Comparative genomics of *Dothideomycete* fungi

Ate van der Burgt

Thesis committee

Promotor
Prof. Dr Pierre J.G.M. de Wit
Professor of Phytopathology
Wageningen University

Co-promotor
Dr Jérôme A.R. Collemare
Scientist, Laboratory of Phytopathology
Wageningen University

Other members
Dr Theo A.J. van der Lee, Wageningen University and Research Centre
Prof. Dr Dick de Ridder, Wageningen University
Dr Pierre Rouzé, Ghent University, Belgium
Prof. Dr Han A.B. Wösten, Utrecht University

This research was conducted under the auspices of the Graduate School of Experimental Plant Sciences.

Comparative genomics of *Dothideomycete* fungi

Ate van der Burgt

Thesis
submitted in fulfilment of the requirements for the degree of doctor
at Wageningen University
by the authority of the Rector Magnificus
Prof Dr M.J. Kropff,
in the presence of the
Thesis Committee appointed by the Academic Board
to be defended in public
on Monday 12 May 2014
at 4 p.m. in the Aula.

Ate van der Burgt
Comparative genomics of *Dothideomycete* fungi,
178 pages.

PhD thesis, Wageningen University, Wageningen, NL (2014)
With references, with summaries in Dutch and English

ISBN 978-90-6173-905-6

Contents

Chapter 1

General introduction

Introduction

The fungal kingdom has been both beloved and reviled since ancient human civilization. Many species of fungi have been within living memory consumed or used as active agent in the processing or fermentation of food products, thereby positively influencing their tenability and taste [1]. Examples are edible mushrooms like champignon (Agaricus bisporus), shiitake (Lentinula edodes) and black truffle (Tuber melanosporum), the leavening of bread by the baker's yeast Saccharomyces cerevisiae, production of soy sauce from a fermented paste of soybeans and various Aspergillus species and the fermentation of sugars into alcohol by various species of yeast during the production of beer and wine. In the 1930s, the antibiotic penicillin was discovered and isolated from the fungus Penicillium chrysogenum, yielding humanity a life-saving medicine against bacterial infections [2]. During the molecular biology revolution of the twentieth century this was followed by discoveries of an arsenal of additional pharmacologically active drugs, biopesticides, fine chemicals and enzymes with industrial applications [3]. Many of these are produced by fungi in large-scale industrial fermentation processes. A contemporary example of a fungus as a biological control organism is the application of spores of the entomopathogenic fungus Beauveria bassiana in the reduction of malaria parasites [4].

However, fungi are also notorious opportunistic pathogens of humans, causing serious discomfort up to fatal infections [5],[6]. Some plant-infecting fungi endanger food safety by the production of mycotoxins that cause serious health problems to human and cattle after their unintentional consumption [7],[8]. Most reknown are the fungi as causal agents of dramatic plant diseases [9]. They endanger global food supply by causing devastating losses to virtually all crops including staple foods as wheat, rice, maize and potato.

Some easy to culture, fast growing and predominantly single-cellular and filamentous fungi are used as simple models to study fundamental but highly complex regulatory processes of the eukaryotic cell, including cell division [10], circadian rhythms [11],[12],[13], apoptosis [14],[15] RNA interference [16] and epigenetics [17]. Because of their scientific and economic importance to industry, medicine and agriculture, many fungi are being studied from a biotechnological, molecular or genetic perspective (see Table 1). Many have been adopted as model organisms, including the famous budding yeast *S. cerevisiae [18]* and the fission yeast *Schizosaccharomyces pombe* [19].

The development of a technique called Sanger-sequencing enabled DNA sequencing from small molecules to complete genome sequences [20],[21]. Maturating and upscaling of this technique to the level of complete genomes spanned two decades. The genome of *S. cerevisiae* was the first eukaryotic genome to become completely sequenced [22], soon followed by the model fungal saprobes *Neurospora crassa* [23] and *Aspergillus nidulans* [24] and the rice blast pathogen *Magnaporthe oryzae* [25] (Table 1). The sequencing of the first eukaryotic genome 20 years ago was a milestone in genome research, but analogous breakthroughs were achieved in understanding biology at the level of the transcriptome, the proteome and the interplay between them all. The invention and consecutive up-scaling of various techniques that produced biological data sets at an unprecedented speed and magnitude, gave birth to the so-called omics era in life sciences [26]. Simultaneously, the scientific discipline bioinformatics was born and co-evolved with novel techniques and demands and became a major discipline in biology. Bioinformatics does not only facilitate storing, retrieving and organizing biological data, but is also crucial for exploring novel scientific concepts by developing novel methods and software tools, typically by integrating various data sources. It can address and answer many questions that could not be posed and solved previously by lack of analytical tools.

Table 1: Importance and application of the earliest sequenced fungal genomes

Fungal species	Year	Ref	Class | Order 1	importance
Saccharomyces cerevisae	1996	[22]	Saccharomycetes | Saccharomycetales	Baker's yeast, eukaryotic model organism
Schizosaccharomyces pombe	2002	[27]	Schizosaccharomycetes | Schizosaccharomycetales	Fission yeast, eukaryotic model organism
Neurospora crassa	2003	[23]	Sordariomycetes | Sordariales	Eukaryotic model organism
Candida albicans	2004	[28]	Saccharomycetes | Saccharomycetales	Human pathogen
Magnaporthe oryzae	2005	[25]	Sordariomycetes | Magnaporthales	Important pathogen on rice
Aspergillus nidulans *Aspergillus fumigatus* *Aspergillus oryzae*	2005	[24]	Eurotiomycetes | Eurotiales	Classical genetics; industrial production of enzymes and chemicals
Cryptococcus neoformans	2005	[29]	Tremellomycetes | Tremellales	Opportunistic human pathogen
Ustilago maydis	2006	[30]	Ustilaginomycetes | Ustilaginales	Biotrophic pathogen on maize
Stagonospora nodorum	2007	[31]	Dothideomycetes | Pleosporales	Important pathogen on wheat; first sequenced Dothideomycete
Aspergillus niger	2007	[32]	Eurotiomycetes | Eurotiales	Producer of chemicals & enzymes
Penicillium chrysogenum	2008	[33]	Eurotiomycetes | Eurotiales	Producer of antibiotics like penicillin
Fusarium graminearum *Fusarium verticillioides* *Fusarium oxysporum*	2010	[34]	Sordariomycetes | Hypocreales	Important pathogens on maize, etc.
powdery mildew species *Blumeria graminis* *Erysiphe pisi* *Golovinomyces orontii*	2010	[35]	Leotiomycetes | Erysiphales	Important biotrophic pathogens on various species
Tuber melanosporum	2010	[36]	Pezizomycetes | Pezizales	Edible truffle
Zymoseptoria tritici	2011	[37]	Dothideomycetes | Capnodiales	Pathogen on wheat
Verticillium dahlia *Verticillium albo-atrum*	2011	[38]	Sordariomycetes | Glomerellales	Important pathogen on many plant species
Sclerotinia sclerotiorum *Botrytis cinerea*	2011	[39]	Leotiomycetes | Helotiales	Important necrotrophic pathogen on many fruits

[1] classification according to NCBI Taxonomy (25/01/2014)

Next-generation sequencing

At the start of this PhD thesis project, tens of fungal genome sequences were available, a number that since has rapidly increased. The technological breakthrough of next-generation sequencing (NGS) have reduced the effort, duration and costs of genome projects [40]. These techniques were realized in commercial products starting from 2005. The term NGS comprises the on-going succession of traditional Sanger sequencing by various novel techniques, that are often known by the company name that has commercialized them. The reduction in sequencing costs per base has allowed increasing sequencing depth and has consequently increased the average sequence coverage of draft genome assemblies compared to those produced by the first-generation Sanger sequencing technology. To date, techniques that have played a major role are: massively parallel signature sequencing [41], 454 pyrosequencing [42],[43] and Illumina/Solexa sequencing [44] (for reviews see [45],[46]). Novel techniques are being developed that complement current shortcomings and might develop in their potential successors. The currently most promising pioneering technique released in a commercial product is single molecule real time sequencing (SMRT), but commonly named PacBio sequencing [47]. Various other promising techniques are in various stages of development upon commercial release. The advent of NGS techniques resulted in hundreds of sequenced fungal genomes that are currently available. The initiation and current realization of the 1000 Fungal Genomes project [48],[49] is the hallmark testimony that the omics era is not at its end yet.

Genomic and comparative genomics analyses of fungi

Availability of the genome sequence of an organism facilitate in depth genome analyses. A typical starting point in the analysis of genome sequences is the identification of its repetitive sequences and the annotation of its genes. Once the genomes of multiple organisms have been sequenced, comparative genomics can be exploited to find shared and unique genes, traits or other genomic features that might explain their commonalities and differences in morphology, metabolic potential, lifestyle or niche adaptation. From that perspective, fungi represent a taxonomic kingdom of particular interest. They have different life styles, thrive in a multitude of different environments and are diverse at the macroscopic level (development/morphology). From genome sequencing, it became evident that the same variation was also true for their genomes [25],[50]. Comparison of fungal genomes revealed enormous variability of all genomic features: genome size, chromosome number, repetitive DNA content, gene content and absence of gene colinearity (synteny). This variability is already realized in phylogenetically related fungi at the class, family [50],[51] and sometimes even genus level [24]. With respect to their gene content, low sequence similarity at the proteome level and large variability in gene family sizes were observed. In addition, a large number of orphan genes, defined as genes that lack obvious homologs in all other fungal species, was also reported [52],[53]. The societal, ecological and industrial importance of fungi and the exciting observations made in the sequenced fungal genomes have turned fungi into interesting study objects for bioinformaticians interested in genomics and comparative genomics analyses.

Genomics and comparative genomics analyses of several Dothideomycete fungi

In the last decade, most sequencing efforts were focussed on the filamentous *Ascomycete* fungi, most notably on the classes of *Dothideomycetes*, *Eurotiomycetes*, *Leotiomycetes* and *Sordariomycetes* (Table 1) [49]. Many of these fungi are devastating pathogens of both herbaceous plants and trees. Sequencing and gene annotations were predominantly carried out in community sequencing consortia like the Fungal Genome Initiative (BROAD institute [54],[55]) and the Fungal Genomics Program (Joint Genome Institute [56],[57]). This thesis has focused on analysing several *Dothideomycetes*. *Cladosporium fulvum* (pathogen of tomato) [58], *Dothistroma septosporum* (pathogen of pine) [59], and *Zymoseptoria tritici* (pathogen of wheat and some grasses; synonymous to *Mycosphaerella graminicola*) [37] are pathogens with a narrow host range. Because of the global importance of wheat and the devastative outcome of its infection, *Z. tritici* is regarded as one of the major fungal pathogens [60]. Despite their pathogenicity on distinct hosts, they start an infection cycle in a similar fashion by entering their hosts *via* open stomata and consequently colonizing apoplastic space, initially without causing any damage.. *C. fulvum* is a biotrophic fungus that on susceptible cultivars (compatible interaction) produces conidiophores that emerge from stomata that sporulate within 14 days post inoculation (dpi). During interaction with its host, it secretes effector proteins that modulate host defence responses and physiology [61]. Compatibility between *C. fulvum* and tomato is dictated by presence or absence of species- or strain-specific avirulence (*Avr*) effector genes of *C. fulvum* and corresponding resistance (*R*) genes of tomato [62]. This makes *C. fulvum* a model organism complying with the gene-for-gene interaction [63]. *D. septosporum*, on the other hand, has a hemi-biotrophic lifestyle on pine that typically spans several weeks to a month [64]. From this fungus no effector genes have been identified yet, but it is well known for the secretion of an aflatoxin-related secondary metabolite called dothistromin [65]. Although dothistromin is not essential for pathogenicity [66], recent observations suggest it to be a virulence factor (Kabir MS and Bradshaw RE, unpublished data). *Z. tritici* is a hemi-biotrophic wheat pathogen that is difficult to control, as most of the cultured wheat cultivars lack resistance genes and the fungus has become tolerant to the once effective azole fungicide [67]. The fungus evades basal host defence during the initial latent period (12-20 days), followed by a rapid switch to necrotrophic growth followed by production of pycnidia carrying the asexual pycnidiospores which completes its life cycle [37]. A comparative genomics approach would allow for exploring the gene content of these fungi and relate it to their lifestyle and ecological niche.

At the start of this thesis project, EST analyses [68], and genome sequencing, annotation and analyses [37] of *Z. tritici* had been performed. This experience was of great utility in performing comparable and additional analyses on the novel sequenced genomes of *C. fulvum* and *D. septosporum*, as described in chapter 2. Remarkably, these comparisons revealed occurrence of a surprisingly high similarity at the protein level combined with striking differences at the DNA level. Most strikingly, the genome of *C. fulvum* appeared to be at least twice as large, which is solely attributable to a much larger content in repetitive sequences. Genomics and comparative genomics analyses of these three fungi revealed various intriguing biological phenomena that led to further analyses as described in chapters 4 to 6.

Genome annotation (1): gene structure prediction

After genome sequencing and assembly, the subsequent *in silico* analyses aim at the identification of the protein-coding genes. In the last two decades, an array of gene identification algorithmic software has been developed. The algorithms are categorized in *ab initio* supervised, *ab initio* unsupervised and (supervised) alignment-based gene predictors, which are implemented in tools such as Augustus [69], GeneMark-ES [70] and TWINSCAN 2.0α [71],[72], respectively. Examples of software chosen here are those which are most suited for the annotations of fungal genomes. Most of the fungal model organism's genomes were sequenced, annotated and published in turn of the 21th century (Table 1). The earliest gene prediction software suited for fungal species comprised adjusted variants of tools developed earlier for the higher eukaryotic model organisms (mainly *Drosophila melanogaster*, *Caenorhabditis elegans*, *Arabidopsis thaliana* and *Homo sapiens*). But, soon it became clear that they were poor predictors of gene models in fungi which have exon-intron structures that are significantly different from higher eukaryotes.

Overall gene prediction software are far from flawless; state of the art eukaryotic gene prediction tools produce a substantial number of errors in gene catalogues, and as a consequence in public sequence repositories [73],[74],[75]. The major types of errors include (i) *bona fide* genes that are not predicted (false negative genes), (ii) non-existing genes that are predicted as *mala fide* genes (false positive genes) and (iii) prediction of incorrect intron-exon structures. In renown genome centres focused on sequencing and analyses of higher eukaryotes, accuracy of gene prediction improved by evidence acquired from large expression data sets, which were for a long time sparsely available for fungal model organisms. In addition, a substantial number of genes from the gene catalogues of model organisms were manually checked and improved.

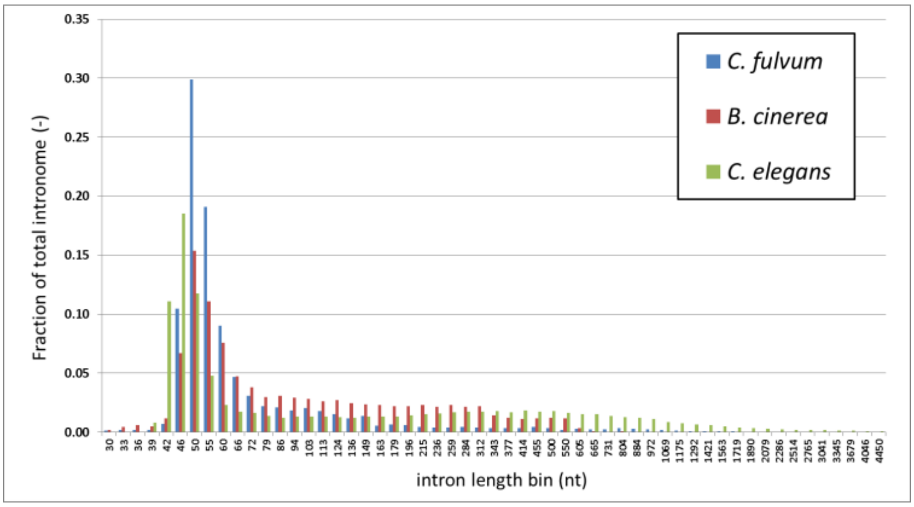


Figure 1: Intron length distribution of Cladosporium fulvum, Botrytis cinerea and Caenorhabditis elegans

Length bins where started at 30nt; consecutive bins start at a 10% size increase of previous bin.

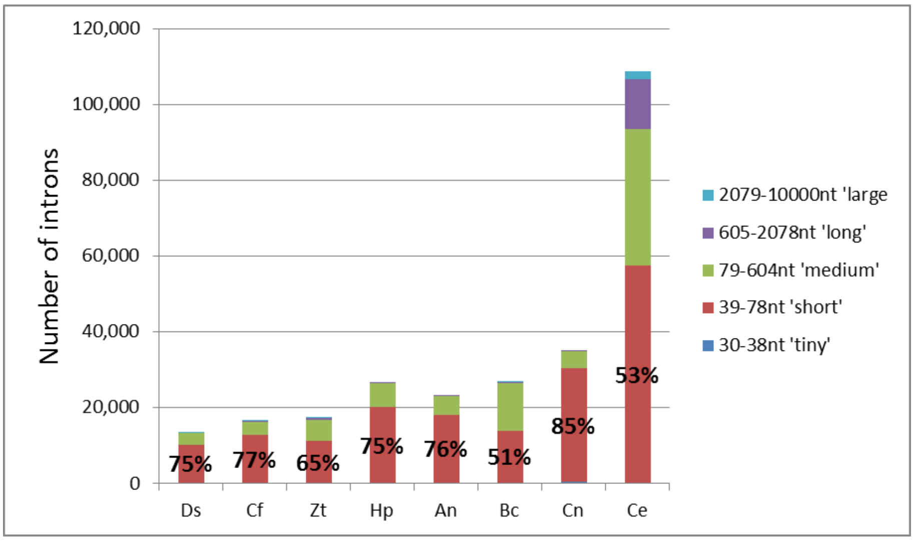


Figure 2: Overall intron length distribution and total intronome size of various species

Length bins were defined as explained in the legend of Figure 1. Species shown are Dothistroma septosporum (Ds), Cladosporium.fulvum (Cf), Zymoseptoria tritici (Zt), Hysterium pulicare (Hp), Aspergillus nidulans (An), Botrytis. cinerea (Bc), Cryptococcus neoformans (Cn) and Caenorhabditis elegans (Ce).

The genomes of filamentous fungi have been described as 'compact genomes with numerous small introns`[72],[70]. A distribution of fungal intron length typically peaks sharply around 50 to 54 nucleotides with 51 to 85% of all introns varying only within 39 to 78 nucleotides in length (Figure 1). In total the gene catalogue of a typical filamentous Ascomycete fungus comprises 10,000 to 15,000 genes carrying between 15,000 and 30,000 introns (Figure 2). For *Dothideomycetes* and *Eurotiomycetes* (*A. nidulans)*, these short introns account for up to 75% of their intronome. In *Leotiomycetes* (*B. cinerea*) and *Sordariomycetes* (data not shown), a considerable number of introns falls within the medium length range (79 to 604nt). It was shown that short and medium-sized introns have distinct characteristics [70], of which the difference in optimal spacing between branchpoint and acceptor site is the most predominant one. Existence of a considerable medium-sized fraction seems to be limited to certain subphyla and classes of fungi. Gene catalogues predating this observation might contain a relative high error rate in their predicted introns. Finally, tiny (<39nt) and long introns (>604nt) are rare in fungi. This is in sharp contrast to plants and animals, which contain much longer and on average more introns, as is exemplified by the intronome of *C. elegans*. It is also in sharp contrast to Hemiascomycete yeast species like *S. cerevisae* which contains only 282 introns in its 6,607 protein-coding genes (sequence release R64-1-1 2011/02/03 [18]).

As outlined above, it can be anticipated that fungal gene catalogues contain errors. Besides systematic errors of gene prediction software, additional circumstances can be identified that make fungal genomes extra vulnerable in respect to errors in their structural gene annotation. The available fungal genome sequence data was often supported by low coverage which seriously hampers accurate gene predictions. Indeed, an update report of the Human Genome Project indicated an estimation of at least 1% sequencing error rate at an average ten-fold coverage [76]. A typical fungal genome produced in the pre NGS era is supported by similar or even lower sequence coverage, which is likely to have a low reliability in both its overall assembly and of individual nucleotide calls [77]. An incorrectly predicted protein sequence is often recognized once aligned to its orthologs from other fungi, although the opposite does also occur, when the putative ortholog is wrongly predicted and will not align properly to the target gene. Manual and semi-automated inspection of the gene catalogues of *C. fulvum* and *D. septosporum* as described in chapter 2 revealed a considerable number of gene models that likely contain errors based on the phenomena described above.

The wealth of available fungal genome data that have recently become available allow the development of unsupervised methods to improve or assess gene annotation quality at the individual gene level. A method that predicts gene models by making use of evidence obtained from other genome sequences is called evidence- based or alignment-based gene prediction. In **chapter 3** we describe a novel alignment-based fungal gene prediction method (ABFGP) that is particularly suitable for plastic genomes like those of fungi. Usefulness of the method was shown by revisiting the annotations of *C. fulvum* and *D. septosporum* and of four other fungal genomes from the first-generation sequencing era. Thousands of gene models were revised in each of the gene catalogues, highlighting different types of errors in different annotation pipelines.

The ABFGP method was particularly efficient in identifying sequence errors (SEs) and/or disruptive mutations (DMs) that caused truncated and erroneous gene models. In **chapter 4**, we revisited the same fungal gene catalogues as those in chapter 3, and aimed at identifying pseudogenes caused by DMs. A dataset of fungal pseudogenes which are listed as *bona fide* genes in current gene catalogues is composed and subsequently analysed. It is enriched for gene models that contain false introns circumventing real DMs that otherwise would have caused the prediction of truncated genes.

Genome annotation (2): repeat identification

Availability of the first eukaryotic genomes showed that many contained large amounts of repetitive sequences. Repetitive DNA sequences, also indicated as repeats , can be subcategorized in low-complexity sequence, tandem repeats and interspersed repeats. Interspersed repeats comprise all classes of retrotransposons and DNA transposons: sequences that can (semi-) autonomously move or copy themselves throughout genomic DNA sequences. Many distinct classes of transposons have been characterized and studied in detail in fungi (for reviews see [78],[79],[80]). Evidence is accumulating that transposons are a major driving force causing genomic diversification in fungi: transposon-insertion mediated gene knock-out, altered expression of individual genes as well as altered genome organisation by chromosomal recombination (for review see [81],[82]).

Although low-complexity repeat sequences are not strictly repetitive, they are often taken along in the same container definition of `repetitive DNA`. For manual and automatic genome annotation and gene prediction, repetitive DNA represents a disturbing component. Therefore it should be identified and subsequently ignored or removed from further analyses. Recognition of repetitive sequence is quite straightforward, with a whole arsenal of dedicated software aiming at repeats duplicated in tandem [83], interspersed repeats in general (e.g. [84],[85],[86],[87]) or a specific class of repetitive sequence in particular (e.g. [88]).

The protocol employed for repeat-identification and/or repeat-masking has a major impact on the quality and content of the generated gene catalogue. The most simple protocol involves gene prediction prior to repeat-identification, without successive integration of data. This will result in numerous (parts) of protein-coding domains of transposon sequences to end up in the gene catalogue, while another involves identification of interspersed repeats and low-complexity sequence, and subsequent repeat masking prior to gene prediction. This is justified with the general observation that transposable elements tend to be located in the intergenic space. The exception to this rule is of course a transposon-insertion mediated pseudogenization event. But, this is expected to occur in low frequencies at the whole-genome scale, and in addition it comprises pseudogenes, and not genes. However, when low complexity sequences are masked too aggressively, parts of exons or introns of protein-coding genes might be masked as well, resulting in erroneous gene models at these loci.

These protocols represent the extremes of options, but might have been adopted in the earliest annotated fungal genomes. This is illustrated by a simple and sensitive BLASTP on NCBI's NR protein database, limited to fungal hits only (e≤1e-20, limited to txid:4890, performed on 2013/01/10). As query protein sequences, presumably full-length accessions of the most abundantly occurring retrotransposons in fungi were randomly chosen: Ty1/Copia (unnamed from *Pyrenophora tritici-repentis*, KF418198.1) and Ty3/Gypsy (Yeti from *Podospora anserina*, AJ272171.1 [89]). Numerous from hundreds of these significant hits (278/394 and 265/302) comprised annotated proteins of fungal species (uncharacterized, unnamed, hypothetical, conserved, unknown function). In fact, only the minority of accessions (93/394 and 27/302) is properly annotated as of presumably transposable origin (keywords: gag, pol, polyprotein, retro, transposon, polymerase).

Fortified with these prior experiences, extra care was required during the annotation of the highly repetitive genome of *C. fulvum*. During the critical assessment of repetitive sequences and structural gene annotation in the genome of *C. fulvum* we discovered something fascinating. Several abundantly occurring classes of unknown interspersed repeats were detected within the boundaries of annotated gene models, and turned out to exactly correspond to predicted and EST-supported introns. Some of the introns were near-identical, and all were present in unrelated genes.

In **chapter 5**, we describe explorative genomics and comparative genomics analyses that revealed the presence of Introner-Like Elements (ILEs) in many genes of various *Dothideomycete* fungi, including *Z. tritici* whose genome sequence has been publicly available for many years. These ILEs were subsequently characterized in detail showing that they are all normally spliced and generate highly structured and thermodynamically stable RNA fold-back loops. In addition, intron gain and loss analyses were performed showing that ILEs correspond to recent events of intron gains in genes. In **chapter 6**, we provide additional evidence that ILE multiplication strongly dominates over other types of intron duplication. The observed high rate of ILE multiplication followed by rapid sequence degeneration led us to the hypothesis that multiplication of ILEs has been the major cause and mechanism of intron gain in fungi.

Non-coding RNA genes: in silico prediction of miRNA hairpins

In the last two decades, awareness of the occurrence and biological relevance of non-coding RNA (ncRNA) genes has increased. Most crucial classes of ncRNAs for (eukaryotic) life are transfer RNAs (tRNAs), ribosomal RNAs (rRNAs), micro RNAs (miRNAs) and the small nucleolar RNAs (snoRNA) that form the catalytic part of the spliceosome. These and other have in common that they are all encoded in the genome as RNA genes (for review see [90]). Many of them have characteristic and stable secondary structures required for their biological function, as best exemplified by tRNAs and rRNAs. Accurate prediction of most classes of ncRNA genes is even more challenging as the prediction of protein-coding genes [91], especially because some of them evolve fast at the sequence level, while maintaining their structure [92].

After the initial discovery of miRNAs in the worm *C. elegans* in the early 1990s [93] their importance to (multicellular) eukaryotic life was not immediately recognized. In the early 2000s, instigated hand in hand with the prospering of NGS, miRNAs were rediscovered and established as a distinct, vital and evolutionarily ancient component of genetic regulation in animals [94],[95],[96] and plants [97]. In addition, retroviral human pathogens were discovered to encode miRNA hairpin genes that encode mature miRNAs that modulate host gene expression ([98]; for review see [99]). Mature miRNAs are small non-coding RNA molecules of ~21nt length which function in transcriptional and post-transcriptional regulation of gene expression (for reviews see [100],[101]). The miRNA is encoded on the DNA as a non-coding gene that is transcribed as a pri-miRNA transcript, in which the mature miRNA is embedded as a near-perfect palindromic repeated sequence. This palindromic sequence folds into a thermodynamically stable hairpin structure that is recognized and further processed by a pathway including Dicer nucleases that will excise the mature miRNA molecule. The mature miRNA will bind to a nuclease called Argonaute guiding the miRNA to a complementary sequence present in of one or more mRNAs (for review see [102]). Interaction by direct base pairing results in either cleavage by the Argonaute nuclease, or affects the level and/or efficiency of mRNA translation. After initial discovery and classifications of miRNAs in eukaryotes, research is shifting towards understanding the evolution and functional diversification of miRNAs [103] .

Although core components are shared between plants and animals, striking differences exist in the miRNA pathway, function and repertoire. This suggests that miRNA repertoires and even components of the miRNA pathway itself have evolved independently in the two kingdoms with different modes of action [104]. This might suggest that other eukaryotes might have evolved equivalent, yet slightly different pathways for gene regulation. This hypothesis is further supported by presence of the same core components in many single-cellular eukaryotes, including fungi [105],[106]. Initial experiments did not conclusively explain the *in vivo* role of both conserved copies of dicers in Ascomycete fungi [107],[108]. Opposite to its essentiality in multicellular eukaryotes, a few lineages of unicellular eukaryotes appear to have lost the RNAi machinery independently and completely [104]. Prior to the NGS era, large-scale *de novo* miRNA discovery was popular among bioinformaticians, because the miRNA hairpin structure can be searched for *in silico* fairly easily in DNA sequences (reviewed in [109]). Each of these methods has its pro's and con's, but generally all struggle with the problem of balancing between good sensitivity and good specificity. In one particular study, conserved secondary structures in six *Aspergillus* genomes were examined, which yielded not a single convincing plant- or animal-like miRNA [110]. However, as McGuire and co-authors correctly concluded, putatively existing fungal miRNA hairpins[1] might have slightly different properties from those of plants and mammals.

Therefore, a method that can be tailored for various types of miRNA hairpins and for which exhaustive sets of genomic hairpins can be monitored, filtered and enriched for putative miRNA hairpins, is needed to assess miRNA presence in previously unexplored organisms like fungi. In **chapter 7**, we describe a new strategy for miRNA hairpin prediction using statistical distributions of observed biological variation of properties (descriptors) of known miRNA hairpins. We show that the method outperforms miRNA prediction by previously conventional methods that usually apply threshold filtering. Although chapter 7 is not targeted on fungi, the study is relevant in the context of having a flexible method of finding evidence for a putative miRNA-like pathway in fungi.

Challenges for current and future bioinformatics analyses of fungi

As is nowadays soberly admitted but factually is 'the elephant in the room', several factors have complicated the initial genomic and comparative genomics analyses of fungi. An incomplete exploration of some of the largest problems is shortly summarized here. As outlined above, and although few studies emphasise this problem, it is well accepted in the fungal community that gene annotation of most fungal genomes is still of low quality. A recent study quantified putative errors in protein-coding genes of higher eukaryotic model organism [112],[75], and predicted a considerable number of proteins to be incorrect. In the reannotation of the fungus *F. graminearum*, 1,770 gene models were corrected and thus recognized as previously incorrectly annotated [113]. Many genomes were sequenced at low coverage due to the expensive Sanger technology, yielding reference sequences with an unknown but high number of sequencing errors [114]. The large evolutionary distances and variability among fungi along with the sparsity of available genome sequences imposes a general problem of under-sampling, which complicates or even negatively influences the outcome of comparative genomics analyses [115]. This includes especially incorrect orthology assignment and false detections of horizontal gene transfer events.

Recognition, correction and description of erroneous predictions of a few genes or a gene family are often mentioned in publications, but is rarely followed by traceable updates of public data-repositories, maintaining the risk of propagating errors by comparative genomics analyses. Finally, the scientific community interested in studying a particular fungal species is often rather small, unlike the well-organized communities working on famous model organisms like the yeast *S. cerevisae*, the worm *C. elegans*, the plant *A. thaliana* and the mammal *H. sapiens*. Quality improving steps that involve human interference, let alone discovering and traceably correcting earlier introduced errors, are marginally performed. Any large-scale dataset of *in silico* predictions of biological properties is intrinsically prone to contain errors. Besides dealing with the on-going scale up in the omics era, the challenge for current and future bioinformaticians is to increase the accuracy of their predictions. This will be the discussion theme in **chapter 8**.

[1] the first fungal miRNA-like RNAs were described to occur in *N. crassa* [111] after finishing the study described in chapter 7.

References

1. Ross RP, Morgan S, Hill C: Preservation and fermentation: past, present and future. International journal of food microbiology 2002, 79(1-2):3-16.

2. Fleming A: The antibacterial action of a Penicillium, with special reference to their use in the isolation of B. influenzae. *Br J Exp Pathol* 1929, 10:222-236.

3. Buchholz K, Collins J: **The roots--a short history of industrial microbiology and biotechnology**. *Applied microbiology and biotechnology* 2013, **97**(9):3747-3762.

4. Blanford S, Shi W, Christian R, Marden JH, Koekemoer LL, Brooke BD, Coetzee M, Read AF, Thomas MB: **Lethal and pre-lethal effects of a fungal biopesticide contribute to substantial and rapid control of malaria vectors**. *PloS one* 2011, **6**(8):e23591.

5. Ponton J, Ruchel R, Clemons KV, Coleman DC, Grillot R, Guarro J, Aldebert D, Ambroise-Thomas P, Cano J, Carrillo-Munoz AJ *et al*: **Emerging pathogens**. *Medical mycology : official publication of the International Society for Human and Animal Mycology* 2000, **38 Suppl 1**:225-236.

6. Richardson M, Lass-Florl C: **Changing epidemiology of systemic fungal infections**. Clinical microbiology and infection : the official publication of the European Society of Clinical Microbiology and Infectious Diseases 2008, **14 Suppl 4**:5-24.

7. Bhatnagar D, Yu J, Ehrlich KC: **Toxins of filamentous fungi**. *Chemical immunology* 2002, **81**:167-206.

8. Richard JL: **Some major mycotoxins and their mycotoxicoses--an overview**. *International journal of food microbiology* 2007, **119**(1-2):3-10.

9. Strange RN, Scott PR: **Plant disease: a threat to global food security**. *Annu Rev Phytopathol* 2005, **43**:83-116.

10. Piel M, Tran PT: **Cell shape and cell division in fission yeast**. *Current biology : CB* 2009, **19**(17):R823-827.

11. Kippert F: **Cellular signalling and the complexity of biological timing: insights from the ultradian clock of Schizosaccharomyces pombe**. *Philosophical transactions of the Royal Society of London Series B, Biological sciences* 2001, **356**(1415):1725-1733.

12. Froehlich AC, Liu Y, Loros JJ, Dunlap JC: White Collar-1, a circadian blue light photoreceptor, binding to the frequency promoter. *Science* 2002, 297(5582):815-819.

13. Eelderink-Chen Z, Mazzotta G, Sturre M, Bosman J, Roenneberg T, Merrow M: **A circadian clock in Saccharomyces cerevisiae**. *Proceedings of the National Academy of Sciences of the United States of America* 2010, **107**(5):2043-2047.

14. Low CP, Liew LP, Pervaiz S, Yang H: **Apoptosis and lipoapoptosis in the fission yeast Schizosaccharomyces pombe**. *FEMS yeast research* 2005, **5**(12):1199-1206.

15. Carmona-Gutierrez D, Eisenberg T, Buttner S, Meisinger C, Kroemer G, Madeo F: **Apoptosis in yeast: triggers, pathways, subroutines**. *Cell death and differentiation* 2010, **17**(5):763-773.

16. Chang SS, Zhang Z, Liu Y: **RNA interference pathways in fungi: mechanisms and functions**. *Annual review of microbiology* 2012, **66**:305-323.

17. Aramayo R, Selker EU: **Neurospora crassa, a model system for epigenetics research**. *Cold Spring Harbor perspectives in biology* 2013, **5**(10):a017921.

18. SGD: Saccharomyces genome database [http://www.yeastgenome.org]

19. pombase: the scientific resource for fission yeast [http://www.pombase.org/]

20. Sanger F, Coulson AR: A rapid method for determining sequences in DNA by primed synthesis with DNA polymerase. *Journal of molecular biology* 1975, 94(3):441-448.

21. Sanger F, Nicklen S, Coulson AR: **DNA sequencing with chain-terminating inhibitors**. *Proceedings of the National Academy of Sciences of the United States of America* 1977, **74**(12):5463-5467.

22. Goffeau A, Barrell BG, Bussey H, Davis RW, Dujon B, Feldmann H, Galibert F, Hoheisel JD, Jacq C, Johnston M *et al*: **Life with 6000 genes**. *Science* 1996, **274**(5287):546, 563-547.

23. 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.

24. Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, Batzoglou S, Lee SI, Basturkmen M, Spevak CC, Clutterbuck J *et al*: **Sequencing of Aspergillus nidulans and comparative analysis with A. fumigatus and A. oryzae**. *Nature* 2005, **438**(7071):1105-1115.

25. 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.

26. Methe BA, Lasa I: Microbiology in the 'omics era: from the study of single cells to communities and beyond. *Current opinion in microbiology* 2013, 16(5):602-604.

27. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, Sgouros J, Peat N, Hayles J, Baker S *et al*: **The genome sequence of Schizosaccharomyces pombe**. *Nature* 2002, **415**(6874):871-880.

28. Jones T, Federspiel NA, Chibana H, Dungan J, Kalman S, Magee BB, Newport G, Thorstenson YR, Agabian N, Magee PT *et al*: **The diploid genome sequence of Candida albicans**. *Proceedings of the National Academy of Sciences of the United States of America* 2004, **101**(19):7329-7334.

29. Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA *et al*: **The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans**. *Science* 2005, **307**(5713):1321-1324.

30. Kamper J, Kahmann R, Bolker M, Ma LJ, Brefort T, Saville BJ, Banuett F, Kronstad JW, Gold SE, Muller O *et al*: **Insights from the genome of the biotrophic fungal plant pathogen Ustilago maydis**. *Nature* 2006, **444**(7115):97-101.

31. 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.

32. Pel HJ, de Winde JH, Archer DB, Dyer PS, Hofmann G, Schaap PJ, Turner G, de Vries RP, Albang R, Albermann K *et al*: **Genome sequencing and analysis of the versatile cell factory Aspergillus niger CBS 513.88**. *Nature biotechnology* 2007, **25**(2):221-231.

33. van den Berg MA, Albang R, Albermann K, Badger JH, Daran JM, Driessen AJ, Garcia-Estrada C, Fedorova ND, Harris DM, Heijne WH *et al*: **Genome sequencing and analysis of the filamentous fungus Penicillium chrysogenum**. *Nature biotechnology* 2008, **26**(10):1161-1168.

34. 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.

35. Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stuber K, Ver Loren van Themaat E, Brown JK, Butcher SA, Gurr SJ *et al*: **Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism**. *Science* 2010, **330**(6010):1543-1546.

36. Martin F, Kohler A, Murat C, Balestrini R, Coutinho PM, Jaillon O, Montanini B, Morin E, Noel B, Percudani R *et al*: **Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis**. *Nature* 2010, **464**(7291):1033-1038.

37. Goodwin SB, M'Barek S B, Dhillon B, Wittenberg AH, Crane CF, Hane JK, Foster AJ, Van der Lee TA, Grimwood J, Aerts A *et al*: **Finished genome of the fungal wheat pathogen Mycosphaerella graminicola reveals dispensome structure, chromosome plasticity, and stealth pathogenesis**. *PLoS genetics* 2011, **7**(6):e1002070.

38. Klosterman SJ, Subbarao KV, Kang S, Veronese P, Gold SE, Thomma BP, Chen Z, Henrissat B, Lee YH, Park J *et al*: **Comparative genomics yields insights into niche adaptation of plant vascular wilt pathogens**. *PLoS pathogens* 2011, **7**(7):e1002137.

39. Amselem J, Cuomo CA, van Kan JA, Viaud M, Benito EP, Couloux A, Coutinho PM, de Vries RP, Dyer PS, Fillinger S *et al*: **Genomic analysis of the necrotrophic fungal pathogens Sclerotinia sclerotiorum and Botrytis cinerea**. *PLoS genetics* 2011, **7**(8):e1002230.

40. Shendure J, Mitra RD, Varma C, Church GM: **Advanced sequencing technologies: methods and goals**. *Nature reviews Genetics* 2004, **5**(5):335-344.

41. Brenner S, Johnson M, Bridgham J, Golda G, Lloyd DH, Johnson D, Luo S, McCurdy S, Foy M, Ewan M *et al*: **Gene expression analysis by massively parallel signature sequencing (MPSS) on microbead arrays**. *Nature biotechnology* 2000, **18**(6):630-634.

42. Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z *et al*: **Genome sequencing in microfabricated high-density picolitre reactors**. *Nature* 2005, **437**(7057):376-380.

43. Wheeler DA, Srinivasan M, Egholm M, Shen Y, Chen L, McGuire A, He W, Chen YJ, Makhijani V, Roth GT *et al*: **The complete genome of an individual by massively parallel DNA sequencing**. *Nature* 2008, **452**(7189):872-876.

44. Bentley DR, Balasubramanian S, Swerdlow HP, Smith GP, Milton J, Brown CG, Hall KP, Evers DJ, Barnes CL, Bignell HR *et al*: **Accurate whole human genome sequencing using reversible terminator chemistry**. *Nature* 2008, **456**(7218):53-59.

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

46. Metzker ML: **Sequencing technologies - the next generation**. *Nature reviews Genetics* 2010, **11**(1):31-46.

47. Chin CS, Alexander DH, Marks P, Klammer AA, Drake J, Heiner C, Clum A, Copeland A, Huddleston J, Eichler EE *et al*: **Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data**. *Nature methods* 2013, **10**(6):563-569.

48. **1kFG: 1000 fungal genomes project** [http://1000.fungalgenomes.org/home/]

49. Grigoriev IV, Nikitin R, Haridas S, Kuo A, Ohm R, Otillar R, Riley R, Salamov A, Zhao X, Korzeniewski F *et al*: **MycoCosm portal: gearing up for 1000 fungal genomes**. *Nucleic acids research* 2014, **42**(1):D699-704.

50. Galagan JE, Henn MR, Ma LJ, Cuomo CA, Birren B: **Genomics of the fungal kingdom: insights into eukaryotic biology**. *Genome research* 2005, **15**(12):1620-1631.

51. Ohm RA, Feau N, Henrissat B, Schoch CL, Horwitz BA, Barry KW, Condon BJ, Copeland AC, Dhillon B, Glaser F *et al*: **Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi**. *PLoS pathogens* 2012, **8**(12):e1003037.

52. Mannhaupt G, Montrone C, Haase D, Mewes HW, Aign V, Hoheisel JD, Fartmann B, Nyakatura G, Kempken F, Maier J *et al*: **What's in the genome of a filamentous fungus? Analysis of the Neurospora genome sequence**. *Nucleic acids research* 2003, **31**(7):1944-1954.

53. Ekman D, Elofsson A: **Identifying and quantifying orphan protein sequences in fungi**. *Journal of molecular biology* 2010, **396**(2):396-405.

54. **FGI: Fungal Genome Initiative** [http://www.broadinstitute.org/scientific-community/science/projects/fungal-genome-initiative/fungal-genome-initiative]

55. Cuomo CA, Birren BW: The fungal genome initiative and lessons learned from genome sequencing. *Methods in enzymology* 2010, 470:833-855.

56. **FGP: Fungal Genomics Program** [http://genome.jgi.doe.gov/programs/fungi/index.jsf]

57. Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky S, Nikitin R, Ohm RA *et al*: **The genome portal of the Department of Energy Joint Genome Institute**. *Nucleic acids research* 2012, **40**(Database issue):D26-32.

58. de Wit PJ, Joosten MH: **Avirulence and resistance genes in the Cladosporium fulvum-tomato interaction**. *Current opinion in microbiology* 1999, **2**(4):368-373.

59. Barnes I, Crous PW, Wingfield BD, Wingfield MJ: Multigene phylogenies reveal that red band needle blight of Pinus is caused by two distinct species of Dothistroma, D-septosporum and D-pini. *Stud Mycol* 2004(50):551-565.

60. Dean R, Van Kan JA, Pretorius ZA, Hammond-Kosack KE, Di Pietro A, Spanu PD, Rudd JJ, Dickman M, Kahmann R, Ellis J *et al*: **The Top 10 fungal pathogens in molecular plant pathology**. *Molecular plant pathology* 2012, **13**(4):414-430.

61. De Wit PJ, Mehrabi R, Van den Burg HA, Stergiopoulos I: **Fungal effector proteins: past, present and future**. *Molecular plant pathology* 2009, **10**(6):735-747.

62. Wulff BB, Chakrabarti A, Jones DA: **Recognitional specificity and evolution in the tomato-Cladosporium fulvum pathosystem**. *Molecular plant-microbe interactions : MPMI* 2009, **22**(10):1191-1202.

63. Dewit PJGM, Joosten MHAJ, Honee G, Wubben JP, Vandenackerveken GFJM, Vandenbroek HWJ: **Molecular Communication between Host-Plant and the Fungal Tomato Pathogen Cladosporium-Fulvum**. *Anton Leeuw Int J G* 1994, **65**(3):257-262.

64. Bradshaw RE: Dothistroma (red-band) needle blight of pines and the dothistromin toxin: a review. *Forest Pathol* 2004, 34(3):163-185.

65. Schwelm A, Bradshaw RE: Genetics of dothistromin biosynthesis of Dothistroma septosporum: an update. *Toxins* 2010, 2(11):2680-2698.

66. Schwelm A, Barron NJ, Baker J, Dick M, Long PG, Zhang S, Bradshaw RE: **Dothistromin toxin is not required for dothistroma needle blight in Pinus radiata**. *Plant Pathol* 2009, **58**(2):293-304.

67. Cools HJ, Fraaije BA: Update on mechanisms of azole resistance in Mycosphaerella graminicola and implications for future control. *Pest management science* 2013, 69(2):150-155.

68. Kema GH, van der Lee TA, Mendes O, Verstappen EC, Lankhorst RK, Sandbrink H, van der Burgt A, Zwiers LH, Csukai M, Waalwijk C: **Large-scale gene discovery in the septoria tritici blotch fungus Mycosphaerella

graminicola with a focus on in planta expression. *Molecular plant-microbe interactions : MPMI* 2008, **21**(9):1249-1260.

69. Stanke M, Tzvetkova A, Morgenstern B: AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. *Genome biology* 2006, 7 Suppl 1:S11 11-18.

70. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M: **Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training**. *Genome research* 2008, **18**(12):1979-1990.

71. Korf I, Flicek P, Duan D, Brent MR: **Integrating genomic homology into gene structure prediction**. *Bioinformatics* 2001, **17 Suppl 1**:S140-148.

72. Tenney AE, Brown RH, Vaske C, Lodge JK, Doering TL, Brent MR: **Gene prediction and verification in a compact genome with numerous small introns**. *Genome research* 2004, **14**(11):2330-2335.

73. Odronitz F, Kollmar M: Drawing the tree of eukaryotic life based on the analysis of 2,269 manually annotated myosins from 328 species. *Genome biology* 2007, 8(9):R196.

74. Keller O, Odronitz F, Stanke M, Kollmar M, Waack S: Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. *BMC bioinformatics* 2008, 9:278.

75. Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Szlama G, Szarka E, Trexler M, Banyai L, Patthy L: **MisPred: quality control of gene predictions and public databases**. *Febs J* 2013, **280**:543-543.

76. Weber JL, Myers EW: **Human whole-genome shotgun sequencing**. *Genome research* 1997, **7**(5):401-409.

77. Staats M, van Kan JA: **Genome update of Botrytis cinerea strains B05.10 and T4**. *Eukaryotic cell* 2012, **11**(11):1413-1414.

78. Oliver R: **Transposons in Filamentous Fungi**. *Molecular Biology of Filamentous Fungi* 1992:3-11.

79. Daboussi MJ, Capy P: **Transposable elements in filamentous fungi**. *Annual review of microbiology* 2003, **57**:275-299.

80. Muszewska A, Hoffman-Sommer M, Grynberg M: **LTR retrotransposons in fungi**. *PloS one* 2011, **6**(12):e29425.

81. Wostemeyer J, Kreibich A: Repetitive DNA elements in fungi (Mycota): impact on genomic architecture and evolution. *Current genetics* 2002, 41(4):189-198.

82. Feschotte C, Pritham EJ: **DNA transposons and the evolution of eukaryotic genomes**. *Annual review of genetics* 2007, **41**:331-368.

83. Benson G: Tandem repeats finder: a program to analyze DNA sequences. *Nucleic acids research* 1999, 27(2):573-580.

84. Volfovsky N, Haas BJ, Salzberg SL: **A clustering method for repeat analysis in DNA sequences**. *Genome biology* 2001, **2**(8):RESEARCH0027.

85. Bao Z, Eddy SR: Automated de novo identification of repeat sequence families in sequenced genomes. *Genome research* 2002, 12(8):1269-1276.

86. Price AL, Jones NC, Pevzner PA: **De novo identification of repeat families in large genomes**. *Bioinformatics* 2005, **21 Suppl 1**:i351-358.

87. **RepeatMasker** [http://repeatmasker.org]

88. Han Y, Wessler SR: MITE-Hunter: a program for discovering miniature inverted-repeat transposable elements from genomic sequences. *Nucleic acids research* 2010, 38(22):e199.

89. Hamann A, Feller F, Osiewacz HD: Yeti--a degenerate gypsy-like LTR retrotransposon in the filamentous ascomycete Podospora anserina. *Current genetics* 2000, 38(3):132-140.

90. Eddy SR: **Non-coding RNA genes and the modern RNA world**. *Nature reviews Genetics* 2001, **2**(12):919-929.

91. Rivas E, Klein RJ, Jones TA, Eddy SR: **Computational identification of noncoding RNAs in E. coli by comparative genomics**. *Current biology : CB* 2001, **11**(17):1369-1373.

92. Savill NJ, Hoyle DC, Higgs PG: RNA sequence evolution with secondary structure constraints: comparison of substitution rate models using maximum-likelihood methods. *Genetics* 2001, 157(1):399-411.

93. Lee RC, Feinbaum RL, Ambros V: The C. elegans heterochronic gene lin-4 encodes small RNAs with antisense complementarity to lin-14. *Cell* 1993, 75(5):843-854.

94. Lee RC, Ambros V: An extensive class of small RNAs in Caenorhabditis elegans. *Science* 2001, 294(5543):862-864.

95. Lau NC, Lim LP, Weinstein EG, Bartel DP: An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans. *Science* 2001, 294(5543):858-862.

96. Lagos-Quintana M, Rauhut R, Lendeckel W, Tuschl T: **Identification of novel genes coding for small expressed RNAs**. *Science* 2001, **294**(5543):853-858.

97. Reinhart BJ, Weinstein EG, Rhoades MW, Bartel B, Bartel DP: **MicroRNAs in plants**. *Genes & development* 2002, **16**(13):1616-1626.

98. Pfeffer S, Zavolan M, Grasser FA, Chien M, Russo JJ, Ju J, John B, Enright AJ, Marks D, Sander C *et al*: **Identification of virus-encoded microRNAs**. *Science* 2004, **304**(5671):734-736.

99. Kincaid RP, Sullivan CS: **Virus-encoded microRNAs: an overview and a look to the future**. *PLoS pathogens* 2012, **8**(12):e1003018.

100. Bartel DP: MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004, 116(2):281-297.

101. Krol J, Loedige I, Filipowicz W: **The widespread regulation of microRNA biogenesis, function and decay**. *Nature reviews Genetics* 2010, **11**(9):597-610.

102. Bartel DP: MicroRNAs: target recognition and regulatory functions. *Cell* 2009, 136(2):215-233.

103. Cuperus JT, Fahlgren N, Carrington JC: **Evolution and functional diversification of MIRNA genes**. *The Plant cell* 2011, **23**(2):431-442.

104. Shabalina SA, Koonin EV: **Origins and evolution of eukaryotic RNA interference**. *Trends in ecology & evolution* 2008, **23**(10):578-587.

105. Nakayashiki H: **RNA silencing in fungi: mechanisms and applications**. *FEBS letters* 2005, **579**(26):5950-5957.

106. Nakayashiki H, Kadotani N, Mayama S: **Evolution and diversification of RNA silencing proteins in fungi**. *Journal of molecular evolution* 2006, **63**(1):127-135.

107. Catalanotto C, Pallotta M, ReFalo P, Sachs MS, Vayssie L, Macino G, Cogoni C: **Redundancy of the two dicer genes in transgene-induced posttranscriptional gene silencing in Neurospora crassa**. *Molecular and cellular biology* 2004, **24**(6):2536-2545.

108. Kadotani N, Nakayashiki H, Tosa Y, Mayama S: One of the two Dicer-like proteins in the filamentous fungi Magnaporthe oryzae genome is responsible for hairpin RNA-triggered RNA silencing and related small interfering RNA accumulation. *The Journal of biological chemistry* 2004, 279(43):44467-44474.

109. Berezikov E, Cuppen E, Plasterk RH: **Approaches to microRNA discovery**. *Nature genetics* 2006, **38 Suppl**:S2-7.

110. McGuire AM, Galagan JE: **Conserved secondary structures in Aspergillus**. *PloS one* 2008, **3**(7):e2812.

111. Lee HC, Li L, Gu W, Xue Z, Crosthwaite SK, Pertsemlidis A, Lewis ZA, Freitag M, Selker EU, Mello CC *et al*: **Diverse pathways generate microRNA-like RNAs and Dicer-independent small interfering RNAs in fungi**. *Molecular cell* 2010, **38**(6):803-814.

112. Nagy A, Hegyi H, Farkas K, Tordai H, Kozma E, Banyai L, Patthy L: **Identification and correction of abnormal, incomplete and mispredicted proteins in public databases**. *BMC bioinformatics* 2008, **9**:353.

113. Wong P, Walter M, Lee W, Mannhaupt G, Munsterkotter M, Mewes HW, Adam G, Guldener U: **FGDB: revisiting the genome annotation of the plant pathogen Fusarium graminearum**. *Nucleic acids research* 2011, **39**(Database issue):D637-639.

114. Churchill GA, Waterman MS: The accuracy of DNA sequences: estimating sequence quality. *Genomics* 1992, 14(1):89-98.

115. Garcia S, Herrera F: Evolutionary Undersampling for Classification with Imbalanced Datasets: Proposals and Taxonomy. *Evolutionary computation* 2009, 17(3):275-306.

Chapter 2

The genomes of the fungal plant pathogens *Cladosporium fulvum* and *Dothistroma septosporum* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry.

de Wit PJGM, van der Burgt A, Ökmen B, Stergiopoulos I, Abd-Elsalam KA, Aerts AL, Bahkali AH, Beenen HG, Chettri P, Cox MP, Datema E, de Vries RP, Dhillon B, Ganley AR, Griffiths SA, Guo Y, Hamelin RC, Henrissat B, Kabir MS, Jashni MK, Kema G, Klaubauf S, Lapidus A, Levasseur A, Lindquist E, Mehrabi R, Ohm RA, Owen TJ, Salamov A, Schwelm A, Schijlen E, Sun H, van den Burg HA, van Ham RC, Zhang S, Goodwin SB, Grigoriev IV, Collemare J, Bradshaw RE.

