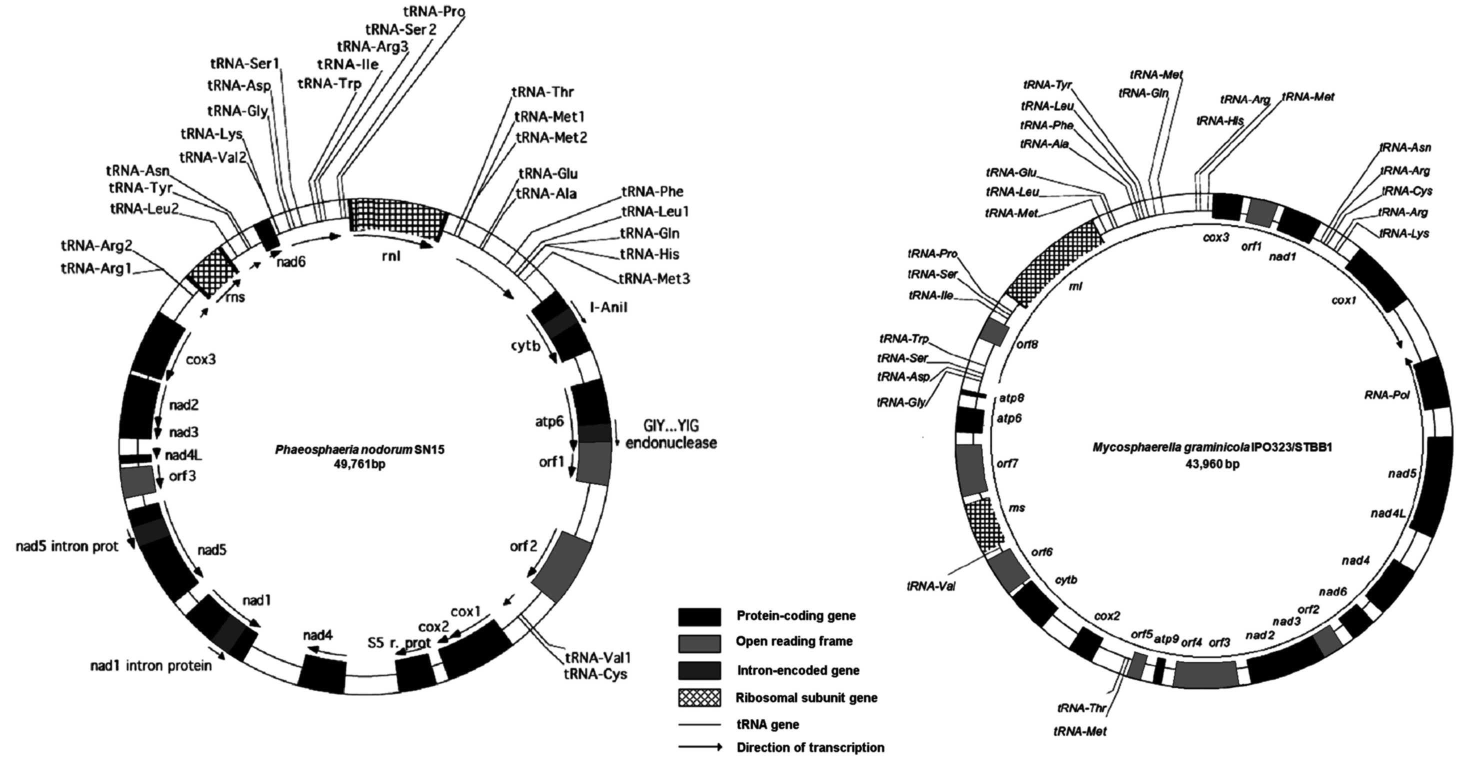


Fig. 9.6. Published and annotated mitochondrial genomes of the Dothideomycetes *Phaeosphaeria nodorum* and *Mycosphaerella graminicola* (adapted from Hane et al. 2007 and Torriani et al. 2008)

Table 9.5. Comparison of tRNAa gene clusters flanking the *rnl* gene in dothideomycete and ascomycete species (adapted from Hane et al. 2007; Torriani et al. 2008)

Species	Class	5’ Upstream[a,b,c]	rnl	3’ Downstream[b,c]	Genbank accession
Mycosphaerella graminicola	Dothideomycetes	$\mathrm{GDS^1WIS^2P}$	rnl	$\mathrm{M^1L^1EAFL^2YQM^2HRM^3}$	EU090238
Phaeosphaeria nodorum	Dothideomycetes	$\mathrm{VKGDS^1WIRS^2P}$	rnl	$\mathrm{TM^1M^2EAFLQHM^3}$	EU053989
Aspergillus niger	Eurotiomycetes	$\mathrm{KGDS^1WIS^2P}$	rnl	$\mathrm{TEVM^1M^2L^1AFL^2QM^1H}$	DQ207726
A. tubingensis	Eurotiomycetes	$\mathrm{KGDS^1WIS^2P}$	rnl	$\mathrm{TEVM^1M^2L^1AF^2QLM^3H}$	DQ217399
Epidermophyton floccosum	Eurotiomycetes	$\mathrm{KGDS^1IWS^2P}$	rnl	$\mathrm{TEVM^1M^2L^1AFL^2QM^3H}$	AY916130
Penicillium marneffei	Eurotiomycetes	$\mathrm{RKG^1G^2DS_1WIS^2P}$	rnl	$\mathrm{TEVM^1M^2L^1AFL^2QM^3H}$	AY347307
Verticillium dahliae	Sordariomycetes	$\mathrm{KGDS{*}VW{*}R{*}P^1{*}P^2}$	rnl	$\mathrm{TE^1M^1M^2L^1AFL^2QHM^3}$	DQ351941
Metarhizium anisopliae	Sordariomycetes	$\mathrm{YDS^1N{*}G{*}LIS^2W}$	rnl	$\mathrm{TEM^1M^2L^1AFKL^2QHM^3}$	AY884128
Lecanicillium muscarium	Sordariomycetes	GVISW*P	rnl	$\mathrm{TE^1M^1M^2L^1E^2FKL^2QHM^3}$	AF487277
Fusarium oxysporum	Sordariomycetes	VISWP	rnl	$\mathrm{TEM^1M^2L^1AFKL^2QHM^3}$	AY945289
Trichoderma reesii	Sordariomycetes	ISWP	rnl	$\mathrm{TEM^1M^2L^1AFKL^2QHM^3}$	AF447590
Podospora anserina	Sordariomycetes	ISP	rnl	$\mathrm{TEIM^1L^1AFL^2QHM^2}$	X55026

[a] Asterisks indicate where functional genes interrupt the tRNA gene sequence.

[b] The numbers 1, 2, 3 indicate the presence of more tRNA genes for the same amino acid in the consensus sequences.

[c] Capital letters refer to tRNA genes for: R, arginine; K, lysine; G, glycine; D, aspartic acid; S, serine; W, tryptophan; I, isoleucine; P, proline; T, threonine; E, glutamic acid; V, valine; L, leucine; A, alanine; F, phenylalanine; Q, glutamine; H, histidine; Y, tyrosine; N, asparagine.

and therefore have no intron-encoded proteins. *P. nodorum* has 4 such intronic genes which encode for GIY-YIG and LAGLIDADG-type endonucleases. Both *P. nodorum* and *M. graminicola* have additional open-reading frames of unknown function (3 and 8 respectively). The expression of the ORFs *orf5*, *orf6* and *orf8* of *M. graminicola* has been confirmed by EST sequencing (Torriani et al. 2008).

C. Repetitive DNA

P. nodorum was initially predicted to have 26 different families of long interspersed repeats, which make up 6.2% of the nuclear genome (Table 9.2; Hane et al. 2007). The number of repeat families has since been revised to 25, of which 17 have been characterised (Hane and Oliver 2008, 2010). A significant proportion of these are non-transposon-derived, instead being highly replicated copies of endogenous genes and gene clusters. These genes include a telomere associated RecQ helicase, ubiquitin conjugating enzyme, *Rad5* and *Rad6* homologues and several genes of unknown function. These endogenous gene-rich repeats were among the largest and most abundant types of repeats in *P. nodorum*, on a scale eclipsing most transposable element repeat families.

The *M. graminicola* genome is comparable in size to *P. nodorum*, (39.7 and 37.1 Mb, respectively) but its repetitive content is much larger at 18% (Table 9.2). The repetitive content is almost doubled within the dispensable chromosomes (30%; see Section II.B) compared to its core chromosomes (15.9%). Details of the types of repeats present in the *M. graminicola* genome have not been published.

The genome of *L. maculans* is significantly larger (45 Mb) than *P. nodorum* and *M. graminicola* (Rouxel et al. 2011). This is due to the higher content of repetitive DNA making up 34% of the whole genome (Table 9.2). Unlike *P. nodorum*, the most abundant repeat families are class II retroelements, with nine families making up 27.26% of the genome. The next most abundant is class I DNA transposons, with nine repeat families making up 2.64% of the genome. There are 11 other repeat families, including rDNA and short telomeric repeats, which together make up the remaining 4.12%. Rouxel et al. (2011) propose that genome invasion by retrotransposons on such a vast scale caused the formation of distinct regions of consistent G:C content (syn. isochores; see Section II.C).