PLoS Genetics 2012, 8(11):e1003088.

doi: 10.1371/journal.pgen.1003088

This article and its Supporting Information are available from:
http://dx.plos.org/10.1371/journal.pgen.1003088

Abstract

We sequenced and compared the genomes of the *Dothideomycete* fungal plant pathogens *Cladosporium fulvum (Cfu)* (syn. *Passalora fulva*) and *Dothistroma septosporum (Dse)* that are closely related phylogenetically, but have different lifestyles and hosts. Although both fungi grow extracellularly in close contact with host mesophyll cells, *Cfu* is a biotroph infecting tomato, while *Dse* is a hemibiotroph infecting pine. The genomes of these fungi have a similar set of genes (70% of gene content in both genomes are homologs), but differ significantly in size (*Cfu* >61.1 Mb; *Dse* 31.2 Mb), which is mainly due to the difference in repeat content (47.2% in *Cfu* versus 3.2% in *Dse*). Recent adaptation to different lifestyles and hosts is suggested by diverged sets of genes. *Cfu* contains an α-tomatinase gene that we predict might be required for detoxification of tomatine, whilst this gene is absent in *Dse*. Many genes encoding secreted proteins are unique to each species and the repeat-rich areas in *Cfu* are enriched for these species-specific genes. In contrast, conserved genes suggest common host ancestry. Homologs of *Cfu* effector genes, including *Ecp2* and *Avr4,* are present in *Dse* and induce a Cf-Ecp2- and Cf-4-mediated hypersensitive response, respectively. Strikingly, genes involved in production of the toxin dothistromin, a likely virulence factor for *Dse,* are conserved in *Cfu*, but their expression differs markedly with essentially no expression by *Cfu in planta*. Likewise, *Cfu* has a carbohydrate-degrading enzyme catalog that is more similar to that of necrotrophs or hemibiotrophs and a larger pectinolytic gene arsenal than *Dse,* but many of these genes are not expressed *in planta* or are pseudogenized. Overall, comparison of their genomes suggests that these closely related plant pathogens had a common ancestral host but since adapted to different hosts and lifestyles by a combination of differentiated gene content, pseudogenization and gene regulation.

Author summary

We compared the genomes of two closely related pathogens with very different lifestyles and hosts: *C. fulvum* (*Cfu*) a biotroph of tomato and *D. septosporum* (*Dse*) a hemibiotroph of pine. Some differences in gene content were identified that can be directly related to their different hosts, such as the presence of a gene involved in degradation of a tomato saponin only in *Cfu*. However, in general the two species share a surprisingly large proportion of genes. *Dse* has functional homologs of *Cfu* effector genes, whilst *Cfu* has genes for biosynthesis of dothistromin, a toxin probably associated with virulence in *Dse*. *Cfu* also has an unexpectedly large content of genes for biosynthesis of other secondary metabolites and degradation of plant cell walls compared to *Dse*, contrasting with its host preference and lifestyle. However, many of these genes were not expressed *in planta* or were pseudogenized. These results suggest that evolving species may retain genetic signatures of the host and lifestyle preferences of their ancestor, and that evolution of new genes, gene regulation and pseudogenization are important factors in adaptation.

Introduction

Cladosporium fulvum and *Dothistroma septosporum* are two related fungal species belonging to the class of *Dothideomycetes*. *C. fulvum* is a biotrophic pathogen of tomato that has served as a model system for plant-microbe interactions since its first effector gene, *Avr9*, was cloned in 1991 [1]. It is not related to species in the genus *Cladosporium sensu strictu*, and has recently been renamed *Passalora fulva* [2]. However, to be consistent with past literature it will be referred to here as *C. fulvum*. Phylogenetic analyses based on sequences of the internal transcribed spacer (ITS) region of the ribosomal DNA revealed that *C. fulvum* is closely related to *D. septosporum* and other *Dothideomycete* fungi such as species of *Mycosphaerella* isolated from eucalyptus [3]. *D. septosporum* is an economically important hemibiotrophic pathogen of pine species that is well known for its production of an aflatoxin-like toxin, dothistromin [4]. A taxonomic revision also occurred for this species: prior to 2004 the name *Dothistroma pini* (syn. *D. septosporum* syn. *D. septospora*) was widely used. The revision involved a split into two species: the best-studied and most widespread species was named *D. septosporum,* and a less common species retained the name of *D. pini* [5].

The disease caused by *C. fulvum,* leaf mold of tomato, likely originates from South America, the center of origin of tomato [6]. The first outbreak of the disease was reported in South Carolina, USA, in the late 1800s [7]. Since then, disease outbreaks have occurred worldwide in moderate temperature zones with high relative humidity. The disease was of high economic importance during the first half of the 20th century, but its importance waned after introgression of *Cf* (for *C. fulvum*) resistance genes by breeders into tomato cultivars began providing effective control [8]. However, recent outbreaks have been reported in countries where tomato cultivars lacking *Cf* resistance genes are grown, and in areas where intensive year-round cultivation of resistant tomato plants led to fungal strains overcoming *Cf* genes [9,10].

In contrast, the foliar forest pathogen *D. septosporum* (Dorog.) Morelet has a relatively recent history and has been less intensively studied than *C. fulvum*. *D. septosporum* infects over 70 species of pine, as well as several minor hosts including some *Picea* species [11]. During the 1960s-1980s, Dothistroma needle blight (DNB) was largely a problem of Southern hemisphere pine plantations, where primary control was achieved by fungicide applications or planting of resistant species (reviewed in [12]). Since the early 1990s DNB incidence has increased greatly in the Northern hemisphere, with some epidemics causing unprecedented levels of mortality [13,14]. In northwest British Columbia, disease outbreaks are correlated with summer rainfall levels, suggesting that climate change could have unpredictable and severe effects on DNB outbreaks in forests. [15].

Infection in both the *C. fulvum*-tomato and *D. septosporum*-pine pathosystems starts with conidia that germinate on the leaf surface and produce runner hyphae that enter the host through open stomata. Subsequently, the fungi colonize the apoplastic space between mesophyll cells. In the case of *C. fulvum* conidiophores emerge from stomata 10-14 days later producing massive amounts of conidia that can re-infect tomato [16,17,18](Fig. 1A-D). *D. septosporum* produces conidia several weeks after infection, on conidiomata that erupt through the needle epidermis where they can be spread to other pines by rain-splash [19,20](Fig. 1E-H). Whilst *C. fulvum* is considered a biotroph, *D. septosporum* is assumed to be a hemibiotroph based on similarities of its lifecycle to other *Dothideomycete* fungi.

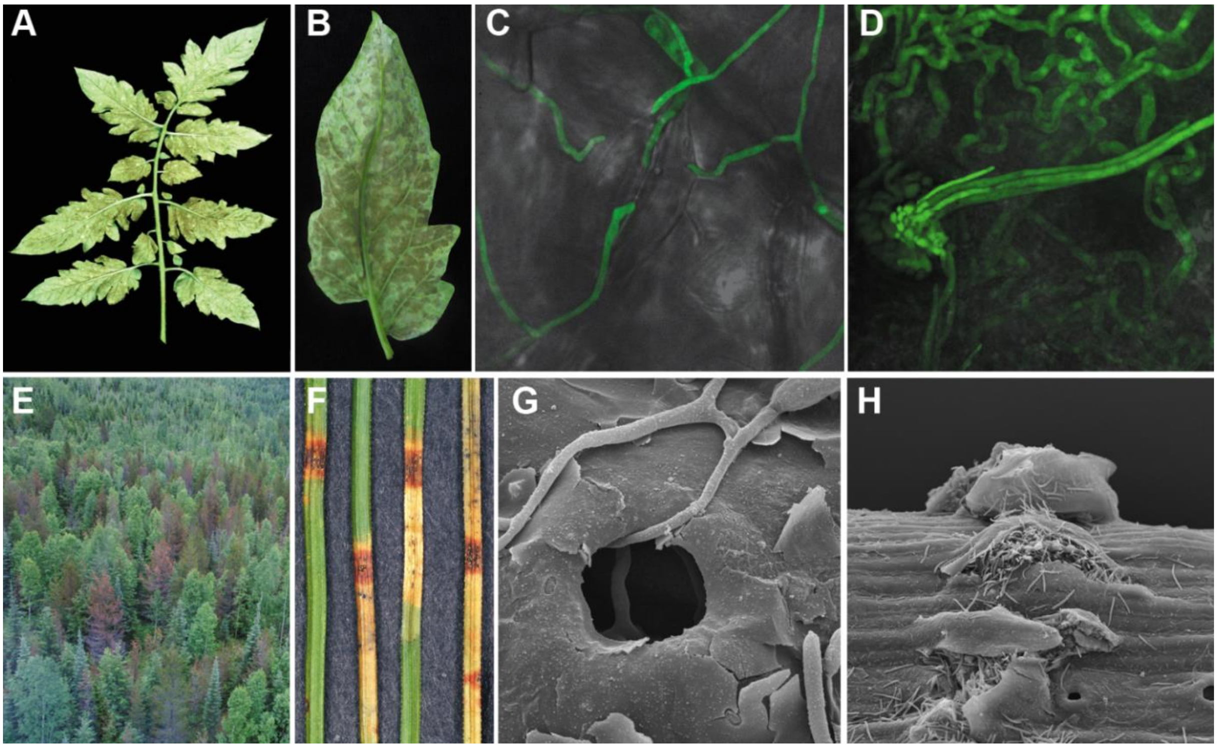


Figure 1. Symptoms caused by *Cladosporium fulvum* and *Dothistroma septosporum* on their host plants.

Disease symptoms of *Cladosporium fulvum* on tomato (A-D). **A)** *C. fulvum* sporulating on the lower side of a tomato (*Solanum lycopersicum*) leaf two weeks post inoculation; **B)** Close-up of *C. fulvum* sporulating on the lower side of a single leaflet two weeks post inoculation; **C)** Runner hyphae of *C. fulvum* (GFP-transgenic strain) at the surface of the leaf; two of them are penetrating a stoma of tomato four days post inoculation; **D)** Conidiophores of *C. fulvum* (GFP-transgenic strain) emerging from a stoma at 10 days post inoculation. Disease symptoms of *Dothistroma septosporum* on pine (E-H). **E)** Mortality of mature lodgepole pines (*Pinus contorta* var. *latifolia*) in northwest British Columbia, Canada caused by *D. septosporum*; **F)** Red band lesions with conidiomata on *Pinus radiata* needles; **G)** Penetration of hypha into stoma of *P. radiata* needle 4 weeks post inoculation; **H)** Eruption of conidiospores through epidermis of pine needle 8 weeks post inoculation.

There is no evidence that *C. fulvum* has an active sexual cycle, although both mating type idiomorphs occur in its global population [21]. Although its lifecycle is also predominantly asexual, *D. septosporum* is known to be sexually active in some parts of the world. The sexual stage *Mycosphaerella pini* Rostr. (syn *Scirrhia pini* Funk & Parker) has been reported in some forests in Europe and North America but has not yet been found in other regions, such as South Africa or the United Kingdom, even though both mating types are known to be present [22]. The rare sightings of the sexual stage are due partly to difficulties in identification, but also reflect findings from population studies that show mixed modes of reproduction with a significant clonal component [23,24]. So far, attempts to induce a sexual cycle between opposite mating types of *D. septosporum* in culture in our laboratory or others (Brown, unpublished data) have failed. Further research is required to determine environmental conditions conducive to sexual reproduction. The *D. septosporum* isolate whose genome was sequenced is derived from a clonal population with a single mating type that was introduced into New Zealand in the 1960s [22,25].

The *C. fulvum*-tomato interaction complies with the gene-for-gene model [2,26]. During infection *C. fulvum* secretes effector proteins into the apoplast of tomato leaves which function not only as virulence factors, but also as avirulence (Avr) factors when recognized by corresponding tomato Cf resistance proteins. This recognition leads to Cf-mediated resistance that often involves a hypersensitive response (HR) preventing further ingress of the fungus into its host plant tomato [8]. To date many cysteine-rich effectors have been cloned from *C. fulvum,* including Avr2, Avr4, Avr4E and Avr9, that can trigger Cf-2-, Cf-4-, Cf-4E-, and Cf-9-mediated resistance, respectively, and Ecps (*e*xtra*c*ellular *p*roteins) like Ecp1, Ecp2, Ecp4, Ecp5 and Ecp6 that trigger Cf-Ecp-mediated resistance [27,28,29]. Specific functions for some *C. fulvum* effectors have been determined: Avr4 is a chitin-binding protein that protects fungi against the deleterious effects of plant chitinases [30,31]; Ecp2 is a virulence factor that occurs in many fungi [32,33] and Ecp6 sequesters chitin fragments released from fungal cell walls by chitinases during infection thereby dampening their potential to induce pathogen-associated molecular pattern (PAMP)-triggered immunity [28]. Initially, the Avr and Ecp effectors seemed unique to *C. fulvum,* but in recent years homologs of Avr4, Ecp2 and Ecp6 with functions in virulence have been found in other fungal genomes, including members of the *Dothideomycetes* [27,28,33].

Whilst most studies of *C. fulvum* have focused on effectors and their interactions with components in both resistant and susceptible plants, studies of *D. septosporum* have instead focused on dothistromin, a toxin produced by the fungus that accumulates in infected pine needles. Dothistromin is a broad-spectrum toxin with structural resemblance to a precursor of the highly toxic and carcinogenic fungal metabolite, aflatoxin [34]. Although dothistromin is not essential for pathogenicity [35], recent observations suggest it to be a virulence factor, affecting lesion size and spore production (Kabir and Bradshaw, unpublished data). Some dothistromin biosynthetic genes were identified in *D. septosporum* but unexpectedly they were in several mini-clusters rather than in one co-regulated cluster of genes as reported for aflatoxin producing species of *Aspergillus* [36,37,38]. The similarity of dothistromin to aflatoxin enabled predictions to be made about other *D. septosporum* genes involved in dothistromin production [39]: the complete set of dothistromin genes will help us to understand the evolution of dothistromin and aflatoxin gene clusters.

Here we report the sequence and comparison of the genomes of *C. fulvum* and *D. septosporum* which have very similar gene contents but differ significantly in genome size as a result of different repeat contents. We found unexpectedly high levels of similarity in genes previously studied in one or other of these fungi, including those encoding Avr and Ecp effectors of *C. fulvum*, and dothistromin toxin genes of *D. septosporum*. Surprisingly, compared to *D. septosporum*, *C. fulvum* has higher numbers of genes normally associated with a necrotrophic or hemibiotrophic lifestyle such as genes for carbohydrate-degrading enzymes and secondary metabolite biosynthesis. However, in *C. fulvum* some of these genes were lowly or not expressed *in planta* and others were pseudogenized. Other *C. fulvum* genes that are absent in *D. septosporum* are putatively involved in virulence on its host plant tomato, such as the α-tomatinase gene. We suggest that regulation of gene expression and pseudogenization, in addition to evolution of new genes, are important traits associated with adaptation to different hosts and lifestyles of the two fungi that, however, also retained some signatures of their common ancestral host.

Results and Discussion

C. fulvum and *D. septosporum* are closely related species with very different genome sizes

The 30.2 Mb genome of *D. septosporum* (http://genome.jgi.doe.gov/Dotse1/Dotse1.home.html; GenBank AIEN00000000) was sequenced at 34-fold coverage (Table S1) and then assembled into 20 scaffolds (>2 kb), 14 of which have telomere sequences at one or both ends and mostly match chromosome sizes estimated from pulsed-field gel electrophoresis (Table S2; [36]). The excellent assembly of the *D. septosporum* genome was facilitated by its very low repeat content of only 3.2% (Table 1; Table 2; Protocol S1). In contrast*,* the repeat-rich genome of *C. fulvum* (http://genome.jgi-psf.org/Clafu1/Clafu1.home.html; GenBank number to be provided) was very difficult to assemble. Fourteen 2 kb paired-end or shotgun 454 sequencing runs for *C. fulvum* resulted in a 21-fold coverage of the 61.1 Mb assembly in 2664 scaffolds >2 kb (Table 1) with a total repeat content of 47.2% (Table 2). The sequencing strategy was initially based on the assumption of a genome size of around 40 Mb, but soon it appeared that the *C. fulvum* genome was much larger due to the high repeat content. Problems with the assembly are not caused by the sequencing coverage of *C. fulvum* because it is estimated to be sufficiently high for a good coverage of the gene encoding areas. Instead, they are a consequence of its high repeat content. An estimated additional 26 Mb of *C. fulvum* DNA reads could not be assembled as they were predominantly repeat sequences (Fig. 2). In the remainder of the manuscript we refer to chromosomes (1 to 14) for *D. septosporum* and scaffolds for *C. fulvum.* Summary statistics for the two genomes are shown in Table 1 and at the Joint Genome Institute (JGI) Genome portal (jgi.doe.gov/fungi) [40]. The *C. fulvum* and *D. septosporum* genomes are predicted to encode approximately 14 and 12.5 thousand gene models, respectively. Nevertheless, the *C. fulvum* and *D. septosporum* genomes share more than 6,000 homologous gene models with at least 80% similarity at the predicted amino acid level, whereas this number drops to 3,000 gene models this similar when comparing *C. fulvum* with other closely related *Dothideomycete* species such as *Mycosphaerella graminicola* and *M. fijiensis* (Fig. 2). Similarly, most introner-like element clusters found in *C. fulvum* and *D. septosporum* are closely related, more than to elements in other *Dothideomycetes* [41].

Table 1. Cladosporium fulvum and Dothistroma septosporum genome statistics.

Species	Genome assembly size (Mb)	Number of predicted gene models	Sequencing coverage depth	Number of scaffolds[a]	Scaffold L50 / N50[b] (Mb)	% GC content
Cladosporium fulvum	61.1	14127	21	2664 >2kb 279 >50kb	250/0.06	48.8
Dothistroma septosporum	30.2	12580	34	20 >2kb 14 >50kb	5.0/2.6	53.1

[a] For details on statistics see Supporting Protocol S1.

[b] L50 is defined as the smallest number of scaffolds that make up 50% of the genome; N50 is defined as the size (Mb) of the smallest of the L50 scaffolds.

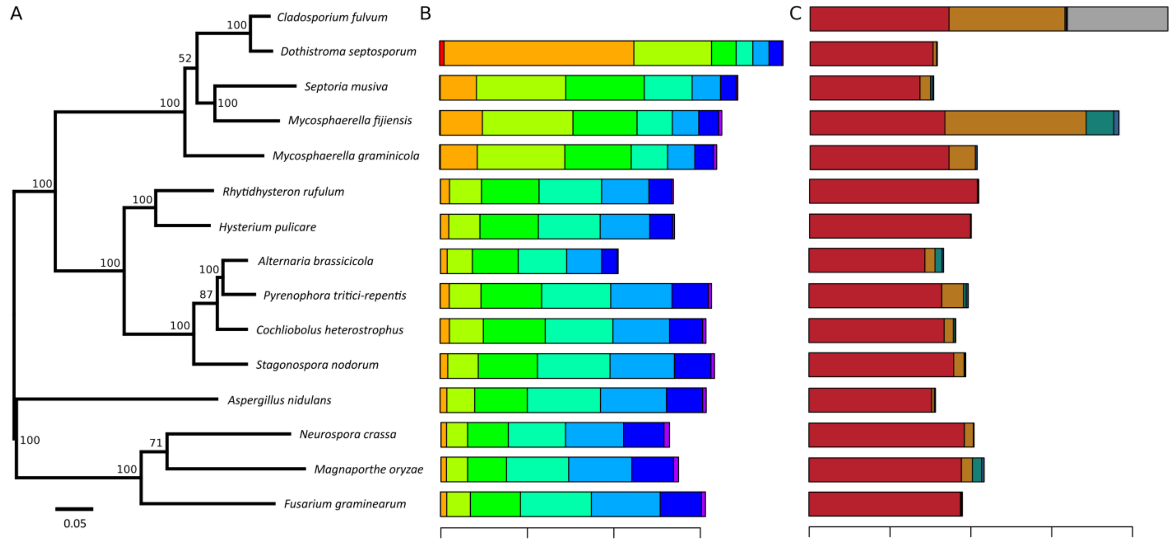


Figure 2. Species phylogeny, amino acid similarity and repeat content.

A) Maximum likelihood phylogenetic tree based on 51 conserved protein families showing evolutionary relationships of *Cladosporium fulvum* and *Dothistroma septosporum*. Branch lengths are indicated by the bar (substitutions/site); bootstrap values are shown as percentage. **B)** Genome-wide amino acid similarity of homologous proteins between *C. fulvum* and other sequenced fungal species. A pair of proteins is only reported as homologous when the predicted similarity (blastp) spans at least 70% of their lengths and their length difference is at most 20%. Axis indicates number of homologous proteins. Bar shading indicates similarity: red, 91-100%; orange, 81-90%; light green, 71-80%; medium green, 61-70%; turquoise, 51-60%; light blue, 41-50%; dark blue, 31-40%; and purple, 0-30%. Homologous proteins with high amino acid similarity are likely orthologs, whereas for those with lower similarity this relation cannot be inferred. **C)** Repeat content of *C. fulvum*, *D. septosporum* and other sequenced fungal species. Bar shading indicates repeat class: red, unique non-repeat regions; brown, repeat elements; green, continuous tracts of N characters; blue, duplicated regions; and grey, poorly assembled regions (*C. fulvum* only). Axis indicates number of nucleotides (Mb).

Phylogenetic analysis of *C. fulvum* and *D. septosporum* genomes in the context of nine other *Dothideomycetes* (Ohm et al., unpublished data) confirms that these two species are the most closely related of the sampled species (Fig. 2), as was inferred earlier from ITS [3] and mating type sequences [21]. This gives us two very closely related genomes with drastically different genome sizes mostly due to the greatly increased repeat content of *C. fulvum*.

C. fulvum and *D. septosporum* differ in content and classes of repeats that are affected by repeat-induced point mutation

The massive increase in repetitive elements in *C. fulvum* might result from expansion of one or more repeat families that are also present in *D. septosporum*. Therefore, we classified the different repeat families in *D. septosporum* and compared them with those in *C. fulvum.* This revealed that some of the repetitive element families present in *D. septosporum* have expanded in *C. fulvum* (Table 2). This is most remarkable for the Class I retrotransposons which comprise over 90% of the repetitive fraction in *C. fulvum* and together account for over 26 Mb of the assembled genome. Retrotransposons are also highly abundant in the large repeat-rich genome of the hemibiotrophic sexual pathogen *Mycosphaerella fijiensis* (Kema et al., unpublished data). Both Copia and Gypsy LTR retroelements are expanded in *C. fulvum* compared to the *D. septosporum* genome, whereas LINEs are detected only in *C. fulvum* (Table 2). Some other fungal species that are closely related to each other, but have a different lifestyle, also differ in repeat content, such as the Leotiomycetes of which *Botrytis cinerea* (<1% repeats) and *Sclerotinia* *sclerotiorum* (7% repeats) are necrotrophs, while *Blumeria graminis* f. sp. *hordei* (64% repeats) is an obligate biotroph [42,43]. The latter species is particularly enriched in Class I elements and one of several biotrophs that show expansion of genome size associated with high repeat content [43,44].

Table 2. Repetitive elements in the *Cladosporium fulvum* and *Dothistroma septosporum* genomes.

		Cladosporium fulvum			*Dothistroma septosporum*		
Repeat Class	**Repeat type**	**Total bases covered**	**Percent of genome**[a]	**Percent of repetitive fraction**[a]	**Total bases covered**	**Percent of genome**	**Percent of repetitive fraction**[a]
Class I	Ty1-Copia LTR	3,208,305	5.3	11.1	43,598	0.1	4.5
Retrotransposon s	Ty3-Gypsy LTR	13,557,457	22.2	47	178,860	0.6	18.6
	Misc. LTR	0	0	0	167,893	0.6	17.5
	LINE	9,464,476	15.5	32.8	0	0	0
	Sub-total	26,230,238	42.9	90.9	390,351	1.3	40.6
Class II	DDE-1	960,373	1.6	3.3	0	0	0
DNA transposons	hAT	144,765	0.2	0.5	0	0	0
	Helitron	14,901	0.02	0.1	384,233	1.3	40
	Mariner	96,124	0.2	0.3	57,646	0.2	6
	MITE	98,047	0.2	0.3	2,165	0.01	0.2
	MuDR_A_B	50,899	0.1	0.2	0	0	0
	Sub-total	1,365,109	2.2	4.7	444,044	1.5	46.2
Unclassified	Unclassified	1,255,529	2.1	4.4	127,027	0.4	13.2
TOTAL		28,850,876	47.2		961,422	3.2	

[a]Numbers refer to repeats present in assembled genome.

In contrast to the retroelements, Class II DNA transposons comprise only a small percentage (4.7%) of the overall repetitive elements in *C. fulvum*, but 46.2% of the repeats in *D. septosporum,* although they make up only a small portion of the genome overall. Interestingly, helitron-like DNA transposons comprise 40% of all repeats in *D. septosporum* and are 25.8-fold higher in terms of sequence coverage than in *C. fulvum*, whereas the DDE-1, hAT, and MuDR_A_B DNA transposons present in *C. fulvum* are not present in *D. septosporum*. Helitrons are transposons that replicate by a rolling-circle mechanism and are found in a wide range of eukaryotes, including the white rot fungus *Phanerochaete chrysosporium* [45], and are thought to have a role in genome evolution [46]. Helitron-like repeats are particularly abundant on *D. septosporum* chromosomes 3, 6 and 11 (Fig. 3) and usually occur in clusters, both with other helitron-like repeats and with other types of repetitive elements.

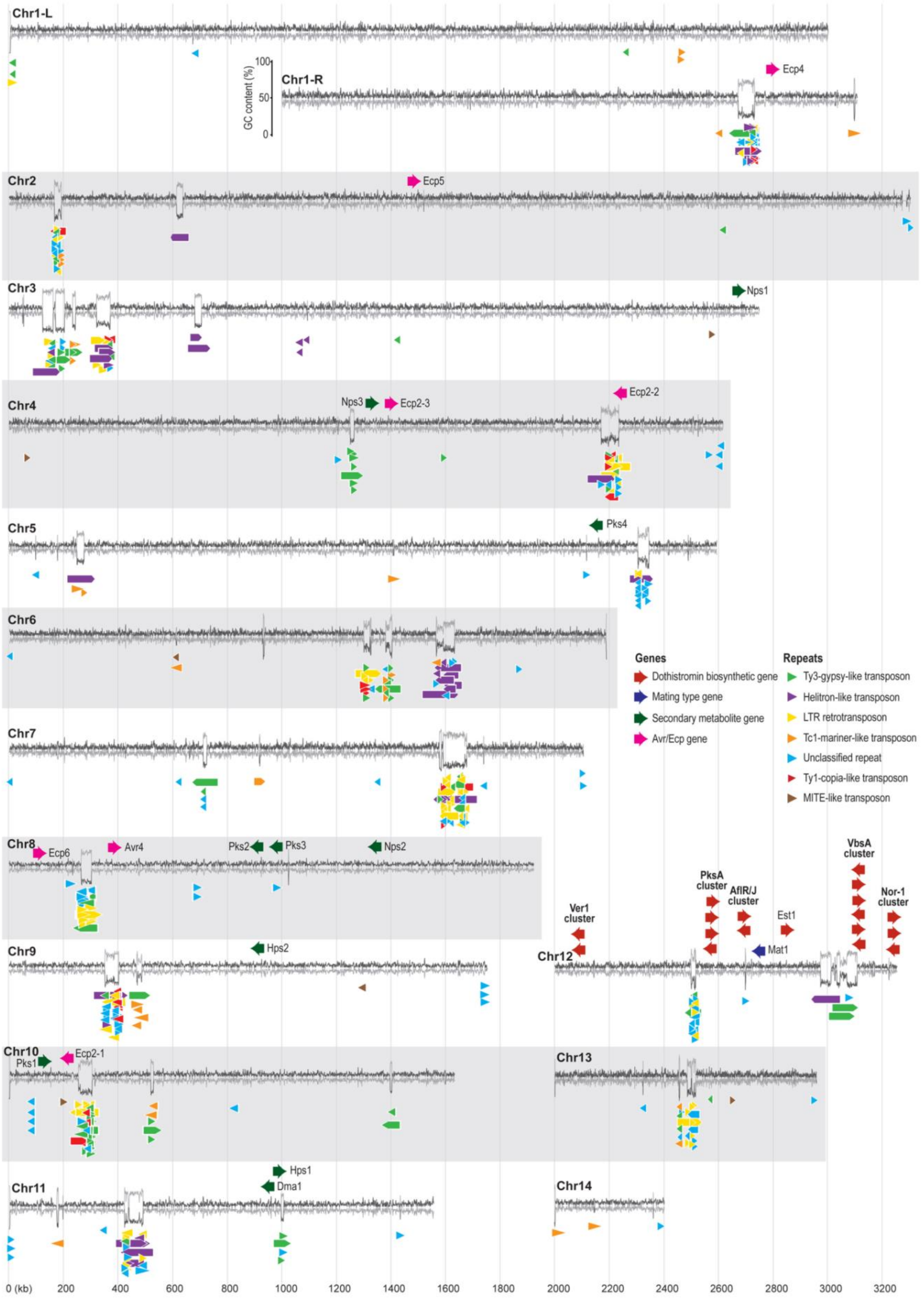


Figure 3. Organization of repeats and pathogenicity-related genes in the Dothistroma septosporum genome.

The fourteen chromosomes from the *D. septosporum* genome assembly are shown as GC (dark grey line) and AT (pale grey line) content (%) plots made from a 500-bp sliding window using Geneious (www.geneious.com). All chromosomes have telomere sequence at both ends except chromosomes 2, 11 and 14 which have telomere sequences only at the left end as shown in the Figure. Chromosome 1 has been split into two parts (L, R) because of its length, and the GC/AT content scale is shown beside the right arm of this chromosome. The positions of putative *Avr* and *Ecp* effector, secondary metabolite, dothistromin biosynthesis, and mating type genes are shown above the GC/AT content plot, while the positions of repeats (>200 bp) are shown below the plot. Color-coding of the gene and repeat types is indicated in the legend. The presence of repeat clusters at one or two sites per chromosome for most chromosomes is evident, and these coincide with regions of high AT content. The chromosome sizes are to scale, as indicated by the vertical pale grey lines, with the values (in kb) shown at the bottom; neither the genes nor the repeats are drawn to scale.

The organization of repeats in *D. septosporum* is striking in that for the majority of the chromosomes, most repeats are localized into just one or two large regions containing a mixture of repeat element types (Fig. 3), although other small repeat clusters also occur. In many eukaryotes, centromeres are characterized by repetitive DNA [47], and therefore we propose that some of the larger complex repeat regions are centromeres, in line with similar suggestions made for other fungal genomes [48], although experimental confirmation is required. The absence of any repeat cluster from chromosome 14, along with the observation that it harbors only one telomere, suggests that it is a chromosome fragment.

Repeats in fungi are affected by *r*epeat-*i*nduced *p*oint mutation, also referred to as RIP, a defense mechanism employed by fungi to suppress transposable element activity that was first described in *Neurospora crassa* [49]. RIP is a process by which DNA accumulates G:C to A:T transition mutations. It occurs during the sexual stage in haploid nuclei after fertilization but prior to meiotic DNA replication. Clear evidence of RIP was found in both the *C. fulvum* and *D. septosporum* genomes (Table 3) and is mainly confined to repeat-rich regions. In total 25.9 Mb were RIP'd in *C. fulvum* and 1.1 Mb in *D. septosporum*, which represent 42.4% and 3.7% of their genomes, respectively. RIP occurred mainly on large repeated sequences (≥ 500 nucleotides) that represent 97.2% of all repeats in *C. fulvum* and 98.0% in *D. septosporum* (Table 3). The high rate of RIP in repeat regions is in the same range as that seen in other *Dothideomycetes* such as *S. nodorum* (97.2%; Table S3) [50]. Although RIP is present at high levels in *C. fulvum*, we propose that it has not been able to prevent transposon expansion possibly due to very rare sexual activity.

Of the RIP'd loci, *C. fulvum* has almost none (0.5%) and *D. septosporum* little (16.9%) outside the main classified repeat regions. This is strikingly different from *N. crassa* (Table S3), where 35.2% of all RIP'd loci are predicted to be non-repeat-associated. For *N. crassa* it has been shown that even single gene duplication events are prey to the RIP machinery, thereby exemplifying its efficiency and sensitivity [49]. Clearly such sensitivity is not applicable to *C. fulvum* and *D. septosporum,* neither for three other studied *Dothideomycetes* (Table S3). In the *Dothideomycete* phytopathogenic fungus *Leptosphaeria maculans*, RIP slippage is found in regions adjacent to repetitive elements. In that species RIP has occurred in genes encoding small secreted proteins, such as the effector genes *AvrLm6* [51] and *AvrLm1* [52] that are located in repeat-rich regions of the genome [53]; mutations in these genes caused by the RIP process enabled the fungus to overcome *Lm6* and *Lm1*-mediated resistance, respectively. However, we found no evidence of RIP slippage into the known effector genes of *C. fulvum* and related effector genes in *D. septosporum.*

Table 3. Occurrence of Repeat-Induced Point Mutation (RIP) signatures and repeats in *Cladosporium fulvum* and *Dothistroma septosporum.*

	Cladosporium fulvum		*Dothistroma septosporum*	
	Bases (kb)	**Loci**[a]	**Bases (kb)**	**Loci**[a]
RIP'd sequence in genome	25,882 (42.4%)[b]	5447	1,114 (3.7%)[b]	65
Repeat sequence ≥500nt in genome[c]	27,170 (44.5%)[b]	7101	798 (2.6%)[b]	133
Repeat sequence ≥500nt with RIP signature[d]	26,397 (97.2%)[e]	6506 (91.6%)[e]	782 (98.0%)[e]	114 (85.7%)[e]

[a] Loci are defined as consecutive blocks of sequence assigned as repeats or RIP'd regions.

[b] Percentage of the genome is indicated in brackets.

[c] Only classified repeats (as shown in Table 2) were considered.

[d] Repeated sequence ≥500nt that (at least partially) overlaps with RIP'd sequence.

[e] Percentage of classified repeats is indicated in brackets.

The *C. fulvum* and *D. septosporum* genomes show extensive intrachromosomal rearrangements.

One way to assist the assembly of a fragmented genome is to use synteny with a well-assembled genome of a closely related species to order the scaffolds [54]. We attempted to use the *D. septosporum* genome to improve the *C. fulvum* assembly in this way. However, although it was possible to map *C. fulvum* scaffolds onto the assembled *D. septosporum* genome (Fig. S2A), individual *C. fulvum* scaffolds are not collinear along their length, but have only short blocks of synteny to different parts of the *D. septosporum* chromosomes. The syntenic regions of the *C. fulvum* and *D. septosporum* genomes are associated with just 461 of the *C. fulvum* scaffolds (Table 4). In contrast, the remaining >4,000 *C. fulvum* scaffolds are non-syntenic. A more detailed analysis with the ten largest *C. fulvum* scaffolds (two are shown in Fig. S2B) revealed that they each match primarily to only one *D. septosporum* chromosome, suggesting predominantly intrachromosomal rearrangements (mesosynteny), as described for other *Dothideomycete* fungi [55] (Ohm et al., unpublished data; Condon et al., unpublished data). As found in other fungi [42,56] non-syntenic regions are repeat-rich; for *C. fulvum* 79.7% of the repeat sequences are present in non-syntenic regions (Table 4).

Table 4. Syntenic and non-syntenic regions between *Cladosporium fulvum* and *Dothistroma septosporum* are unevenly distributed over the *C. fulvum* scaffolds.

Feature	Syntenic	Non-syntenic	Total	Syntenic %	Non-syntenic %
Number of scaffolds	461	4,404	4,865	9.5	90.5
Number of repeats	2,090[a]	6,234	8,324	25.1	74.9
Mb in scaffolds	37.4[b]	23.7	61.1	61.2	38.8
Mb in repeats	5.6[c]	21.9	27.5	20.3	79.7
Mb in whole genome	22.3[b]	38.8	61.1	36.5	63.5

[a]Number of repeat regions on syntenic vs. non-syntenic scaffolds.

[b]A syntenic scaffold is one that contains at least a single syntenic block, but may not be syntenic along its entire length. Total syntenic scaffold size (37.4 Mb) is therefore larger than total syntenic size in whole genome (22.3 Mb).

[c]Summed repeat length on syntenic vs. non-syntenic scaffolds.

Non-syntenic, repeat-rich regions are enriched in genes encoding secreted proteins

Secreted proteins are important for communication of plant-pathogenic fungi with their hosts. They comprise not only enzymes required for penetration and growth on plant cell walls, but also proteins needed to compromize the basal defence system of plants by either suppressing or attacking it, as has been reported for several fungal effector proteins [57]. The percentage of proteins predicted to be secreted is similar for both *C. fulvum* (8.5%) and *D. septosporum* (7.2%), and in the same range as that predicted for other *Dothideomycete* fungi such as *M. graminicola* (9.1%) and *S. nodorum* (10.8%) (jgi.doe.gov/fungi) [40,50,58] (Ohm et al., unpublished data).

Genes encoding secreted proteins including effectors are subject to evolutionary selection pressure imposed by environmental and host plant factors [57], and they often show a high level of diversification. Repeat-rich, gene-poor regions have been proposed to contain genes involved in adaptation to new host plants. For example, in some *Phytophthora* species and in *L. maculans* significantly higher proportions of *in planta*-induced species-specific effector genes encoding secreted proteins are found in repeat-rich compared to repeat-poor regions [50][59] and in pathogenic strains of *Pyrenophora tritici-repentis* transposable elements are associated with effector diversification (Ciuffetti et al., unpublished data). We hypothesized that we would find more genes encoding secreted proteins in repeat-rich regions that are less syntenic between the *C. fulvum* and *D. septosporum* genomes than in repeat-poor syntenic regions.

Table 5. Location of gene models of *Cladosporium fulvum* in regions syntenic or non-syntenic with the *Dothistroma septosporum* genome.

	Number of gene models[a]			Percentage of gene models[b]	
	Total	Syntenic	Non-syntenic	Syntenic	Non-syntenic
All proteins[c]	14,127	9,890	4,237	70	30
BDBH[d]	11,092	8,900	2,192	89.9[g]	51.7[g]
Secreted proteins[e]	1,195	754	441	7.6[h]	10.4[h]
Secreted Cys-rich proteins[f]	271	151	120	1.5	2.8

[a] Values are numbers of gene models located in regions of the *C. fulvum* genome syntenic or non-syntenic with the *D. septosporum* genome as described in materials and methods.

[b] For all proteins the percentage of gene models represents the fraction of all gene models present in the *C. fulvum* genome; for other categories (BDBH, secreted proteins, secreted Cys-rich proteins) the percentage of gene models represents the fraction of gene models present in syntenic and non-syntenic regions.

[c] All proteins encoded by predicted gene models in the *C. fulvum* genome.

[d] Bi-directional best BLAST hit between *C. fulvum* and *D. septosporum* proteins with at least 50% (global) pairwise amino acid similarity and at least 60% coverage by overlap-corrected blastp HSPs.

[e] Gene models predicted to encode secreted proteins.

[f] Secreted small cysteine-rich proteins contain less than 300 amino acids of which at least four are cysteines.

[g] The mean amino acid similarities of all protein gene models in syntenic regions and non-syntenic regions are 85.2% and 65.1%, respectively.

[h] The mean amino acid similarities of secreted protein gene models in syntenic regions and non-syntenic regions are 81.1% and 60.7%, respectively.

We therefore compared the number of genes and their similarity at the nucleotide and protein levels in syntenic and non-syntenic regions of these two genomes (Table 5) using *C. fulvum* as the reference sequence due to its higher overall content of repeat elements in non-syntenic regions. The regions syntenic between *C. fulvum* and *D. septosporum* representing 22.3 Mb of the *C. fulvum* genome contain 70% of all predicted genes whereas 30% of the genes are located in the non-syntenic repeat-rich regions representing 38.8 Mb of the *C. fulvum* genome (Table 5). The syntenic regions contain most of the homologous genes that encode proteins with the highest level of conservation between the two genomes, whereas the proteins encoded by genes located in the non-syntenic repeat-rich regions are less conserved. In syntenic regions, 89.9% of gene models have a bi-directional best BLAST hit (BDBH) to a *D. septosporum* gene model, with a mean predicted amino acid similarity of 85.2%, compared to non-syntenic with only 51.7% of gene models with BDBH and 65.1% amino acid similarity (Table 5). As expected, we found the repeat-rich non-syntenic regions to have higher proportions of gene models encoding secreted proteins (10.4%, with a mean predicted amino acid similarity of 60.7%) and small secreted cysteine-rich proteins (2.8%) than in syntenic regions (7.6%, with a mean amino acid similarity of 81.1%, and 1.5% respectively) (Table 5), as has been reported for *L. maculans* [53].

C. fulvum and *D. septosporum* share functional effectors

Some *C. fulvum* effector homologs have previously been reported to occur in other *Dothideomycete* species including *M. fijiensis, M. graminicola,* and several *Cercospora* species [33] but in the *D. septosporum* genome we found the highest number of *C. fulvum* effector homologs discovered to date, including *Avr4, Ecp2-1, Ecp2-2, Ecp2-3, Ecp4, Ecp5* and *Ecp6*. Of those, Avr4, Ecp2-1 and Ecp6 are core effectors [33] and show the highest identity (51.7%, 59.8% and 68.6% amino acid identity, respectively) with those present in *C. fulvum,* whilst *Ecp4* and *Ecp5* are pseudogenized. We were interested to know whether the *D. septosporum* effectors would be functional in triggering a Cf-mediated hypersensitive response (HR). Therefore we inoculated tomato plants MM (cv. Moneymaker) carrying the Cf-Ecp2 resistance trait with *Agrobacterium tumefaciens* expressing potato virus X (PVX) containing *D. septosporum Ecp2-1* and used PVX-containing *C. fulvum Ecp2-1* as a positive control. *D. septosporum Ecp2-1* triggered a Cf-Ecp2-1-mediated HR (Fig. 4A), whilst MM tomato plants lacking Cf-Ecp2 did not show any HR when inoculated with PVX containing *Ds-Ecp2-1* (results not shown). We also showed that the *D. septosporum* homolog of *C. fulvum Avr4* is functional in triggering a Cf-4-mediated HR in *Nicotiana benthamiana* as determined with an *Agrobacterium* transient transformation assay (Fig. 4B). This is remarkable because *D. septosporum* infects a gymnosperm which is only distantly related to tomato, but apparently produces effectors that can be recognized by tomato Cf resistance proteins. It would be interesting to examine whether gymnosperms carry functional homologs of the well-studied *Cf* tomato resistance gene homologs [33,60] or other major *R* genes that could confer resistance to *Dothistroma* spp.. Major *R* genes have been shown to be involved in resistance of some pine species to *Cronartium* spp. rust pathogens [61,62] and are thought to function in a gene-for-gene manner [63].

In *C. fulvum,* adaptation to resistant tomato cultivars is sometimes associated with deletion of effector genes [64]. Presence of repeats or location near a telomere can cause repeat-associated gene deletion [65]. We analyzed the location of all cloned *C. fulvum* effector genes in its genome. Many scaffolds containing an effector gene are very small (Fig. 5), suggesting that they are surrounded by large repeats hampering assembly into larger scaffolds. The location of the *C.*

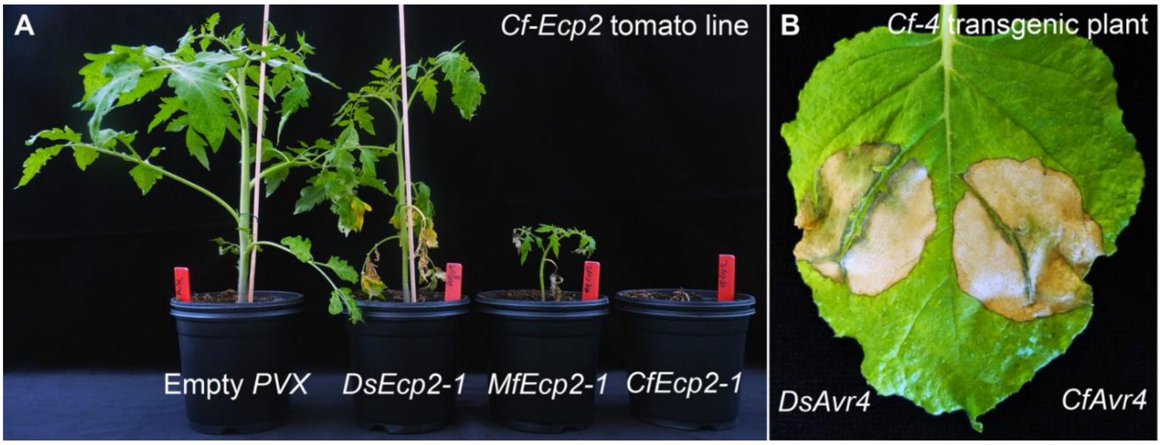


Figure 4. Recognition of *Dothistroma septosporum* effectors by tomato Cf receptors.

A) *Ds-Ecp2-1*, the *D. septosporum* ortholog of *Cf-Ecp2-1,* was cloned into pSfinx (PVX vector). Tomato plants were inoculated with *Agrobacterium tumefaciens* transformants expressing *PVX::DsEcp2-1*. A hypersensitive response (HR) was induced in the tomato line carrying the *Cf-Ecp2* resistance gene (MM-Cf-Ecp2). Empty vector was used as a negative control and caused only mosaic symptoms. Pictures were taken at four weeks post inoculation. **B)** The *C. fulvum* avirulence gene *Avr4* (*Cf-Avr4*) and its ortholog in *D. septosporum (Ds-Avr4*) were heterologously expressed in *Cf-4* transgenic *Nicotiana benthamiana* using the *A. tumefaciens* transient transformation assay (ATTA). Expression of *Cf-Avr4* and *Ds-Avr4* results in an HR demonstrating that Ds-Avr4 is recognized by the tomato Cf-4 receptor. Picture was taken at six days post inoculation.

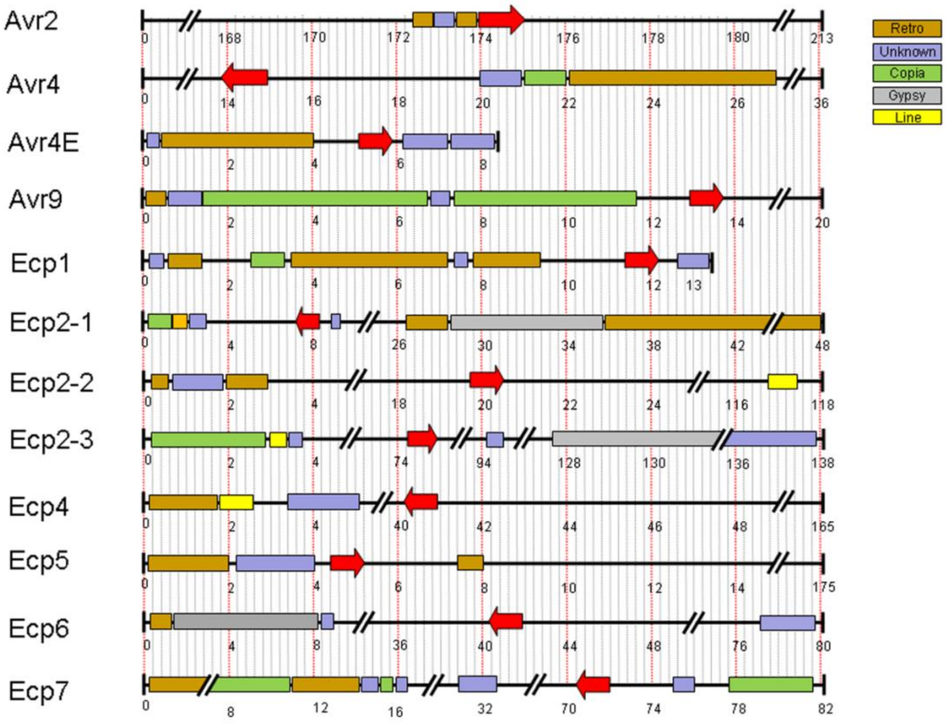


Figure 5. Repetitive regions flanking known effectors of Cladosporium fulvum.

Scaffolds harboring sequenced *C. fulvum* effector genes with flanking repeats. Repeat regions longer than 200 bp are shown, with different types indicated in different colour code. The effector genes are depicted by red arrows and sizes are in kb. The sizes of scaffolds range from 8 to 213 kb but some are shortened to fit the figure due to differences in size.

fulvum effectors is shown in Fig. 5 and the types of flanking repeats are detailed in Table S4. The well-characterized effector gene *Avr9* is located on a very small (20 kb) scaffold (Fig. 5) and is likely flanked on both sides by repeats; on one side there are 11 kb of repeats on the scaffold and on the other side probably also repeats just outside the region shown that prevented further scaffold assembly. This suggests that the absolute correlation found between deletion of the *Avr9* gene in *C. fulvum* and overcoming Cf-9-mediated resistance [64] is most likely due to the close proximity of *Avr9* to large, unstable repeat regions. As well as causing deletions, transposons can contribute to genome plasticity by mutation due to transposition into coding sequences. During co-evolution transposons have inserted into effector genes causing their inactivation and overcoming Cf-mediated resistance in *C. fulvum*, as has been reported for inactivation of both *Avr2* [64] and *Avr4E* [66]. The *C. fulvum* homologous effector genes present in *D. septosporum* are also often in close proximity to repeat-rich areas that may represent centromeres (Fig. 3), but the biological significance of this is not yet clear. Pseudogenization of two *D. septosporum* effector genes, *Ecp4* and *Ecp5*, homologous to those reported for *C. fulvum* [67], could point to host adaptation in the DNB fungus at the pine genus, species or cultivar level. Future population analysis of both fungal strains and host genotypes will reveal the mechanism behind this phenomenon.

New hydrophobin genes in *C. fulvum*

Another class of well-studied *C. fulvum* small cysteine-rich secreted proteins are the hydrophobins. These amphipathic proteins are implicated in developmental processes in filamentous fungi and are localized on the outer surface of fungal cell walls [68]. They are divided into class I and class II hydrophobins based on sequence differences that also correlate with their different solubility [68].

Six hydrophobin genes (*Hcf-1 to Hcf-6*) had previously been identified from *C. fulvum* [66, 67]. We identified five additional hydrophobin genes in the *C. fulvum* genome [two class I (Cf187601 and Cf189770) and three class II (Cf197052, Cf188363 and Cf183780)] (Fig. S3), which makes *C. fulvum* the Ascomycete species with the largest number of hydrophobin genes reported so far. In the *D. septosporum* genome only four hydrophobin genes were found, one of which (Ds75009) is predicted to encode a class II hydrophobin and was highly expressed both in culture and *in planta*. Based on EST data the 11 *C. fulvum* hydrophobin genes show a range of different expression patterns. Of the six class I *C. fulvum* hydrophobins two were only expressed in culture [Cf184635 (*Hcf-2*) and Cf189850 (*Hcf-4*)], three were expressed both in culture and *in planta* [Cf189770, Cf187601 and Cf193176 (*Hcf-1*)], and one was not expressed in culture or *in planta* [Cf184193 (*Hcf-3*)]. Three of the class II *C. fulvum* hydrophobins were only expressed in culture [Cf197052, Cf188363 and Cf193013 (*Hcf-5*)], whilst Cf193331 (*Hcf-6*) and Cf183780 were expressed neither in culture nor *in planta*. None of the *C. fulvum* hydrophobin genes were expressed *in planta* only. It has been proposed that hydrophobins may act as ‘stealth’ factors, preventing the invading fungus from detection by its host plant [69] or protecting it against deleterious effects of plant chitinases and β-1,3 glucanases as reported for *C. fulvum* [70]. Early functional studies focused on the hydrophobin genes *Hcf-1* (Cf193176) and *Hcf-2* (Cf184635). Knocking-down expression of *Hcf-1, Hcf-2*, or both genes by homology-dependent gene silencing did not compromise virulence [71,72]; a similar result was reported for knock-down mutants of class I *Hcf-3* and *Hcf-4* and class II *Hcf-6* genes [73]. The phylogenetic tree (Fig. S3) shows that the four class I genes (*Hcf-1* to *Hcf-4*) are paralogs, suggesting functional redundancy that might explain the lack of a phenotype; functional redundancy may also exist between different classes.

It would be interesting to examine the role in virulence of the two most similar hydrophobin class I and class II genes of *C. fulvum* and *D. septosporum* (Cf 189770/Ds67650 and Cf197052/Ds75009, respectively) either by knock-out or knock-down strategies.

Carbohydrate active enzyme gene and expression profiles reflect adaptation to different host plants.

Because *C. fulvum* and *D. septosporum* have very different plant hosts and pathogenic lifestyles, we expected that their capacity to degrade carbohydrates would also differ and that this might be reflected in their gene complements and expression profiles. We compared numbers of genes predicted to encode carbohydrate-active enzymes (CAZymes) [74] in these two fungi to those in other fungi representative of different lifestyles. As seen for grouped families of CAZyme genes in Table 6 (e.g. GH family of glycoside hydrolases), both *C. fulvum* and *D. septosporum* have gene numbers in the same range as hemibiotrophic and necrotrophic fungi, and many more than the obligate biotroph *B. graminis* f. sp. *hordei*. Despite this, both *C. fulvum* and *D. septosporum* have fewer predicted cellulolytic enzyme genes (e.g. GH6, GH7) as well as fewer genes classified in carbohydrate binding module gene families (e.g. CBM1) than most of the other fungi shown except for *M. graminicola* (Tables 6, S5). The reduced number of predicted genes for cell wall-degrading enzymes in *M. graminicola* was hypothesized to represent an adaptation to avoid host defenses during stealth pathogenicity [58], which also may apply to *C. fulvum* and *D. septosporum*. However it is known that even a small number of genes can enable high levels of enzymatic activity, as has been shown for the strongly cellulolytic fungus *Trichoderma reesei* [75].

Next we focused on CAZyme gene families that appear to differ in gene number between *C. fulvum* and *D. septosporum.* Because small differences in gene number could be due to mis-annotation, only families that differed by two or more genes were considered and examples of these are shown in Table 6 (full data in Table S5). Potentially interesting is the expansion of genes associated with pectin degradation in *C. fulvum*. For example in the GH28 family that includes many pectinolytic enzymes, *C. fulvum* has 15 genes whilst *D. septosporum* has only four. A higher pectinolytic activity in *C. fulvum* is concordant with the higher pectin content of its host, tomato, compared to the pine host of *D. septosporum* [76,77], but larger numbers of genes encoding pectin-degrading enzymes have generally been associated with a necrotrophic rather than a biotrophic lifestyle in fungi [78]. High pectinolytic activity is observed in fungi such as *Botrytis cinerea* [79,80] that invades soft, pectin-rich plant tissues causing a water-soaked appearance of the infected tissues [42]. However, during colonization of tomato leaves by *C. fulvum* this type of symptom is never observed [79,80]. Instead of contributing to the destruction of host cell walls, the *C. fulvum* pectinolytic enzymes may facilitate local modification of primary cell walls of mesophyll cells allowing the fungus to thrive in the apoplast of tomato leaves, as suggested for the ectomycorrhizal fungus *Laccaria bicolor* that thrives on plant roots [81].

Table 6. Comparison of selected CAZy gene families between *Cladosporium fulvum, Dothistroma septosporum* and five other Ascomycetes.

CAZy families[a]	Predicted function[b]	Cf [c]	Ds	Mg	Sn	Mo	Ss	Bg
		B	HB	HB	N	HB	N	B
GH family	**FCW/energy/PCW-C/H/HP/pectin**	**274**	**201**	**191**	**289**	**268**	**223**	**63**
GH3	PCW/FCW	19 (15, 8)	12 (12, 12)	16	16	18	13	1
GH5	PCW/FCW	16 (12, 2)	12 (11, 12)	9	18	13	14	3
GH6	PCW-C	0	0	0	4	3	1	0
GH7	PCW-C	2 (0, 0)	1 (1, 1)	1	5	5	3	2
GH10	PCW-H	2 (2, 1)	1 (1, 1)	2	7	7	2	0
GH28	PCW-pectin	15 (8, 0)	4 (4, 4)	2	4	3	17	0
GH31	PCW-H	15 (12, 0)	10 (9, 8)	7	11	6	6	1
GH32	Energy	4 (2, 1)	2 (2, 2)	4	4	4	1	0
GH35	PCW-H	6 (3, 0)	3 (3, 1)	2	4	0	4	0
GH39	PCW-H	2 (1, 0)	0	1	1	1	0	0
GH43	PCW-HP	22 (14, 3)	11 (9, 9)	10	15	20	4	0
GH78	PCW-pectin	6 (2, 2)	1 (1, 1)	2	4	3	4	1
GH88	PCW-pectin	2 (1, 0)	0	0	1	1	0	0
GH95	PCW-pectin	2 (2, 0)	0	0	2	1	1	0
PL family	PCW-pectin	9	4	3	10	5	5	0
PL1	PCW-pectin	3 (2, 0)	1 (1, 1)	2	4	2	4	0
PL3	PCW-pectin	3 (1, 0)	0	1	2	1	0	0
CE family	FCW/PCW-H/HP/pectin	35	23	18	53	54	32	10
CE5	PCW-H	11 (7, 1)	4 (3, 4)	6	11	18	8	2
CBM family	FCW/energy/PCW	28	24	21	77	113	65	14
CBM1	PCW	0	1 (1, 1)	0	13	22	19	0

[a] Predicted CAZymes were identified using the carbohydrate-active enzymes database tools (www.cazy.org). GH, glycoside hydrolases; PL, polysaccharide lyases; CE, carbohydrate esterases; CBM, carbohydrate binding modules. Pathogenic lifestyle of the fungi is abbreviates as B (biotroph), HB (hemi-biotroph) and N (necrotroph). Total gene numbers in these families are shown in bold. Families that are discussed in the text and differ greatly in copy number between *Cf* and *Ds* are also shown. Additional data are in Supporting Table S5. [b] Where known, CAZy functions are shown as plant cell wall (PCW) or fungal cell wall (FCW) degrading and modifying enzymes, or energy-related, with substrate preferences for PCWs of cellulose (C), hemicellulose (H), hemicellulose or pectin side-chains (HP) or pectin, using classifications as shown [42].

[c] Fungal species and their pathogenic lifestyles are shown: Cf, *C. fulvum*; Ds, *D. septosporum*; Mg, Mycosphaerella graminicola; Sn, Stagonospora nodorum; Mo, Magnaporthe oryzae; Ss, Sclerotinia sclerotiorum; Bg, Blumeria graminis. Numbers in parenthesis for Cf and Ds are number of genes expressed (first number: in culture; second number: in planta) under conditions described in Tables S12 and S13.

Although *C. fulvum* has a large arsenal of pectinolytic genes compared to *D. septosporum*, not all of them appear to be functional. For example, two of the six GH78 and one of the two GH88 pectinolytic genes are pseudogenized in *C. fulvum*, whilst the corresponding *D. septosporum* families do not contain pseudogenes. Another constraint to function is that gene expression appears to be tightly regulated. As shown in Table 6, none of the 15 *C. fulvum* GH28 genes appear to be expressed *in planta*, whilst all four *D. septosporum* GH28 genes are expressed. Indeed in all gene families with predicted pectinolytic function shown in Table 6 (GH28, GH78, GH88, GH95, PL1, PL3), expression *in planta* was only detected in 2 of the 31 *C. fulvum* genes, whilst all 6 genes in these pectinolytic gene families were expressed in *D. septosporum*. It is possible that *C. fulvum* pectinases are only expressed very locally to modulate complex primary cell wall structures. The location and accessibility of pectin structures embedded in the cell wall is an important consideration for its enzymatic degradation. For instance, the Basidiomycete *Schizophyllum commune* grows predominantly on beech and birch wood which is poor in pectin [82]. However, the pectin in these cell walls is concentrated around the bored pits that are used by *S. commune* to enter the wood, explaining why this fungus contains a higher number of pectinase genes than would be expected based on the overall host pectin content. Differences in pectinolytic gene content and expression between *C. fulvum* and *D. septosporum* may therefore be related to their different strategies of host invasion and subsequent colonization.

In addition to increased numbers of pectinolytic genes compared to *D. septosporum*, *C. fulvum* has more genes for enzymes that degrade hemicelluloses (e.g. families GH31, GH35 and GH39) [83] and hemicellulose-pectin complexes (GH43) (Table 6). It also contains 11 genes (compared to 4 in *D. septosporum*) encoding CE5 enzymes; these include cutinases that are required for early recognition and colonization of the host by fungal pathogens [84,85]. The presence of so many genes encoding enzymes for plant cell wall and cuticle degradation in a biotrophic fungus like *C. fulvum* that enters its host *via* stomata is unexpected. However, the number of cutinase genes, and other secreted lipase genes is particularly low in the *D. septosporum* genome compared to other *Dothideomycetes*, a feature shared with the other tree pathogens *Mycopshaerella populorum* and *M. populicola* (Ohm et al., unpublished data). Overall our comparison shows a similar complement of CAZy genes between *C. fulvum* and *D. septosporum*, but an increased number of particular CAZyme families in *C. fulvum* including genes encoding pectin- and hemicellulose-degrading enzymes. However, a large proportion of genes in the *C. fulvum* CAZyme families lack expression *in planta* and some genes are pseudogenized.

C. fulvum and *D. septosporum* share a broad range of carbohydrate substrates

A second aspect of carbohydrate metabolism that we considered was a comparison of growth on defined and complex carbon substrates (Figs 6 and S4; www.fung-growth.org). It was anticipated that growth profiles could illuminate differences between pathogens with dicot and gymnosperm hosts and show correlations with their respective gene complements. In a study of polysaccharide hydrolysis activities of many fungal pathogens, King et al. [86] showed preferential substrate utilization based on host specificity (dicot or monocot). In general *D. septosporum* grows more slowly on minimal control medium [87] than *C. fulvum*, but surprisingly overall the growth profiles of the two fungi are similar on most substrates (Figs. 6 and S4; Table S6) and both appear to utilize a broader range of substrates than *M. graminicola* (Fig. 6). This is not only the case for the oligomeric and polymeric carbon substrates, requiring CAZymes for degradation, but also for monomeric carbon substrates, suggesting a diverse and efficient carbon catabolism in *C. fulvum* and *D. septosporum.* The good growth of *D. septosporum* on sucrose is particularly striking, suggesting that it can utilize sucrose available in apoplastic fluid during its early biotrophic colonization phase.

In terms of complex carbon sources, *D. septosporum* shows a slightly better capacity than *C. fulvum* to utlise apple and citrus pectin (Figs 6 and S4). This seems to contradict the higher pectinolytic gene numbers in *C. fulvum* compared to *D. septosporum,* but is supported by the expression of fewer *C. fulvum* pectinolytic genes during infection of tomato when compared to the *D. septosporum* pectinolytic genes during infection of pine needle (Table 6). Interestingly, good growth on pectin is also observed for *M. graminicola*, despite an even lower number of putative pectinases than *D. septosporum*. This suggests that regulation of expression is a more dominant factor in pectin degradation by these plant pathogens than the number of pectinase-encoding genes in their genomes. In contrast, pectinase gene numbers correlate well with growth profiles of *Aspergillus ndulans*, *Aspergillus oryzae* and *Aspergillus niger* [88]. Compared to growth on controls lacking a carbon source, *D. septosporum* also showed slightly better growth than *C. fulvum* on lignin. This would be consistent with the higher proportion of lignin in pine needles, estimated to be 25-30% of dry weight [89], compared to less than 10% in dicots [90]. However due to the very slow growth of both fungi and the non-uniform growth habit of *D. septosporum* on these media, firm conclusions about their abilities to utilize lignin cannot be made.

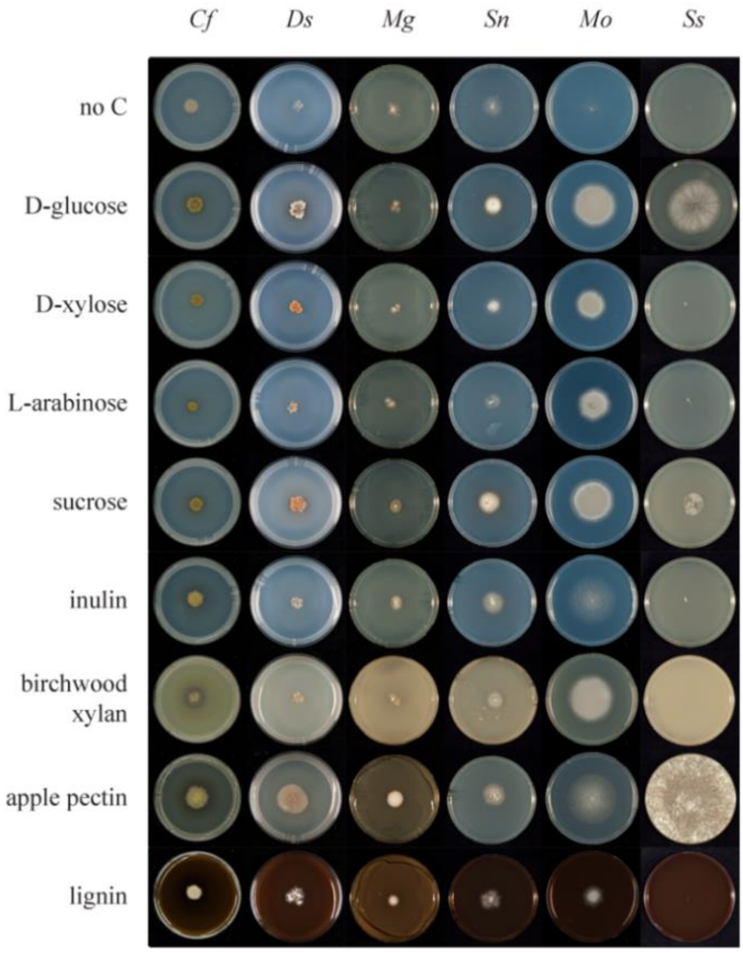


Figure 6. Comparative growth profiling of fungi on various carbohydrate substrates.

Growth on different substrates was compared between 6 fungi on 9 media to highlight differences. *Cf*, *C. fulvum*; *Ds*, *D. septosporum*; *Mg, M. graminicola*; *Sn*, *S. nodorum*; *Mo*, *Magnaporthe oryzae*; *Ss*, *Sclerotinia sclerotiorum*. D-Glucose, D-xylose, L-arabinose and sucrose were added at a final concentration of 25 mM. Birchwood xylan, apple pectin and lignin were added at a final concentration of 1% (w/v).

Adaptations for coping with chemical and structural defences

Tomato plants produce the antimicrobial saponin, tomatine. The tomato pathogen *Fusarium oxysporum* produces α-tomatinase, which functions as a virulence factor as it degrades tomatine into the non-toxic compounds tomatidine and lycotetraose [91]. A gene predicted to encode α-tomatinase, classified as a GH10 enzyme, was found in the *C. fulvum* genome (JGI ID 188986) but is absent from the *D. septosporum* genome. Another gene found only in *C. fulvum* shows predicted similarity to the GH5 family enzyme hesperidin 6-O-α-L-rhamnosyl-β-glucosidase that can degrade hesperidin [77]. Hesperidin occurs most abundantly in citrus fruits [92] and is a member of the flavonoid group of compounds that is well known for its antimicrobial activity. Flavonoid-degrading enzymes such as hesperidin 6-O-α-L-rhamnosyl-β-glucosidase might enable *C. fulvum* to detoxify hesperidin or related compounds present in tomato.

Chemical defence molecules in pine needles include antimicrobial monoterpenes. Thus it is expected that *D. septosporum* is adapted to tolerate or degrade these compounds whilst *C. fulvum* is not. Recent work on the pine pathogen *Grosmannia clavigera* revealed several classes of genes that are upregulated in response to terpene treatment [93]. After 36 hrs, major classes of upregulated genes included those involved in beta-oxidation as well as mono-oxygenases and alcohol/aldehyde dehydrogenases that may be involved in activating terpenes for beta-oxidation. A drug transporter, GLEAN_8030, was functionally analyzed and found to be required for tolerance of the fungus against terpenes, enabling *G. clavigera* to grow on media containing these compounds. A search for three of these genes showed that both *C. fulvum* and *D. septosporum* genomes share a similar gene complement to each other, including a GLEAN_8030 homolog (Table S7). However, since these genes have not all been functionally characterized in *G. clavigera* and all are predicted to encode proteins involved in general metabolic processes, further work is required to determine the roles of the homologs found in both *C. fulvum* and *D. septosporum.*

As well as chemical mechanisms, plants employ basal structural defence mechanims including lignification of cell walls [94,95]. Due to the abundance of lignin in pine needles that block access to usable cellulose, fungal pathogens and saprophytes living on pines have a particularly challenging environment [96]. For *D. septosporum* to complete its lifecycle, degradation of pine needle tissue must occur so that conidiophores bearing conidia can erupt through the epidermis (Fig. 1H), which contains lignin [97]. This is in contrast to *C. fulvum* whose conidiophores emerge from tomato leaves through stomatal pores (Fig. 1D). Thus, we investigated genes that may be involved in lignin degradation.

Some saprophytic fungi utilize oxidoreductases, particularly class-II peroxidases such as lignin peroxidases, manganese peroxidases and laccases, and a number of H_2O_2-producing enzymes to achieve lignin breakdown [98,99]. However, the number of genes encoding oxidoreductases in *D. septosporum* is no higher than those of other *Dothideomycetes* (*C. fulvum, M. graminicola* and *S. nodorum*) that infect plants with lower levels of lignin (Table S8). *D. septosporum* appears to have a similar complement of laccase genes as *C. fulvum* and only one distant relative of a class-II peroxidase, missing in *C. fulvum*, but also present in *M. graminicola* and *S. nodorum*. Interestingly, the classical Ascomycete laccases found in *C. fulvum, D. septosporum* and *M. graminicola* bear a carbohydrate-binding domain (CBM20, putative starch binding domain). This type of laccase is only found in Dothideomycetes but the significance of this novel modular structure is unclear. Brown-rot saprophytes such as *Serpula lacrymans* have a reduced complement of ligninolytic genes compared to lignin-degrading white-rot fungi and are proposed to initially weaken lignocellulose complexes by non-enzymatic use of hydroxyl radicals [81], prior to enzymatic assimilation of accessible carbohydrates [81]. It is likely that *D. septosporum* uses a similar strategy to breach the lignin-rich components of pine needles, as complete degradation of this polymer is not required to complete its life cycle.

The secondary metabolite gene complement of *C. fulvum* and *D. septosporum*

Secondary metabolites (SMs) are important compounds for the colonization of specific ecological niches by fungi. In particular, plant-pathogenic fungi can produce non-specific and host-specific toxic SMs [100]. SMs also include mycotoxins that contaminate food and feed and are harmful to mammals [100]. The only known SMs produced by *C. fulvum* and *D. septosporum* are cladofulvin and dothistromin, respectively [101,102]; both compounds are anthraquinone pigments. In fungi, SM biosynthetic pathways often involve enzymes encoded in gene clusters [103] and always require the activity of at least one of four key enzymes: polyketide synthase (PKS), non-ribosomal peptide synthetase (NRPS), terpene cyclase (TC) or dimethylallyl tryptophan synthase (DMATS) [104]. It has been suggested that loss of SM biosynthetic pathways is associated with biotrophy [43]. We searched for SM gene pathways in both genomes. Surprisingly, the biotroph *C. fulvum* has twice the number of key genes (23 in total) compared to the hemibiotroph *D. septosporum* (11 in total) (Table 7), of which 14 and 9, respectively, are organized into gene clusters along with other SM-related genes. The numbers of key SM enzyme-encoding genes are comparable to those of *M. graminicola*, but are lower than those in most other sequenced *Dothideomycetes* (Ohm et al., unpublished data). Like all *Ascomycetes* [105], the majority of key SM enzymes in *C. fulvum* and *D. septosporum* are PKSs, NRPSs and hybrid PKS-NRPSs. Annotation of all key SM genes was manually checked and two truncated (*Pks4* and *Nps1*) and five pseudogenized (*Pks9*, *Hps2*, *Nps5*, *Nps7* and *Nps10*) genes were found in the *C. fulvum* genome, while all *D. septosporum* genes except *Pks4* (truncated) are predicted to encode functional enzymes. Overall, the number of predicted functional pathways suggests that *C. fulvum* and *D. septosporum* can produce at least 14 and 10 different SMs, respectively.

Table 7. Key secondary metabolism genes in Ascomycete genomes.

Fungal species	Lifestyle	PKS	NRPS	Hybrid	TC	DMATS	Total
Cladosporium fulvum	Biotroph	10	10	2	0	1	23
Dothistroma septosporum	Hemibiotroph	5	3	2	0	1	11
Mycosphaerella graminicola	Hemibiotroph	11	6	2	1	0	20
Stagonospora nodorum	Necrotroph	22	10	2	2	2	38
Magnaporthe oryzae	Hemibiotroph	22	8	10	3	3	46
Fusarium graminearum	Necrotroph	13	12	2	3	0	30
Aspergillus nidulans	Saprophyte	26	10	2	5	2	45
Neurospora crassa	Saprophyte	6	3	1	1	1	12

Numbers of predicted polyketide synthase (PKS), non-ribosomal peptide synthetase (NRPS), hybrid PKS-NRPS (Hybrid), terpene cyclase (TC) and dimethylallyl tryptophan synthase (DMATS) genes are shown. Data are from this study and from Collemare et al. 2008 [105].

Surprisingly, only three of the key SM genes are predicted to belong to biosynthetic pathways shared between the two species (Table S9) suggesting a diverse SM repertoire. This is much lower than expected given the overall level of similarity in gene content between the two genomes, and suggests that this SM repertoire is under strong selection. The three common genes are predicted to be involved in production of a pigment related to melanin (*Pks1*), a siderophore (*Nps2*) and dothistromin (*PksA*) based on similarities to other characterized genes. In addition in *C. fulvum*, the three other functional non-reducing PKS enzymes are candidates for production of cladofulvin.

The genomic locations of the 11 biosynthetic SM genes in *D. septosporum* do not show any enrichment at sub-telomeric positions, as reported for *Aspergillus* spp. and *Fusarium graminearum* [106,107], or near putative centromeres (Fig. 3). However, 8 out of the 11 genes are located on chromosomes smaller than 2 Mb (chromosomes 8 to 12; Fig. 3). The genomic regions immediately surrounding all 11 *D. septosporum* SM genes are conserved in the *C. fulvum* genome, although 8 of them lack the key SM gene itself and sometimes putative accessory genes also (Fig. S5). Reciprocally, only 9 *C. fulvum* SM genomic regions out of 23 are conserved in *D. septosporum* with 6 of these lacking the key SM gene, suggesting either gain or loss of SM genes has occurred. For two of the regions where flanking genes are conserved but SM gene(s) are missing in *C. fulvum* (regions corresponding to those surrounding *Pks3* and *Nps3* in *D. septosporum*), the presence of repeats suggests that SM gene loss may have occurred in *C. fulvum* (Fig. S5). The *C. fulvum*-specific SM loci *Pks5*, *Pks6*, *Nps5*/*Dma1* and *Nps9* include many transposable elements and genes that have similarity to genes scattered in the *D. septosporum* genome, often on the same chromosome, leading to the hypothesis that these SM loci were assembled by gene relocation as recently proposed for the fumonisin gene cluster in *F. verticillioides* [108].

Dothistromin toxin genes are present in both genomes

Analysis of the *D. septosporum* 1.3 Mb chromosome 12 revealed that the three previously identified mini-clusters of dothistromin genes [38] are widely dispersed, confirming fragmentation of this gene cluster (Figs. 3 and 7). Candidates for additional dothistromin genes, previously predicted based on aflatoxin pathway genes [39], are also present. Although three of these genes (*OrdB, AvnA, HexB*) are located in the published *VbsA* mini-cluster, the others are dispersed over different regions of chromosome 12 as shown in Fig. 7. The end of the *Nor1* gene cluster (*Nor1, AdhA, VerB*) is less than 10 kb from one predicted telomere, whilst *Ver1* (previously called *dotA* [35]) is only 81 kb from the other. As expected, a gene similar to the aflatoxin *AflR* regulatory gene is present and, like in aflatoxin-producing species of *Aspergillus*, is divergently transcribed with an adjacent *AflJ* regulatory gene candidate. Functional analysis of these genes is in progress.

Although *C. fulvum* is not known to produce dothistromin, the complete set of predicted dothistromin genes is present in its genome, encoding proteins with amino acid identities ranging from 49% (AflJ) to 98% (Ver1) when compared with those of *D. septosporum* (Table S10). The arrangement of predicted dothistromin genes in *C. fulvum* reveals a high level of synteny with some rearrangements. With the exception of the *Ver1* gene cluster, the mini-clusters contain the same genes in the same orientations in the two species (Fig. 7A). The three mini-clusters on *C. fulvum* scaffold 130775 are much closer together than in *D. septosporum,* but are still separated from each other by considerable distances (approximately 24 kb between *Est1* and the *VbsA* gene cluster, and 40 kb between the *VbsA* and *Nor1* gene clusters). A comparison of the relative locations of the mini-clusters in the two species suggests inversions (*AflR/J* and *VbsA* gene clusters) as well as rearrangements over relatively small (*VbsA*-*Nor1*) and large (*Ver1*-*AflR/J*) distances. This is consistent with the overall pattern of intrachromosomal rearrangements observed between these two genomes.

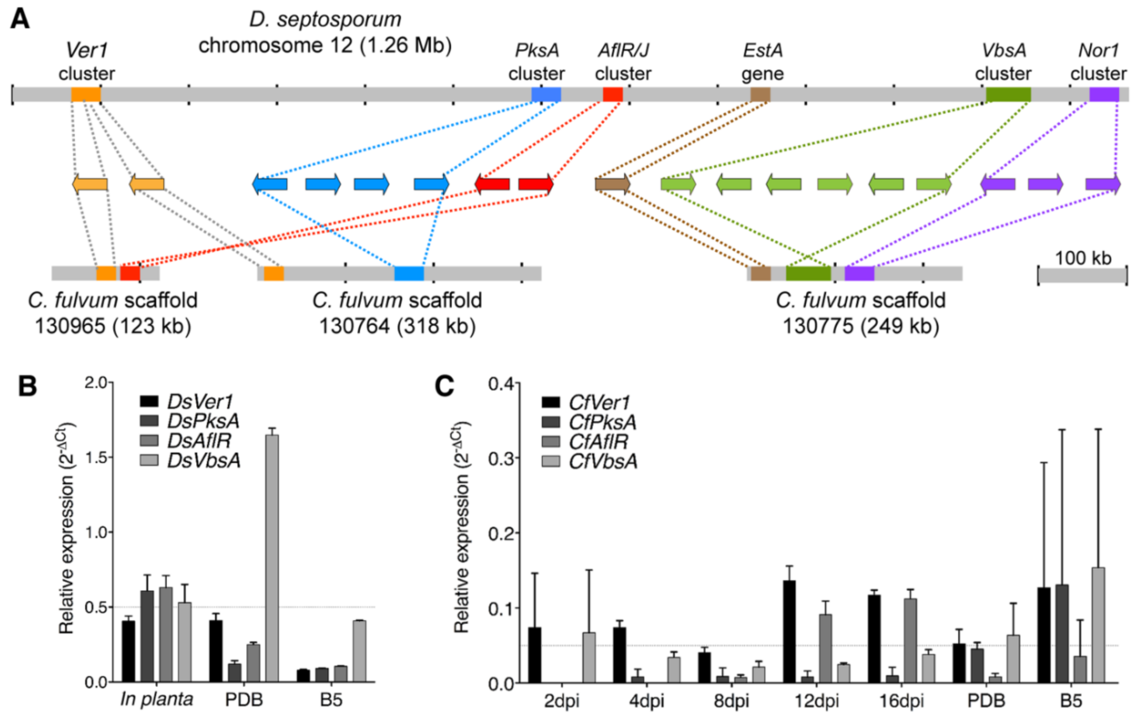


Figure 7. Arrangement of predicted dothistromin genes in *Dothistroma septosporum* and *Cladosporium fulvum*.

A) Predicted dothistromin genes within the labeled clusters (left to right) are: *Ver1, DotC* (*Ver1* cluster); *PksA, CypA, AvfA, MoxA* (*PksA* cluster); *AflR, AflJ* (*AflR/J* cluster); *OrdB, AvnA, HexB, HexA, HypC, VbsA* (*VbsA* cluster); *Nor1, AdhA, VerB* (*Nor1* cluster). Positions of mini-clusters are approximate and they are not drawn to scale. Dothistromin genes within the published *D. septosporum PksA* and *VbsA* clusters [36,38] and the newly discovered *AflR/J* and *Nor-1* clusters are found in the same order and orientation in *C. fulvum*. **B)** Expression of dothistromin biosynthetic genes (*Ver1, PksA, VbsA*) and regulatory gene (*AflR*) was determined in *D. septosporum* by quantitative PCR. Mean expression and standard deviations are shown for at least 3 biological replicates relative to beta-tubulin expression. In *D. septosporum* all genes but *DsVbsA* are expressed more highly *in planta* (late stage sporulating lesions from a forest sample) than in culture (PDB or B5 media) as highlighted by the dashed-grey line. **C)** Expression of *C. fulvum* genes is shown as for (B), revealing that expression is not higher during tomato infection than in culture (dashed-grey line). Note the different scales for expression, which reveal a much lower level of transcription both *in planta* and in PDB medium compared to *D. septosporum*.

Given the presence of the dothistromin biosynthetic pathway genes, we tested whether dothistromin is produced by *C. fulvum*. However, no dothistromin was detected by HPLC analysis of extracts from *C. fulvum* PDB cultures, which is a condition favorable to dothistromin production by *D. septosporum*. Despite the lack of dothistromin production under these conditions, a strong evolutionary constraint on dothistromin biosynthetic genes was seen by analyzing the ratio of non-synonymous to synonymous mutations (Ka/Ks) between *C. fulvum* and *D. septosporum*. The low Ka/Ks ratios seen for dothistromin genes (range 0.018-0.169) are indicative of purifying selection [109] and did not differ from the distribution observed for four housekeeping genes (*Tub1, Eif3b, Pap1, Rps9*; range 0.003-0.073) (P = 0.561). Evidence for purifying selection was also shown for aflatoxin pathway genes in *Aspergillus flavus* and *A. nomius* [110]. On the basis of this we propose that *C. fulvum* might produce dothistromin, or a metabolite related to dothistromin, under certain environmental conditions when it is required.

Regulation of secondary metabolite biosynthetic pathways suggests lifestyle adaptation at the transcriptome level

Many fungal SM biosynthetic pathways are cryptic, meaning that they are not expressed in wild-type strains under laboratory conditions. However, manipulation of genetic regulatory pathways or environmental conditions has shown that some of these cryptic pathways are functional [111,112]. As seen for other gene families such as CAZyme genes, *C. fulvum* appears to be more economical in its gene expression than *D. septosporum*, particularly *in planta*. In *C. fulvum*, EST support was obtained from *in vitro* conditions for all key SM genes except *Hps2*, *Nps7* and *Nps10,* which are pseudogenized. The two truncated genes (*Pks4* and *Nps1*) and the pseudogenized *Nps5* genes also have EST support but the resulting proteins are unlikely to be functional. However, no evidence for *in planta* expression could be obtained for any of the *C. fulvum* key SM genes from this EST library. In contrast, all *D. septosporum* key SM genes have EST support from both *in vitro* and *in planta* libraries, with the unique *DsPKS2* being one of the most highly expressed genes during pine needle infection.

Differences in dothistromin pathway regulation were confirmed by quantitative PCR. In *D. septosporum*, *Ver1*, *PksA*, *AflR* and *VbsA* show higher expression during pine infection than in controlled culture conditions used to induce dothistromin production (Fig. 7B). In contrast, the same genes show a low expression level in *C. fulvum* during infection and *in vitro* (Fig. 7C). Because no dothistromin could be detected in liquid culture, this low expression likely represents background transcription with no biological relevance. Such an expression pattern is significantly different from the up-regulation of *Avr4* and *Avr9* genes during tomato infection (Fig. S6).

SM production is associated with development in fungi and involves common regulators [113]. We searched the genomes of *C. fulvum* and *D. septosporum* for conserved regulators of development and SM production and, based on predicted protein sequences, found clear homologs for most of these genes in both fungi (Table S11). The two species appear to lack a PpoB oxygenase, but *PpoA* and *PpoC* are sufficient to produce all *psi* factors (oxylipins) identified in *Aspergillus* species [114]. In addition, *C. fulvum* lacks clear homologs of the G-protein regulators FlbA and RgsA, while possible homologs are found in *D. septosporum*. In *Aspergillus* species both proteins are negative regulators of G-protein signaling pathways. Neither *C. fulvum* nor *D. septosporum* have a homolog of *BrlA*, an essential regulator of conidiation in *Aspergillus* species [115], suggesting that they use another regulator for this role. Future studies analyzing expression of the SM genes, and the roles of regulatory genes, will help to determine fundamental differences in how *C. fulvum* and *D. septosporum* differentially regulate their SM gene expression.

Conclusion

We embarked upon a comparative genomics analysis of *C. fulvum* and *D. septosporum* to test for differences that might explain their host specificity and lifestyles. The comparison revealed surprising similarities, such as the presence of dothistromin toxin genes in *C. fulvum* and functional *Avr4* and *Ecp2* effector genes in *D. septosporum*. However, the genome sizes of the two fungi are remarkably different, mainly due to a vast expansion of transposable elements in *C. fulvum*, and show several key differences in gene content. Adaptation of *C. fulvum* to its host plant tomato is exemplified by the specific presence of a gene encoding α-tomatinase, likely involved in degradation of tomatine. In contrast, the dothistromin gene cluster is present in both fungi, but while it is strongly expressed in *D. septosporum* at later stages of pine needle infection, it is lowly or not expressed in *C. fulvum* during infection of tomato leaves. Both fungi contain additional key SM genes, but the majority of these are not in common, contrasting with the high degree of homology between the two genomes. We suggest that this lack of conservation of key SM genes in the *C. fulvum* and *D. septosporum* genomes is a consequence of different evolutionary pressures that result from their different lifestyles, either as a pathogen inside their host or possibly as a saprophyte outside their host.

Another key difference between the two fungi during pathogenesis concerns their differential gene regulation. Gene expression in *C. fulvum* is strictly regulated *in planta,* with many SM, hydrophobin and CAZy genes not expressed, while expression in *D. septosporum* is more constitutive. This differential regulation of expression may be crucial in determining differentiation between these fungi despite very similar gene profiles. Furthermore, this expression pattern is consistent with a biotrophic lifestyle without gene loss. Finally we suggest that the higher repeat content of the *C. fulvum* genome, along with evidence for gene pseudogenization (van der Burgt et al., unpublished data) has facilitated the evolution of different lifestyles between *C. fulvum* and its sister species *D. septosporum*. Overall, our comparison of the two genomes suggests that even closely related plant pathogens can adapt to very different hosts and lifestyles by differentiating gene content and regulation, whilst retaining genetic signatures of a common ancestral way of life.

Materials and Methods

Fungal strains and growth conditions

The fungal strains of *C. fulvum* (race 0WU; CBS131901) and *D. septosporum* (strain NZE10; CBS128990) were isolated from tomato growing in an allotment garden in Wageningen, The Netherlands, in 1985, and from a needle from an eight-year-old *Pinus radiata* tree on the West Coast of the South Island of New Zealand in 2005, respectively. Monospore cultures, whose identities were confirmed by ribosomal ITS sequencing, were used throughout. Unless specified otherwise, cultures of these fungi were maintained on potato dextrose agar (PDA) or potato dextrose broth (PDB) media (*C. fulvum*) or Dothistroma Medium (DM; 5% w/v malt extract, 2.8% w/v nutrient agar or nutrient broth) at 22°C prior to use. Cultures were maintained for long-term storage in closed vials as −80°C stocks in 20% glycerol.

Tomato Infections

Conidia of *C. fulvum* were harvested from two-week old PDA plates with distilled water. The conidial suspension was filtered through Calbiochem(R) Miracloth (EMD Millipore Chemicals, Philadelphia, PA) and washed once with water prior to calibration to 5x10^5 conidia/mL. Five-week-old tomato Heinz plants were sprayed on the lower side of the leaves with the conidial suspension (10 mL per plant). The plants were kept at 100% relative humidity for 48 h. The plastic-covered cages were then opened to grow the plants under regular greenhouse conditions (70% relative humidity, 23-25°C during daytime and 19-21°C at night, light/dark regime of 16/8 h, and 100 W/m^2 supplemental light when the sunlight influx intensity was less than 150 W/m^2). The 4th composite leaves of infected tomato plants were harvested at 2, 4, 8, 12 and 16 dpi, and immediately frozen in liquid nitrogen.

Phylogenetic comparison of fungal species

To highlight the phylogenetic relationships of *C. fulvum* and *D. septosporum* with *Dothideomycetes* and other fungi relevant to this study conserved protein families were predicted by use of the MCL Markov clustering program [116] with pairwise blastp protein similarities and an inflation factor of 4. From this multi-gene family set, 51 orthologous groups of genes were identified. Predicted protein sequences were concatenated, aligned using MAFFT 6.717b [117] and a species tree calculated using RAxML 7.2.8 [118]. We also determined protein homology data based on bidirectional best hits when comparing the proteomes of eleven *Dothideomycete* species (*Alternaria brassicicola*, *C. fulvum*, *Cochliobolus heterostrophus*, *D. septosporum*, *Hysterium pulicare*, *Mycosphaerella fijiensis*, *Mycosphaerella graminicola*, *Pyrenophora tritici-repentis*, *Rhytidhysteron rufulum*, *Septoria musiva* and *Stagonospora nodorum*), together with four out-group species (*Aspergillus nidulans*, *Fusarium graminearum*, *Neurospora crassa* and *Magnaporthe grisea*).

Repetitive sequences and transposable elements

Repeat sequences in both genomes were identified using RECON [119]. To group repetitive elements together into different families the default RECON output was parsed to include families with 10 or more elements. The parsed RECON repeat library was used to determine the extent of the repetitive fraction in the *D. septosporum* and *C. fulvum* genomes using RepeatMasker [120] and to annotate repetitive families and identify structural features, such as Long Terminal Repeats (LTRs) and Terminal Inverted Repeats (TIRs), using BLAST.

Repeat-induced point mutation (RIP)

Sequences that had undergone Repeat-Induced Point mutation (RIP) were identified according to the composite RIP index (CRI) method [121]. The CRI was calculated for each 500 nt sequence window which was shifted by a 25 nt step. Sequences were identified as having been subjected to RIP when the RIP product, RIP substrate and composite RIP indices were at least 1.2, at most 0.8 and at least 1.0 respectively. As a final constraint, series of overlapping sequence windows had to exceed 750 nt in length and the CRI value of any of the windows peaked to 1.5 in order to be scored as a RIP'd locus.

Syntenic and non-syntenic regions.

Syntenic regions shared between *C. fulvum* and *D. septosporum* were detected *ab initio* on their repeat masked genome sequences using promer [122], blastp and a suite of custom made python scripts. A script called blastpmer obtained all translated ORFs above a threshold nucleotide length from both query and subject genomes, performed a blastp on these ORFs, and subsequently filtered on expect value and high-scoring segment pair (HSP) length. Protein matches (using *C. fulvum* as query and *D. septosporum* as subject) were obtained with promer (--maxmatch) and blastpmer (–ORF 500 nt –HSP 250 nt –expect 1e-9). Both genomes were masked for these protein matches before being subjected to a second round of searching for weaker and shorter protein similarities, again using promer (–b 50 –c 15 –l 5 --maxmatch) and blastpmer (–ORF 300nt –HSP 110 nt –e 1e-7). These four searches yielded 57,270, 44,865, 1,864 and 2,367 matches, respectively, many of which were redundant and overlapping. This large set was reduced to 24,480 unique matches by removing all except the best alignment for each unique genomic locus. This step removed overlapping alignments with different phases or orientations, and excluded suboptimal alignments caused by paralogs and common protein domains. The product of amino acid similarity and match length was employed as a final alignment quality score. Matches were ordered by query scaffold position and joined into linked syntenic regions according to the following criteria: (i) adjacent matches were identical on the query and subject scaffolds, (ii) matches had the same strand orientation, and (iii) maximum and average nucleotide distance between adjacent matches on the query and subject scaffolds were <10 kb and <5 kb, respectively. This step resulted in a reduction to 1,875 collinear match regions, of which 1,277 were >5 kb. For comparison of protein coding genes in syntenic versus non-syntenic areas, gene models were classified as syntenic if they overlapped with any of the 1,875 collinear syntenic areas. Thus, subsets of 9,890 syntenic and 4,237 non-syntenic genes were inferred for *C. fulvum*.

In order to investigate mesosynteny on a whole genome scale a refined synteny dataset was created with correction for inversions and rearrangements, and removal of spurious, small alignments. Match regions were compared and merged further if (i) adjacent groups had opposite orientations, or (ii) groups with identical query and subject scaffolds were separated by at least one (group of) matches on a conflicting subject scaffold, but maximum and average nucleotide distances between match regions were at most 20 kb or on average <10 kb apart, and finally (iii) match regions <5 kb were rejected. The final refined dataset contained 1,103 syntenic regions between 5 and 226 kb (average 22,194 bp), representing 22,700 matches from the original 24,480 unique matches.

C. fulvum and *D. septosporum*-specific proteins

To identify potential *C. fulvum* and *D. septosporum*-specific proteins, the total protein set from both fungi was used in comparative blastMatrix [123] searches against sequences from the nine additional members of the *Dothideomycetes* listed in the phylogenetics section.

C. fulvum and *D. septosporum* secretome analysis

Initially, subcellular localizations for all *C. fulvum* and *D. septosporum* proteins were predicted using WoLF PSORT (wolfpsort.org; [124]), resulting in identification of 1886 putative extracellular *C. fulvum* proteins and 1591 putative extracellular *D. septosporum* proteins. Only proteins containing a signal peptide and a signal peptide cleavage site, but lacking transmembrane (TM) domains or proteins containing a single TM that overlaps with the secretion signal, were selected. Signal peptides and cleavage sites were predicted using SignalP version 3.0 [125], where a final D-Score cut-off of 0.5 was used to increase specificity while retaining sensitivity. Subsequently, all proteins with signal peptides (1886 and 1591 for *C. fulvum* and *D. septosporum,* respectively) were analyzed for the presence of TM domains using the web servers Phobius [126] and TMHMM (version 2.0; [127]). The servers identified different, partially overlapping, sets of proteins with putative TM domains. On average Phobius detected 22% more TM domain proteins than did TMHMM, and about 75% of the predictions were shared between the servers. For further analyses, all proteins with putative TM domains as predicted by either of the two servers were removed from the dataset. Then, the proteins that contain a putative mitochondrial targeting signal as predicted by TargetP version 1.1 [128] were removed. Finally, proteins containing a potential GPI-anchor signal as predicted by the PredGPI web-service were discarded [129].

Functional analysis of *D. septosporum* Avr4 by *A. tumefaciens*-mediated transient gene expression in *N. benthamiana*

A *C. fulvum Avr4* (*Cf-Avr4*) gene homolog was identified in the genome of *D. septosporum* (*Ds-Avr4*) by blastp, with an E-value of 1x10^{-4}*.* To determine Cf-4-mediated HR-inducing ability of Ds-Avr4 of *D. septosporum* the *Agrobacterium tumefaciens*-mediated transient gene expression (ATTA) method was performed in *N. benthamiana* as described by Van der Hoorn et al. [130]. The *Cf-Avr4* and *Ds-Avr4* genes were each fused to a PR-1A signal peptide sequence [131] for secretion into the apoplast. Subsequently a Gateway cloning strategy was performed to clone them into a pK2GW7 binary expression vector [132] containing the CaMV 35S promoter. *A. tumefaciens* (strain GV3101) was finally transformed with pK2GW7 binary vectors containing *Cf-Avr4* or *Ds-Avr4* genes by electroporation. Agroinfiltration of *Cf-4* transgenic *N. benthamiana* leaves with *Cf-Avr4-* and *Ds-Avr4*-containing *A. tumefaciens* clones was performed as described by van der Hoorn et al. [130]. Photographs were taken at six days post inoculation.

Heterologous expression of the *D. septosporum* Ecp2 (Ds-Ecp2-1) gene in MM-Cf-Ecp2 tomato plants

Three *D. septosporum* homologs of *C. fulvum Ecp2* genes (*Ds-Ecp2-1, Ds-ecp2-2* and *Ds-Ecp2-3*) were identified as described for *Avr4*. A binary Potato Virus X (PVX)–based vector, pSfinx, was used for transient expression of the Cf-Ecp2-1 ortholog, *Ds-Ecp2-1,* in MM-Cf-Ecp2 tomato lines based on methodology described by Hammond-Kosack et al. [131]. The recombinant viruses were obtained by cloning *Ds-Ecp2-1* (an intron-less gene), encoding the mature protein, downstream of the PR-1A signal sequence for secretion into the apoplast and under the control of the CaMV 35S promoter. Recombinant *pSfinx::Ecp2-1* and *pSfinx::Empty* viruses corresponding to the *C. fulvum Ecp2* (*Cf-Ecp2-1*) were published before [33]. *A. tumefaciens* (GV3101) was transformed with *pSfinx::Ds-Ecp2-1* construct by electroporation. *A. tumefaciens* strains containing the *pSfinx* constructs for the expression of Cf-Ecp2-1 and Ds-Ecp2-1 proteins were inoculated on MM-Cf-*Ecp2* tomato lines containing the cognate *R* gene and MM-Cf-0 tomato lines that contain no *R* genes mediating recognition of the Ecp2-1 effector. Photographs were taken four weeks post inoculation.

Analyses of hydrophobin-encoding genes

All six previously reported hydrophobin genes from *C. fulvum* [133] were found in the automated gene predictions performed on the genome sequence. Five of the hydrophobins (Hcf-1 to Hcf-5) are predicted to contain an interpro motif common in fungal hydrophobins (IPR001338), while Hcf-6 has an interpro motif, which is restricted to *Ascomycetes* only (IPR010636). To identify putative hydrophobin-encoding genes in other genomes, all secreted gene models of *C. fulvum, D. septosporum* and *M. graminicola* were computationally annotated using Interpro scan and Gene Ontology terms. Then, gene models with IPR001338 and IPR010636 Interpro scan terms were identified as putative hydrophobin candidates. Also, a HMM profile search (which was built based on the conserved cysteine motifs in class I hydrophobins) was performed to identify missed hydrophobins by standard similarity searches. In this way five additional hydrophobin genes were identified in the *C. fulvum* genome. Hydrophobin sequences were aligned with ClustalW and edited in GeneDoc software. Then a consensus phylogenetic tree of predicted hydrophobin amino acid sequences was constructed using MEGA5 software [134] performing the minimum-evolution algorithm with default parameters and 1000 bootstrap replications.

Analyses of carbohydrate-active (CAZy) enzymes

The carbohydrate-active enzyme catalogs of *C. fulvum* and *D. septosporum* were compared with the corresponding catalogs from other *Dothideomycete* fungi (Ohm et al., unpublished data). The boundaries of the carbohydrate-active modules and associated carbohydrate-binding modules of the proteins encoded by each fungus in the comparison were determined using the BLAST and HMM-based routines of the Carbohydrate-Active-EnZymes database ([74]; www.cazy.org). For determining the growth profiles on different carbohydrate substrates Aspergillus minimal medium [87] adjusted to pH 6.0 and containing 1.5% agar (Invitrogen, 30391-049) was used. Carbon sources were added at concentrations as indicated in the text and using standard methods as described at www.fung-growth.org. Duplicate plates were inoculated with 2 µl of a suspension containing 500 conidia/µl. Cultures were grown at 22-25°C for two weeks for *C. fulvum* and four weeks for *D. septosporum*, and representative plates were photographed.

Secondary metabolite gene analysis

Genes encoding polyketide synthases (PKSs), non-ribosomal peptide synthases (NRPSs), hybrids of PKS and NRPS, terpene cyclases (TCs) and dimethylallyl tryptophan synthases (DMATSs) were sought in the two genomes using tblastn/blastp and several Ascomycete protein sequences as queries (Ace1 for PKS and hybrids; MGG_00022.7 protein for NRPS; tri5, cps/ks, all TCs from *B. cinerea* for TCs; Dma1 from *Claviceps purpurea* for DMATSs). For each tblastn/blastp hit, functional annotation was confirmed by searching for conserved domains (CDS at NCBI, InterproScan) and performing blastp analysis at NCBI and InterproScan. The locus of each key gene was analyzed for genes that could potentially be involved in a biosynthetic pathway. Functional annotation of downstream and upstream genes was confirmed using blastp at NCBI. In addition, homologs to genes that were shown to be involved in the regulation of fungal development and secondary metabolism were sought using tblastn/blastp with the sequences of the characterized proteins as queries.

Ka/Ks calculations were carried out to estimate evolutionary constraints on putative dothistromin genes (*PksA, VbsA, Ver1, HexA, AvfA, CypA* and *MoxA*) in comparison to four housekeeping genes (*Tub1* JGI PIDs Cf-186859 Ds-68998*, Eif3b* Cf-190521 Ds-75033*, Pap1* Cf-190301 Ds-180959 and *Rps9* Cf-196996 Ds-92035). DNA sequences from *D. septosporum* and *C. fulvum* were aligned with the codon-aware multiple sequence alignment software, RevTrans [135]. Sequence alignments were trimmed in codon units to remove missing data across both species with the sequence editor, Jalview [136]. The non-synonymous/synonymous amino acid ratio (Ka/Ks or ω) was obtained using the Ka/Ks Calculator [137] with the algorithm of Nei and Gojobori [138]. Statistical differences between Ka/Ks values for dothistromin and housekeeping genes were determined using Student's two-sided t test [139]. For determination of dothistromin production, previously published extraction and hplc methods were followed [140].

Quantitative PCR

For quantification of dothistromin gene expression in *D. septosporum*, RNA was extracted from sporulating lesions on *Pinus radiata* needles collected from a forest in New Zealand (*in planta* sample) or grown in PDB or B5 [141] broths for 6 days as described previously [140]. cDNA synthesis and relative quantitative RT-PCR were carried out using primers and methods described earlier [140], with three biological replicates and two technical replicates. For *C. fulvum*, similar protocols were followed except that tomato infections, RNA extraction and cDNA synthesis followed the protocols of [142] and four biological replicates were used. Oligonucleotides were designed with Primer3Plus [143] and are shown in Table S14. Their efficiency and specificity were tested on a genomic DNA dilution series. For both species, quantitative PCR was performed with the Applied Biosystems 7300 Real-Time PCR system (Applied Biosystems, USA) using the default parameters. Raw data were analyzed using the $2^{-\Delta Ct}$ method [144].

Acknowledgements

Mattias de Hollander and Konstantinos Billis (Wageningen University, Netherlands) are thanked for initial analysis on the *C. fulvum* genome during their MSc thesis projects at Wageningen University. Ben Doherty (Ministry of Agriculture and Forestry, New Zealand) is thanked for collection of *D. septosporum* isolate NZE10 and the following are thanked for provision of photographs for Fig. 1: E, Harry Kope (BC Ministry of Natural Resource Operations, Canada); F, Margaret Dick (Scion, New Zealand); G.H, Doug Hopcroft (Massey University, New Zealand).

References.

1. van Kan JAL, van den Ackerveken GFJM, de Wit PJGM (1991) Cloning and characterization of cDNA of avirulence gene *Avr9* of the fungal pathogen *Cladosporium fulvum,* causal agent of tomato leaf mold. Mol Plant Microbe Interact 4: 52-59.
2. Thomma BPHJ, Van Esse HP, Crous PW, De Wit PJGM (2005) *Cladosporium fulvum* (syn. *Passalora fulva*), a highly specialized plant pathogen as a model for functional studies on plant pathogenic Mycosphaerellaceae. Mol Plant Pathol 6: 379-393.
3. Goodwin SB, Dunkle LD, Zismann VL (2001) Phylogenetic analysis of *Cercospora* and *Mycosphaerella* based on the internal transcribed spacer region of ribosomal DNA. Phytopathology 91: 648-658.
4. Bradshaw RE, Zhang SG (2006) Biosynthesis of dothistromin. Mycopathologia 162: 201-213.
5. Barnes I, Crous PW, Wingfield BD, Wingfield MJ (2004) Multigene phylogenies reveal that red band needle blight of *Pinus* is caused by two distinct species of *Dothistroma, D. septosporum* and *D. pini*. Stud Mycol 50: 551-565.
6. Jenkins JA (1948) The origin of the cultivated tomato. Economic Botany 2: 379-392.
7. Cooke MC (1883) New American fungi. Grevillea XII: 32.
8. De Wit PJGM (1992) Molecular characterization of gene-for-gene systems in plant-fungus interactions and the application of avirulence genes in control of plant-pathogens. Annu Rev Phytopathol 30: 391-418.
9. Enya J, Ikeda K, Takeuchi T, Horikoshi N, Higashi T, et al. (2009) The first occurrence of leaf mold of tomato caused by races 4.9 and 4.9.11 of *Passalora fulva* (syn. *Fulvia fulva*) in Japan. J Gen Plant Pathol 75: 76-79.
10. Iida Y, Iwadate Y, Kubota M, Terami F (2010) Occurrence of a new race 2.9 of leaf mold of tomato in Japan. J Gen Plant Pathol 76: 84-86.
11. Bednarova M, Palovcikova D, Jankovsky L (2006) The host spectrum of Dothistroma needle blight *Mycosphaerella pini* E. Rostrup - new hosts of Dothistroma needle bight observed in the Czech Republic. J Forest Sci 52: 30-36.
12. Bradshaw RE (2004) Dothistroma (red-band) needle blight of pines and the dothistromin toxin: a review. Forest Pathol 34: 163-185.
13. Woods AJ, Coates KD, Hamann A (2005) Is an unprecedented Dothistroma needle blight epidemic related to climate change? Bioscience 55: 761-769.
14. Brown A, Webber JF (2008) Red band needle blight of conifers in Britain. Edinburgh, UK: Forestry Commission. 1-8 p.
15. Woods A (2011) Is the health of British Columbia's forests being influenced by climate change? If so, was this predictable? Can J Plant Pathol 33: 117-126.
16. Lazarovits G, Higgins VJ (1976) Ultrastucture of susceptible, resistant and immune reactions of tomato to races of *Cladosporium fulvum*. Can J Bot 54: 235-249.
17. Lazarovits G, Higgins VJ (1976) Histological comparison of *Cladosporium fulvum* race 1 on immune, resistant and susceptible tomato varieties. Can J Bot 54: 224-234.
18. De Wit PJGM (1977) A light and scanning-electron microscopic study of the infection of tomato plants by virulent and avirulent races of *Cladosporium fulvum*. Neth J Plant Pathol 83: 109-122.
19. Muir JA, Cobb JFW (2005) Infection of radiata and bishop pine by *Mycosphaerella pini* in California. Can J Forest Res 35: 2529-2538.
20. Gadgil PD (1967) Infection of *Pinus radiata* needles by *Dothistroma pini*. N Z J Bot 5: 498-503.
21. Stergiopoulos I, Groenewald M, Staats M, Lindhout P, Crous PW, et al. (2007) Mating-type genes and the genetic structure of a world-wide collection of the tomato pathogen *Cladosporium fulvum*. Fungal Genet Biol 44: 415-429.
22. Groenewald M, Barnes I, Bradshaw RE, Brown AV, Dale A, et al. (2007) Characterization and distribution of mating type genes in the Dothistroma needle blight pathogens. Phytopathology 97: 825-834.
23. Tomšovský M, Tomešová V, Palovčíková D, Kostovčík M, Rohrer M, et al. (2012) The gene flow and mode of reproduction of *Dothistroma septosporum* in the Czech Republic. Plant Pathol DOI: 10.1111/j.1365-3059.2012.02625.x.
24. Dale AL, Lewis KJ, Murray BW (2011) Sexual reproduction and gene flow in the pine pathogen *Dothistroma septosporum* in British Columbia. Phytopathology 101: 68-76.
25. Hirst P, Richardson TE, Carson SD, Bradshaw RE (1999) *Dothistroma pini* genetic diversity is low in New Zealand. N Z J Forest Sci 29: 459-472.