While this data indicates vast diversity in the repeat content of each Dothideomycete genome, each analysis has used different software and detection criteria for the identification of repetitive DNA. *P. nodorum* repeats were predicted *de novo* via RepeatScout (Price et al. 2005), requiring each repeat family to have a minimum full length of 200 bp, at least 75% identity and a minimum of 10 full-length copies. *L. maculans* repeats were predicted via the REPET pipeline (Bao and Eddy 2002; Edgar and Myers 2005; Quesneville et al. 2005; Rouxel et al. 2011). Details of the method and criteria for prediction of repeats in *M. graminicola* have not been published. Due to these differences in methodology these results are not directly comparable.

1. Repeat-Induced Point Mutation

The repetitive DNA of many fungal genomes is subject to a genome defence mechanism specific to fungi called a repeat-induced point mutation (RIP). RIP was first observed in *Neurospora crassa* (Selker et al. 1987) and has since been experimentally demonstrated to occur in many other filamentous fungi (Hane and Oliver 2008). It is likely that RIP protects the genome against transposon replication by introducing mutations in repetitive DNA that are likely to introduce stop codons in transposon genes.

Prior to meiosis, RIP converts cytosine bases to thymine within similar regions of DNA of a sufficient length. In *N. crassa* the requirements for RIP are a shared sequence identity of at least 80% over at least 400 bp (Watters et al. 1999). The base adjacent to the mutated cytosine is biased towards adenine (i.e., CpA dinucleotides become TpA) in most species. The mutation of CpA to TpA has a high probability of introducing TAG or TAA stop codons within the repetitive region on both strands of the affected DNA molecule. RIP can be detected by comparing the transition to transversion ratio (Tn:Tv) of retroelement sequences. Transitions are mutations between two purines bases (i.e., A→G and G→A) or two pyrimidines (i.e., C→T and T→C) whereas transversions are mutations from a purine to pyrimidine, or vice versa (i.e., C→A, A→C, G→T or T→G).

RIP has been observed in *L. maculans.* After the introduction of multiple copies of an exogenous hygromycin resistance gene and subsequent sexual crossing, RIP-like sequence mutations were detected within resistance gene regions in hygromycin-sensitive progeny (Idnurm and Howlett 2003). Whole genome analysis indicated that RIP is highly active in *L. maculans*, with only 19 out of 42 222 EST sequences aligning to repetitive regions and no detectable expression of transposon open-reading frames by microarray (Rouxel et al. 2011).

RIP was bioinformatically predicted in the whole genome of *M. graminicola* (Goodwin et al. 2011).

As RIP mutations involve C→T transitions, the Tn:Tv ratio of RIP-affected repeats was expected to be higher than non-RIP-affected regions. The Tn:Tv of selected repetitive regions were reported by Goodwin et al. (2011) to be significantly higher (ranging from 25.3 to 42.5) than that of a non-repetitive control set of mutated sequences (1.0). Furthermore, coding regions from transposons present in high copy number were found to contain numerous stop codons as would be expected after RIP.

The repeats of *P. nodorum* were analysed using recently developed software for the analysis of RIP mutation: RIPCAL (Hane and Oliver 2008). Traditional methods of predicting and quantifying RIP have relied on the use of ratios of di-nucleotide frequencies, or RIP indices. While indices are useful when searching for RIP within single-copy sequences, with the availability of whole-genome sequences it is now more appropriate to use alignment-based methods on all repeats within a repeat family. RIPCAL extracts repeat sequences from a whole genome and uses CLUSTALW to generate multiple alignments of each repeat family. Mutations were compared within each repeat family, quantified and their RIP-like identity determined. As in most species, the dominant form of RIP mutation in *P. nodorum* was predicted to be CpA→TpA. The high-copy number repeats of P. *nodorum* were predicted to be affected by RIP to varying degrees.

The RIPCAL package also includes a tool to reverse the effects of RIP: deRIP (Hane and Oliver 2010). The deRIP tool identifies RIP-like mutation sites within a repeat family alignment and generates an alignment consensus with modifications that reverse RIP-affected regions to their presumed pre-RIP states. This allows for accurate characterisation of the origin of a RIP-affected sequence via similarity searches. For several repeat families of *P. nodorum*, BLAST searches did not retrieve hits which could be used to identify the role or origin of that repeat. However after deRIP was applied, the origins of 5 previously uncharacterised repeat families were identified. This increased the proportion of characterised repeats in *P. nodorum* from 65% to 88%. Intriguingly, while some of these were heavily RIP-degenerated transposable elements, others were endogenous *P. nodorum* genes or gene clusters (Hane and Oliver 2008; see Section III.C).

V. Pathogenicity

A. Mechanisms of Pathogenicity

The three published genome analyses of the Dothideomycetes *P. nodorum*, *L. maculans* and *M. graminicola* each describe the mechanisms of pathogenicity in each species (Hane et al. 2007; Rouxel et al. 2011; Goodwin et al. 2011). These closely related fungi have evolved remarkably diverse strategies for survival within their respective hosts. Indeed, a recent Ascomycete-wide genome comparison reported that a significant proportion (8–11%) of the gene content of Dothideomycetes is specific to class or genera (Schoch et al. 2009). While no published work has yet compared the gene contents of these three species in detail, individually each genome analysis highlights the major differences between these species.

During wheat infection, *P. nodorum* expresses genes encoding for a battery of cell wall-degrading enzymes (CWDEs) which act predominantly upon xylan and cellulose and also include enzymes degrading carbohydrates and proteins (Hane et al. 2007; IpCho 2010). Membrane transporter proteins, particularly those of carbohydrates, are also important during infection. During the initial stages of infection (3 dpi), *P. nodorum* highly expresses genes involved in ribogenesis and protein localisation. This is followed (5 dpi) by high expression of genes involved in nutrient assimilation and catabolism which continues into the late stages of infection (7–10 dpi). *P. nodorum* infection is inferred by the authors to involve the secretion of pathogenicity effectors and cell wall degrading enzymes into the extracellular space. These induce necrosis and disrupt the cell walls of neighbouring host cells, either causing nutrient leakage or facilitating fungal penetration. Enzymes degrading plant carbohydrates, sugars and proteins are also secreted, producing simple metabolites which are imported into the fungal cell by membrane transport proteins.

Studies of the genome content in *L. maculans* have chiefly focussed on small secreted proteins (SSPs), in an effort to discover and characterise avirulence (AVR) effector genes (Rouxel et al. 2011). AVR genes of *L. maculans* are located within or adjacent to A:T-rich isochores and some of these are purported to generate sequence mutations by exploiting 'RIP leakage' (see Section III.D.3). Among the SSPs, 60% located in A:T-rich isochores and 73% in GC-equilibrated isochores were found to contain RxLR amino acid motifs, which facilitate translocation into the host cell (Kale et al. 2010). Gene functions of genes within A:T-rich isochores were also compared by Rouxel et al. (2011) against those of genes within G:C-equilibrated isochores. A:T-rich isochore genes were comparatively deficient in genes involved in cell communication, transmembrane and vesicle-mediated transport, assembly of cellular components, gene regulation, translation, carbon metabolism, cell growth, sporulation and sexual reproduction. Significant enrichment was observed within A:T-rich isochore genes for functions relating to catabolic processes, response to chemical and biotic stimuli, cell wall metabolic processes, cell cycle processes and microtubule-based processes (Rouxel et al. 2011).

L. maculans and *M. graminicola* both undergo a latent biotrophic phase prior to their necrotrophic phase (Howlett et al. 2001; Palmer and Skinner 2002), in contrast to *P. nodorum* which lacks a biotrophic phase (Oliver 2009). This significant difference in pathogenic lifestyle may be reflected in their respective genome contents. Although this difference has not been addressed in the genome analysis of *L. maculans*, it has been examined in detail for *M. graminicola*. The genome of *M. graminicola* is significantly depleted in CWDE genes targeting cellulose and xylan, carbohydrate binding genes and carbohydrate metabolism genes relative to other phytopathogenic fungi and there are significantly more genes involved in protein metabolism (Goodwin et al. 2011). This lack of CWDE and carbohydrate metabolism is unusual in a cereal pathogen and Goodwin et al. (2011) suggest that this assists *M. graminicola* to evade detection by host defences during its biotrophic phase. Furthermore, Goodwin et al. (2011) proposed that *M. graminicola* does not access nutrients from the plant cytoplasm during its biotrophic phase, instead metabolising proteins within the apoplastic fluid and intracellular space. However, after switching to its necrotrophic phase, *M. graminicola* may express pathogenicity effectors and CWDEs in a similar manner to *P. nodorum*. This 'stealth biotrophy', as coined by Goodwin et al. (2011), suggests that