26. Joosten MHAJ, De Wit PJGM (1999) The tomato-*Cladosporium fulvum* interaction: A versatile experimental system to study plant-pathogen interactions. Annu Rev Phytopathol 37: 335-367.
27. De Wit PJGM, Joosten MHAJ, Thomma BPHJ, Stergiopoulos I (2009) Gene-for-gene models and beyond: the *Cladosporium fulvum*-tomato pathosystem. The Mycota V: 135-156.
28. De Jonge R, Van Esse HP, Kombrink A, Shinya T, Desaki Y, et al. (2010) Conserved fungal LysM effector Ecp6 prevents chitin-triggered immunity in plants. Science 329: 953-955.
29. De Kock MJD, Brandwagt BF, Bonnema G, De Wit PJGM, Lindhout P (2005) The tomato Orion locus comprises a unique class of Hcr9 genes. Mol Breeding 15: 409-422.
30. van Esse HP, Bolton MD, Stergiopoulos I, de Wit PJGM, Thomma BPHJ (2007) The chitin-binding *Cladosporium fulvum* effector protein Avr4 is a virulence factor. Mol Plant Microbe Interact 20: 1092-1101.
31. van den Burg HA, Harrison SJ, Joosten MHAJ, Vervoort J, de Wit PJGM (2006) *Cladosporium fulvum* Avr4 protects fungal cell walls against hydrolysis by plant chitinases accumulating during infection. Mol Plant Microbe Interact 19: 1420-1430.
32. Stergiopoulos I, Kourmpetis YAI, Slot JC, Bakker FT, De Wit PJGM, et al. (2012) *In silico* characterization and molecular evolutionary analysis of a novel superfamily of fungal effector proteins. Mol Biol Evol: doi: 10.1093/molbev/mss1143.
33. Stergiopoulos I, Van den Burg HA, Ökmen B, Beenen H, Kema GHJ, et al. (2010) Tomato Cf resistance proteins mediate recognition of cognate homologous effectors from fungi pathogenic on dicots and monocots. Proc Natl Acad Sci U S A 107: 7610-7615.
34. Shaw GJ, Chick M, Hodges R (1978) A ^{13}C-NMR study of the biosynthesis of the anthraquinone dothistromin by *Dothistroma pini*. Phytochemistry 17: 1743-1745.
35. Schwelm A, Barron NJ, Baker J, Dick M, Long PG, et al. (2009) Dothistromin toxin is not required for dothistroma needle blight in *Pinus radiata*. Plant Pathology 58: 293-304.
36. Bradshaw RE, Jin HP, Morgan BS, Schwelm A, Teddy OR, et al. (2006) A polyketide synthase gene required for biosynthesis of the aflatoxin-like toxin, dothistromin. Mycopathologia 161: 283-294.
37. Schwelm A, Barron NJ, Zhang S, Bradshaw RE (2008) Early expression of aflatoxin-like dothistromin genes in the forest pathogen *Dothistroma septosporum*. Mycological Research 112: 138-146.
38. Zhang S, Schwelm A, Jin HP, Collins LJ, Bradshaw RE (2007) A fragmented aflatoxin-like gene cluster in the forest pathogen *Dothistroma septosporum*. Fungal Genet Biol 44: 1342-1354.
39. Schwelm A, Bradshaw RE (2010) Genetics of dothistromin biosynthesis of *Dothistroma septosporum*: an update. Toxins 2: 2680-2698.
40. Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, et al. (2012) The genome portal of the Department of Energy Joint Genome Institute. Nucl Acids Res 40: D26-32.
41. van der Burgt A, Severing E, de Wit P, Collemare J (2012) Birth of new spliceosomal introns in fungi by multiplication of introner-like elements. Curr Biol 22: 1260-1265.
42. Amselem J, Cuomo CA, van Kan JAL, Viaud M, Benito EP, et al. (2011) Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. PLoS Genetics 7: e1002230.
43. Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, et al. (2010) Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. Science 330: 1543-1546.
44. Schmidt SM, Panstruga R (2011) Pathogenomics of fungal plant parasites: what have we learnt about pathogenesis? Curr Opin Plant Biol 14: 392-399.
45. Poulter RTM, Goodwin TJD, Butler MI (2003) Vertebrate helentrons and other novel helitrons. Gene 313: 201-212.
46. Kapitonov VV, Jurka J (2007) Helitrons on a roll: eukaryotic rolling-circle transposons. Trends Genet 23: 521-529.
47. Lamb JC, Theuri J, Birchler JA (2004) What's in a centromere? Genome Biol 5: 239.
48. Roy B, Sanyal K (2011) Diversity in requirement of genetic and epigenetic factors for centromere function in fungi. Eukaryot Cell 10: 1384-1395.
49. Freitag M, Williams RL, Kothe GO, Selker EU (2002) A cytosine methyltransferase homologue is essential for repeat-induced point mutation in *Neurospora crassa*. Proc Natl Acad Sci U S A 99: 8802-8807.
50. Hane JK, Lowe RGT, 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.

51. Fudal I, Ross S, Brun H, Besnard A-L, 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.
52. 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.
53. Rouxel T, Grandaubert J, Hane JK, Hoede C, van de Wouw AP, et al. (2011) Effector diversification within compartments of the *Leptosphaeria maculans* genome affected by Repeat-Induced Point mutations. Nat Commun 2: 202.
54. Nowrousian M, Stajich JE, Chu M, Engh I, Espagne E, et al. (2010) *De novo* assembly of a 40 Mb eukaryotic genome from short sequence reads: *Sordaria macrospora*, a model organism for fungal morphogenesis. PLoS Genetics 6: e1000891.
55. Hane JK, Rouxel T, Howlett BJ, Kema GHJ, Goodwin SB, et al. (2011) A novel mode of chromosomal evolution peculiar to filamentous Ascomycete fungi. Genome Biol 12: R45.
56. Thon MR, Pan HQ, Diener S, Papalas J, Taro A, et al. (2006) The role of transposable element clusters in genome evolution and loss of synteny in the rice blast fungus *Magnaporthe oryzae*. Genome Biol 7: R16.
57. Stergiopoulos I, De Wit PJGM (2009) Fungal effector proteins. Annu Rev Phytopathol 47: 233-263.
58. 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 Genetics 7: e1002070.
59. Raffaele S, Farrer RA, Cano LM, Studholme DJ, MacLean D, et al. (2010) Genome evolution following host jumps in the Irish potato famine pathogen lineage. Science 330: 1540-1543.
60. Rivas S, Thomas CM (2005) Molecular interactions between tomato and the leaf mold pathogen *Cladosporium fulvum*. Annu Rev Phytopathol 43: 395-436.
61. Wilcox PL, Amerson HV, Kuhlman EG, Liu BH, O'Malley DM, et al. (1996) Detection of a major gene for resistance to fusiform rust disease in loblolly pine by genomic mapping. Proc Natl Acad Sci U S A 93: 3859-3864.
62. Liu JJ, Ekramoddoullah AKM (2011) Genomic organization, induced expression and promoter activity of a resistance gene analog (PmTNL1) in western white pine (*Pinus monticola*). Planta 233: 1041-1053.
63. Nelson CD, Kubisiak TL, Amerson HV (2010) Unravelling and managing fusiform rust disease: a model approach for coevolved forest tree pathosystems. Forest Pathol 40: 64-72.
64. van den Ackerveken GFJM, van Kan JAL, de Wit PJGM (1992) Molecular analysis of the avirulence gene *Avr9* of the fungal tomato pathogen *Cladosporium fulvum* fully supports the gene-for-gene hypothesis. Plant J 2: 359-366.
65. Chuma I, Isobe C, Hotta Y, Ibaragi K, Futamata N, et al. (2011) Multiple translocation of the AVR-Pita effector gene among chromosomes of the rice blast fungus *Magnaporthe oryzae* and related species. PLoS Pathogens 7: e1002147.
66. Westerink N, Brandwagt BF, De Wit PJGM, Joosten MHAJ (2004) *Cladosporium fulvum* circumvents the second functional resistance gene homologue at the Cf-4 locus (Hcr9-4E) by secretion of a stable avr4E isoform. Mol Microbiol 54: 533-545.
67. Laugé R, Goodwin PH, De Wit PJGM, Joosten MHAJ (2000) Specific HR-associated recognition of secreted proteins from *Cladosporium fulvum* occurs in both host and non-host plants. Plant J 23: 735-745.
68. Wessels JGH (1994) Developmental regulation of fungal cell-wall formation. Annu Rev Phytopathol 32: 413-437.
69. Templeton MD, Rikkerink EHA, Beever RE (1994) Small, cysteine-rich proteins and recognition in fungal-plant interactions. Mol Plant Microbe Interact 7: 320-325.
70. Joosten MHAJ, Verbakel HM, Nettekoven ME, van Leeuwen J, van der Vossen RTM, et al. (2005) The phytopathogenic fungus *Cladosporium fulvum* is not sensitive to the chitinase and β-1,3-glucanase defence proteins of its host, tomato. Physiol Mol Plant Pathol 46: 45-59.
71. Spanu P (1997) HCF-1, a hydrophobin from the tomato pathogen *Cladosporium fulvum*. Gene 193: 89-96.
72. Whiteford JR, Spanu PD (2001) The hydrophobin HCf-1 of *Cladosporium fulvum* is required for efficient water-mediated dispersal of conidia. Fungal Genet Biol 32: 159-168.
73. Lacroix H, Whiteford JR, Spanu PD (2008) Localization of *Cladosporium fulvum* hydrophobins reveals a role for HCf-6 in adhesion. FEMS Microbiol Lett 286: 136-144.

74. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, et al. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. Nucl Acids Res 37: D233-238.
75. Martinez D, Berka RM, Henrissat B, Saloheimo M, Arvas M, et al. (2008) Genome sequencing and analysis of the biomass-degrading fungus *Trichoderma reesei* (syn. *Hypocrea jecorina*). Nat Biotechnol 26: 553-560.
76. Putoczki TL, Gerrard JA, Butterfield BG, Jackson SL (2008) The distribution of un-esterified and methyl-esterified pectic polysaccharides in *Pinus radiata*. IAWA J 29: 115-127.
77. Ridley BL, O'Neill MA, Mohnen DA (2001) Pectins: structure, biosynthesis, and oligogalacturonide-related signaling. Phytochemistry 57: 929-967.
78. Sprockett DD, Piontkivska H, Blackwood CB (2011) Evolutionary analysis of glycosyl hydrolase family 28 (GH28) suggests lineage-specific expansions in necrotrophic fungal pathogens. Gene 479: 29-36.
79. Ten Have A, Mulder W, Visser J, van Kan JAL (1998) The endopolygalacturonase gene Bcpg1 is required for full virulence of *Botrytis cinerea*. Mol Plant Microbe Interact 11: 1009-1016.
80. van Kan JAL (2006) Licensed to kill: the lifestyle of a necrotrophic plant pathogen. Trends Plant Sci 11: 247-253.
81. Eastwood DC, Floudas D, Binder M, Majcherczyk A, Schneider P, et al. (2011) The plant cell wall-decomposing machinery underlies the functional diversity of forest fungi. Science 333: 762-765.
82. Ohm RA, de Jong JF, Lugones LG, Aerts A, Kothe E, et al. (2010) Genome sequence of the model mushroom *Schizophyllum commune*. Nat Biotechnol 28: 957-U910.
83. van den Brink J, de Vries RP (2011) Fungal enzyme sets for plant polysaccharide degradation. Appl Microbiol Biotechnol 91: 1477-1492.
84. Lee MH, Chiu CM, Roubtsova T, Chou CM, Bostock RM (2010) Overexpression of a redox-regulated cutinase gene, MfCUT1, increases virulence of the brown rot pathogen *Monilinia fructicola* on *Prunus* spp. Mol Plant Microbe Interact 23: 176-186.
85. Skamnioti P, Gurr SJ (2008) Cutinase and hydrophobin interplay: A herald for pathogenesis? Plant Signal Behav 3: 248-250.
86. King BC, Waxman KD, Nenni NV, Walker LP, Bergstrom GC, et al. (2011) Arsenal of plant cell wall degrading enzymes reflects host preference among plant pathogenic fungi. Biotechnol Biofuels 4: 4.
87. de Vries RP, Burgers K, van de Vondervoort PJI, Frisvad JC, Samson RA, et al. (2004) A new black *Aspergillus* species, *A. vadensis*, is a promising host for homologous and heterologous protein production. Appl Environ Microbiol 70: 3954-3959.
88. Coutinho PM, Andersen MR, Kolenova K, vanKuyk PA, Benoit I, et al. (2009) Post-genomic insights into the plant polysaccharide degradation potential of *Aspergillus nidulans* and comparison to *Aspergillus niger* and *Aspergillus oryzae*. Fungal Genet Biol 46: S161-S169.
89. Berg B, Deanta RC, Escudero A, Gardenas A, Johansson MB, et al. (1995) The chemical composition of newly shed needle litter of scots pine and some other pine species in a climatic transect. X long-term decomposition in a scots pine forest. Can J Bot 73: 1423-1435.
90. Vogel J (2008) Unique aspects of the grass cell wall. Curr Opin Plant Biol 11: 301-307.
91. Pareja-Jaime Y, Roncero MIG, Ruiz-Roldán MC (2008) Tomatinase from *Fusarium oxysporum* f. sp. *lycopersici* is required for full virulence on tomato plants. Mol Plant Microbe Interact 21: 728-736.
92. Barthe GA, Jourdan PS, McIntosh CA, Mansell RL (1988) Radioimmunoassay for the quantitative determination of hesperidin and analysis of its distribution in *Citrus sinensis*. Phytochemistry 27: 249-254.
93. DiGuistini S, Wang Y, Liao NY, Taylor G, Tanguay P, et al. (2011) Genome and transcriptome analyses of the mountain pine beetle-fungal symbiont *Grosmannia clavigera*, a lodgepole pine pathogen. Proc Natl Acad Sci U S A 108: 2504-2509.
94. Hammerschmidt R, Bonnen AM, Bergstrom GC (1983) Association of lignification with non-host resistance of cucurbits. Phytopathology 73: 829-829.
95. Xu L, Zhu L, Tu L, Liu L, Yuan D, et al. (2011) Lignin metabolism has a central role in the resistance of cotton to the wilt fungus *Verticillium dahliae* as revealed by RNA-Seq-dependent transcriptional analysis and histochemistry. J Exp Bot 62: 5607–5621.
96. Berg B, Wessen B, Ekbohm G (1982) Nitrogen level and decomposition in scots pine needle litter. Oikos 38: 291-296.
97. Lukjanova A, Mandre M (2008) Anatomical structure and localisation of lignin in needles and shoots of Scots pine (*Pinus sylvestris* L.) growing in a habitat with varying environmental characteristics. Forest Stud 49: 37-46.

98. Kirk TK, Farrell RL (1987) Enzymatic "combustion": the microbial degradation of lignin. Annu Rev Microbiol 41: 465-505.
99. Ruiz-Dueñas FJ, Martínez AT (2009) Microbial degradation of lignin: how a bulky recalcitrant polymer is efficiently recycled in nature and how we can take advantage of this. Microb Biotechnol 2: 164-177.
100. Collemare J, Lebrun M-H (2012) Fungal secondary metabolites: ancient toxins and novel effectors in plant-microbe interactions. In: Martin F, Kamoun S, editors. Effectors in Plant-Microbe Interactions: Blackwell Publishing Ltd. pp. 379-402.
101. Davies DG, Hodge P (1974) Chemistry of quinones. 5. Structure of cladofulvin, a bianthraquinone from *Cladosporium fulvum* Cooke. J Chem Soc-Perkin Trans 1: 2403-2405.
102. Bradshaw RE, Bhatnagar D, Ganley RJ, Gillman CJ, Monahan BJ, et al. (2002) *Dothistroma pini*, a forest pathogen, contains homologs of aflatoxin biosynthetic pathway genes. Appl Environ Microbiol 68: 2885-2892.
103. Keller NP, Hohn TM (1997) Metabolic pathway gene clusters in filamentous fungi. Fungal Genet Biol 21: 17-29.
104. Keller NP, Turner G, Bennett JW (2005) Fungal secondary metabolism - From biochemistry to genomics. Nat Rev Microbiol 3: 937-947.
105. Collemare J, Billard A, Boehnert HU, Lebrun MH (2008) Biosynthesis of secondary metabolites in the rice blast fungus *Magnaporthe grisea*: the role of hybrid PKS-NRPS in pathogenicity. Mycol Res 112: 207-215.
106. Galagan JE, Calvo SE, Cuomo C, Ma LJ, Wortman JR, et al. (2005) Sequencing of *Aspergillus nidulans* and comparative analysis with *A. fumigatus* and *A. oryzae*. Nature 438: 1105-1115.
107. Cuomo CA, Gueldener 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.
108. Khaldi N, Wolfe KH (2011) Evolutionary origins of the fumonisin secondary metabolite gene cluster in *Fusarium verticillioides* and *Aspergillus niger*. Internat J Evol Biol 2011: 423821.
109. Yang Z, Bielawski JP (2000) Statistical methods for detecting molecular adaptation. Trends Ecol Evol 15: 496–503.
110. Ehrlich KC, Yu JJ, Cotty PJ (2005) Aflatoxin biosynthesis gene clusters and flanking regions. J Appl Microbiol 99: 518-527.
111. Bergmann S, Funk AN, Scherlach K, Schroeckh V, Shelest E, et al. (2010) Activation of a silent fungal polyketide biosynthesis pathway through regulatory cross talk with a cryptic nonribosomal peptide synthetase gene cluster. Appl Environ Microbiol 76: 8143-8149.
112. Brakhage AA, Schroeckh V (2011) Fungal secondary metabolites - Strategies to activate silent gene clusters. Fungal Genet Biol 48: 15-22.
113. Bayram O, Braus GH (2011) Coordination of secondary metabolism and development in fungi: the velvet family of proteins. FEMS Microbiol Rev 35: 1-24.
114. Tsitsigiannis DI, Keller NP (2006) Oxylipins act as determinants of natural product biosynthesis and seed colonisation in *Aspergillus nidulans*. Mol Microbiol 59: 882-892.
115. Adams TH, Boylan MT, Timberlake WE (1988) BrlA is necessary and sufficient to direct conidiophore development in *Aspergillus nidulans*. Cell 54: 353-362.
116. Enright AJ, Van Dongen S, Ouzounis CA (2002) An efficient algorithm for large-scale detection of protein families. Nucl Acids Res 30: 1575-1584.
117. Katoh K, Toh H (2008) Recent developments in the MAFFT multiple sequence alignment program. Briefings Bioinf 9: 286-298.
118. Stamatakis A, Hoover P, Rougemont J (2008) A rapid bootstrap algorithm for the RAxML Web servers. Syst Biol 57: 758-771.
119. Bao ZR, Eddy SR (2002) Automated *de novo* identification of repeat sequence families in sequenced genomes. Genome Res 12: 1269-1276.
120. Smit AFA, Hubley R, Green P (1996-2010) RepeatMasker Open-3.0.
121. Lewis ZA, Honda S, Khlafallah TK, Jeffress JK, Freitag M, et al. (2009) Relics of repeat-induced point mutation direct heterochromatin formation in *Neurospora crassa*. Genome Res 19: 427-437.
122. Delcher AL, Salzberg SL, Phillippy AM (2003) Using MUMmer to identify similar regions in large sequence sets. Curr Protocol Bioinf 10.3.1–10.3.18.
123. Park J, Park B, Jung K, Jang S, Yu K, et al. (2008) CFGP: a web-based, comparative fungal genomics platform. Nucl Acids Res 36: D562-D571.

124. Horton P, Park K-J, Obayashi T, Fujita N, Harada H, et al. (2007) WoLF PSORT: protein localization predictor. Nucl Acids Res 35: W585-W587.

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

126. Kaell L, Krogh A, Sonnhammer ELL (2007) Advantages of combined transmembrane topology and signal peptide prediction - the Phobius web server. Nucl Acids Res 35: W429-W432.

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

128. Emanuelsson O, Nielsen H, Brunak S, von Heijne G (2000) Predicting subcellular localization of proteins based on their N-terminal amino acid sequence. J Mol Biol 300: 1005-1016.

129. Pierleoni A, Martelli PL, Casadio R (2008) PredGPI: a GPI-anchor predictor. BMC Bioinf 9: 392.

130. van der Hoorn RAL, Laurent F, Roth R, de Wit PJGM (2000) Agroinfiltration is a versatile tool that facilitates comparative analyses of Avr9/Cf-9-induced and Avr4/Cf-4-induced necrosis. Mol Plant Microbe Interact 13: 439-446.

131. Hammond-Kosack KE, Staskawicz BJ, Jones JDG, Baulcombe DC (1995) Functional expression of a fungal avirulence gene from a modified Potato-Virus-X genome. Mol Plant Microbe Interact 8: 181-185.

132. Karimi M, Inze D, Depicker A (2002) GATEWAY vectors for *Agrobacterium*-mediated plant transformation. Trends Plant Sci 7: 193-195.

133. Whiteford JR, Spanu PD (2002) Hydrophobins and the interactions between fungi and plants. Mol Plant Pathol 3: 391-400.

134. Tamura K, Peterson D, Peterson N, Stecher G, Nei M, et al. (2011) MEGA5: Molecular Evolutionary Genetics Analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731-2739.

135. Wernersson R, Pedersen AG (2003) RevTrans: Multiple alignment of coding DNA from aligned amino acid sequences. Nucl Acids Res 31: 3537-3539.

136. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ (2009) Jalview version 2 — a multiple sequence alignment editor and analysis workbench. Bioinformatics 25: 1189-1191.

137. Zhang Z, Li J, Zhao XQ, Wang J, Wong GK, et al. (2006) KaKs Calculator: Calculating Ka and Ks through model selection and model averaging. Genomics Proteomics Bioinf 4: 259-263.

138. Nei M, Gojobori T (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. Mol Biol Evol 3: 418-426.

139. Student (1908) The probable error of a mean. Biometrika 6: 1-25.

140. Chettri P, Calvo AM, Cary JW, Dhingra S, Guo YA, et al. (2012) The *veA* gene of the pine needle pathogen *Dothistroma septosporum* regulates sporulation and secondary metabolism. Fungal Genet Biol 49: 141-151.

141. Gamborg OL, Miller RA, Ojima K (1968) Nutrient requirements of suspension cultures of soybean root cell Exp Cell Res 50: 151-158.

142. van Esse HP, Van 't Klooster JW, Bolton MD, Yadeta KA, van Baarlen P, et al. (2008) The *Cladosporium fulvum* virulence protein Avr2 inhibits host proteases required for basal defense. Plant Cell 20: 1948-1963.

143. Untergasser A, Nijveen H, Rao X, Bisseling T, Geurts R, et al. (2007) Primer3Plus, an enhanced web interface to Primer3. Nucl Acids Res 35: W71-75.

144. Livak KJ, Schmittgen TD (2001) Analysis of relative gene expression data using real time quantitative PCR and the 2-ΔΔCT method. Methods 25: 402-408.

Supporting Information

Figure S1. Amino acid similarity to *Dothistroma septosporum.* Genome-wide amino acid similarity of homologous proteins between *C. fulvum* and other sequenced fungal species. A pair of proteins is only reported as homologous when the predicted similarity (blastp) spans at least 70% of their lengths and their length difference is at most 20%. Axis indicates number of homologous proteins. Bar shading indicates similarity: red, 91-100%; orange, 81-90%; light green, 71-80%; medium green, 61-70%; turquoise, 51-60%; light blue, 41-50%; dark blue, 31-40%; and purple, 0-30%. Homologous proteins with high amino acid similarity are likely orthologs, whereas for those with lower similarity this relation cannot be inferred. Species abbreviations: Cf, *Cladosporium fulvum*; Sm, *Septoria musiva*; Mf, *Mycosphaerella fijiensis*; Mg, *Mycosphaerella graminicola*; Rr, *Rhytidhysteron rufulum*; Hp, *Hysterium pulicare*; Ab, *Alternaria brassicicola*; Pt, *Pyrenophora tritici-repentis*; Ch, *Cochliobolus heterostrophus*; Sn, *Stagonospora nodorum*; An, *Aspergillus nidulans*; Nc, *Neurospora crassa*; Mo, *Magnaporthe oryzae*; Fg, *Fusarium graminearum*.

Figure S2. Synteny of Cladosporium fulvum and Dothistroma septosporum genomes.
A) Whole-genome synteny, computed using the JGI synteny browser (available at http://genome.jgi-psf.org/Dotse1/Dotse1.home.html) and shown as pairwise alignment blocks. The 14 main *D. septosporum* chromosomes are shown with colored blocks indicating regions of synteny with *C. fulvum,* computed with a 1 kb cut-off. The different colored blocks represent the different *C. fulvum* scaffolds; colour keys were generated independently for the 14 chromosomes and are as shown on the JGI synteny browser. **B)** Synteny between two examples of large scaffolds of *C. fulvum* (top, Cf2, 526 kb; bottom, Cf5, 315 kb) with corresponding regions of *D. septosporum* chromosomes 2 (3.3 Mb) and 4 (2.6 Mb). In each case the top line (eg Cf 2) shows blocks indicating regions of synteny with *D. septosporum,* computed with a 50-bp cut-off. Different colours represent matches to different chromosomes, indicating that each *C. fulvum* scaffold predominantly aligns with just one *D. septosporum* chromosome. Slanting lines connecting Cf and Ds scaffolds show matching regions between the two, illustrating that the syntenic blocks are not collinear but show fragmentation and different degrees of dispersal, consistent with intrachromosomal rearrangements.

Figure S3. Hydrophobin genes in *Dothideomycete* species. Consensus phylogenetic tree of predicted class I and class II hydrophobins from *Cladosporium fulvum* (Cf), *Dothistroma septosporum* (Ds), and *Mycosphaerella graminicola* (Mg). Amino acid sequences of hydrophobins were aligned using ClustalW2 and the phylogenetic tree was constructed using the minimum-evolution method of MEGA5 with 1000 bootstraps. Bootstrap values less than 90% are not shown. Scale bar shows the genetic distance (substitutions per site). The letters c and p in brackets indicate whether expression was detected in culture or *in planta*, respectively, in EST data. Asterisks indicate hydrophobin proteins for which no hydrophobin conserved domain could be identified using InterproScan. However, the phylogenetic tree clearly shows that Ds67650 belongs to class I, and Mg965336 and Cf183780 belong to class II hydrophobins.

Figure S4. Growth profile assays for *Cladosporium fulvum* and *Dothistroma septosporum*. The fungi were grown on 32 solid agar media containing well-defined or complex carbohydrate substrates as detailed at www.fung-growth.org. **A)** *C. fulvum* grown for 2 weeks and **B)** *D. septosporum* grown for 4 weeks, both in the dark at 22-25°C.

Figure S5. Synteny of secondary metabolism loci between *Dothistroma septosporum* and *Cladosporium fulvum*. Gene clusters of *D. septosporum*- **(A)** and *C. fulvum*-specific **(B)** key genes were predicted based on functional annotations. Synteny of the borders of each gene cluster was checked in the other species. All *D. septosporum*-specific gene clusters are located at conserved loci in *C. fulvum*. Conversely, only 6 *C. fulvum*-specific gene clusters are located at conserved loci in *D. septosporum*. For *Cf-Pks5*, *Cf-Pks6* and *Cf-Nps5*, the border genes are scattered over one or more chromosomes in *D. septosporum*. Black arrows represent key genes; dark grey arrows represent putative accessory genes; light grey arrows represent putative border genes; outlined arrows indicate genes present in only one species; black triangles represent transposable elements. The *C. fulvum* scaffold numbers and *D. septosporum* chromosome numbers are indicated for syntenic loci. Loci are not drawn to scale.

Figure S6. Expression of *Cladosporium fulvum* effector genes *Avr4* and *Avr9* during infection of tomato. Expression of *Avr4* and *Avr9* was measured by quantitative PCR during tomato infection and in two *in vitro* conditions (PDB and B5 media). Expression was calibrated using the tubulin gene according to the $2^{-\Delta Ct}$ method [136]. Expression was not detected *in vitro* for *Avr9* and weakly in B5 medium only for *Avr4*. Induction of the expression of both effectors during tomato infection is highlighted by the grey-dashed curve. Relative expression of actin is shown as the control for calibration.

Protocol S1 Genome sequencing, assembly and annotation methods.

Table S1 Cladosporium fulvum and Dothistroma septosporum sequence statistics.

Table S2 *Dothistroma septosporum* genome scaffolds.

Table S3 Overview of Repeat-Induced Point Mutations (RIP) in *Cladosporium fulvum, Dothistroma septosporum* and other related *Dothideomycete* fungi. *Neurospora crassa* is used as a reference.

Table S4 Repetitive regions flanking known effectors of *Cladosporium fulvum* and *Dothistroma septosporum.*

Table S5 Comparison of CAZy gene numbers in *Cladosporium fulvum* and *Dothistroma septosporum*.

Table S6 Growth diameters of *Cladosporium fulvum, Dothistroma septosporum* and other fungi on various carbon sources.

Table S7 Putative monoterpene-degrading genes.

Table S8 Comparison of oxidoreductase gene numbers in Cladosporium fulvum, Dothistroma septosporum, Mycosphaerella graminicola and Stagonospora nodorum.

Table S9 Key secondary metabolism enzyme identifiers in *Cladosporium fulvum* and *Dothistroma septosporum*.

Table S10 Putative dothistromin genes in *Cladosporium fulvum* and *Dothistroma septosporum*.

Table S11 Regulatory genes involved in development and secondary metabolism of *Cladosporium fulvum* and *Dothistroma septosporum*.

Table S12 Conditions for *Cladosporium fulvum* EST libraries.

Table S13 Conditions for *Dothistroma septosporum* EST libraries.

Table S14 Primers used for reverse transcription quantitative PCR.

Chapter 3

Automated alignment-based curation of gene models in filamentous fungi.

van der Burgt A, Severing E, Collemare J, de Wit PJGM.

BMC Bioinformatics 2014, 15(1):19.

doi: 10.1186/1471-2105-15-19

This article is available from: http://www.biomedcentral.com/1471-2105/15/19
Additional Files are available from:
http://www.biomedcentral.com/1471-2105/15/19/additional

Abstract

Background: Automated gene-calling is still an error-prone process, particularly for the highly plastic genomes of fungal species. Improvement through quality control and manual curation of gene models is a time-consuming process that requires skilled biologists and is only marginally performed. The wealth of available fungal genomes has not yet been exploited by an automated method that applies quality control of gene models in order to obtain more accurate genome annotations.

Results: We provide a novel method named alignment-based fungal gene prediction (ABFGP) that is particularly suitable for plastic genomes like those of fungi. It can assess gene models on a gene-by-gene basis making use of informant gene loci. Its performance was benchmarked on 6,965 gene models confirmed by full-length unigenes from ten different fungi. 79.4% of all gene models were correctly predicted by ABFGP. It improves the output of ab initio gene prediction software due to a higher sensitivity and precision for all gene model components. Applicability of the method was shown by revisiting the annotations of six different fungi, using gene loci from up to 29 fungal genomes as informants. Between 7,231 and 8,337 genes were assessed by ABFGP and for each genome between 1,724 and 3,505 gene model revisions were proposed. The reliability of the proposed gene models is assessed by an a posteriori introspection procedure of each intron and exon in the multiple gene model alignment. The total number and type of proposed gene model revisions in the six fungal genomes is correlated to the quality of the genome assembly, and to sequencing strategies used in the sequencing centre, highlighting different types of errors in different annotation pipelines. The ABFGP method is particularly successful in discovering sequence errors and/or disruptive mutations causing truncated and erroneous gene models.

Conclusions: The ABFGP method is an accurate and fully automated quality control method for fungal gene catalogues that can be easily implemented into existing annotation pipelines. With the exponential release of new genomes, the ABFGP method will help decreasing the number of gene models that require additional manual curation.

Background

In the past decade, numerous fungal genomes of importance to medicine, agriculture and industry have been sequenced [1,2] and continuous innovations in next generation sequencing technology will spur this number to rapidly increase further. Once sequenced and assembled, genomes are annotated through an automated gene-calling pipeline, which is still an error-prone process, particularly for the highly plastic and diverse genomes of fungal species.

Most gene annotation pipelines integrate different gene prediction algorithms to increase the accuracy of the annotation [3]. These algorithms include ab initio supervised, ab initio unsupervised and (supervised) alignment-based gene predictors, which are implemented in tools such as Augustus [4], GeneMark-ES [5] and TWINSCAN 2.0α [6], respectively. Augustus is one of the most frequently employed and best performing ab initio supervised gene prediction tools that offers parameterizations for several dozens of fungi [4]. For species lacking a provided parameterization, a considerable manual input is required to obtain such species-specific parameterization by training the algorithm with a large sample (~1000) of correct gene models [5]. Thus, its applicability is limited to only those species for which parameterization is available [5,6]. GeneMark-ES-2 is an ab initio unsupervised gene predictor iteratively training itself on the input genome sequence alone that outperformed Augustus [5], but is reported to be relatively inaccurate in predicting single exon genes [5]. A hybrid strategy between ab initio and alignment- (or evidence) based gene prediction is currently implemented in several tools. Updated versions of Augustus integrate evidence obtained from unigene alignments [4], protein multiple sequence alignments [7] and intron- and exon-hints acquired from RNA-Seq data, which greatly improved their prediction accuracy. To our knowledge, alignment-based gene prediction in fungi using genomic data alone has only been successfully applied using TWINSCAN 2.0 α, which was specifically adapted and trained to *Cryptococcus neoformans* [6]. In that case, the whole-genome DNA alignment of two strains of this fungus, whose genomes are largely syntenic and exhibit around 95% nucleotide identity in coding regions, served as input. The reported ~60% gene accuracy clearly outperformed non-alignment-based ab initio gene prediction software [6]. TWINSCAN 2.0 α requires extensive species-specific training and parameterization, offering a tailor-made solution for a defined pair of related species only. Most importantly, the approach taken in TWINSCAN 2.0 α is difficult to apply to fungal genomes because of their high plasticity [8-10]. The absence of conserved regions exhibiting macro- or even meso-synteny between related fungal genomes [8] severely hampers the construction of whole-genome DNA alignments. Besides reshuffled gene orders, a highly variable gene content is also observed among fungi with a large number of genes showing a discontinuous distribution in the fungal tree of life. This is caused by frequent gene, gene-cluster, segmental and whole chromosome duplications, losses or horizontal transfers, which have created complex variation in both gene family expansion and reduction [8,10]. Although homologous gene loci can often be inferred easily between distantly related fungi, annotation of fungal genomes by classical alignment-based gene prediction tools is problematic. In recent years, ensemble predictors have been developed to weigh and combine similarity evidence and the predictions made by various other tools into a single, more accurate gene model [11,12]. However, it “often requires significant effort in implementation to cast comparative information into a form compatible with the existing gene models” [13].

Because none of the available gene prediction tools were specifically developed for fungal genomes, automatic gene annotation of fungi often yields a relatively high fraction of incorrect gene models. These can only be revised through a time-consuming process of quality control and manual curation by skilled biologists or bioinformaticians, but this is often only marginally performed. Manual curation usually involves comparative analyses with tools that can accurately identify a spliced gene structure in a target DNA sequence using a homologous protein sequence as a so-called "informant" sequence (e.g. GeneWise [14], Scipio [15], etc.). However, a large proportion of gene models and derived protein sequences in current fungal sequence releases contain errors, and a manual curator can easily propagate existing errors when using incorrectly predicted informant protein(s). A typical example of the marginal quality of fungal gene catalogues is exemplified by the re-annotation of the *Fusarium graminearum* genome [16]. In the new version, 1,770 gene models were revised by using various new gene predictors, exploiting expression data, performing extensive manual curation and evidence-based selection of the best gene model from alternative predictions [16]. Despite this effort, recent RNA-Seq data provided experimental proof for at least another 655 incorrectly predicted gene models in the latest version of the *F. graminearum* annotation [17].

We have now entered an era in which genome sequencing of clusters of related fungi will be performed on a massive scale. Subsequent gene prediction on these genomes will require automation with very little manual inspection [6]. Although gene prediction software suitable for fungal genomes has become more accurate over the last decade, they are still error-prone. A method that facilitates or automates the process of curating gene models is therefore needed to increase the accuracy of the catalogues of predicted genes in sequenced fungal genomes. Here, we present a novel gene-by-gene method for alignment-based gene prediction that is particularly suitable for the plastic genomes of fungi. Our method, called alignment-based fungal gene prediction (ABFGP), (i) provides improved accuracy of predicted gene models, (ii) is species-independent, (iii) does not require partial or whole-genome DNA alignments, (iv) does not require supervision and (v) can use a variable number of informant genes. We demonstrate the accuracy and versatility of the ABFGP method by re-annotating the genomes of a selection of six sequenced Ascomycete fungi.

Results

The alignment-based fungal gene prediction (ABFGP) method

The ABFGP method re-annotates gene models on a gene-by-gene basis by using informants, which differ from regular alignment-based approaches that require a whole-genome DNA alignment. An ABFGP informant refers to the genomic locus at which an homologous gene is encoded that may support revision of the target gene locus. First, a similarity matrix of predicted protein sequences from several fungal species is obtained (Figure 1; Additional file 1). From this matrix, bi-directional best hits (BDBH) with sufficient overlap between both annotated proteins are selected, representing most likely orthologous informant gene loci. Subsequently, the genomic loci that encodes these proteins - not the predicted proteins themselves - are used as informants to avoid propagation of errors in the gene structures. Other resources can be used to find informant gene loci such as unigene datasets or any alternative homology search (Figure 1).

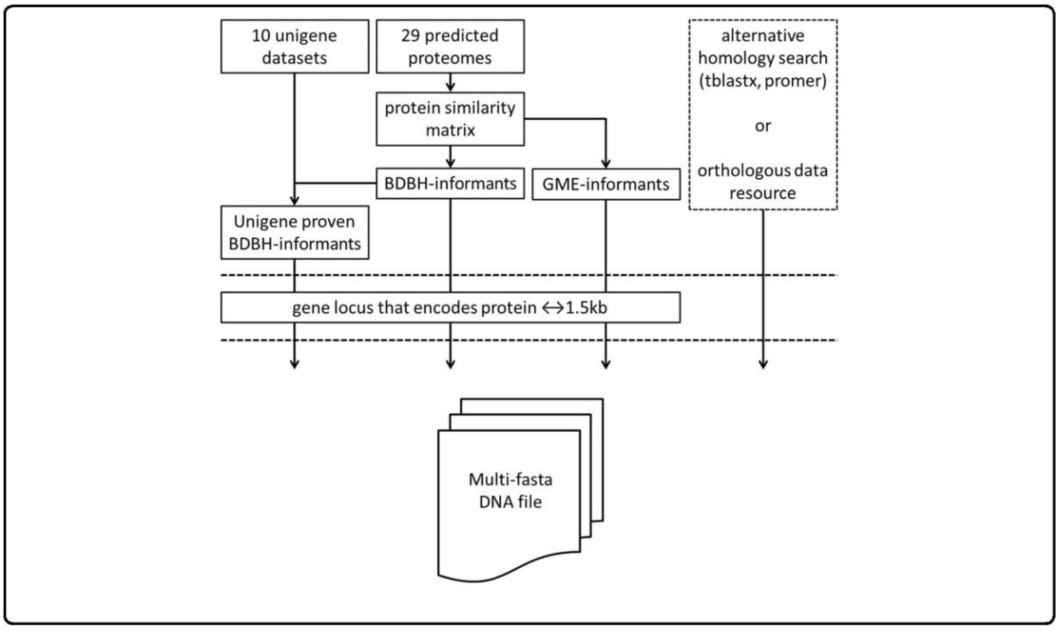

10 unigene datasets
29 predicted proteomes
protein similarity matrix
BDBH-informants
GME-informants
Unigene proven BDBH-informants
alternative homology search (tblastx, promer)
or
orthologous data resource
gene locus that encodes protein ↔1.5kb
Multi-fasta DNA file

Figure 1 Flow diagram of informant gene selection for the alignment-based fungal gene prediction (ABFGP) method.

The output of an ABFGP execution is a GFF file containing the predicted gene model and several features that assist manual inspection of the predicted gene model. Input for ABFGP is a list of orthologous gene loci, of which one is assigned as the target locus to be re-annotated, and all others serve as informants. This resulting list of gene encoding loci is provided as a multi fasta file. A second input option provides additional functionality, where each (informant) gene locus is a folder that contains the genomic locus (fasta format), optionally its currently annotated gene model (gff format) and unigenes aligned to this locus (gff format). A provided unigene is used as an additional informant, from which spliced alignments are exploited as guidance to infer intron-exon boundaries to enhance the prediction performance. A provided gene structure is used to speed up similarity searches by prioritization and to visualize differences between current annotation and the ABFGP prediction. Optionally, the exons of provided genes can be used as prior knowledge to facilitate detection of poorly conserved parts of the gene.

The ABFGP method is an automated workflow that includes all steps typically undertaken when performing manual annotation of a predicted gene model. It comprises nucleotide and protein similarity searches (BLAST, ClustalW and HMM) to build (pairwise) alignments, motif searches (SignalP and TMHMM) and degenerate Position Specific Scoring Matrix (PSSM) searches to identify elements of gene structure [18] including splice sites, branch points, polypyrimidine tracks and translational start sites. A flow diagram of the consecutive steps undertaken in the ABFGP method is presented in Figure 2. Graph-theory is used to translate pairwise alignments of sequences, open reading frames (ORFs), sequence elements or positional attributes to multiple alignments of these entities. The gene similarity graph is an estimation of the gene tree and is used to favor, demote or remove nodes and edges from the ORF similarity graph. Inconsistencies or missing data in series of multiple aligned ORFs trigger a more sensitive HMMER protein search, which can identify missing ORFs of target or informant genes or can recognize lower similarity. The ABFGP method accurately predicts intron-exon boundaries by exploiting ORF (dis)continuity surrounding intron presence-absence patterns [19]. In contrast to ab initio gene prediction software, the ABFGP method is able to cope with sequence errors (SEs) and true disruptive mutations (DMs), and recognizes those as inconsistencies in coding region continuity. A quality check on the similarity graph is performed at various stages during ABFGP execution, which can result in removal of an informant once recognized as too distinct. In case the target gene locus is distinct from all informants (which are all homologous to each other), they are all removed. Once the number of informants drops below a user-adjustable threshold (by default set at four), execution of the method is aborted. Finally, an a posteriori introspection procedure is applied to each intron and exon in the predicted gene model which assigns a reliability label ('ok' or 'doubtful') to the predicted gene model.

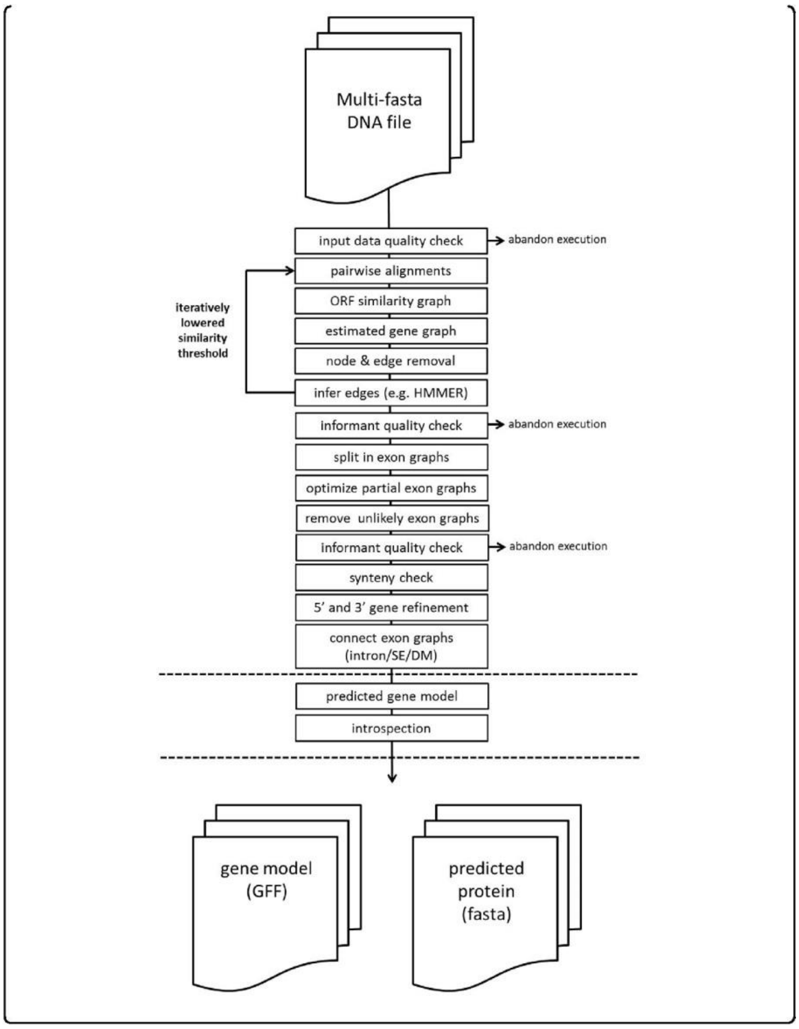
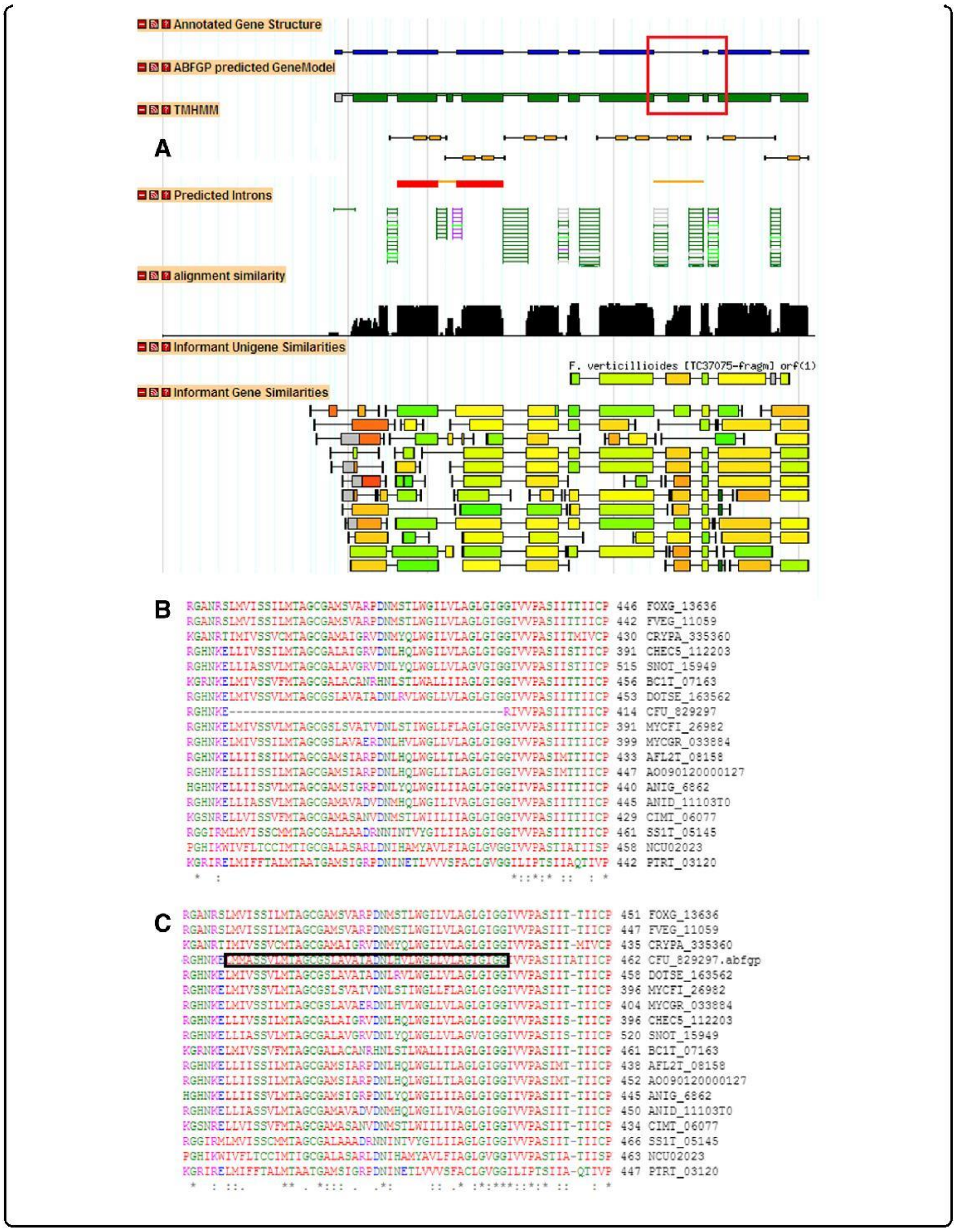


Figure 2 Flow diagram of the ABFGP method.

An example of a re-annotated gene model by ABFGP is given in Figure 3A. It illustrates the predicted gene model at the genomic locus that encodes a Major Facilitator Superfamily (MFS) transporter (Cf189922) in the Cladosporium fulvum genome. In this example ABFGP proposes two revisions compared to the originally annotated gene model. Introns (orange) and exons (red) with revised nucleotides are indicated in a separate track. Both revisions involve inclusion of novel exons that split up one intron into two smaller ones. The multiple protein sequence alignment around the second proposed revised site is shown for the unrevised (Figure 3B) and revised (Figure 3C) model. The improved continuity and quality of the sequence alignment suggest that the proposed revision is most likely correct. Moreover, TMHMM prediction performed on a 3-frame translation of the complete locus assigns two transmembrane helices in the revised exon, which is consistent with the secondary structure of the proteins encoded by the informant gene loci (data not shown). Finally, the additional exon is supported by a partial unigene aligned to the informant gene locus of Fusarium verticillioides (TC27075). A more detailed description of the ABFGP method is provided in Additional file 2.


Annotated Gene Structure
ABFGP predicted GeneModel
TMHMM
A
Predicted Introns
alignment similarity
Informant Unigene Similarities
F. verticillioides [TC37075-fragm] orf(1)
Informant Gene Similarities
B
C

```
RGANRSLMVISSILMTAGCGAMSVARPDNMSTLWGILVLAGLGIGGIVVPASIITTIICP 446 FOXG_13636
RGANRSLMVISSILMTAGCGAMSVARPDNMSTLWGILVLAGLGIGGIVVPASIITTIICP 442 FVEG_11059
KGANRTIMIVSSVCMTAGCGAMAIGRVDNMYQLWGILVLAGLGIGGIVVPASIITMIVCP 430 CRYPA_335360
RGHNKELLIVSSILMTAGCGALAIGRVDNLHQLWGILVLAGLGIGGIVVPASIISTIICP 391 CHEC5_112203
RGHNKELLIASSVLMTAGCGALAVGRVDNLYQLWGLLVLAGVGIGGIVVPASIISTIICP 515 SNOT_15949
KGRNKELMIVSSVFMTAGCGALACANRHNLSTLWALLIIAGLGIGGIVVPASIITTIICP 456 BC1T_07163
RGHNKELMIVSSVLMTAGCGSLAVATADNLRVLWGLLVLAGLGIGGIVVPASIITTIICP 453 DOTSE_163562
RGHNKE---------------------------------------RIVVPASIITTIICP 414 CFU_829297
RGHNKELMIVSSVLMTAGCGSLSVATVDNLSTIWGLLFLAGLGIGGIVVPASIITTIICP 391 MYCFI_26982
RGHNKELMIVSSILMTAGCGSLAVAERDNLHVLWGLLVLAGLGIGGIVVPASIITTIICP 399 MYCGR_033884
RGHNKELLIISSILMTAGCGAMSIARPDNLHQLWGLLTLAGLGIGGIVVPASIMTTIICP 433 AFL2T_08158
RGHNKELLIISSILMTAGCGAMSIARPDNLHQLWGLLTLAGLGIGGIVVPASIMTTIICP 447 AO090120000127
HGHNKELLIISSVLMTAGCGAMSIGRPDNLYQLWGILIIAGLGIGGIIVPASIITTIICP 440 ANIG_6862
RGHNKELLIASSVLMTAGCGAMAVADVDNMHQLWGILIVAGLGIGGIVVPASIITTIICP 445 ANID_11103T0
KGSNRELLVISSVFMTAGCGAMASANVDNMSTLWIILIIAGLGIGGIVVPASIITTIICP 429 CIMT_06077
RGGIRMLMVISSCMMTAGCGALAAADRNNINTVYGILIIAGLGIGGIVVPASIITTIICP 461 SS1T_05145
PGHIKWIVFLTCCIMTIGCGALASARLDNIHAMYAVLFIAGLGVGGIVVPASTIATIISP 458 NCU02023
KGRIRELMIFFTALMTAATGAMSIGRPDNINETLVVVSFACLGVGGILIPTSIIAQTIVP 442 PTRI_03120
 *  :                                       *::*:* ::  : *
```

```
RGANRSLMVISSILMTAGCGAMSVARPDNMSTLWGILVLAGLGIGGIVVPASIIT-TIICP 451 FOXG_13636
RGANRSLMVISSILMTAGCGAMSVARPDNMSTLWGILVLAGLGIGGIVVPASIIT-TIICP 447 FVEG_11059
KGANRTIMIVSSVCMTAGCGAMAIGRVDNMYQLWGILVLAGLGIGGIVVPASIIT-MIVCP 435 CRYPA_335360
RGHNKELMMASSVLMTAGCGSLAVATADNLHVLWGLLVLAGIGIGGIVVPASIITATIICP 462 CFU_829297.abfgp
RGHNKELMIVSSVLMTAGCGSLAVATADNLRVLWGLLVLAGLGIGGIVVPASIIT-TIICP 458 DOTSE_163562
RGHNKELMIVSSVLMTAGCGSLSVATVDNLSTIWGLLFLAGLGIGGIVVPASIIT-TIICP 396 MYCFI_26982
RGHNKELMIVSSILMTAGCGSLAVAERDNLHVLWGLLVLAGLGIGGIVVPASIIT-TIICP 404 MYCGR_033884
RGHNKELLIVSSILMTAGCGALAIGRVDNLHQLWGILVLAGLGIGGIVVPASIIS-TIICP 396 CHEC5_112203
RGHNKELLIASSVLMTAGCGALAVGRVDNLYQLWGLLVLAGVGIGGIVVPASIIS-TIICP 520 SNOT_15949
KGRNKELMIVSSVFMTAGCGALACANRHNLSTLWALLIIAGLGIGGIVVPASIIT-TIICP 461 BC1T_07163
RGHNKELLIISSILMTAGCGAMSIARPDNLHQLWGLLTLAGLGIGGIVVPASIMT-TIICP 438 AFL2T_08158
RGHNKELLIISSILMTAGCGAMSIARPDNLHQLWGLLTLAGLGIGGIVVPASIMT-TIICP 452 AO090120000127
HGHNKELLIISSVLMTAGCGAMSIGRPDNLYQLWGILIIAGLGIGGIIVPASIIT-TIICP 445 ANIG_6862
RGHNKELLIASSVLMTAGCGAMAVADVDNMHQLWGILIVAGLGIGGIVVPASIIT-TIICP 450 ANID_11103T0
KGSNRELLVISSVFMTAGCGAMASANVDNMSTLWIILIIAGLGIGGIVVPASIIT-TIICP 434 CIMT_06077
RGGIRMLMVISSCMMTAGCGALAAADRNNINTVYGILIIAGLGIGGIVVPASIIT-TIICP 466 SS1T_05145
PGHIKWIVFLTCCIMTIGCGALASARLDNIHAMYAVLFIAGLGVGGIVVPASTIA-TIISP 463 NCU02023
KGRIRELMIFFTALMTAATGAMSIGRPDNINETLVVVSFACLGVGGILIPTSIIA-QTIVP 447 PTRI_03120
 *  : ::.      ** . *::: .   .*:      :: .* :*:***::*:* ::   : *
```

Figure 3 ABFGP-based curation of the MFS transporter-encoding gene Cf189922 of Cladosporium fulvum.

A. Selected tracks of the GFF results obtained by applying ABFGP on the Cf189922 gene locus using 17 fungal informant genes. The annotated (blue) and the ABFGP-predicted gene model (green and grey) are shown on top. The grey part of the ABFGP prediction indicates an intron-exon boundary with status ‘doubtful’. Below are indicated the introns (orange) and exons (red) that were revised; the red box highlights the site of the second revision. The intron evidence track lists intron-exon boundaries obtained from informants; the colours used in the informant gene similarity track represent a measure for pairwise amino acid similarity. The alignment similarity track represents a summed representation of the inferred multiple sequence alignment of all informants.

B. Multiple protein sequence alignment of currently annotated gene models of Cf189922 and its informants. Sequence is restricted to the red box shown in panel A.

C. Multiple protein sequence alignment of the ABFGP-revised gene model of Cf189922 and its informants. Sequence is restricted to the red box shown in panel A. The proposed revision is highlighted in the black box.

Benchmarking of the ABFGP method

To benchmark the performance of the ABFGP method, we selected genes from ten different fungi, for which their intron-exon structure is confirmed by full-length unigenes. Of those, 6,965 genes have at least four reliable informant gene loci and passed all selection criteria (Additional file 3; an excel file with all gene identifiers is available at http://tinyurl.com/k9qft5o). Using this dataset, the ABFGP method achieves an overall gene sensitivity of 79.4% (Table 1; Additional file 4), meaning that on average 79 out of 100 gene models are predicted correctly without a single nucleotide error in their overall intron-exon structure.

Table 1 Benchmarking of the ABFGP performance on validated genes compared to GeneMark-ES

Species		**10 pooled species**[1]	***Magnaporthe oryzae***[2]		***Fusarium verticillioides***	
Method		**ABFGP**	**ABFGP**	**GeneMark-ES**	**ABFGP**	**GeneMark-ES**
# unigenes		6965	956	169	1154	327
Intron	Sn	91.16	91.5	89.3	92.2	90.7
	Pr	97.08	97.4	90.5	98.2	94.3
Exon	Sn	88.54	89.1	88	90.4	85.4
	Pr	98.91	99.4	89.1	98.9	87.9
Nucleotide	Sn	98.75	98.3	98.2	99.3	98.8
	Pr	99.08	99.3	97.1	99	97.1
Gene[3]	Sn	79.4 (5,533)	81.7 (781)	n.a.	82.1 (947)	n.a.

Sensitivity (Sn) and precision (Pr) of the gene model components (introns, exons, nucleotides) are expressed in percentages. Sn is calculated as true positives divided by: (true positives + false negatives); and Pr as true positives divide by: (true positives + false positives) [3].

[1] A list of all ten fungal species and results per species are provided in Additional file 4.

[2] Formerly named Magnaporthe grisea.

[3] The gene sensitivity is the percentage of gene models that is predicted without a single error. Total number of correctly predicted gene models is indicated in between brackets. Gene sensitivity was not provided for GeneMark-ES.

The ABFGP method applied to a set of available fulllength unigenes from Magnaporthe oryzae and Fusarium verticillioides was compared to GeneMark-ES [5], which was previously used on a smaller set of unigenes from these two fungi (Table 1). The ABFGP method performed better than GeneMark-ES on all gene components (exons, introns and nucleotides), in terms of sensitivity but most noticeably in terms of precision. The gene sensitivity achieved by ABFGP was 81.7% and 82.1% for the unigenes of *M. oryzae* and *F. verticilloides*, respectively. The results of this benchmarking show that the ABFGP method can confidently be applied to improve gene models in fungal genomes.

ABFGP as a tool to curate gene models of six annotated fungal genomes

To illustrate its versatility, we applied the ABFGP method on the gene catalogue of six different fungal species previously sequenced and annotated at the BROAD [2] and JGI institutes [1]: Botrytis cinerea, Cladosporium fulvum, Dothistroma septosporum, Mycosphaerella fijiensis, Verticillium dahliae and Zymoseptoria tritici (Table 2). ABFGP was performed after selection of eligible genes (BDBH category) based on informant genes retrieved from a set of 29 fungal genomes (Additional file 1). A second, much smaller set of genes was compiled from informants that suggested species-specific variation or gene models with errors (GME). Between 7,285 and 8,504 annotated gene loci per species were eligible for ABFGP using these criteria. For 0.4-2.0% of these, ABFGP was aborted during execution because the number of representative informants

Table 2 Gene models in six fungal species re-annotated by the ABFGP method

Species	*Botrytis cinerea*		*Cladosporium fulvum*		*Dothistroma septosporum*	
Sequence technology	Sanger		454		Illumina/454/ Sanger	
Fold genome coverage	4.5		21		34	
# Annotated genes	16,448		14,127		12,580	
Annotation pipeline[2]	BROAD		GeneMark-ES		JGI	
Annotation year[3]	2005		2009		2010	
Reference	[21]		[20]		[20]	
Total eligible gene models	8,503		7,574		8,090	
Bi Directional Best Hit	7,165		6,990		7,511	
Gene Model Error	1,338		584		579	
Confirmed/unchanged[4]	4,832	57%	5,823	77%	6,249	77%
Revised[4]	3,505	41%	1,724	23%	1,770	22%
Bi Directional Best Hit[5]	2,481	35%	1,304	19%	1,404	19%
Gene Model Error[5]	1,024	77%	420	72%	366	63%
Aborted[4]	166	2,0%	27	0,4%	71	0,9%

Species	*Mycosphaerella fijiensis*		*Verticillium dahliae*		*Zymoseptoria tritici* [1]	
Sequence technology	Sanger		Sanger		Sanger	
Fold genome coverage	7.1		7.5		8.9	
# Annotated genes	10,313		10,535		10,952	
Annotation pipeline[2]	JGI		BROAD		JGI	
Annotation year[3]	≤2008		2008		2008	
Reference	n.a.		[22]		[23]	
Total eligible gene models	7,283		8,362		7,893	
Bi Directional Best Hit	6,773		7,814		7,317	
Gene Model Error	510		548		576	
Confirmed/unchanged[4]	4,775	66%	5,390	64%	5,262	67%
Revised[4]	2,456	34%	2,870	34%	2,553	32%
Bi Directional Best Hit[5]	2,064	30%	2,511	32%	2,137	29%
Gene Model Error[5]	392	77%	359	66%	416	72%
Aborted[4]	52	0,7%	102	1,2%	78	1,0%

[1] Formerly named Mycosphaerella graminicola.

[2] Sequencing centre which sequenced and annotated this genome (BROAD institute or Joint Genome Institute); *C. fulvum* was sequenced at Wageningen University and annotated using GeneMark-ES version 2.2 [20].

[3] Estimated year the gene calling was performed.

[4] Number and percentage of all gene models in this category.

[5] Number and percentage of revised gene models in this category.

dropped below four during the integrated quality assessment. For the remaining loci, the gene models predicted by ABFGP were compared to their current annotations. As expected, the currently available annotation and the ABFGP-predicted structures of a large fraction of the gene models (on average 68%) were identical (Table 2). Predicting the same intron-exon-structure by two independent methods is a strong indication that the predicted gene models are correct. The ABFGP method proposed at least a minor revision for 22% to 41% of the assessed genes in a given species. Among those, the GME category of genes is highly overrepresented, with 62% to 75% of them being revised versus only 19% to 34% from the BDBH category. The lowest number of revisions is proposed for the most recently annotated genomes of the fungi *C. fulvum* and *D. septosporum* [20], and the highest number for B. cinerea. This is likely due to the fact that the genome assembly of B. cinerea has a low sequence coverage produced by Sanger technology, and its annotation was performed by older, less accurate gene predictor software (Table 2).

Reliability of the ABFGP-predicted gene models

The ABFGP method confirmed 57 to 77% of the previously reported annotated gene models and proposed revisions for the remaining in six fungal species (Table 2). The overall quality of the revised predictions is supported by high accuracy as shown by the benchmarking on unigenes (Table 1). To address the reliability at the level of individual genes, the ABFGP method was equipped with an *a posteriori* introspection module. Each intron and exon in the multiple gene model alignment was evaluated on a series of stringent criteria (e.g. alignment quality, length variance, splice site score, etc.) and was labelled to indicate the likelihood of its correctness: gene models were labelled as ‘ok’ only if all individual introns and exons received this label, and were labelled as ‘doubtful’ in case one or more introns or exons received this label (Table 3). Of the confirmed gene models on average 86.4% was labelled ‘ok’ and 13.6% ‘doubtful’, whereas of the revised gene models 66.1% was labelled ‘ok’ and 33.9% ‘doubtful’. The introspection procedure was also applied to the benchmark set of 6,965 genes supported by unigenes and resulted in 5,016 true positives (72.2%), 496 true negatives (7.1%), 533 false negatives (7.7%) and 899 false positives (12.9%). This analysis shows that the introspection procedure is quite accurate, and that the majority of ABFGP-revised models of the re-annotated genomes is reliable.

Types of revisions proposed by the ABFGP method

The most conspicuous differences between the annotated and ABFGP-predicted gene models are summarized in Table 4. Major revisions proposed by the ABFGP method comprise corrections of falsely fused and split gene models in current annotations. *B. cinerea* appears enriched for both incorrectly merged and split genes and *C. fulvum* for incorrectly merged genes. Up to 19% of the revisions proposed by the ABFGP method are due to SEs and/or DMs, which were particularly often encountered in genes of *B. cinerea*, *C. fulvum* and *V. dahliae*. Other revisions involve boundary changes, removal and addition of exons and introns in predicted gene models. Additional exons are more rarely predicted, but they are frequently occurring as internal revisions (as shown in Figure 3 for *C. fulvum*) of genes in *B. cinerea* and *V. dahliae*. ABFGP frequently removed stopless 3n introns in the gene models of *M. fijiensis* and *Z. tritici*. The proposed revisions resulted mainly in a decrease of the average intron length: -42, -35, -30, -29, -6 and +1 nucleotides for *M. fijiensis, B. cinerea*, *C. fulvum, Z. tritici, V. dahliae* and *D. septosporum*, respectively.

Table 3 Introspection of results obtained by the ABFGP method

Species	*Botrytis cinerea*		*Cladosporium fulvum*		*Dothistroma septosporum*		Pooled unigenes[2]
Total number of assessed genes[3]	8,337		7,547		8,019		6,965
Confirmed/unchanged	4,832		5,823		6,249		Correct
Labeled 'ok'[4]	3,942	82%	5,186	89%	5,505	88%	5,015 (TP)
Labeled 'doubtful'[4]	890	16%	637	11%	744	12%	533 (FN)
Revised	3,505		1,724		1,770		Incorrect
Labeled 'ok'[4]	2,137	61%	1,160	67%	1,209	68%	899 (FP)
Labeled 'doubtful'[4]	1,368	29%	564	33%	561	32%	496 (TN)

Species	*Mycosphaerella fijiensis*		*Verticillium dahliae*		*Zymoseptoria tritici*[1]	
Total number of assessed genes[3]	7,231		8,260		7,815	
Confirmed/unchanged	4,775		5,390		5,262	
Labeled 'ok'[4]	4,216	88%	4,536	84%	4,539	84%
Labeled 'doubtful'[4]	559	12%	854	16%	723	16%
Revised	2,456		2,870		2,553	
Labeled 'ok'[4]	1,730	70%	1,864	65%	1,734	68%
Labeled 'doubtful'[4]	726	30%	1,006	35%	819	32%

[1] Formerly named Mycosphaerella graminicola.

[2] Correctly predicted gene models (benchmarked on the full-length unigenes) that were labelled by the introspection procedure as 'ok' are true positives (TP) and labelled 'doubtful' are false negatives (FN). Genes that were incorrectly predicted and were labelled 'ok' are false positives (FP) and labelled 'doubtful' are true negatives (TN). [3] Total eligible number of genes minus number of genes aborted during execution (Table 2).

[4] Number and percentage of genes that are labelled 'ok' and 'doubtful' by the introspection procedure in each category.

Increasing the number of informants improves performance of the ABFGP method

ABFGP performance decreased when using fewer informants or when closely related informants are not available (data not shown). For the curation of a particular gene model, the most closely related fungal species failed to provide informants for 7 to 19% of selected loci (Additional File 5). Conversely, fungal species that provided the lowest number of informants still contributed 16 to 38% of informant loci. In addition, in some cases, fungal species that provided most of the informant loci are not always the closest relatives. For example, *M. fijiensis*, the closest relative of *Z. tritici*, is not among the top three species that provided the highest number of informants (Additional File 5). Similarly, *N. haematococca* and *M. oryzae* provide more informants than *V. albo-atrum* for the curation of *V. dahliae*. For *C. fulvum* and *M. fijiensis*, it is striking that fungi that belong to a different taxonomic class are in the top three species that provided the highest number of informants. Our results show that the six studied fungal gene catalogues differ in quality. Because all informant catalogues were predicted by the same genome sequence centres (see Additional file 1), similar error rates are expected to occur in their gene models. An unexpected low contributor to the pool of informants could be explained by a slightly higher error rate in its gene catalogue. In addition, many genes show a discontinuous distribution in the fungal tree of life [8],[10]. This underlines the importance of selecting informants from a wide phylogenetic spectrum of species rather than from a small set of closely related species.

Table 4 Types of revisions in annotated gene models made by the ABFGP method

Species	*Botrytis cinerea*		*Cladosporium fulvum*		*Dothistroma septosporum*	
Total revised genes[2]	3,473		1,721		1,761	
Genes containing SE and/or DMs[3]	353		333		176	
Genes split by ABFGP	195		183		62	
Genes merged by ABFGP	102		12		16	
Total annotated exons	12967		5675		5211	
Unrevised	5970		2372		2078	
Boundary revision[4]	4851		2274		2341	
5’ or 3’ removed (−) / added (+)[5,6]	−783	617	−451	252	−265	224
Internal removed (−) / added (+)[5,7]	−51	616	−20	98	−24	35
Total annotated introns	9459		3947		3438	
Unrevised	4907		2019		1740	
Boundary revision[8]	1799		692		727	
Stopless 3n removed (−) / added (+)[9]	−447	189	−166	146	−365	146

Species	*Mycosphaerella fijiensis*		*Verticillium dahliae*		*Zymoseptoria tritici*[1]	
Total revised genes[2]	2,448		2,865		2,552	
Genes containing SE and/or DMs[3]	127		515		66	
Genes split by ABFGP	91		94		130	
Genes merged by ABFGP	27		19		28	
Total annotated exons	7525		10709		8316	
Unrevised	2593		4956		3116	
Boundary revision[4]	3230		4355		3357	
5’ or 3’ removed (−) / added (+)[5,6]	−297	341	−529	333	−415	335
Internal removed (−) / added (+)[5,7]	−59	74	−66	346	−76	75
Total annotated introns	5058		7838		5753	
Unrevised	2276		4048		2836	
Boundary revision[8]	889		1738		839	
Stopless 3n removed (−) / added (+)[9]	−1032	99	−331	244	−953	130

[1] Formerly named *Mycosphaerella graminicola*.

[2] The total number of revisions can exceed the total number of revised genes because a gene model can contain more than one revision. [3] Genes for which the revision(s) include sequence errors or mutations.

[4] Exons with a different start and/or end coordinate when comparing both gene models.

[5] Exons incorporated in only one of both gene models (not in the ABFGP model/only in the ABFGP model). [6] Omitted and additional exons in recognized false gene splits and fusions were not counted.

[7] (Large) intron in one gene model, split into two smaller introns with intermediary (small) exon in the other gene model. [8] Introns with a different donor and/or acceptor site when comparing both gene models.

[9] Stopless 3n introns incorporated in only one of both gene models (not in the ABFGP model/only in the ABFGP model).

Discussion

The ABFGP method accurately predicts intron-exon structures of protein-encoding genes in fungi

The ABFGP method can accurately re-annotate the intron-exon structure in a gene-by-gene fashion when a gene locus is provided with sufficient informants. GeneMark-ES was chosen as a state of the art ab initio gene predictor, and we have shown that the ABFGP method improves the quality of the gene models. This is explained by a higher precision (Table 1), which means that a lower number of false positives are reported by ABFGP. Indeed, in general, evidence- or alignment-based methods are less prone to wrongly assign additional exons [3], because they are only predicted when supported by informants. Predicting introns in compact genomes with numerous small introns is challenging [5], yet ABFGP achieves both a high sensitivity (91.2%) and precision (97.3%) (Table 1). This is achieved by exploiting abundantly occurring intron presence-absence patterns [19]. SEs and/or DMs can be confidentially recognized as discontinuities when compared with exonic sequences of informant genes. Finally, lack of synteny in distantly related fungi facilitates recognition of false gene fusions, which is a frequently observed error made by ab initio gene predictors [5,16]. Adjacent genes with the same orientation are prone to be falsely fused to the target gene, but this is minimized in the ABFGP method because of the shuffled gene order in informant genomes. Whole-genome alignment-based gene prediction benchmarked on a test set of 1,483 genes from two strains of *C. neoformans* achieved 88% and 89% exon sensitivity and precision, respectively, resulting in an overall gene sensitivity of ~60% [6], which is low considering the high conservation between the two genomes. This shows that the gene-by-gene approach by the ABFGP method is more powerful, even by making use of informant genes from evolutionary distant fungal species. The benchmark test showed uniform performance on unigenes from ten selected species (Additional file 4). Yet, this performance was, in case of *D. septosporum*, achieved with generic PSSMs that were not derived from its own splice sites. Species-specific parameterization of gene properties was indicated as crucial for the performance of ab initio supervised [4], unsupervised [5] as well as the alignment-based gene prediction methods [6]. We speculate that in the ABFGP method, the number of informants compensates for the absence of species-specific parameterization.

ABFGP as a genome-wide annotation assessment tool

Between 7,205 and 8,270 gene models of six fungal genomes were automatically assessed by the ABFGP method. Between 1,724 and 3,505 (on average 2,480) of these gene models were proposed to be incorrect and needed revision. A more stringent indication of correct revisions is obtained by counting only those revised gene models that were labelled 'ok' (Table 2), corrected for the observed error rate of the ABFGP method (based on 79% gene sensitivity). This yields an estimated revision of between 1,362 and 2,769 gene models for each fungal species. These numbers are in the same range as those obtained in a recent genome-wide re-annotation effort of the *F. graminearum* genome, which was based on predictions by a suite of gene predictors, using expression data and followed by extensive manual curation [16]. In that case, 1,770 gene models were revised, 691 new gene models were added and 286 gene models were removed. Yet, a recent study using RNA-Seq data revised another 655 gene models [17], showing that the quality-improving manual curation effort was not yet exhaustive. Their analysis [16] and ours independently show that thousands of genes are still wrongly annotated in gene catalogues of many published fungal genomes. Interestingly, the same types of revision were reported (false gene splits and fusions, novel introns and a decrease in average intron length) as those proposed by the ABFGP method.

Types of revision are often related to the annotation pipelines used (Table 2). For example, inclusion of new exons represents a rare class of revisions, except in the two genomes that were annotated at the BROAD institute. In contrast, prediction of too many stopless 3n introns was observed in the genomes of M. fijiensis and Z. tritici that were sequenced at the JGI. The lowest number of revised gene models was proposed for *C. fulvum* and *D. septosporum*, which represent the most recently sequenced and independently annotated genomes [20]. We speculate that this might reflect the steady increase in accuracy of ab initio gene prediction software. In this study six different fungi from three distinct phylogenetic classes were re-annotated, using informants from five classes of Ascomycota and two unrelated Basidiomycota. This shows that the ABFGP method is species-independent and can be applied to a wide variety of fungal genomes.

Genome-wide re-annotation by the ABFGP method did not capture the complete gene catalogues (Table 2) which is mainly due to the stringent criteria that were chosen to obtain the most likely orthologous informant genes (see Methods). This effect is most obvious for informant genes obtained from poorly annotated genomes. Performance for those genes can be improved, besides lowering this threshold, by expanding beyond using annotated genes only. An informant locus can be any genomic region that has ample sequence similarity to the target protein or locus. TBLASTN or TBLASTX could be used to detect loci that failed to be recognized and annotated as protein-coding genes or were poorly annotated (see Figure 1). Loci that are obtained directly from a (non-annotated) genomic sequence could be used as an additional resource for informants that would simultaneously increase the number of eligible target genes and prediction performance of ABFGP. The reverse strategy could also be employed by using the ABFGP method to generate de novo gene models in the target genome that lack predicted gene models but have significant sequence similarity to predicted proteins in other species. However, a general limitation of de novo evidence-based gene prediction, including the ABGFP method, is that annotation of species-specific or fast evolving genes is not possible by any prediction method. The ABFGP method follows an alternative approach to the various other ensemble predictors, because it derives its evidence directly from genomic informant sequences. Moreover, it proposes revised gene models that include SEs and/or DMs. This makes the ABFGP method complementary to other ensemble predictors, because these occur frequently in the gene catalogues of these fungal genomes [24].

Sequence errors and disruptive mutations in fungal genes

Presumed inconsistent gene models were revised in 70 to 83% of all cases (Table 2), of which on average 55% were labelled by the introspection procedure as ‘ok’ for all introns and exons. Among these revisions was an unexpected high number of gene models containing SEs and/or DMs. Because ab initio gene prediction software does not allow in-frame stops or frame-shifts causing indels, (pseudo)genic regions with strong coding signals will often be predicted to be truncated or split gene model(s). Of the six studied fungi, most revisions were proposed for B. cinerea, likely because its Sanger sequenced genome assembly is supported by 4.5× coverage only [21], and its annotation was performed several years ago. Recently, resequencing of B. cinerea using Illumina, supplemented with some additional small Sanger reads, resulted in a new assembly with 50× coverage [25]. This new sequence not only revealed 31,275 SEs (personal communication Dr. Martijn Staats), but also a considerable number of assembly errors in the original reference sequence, of which many were located in coding regions that contained annotated, yet apparently fragmented genes (personal communication Dr. Jan van Kan). This could be an explanation for the higher frequency (2.0% versus 0.4-1.2% for the other five fungi, Table 2) of abandoned executions by the ABFGP method. However, a considerable fraction of inconsistencies observed in coding regions were confirmed by resequencing, indicating that they were not SEs but true DMs. Additional studies on DMs in these six fungal species suggest that pseudogenization is very common in fungi [24]. Our results show that many fungal gene catalogues still contain numerous unidentified truncated and erroneous gene models due to SEs and/or DMs, that are readily detected by the ABFGP method.

Introspection of proposed gene model revisions

The introspection module for assessing gene model correctness is a useful extension of the ABFGP method as it helps to prioritize gene models that still need manual curation. For the six fungal genomes, between 3,942 and 5,505 genes were suggested to not require additional manual curation (Table 3). Based on the benchmarked performance of the introspection procedure using the unigene dataset, the error rate of genes incorrectly labelled as 'ok' is estimated to be 12.9%. This accounts for only 500 to 700 models out of 4,000 to 5,500 that contain errors. For gene models that were recognized as 'doubtful', the ABFGP method provides a GFF-track that shows the doubtful parts of the predicted gene model that require manual curation. However, the introspection module still needs further improvement because 20.6% of the gene models is incorrectly labelled: 12.9% is labelled as 'ok' but do contain (small) errors and 7.7% is labelled as 'doubtful' whereas the gene models are correct. Lowering the number of false positives can possibly be achieved by including ab initio gene model prediction in the ABFGP method, which would allow better detection of species-specific variation of genic regions. This would further increase the efficiency of the ABFGP method as an automated and accurate method for gene model curation.

Conclusions

Availability of an accurate gene catalogue of an organism is a prerequisite and starting point for functional analyses of its genes. Obtaining such a catalogue with minimal manual input is still a major challenge. The ABFGP method is a useful tool to integrate into existing gene annotation pipelines because it can assess and improve gene models with great accuracy in a fully automated manner. The concept of gene-by-gene alignment-based gene prediction exploits the availability of dozens of sequenced fungal genomes, which is particularly useful for annotating novel genomes of these plastic organisms. The possibility of the ABFGP introspection procedure at the gene and intron-exon level helps to decrease the number of gene models that still require manual curation. Because fungal genome sequencing is undertaken at an accelerating pace [1], both quality and number of informant gene loci are expected to increase in the coming years, which will disclose more target gene loci in genomes and also increase the efficiency and reliability of the ABFGP method.

Methods

Sequences, annotations and third party software used

Genomes, proteomes and annotations of 29 fungal species were downloaded from the Fungal Genome Initiative of the BROAD Institute [2] and the Fungal Genomics Program of the Joint Genome Institute (JGI) [1] (Additional file 1). Available unigenes from ten fungal species were downloaded from the JGI and The Gene Index Project (http://compbio.dfci.harvard.edu/tgi/). The ABFGP method uses several third party applications: BLAST 2.2.8, ClustalW 2.0.12, HMMER 2.3.2, SignalP 3.0, TMHMM 2.0, transeq, getorf and tcode from EMBOSS 6.2.0.

Full-length unigenes

Datasets of assembled unigenes (Additional file 1) were aligned to their genomes using GeneSeqer (October 2005) and for each unigene the obtained intron-exon structure of its coding sequence was compared to its annotated gene model. For benchmarking the ABFGP method only those unigenes that were full-length were selected.

Informant selection

An all-versus-all similarity matrix was created between all proteins from the 29 predicted proteomes using BLASTP. From this matrix, informant proteins from different fungi were selected for each target protein by applying the following criteria: the protein must represent (i) the bi-directional best hit (BDBH) in the informant's proteome, (ii) the alignment must span at least 70% of the length of both target and informant protein, (iii) the relative difference in length between target and informant protein must be below 50% (calculated from ii) and (iv) the alignment's bitscore between target and informant protein must be at least 10% of the bitscore of the proteins when compared to themselves. As a final criterion, at least four informant proteins must be available for a target protein, and the total number of informants was limited to the 19 most similar informants (based on bitscore). This dataset of genes eligible for ABFGP is referred to as BDBH. A second category was created by lowering the requirement of length coverage to 25% and increasing length difference to 300%, followed by filtering for target proteins that were linked to either consistently longer or shorter informant proteins. Consistent protein length variation putatively indicates species-specific variation or that the corresponding gene model contains major errors (this dataset is referred to as GME). For both categories, target and informant proteins were loaded into ABFGP as DNA sequence of their genomic locus flanked by an additional 1.5 kb of sequence on both sides of the gene's start and stop codon. Unigenes aligned to these gene loci were taken along as additional informants. In the benchmark that uses unigenes, informants were selected only by the BDBH approach and full-length unigene data aligned to the target gene locus were discarded; the parameters `–abinitio` and `–benchmark` were used to discard the unigene of the target locus and annotated gene models as hints. In all benchmark analyses, sensitivity and precision are calculated according as described by Picardi and Pesole [3], in which specificity is an alias for precision.

Position Specific Scoring Matrices of genic elements

Definitions of donor site, acceptor site, branch point and polypyrimidine tracks were chosen according to [18]. Generic fungal PSSMs (Additional file 2) for the canonical donor (n = 571,185), the non-canonical GC donor (n = 2,428) and the canonical acceptor (n = 576,021) were derived from all splice sites without any nonambiguous nucleotide in 25 annotated genomes (excluding the annotations of Cladosporium fulvum, Coccidioides posadasii, Dothistroma septosporum, Nectria haematococca and Trichoderma atroviride, which were added as target and/or informant species in a later stage of the analyses).

Access to the method and data

A technical explanation of the ABFGP method, and its GFF visualization is provided in Additional file 2. The source code of the ABFGP method is available (see Availability and requirements). Other datasets are available upon request by the corresponding authors: the complete list of unigene identifiers used for the benchmark analyses (.xls), the predicted gene models from the benchmark that uses unigenes (GFF files) and the genome-wide re-annotation of the six fungi (fasta and simplified GFF files).

Availability and requirements

Project name: ABFGP

Project home page: https://github.com/atevanderburgt/ABFGP

Operating system: Linux, Unix Programming language: Python

Other requirements: Python 2.6 or higher Licence: GNU GPL

Any restrictions to use by non-academics: None

Additional files

Additional file 1: Fungal genomes used for alignment-based fungal gene prediction. Fungal genomes and their phylogeny used in this study.

Additional file 2: Explanation of the ABFGP method. In-depth explanation of the ABFGP method.

Additional file 3: Determination of a dataset from ten fungi for benchmarking the ABFGP method. Determination of a dataset of 6,965 experimentally validated genes models from ten fungal genomes for benchmarking the performance of the ABFGP method.

Additional file 4: Benchmarking results of ABFGP performance. Benchmarking results of ABFGP performance on 6,965 experimentally validated gene models from ten fungal species.

Additional file 5: Rank of species providing informant gene loci used for the six re-annotated gene catalogues. Top three and bottom two species that provided the highest number of informants for the re-annotation of the gene catalogues of six fungal species.

Acknowledgements

Ate van der Burgt, Jerome Collemare and Pierre JGM de Wit were funded in part by the laboratory of Phytopathology of Wageningen University and a grant from The Royal Netherlands Academy of Arts and Sciences. Edouard Severing and Ate van der Burgt were also funded in part by the Applied Bioinformatics Group of Wageningen University. The authors would like to thank Dr. Jan A.L. van Kan and Dr. Martijn Staats for sharing unpublished data on the re-sequenced genome of *Botrytis cinerea*.

References

1. Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky S, Nikitin R, Ohm RA, et al: The genome portal of the Department of Energy Joint Genome Institute. Nucleic Acids Res 2012, 40(Database issue):D26–D32.
2. Cuomo CA, Birren BW: The fungal genome initiative and lessons learned from genome sequencing. Methods Enzymol 2010, 470:833–855.
3. Picardi E, Pesole G: Computational methods for ab initio and comparative gene finding. Methods Mol Biol 2010, 609:269–284.
4. Stanke M, Tzvetkova A, Morgenstern B: AUGUSTUS at EGASP: using EST, protein and genomic alignments for improved gene prediction in the human genome. Genome Biol 2006, 7(1):S11. 11–18.
5. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M: Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res 2008, 18(12):1979–1990.
6. Tenney AE, Brown RH, Vaske C, Lodge JK, Doering TL, Brent MR: Gene prediction and verification in a compact genome with numerous small introns. Genome Res 2004, 14(11):2330–2335.
7. Stanke M, Schoffmann O, Morgenstern B, Waack S: Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. BMC Bioinformatics 2006, 7:62.
8. Ohm RA, Feau N, Henrissat B, Schoch CL, Horwitz BA, Barry KW, Condon BJ, Copeland AC, Dhillon B, Glaser F, et al: Diverse lifestyles and strategies of plant pathogenesis encoded in the genomes of eighteen Dothideomycetes fungi. PLoS Pathog 2012, 8(12):e1003037.
9. Oliver R: Genomic tillage and the harvest of fungal phytopathogens. New Phytol 2012, 196(4):1015–1023.
10. Raffaele S, Kamoun S: Genome evolution in filamentous plant pathogens: why bigger can be better. Nat Rev Microbiol 2012, 10(6):417–430.
11. Liu Q, Mackey AJ, Roos DS, Pereira FC: Evigan: a hidden variable model for integrating gene evidence for eukaryotic gene prediction. Bioinformatics 2008, 24(5):597–605.
12. Bernal A, Crammer K, Pereira F: Automated gene-model curation using global discriminative learning. Bioinformatics 2012, 28(12):1571–1578.
13. Liu Q, Crammer K, Pereira FC, Roos DS: Reranking candidate gene models with cross-species comparison for improved gene prediction. BMC Bioinformatics 2008, 9:433.
14. Birney E, Clamp M, Durbin R: GeneWise and Genomewise. Genome Res 2004, 14(5):988–995.
15. Keller O, Odronitz F, Stanke M, Kollmar M, Waack S: Scipio: using protein sequences to determine the precise exon/intron structures of genes and their orthologs in closely related species. BMC Bioinformatics 2008, 9:278.
16. Wong P, Walter M, Lee W, Mannhaupt G, Munsterkotter M, Mewes HW, Adam G, Guldener U: FGDB: revisiting the genome annotation of the plant pathogen Fusarium graminearum. Nucleic Acids Res 2011, 39(Database issue):D637–D639.
17. Zhao C, Waalwijk C, de Wit PJ, Tang D, van der Lee T: RNA-Seq analysis reveals new gene models and alternative splicing in the fungal pathogen Fusarium graminearum. BMC Genomics 2013, 14(1):21.
18. Kupfer DM, Drabenstot SD, Buchanan KL, Lai H, Zhu H, Dyer DW, Roe BA, Murphy JW: Introns and splicing elements of five diverse fungi. Eukaryot Cell 2004, 3(5):1088–1100.
19. Nielsen CB, Friedman B, Birren B, Burge CB, Galagan JE: Patterns of intron gain and loss in fungi. PLoS Biology 2004, 2(12):e422.
20. de Wit PJ, van der Burgt A, Okmen B, Stergiopoulos I, Abd-Elsalam KA, Aerts AL, Bahkali AH, Beenen HG, Chettri P, Cox MP, et al: The genomes of the fungal plant pathogens Cladosporium fulvum and Dothistroma septosporum reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. PLoS Genetics 2012, 8(11):e1003088.
21. Amselem J, Cuomo CA, van Kan JA, Viaud M, Benito EP, Couloux A, Coutinho PM, de Vries RP, Dyer PS, Fillinger S, et al: Genomic analysis of the necrotrophic fungal pathogens Sclerotinia sclerotiorum and Botrytis cinerea. PLoS Genetics 2011, 7(8):e1002230.
22. Klosterman SJ, Subbarao KV, Kang S, Veronese P, Gold SE, Thomma BP, Chen Z, Henrissat B, Lee YH, Park J, et al: Comparative genomics yields insights into niche adaptation of plant vascular wilt pathogens. PLoS Pathogens 2011, 7(7):e1002137.

23. Goodwin SB, M'Barek SB, Dhillon B, Wittenberg AH, Crane CF, Hane JK, Foster AJ, Van der Lee TA, Grimwood J, Aerts A, et al: Finished genome of the fungal wheat pathogen Mycosphaerella graminicola reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genetics 2011, 7(6):e1002070.
24. van der Burgt A, Karimi M, Bahkali AH, de Wit PJ: Pseudogenization in pathogenic fungi with different host plants and lifestyles might reflect their evolutionary past. Mol Plant Pathol. in press 2013, 15:133–144.
25. Staats M, van Kan JA: Genome update of Botrytis cinerea strains B05.10 and T4. Eukaryot Cell 2012, 11(11):1413–1414.

Chapter 4

Pseudogenization in pathogenic fungi with different host plants and lifestyles might reflect their evolutionary past.

van der Burgt A, Karimi Jashni M, Severing E, Collemare J, de Wit PJGM.

Molecular Plant Pathology 2014, 15(2):133–144.

doi: 10.1111/mpp.12072

Supporting Information is available from:
http://onlinelibrary.wiley.com/doi/10.1111/mpp.12072/suppinfo

Summary

Pseudogenes are genes with significant homology to functional genes but contain disruptive mutations (DMs) leading to production of non- or partially functional proteins. Little is known about pseudogenization in fungi. Here we report on identification of DMs causing pseudogenes in the genomes of the fungal plant pathogens *Botrytis cinerea*, *Cladosporium fulvum, Dothistroma septosporum, Mycosphaerella fijiensis, Verticillium dahliae* and *Zymoseptoria tritici*. In these fungi we have identified 1740 gene models containing 2795 DMs obtained by an alignment-based gene prediction method. The contribution of sequencing errors to DMs was minimized by analyses of resequenced genomes to obtain a refined data set of 924 gene models containing 1666 true DMs. The frequency of pseudogenes varied from 1 to 5% in the gene catalogues of these fungi, being the highest in the asexually reproducing fungi *C. fulvum* (4.9%), followed by *D. septosporum* (2.4%) and *V. dahliae* (2.1%). The majority of pseudogenes do not represent recent gene duplications, but members of multi-gene families and unitary genes. In general there was no bias for pseudogenization of specific genes in the six fungi. Single exception are those encoding secreted proteins including proteases which appeared more pseudogenized in *C. fulvum* than in *D. septosporum*. Most pseudogenes present in these two phylogenically closely related fungi are not shared suggesting that they are related to adaptation to a different host (tomato versus pine) and lifestyle (biotroph versus hemibiotroph).

Introduction

Pseudogenes show homology to functional genes, but contain disruptive mutations (DMs) leading to non- or partially functional proteins [1]. A pseudogenization event caused by a single DM can result in a premature stop codon, frameshift, defective splice junction or distortion of regulatory sequences required for transcription [1], [2]. Similarly, a transposon insertion dramatically alters gene continuity, but also represents a single DM event leading to pseudogenization. Most eukaryotic pseudogenes are disabled copies of duplicated parental genes [3] and their majority will eventually disappear, while some will evolve new functions and might become fixed in an organism [4]. Unitary pseudogenes are single copy genes that may become non-functional through loss-of-function (LOF) variation caused by various types of mutations [5],[2]. A residual biological function might develop for genes encoding multi-domain proteins that have lost only one or a few of their functional domains. However, when a lost domain in a unitary pseudogene is essential and is not compensated for by another protein, the LOF variant will affect the performance of an organism [2]. LOF variants and unitary pseudogenes have been reported to cause several inheritable human diseases [2]. However, in some cases an organism might also profit from pseudogenization as for pathogens and commensals that need to adapt and co-evolve with their hosts.

When only few DMs are present, pseudogenes still bear all hallmarks of a protein-encoding gene and *ab initio* gene prediction software will likely predict gene models at these loci, also in the case of absence of splice sites or presence of a premature stop. Therefore, DMs will often cause erroneous gene model predictions. This is also true for sequence errors (SEs) in genomic sequences that introduce in-frame stops by erroneous base calling or distortion of reading frames by insertions or deletions (indels). Thus, SEs can cause incorrect assignment of DMs and incorrect assignment of pseudogenes that contain them [5].

The extent of pseudogenization in (plant pathogenic) fungi has not been studied on a whole genome scale yet, although numerous reports describe individual genes that were subjected to pseudogenization. Selection pressure imposed on plant pathogenic fungi by plant disease resistance genes has led to rapid development of pseudogenes of which the parental genes encode effectors that are recognized by matching resistance gene-encoded receptor-like proteins [6],[7]. Repeat-induced point mutations (RIP) can cause pseudogenization by introducing premature stop codons by C to T and G to A transitions. RIP occurs in sexually active fungi mainly belonging to the *Ascomycetes,* where it was first discovered in *Neurospora crassa* [8]. Genes directly adjacent to repeats are at risk for being pseudogenized when RIP activity slightly protrudes the repeat locus boundaries. This has been shown in the oil seed rape pathogen *Leptosphaeria maculans* where pseudogenization of the *AvrLm1* effector gene is caused by RIP [9].

Here we report on identification of DMs causing pseudogenes in the fungal plant pathogens Botrytis cinerea, Cladosporium fulvum, Dothistroma septosporum, Mycosphaerella fijiensis, Verticillium dahlia and Zymoseptoria tritici. From these six fungi we have identified many DMs obtained by an alignment-based fungal gene prediction method [10]. The frequency of pseudogenes was highest in in the gene catalogues of the phylogenetically related *C. fulvum* (4.9%) and *D. septosporum* (2.4%). There was no clear bias for pseudogenization of specific genes in these two fungi except for those encoding secreted proteins including proteases and genes involved in production of secondary metabolites like dothistromin. The biotrophic tomato pathogen *C. fulvum* shares many genes with the hemi-biotrophic pine pathogen *D. septosporum*, but the gene set affected by pseudogenization in the two fungi is not shared. A possible role of pseudogenization and eventually gene loss in adaptation to a different hosts and lifestyle is discussed.

Results

The genomes of *C. fulvum* and *D. septosporum* have recently been released [11]. The Alignment-Based Fungal Gene Prediction (ABFGP) method [10] was applied to six fungal genomes in order to identify disruptive mutations (DMs) that would cause pseudogenization. Gene models predicted by ABFGP represent exons which are chained by both introns and DMs. The ABFGP method recognized DMs in genes which resulted in frame shifts (non-3n indels) or would lead to an in-frame stop codon when compared with homologous informant genes from several different fungi lacking the DMs. In multiple protein sequence alignments, the DMs are recognized as extension of conservation (i) throughout annotated introns, (ii) upstream of annotated start codons or (iii) downstream of annotated stop codons (Figure 1). In all cases, high sequence similarity is shared with corresponding exonic parts of informant genes. Predicted DMs coincided predominantly with incorrectly predicted introns (1a,1c), truncated predicted proteins (1a,1b) and rarely in a single gene splitted into two gene models (1c).

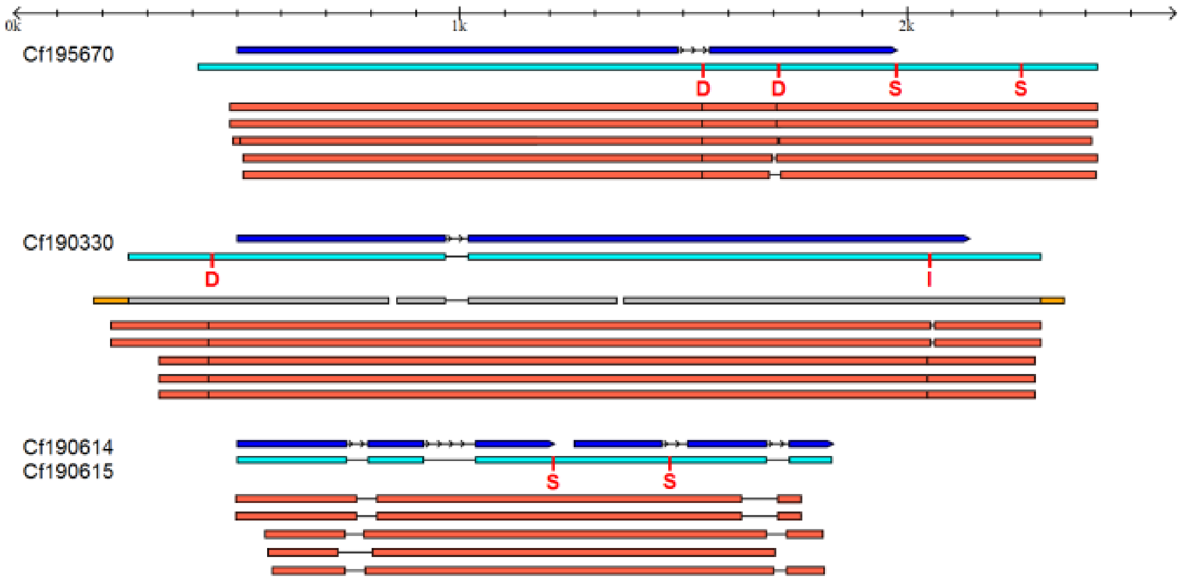


Figure 1: Examples of *ab initio*-predicted gene models that should have been designated as pseudogenes according the ABFGP method.

Three samples of GeneMark-ES- predicted gene models (blue) compared with ABFGP-predicted [10] pseudogene models (cyan) containing disruptive mutations (DMs, marked red). DMs are labelled as insertion, deletion or in-frame stop. Available EST data were manually annotated as CDS (grey) or UTR (orange). *Cf195670*: gene encoding an unknown protein. *Cf190330*: gene encoding an oxidoreductase. *Cf190614/Cf190615*: gene presumably encoding a glycosyl hydrolase. The genes are taken from the gene catalogue of *Cladosporium fulvum* (de Wit *et al.*, 2012). Five randomly chosen blastx similarities (brown) against proteins from the nr and Trembl database indicate approximate coding regions and in all cases support the pseudogene model extensions and the false gene split for Cf190614 and Cf190615.

Table 1. Biological properties, genome sizes and gene content of six fungal species.

Species	Mode of reproduction	Life style	Size (Mb)	Genomic coverage (fold)	Reference	No. of genes	Studied genes
Botrytis cinerea	Sexual	Necrotroph	43.4	4.5	[34]	16448	8504
Cladosporium. fulvum	Asexual [1]	Biotroph	61.1	21.1	[11]	14127	7575
Dothistroma. septosporum	Asexual [2]	Hemi-biotroph	31.2	34.2	[11]	12580	8091
Mycosphaerella fijiensis	Sexual	Hemi-biotroph	74.4	7.1	JGI [4]	10313	7285
Verticillium dahliae	Asexual [3]	Hemi-biotroph	34.4	7.5	[35]	10535	8362
Zymoseptoria tritici	Sexual	Hemi-biotroph	40.3	8.9	[36]	10952	7904

[1] sexual stage unknown (Thomma et al., 2005; Stergiopoulos et al., 2007)

[2] For *D. septosporum*, both mating types have been reported, but reproduction is predominantly asexually [17]. The sequenced *D. septosporum* NZE10 was isolated from a population in New Zealand that contains only one mating type and only reproduces asexually since its introduction in the 1960s.

[3] For *V. dahliae*, both mating types have been reported, but reproduction is predominantly asexually (Usami et al., 2009).

[4] Unpublished; http://genome.jgi.doe.gov/Mycfi1/Mycfi1.home.htm

Genes with predicted disrupted mutations.

Around 8,000 predicted gene models for each of the six selected fungi were assessed by ABFGP [10] using informant genes from up to 28 different fungal species (Table S1, see Supporting Information). Biological properties and genome statistics of the six fungi belonging to the class of *Ascomycetes* are shown in Table 1. From this data set, we retrieved the gene models with predicted DMs resulting in a subset of 1713 genes (ranging from 68 to 567 affected genes per species) containing 2762 DMs in total for the six fungal species.

The number of SEs occurring in sequenced genomes is expected to be inversely related to genome coverage. This renders the prediction of DMs in *Z. tritici, V. dahliae, M. fijiensis* and *B. cinerea* (in decreasing order of genome coverage) more unreliable than in the genomes of *C. fulvum* and *D. septosporum,* which have been sequenced using next generation sequencing techniques at 21-fold and 34-fold coverage, respectively [11]. From those genomes with low coverage DMs that could not be confirmed by resequencing (or sequencing related isolates) were scored as incorrect and accordingly were removed from the DM data set (Method S1, see Supporting Information). This accounted for 39 (34%), 453 (54%), 105 (46%) and 363 (72%) SEs, in the four fungi which indeed correlate with sequence coverage. A 100nt window surrounding a predicted DM in *C. fulvum* was inspected in the genome assembly for coverage, correct base calling and presence of polypyrimidine tracts. No indication for sequence errors was observed (data not shown). This refinement yielded a final set of 1.662 presumed true DMs in 924 genes which were used throughout this analysis (Table 2). The predicted ancestral protein products of the 924 genes are provided in Datafile S1 (see Supporting Information).

Table 2. High quality set of 1.666 presumed true disruptive mutations (DMs) in 924 genes.

Species	All DMs	Genes with DMs	Genes (%)	Substitutions	Indels
Botrytis cinerea	130	82	1.0	66	64
Cladosporium. fulvum	565	372	4.9	256	309
Dothistroma. septosporum	497	194	2.4	247	250
Mycosphaerella fijiensis	97	60	0.8	47	50
Verticillium dahliae	308	173	2.1	121	187
Zymoseptoria tritici	69	43	0.5	34	35
Total	**1666**	**924**		**771**	**895**

As DMs recognized by ABFGP are located in exons of their functional homologs, we conclude that DMs are present in mature mRNAs and not in the introns. For five out of six of the studied fungi we aligned available unigene data to their genomes to verify whether predicted DMs overlapped with exons or introns (Table S2, see Supporting Information). Many of the identified pseudogenes appeared to be expressed at 72% and 74% for *C. fulvum* and *D. septosporum*, respectively. In total 572 DMs were covered by ESTs confirming that they occurred in exons. In all cases where a DM was overlapping with a predicted intron (like the first deletion in Cf195670 in Figure 1), EST data indicated absence of splicing. Only eleven DMs (1.9%) matched to introns and are therefore wrongly predicted as DMs. The latter number reflects the false discovery of DMs by ABFGP. Interestingly, three out of these eleven wrongly predicted DMs matched to alternatively spliced transcripts with intron retention around the DM site.

Although examination of unigene data indicated at least 98% accuracy in appointing DMs by ABFGP, we decided to closer examine and experimentally confirm several of them. DMs were not chosen at random, but all predicted DMs in a particular class of genes in *C. fulvum*, namely secreted proteases, were selected. Five protease genes with predicted DMs (Figure 2) were resequenced in the type strain and in six additional isolates of *C. fulvum* originating from different parts of the world (Table S3 and S4, see Supporting Information). All DMs were confirmed and appeared identical in all seven isolates analyzed: two collected in The Netherlands , two collected in Cuba and two collected in Japan. Seven out of eight DMs coincided with introns predicted by GeneMark-ES [12], which all were in conflict with observed expression data. This suggests that the predicted introns are incorrect and represent DMs. To validate this, cDNA libraries from the *C. fulvum* sequenced reference strain (CBS131901) grown in different conditions were analyzed (Table S3, see Supporting Information). The results confirmed that, except for Cf189824, all genes were clearly expressed and in none of the tested growth conditions support for splicing of any of the wrongly predicted introns was observed (data not shown). For the second DM leading to protein truncation of Cf186241, the cDNA covered the complete ancestral protein, suggesting that the parental gene locus once produced a functional transcript. All genes encode proteins with crucial functional domains interrupted by or downstream of the first encountered DM (Figure 2). Based on these results we conclude thatnone of them produce mRNAs that can be translated in a functional protease.

Analysis of 1.662 disruptive mutations in 924 genes

The 1662 DMs identified in 924 genes could be subcategorized as nucleotide substitutions (46%) and indels (54%) (Figure 3a). Indels were based on the DNA sequences of informant genes estimated to represent nucleotide deletions (30%) and nucleotide insertions (24%). The frequencies of these subcategories appeared fairly similar for the different species; they varied from 39 to 50% for substitutions (Figure 3b). The point mutations leading to the stop codons TAG, TGA and TAA accounted for 49, 27 and 23% of in-frame stops, respectively (Figure 3c). These frequencies are as expected based on the notion that transitions occur more frequently than transversions and observed codon usage in *C.fulvum* and *D. septosporum* (Method S2, see Supporting Information). We conclude that the observed type of mutations result from random DNA mutations. Remarkably, only fourteen pseudogene models contained long stretches of N-nucleotides which might represent repetitive sequence due to inserted transposons as will be discussed later.

Pseudogenes with DMs are evenly distributed over the genome.

If transposon insertion or repeat-induced point mutation (RIP) would play a significant role in the creation of pseudogenes, they would occur more frequently in direct vicinity of repeats that might have undergone RIP. Other biased genomic distributions of pseudogenes could point to preference of specific chromosomes, specific parts of chromosomes or gene clusters. Only 105 (11%) of the pseudogenes are located within a distance of 1-kb of a repeat or scaffold end (Figure S1, see Supporting Information), and only 32 pseudogenes (3.4%) are located close to repeat areas that have undergone RIP (Figure S2, see Supporting Information). Only 14 pseudogenes embedded a repeat within their coding sequence (Datafile S1, see Supporting Information), which represent most likely genes inactivated by transposon insertion. On average, pseudogenes were 26.3-kb apart from repeats, and for the extremely repeat-dense *C. fulvum* genome [11] the average distance was reduced to 14.5-kb. Therefore we conclude that presence of repeats and RIP activity were of minor importance on the evolution of pseudogenes genes that we have studied here. The pseudogenes did not only lack a positional bias towards repeats, also no general trends for chromosome enrichment neither a positional enrichment towards other pseudogenes could be observed. In general, pseudogenes are evenly distributed over the chromosomes of *Z. tritici* and *D. septosporum*.(Table S5 and S6, see Supporting Information). No enrichment on the dispensable chromosomes of *Z. tritici* was observed.

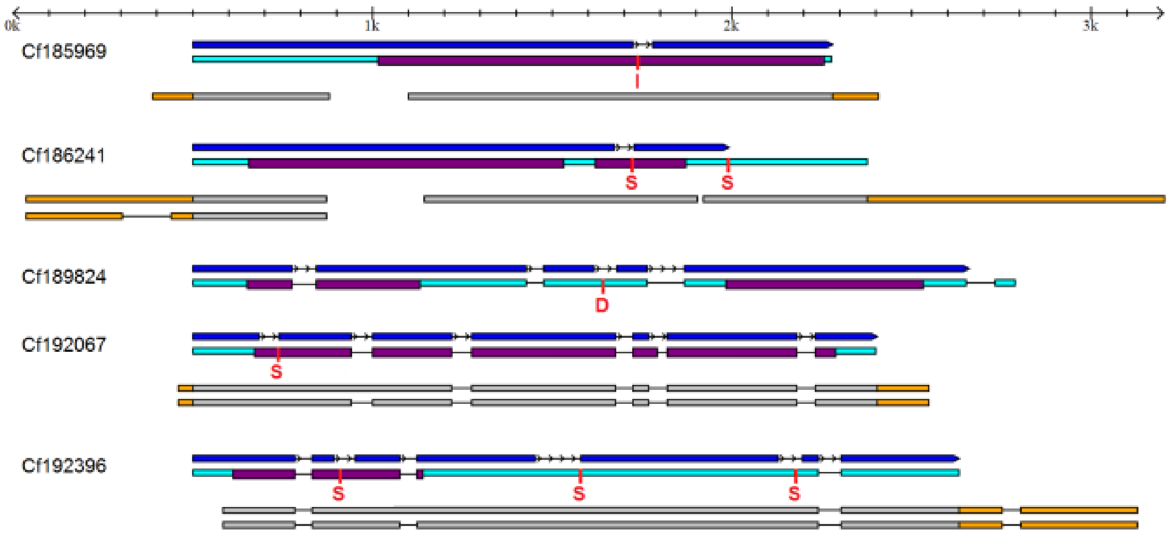


Figure 2: Pseudogenization of genes in *Cladosporium fulvum* encoding secreted proteases.