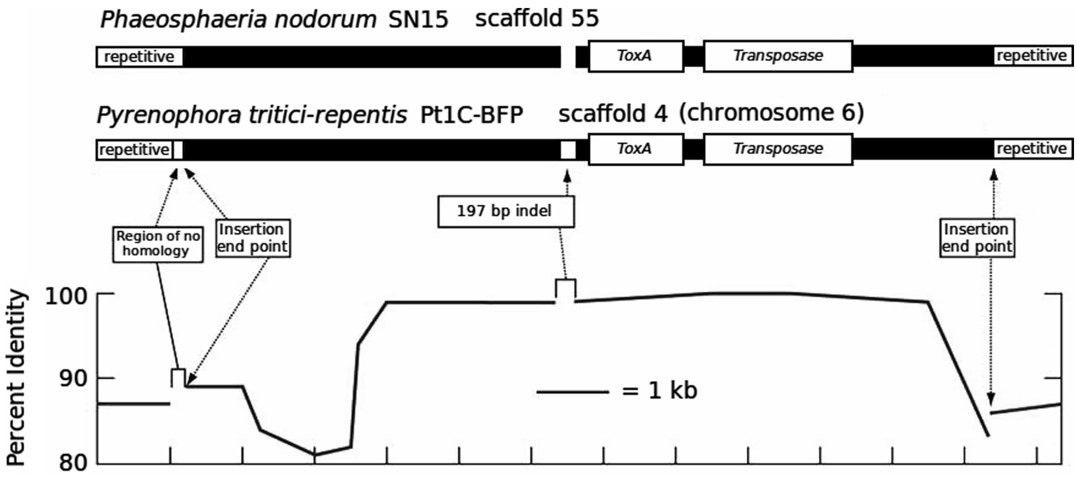


Fig. 9.7. The laterally transferred region containing *ToxA* loci in *P. nodorum* and *P. tritici-repentis*. The sequence of *P. nodorum* scaffold 55 compared with the *P. tritici-repentis* scaffold 4 contains regions of high sequence similarity flanking the ToxA and transposase genes. The *P. nodorum* sequence is flanked by AT-rich sequence that is multiply repeated in the *P. nodorum* genome. The percent similarities between the *P. nodorum* and *P. tritici-repentis* sequences are shown below (adapted from Friesen et al. 2006)

M. graminicola may have originally evolved from an endophytic species.

1. Lateral Gene Transfer

Lateral gene transfer (LGT, syn. horizontal gene transfer, HGT) involves the incorporation of the genetic material of one species into the genome of another. There are a number of reported cases of LGTs in the Dothideomycetes involving pathogenicity-related genes (see Chapters 10, 13 in this volume). The best documented among these is the LGT of the necrotrophic effector gene *ToxA* between *P. nodorum* and *P. tritici-repentis* (Friesen et al. 2006). *ToxA* encodes a 13.2-kDa protein which induces necrosis in *Triticum aestivum* (wheat) cultivars possessing the dominant allele of the susceptibility gene *Tsn1* (Faris et al. 1996). After the sequencing of the *P. nodorum* genome, it was observed that the sequence of the gene *SNOG_16571.1* was 99.7% similar to the previously characterised *ToxA* gene in *P. tritici-repentis*. Friesen et al. (2006) used sequences flanking the *P. nodorum ToxA* sequence to design primer pairs for the purpose of sequencing the corresponding region in *P. tritici-repentis*. An 11-kb region containing the *ToxA* locus was conserved between the two species and contained an hAT family transposase gene adjacent to *ToxA*, which together with *ToxA* was flanked by highly repetitive regions (Fig. 9.7). PCR screening for *ToxA* in globally representative samples revealed a lower incidence of *ToxA* (40%; Stukenbrock and McDonald 2007) in *P. nodorum* isolates compared to 80% in *P. tritici-repentis* (Friesen et al. 2006). Furthermore, *P. nodorum* had been described as a pathogen since at least the year 1889, whereas *P. tritici repentis* was initially identified in 1902 as a saprophyte of grass and in 1928 of wheat. It was only later described as the causal agent of tan spot in 1941. The recent emergence of *P. tritici repentis* as a plant pathogen and rise to prominence as the most important pathogen of crops in Australia, the low level of diversity in isolates of *P. tritici repentis* relative to *P. nodorum*, the presence of a proximal transposase (Fig. 9.7) and flanking repetitive DNA (Fig. 9.7), high sequence conservation at the nucleotide level (Fig. 9.7) and the general lack of sequence conservation between species for other proteinaceous pathogenicity effectors, all suggested that a LGT has occurred between these two species (Friesen et al. 2006). This was the first published evidence of a LGT that was linked to evolution of virulence in a eukaryote.

Besides *ToxA*, several other pathogenicity effector genes of Dothideomycetes have been predicted to have been acquired via LGT events (Oliver and

Solomon 2010). The *Ace1* polyketide synthase/non-ribosomal peptide synthase hybrid gene of *M. grisea* has homologues in several Pezizomycotina species, including *P. nodorum* (Khaldi et al. 2008). Phylogenetic evidence suggests that LGT is the probable explanation for the acquisition of *Ace1* in *P. nodorum*. The T-toxin biosynthesis gene cluster of *C. heterostrophus* is another LGT candidate, residing within 1.2 megabases of sequence which is present in race T but absent in race O (Turgeon and Baker 2007). The 6-methyl-salicylic acid biosynthesis gene clusters of the Dothideomycetes, Eurotiomycetes, Lecanoromycetes and Sordariomycetes are also purported to have been acquired via LGT from the Actinobacteria (Kroken et al. 2003; Schmitt and Lumbsch 2009).

A recent study predicting probable LGTs of prokaryotic genes into fungal genomes found that LGTs were much more prevalent in the Pezizomycotina (filamentous Ascomycetes, Fig. 9.1), relative to non-filamentous Ascomycetes and other fungal phyla (Marcet-Houben and Gabaldon 2010). They speculated that the differences in genome size and gene density between filamentous Pezizomycotina species (larger genomes, lower gene densities) and the non-filamentous Saccharyomycotina (smaller genomes, higher gene densities) may account for their different rates of LGT. However, due to the diversity of lifestyles and genome compositions between Pezizomycotina species, Marcet-Houben and Gabaldon (2010) were unable to determine any further shared characteristics which may predispose Pezizomycotina species towards LGT. *P. nodorum* and *M. fijiensis* represented the Dothideomycetes used in this study, both having 23 predicted prokaryotic LGT events. One notable example of prokaryotic LGT to these Dothideomycetes mentioned by the authors was the transfer of a bacterial catalase from the bacterium *Psuedomonas syringae*. This catalase was predicted to decompose reactive oxygen species and potentially plays a role in the fungal evasion of host plant defences. This gene was also detected in the Leotiomycete *Botrytis cinerea*, and phylogenetic considerations suggest that *B.cinerea* obtained its copy of the gene from the Dothideomycetes.

2. Generation of Diversity via RIP

Avirulence (AVR) effectors are gene products which play a role in the pathogenesis of some Dothideomycete species, most notably *L. maculans* (Fudal et al. 2009; Rouxel et al. 2011; Van de Wouw et al. 2010) and *P. fulvum* (Wulff et al. 2009). AVR effectors are disadvantageous to a pathogen if the host plant possesses resistance (R) genes which are capable of recognising them thereby activating the host plant's defences. Therefore fungi which employ AVR effectors are under selective pressure to mutate and adapt.

The AVR genes *AvrLm6* and *LymCys1* of *L. maculans* have been observed to exhibit RIP-like polymorphism between isolates (Fudal et al. 2009; Van de Wouw et al. 2010). This is unusual as only one copy of both of these genes is found in the *L. maculans* genome and RIP is triggered by the alignment of repetitive DNA (2 or more copies). However RIP acting upon repetitive DNA has been reported to 'leak' into neighbouring non-repetitive regions within a limited range (Irelan et al. 1994). Fudal et al. (2009) proposed that certain AVR genes of *L. maculans* have accumulated RIP mutations from their neighbouring repetitive DNA regions. Thus RIP could be being exploited by the fungus as a means of accelerating the rate of mutation in certain genes (Irelan et al. 1994; Fudal et al. 2009). This was supported by Van de Wouw et al. (2010), who observed a gradient in the frequency of RIP-like mutations between AVR genes of different isolates which is dependent upon the distance from the neighbouring repeat.

Other pathogenicity effectors of the Dothideomycetes are also frequently associated with neighbouring repetitive DNA or A:T-rich regions. These include the effector genes *PtrToxA* of *P. tritici-repentis* and *SnToxA* and *SnTox3* of *P. nodorum* which are next to transposable elements (Friesen et al. 2006; Liu et al. 2009). The dispensable chromosomes of *M. graminicola* (see Section II.B) contain clusters of genes encoding secreted proteins, which are also frequently transposon-associated (Goodwin et al. 2011). Whether association with repetitive DNA also generates effector diversity in these species remains to be determined.

VI. Conclusions

The Dothideomycetes are an ecologically diverse class of Fungi of which many species are pathogens of economically important crops. Significant phytopathogenic species share the ability to rapidly adapt to host plant defences and agricultural practices by mutation or lateral gene transfer. Highly virulent isolates of several species have rapidly emerged and spread over vast distances. While the attributing factor for this remarkable adaptability is currently unknown, available genome data has revealed several complex mechanisms geared towards generating diversity at the genomic level. Dothideomycete genome structure is remarkably plastic, undergoing major rearrangement whilst retaining gene content within the same chromosomes. They are believed to readily acquire new genes via lateral gene transfer, many of which are involved in pathogenicity. Fungal genome defences against transposons may have also been hijacked in order to accelerate mutation rates in pathogenicity effector genes. Significant proportions of Dothideomycete genes are unique to the class or respective genera, indicating adaptations to highly specialised ecological niches. It is therefore unsurprising that each of the Dothideomycetes outlined here employ diverse pathogenicity strategies specifically tailored to their host and lifestyle.