Gene models and predicted PFAM domains (purple) of five pseudogenized secreted proteases of *Cladosporium fulvum*;.for explanation of colours and symbols see Figure 1. For *Cf192067*, its fourth exon exon is incorrectly predicted by ABFGP. For *Cf189824*, an additional 3' exon is predicted by high-confidence sequence alignment to informant genes (data not shown).

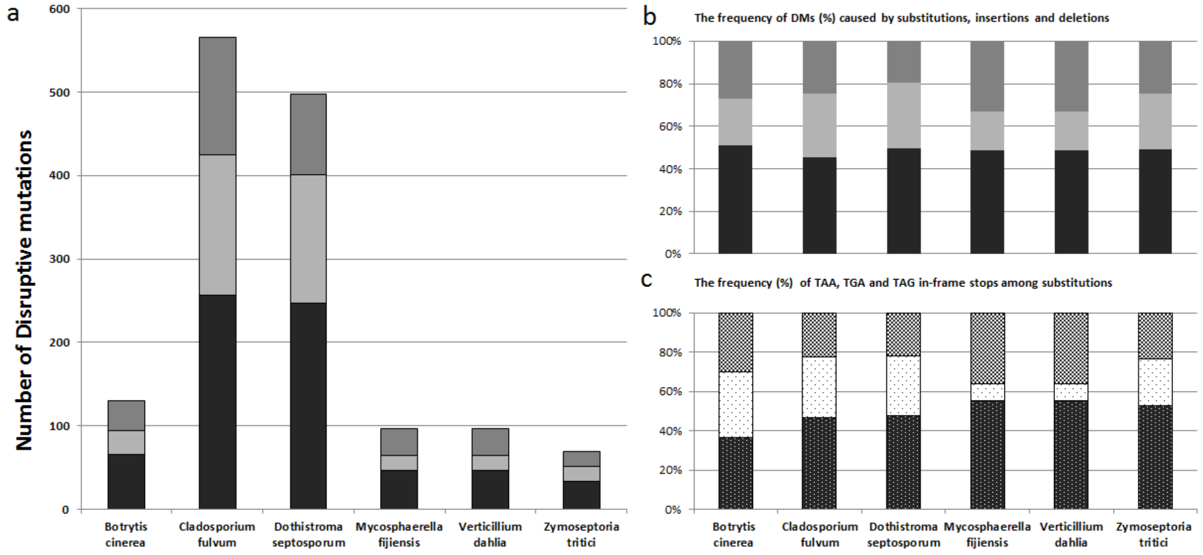


Figure 3: Disruptive mutations (DMs), type of DMs and their frequency.

A. Numbers of DMs caused by substitutions (black), insertions (dark grey) and deletions (light grey) after removal of sequencing errors in the six different fungi. B. The frequency of DMs (%) caused by substitutions, insertions and deletions in six different fungi**.** C. The frequency (%) of TAG (dark), TGA (white) and TAA (grey) in-frame stop codons observed in DMs representing substitutions from the six different fungi.

We observed a median distance of pseudogene per 147-kb, with the exception of *C. fulvum*, where this number was on average one pseudogene per 34.8-kb (Figure S3, see Supporting Information). The observed median and average inter-pseudogene distances indicate that pseudogenes do not tend to cluster together, although occasionally (nearly) adjacent gene pairs were pseudogenized. In *C. fulvum* and *D. septosporum*, the species with the most pseudogenes, in total 23 pairs of directly adjacent pseudogenes were observed (Table S7 and S8, see Supporting Information). This is slightly more than what could be expected based on chance only (data not shown); therefore all pairs were inspected for being member of a gene cluster. In the pseudogene-rich *C. fulvum* some clear examples of functionally related, adjacent pseudogenes were found: a quartet of four adjacent pseudogenes which are involved in carbohydrate metabolism (Cf186934- Cf186937) and a triplet that encoded a putative chitinase, amino acid transporter and phosphodiesterase/alkaline phosphatase (Cf191135-Cf191137), respectively (Table S8, see Supporting Information).

A bias for pseudogenization of members of multi-gene families and secreted proteins

For each gene and pseudogene, the (global) amino acid similarity to their most similar protein-encoding homolog in the complete protein catalogue was determined (Figure 4a). Additionally, the total number of potential homologs was counted to express membership and size of a multi-gene family. In total 682 pseudogenes, representing 74% of all DM-containing pseudogenes, share 45% to 75% similarity with at least a single homologous, non-pseudogenized protein which is more than the genomic average. The majority of this class of pseudogenes has more than one homolog (Figure 4b) suggesting that multi-gene families seem more frequently affected by pseudogenization. When comparing the multi-gene family size of this class with all multi-gene families, no significant difference, increase nor decrease, in gene-family size could be observed.

Genes encoding proteins that are less than 45% similar are less affected by pseudogenization (Figure 4a). Based on these findings, we have made an arbitrary distinction between recent gene duplicates (>75% similarity), single copy genes (<45% similarity) and genes that share between 75 and 45% sequence similarity. Remarkably, the set of 924 pseudogenes is not enriched for recent gene duplications, as was expected based on the general observation made in other higher eukaryotes [3]. Figure 4a shows that recent gene duplications do not only occur rarely in the six studied genomes, but they are also not enriched for pseudogenes. At the proposed threshold of at most 45% similarity, 22% of all pseudogenes (8 to 62 per species and 203 in total) can be classified as single-copy, unitary pseudogenes [5],[2]. Pseudogenization of genes belonging to multi-gene families suggests that some members might be redundant. In contrast, pseudogenization of unitary genes have most likely a direct impact on the functional repertoire of an organism.

Because the studied fungi are all plant pathogens which manipulate their host by means of secreted proteins, pseudogenization of genes encoding this class of proteins was studied in more detail. Between one and 51 genes encoding secreted proteins appeared pseudogenized (Figure 5). On average secreted proteins account for around 10% of all proteins in these pathogenic fungi. In *C. fulvum*, pseudogenes encoding secreted proteins are significantly overrepresented, but are significantly underrepresented in *M. fijiensis* (but it should be noted that only small numbers of pseudogenes are present in the latter fungus). Remarkably, the percentage of pseudogenes that used to encode secreted proteins in *C. fulvum* was twice as high as that observed in its close relative *D. septosporum*.

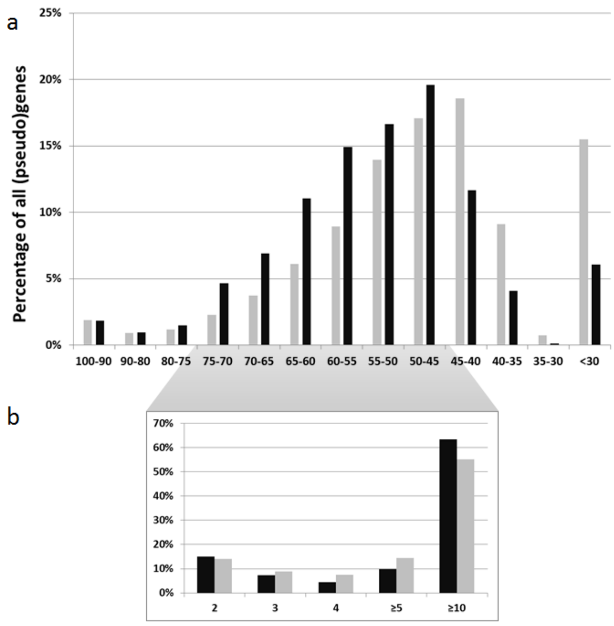


Figure 4a: Pairwise amino acid sequence similarity of proteins encoded by pseudogenes and their most similar functional homolog present in the predicted proteome.

Figure 4b: A bias for pseudogenization of multi-gene families

Gene family size distribution of (pseudo)genes with at least a single functional homolog (between 45 and 75% homology). All proteins (n=74,955) are compared to all pseudogenes (n=924) in six species in the specified range of similarity. Gene family membership threshold are set to a bit score of ≥ 200 (blastp) and similarity ≥ 60% of the protein's length.

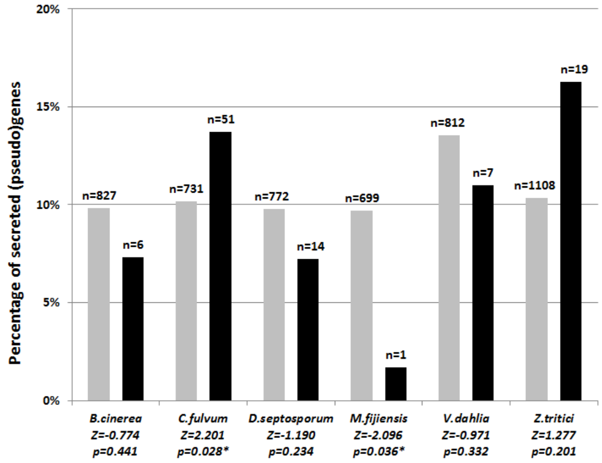


Figure 5: Pseudogenization of genes encoding secreted proteins in six different fungi.

Frequency (%) of genes encoding secreted proteins (grey bars) and of pseudogenes encoding secreted proteins (black bars) for six different fungi. Total number of (pseudo)genes analyzed per species is shown on top of the bars. Two-tailed Z-test statistics were calculated (http://socscistatistics.com); at p<0.05 pseudogenes encoding secreted proteins are significantly enriched in *C. fulvum*, whereas they are significantly underrepresented in *M. fijiensis.*

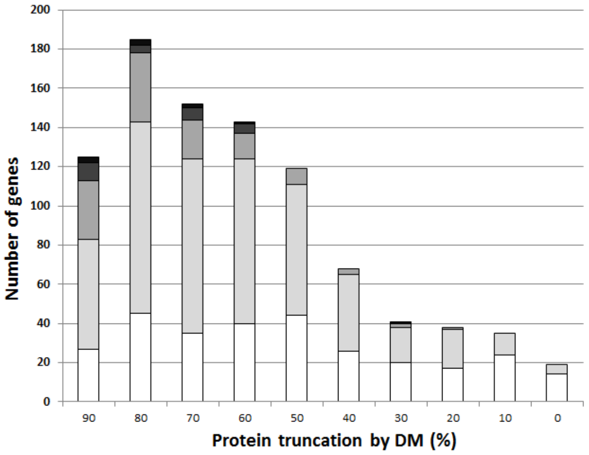


Figure 6: Protein truncation and number of PFAM domains lost by truncation caused by the first 5' disruptive mutation (DM).

Truncation of proteins is expressed as percentage of the total protein length; the number of PFAM domains lost by the first 5' DM are indicated in a greyscale from 0 (white), 1, 2, 3 to ≥4 (black).

Estimation of loss of function among the 924 genes with disruptive mutations

A pseudogene can cause a loss of function (LOF), or a change of function of the encoded protein, but detailed functional analyses are required to draw reliable conclusions. To address this question by an *in silico* approach, we quantified protein length truncation and the number of lost PFAM domains which are located downstream of the most 5′ DM in the gene.(Table S9, see Supporting Information).The results are summarized in Figure 6. 824 of the encoded proteins (89%) are truncated by more than 50% or lost at least one functional domain. In contrast only 75 proteins (8%) are truncated by less than 30% without having lost a single known protein domain. Based on these numbers, we assume that the vast majority of genes with DMs do no longer encode a functional protein, or a protein that no longer fulfils its ancestral function.

Discussion

The ABFGP method recognized many DMs in genes which resulted in frameshifts (non-3n insertions) or in-frame stop codons when compared with functional informant genes from fungi lacking these DMs. Closer inspection of four resequenced genomes showed that a large fraction of DMs appeared SEs. Genes in genomes sequenced with low coverage contain considerable numbers of SEs in regions of protein-encoding genes (*B. cinerea*, *V. dahlia, M. fijiensis and Z. tritici*) which hampered the correct assignment of gene models. Occurrence of thousands of SEs in the original reference genomes was independently shown in *B. cinerea* [13] (Method S1, see Supporting Information). This is likely also the case for many other fungal genomes that have been sequenced in the era of low coverage Sanger sequencing.

Estimation of the extent of pseudogenes in fungi

In this study we identified 924 pseudogenes in the gene catalogue of six different fungi, representing 0.5 to 4.9% of their annotated genes. This number is likely a strong underestimation of the total extent of pseudogenization in these fungi due to the necessity of and method chosen for sampling informant genes. In the following section evidence for underestimation is provided and discussed. The provided examples are all chosen from *C. fulvum* and *D. septosporum,* for which the prior knowledge of higher than expected levels of pseudogenization was used throughout the analyses of their genomes [11].

As a start, only 7300 to 8500 of the annotated genes per species have been searched for the occurrence of pseudogenization of genes that are shared among 28 fungi. Genes that were not eligible to ABFGP are highly divergent, short, clade- or species-specific like effector genes. Only few effectors are shared among fungi and most are species-specific, but they can also be subject to pseudogenization when selection is imposed. A clear example of pseudogenization of a species-specific effector is reported for the *Avr2* gene of the tomato pathogen *C. fulvum* [14]. Other examples are the homologs of the *C. fulvum* effector proteins *Ecp4* and *Ecp5,* which were identified as pseudogenes in *D. septosporum* due to presence of in-frame stop codons [11]. These pseudogenes were not identified in this study because of absence of close homologs in the gene catalogues of the 28 fungi used.

Furthermore, DMs were called by ABFGP [10] in regions supported by strong sequence similarity to exons of informant genes. DMs in regions with poor similarity support, directly adjacent or even in splice sites, translational start sites or promoter regions are not addressed by ABFGP, but their contribution to pseudogenization can be significant. An example is the key secondary metabolite (SM) pseudogene (*Pks9*) in *C. fulvum [11]*, which was not detected in our study, because of a single in-frame stop codon only 16nt upstream of the donor sequence of its third (EST-supported) intron.

Finally, when several DMs or dramatic DMs (e.g. transposon insertions) are present in a gene, gene prediction software is likely to predict a fragmented (like example three in Figure 1), highly truncated or even no gene model at all. This might account for a rather large number of genes, which is emphasized by failure of detection of six out of seven manually annotated key SM pseudogenes in *C. fulvum* [11]. Due to occurrence of DMs, *Nps5* was predicted to be divided over two separate gene models; *Nps7*, *Nps10* and *Hps2* even over three separate gene models. *Nps1* is highly truncated, most likely by the transposon insertion (which concurrently marks the end of its contig), whereas *Pks4* has several adjacent DMs which resulted in a predicted but not existing 864nt intron. In all instances, recruitment of suitable informant gene loci failed for these (fragmented) gene models, explaining why the gene model was not to be among the assessed genes by ABFGP.

In general, the (manual) annotation efforts invested in the gene catalogs of each of the six studied species varies significantly and might have affected the quality of the integral gene catalogue considerably, for instance by prior removal of obvious pseudogenes. This indicates that the practical delimitation to (somewhat properly) annotated gene models might have introduced a methodological bias in this study, resulting in failure of detection of pseudogenes that contain more or dramatic DMs, which is the most obvious category to be resolved by manual curation of a gene catalogue. To further investigate this issue, we decided for an additional experiment which compares the gene catalogs of closely related species. In a pairwise comparison, genes unique to one species can be the result of gene gains in that species, gene losses in the other or due to un-annotated genes because of pseudogenization [2]. Such an analysis was performed on our data set using the closely related *Capnodiales* species *C. fulvum*, *D. septosporum* and *Z. tritici*, and unambiguously identified 674 additional pseudogenes on loci lacking annotated gene models or containing misannotated fragments of longer genes (Method S2, Table S10, Datafile S2, see Supporting Information). In this pseudogene dataset, *C. fulvum* again stands out in terms of total number of pseudogenized genes. Among these genes were several (types of) pseudogenes that were expected. Approximately half of the pseudogenes are listed in the gene catalogue of these species, but as truncated, incorrectly predicted genes. Second, a small proportion was identified as being disrupted by repetitive sequence, which most likely represent transposon insertions and explains why they were not predicted as genes by the gene prediction software. A higher incidence of putative transposon insertion in *C. fulvum* compared to *D. septosporum* is likely correlated to the much higher repetitive content of the first The additionally identified pseudogenes do not alter the lower degree of pseudogenization in *Z. tritici* compared to *C. fulvum* and *D.septosporum*. Therefore, the initial observation that less genes are pseudogenized in sexual versus asexually reproducing fungi remains valid.

Overall, we conclude that the actual number of pseudogenes in these six fungi is considerably higher than that described in this study as assigned by ABFGP [10]. The data set described here represents only a subset of pseudogenes that is listed in current gene catalogs. DMs and SEs together account for a high error-rate in current gene catalogs, which hampers *in silico* comparative genomics analyses.

Pseudogenes occur more frequently in asexually reproducing fungi

Significant difference in the frequency of DMs occur in the six different fungi. The highest frequency of DMs was observed in the two related fungi *C. fulvum* and *D.septosporum.*These large differences in level of pseudogenization might be related to their mode of reproduction (Table 1). Species that, apart from asexual reproduction, reproduce also sexually like *B.* cinerea, *M. fijiensis* and *Z. tritici* show lower numbers of pseudogenes as compared to those that reproduce predominantly asexually like *C. fulvum* [15], [16] *D. septosporum* [17] and *V. dahlia* [18]. Deleterious DMs in sexually reproducing *Ascomycetes* can be either lost or restored after recombination and selection. It is assumed that haploid asexual fungi will initially adapt quicker to a new environment than sexually reproducing relatives. Pseudogenization of genes which are no longer required, not of advantage or even deleterious for a pathogen might enable it to quickly adapt to new environments including new host plants.

The set of 924 pseudogenes that were identified in the six fungal genomes did not show a biased genomic distribution. The only exception to a random distribution over chromosomes are occasional (nearly) adjacent pseudogenes with related functions, suggesting that more than one gene of the same pathway was affected (Table S7 and S8, see Supporting Information). Unexpectedly, no preference for pseudogenes in vicinity of repeats, whether or not affected by RIP, was observed. A relation between repeats and pseudogenes/gene loss has been suggested in powdery mildew fungi due to retrotransposon insertions in the absence of RIP [19], while RIP activity in *Leptosphaeria maculans* clearly affected nearby located genes [9]. Also in fungi that are assumed to reproduce asexually, RIP signatures have been observed [20], as was the case in *C. fulvum* and *D. septosporum* [11]. This indicates that these fungi once were sexually active but lost their ability to reproduce sexually or sexual reproduction occurs only rarely [17]. Extensive RIP activity will dramatically affect continuity of a coding sequence and *ab initio* gene prediction will likely not predict a gene on a RIP-affected locus. Because the data set of 924 pseudogenes were retrieved from catalogues of predicted genes it is expected to be underrepresented for pseudogenes caused by RIP. The same holds true for genes inactivated by transposon insertion; indeed only fourteen gene models with putative transposon insertions were identified. The apparent underrepresentation of pseudogenes by transposon insertion was further addressed by the additional *in silico* experiment on the three *Capnodiales* species discussed before (Table S10, see Supporting Information), where a small number of additional pseudogenes of this type was identified. However, even when these additional pseudogenes are taken along, the contribution of transposon insertion to pseudogenization is of minor impact compared to indels and substitutions.

Pseudogenes in multi-gene families and unitary pseudogenes

Our analyses showed that fungal pseudogenes are not predominantly associated with recent gene duplications, but occur predominantly in multi-gene families. 74% of all pseudogenes have a closest homolog within the 45-75% similarity range and of these 70% belong to multi-gene families of at least 5 members (Figure 4b). One could argue that predominantly (partially) redundant genes become randomly pseudogenized. For example, high-throughput gene knock-out studies in *Schizosaccharomyces pombe [21]*,[22] showed that 17.5% of genes when knocked-out caused a lethal phenotype. Most knock-out mutants gave no or weak phenotypes, suggesting some level of functional redundancy for many genes. However, these conclusion can only be drawn when supported by ecological studies performed on populations enabling comparison the fitness of wild-type and knock-out mutants under different environmental conditions. Therefore, we expect that pseudogenization of members of multiple gene families is likely involved in subtle adaptations of fungi to different environmental conditions, whereas pseudogenization of unitary genes is expected to have more drastic effects on phenotypes including even beneficial ones when they would facilitate adaptation to a new environment. However, attributing pseudogenes in multi-gene family to mere redundancy is not supported by reported functional diversification in gene families. Proteases cluster into several gene families based on sequence similarity in their functional domain(s), yet have very distinctive substrate specificities (Monod *et al.*, 2002, Hedstrom, 2002, Yike 2011). In Figure 4a, they fall in the class of sharing intermediary similarity to a non-pseudogenized homologue and putatively being member of a multi-gene family. Some but not all proteases have been reported to cause tissue necrosis and their pseudogenization might suppress this phenotype [23].

Pseudogenization of genes encoding secreted proteins and key secondary metabolite genes in *C. fulvum* might reflect host adaptation

In *C. fulvum* genes encoding secreted proteins showed a higher frequency of pseudogenization, and among these were five secreted proteases. The eight DMs in these five genes were confirmed by various approaches. It is tempting to speculate that pseudogenization of genes encoding secreted proteases could be related to the lifestyle of *C. fulvum*. Many proteases are known to induce senescence and sometime cell death, which could facilitate some plant pathogenic fungi to kill plants and retrieve nutrients from necrotized or death cells. *C. fulvum* is a biotrophic fungus thriving in the apoplast in close contact with mesophyll cells of tomato leaves where it lives on nutrients released by the host either passively or induced by fungal effectors. Only at very late stages of infection host cells may collapse. The origin and the ancestor of *C. fulvum* is not known, but it is closely related to *D. septosporum*, a pine pathogen that behaves as a hemi-biotroph, killing host cells after a short biotrophic phase [11]. At proteome level they are remarkably similar and their genomes share extended regions of meso-synteny, which accounts for about 70% of all genes and facilitates robust inference of orthology. When comparing the complete pseudogene catalogues of both species, only 22 pairs of closest homologs are pseudogenes in both species (Table S11, see Supporting Information). In these pairs, not a single individual DM is shared (compare Supporting Information 4). Over 85% of the pseudogenes in either of the two species have non-pseudogenized closest homologs in the other species, of which many are indisputable orthologs based on presence in meso-syntenic areas. In a few cases (four and 23, representing 2 and 6%), the pseudogenized gene is absent in one of the two species. These 23 pseudogenes in *C. fulvum* could represent quickly diversifying genes, *C. fulvum*-specific gene gains followed by pseudogenization, or genes which have been lost in *D. septosporum* at an early time point after species divergence, and now might await the same faith in *C. fulvum*. Given the fact that those genes have easily recognizable closest homologs in many more distantly related *Ascomycetes*, the latter hypothesis seems most likely.

These observations suggests that many, if not all, of the observed DM events leading to pseudogenization in *C. fulvum* and *D. septosporum* occurred post speciation. After occurrence of a first DM, the gene’s locus is expected to exhibit neutral evolution, until it is lost from the genome. A discrepancy between the speed at which neutral evolution takes place and the speed at which small genomic segments get lost might explain why relative large numbers of pseudogenes can pile up in two genomes that recently speciated. Pseudogenization of some of those genes could provide an advantage to these fungi in adapting to a new host, and the lack of a sexual cycle could accelerate the process of adaptation for these haploid fungi. The effector genes *Ecp4* and *Ecp5*, known to be involved in virulence of *C. fulvum* on tomato, are pseudogenes in *D. septosporum* [11]. Similarly, two crucial genes of the pathway that produces the toxin dothistromin during infection of pine by *D. septosporum*, are pseudogenes in *C. fulvum* [24]. Apparently selection pressure on loosing certain genes or reversely, the necessity to maintain certain genes is acting on a different gene set in these two fungi, which could reflect adaptation to a new environmental niche. We speculate that pseudogenization of genes involved in secondary metabolite production, secreted proteases and other damaging enzymes might be one of the reasons why *C. fulvum* might have been a hemi-biotrophic tree pathogen like *D. septosporum* but started to live as a biotroph when it became pathogenic on tomato. It would be interesting to determine whether removal of the stop codons in a number of these pseudogenized protease genes would make *C. fulvum* more aggressive and possibly also hemi-biotrophic on tomato. Additionally, further research in the recent ancestry of *C. fulvum* and *D. septosporum* could determine the age of the pseudogenization events and therefor the speed at which pseudogenization takes place.

This study not only supports the generally accepted fact that fungal genomes contain pseudogenes, it demonstrates that some actually have a larger number of pseudogenes than others and that many pseudogenes are still listed as bona fide genes in gene catalogues. Therefore, we argue for the need of providing fungal (functional) gene annotations not only as a gene catalogue but additionally with a pseudogene catalogue counterpart. Comparative genomics studies heavily rely on the predicted gene catalogue, but these can be hampered by the occurrence of pseudogenes that were failed to be identified. Moreover, the pseudogene arsenal of a species might still reflect an echo of the legacy of an earlier, at that point in time, optimal gene repertoire. Especially for plant pathogenic fungi, comparing several complete pseudogene repertoires might reveal interesting facts about their recent evolutionary past that could provide insight in their current host-specificity and pathogenicity.

Experimental procedures

Fungal genomes used in this study

The genomes, proteomes, annotations and available unigenes of five fungal species were downloaded from the Fungal Genome Initiative of the BROAD Institute[25] [26] and the Fungal Genomics Program of the US Department of Energy Joint Genome Institute (JGI) [27] in collaboration with the user community. The data from the *C. fulvum* genome and transcriptome were generated at Wageningen University [11] and are also available on the JGI MycoCosm website [27].

Alignment-based fungal gene prediction

Genes with predicted disruptive mutations (DMs) were obtained by genome-wide gene model assessment of six fungi with the Alignment-Based Fungal Gene Prediction (ABFGP) method [10]. Gene loci from 29 different fungi mainly belonging to the *Ascomycetes*, served as informant DNA sequences for alignment-based assessment of the genes in the six target genomes (Table S1, see Supporting Information).

Distinguishing sequencing errors from true disruptive mutations

SEs among predicted DMs were identified by comparing a 200nt window around each predicted DM with base calling in Illumina-based assemblies from *Z. tritici* strains STIR04_A26b and STIR04_A48b (62-fold versus 8.9-fold coverage of the reference genome *Z. tritici* IPO323) [28], *V. dahliae* strain JR2 (30-fold versus 7.5-fold coverage of *V. dahlia* VdLs.17 genome) [29], *M. fijiensis* strain CIRAD139a (25-fold versus 7.1 fold coverage of *M. fijiensis* CIRAD86 genome) and *B. cinerea* strain B05.10 (50-fold versus 4.5-fold coverage assembly of the same isolate) [13]. All DMs that could not be confirmed as truly occurring in the population of these fungi were removed from the DM data set (Method S1, see Supporting Information). Additionally, DMs that were discovered to be falsely predicted by analyses of EST data (11 DMs, Table S2, see Supporting Information) and gene models containing short, contiguous stretches of n-characters directly adjacent to DMs (7 DMs, data not shown) were removed.

Determining closest protein homolog and gene family size

For each protein, the closest homolog (in the proteome of the fungal species) was determined as the protein with the highest bitscore of concatenated alignments (blastp), requiring the alignment to span at least 60% of the length of both proteins. A simple estimation of gene family size was performed by counting number of proteins with a score of at least 200 bits, requiring the same alignment length.

Third party software

Predicted (pseudo-) protein sequences of the gene loci with DMs were searched for putative secretion signals using SignalP 3.0 [30] and known PFAM protein domains with InterproScan [31]. Unigenes were aligned to their genomes using GenomeThreader version 1.1.1.2 [32].

Confirmation of base calling and mRNA splicing in *C. fulvum* genes

Five *C. fulvum* genes encoding secreted proteases with predicted DMs were selected for confirmation of genome base calling and intron splicing. The original sequences were obtained from the published genome of *C. fulvum* race 0WU (CBS131901) [11]. The sequences at and around a DM site were amplified with primers given in Table S3 (see Supporting Information). The presence of the DMs was analysed in six different *C. fulvum* isolates varying in geographical origin, race and mating type (Table S4, see Supporting Information). From *C. fulvum* CBS131901 total RNA was isolated from mycelium grown under different *in vitro* and *in planta* conditions (Table S3, see Supporting Information) and the amplified cDNA fragments (using the same primer pairs) were evaluated for occurrence of splicing around the DMs.

Repeat identification and RIP analyses

Repeats were determined using mummer (-maxmatch -nosimplify) [33] as segments of at least 250nt that are present in at least five copies in a given genome. RIP analysis was performed as described for *C. fulvum* and *D. septosporum* in de Wit *et al.* (2012).

Acknowledgements

The authors thank Prof. Dr. Bart P.H.J. Thomma (WUR) for providing sequencing data of *V. dahlia* [29], Dr. Jan A.L. van Kan (WUR) for resequencing data of *B. cinerea* [13] and Dr. Gert H. J. Kema (Plant Research International / WUR) for unpublished sequencing data of *M. fijiensis*. Sequence data were primarily prepared for other manuscripts which were recently publically released.

Supporting Information

Table S1: 28 informant species used for ABFGP
Table S2: Expression evidence of pseudogenes and false discovery estimation of DMs by ABFGP measured on unigene data.
Table S3: Sequences of primers used for resequencing disruptive mutations in secreted proteases of *Cladosporium fulvum*.
Table S4: Six strains of *Cladosporium fulvum* used to confirm presence of predicted DMs in secreted proteases.
Table S5: Pseudogene distribution over the chromosomes of *Zymoseptoria tritici*.
Table S6: Pseudogene distribution over the chromosomes of *Dothistroma septosporum*.
Table S7: Directly adjacent pseudogenized gene pairs and triplet in *Dothistroma septosporum.*
Table S8: Directly adjacent pseudogenized gene pairs and triplet in *Cladosporium fulvum*.
Table S9: Functional prediction of 924 pseudogenes by assignment of PFAM domains.
Table S10: Hundreds of additional pseudogenes in the *Capnodiales* species *Cladosporium fulvum*, *Dothistroma septosporum* and *Zymoseptoria tritici* detected based on similarity to known proteins.
Table S11: Twenty two pseudogenes shared between *C. fulvum* and *D. septosporum*

Figure S1: Distance of pseudogenes towards repetitive sequence or scaffold boundary.
Figure S2: Distance of pseudogenes towards RIP'd loci.
Figure S3: Inter-pseudogene distance.

Method S1: Removal of sequence errors from predicted DMs by comparison with next-generation sequencing assemblies of four genomes with low coverage.
Method S2: Observed pattern of substitutions is explained by determinants of random DNA mutation alone.
Method S3: Additional pseudogenes in *Cladosporium fulvum*, *Dothistroma septosporum* and *Zymoseptoria tritici* identified by pairwise comparisons.

Datafile S1: Protein sequences of 924 pseudogenes including 1.662 disruptive mutations.
Datafile S2: GFF annotation of 674 extra pseudogenes found in the *Capnodiales* species *Cladosporium fulvum*, *Dothistroma septosporum* and *Zymoseptoria tritici.*

References

1. Yang L, Takuno S, Waters ER, Gaut BS: Lowly expressed genes in Arabidopsis thaliana bear the signature of possible pseudogenization by promoter degradation. *Mol Biol Evol* 2011, 28(3):1193-1203.
2. Zhang ZD, Frankish A, Hunt T, Harrow J, Gerstein M: Identification and analysis of unitary pseudogenes: historic and contemporary gene losses in humans and other primates. *Genome Biol* 2010, 11(3):R26.
3. Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M: What is a gene, post-ENCODE? History and updated definition. *Genome Res* 2007, 17(6):669-681.
4. Lynch M, Conery JS: The evolutionary fate and consequences of duplicate genes. *Science* 2000, 290(5494):1151-1155.
5. Balasubramanian S, Habegger L, Frankish A, MacArthur DG, Harte R, Tyler-Smith C, Harrow J, Gerstein M: Gene inactivation and its implications for annotation in the era of personal genomics. *Genes Dev* 2011, 25(1):1-10.
6. Westerink N, Brandwagt BF, de Wit PJ, Joosten MH: Cladosporium fulvum circumvents the second functional resistance gene homologue at the Cf-4 locus (Hcr9-4E) by secretion of a stable avr4E isoform. *Mol Microbiol* 2004, 54(2):533-545.
7. Stergiopoulos I, De Kock MJ, Lindhout P, De Wit PJ: Allelic variation in the effector genes of the tomato pathogen Cladosporium fulvum reveals different modes of adaptive evolution. *Mol Plant Microbe Interact* 2007, 20(10):1271-1283.
8. Galagan JE, Selker EU: RIP: the evolutionary cost of genome defense. *Trends Genet* 2004, 20(9):417-423.
9. Gout L, Kuhn ML, Vincenot L, Bernard-Samain S, Cattolico L, Barbetti M, Moreno-Rico O, Balesdent MH, Rouxel T: Genome structure impacts molecular evolution at the AvrLm1 avirulence locus of the plant pathogen Leptosphaeria maculans. *Environ Microbiol* 2007, 9(12):2978-2992.
10. van der Burgt A: ABFGP. 2013.
11. de Wit PJ, van der Burgt A, Okmen B, Stergiopoulos I, Abd-Elsalam KA, Aerts AL, Bahkali AH, Beenen HG, Chettri P, Cox MP *et al*: The genomes of the fungal plant pathogens Cladosporium fulvum and Dothistroma septosporum reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. *PLoS Genet* 2012, 8(11):e1003088.
12. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M: Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. *Genome Res* 2008, 18(12):1979-1990.
13. Staats M, van Kan JA: Genome update of Botrytis cinerea strains B05.10 and T4. *Eukaryotic cell* 2012, 11(11):1413-1414.
14. Luderer R, Takken FL, de Wit PJ, Joosten MH: Cladosporium fulvum overcomes Cf-2-mediated resistance by producing truncated AVR2 elicitor proteins. *Mol Microbiol* 2002, 45(3):875-884.
15. Thomma BP, HP VANE, Crous PW, PJ DEW: Cladosporium fulvum (syn. Passalora fulva), a highly specialized plant pathogen as a model for functional studies on plant pathogenic Mycosphaerellaceae. *Molecular plant pathology* 2005, 6(4):379-393.
16. Stergiopoulos I, Groenewald M, Staats M, Lindhout P, Crous PW, De Wit PJ: Mating-type genes and the genetic structure of a world-wide collection of the tomato pathogen Cladosporium fulvum. *Fungal genetics and biology : FG & B* 2007, 44(5):415-429.
17. Dale AL, Lewis KJ, Murray BW: Sexual reproduction and gene flow in the pine pathogen Dothistroma septosporum in British Columbia. *Phytopathology* 2011, 101(1):68-76.
18. Usami T, Itoh M, Amemiya Y: Asexual fungus Verticillium dahliae is potentially heterothallic. *J Gen Plant Pathol* 2009, 75(6):422-427.
19. Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stuber K, Ver Loren van Themaat E, Brown JK, Butcher SA, Gurr SJ *et al*: Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism. *Science* 2010, 330(6010):1543-1546.
20. Oliver R: Genomic tillage and the harvest of fungal phytopathogens. *The New phytologist* 2012, 196(4):1015-1023.

21. Decottignies A, Sanchez-Perez I, Nurse P: Schizosaccharomyces pombe essential genes: a pilot study. *Genome Res* 2003, 13(3):399-406.
22. Spirek M, Benko Z, Carnecka M, Rumpf C, Cipak L, Batova M, Marova I, Nam M, Kim DU, Park HO *et al*: S. pombe genome deletion project: an update. *Cell cycle* 2010, 9(12):2399-2402.
23. Gilroy EM, Hein I, van der Hoorn R, Boevink PC, Venter E, McLellan H, Kaffarnik F, Hrubikova K, Shaw J, Holeva M *et al*: Involvement of cathepsin B in the plant disease resistance hypersensitive response. *Plant J* 2007, 52(1):1-13.
24. Chettri P, Ehrlich KC, Cary JW, Collemare J, Cox MP, Griffiths SA, Olson MA, de Wit PJ, Bradshaw RE: Dothistromin genes at multiple separate loci are regulated by AflR. *Fungal genetics and biology : FG & B* 2013, 51:12-20.
25. BROAD Fungal Genome Initiative [http://www.broadinstitute.org/scientific-community/science/projects/fungal-genome-initiative/fungal-genome-initiative]
26. Cuomo CA, Birren BW: The fungal genome initiative and lessons learned from genome sequencing. *Methods Enzymol* 2010, 470:833-855.
27. Grigoriev IV, Nordberg H, Shabalov I, Aerts A, Cantor M, Goodstein D, Kuo A, Minovitsky S, Nikitin R, Ohm RA *et al*: The genome portal of the Department of Energy Joint Genome Institute. *Nucleic Acids Res* 2012, 40(Database issue):D26-32.
28. Stukenbrock EH, Bataillon T, Dutheil JY, Hansen TT, Li R, Zala M, McDonald BA, Wang J, Schierup MH: The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen Mycosphaerella graminicola and its wild sister species. *Genome Res* 2011, 21(12):2157-2166.
29. de Jonge R, Bolton M, Kombrink A, van den Berg G, Yadeta K, Thomma BP: Extensive chromosomal reshuffling drives evolution of virulence in an asexual pathogen. *Genome Res* 2013.
30. Emanuelsson O, Brunak S, von Heijne G, Nielsen H: Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc* 2007, 2(4):953-971.
31. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Das U, Daugherty L, Duquenne L *et al*: InterPro: the integrative protein signature database. *Nucleic Acids Res* 2009, 37(Database issue):D211-215.
32. Kurtz S, Gremme G, Brendel V, Sparks ME: Engineering a software tool for gene structure prediction in higher organisms. *Inform Software Tech* 2005, 47(15):965-978.
33. 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(2):R12.
34. Amselem J, Cuomo CA, van Kan JA, Viaud M, Benito EP, Couloux A, Coutinho PM, de Vries RP, Dyer PS, Fillinger S *et al*: Genomic Analysis of the Necrotrophic Fungal Pathogens Sclerotinia sclerotiorum and Botrytis cinerea. *PLoS Genet* 2011, 7(8):e1002230.
35. Klosterman SJ, Subbarao KV, Kang S, Veronese P, Gold SE, Thomma BP, Chen Z, Henrissat B, Lee YH, Park J *et al*: Comparative genomics yields insights into niche adaptation of plant vascular wilt pathogens. *PLoS Pathog* 2011, 7(7):e1002137.
36. Goodwin SB, M'Barek S B, Dhillon B, Wittenberg AH, Crane CF, Hane JK, Foster AJ, Van der Lee TA, Grimwood J, Aerts A *et al*: Finished genome of the fungal wheat pathogen Mycosphaerella graminicola reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. *PLoS Genet* 2011, 7(6):e1002070.

Chapter 5

Birth of new spliceosomal introns in fungi by multiplication of introner-like elements.

van der Burgt A, Severing E, de Wit PJGM, Collemare J.

Current Biology 2012, 22(13):1260-5.

doi: 10.1016/j.cub.2012.05.011

This article and its Supplemental Information are available from:
http://www.sciencedirect.com/science/article/pii/S0960982212005325

Summary

Spliceosomal introns are noncoding sequences that separate exons in eukaryotic genes and are removed from pre-messenger RNAs by the splicing machinery. Their origin has remained a mystery in biology since their discovery [1,2] because intron gains seem to be infrequent in many eukaryotic lineages [3,4]. Although a few recent intron gains have been reported [5,6], none of the proposed gain mechanisms [7] can convincingly explain the high number of introns in present-day eukaryotic genomes. Here we report on particular spliceosomal introns that share high sequence similarity and are reminiscent of introner elements [8]. These elements multiplied in unrelated genes of six fungal genomes and account for the vast majority of intron gains in these fungal species. Such introner-like elements (ILEs) contain all typical characteristics of regular spliceosomal introns (RSIs) [9,10] but are longer and predicted to harbor more stable secondary structures. However, dating of multiplication events showed that they degenerate in sequence and length within 100,000 years to eventually become indistinguishable from RSIs. We suggest that ILEs not only account for intron gains in six fungi but also in ancestral eukaryotes to give rise to most RSIs by a yet unknown multiplication mechanism.

Results and Discussion

Characterization of a New Type of Spliceosomal Intron that Is Able to Multiply in Unrelated Genes

Spliceosomal introns are one of the key innovations of eukaryotes [1]. They are an important component of the eukaryotic gene structure mainly because they enlarge the proteome diversity by alternative splicing and regulate gene expression at the posttranscriptional level [11,12]. Canonical regular spliceosomal introns (RSIs) share GT/AG donor and acceptor sites that are required for their recognition and removal by the spliceosome [10]. Although most RSIs contain branch point sequences and polypyrimidine tracts involved in the splicing mechanism, the sequences of RSIs are usually not conserved [9,10,13]. Since their discovery, the origin of introns has remained a mystery to molecular biologists. Evolution of the eukaryotic gene structure seems to have been predominated by intron loss [4] and the theoretically calculated intron gain rates cannot explain the large number of introns in present-day eukaryotic genomes [14]. However, extensive recent intron gains have been reported in the microcrustacean *Daphnia pulex* [5], the fungus *Mycosphaerella graminicola* [6], and possibly in the green alga *Micromonas pusilla* [8] and the urochordate *Oikopleura dioica* [15]. These reports suggest that the number of introns in a given genome is not only subject to losses but also to substantial gains. In this study, we report on a particular type of spliceosomal introns that show a high level of sequence similarity and some reminiscence of introner elements found in *Micromonas* [8]. These so-called introner-like elements (ILEs) likely originate from multiplication of a discreet number of ancestral elements that are present in related fungal genomes. Although they are typical spliceosomal introns, ILEs are significantly longer and predicted to fold into more stable secondary structures than RSIs. Rigorous intron gain analyses in six fungal species revealed that the vast majority of gained introns are ILEs. By analyzing closely related fungi that diverged less than 100,000 years ago, we could show that the majority of newborn ILEs rapidly degenerate in length, sequence, and stability to become indistinguishable from RSIs. We propose that ILEs are the predecessors of RSIs in these six fungal species. This multiplication mechanism might also be involved in intron gains in other fungi and possibly in ancestral eukaryotic lineages.

Identification of Near-Identical Introns in Six Fungal Genomes

Using a simple BlastN-based method, numerous introns with near-identical sequences could be identified in the intronome of the *Dothideomycete* fungus *Cladosporium fulvum*. Our analysis thoroughly excluded repeated sequences originating from repetitive elements, segmental duplications, or recombinant genes. This observation corroborates recent findings of near-identical intronic sequences in the marine picoeukaryote *Micromonas* [8], the tunicate *Oikopleura dioica* [15], and the *Dothideomycete* fungus *M. graminicola* [6]. In *Micromonas*, such repeat sequences are located within introns and were called introner elements [8]. The analysis was extended to the intronomes of 22 additional Ascomycete fungi and one Basidiomycete. The use of hidden Markov models (HMMs) corresponding to near-identical introns resulted in the identification of 45 to 538 near-identical introns in six different fungi: *C. fulvum*, *Dothistroma septosporum*, *M. graminicola*, *Mycosphaerella fijiensis*, *Hysterium pulicare*, and *Stagonospora nodorum*. In each fungus, these introns could be grouped into one to eight different clusters, each likely originating from a single ancestral element that had been multiplied. Each cluster contains between 10 and 180 members. Thereafter, they will be called ILEs to distinguish them from introner elements found in *Micromonas* [8]. According to expressed sequence tag (EST) support in the different species, ILEs are introns that are spliced out (see Table S1 available online). Although the available EST data for some fungal species is limited, EST support for ILEs is slightly higher than for the entire intronome (Table S1).

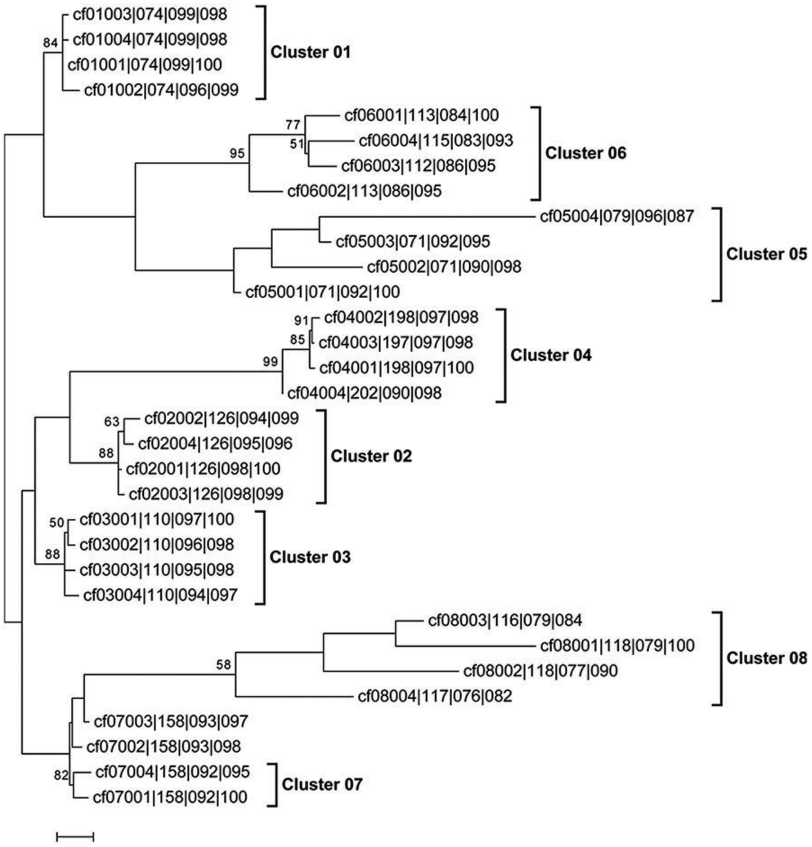


Figure 1. Introner-like Elements Originate from Common Ancestor Elements

The four ILEs with highest HMM expect similarity of each *Cladosporium fulvum* cluster were aligned to construct a maximum likelihood phylogenetic tree. Identifiers contain ILE number|length (nt)|pairwise identity (%)|HMM expect similarity (%). ILE cluster number is indicated next to brackets. The midpoint rooting method was used to estimate the root of the tree. Only bootstrap values over 50 are shown. Scale indicates 0.1 substitutions per site. See also Figure S1 and Table S2.

ILE Clusters Share Common Origins in Several Fungal Species

Phylogenetic analyses for each species showed that most ILE clusters are monophyletic clades, indicating that all elements of a given cluster share the same origin Figure 1; Figure S1). A similarity matrix suggests that most ILE clusters in *M. graminicola* are related to each other and that clusters mf04 and mf05 in *M. fijiensis* are related to cluster cf01 in *C. fulvum* (Figure S1). In addition, the closely related fungi *C. fulvum* and *D. septosporum* share three ILE clusters (cf04/ ds03, cf06/ds02, and cf08/ds04). These results indicate that ILE clusters present in different fungal species originate from the multiplication of a single ancestral element that was present before species divergence. Using HMMs to search the intronomes of closely related fungi not only showed that some of the identified ILE clusters contain additional members (mf04/mf05, cf05, cf06/ds02, mg05) but also revealed initially not identified shared clusters (Table S2). The intronomes of *D. septosporum* and *Septoria musiva* seem to contain less-conserved ILEs belonging to clusters cf07 and cf01/mf04/ mf05, respectively. These results argue for the presence of ancestral ILEs that cannot easily be detected in other fungal intronomes.

ILEs Harbor All Hallmark Features of Spliceosomal Introns

We characterized ILEs in more details to distinguish them from RSIs. Sequence analysis of all ILEs showed that they are genuine spliceosomal introns of which 99% contain canonical acceptor and donor sites (a representative example is given for *C. fulvum* in Figure 2A). In addition, 96% of all ILEs contain a predicted CURAY branch point sequence, 95% contain a 5´ polypyrimidine tract, and 76% contain both 5´ and 3´ polypyrimidine tracts, frequencies that are higher than those reported for fungal RSIs [9]. Introns that lack an identifiable branch point sequence likely contain another noncanonical sequence that can be recruited as a branch point. Indeed, many introns lacking an identifiable branch point sequence have EST support for proper splicing (Table S1). Similar to RSIs [16], ILE clusters with the highest number of members show a preference for insertion in AG/GY sites and in phase 0 of coding sequences. They do, however, show a slight bias for being present in the center of genes in contrast to RSIs, which are more frequently found at the 5´ end of genes (Figure 2B; Figure S2). The biased location of RSIs was reportedly due to intron losses that primarily occur at the 3´ end of genes [17]. Altogether, these hallmark features suggest that ILEs are model RSIs and, more importantly, they can be perfectly spliced by the spliceosome immediately after multiplication and insertion in a new location.

ILEs Are Predicted to be More Stable than RSIs but Are Also Prone to Degeneration

Although ILEs are model spliceosomal introns, they also have particular features that distinguish them from RSIs. Indeed, ILEs are longer than RSIs and show different length distributions with multiple peaks that correspond to the optimum lengths of different clusters (Figure 2C; Figure S2). In a given cluster, ILEs with the lowest identity are predominantly those showing sequence deletions (Figure S3). Further analyses confirmed a positive correlation between pairwise identity and length of all ILEs (Figure 3A). Substitutions and minor insertions are observed over ILEs' full length, but deletions occur less often around the conserved branch point sequence at the 3^0 end (Figure 3B). This bias could be the result of selection pressure to retain splicing features. These observations suggest that ILEs may lose their ability to multiply due to degeneration in length and sequence. Although RSIs have some secondary structures that can

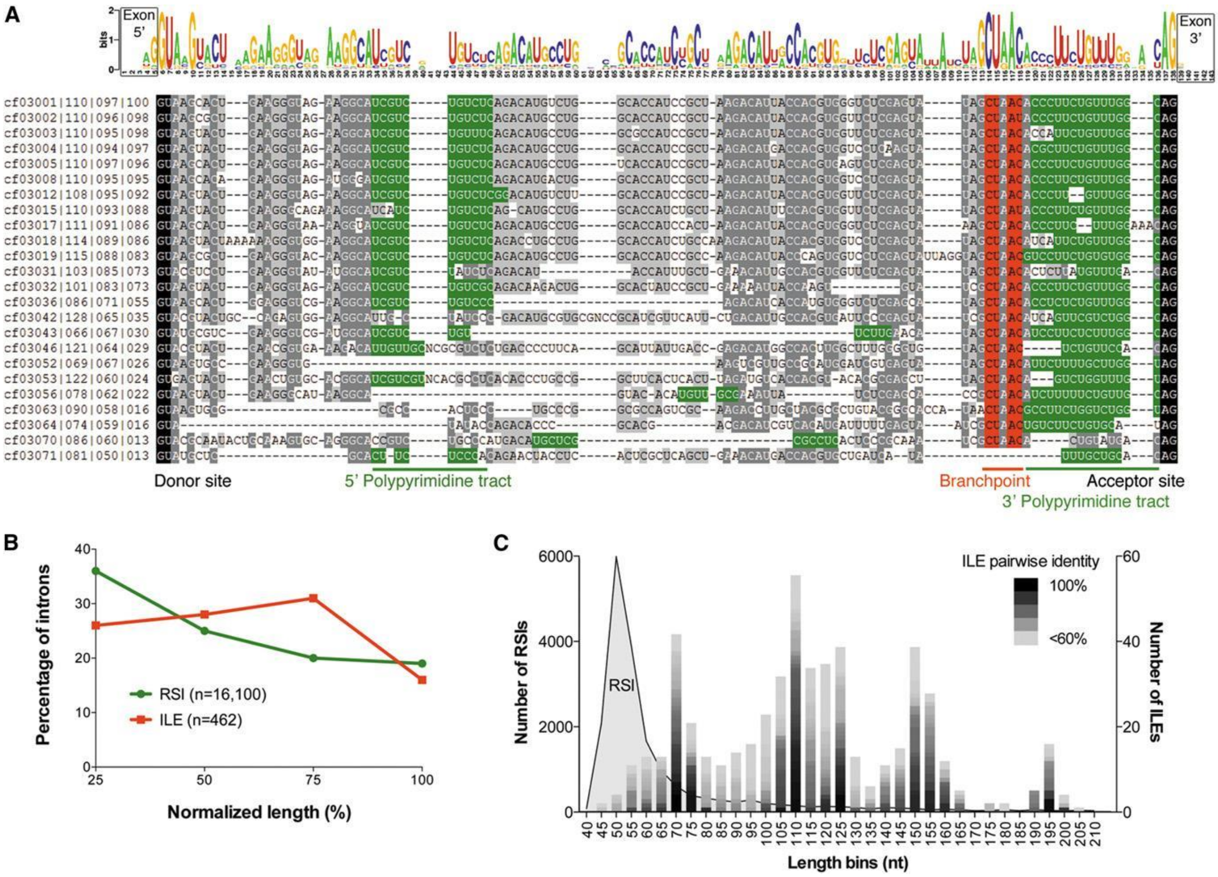


Figure 2. Characteristics of Introner-like Elements

Each panel presents representative results obtained for ILEs from *Cladosporium fulvum*. Similar results were obtained for the five other fungal species. (A) Alignment and consensus sequence of selected ILEs. Identifiers are as shown in Figure 1. Colors show splicing elements typical of RSIs. (B) Distribution of RSIs and ILEs within *C. fulvum* genes. Genes were divided in four sections expressed as percentages. Number of RSIs and ILEs in each section was counted and expressed as a percentage of all RSIs and ILEs, respectively. (C) Length distribution of RSIs and ILEs from *C. fulvum*. Scale numbers on the x axis indicate the shortest length of 5 nt bins. See also Figure S2.

facilitate splicing [12], the lower predicted Gibbs free energy (ΔG) of ILEs suggests a significant greater molecular stability (Figure 3C). Remarkably, the increase in ΔG values of ILEs correlates with their degree of degeneration, until ΔG values of RSIs are eventually reached (Figures 3A and 3C). The low ΔG values of ILEs are explained by their predicted alignment-based secondary structures that often consist of three stem-loops containing many G-U pairs (Figures 4A and 4B). Consistent with the similarity matrix analysis, structure predictions for all *M. graminicola* ILE clusters suggest that they all have a common structure due to stretches of identical nucleotides (Figure 4B; Figure S4). Moreover, alignments of related ILEs revealed many compensatory mutations that conserve hairpin structures. Together, these observations argue for evolutionary constraints that preserve ILE secondary structures. We propose that they are an important feature because such predicted stable secondary structures are known for noncoding RNAs with specific functions [18]. Overall, our results strongly suggest that highly structured ILEs are likely mobile, but they are prone to degenerate mainly through deletions. They seem to gradually evolve to become RSIs that lost the ability to multiply and lack conserved predicted secondary structures. Thus, we hypothesize that RSIs might originate from ILEs.

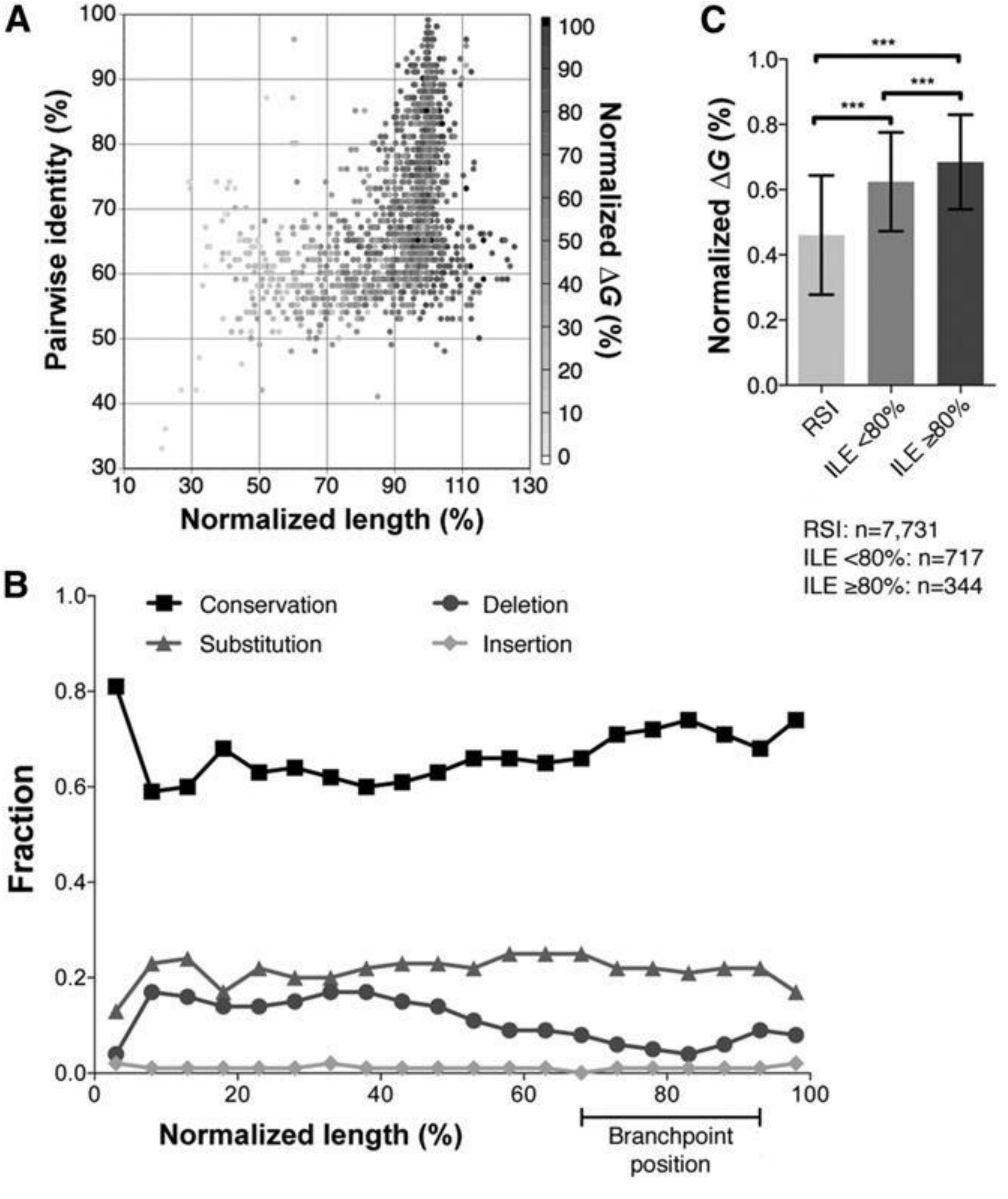


Figure 3. Degeneration of Introner-like Elements

(A) Correlation between pairwise identity and normalized length of all identified fungal ILEs. The gray scale indicates normalized Gibbs free energy (ΔG). (B) Conserved, substituted, deleted, and inserted nucleotide positions within ILEs. ILE length was expressed as percentage and arranged in 5% bins. Numbers of conserved positions, substitutions, deletions, and insertions were counted in each bin and expressed as a fraction. Distribution is shown for all ILEs from *Cladosporium fulvum*, *Dothistroma septosporum*, *Mycosphaerella graminicola*, and Mycosphaerella fijiensis. (C) Mean and SD of normalized ΔG values of all RSIs, ILEs with < 80%, and ILEs with R 80% pairwise identity from all six fungi. A non-parametric Kruskall-Wallis test was carried out (***p < 0.0001), followed by a Dunn's pairwise comparison test at a = 0.05 significance level. See also Figure S3.

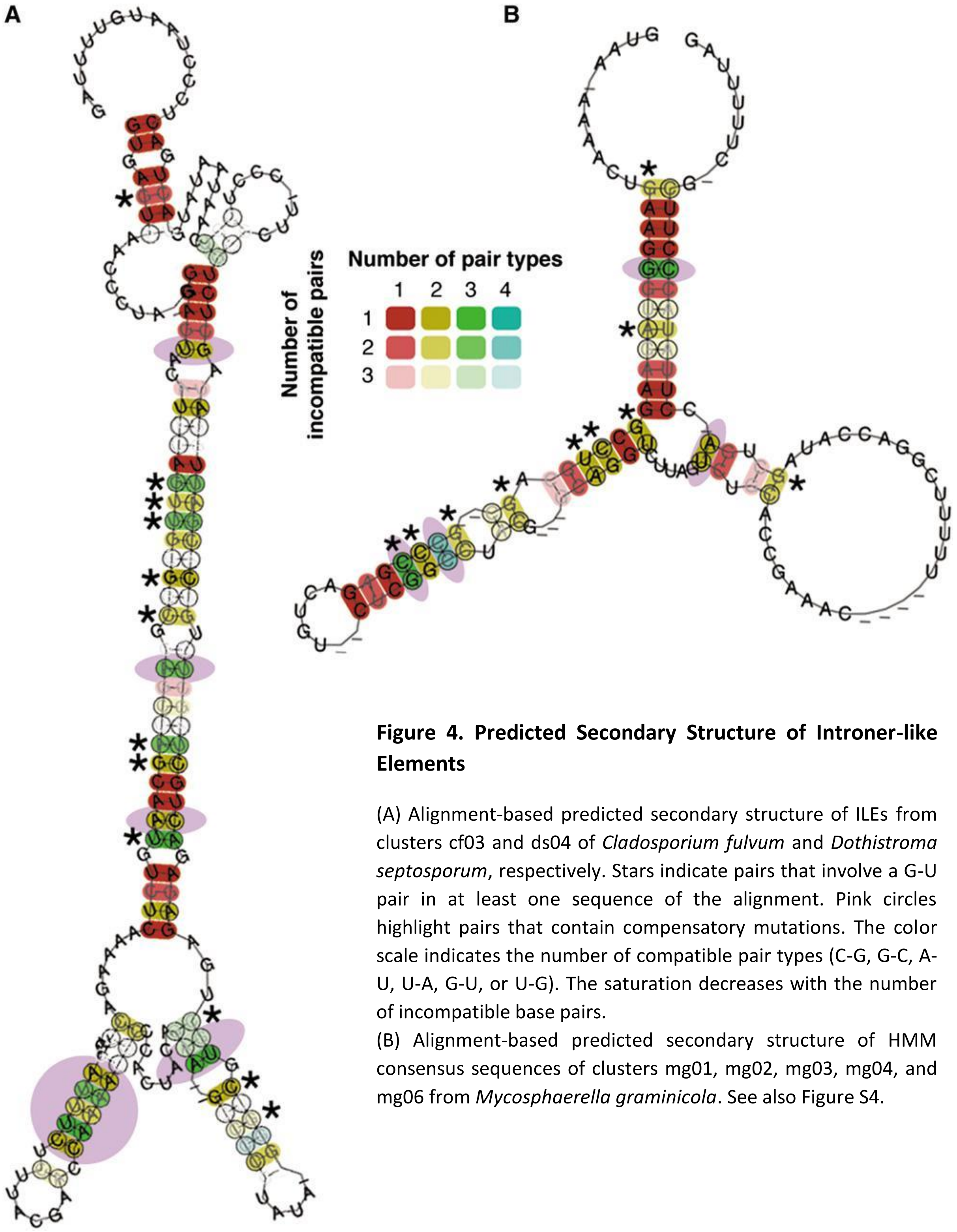


Figure 4. Predicted Secondary Structure of Introner-like Elements

(A) Alignment-based predicted secondary structure of ILEs from clusters cf03 and ds04 of *Cladosporium fulvum* and *Dothistroma septosporum*, respectively. Stars indicate pairs that involve a G-U pair in at least one sequence of the alignment. Pink circles highlight pairs that contain compensatory mutations. The color scale indicates the number of compatible pair types (C-G, G-C, A-U, U-A, G-U, or U-G). The saturation decreases with the number of incompatible base pairs.

(B) Alignment-based predicted secondary structure of HMM consensus sequences of clusters mg01, mg02, mg03, mg04, and mg06 from *Mycosphaerella graminicola*. See also Figure S4.

ILEs Account for the Vast Majority of Intron Gains in Six Fungi

The proposed hypothesis for the origin of RSIs implies that recent intron gains in fungi are mainly the result of ILE multiplication. Previous studies suggested that intron losses prevail over intron gains in most eukaryotic lineages [4], although several extensive gains have been reported as well [5,6]. The balanced rates estimated in fungi are consistent with the presence of active ILEs [4, 14, 17]. By using up to six nodes between *C. fulvum*/*D. septosporum* and the most distant outgroup, the Basidiomycete Cryptococcus neoformans, single intron gains in Dothideomycetes could confidently be assigned as intron presence in only one of all species included in this study

(Figure 5A). Thirteen single gains were identified in both *C. fulvum* and *D. septosporum* when using the maximum number of outgroups, which allowed inspection of 951 orthologs only. As outgroups are removed and more orthologs become available for inspection, the number of single gains increased to 199 (Table S3). Strikingly, on average, 50% of single gains originate from ILEs, irrespective of the number of outgroups used in the analysis (Figure 5B). It is noteworthy that 75% to 90% of ILEs present in a set of orthologs are associated with gains in both *C. fulvum* and *D. septosporum* (Table S3). Up to 351 single gains were inferred in their common ancestor, of which ILEs only represent 10%. Similar numbers of single gains were not only found in *M. graminicola* and *M. fijiensis* but also in *S. musiva* that only contains highly degenerated ILEs (Table S3). From these results, we conclude that due to ILE degeneration, it is essential to compare species with shorter evolutionary distances in order to confidently estimate the contribution of ILEs to intron gains.

Newborn ILEs Become Undistinguishable from RSIs within 100,000 Years

The genomes of *M. graminicola* (three isolates), two sister species S1 and S2 (five and four isolates, respectively), and Septoria passerinii (one isolate) were used because divergence between *M. graminicola*, S1 and S2 was dated to 11 and 22.3 thousand years, respectively [19]. *C. fulvum* and *D. septosporum* were included as an outgroup and, in contrast to the recent report on extensive intron gains in *M. graminicola* [6], polymorphic positions that mainly represent segregating introns within populations were excluded from the analysis. This new data set confirmed that 50% of the single gains in *C. fulvum* and *D. septosporum* originate from ILEs (Figure 5C). ILE contribution to intron gains reaches 90% in *M. graminicola* and S2, which diverged more recently than *C. fulvum* and *D. septosporum*, but only 40% of the single gains in S. passerinii and the common ancestor of *M. graminicola*, S1, and S2, could be ascribed to distinguishable ILEs (Figure 5C; Table S4). ILE contribution to single gains even drops to 6%– 10% in older ancestors. These observations support the hypothesis that many or even potentially all RSIs are degenerated ILEs. Rough dating of the species divergences suggests that extensive intron gains have occurred in all these species during the last 100 thousand years. The dating also shows that all six *M. graminicola* ILE clusters must have multiplied more than 22 thousand years ago, but members of clusters mg05 and mg06 are no longer active (Figure 5D). Certainly, degenerated ILE clusters identified in *D. septosporum* and *S. musiva* lost their ability to multiply a long time ago. Additionally, over this short time frame, average pairwise identity and average length of ILEs have decreased (Figure 5E), showing that ILE clusters can emerge and successfully multiply, but multiplication may also stop due to rapid degeneration. As such, ILE identification becomes difficult in organisms with short generation times in which the last multiplication event has occurred more than 100 thousand years (according to the dated species tree).

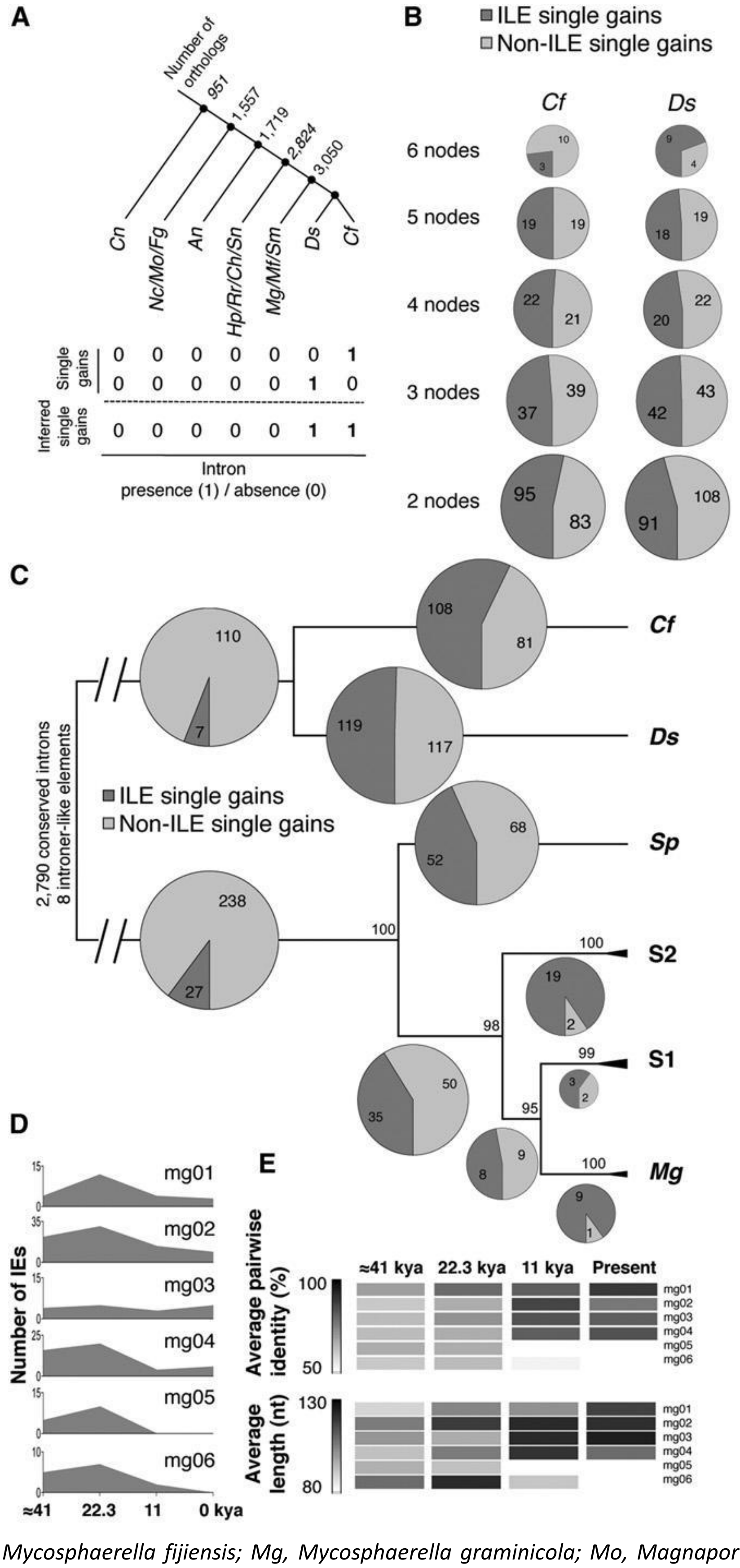


Figure 5. Contribution of Introner-like Elements to Single Intron Gains

(A) Species tree [21] and pattern used to assign single intron gains in Dothideomycetes.

(B) Contribution of ILEs and non-ILEs to single gains in *Cf* and *Ds* when including outgroups of 2 to 6 nodes. Size of the pies is proportional to the total number of single gains.

(C) Contribution of ILEs to single gains in *Mg*, S1, S2, *Sp*, *Cf*, and *Ds*. Inferred gains in the oldest ancestors were determined using other Dothideomycetes as outgroup. Size of the pies is proportional to the total number of single gains. The species tree was constructed using four genes (*TUB1*, *EIF3b*, *PAP1*, and *RPS9*). Numbers at the nodes indicate bootstrap values of 500 replicates. The tree was rooted according to Wang et al. [21]. The molecular clock was calibrated using the dating of S2 speciation [19].

(D) Number of ILEs per *Mg* cluster (mg01 to mg06) that are conserved in all species at each dated node. The dating on the x axis corresponds to the age of the nodes determined in the phylogenetic tree. Kya, thousand years ago.

(E) The average pairwise identity and average length of conserved ILEs is indicated per cluster for each dated node of the phylogenetic tree. *An, Aspergillus nidulans; Cf, Cladosporium fulvum; Ch, Cochliobolus heterostrophus; Cn, Cryptococcus neoformans; Ds, Dothistroma septosporum; Fg, Fusarium graminearum; Hp, Hysterium pulicare; Mf, Mycosphaerella fijiensis; Mg, Mycosphaerella graminicola; Mo, Magnaporthe oryzae; Nc, Neurospora crassa; Rr, Rhytidhysteron rufulum; Sm, Septoria musiva; Sn, Stagonospora nodorum; Sp, Septoria passerini*. See also Tables S3 and S4.

Conclusions

Based on this study, we conclude that ILEs account for most of the intron gains in at least six fungal species. Several mechanisms have been proposed for intron gains, from intron transposition to intronization of exons [7]. Most, however, are supported by few experimental data [7] and none can convincingly explain ILE multiplication. The contribution of these mechanisms to intron gains appears limited in comparison to the ILE multiplication reported here. A recently proposed mechanism involves reverse splicing of introns directly into the genome followed by reverse transcription [3]. This hypothetical mechanism requires fewer steps than the hitherto accepted mechanism of intron transposition but requires RNA-DNA hybridization. This mechanism could apply to ILEs because new ILEs perfectly insert into genes (no sequence deletion or duplication), and it would not involve homologous recombination. Such a step is required in the other proposed mechanisms although it occurs at low frequency in filamentous fungi [20]. ILEs multiplication might be linked to transcription because they are always found on the coding strand of genes as reported for introner elements in *Micromonas* [8].

Our findings show that introns in fungi are highly dynamic because only 51% of introns are conserved from S. passerinii to *C. fulvum* (Figure 5C). Because up to 90% of intron gains in the fungal species included in this study originate from ILEs, we propose that ILEs, too degenerated for detection, represent the species-specific introns in other fungi. Introner elements in *Micromonas* extensively multiplied in thousands of copies, which suggests that intron multiplication could also occur outside the fungal kingdom [8]. However, introner elements differ from ILEs because they lack predicted stable secondary structures. We could not find ILEs in intronomes of other eukaryotic lineages that contain near-identical introns such as *O. dioica* [15]. However, we speculate that active ILEs might have appeared very early in eukaryotes evolution but have become indistinguishable from RSIs. It is also possible that similar ILEs are currently spreading in other not yet studied eukaryotes. Further studies on ILEs are required to increase our understanding of eukaryotic gene structure evolution.

Supplemental Information

Supplemental Information includes four figures, four tables, Supplemental Raw Data, and Supplemental Experimental Procedures and can be found with this article online at doi:10.1016/j.cub.2012.05.011.

Acknowledgments

We thank Jan van Kan, Bart Thomma, and Scott Griffiths for critical reading of the manuscript. HMM profiles and aligned intron windows used in the gain analyses are available on request.

References

1. Koonin, E.V. (2006). The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Biol. Direct 1, 22.
2. Rodrı´guez-Trelles, F., Tarrı´o, R., and Ayala, F.J. (2006). Origins and evolution of spliceosomal introns. Annu. Rev. Genet. 40, 47–76.
3. Roy, S.W., and Irimia, M. (2009). Mystery of intron gain: new data and new models. Trends Genet. 25, 67–73.
4. Csuros, M., Rogozin, I.B., and Koonin, E.V. (2011). A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. PLoS Comput. Biol. 7, e1002150.
5. Li, W., Tucker, A.E., Sung, W., Thomas, W.K., and Lynch, M. (2009). Extensive, recent intron gains in Daphnia populations. Science 326, 1260–1262.
6. Torriani, S.F., Stukenbrock, E.H., Brunner, P.C., McDonald, B.A., and Croll, D. (2011). Evidence for extensive recent intron transposition in closely related fungi. Curr. Biol. 21, 2017–2022.
7. Yenerall, P., Krupa, B., and Zhou, L. (2011). Mechanisms of intron gain and loss in Drosophila. BMC Evol. Biol. 11, 364.
8. Worden, A.Z., Lee, J.H., Mock, T., Rouze´, P., Simmons, M.P., Aerts, A.L., Allen, A.E., Cuvelier, M.L., Derelle, E., Everett, M.V., et al. (2009). Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. Science 324, 268–272.
9. Kupfer, D.M., Drabenstot, S.D., Buchanan, K.L., Lai, H., Zhu, H., Dyer, D.W., Roe, B.A., and Murphy, J.W. (2004). Introns and splicing elements of five diverse fungi. Eukaryot. Cell 3, 1088–1100.
10. Schwartz, S.H., Silva, J., Burstein, D., Pupko, T., Eyras, E., and Ast, G. (2008). Large-scale comparative analysis of splicing signals and their corresponding splicing factors in eukaryotes. Genome Res. 18, 88–103.
11. Le Hir, H., Nott, A., and Moore, M.J. (2003). How introns influence and enhance eukaryotic gene expression. Trends Biochem. Sci. 28, 215–220.
12. Warf, M.B., and Berglund, J.A. (2010). Role of RNA structure in regulating pre-mRNA splicing. Trends Biochem. Sci. 35, 169–178.
13. Roy, S.W., and Irimia, M. (2009). Splicing in the eukaryotic ancestor: form, function and dysfunction. Trends Ecol. Evol. (Amst.) 24, 447–455.
14. Roy, S.W., and Gilbert, W. (2005). Rates of intron loss and gain: implications for early eukaryotic evolution. Proc. Natl. Acad. Sci. USA 102, 5773–5778.
15. Denoeud, F., Henriet, S., Mungpakdee, S., Aury, J.M., Da Silva, C., Brinkmann, H., Mikhaleva, J., Olsen, L.C., Jubin, C., Can˜estro, C., et al. (2010). Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. Science 330, 1381–1385.
16. Qiu, W.G., Schisler, N., and Stoltzfus, A. (2004). The evolutionary gain of spliceosomal introns: sequence and phase preferences. Mol. Biol. Evol. 21, 1252–1263.
17. Nielsen, C.B., Friedman, B., Birren, B., Burge, C.B., and Galagan, J.E. (2004). Patterns of intron gain and loss in fungi. PLoS Biol. 2, e422.
18. Mathews, D.H., Moss, W.N., and Turner, D.H. (2010). Folding and finding RNA secondary structure. Cold Spring Harb Perspect Biol 2, a003665.
19. Stukenbrock, E.H., Bataillon, T., Dutheil, J.Y., Hansen, T.T., Li, R., Zala, M., McDonald, B.A., Wang, J., and Schierup, M.H. (2011). The making of a new pathogen: insights from comparative population genomics of the domesticated wheat pathogen *Mycosphaerella graminicola* and its wild sister species. Genome Res. 21, 2157–2166.
20. Weld, R.J., Plummer, K.M., Carpenter, M.A., and Ridgway, H.J. (2006). Approaches to functional genomics in filamentous fungi. Cell Res. 16, 31–44.
21. Wang, H., Xu, Z., Gao, L., and Hao, B. (2009). A fungal phylogeny based on 82 complete genomes using the composition vector method. BMC Evol. Biol. 9, 195.

Chapter 6

At the origin of spliceosomal introns: Is multiplication of introner-like elements the main mechanism of intron gain in fungi?

Collemare J, van der Burgt A, de Wit PJGM.

Communicative & Integrative Biology 2013, 6(2):e23147.

doi: 10.4161/cib.23147

This article is available from:
http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3609843/pdf/cib-6-e23147.pdf

Summary

The recent discovery of introner-like elements (ILEs) in six fungal species shed new light on the origin of regular spliceosomal introns (RSIs) and the mechanism of intron gains. These novel spliceosomal introns are found in hundreds of copies, are longer than RSIs and harbor stable predicted secondary structures. Yet, they are prone to degeneration in sequence and length to become undistinguishable from RSIs, suggesting that ILEs are predecessors of most RSIs. In most fungi, other near-identical introns were found duplicated in lower numbers in the same gene or in unrelated genes, indicating that intron duplication is a widespread phenomenon. However, ILEs are associated with the majority of intron gains, suggesting that the other types of duplication are of minor importance to the overall gains of introns. Our data support the hypothesis that ILEs' multiplication corresponds to the main mechanism of intron gain in fungi.

The Proposed Mechanisms for Intron Gain Cannot Explain the High Intron Density in Present Day Eukaryotic Genomes

Eukaryotic genes consist of exons that contain the coding sequence, and of introns that are non-coding and are removed from premature mRNA after transcription. The spliceosome machinery, a large ribonucleoprotein that recognizes specific intronic features, catalyzes two consecutive transesterification reactions that result in splicing of the nuclear introns and ligation of adjacent exons [1]. Such a mosaic gene structure is certainly one of the most important features that allowed the appearance of complex organisms during evolution of higher Eukaryotes [2]. Indeed, land plants and animals, including humans, have intron-rich genomes (> 3 introns per kb coding sequence) as compared with more simple organisms such as most fungi (< 3 introns per kb coding sequence) [3],[4]. Yet, more than 30 y after their discovery, the origin of spliceosomal introns is still unknown. Analyses of gain and loss of introns in diverse eukaryotic lineages kept the mystery on introns' origin alive because there was less evidence for gains as compared with losses [4],[5]. In many Eukaryotes, the estimated rates for intron gain and loss cannot explain the high intron density in many present-day genomes. Indeed, a higher intron loss rate would ultimately result in the disappearance of spliceosomal introns. However, some lineages such as fungi have experienced more balanced rates of intron gains and losses [6-7], suggesting that intron gains can still occur to a large extent in present days. In addition to fungi [6-9], extensive recent intron gains have been reported in the micro-crustacean *Daphnia pulex* [10].

Several mechanisms have been proposed for intron gains and have been recently reviewed in detail [11]. The model that has received most support in the scientific community is referred to as intron transposition. It involves reverse splicing of a spliced intron into the mRNA of another gene, followed by reverse transcription and homologous recombination at the gene locus. This model is almost identical to the main mechanism proposed for intron loss by reverse transcription and homologous recombination after intron splicing [11-12]. Observations of intron losses occurring more frequently at the 3' end of the genes support this mechanism [6,12,13]. However, according to these models, the difference in rates of intron gain and loss solely depends on the rate of reverse splicing, which is expected to occur at low frequency [14]. Thus, the balanced rates of intron gain and loss in certain lineages challenge the intron transposition model. Roy and Irimia proposed two new models to resolve this paradox: spliceosomal retrohoming (reverse splicing of an intron directly into DNA followed by reverse transcription) and template switching during reverse transcription [14]. Other mechanisms have also been suggested including: (1) recombination between two paralogs, one containing an intron and the other one intronless (intron transfer); (2) insertion of a transposable element followed by conversion to an intron; (3) intronization of an exon by acquisition of splicing sites; (4) mobilisation and propagation of a self-splicing group II intron from an organelle into the nucleus; (5) insertion during DNA double-strand breaks repair; and finally (6) duplication of a genomic segment that contains cryptic splicing sites [11]. However, only the last mechanism has been experimentally proven [15]. All the other models, including intron transposition, only rely on indirect evidence and fail to describe how the vast majority of introns were gained [11]. It is likely that all proposed mechanisms contribute to intron gains to some extent, but the frequencies at which they occur cannot explain the high number of introns present in numerous Eukaryotes. Therefore, it has been suggested that the mechanism of intron gain in ancestral lineages might differ from those that occur in modern Eukaryotes [5].

Intron Duplication is a Widespread Phenomenon in Fungi

A striking observation in the animal *Oikopleura dioica* [16] and in the alga *Micromonas pusilla* [17] was the presence of introns that are nearly identical at the sequence level. In *M. pusilla*, these near-identical introns are present in thousands of copies and were named introner elements (IE). Near-identical introns were also reported to occur in the fungus *Mycosphaerella graminicola* [8]. Recently, we reported on the occurrence of near-identical introns in five additional fungal species, where they are present in up to five hundred copies [9]. We named these high-copy introns introner-like elements (ILE) to refer to IEs found in *M. pusilla*. Like regular spliceosomal introns (RSIs), ILEs have typical splicing features including canonical acceptor and donor sites, branch point sequence and polypyrimidine tracts, which suggest that they can be spliced by the spliceosome machinery. However, in addition to being present in many near-identical copies, we also found that ILEs have features completely different from RSIs. They are significantly longer and have lower predicted Gibbs free energy (Δ*G*) values that were ascribed to stable predicted secondary structures. A robust gain analysis showed that up to 90% of gained introns are ILEs. Because our data showed that ILEs quickly degenerate in length and sequence to become undistinguishable from RSIs, we hypothesized that non-ILE-associated gains are highly degenerated ILEs. Thus, most RSIs might originate from ILEs in at least six fungal species [9].

In this study, the very first step of the pipeline that was developed to identify ILEs involved a simple BlastN search and clustering method, which retrieved three different types of near-identical introns [9]. Depending on the number of introns with a near-identical sequence and whether they were duplicated within the same gene or in different genes, these multi-copy introns were classified as same gene duplications (SGD; 82 members), low-copy introns (LCI; 302 members) and high-copy introns (1226 members) that were subsequently named ILEs. This search revealed that intron duplication is a widespread phenomenon in fungi because it was found in all species included in the study except *Aspergillus nidulans* (**Table 1**). However, the contribution of each category to the observed duplication events varies. Nine species contain only LCIs, while both SGDs and LCIs are found in five other species. In the latter, SGDs occur less frequently and contribute to 25–54% of the observed duplications (**Table 1**). The remaining six fungal species have all three types of duplicated introns, but they also have a very high number of ILEs (24 to 377), which contribute between 60% and 92% to all duplication events (**Table 1**). Noteworthy, *Rhytidhysteron rufulum*, *Fusarium graminearum* and *Sclerotinia sclerotiorum* contain near-identical introns in high numbers but they correspond to repetitive elements that inserted within RSIs and were not retrieved as ILEs in the subsequent and more stringent steps of ILE identification (**Table 1**) [9].

Table 1. Identification of multi-copy introns in 24 fungal species

Fungal species	Total	SGD[a]	LCI[a]	ILE[a]
Cladosporium fulvum	408	3 (1)	28 (7)	377 (92)
Mycosphaerella graminicola	344	16 (5)	22 (6)	306 (89)
Dothistroma septosporum	322	7 (2)	17 (5)	298 (93)
Hysterium pulicare	188	16 (9)	28 (15)	144 (77)
Mycosphaerella fijiensis	97	14 (14)	22 (23)	61 (63)
Stagonospora nodorum	40	0	16 (40)	24 (60)
Fusarium oxysporum	37	0	37 (100)	0
Coccidioides immitis	24	6 (25)	18 (75)	0
Histoplasma capsulatum	18	0	18 (100)	0
Rhytidhysteron rufulum	17	5 (29)	8 (47)	4[b] (24)
Leptosphaeria maculans	13	0	13 (100)	0
Septoria musiva	13	4 (31)	9 (69)	0
Nectria haematococca	13	7 (54)	6 (46)	0
Fusarium graminearum	12	0	2 (17)	10[b] (83)
Cryptococcus neoformans	12	0	12 (100)	0
Sclerotinia sclerotiorum	10	0	8 (80)	2[b] (20)
Cochliobolus heterostrophus	8	0	8 (100)	0
Botrytis cinerea	8	2 (25)	6 (75)	0
Neurospora crassa	6	0	6 (100)	0
Trichoderma atroviridae	6	0	6 (100)	0
Verticillium albo-atrum	6	0	6 (100)	0
Magnaporthe oryzae	6	2 (33)	4 (67)	0
Verticillium dahliae	2	0	2 (100)	0
Aspergillus nidulans	0	0	0	0
Total	**1610**	**82 (5)**	**302 (19)**	**1226 (76)**

For each intron of a given fungal species, a BlastN analysis was performed using the complete intronome. Then, intron clusters were built by grouping a given intron with its near-identical introns. Introns that were duplicated only within the same gene were classified as same gene duplications (SGD). Near-identical introns found in unrelated genes were classified as low-copy introns (LCI) when a search using hidden Markov models did not increase the number of members by more than 2-fold; they were classified as high-copy introns when this search increased the number of members by more than 2-fold. These high-copy introns were subsequently named introner-like elements (ILE) [9]. [a] Number of introns. Contribution of a duplication type to the total number of duplications is indicated as percentage in brackets. [b] These high-copy introns were not retrieved as ILEs by additional more stringent analyses.

As was done in our previous study on ILEs, the length and stability of the two other types of near-identical introns were measured. The median length of SGDs and LCIs are in the same range as observed for non-duplicated introns (NDI), but ILEs are about twice as long (**Fig. 1A**). The Δ*G* free energy of SGDs and LCIs is not different from that of NDIs, while ILEs have a significantly lower Δ*G* (**Fig. 1B**). These results suggest that different mechanisms might be involved in the duplication of each intron type. SGDs are found in only 11 fungal species and are limited in number (maximum of 16 members in a given species). Fifty percent of these duplication events represent segmental duplication within the same gene because exon sequences on each side of these introns are also duplicated. The other 50% might represent intron transpositions within the same transcript or intron transfers between paralogs. Comparable low numbers were also reported in *Caenorhabditis elegans* in which only three gained introns are SGDs [18]. In *C. neoformans*, a single gene with several putative SGDs was also shown to be most likely the result of a duplication of exonic repeats [19]. The two other types of multi-copy introns are found in different unrelated genes, suggesting that they may represent the same type of introns, but differ in multiplication frequency. They have different characteristics (length and Δ*G*), which suggests that different duplication mechanisms are involved. However, these differences are also consistent with ILE degeneration and LCIs might represent degenerated ILEs. This hypothesis might explain why we could not identify more introns that would have originated from them. Alternatively, LCIs could originate from a low frequency transposition mechanism. Altogether, our results suggest that ILEs are prevailing duplication events in fungi, explaining on average 76% of intron duplications.

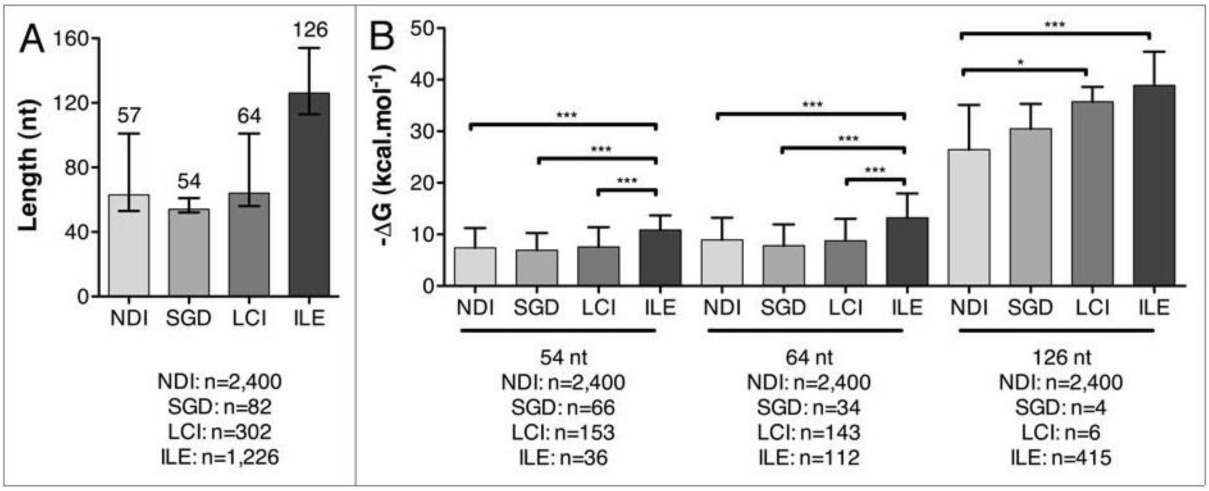


Figure 1. Length and stability of the different types of duplicated introns.

The length and predicted Gibbs free energy (Δ*G*) were measured for non-duplicated intron (NDI), same gene duplications (SGD), low-copy introns (LCI) and introner-like elements (ILE) from 24 fungal species included in this study [9]. (**A**) Median length and interquartile range are plotted for each type of intron. The median length is indicated above the bars. (**B**) Mean and SD of Δ*G* values of introns with a length corresponding to the median of each type of intron. A non-parametric Kruskall-Wallis test was performed ($p < 0.0001$), followed by a Dunn's pairwise comparison test at $\alpha = 0.05$ significance level. Only significant differences are indicated.

Introner-Like Elements Reconcile the Intron Gain Mechanism in Ancestral and Modern Genomes

Based on the observed degeneration, we speculated that ILEs are at the origin of most RSIs in at least six fungal species, which implies that they should be associated with intron gains. Indeed, ILEs can contribute up to 90% of recent intron gains [9]. An intron gain and loss analysis (IGL) in fungal species that contain ILEs showed that gains occur on average 10-fold more frequently than losses (**Table 2**). Remarkably, this is also true in *Septoria musiva*, a species that carries highly degenerated ILEs only, which initially could not be identified as such [9]. In the IGL analysis shown here, up to 50% of the gains are explained by ILEs, while almost none are explained by SGDs or LCIs (**Table 2**). The non-explained gains certainly correspond to more ancient gained introns that cannot be recognized as ILEs because of the high level of degeneration [9].

Our analysis also revealed that introns absent in other species are similar in length to ancestral introns that are conserved in all fungal species included in this study, although with a much lower standard deviation (**Fig. 2A**). Our findings suggest that the majority of new introns originate from ILEs, which subsequently lose their stable secondary structure and shorten toward the optimal intron length, to eventually be lost (**Fig. 2B**). Accordingly in *Aspergillus* species, it was found that lost introns are significantly shorter than conserved introns [7]. Our proposed model for fungal intron birth, life and death is consistent with the high intron dynamics observed in fungi, but also with lower dynamics in higher Eukaryotes, which is most likely related to the different generation times. Intron-rich genomes usually have longer introns [3], which would hamper their loss.

Table 2. Single intron gain and loss analysis in fungal species containing ILEs

Fungal species	Orthologs	Introns	ILEs	Ancestral intron[a]	Single gain[b]	Single loss[b]	SGD at gain positions[c]	LCI at gain positions[c]	ILE at gain positions[c]
Cladosporium fulvum	3050	3483	110	2209	178	20	0	5 (0.028)	95 (0.534)
Dothistroma septosporum	3050	3516	101	2209	199	10	0	2 (0.010)	91 (0.457)
Septoria musiva	2824	2084	-	906	372	60	1 (0.003)	2 (0.005)	-
Mycosphaerella fijiensis	2824	1951	14	906	236	43	0	1 (0.004)	14 (0.059)
Mycosphaerella graminicola	2824	2240	44	906	388	40	0	1 (0.003)	43 (0.111)

Single gains and single losses were determined using only one outgroup clade for each species as described in our previous report [9]. Contribution of same gene duplications (SGD), low-copy introns (LCI) and introner-like elements (ILE) to single gains was determined. [a] Intron position conserved in all analyzed fungal species; [b] Introns that are present or absent only in the considered species; [c] Numbers in brackets are numbers of SGDs, LCIs or ILEs at single gain positions divided by the number of single gains.

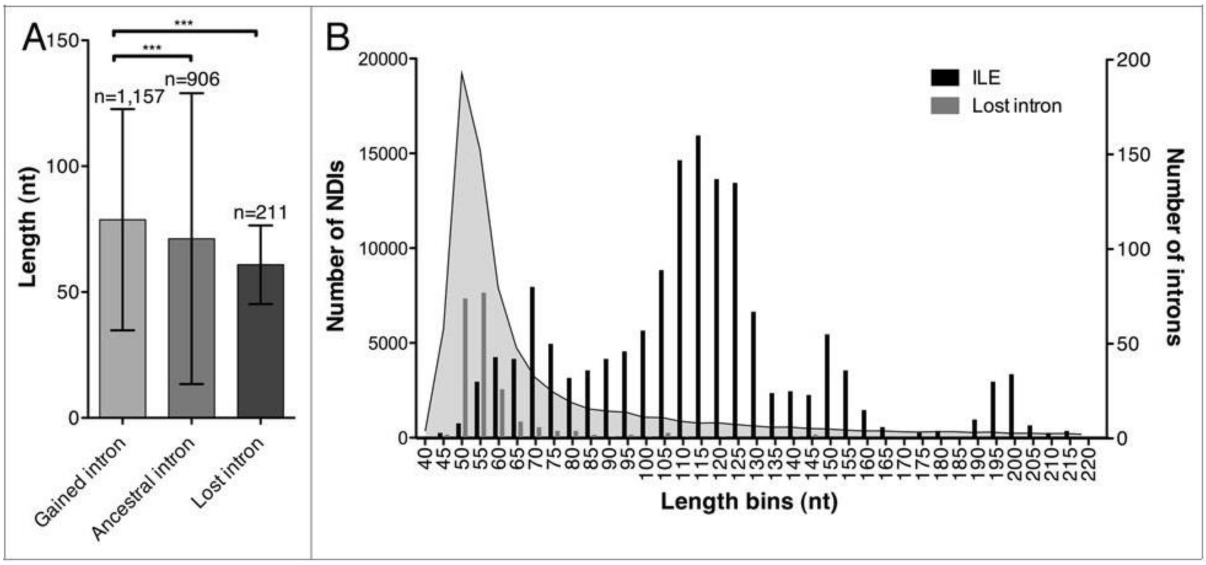


Figure 2. Birth, life and death of spliceosomal introns in fungi.

(**A**) Gained introns are single gains in *Cladosporium fulvum*, *Dothistroma septosporum*, *Mycosphaerella graminicola*, *Mycosphaerella fijiensis* or *Septoria musiva* as determined in **Table 2**. Ancestral introns are conserved among all fungi included in this study. Lost introns are single losses in one of the five fungal species. Length of lost introns that are still present in the other four species was calculated and corrected for outliers using the formula: (sum-max-min)/(length-2). A non-parametric Kruskall-Wallis test was performed ($p < 0.0001$), followed by a Dunn's pairwise comparison test at $\alpha = 0.05$ significance level. Only significant differences are indicated. (**B**) Length distribution of non-duplicated introns (NDIs), introner-like elements (ILEs) and lost introns in the five fungal species listed above.

With the resonance of IEs in *M. pusilla*, it is very likely that genome invasion by introns could have occurred at least once in an ancestral Eukaryotic lineage to give rise to the present-day intron-rich Eukaryotes. This hypothesis suggests that the mechanisms of intron gains in ancestral and modern genomes are still the same. From the results presented above, multiplication of ILEs in fungi and IEs in *M. pusilla* is certainly the main mechanism of intron gain in these species. Because of the high frequency of duplication events, ILE and IE multiplication likely involves a mechanism different from those proposed so far. Yet, spliceosomal retrohoming is the model that would comply best with our observations, but additional concepts are required in this model to take into account ILE specific characteristics. The predicted stable secondary structures of ILEs seem to be under selection pressure as suggested by the many compensatory mutations observed in ILEs [9]. It is tempting to speculate that ILE secondary structures might significantly contribute to the multiplication mechanism. We are now setting up experiments to find evidence for the mobility of ILEs and deciphering the mechanism of their multiplication.

References

1. Will CL, Lührmann R. Spliceosome structure and function. Cold Spring Harb Perspect Biol 2011; 3:a003707; PMID:21441581; http://dx.doi.org/10.1101/cshperspect.a003707.
2. Koonin EV. The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate? Biol Direct 2006; 1:22; PMID:16907971; http://dx.doi.org/10.1186/1745-6150-1-22.
3. Rogozin IB, Carmel L, Csuros M, Koonin EV. Origin and evolution of spliceosomal introns. Biol Direct 2012; 7:11; PMID:22507701; http://dx.doi. org/10.1186/1745-6150-7-11.
4. Csuros M, Rogozin IB, Koonin EV. A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes. PLoS Comput Biol 2011; 7:e1002150; PMID:21935348; http://dx.doi.org/10.1371/journal.pcbi.1002150.
5. Roy SW, Gilbert W. Rates of intron loss and gain: implications for early eukaryotic evolution. Proc Natl Acad Sci USA 2005; 102:5773-8; PMID:15827119; http://dx.doi.org/10.1073/pnas.0500383102.
6. Nielsen CB, Friedman B, Birren B, Burge CB, Galagan JE. Patterns of intron gain and loss in fungi. PLoS Biol 2004; 2:e422; PMID:15562318; http://dx.doi. org/10.1371/journal.pbio.0020422.
7. Zhang LY, Yang YF, Niu DK. Evaluation of models of the mechanisms underlying intron loss and gain in *Aspergillus* fungi. J Mol Evol 2010; 71:364-73; PMID:20862581; http://dx.doi.org/10.1007/s00239-010-9391-6.
8. Torriani SF, Stukenbrock EH, Brunner PC, McDonald BA, Croll D. Evidence for extensive recent intron transposition in closely related fungi. Curr Biol 2011; 21:2017-22; PMID:22100062; http://dx.doi.org/10.1016/j.cub.2011.10.041.
9. van der Burgt A, Severing E, de Wit PJ, Collemare J. Birth of new spliceosomal introns in fungi by multiplication of introner-like elements. Curr Biol 2012; 22:1260-5; PMID:22658596; http://dx.doi.org/10.1016/j.cub.2012.05.011.
10. Li W, Tucker AE, Sung W, Thomas WK, Lynch M. Extensive, recent intron gains in *Daphnia* populations. Science 2009; 326:1260-2; PMID:19965475; http:// dx.doi.org/10.1126/science.1179302.
11. Yenerall P, Zhou L. Identifying the mechanisms of intron gain: progress and trends. Biol Direct 2012; 7:29; PMID:22963364; http://dx.doi.org/10.1186/1745-6150-7-29.
12. Fink GR. Pseudogenes in yeast? Cell 1987; 49:5-6; PMID:3549000; http://dx.doi.org/10.1016/0092-8674(87)90746-X.
13. Roy SW, Gilbert W. The pattern of intron loss. Proc Natl Acad Sci USA 2005; 102:713-8; PMID:15642949; http://dx.doi.org/10.1073/pnas.0408274102.
14. Roy SW, Irimia M. Mystery of intron gain: new data and new models. Trends Genet 2009; 25:67-73; PMID:19070397; http://dx.doi.org/10.1016/j. tig.2008.11.004.
15. Hellsten U, Aspden JL, Rio DC, Rokhsar DS. A segmental genomic duplication generates a functional intron. Nat Commun 2011; 2:454; PMID:21878908; http://dx.doi.org/10.1038/ncomms1461.
16. Denoeud F, Henriet S, Mungpakdee S, Aury JM, Da Silva C, Brinkmann H, et al. Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. Science 2010; 330:1381-5; PMID:21097902; http://dx.doi.org/10.1126/science.1194167.
17. Worden AZ, Lee JH, Mock T, Rouzé P, Simmons MP, Aerts AL, et al. Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes *Micromonas*. Science 2009; 324:268-72; PMID:19359590; http://dx.doi.org/10.1126/science.1167222.
18. Coghlan A, Wolfe KH. Origins of recently gained introns in *Caenorhabditis.* Proc Natl Acad Sci USA 2004; 101:11362-7; PMID:15243155; http://dx.doi. org/10.1073/pnas.0308192101.
19. Sharpton TJ, Neafsey DE, Galagan JE, Taylor JW. Mechanisms of intron gain and loss in *Cryptococcus.* Genome Biol 2008; 9:R24; PMID:18234113; http:// dx.doi.org/10.1186/gb-2008-9-1-r24.

Chapter 7

In silico miRNA prediction in metazoan genomes: balancing between sensitivity and specificity.

van der Burgt A, Fiers MW, Nap JP, van Ham RC.

BMC Genomics 2009, 10:204.

doi: 10.1186/1471-2164-10-204.

This article is available from: http://www.biomedcentral.com/1471-2164/10/204
Additional Files are available from:
http://www.biomedcentral.com/1471-2164/10/204/additional

Abstract

Background: MicroRNAs (miRNAs), short ~21-nucleotide RNA molecules, play an important role in post-transcriptional regulation of gene expression. The number of known miRNA hairpins registered in the miRBase database is rapidly increasing, but recent reports suggest that many miRNAs with restricted temporal or tissue-specific expression remain undiscovered. Various strategies for *in silico* miRNA identification have been proposed to facilitate miRNA discovery. Notably support vector machine (SVM) methods have recently gained popularity. However, a drawback of these methods is that they do not provide insight into the biological properties of miRNA sequences.

Results: We here propose a new strategy for miRNA hairpin prediction in which the likelihood that a genomic hairpin is a true miRNA hairpin is evaluated based on statistical distributions of observed biological variation of properties (descriptors) of known miRNA hairpins. These distributions are transformed into a single and continuous outcome classifier called the *L* score. Using a dataset of known miRNA hairpins from the miRBase database and an exhaustive set of genomic hairpins identified in the genome of *Caenorhabditis elegans*, a subset of 18 most informative descriptors was selected after detailed analysis of correlation among and discriminative power of individual descriptors. We show that the majority of previously identified miRNA hairpins have high *L* scores, that the method outperforms miRNA prediction by threshold filtering and that it is more transparent than SVM classifiers.

Conclusion: The *L* score is applicable as a prediction classifier with high sensitivity for novel miRNA hairpins. The *L*-score approach can be used to rank and select interesting miRNA hairpin candidates for downstream experimental analysis when coupled to a genome-wide set of *in silico*-identified hairpins or to facilitate the analysis of large sets of putative miRNA hairpin loci obtained in deep-sequencing efforts of small RNAs. Moreover, the in-depth analyses of miRNA hairpins descriptors preceding and determining the *L* score outcome could be used as an extension to miRBase entries to help increase the reliability and biological relevance of the miRNA registry.

Background

MicroRNAs (miRNAs) are ~21-nucleotide (nt) short, single stranded RNA molecules involved in post-transcriptional regulation of gene expression [1]. They are present in higher eukaryotes and some viral genomes [2]. Because the miRNA and small-interfering RNA (siRNA) pathways partly overlap, current understanding of miRNA biogenesis has gained from advances made in the field of RNA interference. Mature, functional miRNAs develop from degenerate palindromic repeats with a characteristic hairpin-like secondary structure [1,3,4]. Initially, experimental identification of miRNAs was achieved through direct cloning and sequencing of small RNAs [5,6]. However, such relatively low-throughput screenings were biased towards abundantly or ubiquitously expressed miRNAs [6] and missed many miRNAs with restricted temporal or tissue-specific expression patterns [1]. Recently, strategies using PCR [7], microarrays [8,9] or ultra high-throughput sequencing [10,11] have expanded the list of known miRNAs. Many of these show tissue-specific expression [9,12] or appear to be species-specific [12,13]. In both *Arabidopsis thaliana* [14,15] and *Caenorhabditis elegans* [10,16], for example, high-throughput sequencing of small RNAs shows moderate overlap in detected miRNAs between experiments in each species, indicating that also in well studied genomes many new miRNAs remain to be discovered. In addition, the observation that many miRNA loci exhibit compelling hairpin structures on both sense and antisense strands led to the discovery of anti-sense miRNA transcription [17]. Antisense miRNA transcription and processing yield distinct mature miRNAs. This contributes to the functional diversification of miRNA genes for a considerable fraction of the known miRNA loci [17,18].

Various methods for the *in silico* prediction of miRNAs have been developed to aid in experimental studies of miRNA discovery [19,20]. These methods generally consider the hairpin-like secondary structure of the miRNA precursor, the miRNA hairpin, as the most important characteristic of a miRNA gene. They use RNA secondary structure prediction (RSSP) algorithms, such as RNAfold [21] or Mfold [22], to predict the secondary structure and thermodynamic stability of the RNA hairpin structures. Current bioinformatics approaches for the prediction of miRNAs [19,20] generally include three steps: (1) genome-wide prediction of hairpin structures; (2) filtering or scoring of those hairpins on the basis of their similarity in physical and sequence features to known miRNA hairpins and (3) experimental validation of putative candidates.

A common approach for the first step is to search for hairpin structures using a sliding window and perform RSSP on each window [5,8,23,24]. An improvement in overall calculation time over this approach is to first identify degenerate palindromic sequences and to analyse only these further with RSSP [25,26]. Unfortunately, these approaches detect vast numbers of hairpin structures in complete eukaryotic genomes. Depending on the method used, 1E3 to 4E3 hairpins per Mb of genomic sequence are found, resulting in about 1E7 hairpins identified in the human genome [8,24-27]. The challenge is, therefore, to devise an appropriate filtering method to separate the chaff from the wheat.

Different criteria for filtering candidate miRNA sequences have been proposed, generally with the aim to reduce the search space and/or to increase the specificity of prediction [20]. Evolutionary conservation is considered an important feature of the hairpin sequence [1] and analysis thereof is often used to identify and focus comparisons on the conserved non-coding sequence space in different genomes [5,23]. An evolutionary approach known as phylogenetic shadowing has been used for combined selection and filtering of miRNA candidates [28]. This study revealed a characteristic camel-shaped conservation pattern of putatively orthologous miRNAs that was useful as a criterion for finding conserved miRNA candidates in primate genomes. Other filtering criteria include intragenomic matching of candidate miRNAs and their potential targets [29], evidence for expression, thresholds on structural properties of hairpins, e.g. minimal folding energy (MFE), absence of repetitive or low-complexity sequences, occurrence in introns or intergenic regions [11], or proximity to known miRNA loci [2,24,30].