Acknowledgements This work was supported by the Grains Research and Development Corporation, Barton, ACT, Australia.

References

Andrew M, Peever TL and Pryor BM (2009) An expanded multilocus phylogeny does not resolve morphological species within the small-spored *Altemaria* species complex. Mycologia 101:95–109

Aptroot A, Lücking R, Sipman H, Umana L and Chaves J-L (2008) *Pyrenocarpous lichens* with bitunicate asci. A first assessment of the lichen biodiversity inventory in Costa Rica. Bibl Lichenol 97:1–162

Bao Z and Eddy SR (2002) Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res 12(8):1269–1276

Bearchell SJ, Fraaije BA, Shaw MW and Fitt BD (2005) Wheat archive links long-term fungal pathogen population dynamics to air pollution. Proc Natl Acad Sci 102:5438–5442

Bhathal JS, Loughman R and Speijers J (2003) Yield reduction in wheat in relation to leaf disease from yellow (tan) spot and Septoria nodorum blotch. Eur J Plant Pathol 109:435–443

Boehm EW, Mugambi GK, Miller AN, Huhndorf SM, Marincowitz S, Spatafora JW and Schoch CL (2009) A molecular phylogenetic reappraisal of the Hysteriaceae, Mytilinidiaceae and Gloniaceae (Pleosporomycetidae, Dothideomycetes) with keys to world species. Stud Mycol 64:49–83

Bringans S, Hane JK, Casey T, Tan KC, Lipscombe R, Solomon PS and Oliver RP (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

Brun H, Levivier S, Somda I, Ruer D, Renard M and Chevre AM (2000) A field method for evaluating the potential durability of new resistance sources: application to the *Leptosphaeria maculans–Brassica napus* pathosystem. Phytopathology 90(9):961–966

Cannon SB, Sterck L, Rombauts S, Sato S, Cheung F, Gouzy J, Wang X, Mudge J, Vasdewani J, Schiex T, Spannagl M, Monaghan E, Nicholson C, Humphray SJ, Schoof H, Mayer KF, Rogers J, Quetier F, Oldroyd GE, Debelle F, Cook DR, Retzel EF, Roe BA, Town CD, Tabata S, Van de Peer Y and Young ND (2006) Legume genome evolution viewed through the *Medicago truncatula* and *Lotus japonicus* genomes. Proc Natl Acad Sci USA 103(40):14959–14964

Cooley RN and Caten CE (1991) Variation in electrophoretic karyotype between strains of *Septoria nodorum.* Mol Gen Genet 228(1/2):17–23

Curtis MJ and Wolpert TJ (2004) The victorin-induced mitochondrial permeability transition precedes cell shrinkage and biochemical markers of cell death, and shrinkage occurs without loss of membrane integrity. Plant J 38(2):244–259

Dear PH and Cook PR (1993) Happy mapping: linkage mapping using a physical analogue of meiosis. Nucleic Acids Res 21(1):13–20

Edgar RC and Myers EW (2005) PILER: identification and classification of genomic repeats. Bioinformatics 21 [Suppl 1]:i152–il58

Espagne E, Lespinet O, Malagnac F, Da Silva C, Jaillon O, Porcel BM, Couloux A, Aury JM, Segurens B, Poulain J, Anthouard V, Grossetete S, Khalili H, Coppin E, Dequard-Chablat M, Picard M, Contamine V, Arnaise S, Bourdais A, Berteaux-Lecellier V, Gautheret D, de Vries RP, Battaglia E, Coutinho PM, Danchin EG, Henrissat B, Khoury RE, Sainsard-Chanet A, Boivin A, Pinan-Lucarre B, Sellem CH, Debuchy R, Wincker P, Weissenbach J and Silar P (2008) The genome sequence of the model ascomycete fungus *Podospora anserina.* Genome Biol 9(5):R77

Faris JD, Anderson JA, Francl LJ and Jordahl JG (1996) Chromosomal location of a gene conditioning insensitivity in wheat to a necrosis-inducing culture filtrate from *Pyrenophora tritici-repentis.* Phytopathology 86: 459–463

Fitt BDL, Brun H, Barbetti MJ and Rimmer SR (2006) World-wide importance of *Phoma* stem canker (*Leptosphaeria maculans* and *L. biglobosa*) on oilseed rape (*Brassica napus*). Eur J Plant Pathol 114:3–15

Friesen TL, Stukenbrock EH, Liu Z, Meinhardt S, Ling H, Faris JD, Rasmussen JB, Solomon PS, McDonald BA and Oliver RP (2006) Emergence of a new disease as a result of interspecific virulence gene transfer. Nat Genet 38(8):953–956

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

Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, FitzHugh W, Ma LJ, Smirnov S, Purcell S, Rehman B, Elkins T, Engels R, Wang S, Nielsen CB, Butler J, Endrizzi M, Qui D, Ianakiev P, Bell-Pedersen D, Nelson MA, Werner-Washburne M, Selitrennikoff CP, Kinsey JA, Braun EL, Zelter A, Schulte U, Kothe GO, Jedd G, Mewes W, Staben C, Marcotte E, Greenberg D, Roy A, Foley K, Naylor J, Stange-Thomann N, Barrett R, Gnerre S, Kamal M, Kamvysselis M, Mauceli E, Bielke C, Rudd S, Frishman D, Krystofova S, Rasmussen C, Metzenberg RL, Perkins DD, Kroken S, Cogoni C, Macino G, Catcheside D, Li W, Pratt RJ, Osmani SA, DeSouza CP, Glass L, Orbach MJ, Berglund JA, Voelker R, Yarden O, Plamann M, Seiler S, Dunlap J, Radford A, Aramayo R, Natvig DO, Alex LA, Mannhaupt G, Ebbole DJ, Freitag M, Paulsen I, Sachs MS, Lander ES, Nusbaum C and Birren B (2003) The genome sequence of the filamentous fungus *Neurospora crassa*. Nature 422(6934):859–868

Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, Batzoglou S, Lee SI, Basturkmen M, Spevak CC, Clutterbuck J, Kapitonov V, Jurka J, Scazzocchio C, Farman M, Butler J, Purcell S, Harris S, Braus GH, Draht O, Busch S, D'Enfert C, Bouchier C, Goldman GH, Bell-Pedersen D, Griffiths-Jones S, Doonan JH, Yu J, Vienken K, Pain A, Freitag M, Selker EU, Archer DB, Penalva MA, Oakley BR, Momany M, Tanaka T, Kumagai T, Asai K, Machida M, Nierman WC, Denning DW, Caddick M, Hynes M, Paoletti M, Fischer R, Miller B, Dyer P, Sachs MS, Osmani SA and Birren BW (2005) Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. Nature 438(7071):1105–1115

Gardiner DM, Cozijnsen AJ, Wilson LM, Pedras MS and Howlett BJ (2004) The sirodesmin biosynthetic gene cluster of the plant pathogenic fungus *Leptosphaeria maculans*. Mol Microbiol 53(5):1307–1318

Gardiner DM and Howlett BJ (2005) Bioinformatic and expression analysis of the putative gliotoxin biosynthetic gene cluster of *Aspergillus fumigatus*. FEMS Microbiol Lett 248(2):241–248

Gladieux P, Zhang XG, Afoufa-Bastien D, Valdebenito Sanhueza RM, Sbaghi M and Le Cam B (2008) On the origin and spread of the Scab disease of apple: out of central Asia. PLoS One 3(1):e1455

Goodwin SB, Ben M'Barek S, Dhillon B, Wittenberg A, Crane CF, Van der Lee TAJ, Grimwood J, Aerts A, Antoniw J, Bailey A, Bluhm B, Bowler J, Bristow J, Brokstein P, Canto-Canche B, Churchill A, Conde-Ferràez L, Cools H, Coutinho PM, Csukai M, Dehal P, Donzelli B, Foster AJ, Hammond-Kosack K, Hane J, Henrissat B, Kilian A, Koopmann E, Kourmpetis Y, Kuo A, Kuzniar A, Lindquist E, Maliepaard C, Martins N, Mehrabi R, Oliver R, Platt D, Ponomarenko A, Rudd J, Salamov A, Schwarz J, Shapiro H, Stergiopoulos I, Torriani S, Tu H, de Vries R, Wiebenga A, Zwiers L-H, Grigoriev IV and Kema GHJ (2011) Finished genome of *Mycosphaerella graminicola* reveals stealth pathogenesis and dispensome structure. PLoS Genet (in press)

Groenewald M, Groenewald JZ and Crous PW (2005) Distinct species exist within the *Cercospora apii* morphotype. Phytopathology 95:951–959

Hane JK (2011) Bioinformatic genome analysis of the necrotrophic wheat-pathogenic fungus *Phaeosphaeria nodorum*. PhD thesis, Murdoch University, Murdoch

Hane JK, Lowe RGT, Solomon PS, Tan KC, Schoch CL, Spatafora JW, Crous PW, Kodira C, Birren BW, Galagan JE, Torriani SFF, McDonald BA and Oliver RP (2007) Dothideomycete–plant interactions illuminated by genome sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*. Plant Cell 19(11):3347–3368

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

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

Hane JK, Rouxel T, Howlett BJ, Kema GHJ, Goodwin SB, Oliver RP (2011) A novel mode of chromosomal evolution called Mesosynteny that is peculiar to filamentous Ascomycete fungi. Genome Biology (in press).

Hardwick NV, Jones DR and Slough JE (2001) Factors affecting diseases of winter wheat in England and Wales, 1989–98. Plant Pathol 50:453–462

Hashim I (1998) Disease survey. IRRDB, Kuala Lumpur

Hooker AL, Smith DR, Lim SM and Musson MD (1970) Physiological races of *Helminthosporium maydis* and disease resistance. Plant Dis Rep 54:1109–1110

Howlett BJ, Idnurm A and Pedras MSC (2001) *Leptosphaeria maculans*, the causal agent of blackleg disease of Brassicas. Fungal Genet Biol 33:1–14

Idnurm A and Howlett BJ (2003) Analysis of loss of pathogenicity mutants reveals that repeat-induced point mutations can occur in the Dothideomycete *Leptosphaeria maculans*. Fungal Genet Biol 39(1):31–37

IpCho SV S (2010) Pathogenicity of *Stagonospora nodorum*. PhD thesis, Murdoch University, Murdoch

Irelan JT, Hagemann AT and Selker EU (1994) High frequency repeat-induced point mutation (RIP) is not associated with efficient recombination in *Neurospora*. Genetics 138(4):1093–1103

Kale SD, Gu B, Capelluto DGS, Dou D, Cronin A, Arredondo FD, Feldman E, Fudal I, Rouxel T, Lawrence CB, Shan W and Tyler BM (2010) External phosphatidylinositol-3-phosphate mediates host cell entry by eukaryotic pathogen effectors. Cell 142(2): 284–295

Khaldi N, Collemare J, Lebrun M-H and Wolfe KH (2008) Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi. Genome Biol 9:R18

Kohn M, Kehrer-Sawatzki H, Vogel W, Graves JA and Hameister H (2004) Wide genome comparisons reveal the origins of the human X chromosome. Trends Genet 20(12):598–603

Kroken S, Glass NL, Taylor JW, Yoder OC and Turgeon BG (2003) Phylogenomic analysis of type I polyketide synthase genes in pathogenic and saprobic ascomycetes. Proc Natl Acad Sci USA 100(26): 15670–15675

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

Levings CS, 3rd and Siedow JN (1992) Molecular basis of disease susceptibility in the Texas cytoplasm of maize. Plant Mol Biol 19(1):135–147

Li CX and Cowling WA (2003) Identification of a single dominant allele for resistance to blackleg in *Brassica napus* "Surpass 400". Plant Breed 122:485–488

Liu Z, Faris JD, Oliver RP, Tan KC, Solomon PS, McDonald MC, McDonald BA, Nunez A, Lu S, Rasmussen JB and Friesen TL (2009) SnTox3 acts in effector triggered susceptibility to induce disease on wheat carrying the *Snn3* gene. PLoS Pathog 5(9):e1000581

Luttrell ES (1955) The ascostromatic Ascomycetes. Mycologia 47:511–532

Machida M, Asai K, Sano M, Tanaka T, Kumagai T, Terai G, Kusumoto K, Arima T, Akita O, Kashiwagi Y, Abe K, Gomi K, Horiuchi H, Kitamoto K, Kobayashi T, Takeuchi M, Denning DW, Galagan JE, Nierman WC, Yu J, Archer DB, Bennett JW, Bhatnagar D, Cleveland TE, Fedorova ND, Gotoh O, Horikawa H, Hosoyama A, Ichinomiya M, Igarashi R, Iwashita K, Juvvadi PR, Kato M, Kato Y, Kin T, Kokubun A, Maeda H, Maeyama N, Maruyama J, Nagasaki H, Nakajima T, Oda K, Okada K, Paulsen I, Sakamoto K, Sawano T, Takahashi M, Takase K, Terabayashi Y, Wortman JR, Yamada O, Yamagata Y, Anazawa H, Hata Y, Koide Y, Komori T, Koyama Y, Minetoki T, Suharnan S, Tanaka A, Isono K, Kuhara S, Ogasawara N and Kikuchi H (2005) Genome sequencing and analysis of *Aspergillus oryzae.* Nature 438(7071): 1157–1161

Malkus A, Song Q, Cregan P, Arseniuk E and Ueng PP (2009) Genetic linkage map of *Phaeosphaeria nodorum*, the causal agent of stagonospora nodorum blotch disease of wheat. Eur J Plant Pathol:1–10

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

McLysaght A, Enright AJ, Skrabanek L and Wolfe KH (2000) Estimation of synteny conservation and genome compaction between pufferfish (*Fugu*) and human. Yeast 17(1):22–36

Murray GM and Brennan JP (2009) Estimating disease losses to the Australian wheat industry. Aust Plant Pathol 38:558–570

Nowrousian M, Stajich JE, Chu M, Engh I, Espagne E, Halliday K, Kamerewerd J, Kempken F, Knab B, Kuo HC, Osiewacz HD, Poggeler S, Read ND, Seiler S, Smith KM, Zickler D, Kuck U and Freitag M (2010) *De novo* assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. PLoS Genet 6(4): e1000891

Oliver R (2009) Plant breeding for disease resistance in the age of effectors. Phytoparasitica 37:1–5

Oliver RP (1992) A model system for the study of plant–fungal interactions: Tomato leaf mold caused by *Cladosporium fulvum*.In: Verma DPS (ed) Molecular signals in plant–microbe communications. CRC, Boca Raton, pp 97–106

Oliver RP and Solomon PS (2010) New developments in pathogenicity and virulence of necrotrophs. Curr Opin Plant Biol 13:1–5

Palmer C-L and Skinner W (2002) *Mycosphaerella graminicola*: latent infection, crop devastation and genomics. Mol Plant Pathol 3(2):63–70

Pel HJ, de Winde JH, Archer DB, Dyer PS, Hofmann G, Schaap PJ, Turner G, de Vries RP, Albang R, Albermann K, Andersen MR, Bendtsen JD, Benen JA, van den Berg M, Breestraat S, Caddick MX, Contreras R, Cornell M, Coutinho PM, Danchin EG, Debets AJ, Dekker P, van Dijck PW, van Dijk A, Dijkhuizen L, Driessen AJ, d'Enfert C, Geysens S, Goosen C, Groot GS, de Groot PW, Guillemette T, Henrissat B, Herweijer M, van den Hombergh JP, van den Hondel CA, van der Heijden RT, van der Kaaij RM, Klis FM, Kools HJ, Kubicek CP, van Kuyk PA, Lauber J, Lu X, van der Maarel MJ, Meulenberg R, Menke H, Mortimer MA, Nielsen J, Oliver SG, Olsthoorn M, Pal K, van Peij NN, Ram AF, Rinas U, Roubos JA, Sagt CM, Schmoll M, Sun J, Ussery D, Varga J, Vervecken W, van de Vondervoort PJ, Wedler H, Wosten HA, Zeng AP, van Ooyen AJ, Visser J and Stam H (2007) Genome sequencing and analysis of the versatile cell factory *Aspergillus niger* CBS 513.88. Nat Biotechnol 25(2):221–231

Pennacchio LA (2003) Insights from human/mouse genome comparisons. Mamm Genome 14(7):429–436

Phan HT, Ellwood SR, Hane JK, Ford R, Materne M and Oliver RP (2007) Extensive macrosynteny between *Medicago truncatula* and *Lens culinaris* ssp. culinaris. Theor Appl Genet 114(3):549–558

Ploetz RC (2001) Black sigatoka of banana. The Plant Health Instructor, New York

Price AL, Jones NC and Pevzner PA (2005) *De novo* identification of repeat families in large genomes. Bioinformatics 21 [Suppl 1]:i351–i358

Quesneville H, Nouaud D and Anxolabehere D (2005) Recurrent recruitment of the THAP DNA-binding domain and molecular domestication of the P-transposable element. Mol Biol Evol 22(3):741–746

Rouxel T, Grandaubert J, Hane J, Hoede C, van de Wouw A, Couloux A, Dominguez V, Anthouard V, Bally P, Bourras S, Cozijnsen A, Ciuffetti L, Dimaghani A, Duret L, Fudal I, Goodwin S, Gout L, Glaser N, Kema G, Lapalu N, Lawrence C, May K, Meyer M, Ollivier B, Poulain J, Turgeon G, Tyler BM, Vincent D, Weissenbach J, Amselem J, Balesdent M-H, Howlett BJ, Oliver R, Quesneville H and Wincker P (2011) The patchwork genome of *Leptosphaeria maculans*: effector diversification driven by location within RIP-affected isochores. Nat Commun 2:202

Rouxel T, Penaud A, Pinochet X, Brun H, Gout L, Delourne R, Schmit J and Bealesdent MH (2003) A 10-year survey of populations of *Leptosphaeria maculans* in

France indicates a rapid adaptation towards to *Rlm1* resistance gene of oilseed rape. Eur J Plant Pathol 109:871–881

Schmitt I and Lumbsch HT (2009) Ancient horizontal gene transfer from bacteria enhances biosynthetic capabilities of fungi. PLoS One 4(2):e4437

Schoch CL, Crous PW, Groenewald JZ, Boehm EW, Burgess TI, de Gruyter J, de Hoog GS, Dixon LJ, Grube M, Gueidan C, Harada Y, Hatakeyama S, Hirayama K, Hosoya T, Huhndorf SM, Hyde KD, Jones EB, Kohlmeyer J, Kruys A, Li YM, Lucking R, Lumbsch HT, Marvanova L, Mbatchou JS, McVay AH, Miller AN, Mugambi GK, Muggia L, Nelsen MP, Nelson P, Owensby CA, Phillips AJ, Phongpaichit S, Pointing SB, Pujade-Renaud V, Raja HA, Plata ER, Robbertse B, Ruibal C, Sakayaroj J, Sano T, Selbmann L, Shearer CA, Shirouzu T, Slippers B, Suetrong S, Tanaka K, Volkmann-Kohlmeyer B, Wingfield MJ, Wood AR, Woudenberg JH, Yonezawa H, Zhang Y and Spatafora JW (2009) A class-wide phylogenetic assessment of Dothideomycetes. Stud Mycol 64:1–15

Schoch CL, Shoemaker RA, Seifert KA, Hambleton S, Spatafora JW and Crous PW (2006) A multigene phylogeny of the Dothideomycetes using four nuclear loci. Mycologia 98:1043–1054

Schwartz DC, Li X, Hernandez LI, Ramnarain SP, Huff EJ and Wang YK (1993) Ordered restriction maps of *Saccharomyces cerevisiae* chromosomes constructed by optical mapping. Science 262(5130):110–114

Selker EU, Cambareri EB, Jensen BC and Haack KR (1987) Rearrangement of duplicated DNA in specialized cells of *Neurospora.* Cell 51(5):741–752

Shultz JL, Ray JD and Lightfoot DA (2007) A sequence based synteny map between soybean and *Arabidopsis thaliana.* BMC Genomics 8:8

Silva WPK, Wijesundera RIC, Karunanayake EH, Jayasinghe CK and Priyanka UMS (2000) New hosts of *Corynespora cassiicola* in Sri Lanka. Plant Dis 84:202

Smith DR and White DG (1988) Diseases of corn. In: Sprague GF and Dudley JW (eds) Corn and corn improvement. Agronomy series 18. ASA/CSSA/SSS, Madison, pp 701–766

Solomon PS, Lowe RG T, Tan KC, Waters ODC and Oliver RP (2006) *Stagonospora nodorum*: cause of stagonospora nodorum blotch of wheat. Mol Plant Pathol 7:147–156

Stukenbrock EH and McDonald BA (2007) Geographical variation and positive diversifying selection in the host-specific toxin SnToxA. Mol Plant Pathol 8(3): 321–332

Thomma BPHJ (2003) *Alternaria* spp.: from general saprophyte to specific parasite. Mol Plant Pathol 4:225–235

Torriani SF, Goodwin SB, Kema GH, Pangilinan JL and McDonald BA (2008) Intraspecific comparison and annotation of two complete mitochondrial genome sequences from the plant pathogenic fungus *Mycosphaerella graminicola*. Fungal Genet Biol 45(5): 628–637

Turgeon BG and Baker SE (2007) Genetic and genomic dissection of the *Cochliobolus heterostrophus Tox1* locus controlling biosynthesis of the polyketide virulence factor T-toxin. Adv Genet 57:219–261

Tzeng T-H, Lyngholm L, Ford C and Bronson C (1992) A restriction fragment length polymorphism map and electrophoretic karyotype of the fungal maize pathogen *Cochliobolus heterostrophus.* Genetics 130 (1):81–96

Ullstrup AJ (1972) The impacts of the southern corn leaf blight epidemics of 1970–71. Annu Rev Phytopathol 10:37–50

Van de Wouw AP, Cozijnsen AJ, Hane JK, Brunner PC, McDonald BA, Oliver RP and Howlett BJ (2010) Evolution of linked avirulence effectors in *Leptosphaeria maculans* is affected by genomic environment and exposure to resistance genes in host plants. PLoS Pathogens 6(11):e1001180

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

Wei J, Lui K, Chen J, Luo P and Lee-Stadelmann OY (1988) Pathological and physiological identification of race C of *Bipolaris maydis* in China. Phytopathology 78: 550–554

Wulff BBH, Chakrabarti A and Jones DA (2009) Recognitional Specificity and Evolution in the Tomato–*Cladosporium* fulvum Pathosystem. Mol Plant Microbe Interact 22(10):1191–1202

Zhang Y, Schoch CL, Fournier J, Crous PW, de Gruyter J, Woudenberg JH, Hirayama K, Tanaka K, Pointing SB, Spatafora JW and Hyde KD (2009) Multi-locus phylogeny of Pleosporales: a taxonomic, ecological and evolutionary re-evaluation. Stud Mycol 64:85–1022

Chapter 11: Conclusion

The immediate impact of this Ph. D. project has been advancing the understanding of the necrotrophic wheat pathogen *Phaeosphaeria nodorum* and providing resources for other researchers to build upon that knowledge. The *P. nodorum* genome has been established as a high quality bioinformatic resource upon which additional genomic, transcriptomic, proteomic and metabolomic supporting data can be built upon. A significant proportion of the project involved the refinement of gene and repeat regions and development of genome curation and analysis techniques (Chapters 2, 3, 4 and 7). The *P. nodorum* genome has illuminated the roles that repetitive DNA, repeat-induced point mutation (RIP) and genome structure play in pathogenicity. The analysis of RIP in *P. nodorum* and *Leptosphaeria maculans* provided insight into the history of transposon-invasion and its dramatic effect upon fungal genome evolution (Chapters 4 and 6). Analysis of RIP in *L. maculans* suggests that it has a role in accelerating the rate of mutation in repeat-associated pathogenicity effector genes (Chapter 5). Once established, the *P. nodorum* genome assembly was also used as a point of reference for comparative genomics with other Dothideomycete species (Chapters 6, 8 and 9). This led to the discovery of a novel mode of genome conservation called 'mesosynteny' (Chapter 8) between Pezizomycotina species and particularly prominent between Dothideomycetes. The body of work from Chapters 2 through 9 was summarised in chapter 10.

The *P. nodorum* genome was rapidly accepted and utilised by researchers within the fungal genomics and pathology communities. Mechanisms of pathogenicity in *P. nodorum* have been interrogated at the transcriptomic level, using a microarray specific to the annotated genes of *P. nodorum* [1]. The *P. nodorum* genome has represented the class Dothideomycetes in several phylogenetic [2-4] and comparative genomics studies [5] (Chapters 6, 8 and 9) and has also aided in the characterisation of genes via gene knock-out [6-10].

The sequence region encoding a newly identified effector gene *Tox3* [11] was also identified with the aid of the *P. nodorum* genome. With the availability of a whole-genome sequence for *P. nodorum*, it would be logical to use this resource to rapidly identify additional effector genes. But research in this area has been complicated by a number of issues. Known effector genes tend to encode cysteine-rich secreted proteins of < 100-150 amino acids in length [12, 13]. Gene prediction algorithms used early in the project, such as UNVEIL [14], predict a high number of genes of shorter length. As pathogenicity effectors are not well characterised and represented in online databases it was not possible to identify them by sequence similarity searches. Thus, legitimate effector genes were obscured by background levels of false gene predictions. Additionally, many known and candidate effector genes are associated with repetitive DNA and regions of high A:T nucleotide content (Chapters 5, 6 and 8) [15, 16]. Gene prediction algorithms are usually trained on a set of EST-supported or highly conserved genes, which are expected to be encoded within DNA regions of equilibrated G:C content (Chapter 6). Therefore, candidate effector genes were less likely to be predicted based on these training models built from genes residing in G:C equilibrated DNA regions.

Candidate effector genes could be identified using transcriptomic or proteomic methods. Effector genes are likely to be involved at the early stages of infection, where they assist fungal penetration or induce necrosis [12, 13]. A subset of transcriptome sequences derived from *in planta* cDNA libraries obtained in the early stages of infection would be expected to correspond to effector genes, thus providing supporting evidence for a gene to be considered as an effector candidate. Unfortunately, fungal biomass during the early stages of infection is very low and obtaining sufficient messenger RNA for the generation of cDNA libraries is difficult. The *in-planta* cDNA library presented in Chapter 2 was extracted from the late stages of infection when fungal biomass was greater for this reason [17]. Microarrays can be designed to validate candidate effector genes, however these are relatively costly and rely on accurate gene annotation [18]. The high coverage of transcriptome next-generation sequencing methods such as RNA-seq would be likely to provide the depth of coverage necessary to detect fungal transcripts in early infection. This technology however only became available very late in the course of this project.

Proteomic techniques could also be employed to identify candidate effector genes. Proteomics and proteogenomics involve matching measured molecular weights of peptides to a database predicted molecular weights. This database is derived from predicted proteins in the case of proteomics, and translated open-reading frames in the case of proteogenomics. Prior to either method, it is possible to selectively capture proteins based on their physical properties (Chapter 7). To target effector genes for example, one could isolate small secreted proteins of low molecular weight and toxic activity. The work presented in Chapter 7 on proteogenomic gene discovery and verification was targeted to intra-cellular proteins and thus not able to provide supporting data for candidate effector genes. Nevertheless this work served to lay the technical groundwork for further proteomic [19, 20] and future proteogenomic studies.

A major finding produced by this Ph. D. project was the discovery of mesosynteny. The *P. nodorum* genome was the first Dothideomycete and one of the first

Ascomycete plant pathogens to be sequenced (Chapter 10). Early in the project (Chapter 2) one of the major obstacles to genomic analysis of *P. nodorum* was the lack of appropriate species to which the *P. nodorum* genome could be compared. Later, *P. nodorum* was followed by 6 additional Dothideomycete genomes (*L. maculans*, *Mycosphaerella graminicola*, *M. fijiensis*, *Alternaria brassicicola*, *Pyrenophora tritici-repentis* and *Cochliobolus heterostrophus*), two of which (*L. maculans* and *M. graminicola*) have recently published genome surveys (Chapters 6 and 9). The availability and accuracy of the *P. nodorum* genomic resources have greatly aided research efforts on these related plant pathogens. Genomic comparisons between the sequenced Dothideomycetes revealed a novel mechanism of chromosome sequence conservation, which we have called 'mesosynteny'. A wider comparison across the fungal kingdom revealed that mesosynteny appears to be restricted to the sub-phylum Pezizomycotina (Chapter 8).

Mesosynteny significantly impacts upon the field of fungal genomics in four ways. Firstly, with appropriate species to compare against, mesosynteny could be used as an *in silico* alternative to genome finishing techniques such as physical and genetic mapping [21, 22], haplotype mapping [23] or optical mapping [24, 25] . Mesosyntenic finishing was applied *post facto* to two species, *L. maculans* (Chapter 6) and *M. graminicola* (Chapter 8), and successfully predicted results obtained by traditional genome finishing methods.

Secondly, mesosyntenic rearrangements involve a reordering of the sequence content of a chromosome, without exchange of sequence content between chromosomes. If mesosyntenic rearrangements are a by-product of multiple homologous recombination events (Chapter 8), these might only occur during meiosis. This raises questions as to the rate of mesosyntenic rearrangement and how accurately the currently sequenced genomes represent fungal populations regularly undergoing sexual reproduction. If fungal genomes are constantly changing with each round of meiosis, a sequenced genome would only represent a brief 'snapshot' in time.

Thirdly, mesosynteny may alter the way fungal bioinformaticists search for clusters of conserved microsyntenic genes. Microsynteny is the conservation of gene order over a run of a few genes (i.e. 2-10). Genes within microsyntenic clusters have been observed to share related biological functions [2, 26-30]). It is presumed that the retention of the order of genes with related functions would have been selected. This is possibly because the transcriptional expression of these genes is more easily co-regulated when they are in close proximity. Recent work in *Schizosaccharomyces pombe* has demonstrated that its chromosomes collectively form a three-dimensional structure within the nucleus [31]. *S. pombe* genes with related functions are found located in physical proximity to one another, not just on the same length of DNA but also in three-dimensional space. This observation suggests that gene clusters may be co-regulated if they share similar physical accessibilities to transcription-related complexes across the contours of the chromosome ultra-structure. Given the extent of mesosyntenic rearrangement and lack of microsynteny that has been observed (Chapter 8), it is possible while a group of functionally-related genes are unlikely to be microsyntenic they may still be located nearby to one another. If physical proximity is a major factor in transcriptional regulation then many potentially important "mesosyntenic" gene clusters have thus far been overlooked in bioinformatic surveys of the Pezizomycotina. Indeed most documented microsyntenic clusters do not display perfect microsynteny and contain several rearrangements in gene order and orientation (Chapters 2 and 8).

Finally, in addition to mesosynteny, species of the Pezizomycotina display a number of common attributes, most of which appear to be exaggerated within this taxon - this includes: a preponderance of species with supernumerary or dispensable chromosomes [32, 33]; a distinctive profile of RIP-like mutation [34]; a higher frequency of lateral gene transfer (LGT) [5, 35]; the use of necrotrophic effectors in plant pathogenicity [13] ; reports of LGT of pathogenicity effectors [36, 37]; a relatively high rate of sequence mutation [38]; and a history of sudden, rapid and highly destructive plant disease epidemics [3, 36, 39] . Due to the diverse genomic and lifestyle differences within this taxon, previous studies have been unable to identify a common factor between the Pezizomycotina which would account for these seemingly unrelated attributes. I propose that mesosynteny may be the "missing link" which holds the key to understanding this large and significant group of fungi.

As RIP and mesosynteny both appear to be exlusive to the Pezizomycotina, RIP is a prime suspect to be a cause of mesosynteny. RIP targets large regions of DNA with similar sequence, irreversibly introducing mutations at random and thus reducing the redundancy of sequences within the genome [40]. This is expected to reduce the number of potential sites of homologous recombination [41] between non-sister chromosomes during mitosis (Figure 1). In contrast, it would be expected that there would be no barrier to homologous recombination occurring between meiotically-duplicated sister chromosomes. This may explain the phenomenon of mesosyntenic rearrangement, which is characterised by a higher rate of intra-chromosomal rearrangement (within the same chromosome) than inter-chromosomal rearrangement (between different chromosomes) (Chapter 8).

Mesosyntenic rearrangements quasi-randomly alter the order of genes within the same chromosome. Assuming that proximal gene co-regulation as observed in *S. pombe* extends to all fungi, this might mean that rearranging certain elements of the physical genome structure would be deleterious to the fungus in question. Therefore it would be expected that some combinations of gene orders will be more favourable for organism fitness than others. If mesosyntenic rearrangement is meiotically-associated then, after multiple rounds of sexual reproduction progeny would be expected to exhibit a range of different mesosyntenic rearrangements in gene orders. Under the influence of natural selection, the sexual progeny possessing gene orders that allow for the regulated expression of phenotypes optimal for survival in those conditions would be expected to out-compete the others. Therefore it is possible after multiple rounds of meiosis that selection pressures could influence the relative genomic locations of certain groups of genes. For example, it may be advantageous for the expression of genes involved in synthesis of secondary metabolites such as polyketides and non-ribosomal peptides to be co-regulated. In this case it would be expected that progeny in which these genes are proximally clustered would be more successful and become more numerous after selection. The net effect of multiple selection pressures acting upon multiple groups of genes could mean that genomes exhibiting mesosynteny may be divided into regions containing clusters of ‘useful’ genes and other regions containing ‘non-essential’ genes.

The potential consequences of this are two-fold. Genomes of Pezizomycotina species might eventually become ‘streamlined’ over time if non-essential sequences were to cluster together and were subsequently removed from the genome. Chromosome breakage [42] and transposon-mediated excision [43] are possible mechanisms of sequence removal. This potential for loss of non-essential sequences could explain the dispensable chromosomes that are observed in some species (Chapter 8) [32, 33]. Clusters of 'useful'genes, such as those related to pathogenicity or survivability, might also be selected for among a population. The removal of such regions from a genome might result in a discrete and genetically mobile sequence that is amenable to LGT. This is a possible explanation for the high frequency of LGT within this taxon [5, 35], as well as reports of LGT of pathogenicity effector genes [36, 37].

In summary, the analysis of the *P. nodorum* genome has provided a significant resource from which major advances in understanding of necrotrophic pathogens have been made and can continue to be built upon (Chapters 2, 6 and 9). Other significant outputs of this project were the development of open-source software for the analysis of repeat-induced point mutation in fungal repetitive DNA (Chapters 3 and 4) and in the analysis of proteogenomic data (www.sourceforge.net/projects/cdsmapper, Chapter 7). It has also enabled the discovery of mesosynteny (Chapter 8), a novel pattern of genome conservation, which may open up exciting areas of inquiry in fungal genomics. Considerations on the potential mode-of-action and influence of mesosynteny outlined in this chapter are speculative. But given the potential impact on the understanding of phytopathogen genomics and adaptability, investigation of mesosynteny and its genomic consequences warrants further research investment.

References

1. IpCho SVS, Hane JK, Ahren D, Friesen TL, Solomon PS, Oliver RP: **Comprehensive transcriptomic analysis of the wheat pathogen *Stagonospora nodorum*; gene model validation, effector candidate genes and intensive host regulation of metabolism.** *in press* 2010.
2. Hane JK, Lowe RG, Solomon PS, Tan KC, Schoch CL, Spatafora JW, Crous PW, Kodira C, Birren BW, Galagan JE *et al*: **Dothideomycete plant interactions illuminated by genome**

sequencing and EST analysis of the wheat pathogen *Stagonospora nodorum*. *The Plant cell* 2007, **19**(11):3347-3368.

3. Schoch CL, Crous PW, Groenewald JZ, Boehm EW, Burgess TI, de Gruyter J, de Hoog GS, Dixon LJ, Grube M, Gueidan C *et al*: **A class-wide phylogenetic assessment of Dothideomycetes**. *Studies in mycology* 2009, **64**:1-15S10.
4. Spatafora JW, Sung GH, Johnson D, Hesse C, O'Rourke B, Serdani M, Spotts R, Lutzoni F, Hofstetter V, Miadlikowska J *et al*: **A five-gene phylogeny of Pezizomycotina**. *Mycologia* 2006, **98**(6):1018-1028.
5. Marcet-Houben M, Gabaldon T: **Acquisition of prokaryotic genes by fungal genomes**. *Trends Genet* 2010, **26**(1):5-8.
6. IpCho SV, Tan KC, Koh G, Gummer J, Oliver RP, Trengove RD, Solomon PS: **The transcription factor *StuA* regulates central carbon metabolism, mycotoxin production, and effector gene expression in the wheat pathogen *Stagonospora nodorum***. *Eukaryotic cell* 2010, **9**(7):1100-1108.
7. Tan KC, Heazlewood JL, Millar AH, Thomson G, Oliver RP, Solomon PS: **A signaling-regulated, short-chain dehydrogenase of *Stagonospora nodorum* regulates asexual development**. *Eukaryotic cell* 2008, **7**(11):1916-1929.
8. Lowe RG, Lord M, Rybak K, Trengove RD, Oliver RP, Solomon PS: **Trehalose biosynthesis is involved in sporulation of *Stagonospora nodorum***. *Fungal Genet Biol* 2009, **46**(5):381-389.
9. Li W, Csukai M, Corran A, Crowley P, Solomon PS, Oliver RP: **Malayamycin, a new streptomycete antifungal compound, specifically inhibits sporulation of *Stagonospora nodorum* (Berk) Castell and Germano, the cause of wheat glume blotch disease**. *Pest Manag Sci* 2008, **64**(12):1294-1302.
10. Lowe RG, Lord M, Rybak K, Trengove RD, Oliver RP, Solomon PS: **A metabolomic approach to dissecting osmotic stress in the wheat pathogen *Stagonospora nodorum***. *Fungal Genet Biol* 2008, **45**(11):1479-1486.
11. Liu Z, Faris JD, Oliver RP, Tan KC, Solomon PS, McDonald MC, McDonald BA, Nunez A, Lu S, Rasmussen JB *et al*: **SnTox3 acts in effector triggered susceptibility to induce disease on wheat carrying the *Snn3* gene**. *PLoS pathogens* 2009, **5**(9):e1000581.
12. Friesen TL, Faris JD, Solomon PS, Oliver RP: **Host-specific toxins: effectors of necrotrophic pathogenicity**. *Cellular microbiology* 2008, **10**(7):1421-1428.
13. Oliver RP, Solomon PS: **New developments in pathogenicity and virulence of necrotrophs**. *Current opinion in plant biology* 2010, **13**(4):415-419.
14. Majoros WH, Pertea M, Antonescu C, Salzberg SL: **GlimmerM, Exonomy and Unveil: three *ab initio* eukaryotic genefinders**. *Nucleic acids research* 2003, **31**(13):3601-3604.
15. Fudal I, Ross S, Brun H, Besnard AL, Ermel M, Kuhn ML, Balesdent MH, Rouxel T: **Repeat-induced point mutation (RIP) as an alternative mechanism of evolution toward virulence in *Leptosphaeria maculans***. *Mol Plant Microbe Interact* 2009, **22**(8):932-941.
16. Van de Wouw AP, Cozijnsen AJ, Hane JK, Brunner PC, McDonald BA, Oliver RP, Howlett BJ: **Evolution of linked avirulence effectors in Leptosphaeria maculans is affected by genomic environment and exposure to resistance genes in host plants**. *PLoS pathogens* 2010, **6**(11):e1001180.
17. Lowe RGT: **Sporulation of *Stagonospra nodorum*.** Perth, Australia: Murdoch University; 2006.
18. Stoughton RB: **Applications of DNA microarrays in biology**. *Annual review of biochemistry* 2005, **74**:53-82.
19. Casey T, Solomon PS, Bringans S, Tan KC, Oliver RP, Lipscombe R: **Quantitative proteomic analysis of G-protein signalling in Stagonospora nodorum using isobaric tags for relative and absolute quantification**. *Proteomics* 2010, **10**(1):38-47.
20. Tan KC, Heazlewood JL, Millar AH, Oliver RP, Solomon PS: **Proteomic identification of extracellular proteins regulated by the Gna1 Galpha subunit in Stagonospora nodorum**. *Mycological research* 2009, **113**(5):523-531.
21. Chibana H, Oka N, Nakayama H, Aoyama T, Magee BB, Magee PT, Mikami Y: **Sequence finishing and gene mapping for Candida albicans chromosome 7 and syntenic analysis against the Saccharomyces cerevisiae genome**. *Genetics* 2005, **170**(4):1525-1537.
22. Stajich JE, Wilke SK, Ahren D, Au CH, Birren BW, Borodovsky M, Burns C, Canback B, Casselton LA, Cheng CK *et al*: **Insights into evolution of multicellular fungi from the assembled chromosomes of the mushroom Coprinopsis cinerea (Coprinus cinereus)**.

Proceedings of the National Academy of Sciences of the United States of America 2010, **107**(26):11889-11894.

23. Gale LR, Bryant JD, Calvo S, Giese H, Katan T, O'Donnell K, Suga H, Taga M, Usgaard TR, Ward TJ *et al*: **Chromosome complement of the fungal plant pathogen *Fusarium graminearum* based on genetic and physical mapping and cytological observations**. *Genetics* 2005, **171**(3):985-1001.
24. Ma LJ, Ibrahim AS, Skory C, Grabherr MG, Burger G, Butler M, Elias M, Idnurm A, Lang BF, Sone T *et al*: **Genomic analysis of the basal lineage fungus *Rhizopus oryzae* reveals a whole-genome duplication**. *PLoS genetics* 2009, **5**(7):e1000549.
25. Valouev A, Zhang Y, Schwartz DC, Waterman MS: **Refinement of optical map assemblies**. *Bioinformatics (Oxford, England)* 2006, **22**(10):1217-1224.
26. Gardiner DM, Cozijnsen AJ, Wilson LM, Pedras MS, Howlett BJ: **The sirodesmin biosynthetic gene cluster of the plant pathogenic fungus *Leptosphaeria maculans***. *Molecular microbiology* 2004, **53**(5):1307-1318.
27. Tudzynski B, Holter K: **Gibberellin biosynthetic pathway in *Gibberella fujikuroi*: evidence for a gene cluster**. *Fungal Genet Biol* 1998, **25**(3):157-170.
28. Tully RE, van Berkum P, Lovins KW, Keister DL: **Identification and sequencing of a cytochrome P450 gene cluster from *Bradyrhizobium japonicum***. *Biochimica et biophysica acta* 1998, **1398**(3):243-255.
29. Hamer L, Pan H, Adachi K, Orbach MJ, Page A, Ramamurthy L, Woessner JP: **Regions of microsynteny in *Magnaporthe grisea* and *Neurospora crassa***. *Fungal Genet Biol* 2001, **33**(2):137-143.
30. Kutil BL, Liu G, Vrebalov J, Wilkinson HH: **Contig assembly and microsynteny analysis using a bacterial artificial chromosome library for *Epichloe festucae*, a mutualistic fungal endophyte of grasses**. *Fungal Genet Biol* 2004, **41**(1):23-32.
31. Tanizawa H, Iwasaki O, Tanaka A, Capizzi JR, Wickramasinghe P, Lee M, Fu Z, Noma K: **Mapping of long-range associations throughout the fission yeast genome reveals global genome organization linked to transcriptional regulation**. *Nucleic acids research* 2010, **38**(22):8164-8177.
32. Coleman JJ, Rounsley SD, Rodriguez-Carres M, Kuo A, Wasmann CC, Grimwood J, Schmutz J, Taga M, White GJ, Zhou S *et al*: **The genome of *Nectria haematococca*: contribution of supernumerary chromosomes to gene expansion**. *PLoS genetics* 2009, **5**(8):e1000618.
33. Ma LJ, van der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, Di Pietro A, Dufresne M, Freitag M, Grabherr M, Henrissat B *et al*: **Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium***. *Nature* 2010, **464**(7287):367-373.
34. Clutterbuck AJ: **MATE transposable elements in *Aspergillus nidulans*: evidence of repeat-induced point mutation**. *Fungal Genet Biol* 2004, **41**(3):308-316.
35. Oliver RP, Solomon PS: **Recent fungal diseases of crop plants: is lateral gene transfer a common theme?** *Mol Plant Microbe Interact* 2008, **21**(3):287-293.
36. 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**. *Nature genetics* 2006, **38**(8):953-956.
37. Khaldi N, Collemare J, Lebrun MH, Wolfe KH: **Evidence for horizontal transfer of a secondary metabolite gene cluster between fungi**. *Genome biology* 2008, **9**(1):R18.
38. Wang H, Guo S, Huang M, Thorsten LH, Wei J: **Ascomycota has a faster evolutionary rate and higher species diversity than Basidiomycota**. *Sci China Life Sci* 2010, **53**(10):1163-1169.
39. Liu Z, Ellwood SR, Oliver RP, Friesen TL: ***Pyrenophora teres*: profile of an increasingly damaging barley pathogen**. *Molecular plant pathology* 2010, **12**(1):1-19.
40. Galagan JE, Selker EU: **RIP: the evolutionary cost of genome defense**. *Trends Genet* 2004, **20**(9):417-423.
41. Filippo JS, Sung P, Klein H: **Mechanism of Eukaryotic Homologous Recombination**. *Annual review of biochemistry* 2008, **77**:229-257.
42. Surosky RT, Tye BK: **Construction of telocentric chromosomes in *Saccharomyces cerevisiae***. *Proceedings of the National Academy of Sciences of the United States of America* 1985, **82**(7):2106-2110.
43. Mullany P, Roberts AP, Wang H: **Mechanism of integration and excision in conjugative transposons**. *Cell Mol Life Sci* 2002, **59**(12):2017-2022.