Stringent filtering is performed to attain high specificity, that is, to minimize the number of false positive predictions of miRNA genes [20,25]. Obviously, maximizing specificity increases the number of false negative predictions, that is, a decreased sensitivity [31]. This implies that with stringent filtering, relevant miRNAs will be missed. The success of any filtering procedure depends however on the validity of the underlying assumptions. For example, filtering on evolutionary conservation will miss species-specific or fast evolving miRNAs [12] and low-complexity filtering is likely to miss miRNAs originating from transposable elements [32,33]. Several recent methods employ machine learning techniques such as a Support Vector Machine (SVM) [24,27,34,35] for classification. Such SVMs evaluate differences in hairpin properties between true-positive and true-negative examples of miRNAs for a given taxon to generate a prediction classifier. SVMs have been successfully used for miRNA gene prediction [24,34,35] and for the prediction of 5' Drosha processing sites in miRNA hairpins [26,11]. Although the SVM approach is claimed to outperform earlier methods [35], SVM-based classifications combine many features in a single kernel function and therefore do not provide direct insight into the biological significance of these features. Such insight can only be obtained through expert analyses and dedicated feature selection procedures [19]. Moreover, the set of true-negatives which is required for training an SVM is often very difficult to define.

Here, we present an innovative strategy for miRNA prediction that focuses on attaining optimal sensitivity. We define and combine 40 different filtering criteria and, using a set of genomic hairpins identified in the genome of *Caenorhabditis elegans*, show that 18 of these characteristics capture the biological variation of miRNA features present in sets of known miRNA hairpins. These 18 criteria are used to establish a combined likelihood score *L* that assesses the likelihood that a predicted hairpin structure in a genome contains a genuine miRNA. *L* is a continuous classifier that allows user-adjustable thresholds for sensitivity and specificity in *ab initio* miRNA prediction and miRNA analysis. Good performance of *L* for large sets of hairpins from the genomes of *C. elegans* and four viruses demonstrates the added value of the new analytical strategy for future miRNA discovery and selection.

Results

We have developed and evaluated a new computational strategy for the prediction of candidate miRNAs in DNA sequences. The new approach focuses on high sensitivity in an initial hairpin detection step, followed by a flexible, user-adjustable procedure to balance sensitivity with specificity and selectivity. For all hairpin sequences from an input sequence, a miRNA likelihood score *L* is calculated, given an underlying scoring model based on descriptors of the physical and sequence characteristics of miRNAs (Table 1; [see Additional file 1]). The performance of the strategy was assessed by retrieval of known miRNAs from hairpin structures identified in the genome of *C. elegans*. Appropriate scoring models were derived for various taxonomic sets of known miRNAs.

Descriptor data fit

Most of the descriptors used for miRNA characteristics have been proposed in previous studies [3639], but a few are, to the best of our knowledge, for the first time defined in this study, for example 'GAsurplusCU' (Table 1). In all cases except one, the empirical data showed a good fit to a skewnormal (SN) probability distribution [40] accord-ing to a Chisquare goodness-of-fit test ($p \leq 1.4E-5$; [see Additional file 2]). As example, the frequency distribution, fitted SN distribution and the transformed likelihood distribution function (LDF) are shown for the descriptors MFE and GC content in Figure 1. The only exception to an SN probability distribution fit was found for the descriptor 'P', which is the *p*-value of the MFE of randomized sequence [38]. This fitted best to an exponential distribution corrected for zero values. Results of the data fit for all descriptors from the taxonomic set Metazoa are listed in Additional file 1.

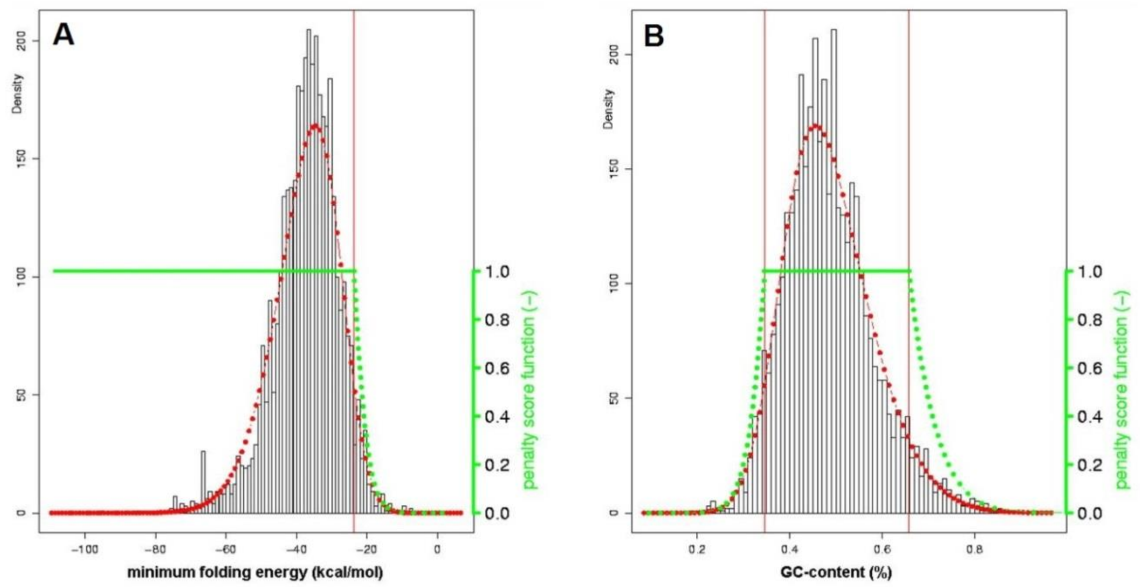


Figure 1. Data fit and likelihood distribution function for two descriptors.

Frequency distribution (black bars), SN-fitted distribution (red curve) and likelihood distribution function (LDF) (green curve) for descriptors MFE (A) and GC-content (B) of the taxonomic set Metazoa (3,902 miRNA hairpins). Red vertical lines mark the upper and lower 5% tails of the distribution.

Table 1: Subset of 18 most informative miRNA hairpin descriptors

Descriptor	Explanation	Bound [a]	Type [b]	Discriminative power [c]	K [d]
bulgeRatio	ratio asymmetrical bulges vs. stem length	↑	str	1.45	0.416
dP	adjusted base pairing propensity (dP)	↓	str	2.28	0.417
largest bulge	longest bulge in stem (nt)	↑	str	1.74	0.343
longest match-stretch	longest match-stretch in stem (nt)	↓	str	1.20	0.336
Looplength	central loop length (nt)	↑	str	1.17 (u)	0.191
max match count	matches in 24 nt	↓	str	2.75	0.477
MFEahl index [39]	MFEahl corrected for GC- Content	↓	str	4.75	0.706
Q [37,39]	Normalized Shannon entropy (Q)	↑	str	3.01	0.844
stem length	stem length	↑↓	str	1.29 (l)	0.404
GAsurplusCU	surplus of GA over CU in sequence	↑↓	seq	1.12 (u) 1.05 (l)	0.195
GsurplusC	surplus of G over C in sequence	↓	seq	1.12	0.995
polyA	longest poly-A stretch (nt)	↑	seq	1.58	0.834
polyNucHairpin	longest mono-nucleotide stretch (nt) in the hairpin	↑	seq	1.64	0.846
polyU	longest poly-U stretch (nt)	↑	seq	1.53	0.540
SCS-di	Di-nucleotide Sequence Complexity (-)	↑	seq	1.77	0.557
SCS-mono	Mono-nucleotide Sequence Complexity (-)	↓	seq	1.60	0.317
GU-match contribution	ratio of GU-matches vs. all matches	↑	mix	1.28	0.173
MFEahl (dG) [37,39]	MFE Adjusted for hairpin length	↓	mix	13.33	0.742

A detailed explanation of all 40 descriptors is given in the additional information [see Additional files 1 and 2].

[a] Boundary; indication of extreme tail of descriptor distribution that was transformed into $S < 1$ fraction. Symbols denote: ↓ lower tail; ↑ upper tail, and; ↓↑ both tails. [b] Type; descriptor based on structural (str), sequence (seq) or both structural and sequence (mix) properties of the hairpin. [c] Discriminative power; expressed at 95% sensitivity, measured on the taxonomic set Metazoa (3,902 miRNA hairpins, positives) and genomic hairpins in *C. elegans* (3,526,115 hairpins, negatives). * Discriminative power of descriptor 'Z' was measured on 25,599 identified hairpins in four viruses. [d] K; highest Cohen's kappa coefficient [41] with another descriptor, measured on the taxonomic set Metazoa with $S < 1$ cut-off of 95%.

Likelihood score S for miRNA hairpin descriptors

The CDF of the fitted distribution was transformed into an LDF with outcome S. The range of S (between 0 and 1) has an S < 1 and an S = 1 fraction which are separated by the cut-off value derived from the 95% confidence interval of the descriptor's CDF. The S < 1 fraction contains miRNAs hairpins with descriptor values in the tail(s) of the distribution. These have a low probability of occurring in true miRNA hairpins. The S = 1 fraction contains miRNA hairpins with values in the remainder of the distribution and corresponds to likely properties of miRNA hairpins. Descriptors were treated differently with respect to the transformation of the tails of the CDF (Table 1; [see Additional file 2]). For example, for the descriptor "minimal folding energy" (MFE) of a miRNA, there is in principle no need to impose a lower bound, even though the fitted distribution (Figure 1A) indicates that very low MFE values occur rarely in known miRNAs. The LDF of the descriptor MFE assigns a score S = 1 for all values below - 23.72 kcal/mol. Higher MFEs are penalized proportional to the LDF and therefore assigned a score S < 1. MFE is an example of a descriptor where the S < 1 fraction represents the upper 5% tail of the confidence interval. For other descriptors, the S < 1 fraction is represented by the lower 5% tail of the distribution (e.g. match ratio) or by both the lower 5% and the upper 5% tail (e.g. GC-content, Figure 1B). Table 1 and Additional file 1 list the unlikely tails for each descriptor.

Correlation of descriptors

The 40 descriptors here defined were either based on the sequence of the hairpin, the structure of the hairpin or on a combination of both (Table 1; [see Additional file 1]). Correlations among these were obvious, for instance, between stem length or GC-content on the one hand and the MFE of a hairpin on the other hand. Correlated descriptors will overemphasize the importance of a more general feature of a miRNA hairpin and affect the usefulness of L. To assess the correlation among descriptors in their S < 1 fractions, we calculated Cohen's kappa coefficient κ [41] for all 780 possible pairs of descriptors, using the miRNA hairpins of the taxonomic set Metazoa [see Additional file 3]. For each descriptor the most strongly correlated descriptor, as determined by the highest observed κ is listed in Table 1 (column κ) and Additional file 3. The highest κ was found between descriptors 'GCratio' and 'GsurplusC' (κ = 0.995), followed by 'D' and 'Q' (κ = 0.844) and the pair 'P' and 'Z' (κ = 0.784).

Discriminative power of descriptors

A discriminative descriptor contributes to the separation of true miRNA hairpins from non-miRNA hairpins. The discriminative power of a descriptor was defined as the ratio of percentages of miRNA hairpins and genomic hairpins that comply with a threshold set to the descriptor's limiting value between S = 1 and S < 1 of the LDF (95% of the CDF). A discriminative power smaller than 1.0 implies that relatively more miRNA hairpins are rejected than genomic hairpins. Higher values are obtained for descriptor values that are typically encountered in miRNA hairpins, but that are less common in collections of genomic, predominantly non-miRNA hairpins. The most discriminative descriptor was MFEahl (13.33) and least discriminative were polyC, polyCstem, polyGstem (0.95) (Table 1; [see Additional file 2]). Figure 2 illustrates the discriminative power of descriptor MFEahl and reveals a substantial difference between the SN-fitted CDFs of miRNA hairpins and randomly selected genomic hairpins. Only 7% of genomic hairpins complied with the criterion of an MFEahl of 0.314, representing 95% of the CDF of metazoan miRNA hairpins. The opposite was true for the least discriminative descriptors: for example for polyC, 99% of the genomic hairpins versus 96% of the metazoan miRNA hairpins had a longest polyC stretch smaller than five (not shown).

Delimiting a subset of most informative descriptors

The correlation and discriminative power of descriptors were used to select a non-redundant subset of most informative descriptors. Descriptors that either correlated with a more selective descriptor, using a threshold for correlation of κ>0.4, or that had a discriminative power smaller than 1.1 were omitted. This resulted in a subset of 18 descriptors, seven of which were sequence related, nine structure related, and two descriptors with mixed properties (Table 1). Remarkably, the descriptors 'GC-content' and the MFE randomization descriptors 'P' [38] and 'Z' [37], which are often used in miRNA prediction studies, were not included in the subset [see Additional file 1]. The latter two ranked among the most selective descriptors, but were excluded because of their strong correlation with the most selective descriptor MFEahl index, in which the MFE is adjusted for hairpin length and GC-content [see Additional file 3].

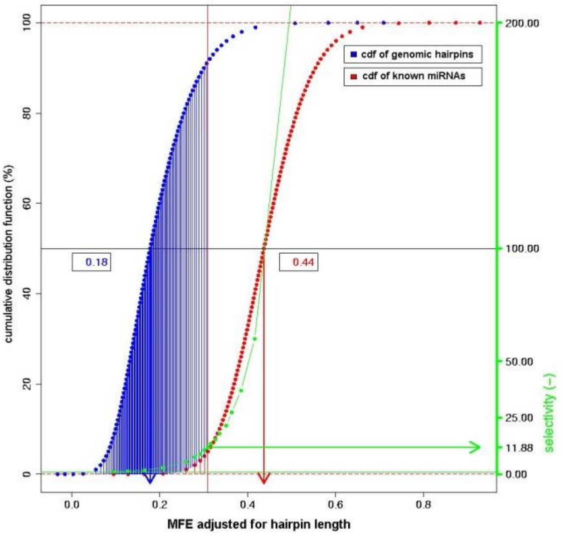


Figure 2. Discriminative power of the descriptor MFEahl.

Red curve represents the CDF of the descriptor MFEahl for the taxonomic set Metazoa (3,902 miRNA hairpins). Blue curve represents the CDF of the SN-fitted distribution of the same descriptor in case of 100,000 randomly selected hairpins from the C. elegans genome. Green curve represents the discriminative power, calculated as sensitivity/(1.0-specificity). The fraction of hairpins in the S < 1 fraction is shaded (S < 1 cut-off at 95% of the CDF of known miRNA hairpins). The discriminative power at 95% sensitivity is shown by a green arrow (13.33). SN-fitted means are shown by red (0.44) and blue (0.18) arrows.

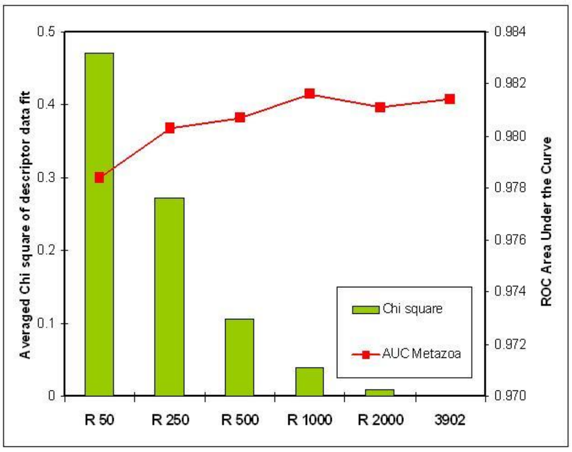


Figure 3. Accuracy of fit and scoring model performance depends on the size of the input set.

AUC performance (red line) and average Chi-square accuracy of fit of 40 descriptors (green bars), using six scoring models that were based on varying sizes of the input set. Input-set sizes are indicated with a prefix 'R' and comprised 50, 250, 500, 1000, 2000, and the complete set (3,902) of metazoan miRNA hairpins. The smaller sets were compiled by randomly selecting miRNA hairpins from the complete set. This was repeated 50 times for each set. The accuracy of fit was then calculated by averaging Chi-square test statistics over all 40 descriptors and the 50 randomly selected subsets of each indicated size. Both AUC performance and Chi-square statistics show a strong dependency on Input set size.

Assessment of scoring model performance

The 18 most selective descriptors (Table 1) were used to define and analyze different scoring models. The CDF of the fitted distribution of all 18 descriptors was transformed into an LDF with outcome S and the L score for a given miRNA sequence was calculated as the product of all S values. Scoring model performance, defined as the power to distinguish (potential) miRNA hairpins from other (or random) genomic hairpins, was compared for different models that were built using varying settings for five parameters (see below). Scoring model performance was measured as AUC performance and, where appropriate, with selectivity measured at two values of sensitivity (95% and 75%), using genomic hairpins in C. elegans and the collection of miRNAs hairpins in the taxonomic set Metazoa.

(1) Size of taxonomic set

As expected, scoring models based on small taxonomic sets had a less accurate data fit, as shown by increasing Chi-square statistics for decreasing set size (Figure 3). Gain in goodness-of-fit and AUC performance saturated with increasing set size. The data presented in Figure 3 indicate that a minimal set size of a thousand miRNA hairpins is required for an accurate data fit (average p(Chi-square) < 0.05). Performance in terms of selectivity appeared to increase beyond this set size, suggesting that prediction performance can be further improved by using larger taxonomic sets.

(2) Composition of taxonomic set

Performance was found to depend on the evolutionary distance between the species contained in a scoring model's taxonomic set and the species for which (miRNA) hairpins are scored by that scoring model. When five taxonomic sets were constructed that comprised equally sized sets of miRNA hairpins from taxa with a decreasing evolutionary distance and diversity relative to human (Figure 4), these showed increasing AUC performance on the miRNA hairpins from human. The opposite trend, i.e. a decrease in AUC performance of the same five scoring models, was observed for miRNA hairpins from Nematoda and Metazoa. Seemingly small differences in AUC values translate to substantial differences in genome-wide counts of positive hairpins. When choosing an arbitrary miRNA detection sensitivity of 75%, the difference between an AUC of 0.9831 (red bar Metazoa- Mammalia in Figure 4) and 0.9806 (red bar Homo sapiens in Figure 4) results in 6,749 or 17.6% fewer remaining genomic hairpins (38,383 and 31,634 hairpins, respectively). The results confirm that a scoring model based on a set that is taxonomically closest to the organism for which the miRNA hairpins are scored, performs best. However, it is noteworthy that taxonomic sets that do not contain human miRNA hairpins (e.g. Mammalia excluding H. sapiens, Figure 4) can yield scoring models with a good performance for identifying human miRNAs.

(3) Composition of descriptor subset

Selection of informative descriptors was found to be a crucial step in the development of scoring models, and three factors that affected scoring model performance were therefore analyzed in detail, i.e. the number of descriptors selected (quantity), discriminative power of descriptors (quality) and correlation between descriptors. Correlation of structural properties that describe RNA molecules is well-known [37]. In a miRNA prediction method such correlations can strongly influence the prediction accuracy and should therefore be dealt with cautiously. Figure 5 illustrates the effect of highly correlated descriptors on scoring model performance, using the strongly correlated descriptors 'D' and 'Q' (κ = 0.84) and a third descriptor 'SCS-di' that has a weak correlation with both 'D' and 'Q' (κ = 0.13 and 0.12, respectively). Scoring models containing pairs of uncorrelated descriptors had a higher selectivity value than the individual descriptors. In contrast, the scoring model based on the strongly correlated descriptors 'D' and 'Q' gave a lower discriminative power than one based on the most discriminative, individual descriptor 'Q' (Figure 5). Although this decrease (-0.14) seems small, the combined effect over all correlated descriptors will have a considerable effect in terms of the absolute number of genomic hairpins that are penalized. It illustrates the necessity of selecting a descriptor subset with as little pairwise correlation as possible in the development of appropriate scoring models.

(4) Parameterization of the LDF

The effect of parameterization on performance was assessed by lowering the default cut-off value (default 95% of the CDF) to 90% and 80% before use in calculation of the transformed likelihood distribution score S. The parameterization caused the number of miRNA hairpins included in the S < 1 fractions to double (90%) or quadruplicate (80%) and increased the penalization of hairpins relative to the default cut-off value. Figure 6 shows that increased AUC performance and selectivity values were obtained by lowering the cut-off value compared to the reference model, with selectivity indexed for the reference model. The increase is explained by the fact that most descriptors gain in discriminative power at a decrease in sensitivity (see Figure 2).

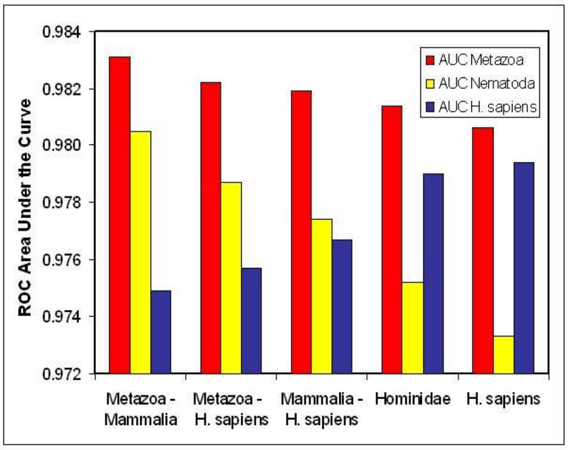


Figure 4. Scoring model performance depends on the taxonomic distance of the input set. AUC performance of five different scoring models that vary in the distance of the taxonomic input set. Area under the ROC curves is measured for the taxonomic sets Metazoa (red), Nematoda (yellow) and H. sapiens (blue) versus 200,000 randomly selected hairpins from the set of 3,526,115 C. elegans hairpins. Scoring models "X – Y" have as taxonomic input set all miRNA hairpins from set X after removal of set Y. From these subsets (and the set Hominidae) 781 miRNA hairpins have been randomly selected. The results presented for the random subsets are averages from 50 independent repeats.

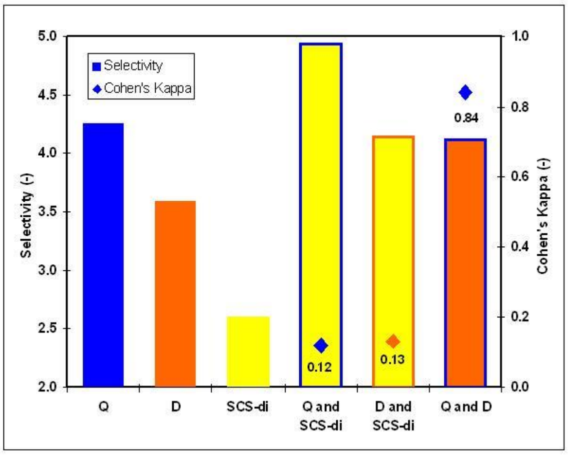


Figure 5. Scoring model performance depends on the correlation of descriptors.

Discriminative power of three individual and three pairs of descriptors. For the descriptor pairs, Cohen's kappa coefficients are also given. Selectivity is expressed at 95% sensitivity on the set of all metazoan miRNA hairpins; specificity is measured on the set of 3,526,115 hairpins in the genome of *C. elegans*.

(5) Weighting of descriptors

The effect of weighting individual descriptors was studied by assigning weights to each of the 18 previously selected descriptors. Weights were equal to the square root of a descriptor's discriminative power as measured at a sensitivity of 95% and ranged from 1.06 for GAsurplusCU to 3.65 for MFEahl (Table 1; [see Additional file 1]). Figure 6 shows that a scoring model with weighted descriptors (W95% and W90%) had a significantly increased AUC performance and selectivity index relative to their unweighted models (95% and 90%). Obviously, when weighted on the basis of their discriminative power, descriptors with a high discriminative power contribute stronger to the overall L score than descriptors with low discriminative power. As such, a better separation between miRNA hairpins and random genomic hairpins is accomplished.

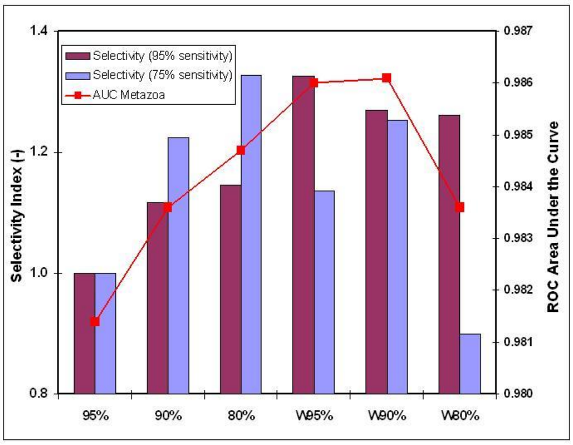


Figure 6. Scoring model performance depends on LDF parameterization and weighting of descriptors.

AUC performance and selectivity of six different scoring models that vary in parameterization of the LDF (95-90-80%) and have no weighted (weight = 1.0) or weighted individual descriptors (W). Weights were adjusted to the square root of the descriptor's discriminative power as measured at a sensitivity of 95% (Table 1). The square root was taken to prevent disproportionate influence of descriptors with high discriminative power. All models have the same input set (3,902 metazoan miRNA hairpins) and are based on the previously selected set of 18 descriptors. Selectivity is expressed at 95% (purple) and 75% (blue) sensitivity on the set of all metazoan miRNA hairpins; specificity is measured on the set of 3,526,115 hairpins in the genome of *C. elegans*. Relative values of selectivity are presented with the initial scoring model taken as index (selectivity of 12.6 at 95% and 74.1 at 75% sensitivity).

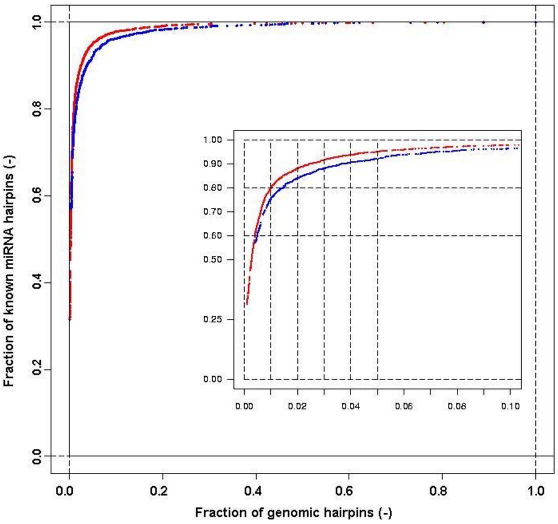


Figure 7. ROC-curve of the L-score classifier of two different scoring models.

ROC curve of the *L-score* classifier of the final scoring model Metazoa (red) and the initial model without weighting and default parameterization (blue). True positives are measured on the taxonomic set Metazoa (3,902 miRNA hairpins), false positives on 500,000 randomly selected genomic hairpins from *C. elegans*.

Finally, combinations of weighting and varying the parameterization of descriptors were tested (models W90% and W80%, Fig. 6). Scoring model W90% had the highest AUC performance on metazoan miRNA hairpins of all tested models. When weighting and parameterization were compared, weighting showed a stronger effect on selectivity measured at 95% sensitivity, whereas parameterization had a stronger effect on selectivity at 75% sensitivity. This implies that both variables have a distinct effect on the shape of the ROC curve.

Building optimal scoring models

The data collected and analyses presented allowed the selection of optimal scoring models, with maximal discriminative power to distinguish true miRNA hairpins from other genomic (or random) hairpins for any given case. In general, significant increases in selectivity were gained by descriptor weighting and parameterization. In terms of choice for a specific scoring model, the taxonomic input set should be sufficiently large and taxonomically as close as possible to the organism of interest. In total, 23 scoring models were built, based on 23 distinct taxonomic sets and the subset of 18 most informative descriptors, with the LDF parameterized at 90% of the CDF and using individual weighting of descriptors. Weighting and parameterization at 90% of the CDF resulted in the highest AUC performance. Table 2 shows the performance gain of these optimal scoring models relative to their non-optimized counterparts. Figure 7 shows the ROC curves for the final and initial scoring model Metazoa. The final scoring model Metazoa was subjected to a 10-fold cross-validation for benchmarking and yielded an AUC of 0.9732. The arbitrary cut-off for L of 1.0e-4 classifies 87.3% (1,774/2,033) of miRNA hairpins correctly as positive and 97.0% of all genomic hairpins of C. elegans as negative. The difference in performance with the non cross-validated AUC performance (0.9874) is due in part to the much smaller taxonomic input set used. The latter was obtained by clustering all Metazoan miRNA hairpins on the basis of sequence similarity. Nearly identical hairpin sequences can have subtle variation in some descriptor values and thereby accurately represent the fact that miRNAs occur in families. Therefore, and to maintain a classifier as selective as possible, we recommend to use the non-clustered variant of the scoring model.

Table 2: Scoring model performance

Scoring model description	**#** [a]	**Wt** [b]	**LDF** [c]	**Chi-Square** [d]	**AUC** [e] **Nematoda**	**AUC** [e] ***H. sapiens***	**AUC** [e] **Metazoa**	**Selectivity** [f] **95%**	**75%**
Default models [g]									
Metazoa	3,902	-	0.95	9.95e-7	0.9760	0.9764	0.9814	12.57	74.06
H. sapiens	781	-	0.95	0.090	0.9733	0.9794	0.9806	11.04	68.90
C. elegans	131	-	0.95	0.405	0.9747	0.9638	0.9735	7.73	41.43
Optimized models [g]									
Metazoa	3902	Y	0.90	9.95e-7	0.9813	0.9848	0.9874	21.92	105.7
H. sapiens	781	Y	0.90	0.090	0.9798	0.9870	0.9871	21.93	87.36
C. elegans	131	Y	0.90	0.405	0.9817	0.9775	0.9835	16.65	75.34

[a] Number of miRNAs in taxonomic set. [b] -; No weighting (weight = 1), Y; weighting individual descriptors by the square root of their discriminative power (Table 1) [c] LDF parameterized at 95% or 90% of the CDF

[d] Goodness-of-fit is evaluated by averaging Chi-square test statistics of all 40 descriptors

[e] Area under the curve of ROC curves measured on the taxonomic sets of known miRNA hairpins from Nematodes (211 miRNA hairpins), *H. sapiens* (781) and Metazoa (3,902) versus 200,000 randomly selected genomic hairpins from *C. elegans*. [f] Selectivity expressed at 95% and 75% sensitivity, with sensitivity measured on the taxonomic set of Metazoa, specificity measured on set of 3,526,115 *C. elegans* genomic hairpins.

[g] Scorings models composed of the subset of 18 most informative descriptors

Combined likelihood score L for miRNA hairpin descriptors

The combined L score for a given miRNA hairpin was calculated as the product of all S values of descriptors considered in a scoring model. In Figure 8, the distribution of the resulting L scores for all miRNA hairpins from the taxonomic set 'Metazoa' is given for two different scoring models in a cumulative L-score plot. For the scoring model Metazoa (red), 31% of the known metazoan miRNA hairpins had an L score of 1.0. This means that for 1,227 miRNA hairpins, the S score of the LDF of each of the 18 descriptors was 1.0. A cumulative L score plot can be used to select a desired level of sensitivity: an arbitrarily chosen sensitivity of 90% is reached at an L of 0.0004 for the optimized scoring model Metazoa (red) and at an L of 0.032 for the initial scoring model Metazoa (blue). This hundred-fold difference is caused by adjusted parameterization and weighting of descriptors in the first model, resulting in a higher overall penalization. This example illustrates that the L score is a relative measure that depends on the scoring model. For all 3,984 miRNA hairpins used in this study, L scores were calculated for all 23 scoring models [see Additional file 4], which were all based on the subset of 18 descriptors. Figure 9 shows an example of a detailed descriptor report for cel-mir-51 for scoring model Metazoa. The report shows the individual descriptor values, positions of these values in the CDF of the descriptors and the transformed S scores. The resulting L score equals 0.057 due to the fact that three descriptors fall outside the 90% range of the CDF. Data for all 91,632 combinations of miRNAs (3,984) and scoring models (23) are available in the accompanying web document µRNALL [42].

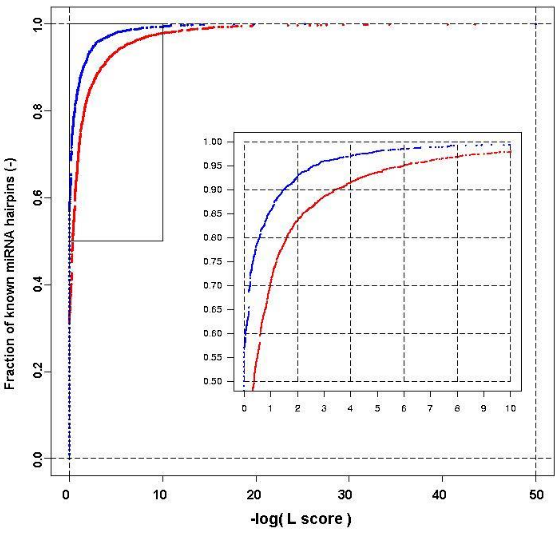
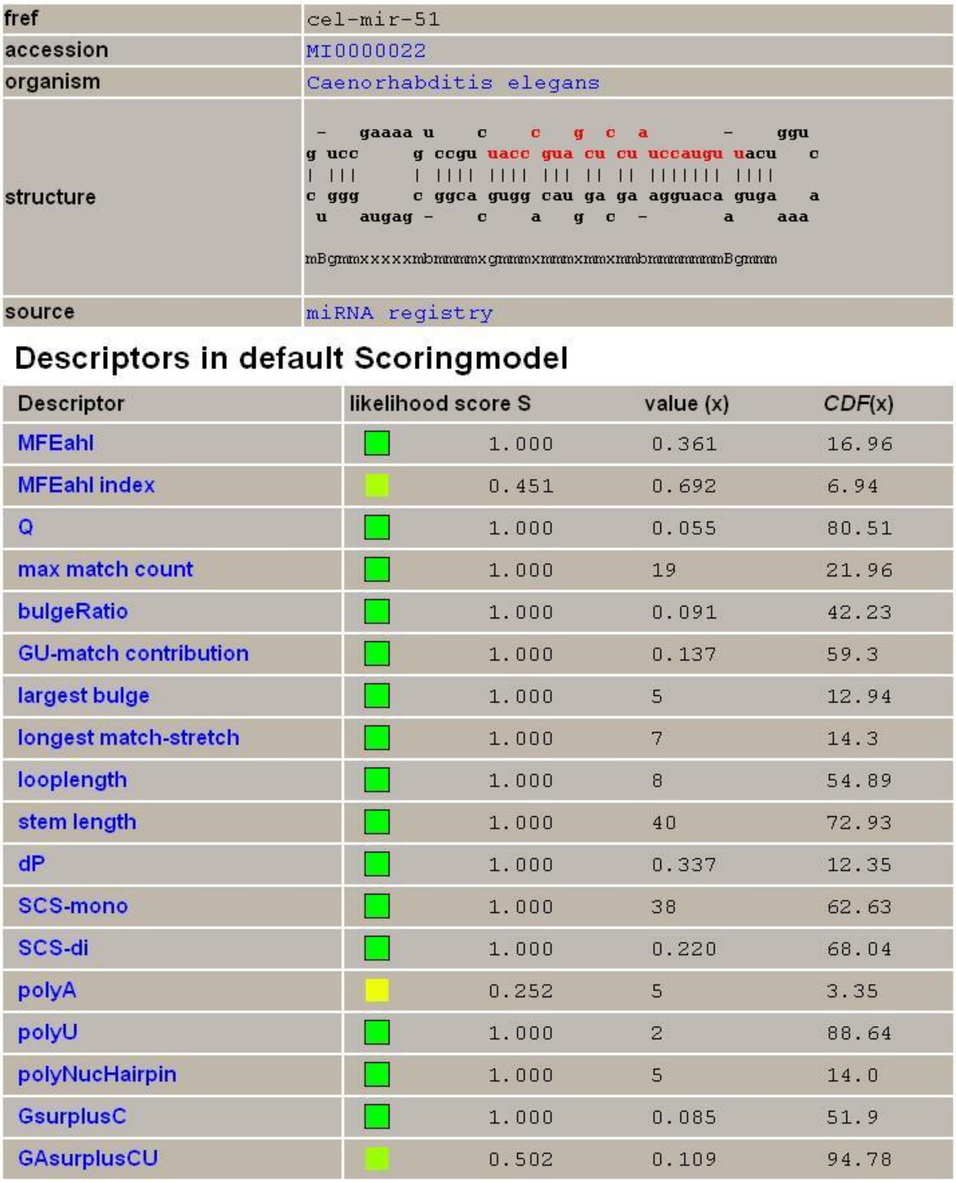


Figure 8. Cumulative *L-score* plot of two different scoring models.

Ratio of miRNA hairpins in the taxonomic set Metazoa (3,902) that have an *L* score of at least a certain value. Data are shown for the final scoring models Metazoa (red) and the initial model without weighting and default parameterization (blue).

Comparison of the L score approach to threshold filtering

Comparison of our *L* score method to binary threshold filtering on miRNA hairpin descriptors showed superior performance of the *L* classifier (Table 3). For sets of miRNA and genomic hairpins that were filtered at the bordering value between the *S* = 1 and *S* < 1 fractions for all 18 descriptors, values for the performance parameters sensitivity, specificity and selectivity were obtained. Next, we kept either sensitivity or specificity constant, assessed at which *L* score this parameter was equalled and compared the other performance parameters. In both cases, our scoring model approach outperformed threshold filtering. When fixed at sensitivity, the scoring model approach achieved 17% better. Sensitivity was measured on all metazoan miRNA hairpins (3,902 hairpins) instead of on *C. elegans* miRNA hairpins (132 hairpins) because of the inaccuracy caused by the limited number of miRNAs in the *C. elegans* set (data not shown).



fref	cel-mir-51
accession	MI0000022
organism	Caenorhabditis elegans
structure	see below
source	miRNA registry

```
 -   gaaaa u    c    c   g  c  a       -    ggu
g ucc     g ccgu uacc gua cu cu uccaugu uacu   c
| |||     | |||| |||| ||| || || ||||||| ||||
c ggg     c ggca gugg cau ga ga agguaca guga   a
 u   augag -    c    a   g  c  -       a    aaa

mBgmmxxxxxmbmmmmxgmmmxmmmxmmxmmbmmmmmmmBgmmm
```

Descriptors in default Scoringmodel

Descriptor	likelihood score S		value (x)	*CDF*(x)
MFEahl	🟩	1.000	0.361	16.96
MFEahl index	🟩	0.451	0.692	6.94
Q	🟩	1.000	0.055	80.51
max match count	🟩	1.000	19	21.96
bulgeRatio	🟩	1.000	0.091	42.23
GU-match contribution	🟩	1.000	0.137	59.3
largest bulge	🟩	1.000	5	12.94
longest match-stretch	🟩	1.000	7	14.3
looplength	🟩	1.000	8	54.89
stem length	🟩	1.000	40	72.93
dP	🟩	1.000	0.337	12.35
SCS-mono	🟩	1.000	38	62.63
SCS-di	🟩	1.000	0.220	68.04
polyA	🟨	0.252	5	3.35
polyU	🟩	1.000	2	88.64
polyNucHairpin	🟩	1.000	5	14.0
GsurplusC	🟩	1.000	0.085	51.9
GAsurplusCU	🟩	0.502	0.109	94.78

Figure 9. Detailed descriptor analyses report for cel-mir-38.

Detailed report for the observed descriptor values of cel-mir-38 in the scoring model Metazoa (*L* score = 0.057). For each descriptor, a color-coded representation of the likelihood score S, the actual value of *S*, the actual observed descriptor value and the position of this value in the CDF of the descriptor are given. Descriptors MFEahl index, polyA and GAsurplusCU are in the S<1 fraction outside 90% of the CDF.

Table 3: Comparison of threshold filtering of miRNA hairpins with the *L*-score classifier

Method	Sensitivity % (number)	Specificity % (number)	Selectivity [c]	*L* score [a]
Threshold filtering	56.8 (2,216)	99.57 (15,049)	133	-
Scoring model Metazoa				
Fixed at specificity	62.0 (2,421)	99.57 (15,048)	145	0.221

[a] Sensitivity measured on the taxonomic set Metazoa (3,902 miRNA hairpins)

[b] Specificity measured on the set of 3,526,115 genomic hairpins in *C. elegans*

[c] Selectivity calculated with sensitivity on metazoan miRNAs and specificity measured on *C. elegans* genomic hairpins. [d] minimal *L* score of the scoring model Metazoa for which the performance of threshold filtering performance is equalled

Predicting hairpins from a genomic sequence

We developed a procedure for predicting hairpin structures from genomic sequences using Vmatch [43]. The algorithm detects degenerate palindromic repeats and is therefore able to recover known miRNA hairpins with very high sensitivity. To benchmark the procedure, we predicted hairpin structures in the genomes *C. elegans* and four viruses: Epstein-Barr virus (EBV), Mareks disease virus (MDV), Human cytomegalovirus (HCMV) and Kaposi sarcoma-associated herpesvirus (KSHV). The number of recovered, known miRNA hairpins, number of identified hairpins and percentage and absolute number of non-overlapping hairpins on unique loci are given for three different *L* score criteria for the scoring model Metazoa, for both miRNA and genomic hairpins (Table 4).

In the four viral genomes, 25,599 hairpins were identified, including all 55 known miRNAs hairpins of these viruses. In *C. elegans*, 3,526,115 hairpins were predicted and only four out of 132 known miRNAs hairpins were missed (cel-mir-262, cel-mir-260, cel-mir-272 and cel-mir-256). When benchmarking the performance of the hairpin identification with the miRBase entries [44] of all metazoan miRNA hairpins, 3,803 out of 3,902 (97.5%) miRNA hairpins were recovered. The hairpin prediction algorithm is independent of sequence context (data not shown). This benchmark is therefore an estimate of the algorithm's good performance on metazoan genomes. Similar performance has been reported for other edit-distance based hairpins detection methods [26]. On a genome scale, all these methods yield around 10,000– 20,000 hairpins per single stranded Mb of sequence. These hairpins are predominantly overlapping and nested.

Analyses of miRNA hairpin candidates in viral genomes

The data in Table 4 show that only 1.3–2.6% of the indentified hairpin loci in four viral genomes have a high *L* score ($L \geq 0.05$). This corresponds to 69–247 loci per genome. In addition, 83–100% of all known miRNA hairpins comply with this threshold for *L*. Out of 6,182 hairpins predicted in the genome of EBV, only 23 hairpins, originating from 20 unique loci, had an *L* score of 1.0. Ten of these were experimentally validated miRNA loci [2] [see Additional file 5]. Further support for our miRNA prediction and scoring method comes from recently discovered miRNAs in the MDV genome [45] that were not included in miRBase 9.0. All five novel miRNAs (mdv1-mir-M9 to mdv1-mir-M13) were present in the set of here predicted hairpins and three of these had *L* score of 1.0 [see Additional file 6]. The 15 unique loci in MDV with an *L* score of 1.0 collapsed into eight unique sequences due to a large inverted repeat. Out of these eight, all five loci that did not overlap with annotated exons corresponded to known miRNA loci.

The two examples show that most hairpins in viral genomes with high *L* score are true miRNA hairpins and that the absolute number of hairpin loci with high *L* scores is small. This allowed us to manually examine the remaining, non-miRNA loci with high *L* scores, using additional filtering criteria for genomic location such as proximity to known miRNAs [2] and intronic position [44]. Among the remaining loci in MDV with $L \geq 0.05$, one appeared to be located in the transcribed strand of an intron and two others closely flanked (0.3 kb) the mdv1-mir-M1 gene in the same orientation [see Additional file 6]. Similarly, out of the remaining ten EBV loci with $L = 1.0$, two candidate miRNAs were located directly upstream and amidst a cluster of eleven known miRNAs in an intronic region of the BART gene [46] [see Additional file 5].

Table 4: Identified (miRNA) hairpins in genomes of C. elegans and four viruses

Organism	Identified miRNA hairpins [a]	Identified genomic hairpins	miRNA hairpins % (number) [b]			genomic hairpin loci %(number) [b]		
			L = 1.0	L ≥ 0.05	L ≥ 1e-5	L = 1.0	L ≥ 0.05	L ≥ 1e-5
C. elegans	128/132	3,526,115	34 (45)	71 (94)	89 (117)	0.1 (3,110)	0.6 (21,313)	2.8 (98,309)
EBV	23/23	6,182	35 (8)	87 (20)	100 (23)	0.3 (20)	2.6 (162)	13 (793)
MDV	8/8	5,374	25 (2)	100 (8)	100 (8)	0.3 (15)	1.3 (69)	6.1 (329)
HCMV	10/11 [c]	9,747	40 (4)	90 (9)	100 (10)	0.3 (30)	2.5 (247)	12 (1,147)

[a] Identified miRNA hairpins/number of known miRNA hairpins in genome(s) in miRBase version 9.0 [44].

[b] Percentage of miRNA or genomic hairpin loci that have at least a certain *L*-score for scoring model Metazoa (absolute numbers between brackets).

[c] MiRNA hcmv-mir-UL148D could not be mapped on the genomic sequence of HCMV [EMBL: X17043].

[d] MiRNA kshv-mir-K12-10b could not be mapped on the genomic sequence of KSHV [EMBL: U75698HCMV].

Table 5: Mining the C. elegans genome for putative miRNA hairpins

	Rejected miRNA hairpins [a]	Clustered [b] (miRNAs/loci)		Similar [b] (miRNAs/loci)	
5 kb flanking sequence	NA	132	15,514	-	-
Similarity to mature miRNA [c]	NA	-	-	132	937
No overlap with exons [d]	2	130	11,197	130	706
L = 1.0	87	45	80	-	-
L >= 1e-5	15	-	-	116	197
Exclude TRF overlap [e]	0	45	74	-	-
Filter on 7 descriptors	6	-	-	116	162
Exclude known miRNA loci	132	0	20	0	64
similarity positioned correctly [f]	18	-	-	0	41

[a] Rejected *C. elegans* miRNA hairpins by this step alone. [b] Remaining cumulative number of *C. elegans* miRNA hairpins; remaining cumulative number of genomic hairpin loci. [c] Similarity to a metazoan mature miRNA of at least 19 nt length with at most three mismatches [d] Cel-mir-354 is fully located in the final exon of Y105E8A.16; cel-mir-356 is largely located in the 3'utr of ZK652.2, but overlaps 6 nt with the final exon. [e] At most 40 nt overlap with a repeat identified by Tandem Repeats Finder [47]. [f] Similarity with mature miRNA may have at most 3 nt overlap with hairpin loop coordinates and at least 8 nt of the stem must separate the similarity from the end of the hairpin. Criteria are set according to the biological model of miRNA maturation from hairpins [3].

Mining the C. elegans genome for putative miRNA hairpins

Using the scoring model Metazoa, the genome of *C. elegans* was searched for potentially novel miRNA hairpins. For the set of 3,525,115 genomic hairpins, a cut-off value $L \geq 0.05$ resulted in a reduction of the number of candidate miRNA loci to 21,158 (0.6%). For $L = 1$, only 3,099 hairpins loci remained, but the sensitivity measured from retrieval of known miRNA hairpins was only 34%. Nevertheless, this shows that filtering criteria in addition to the *L* score are required prior to experimental evaluation of candidate miRNAs. Such criteria may include the use of annotation data or genomic context. It has been suggested, for example, that metazoan miRNAs do not overlap with exons [11]. Indeed, in the annotation used here (Ensembl build 150), manual inspection of the genomic position of 132 known miRNAs of *C. elegans* showed that only two hairpins (2%) overlap with annotated exons (cel-mir-354, cel-mir-356, Table 5). Another very selective criterion is similarity to a known (metazoan) mature miRNA. We used this criterion such that similarity should be present in the stem of the hairpin and cover at least a 19 nt overlap with at most three mismatches. This resulted in a reduction to only 937 hairpin loci (Table 5).

Another property of many miRNAs is their clustered occurrence, with 45 from 132 miRNAs in *C. elegans* separated by less than 5 kb [10]. This criterion limited the number of hairpin loci to 15,514 (Table 5). An example of a less selective criterion is removing all hairpins that fall in highly repetitive genomic areas. A catalogue of repeat regions can be obtained by an algorithm as Tandem repeats finder [47]. Additional selection can be accomplished with threshold filtering using the same (or a subset of) descriptors already used in our scoring model. Besides translating miRNA hairpin properties into a statistic for relative rating, as done in calculating the *L* score, a descriptor can be used as a binary decision criterion: below the threshold, the hairpin is rejected as potential candidate; above it is included. Which descriptor(s) to use for further filtering is in the hands of the individual researcher and may depend on the data studied. In the following, we provide two examples of such filtering using seven individual descriptors: four based on structure (stem length <= 55, loop length <= 40, largest bulge <= 8, max match count >= 17) and three based on sequence complexity (polyNucHairpin <= 8, SCS-mono >= -10, SCS-di <= 0.40). Except for stem length, all individual thresholds fall within their *S* < 1 fractions and separately reject at most three *C. elegans* miRNA hairpins. This ensemble of thresholds excludes low-complexity sequences, captures palindromic repeats with limited degeneracy and should comprise the vast majority of miRNAs. In total, only six *C. elegans* miRNA hairpins fail this filter and the same sensitivity (96%) is achieved on all metazoan miRNA hairpins. Only 30% of the set of genomic hairpins in *C. elegans* passes this filter, representing a selectivity of 3.3 (data not shown).

In the following examples, filtering on *L* score was combined with filtering on either genomic context or on a similarity threshold to known metazoan mature miRNAs (we refer to these protocols as "Clustered" and "Similar"). Goal of these filtering protocols was to achieve sets feasibly sized for manual inspection and/or laboratory evaluation. The protocols comply with characteristics of miRNAs mentioned above: occurrence in clusters and presence in families. "Clustered" and "Similar" resulted in lists of 20 and 64 candidate miRNA loci [see Additional files 7 and 8], respectively, that were manually inspected. The most compelling cases among these are presented in Figure 10 and Figure 11.

Figure 10 shows a cluster of hairpins of which several have *L* = 1, starting 3 kb downstream of cel-mir-76. Figure 11 shows two conspicuous hairpins that share similarity with the two known mature miRNAs cel-mir-269 and cel-mir-266 (hairpins 1,165,306 and 2,047,661, with *L* scores of 0.030 and 0.007, respectively). The multiple alignment of the four hairpin sequences revealed the characteristic camel-shaped conservation pattern that is often observed between related miRNAs [28]: high or perfect conservation in the stem of the pre-miRNA hairpins, low conservation in the hairpin loop and the up-and downstream stem sequences. The 5' seed sequences of both mature miRNAs (position two to seven) are exactly conserved in the novel hairpins, suggesting that hairpins 1,165,306 and 2,047,661 are likely members of the miRNA families to which cel-mir-269 and cel-mir-266 belong [10]. Furthermore, candidate 1,165,306 is located in the 12th intron of the gene F54F11.2, for which the *C. elegans* unigene set (build 28) [48] provided evidence of transcription. Cel-mir-269 and cel-mir-266 were predicted by comparative computational approaches and confirmed by a PCR amplification protocol, but their precise mature miRNA ends are unknown [36,44]. Recent high-throughput sequencing of miRNAs from *C. elegans* [10,16] could not confirm the existence of both miRNAs.

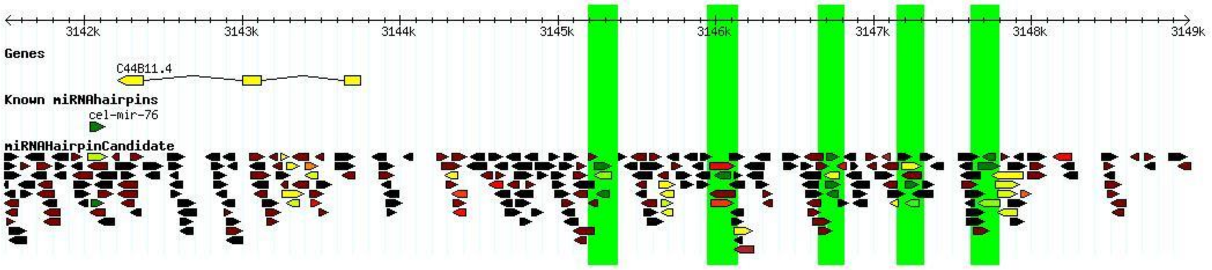
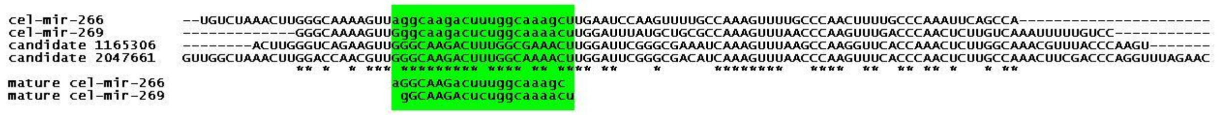


Figure 10. A cluster of candidate miRNA hairpins in *C. elegans* 3 kb upstream of cel-mir-76.

Five candidate miRNA hairpin loci with *L* score = 1 on chromosome III of *C. elegans*, selected by the filtering protocol Clustered. Loci are marked by green bars. Three out of five loci have hairpins with *L* score = 1 on both strands (positive strand: 3145224–3145336, 3146698–3146781, 3147197–3147283 and 3147660–3147798; negative strand: 3145240–3145320, 3145991–3146089, 3146703–3146775 and 3147690–3147767). The *L* score of genomic hairpins is indicated by a color gradient that ranges from dark green (*L* = 1) over yellow (*L* = 1e-4) and red (*L* = 5e-7) to black (*L* = 0).



```
cel-mir-266           --UGUCUAAACUUGGGCAAAAGUUaggcaagacuuuggcaaagcUUGAAUCCAAGUUUUGCCAAAGUUUUGCCCAACUUUUGCCCAAAUUCAGCCA----------------------
cel-mir-269           -------------GGGCAAAAGUUGggcaagacucuggcaaaacuUGGAUUUAUGCUGCGCCAAAGUUUAACCCAAGUUUGACCCAACUCUUGUCAAAUUUUUGUCC-----------
candidate 1165306     --------ACUUGGGUCAGAAGUUGGGCAAGACUUUGGCGAAACUUGGAUUCGGGCGAAAUCAAAGUUUAAGCCAAGGUUCACCAAACUCUUGGCAAACGUUUACCCAAGU-------
candidate 2047661     GUUGGCUAAACUUGGACCAACGUUGGGCAAGACUUUGGCAAAACUUGGAUUCGGGCGACAUCAAAGUUUAACCCAAGUUUCACCCAACUCUUGCCAAACUUCGACCCAGGUUUAGAAC
                                   ** *  * *** ********* **** ** **** **    *      ********   ****  **  ** **  *   * **

mature cel-mir-266                            aGGCAAGacuuuggcaaagc
mature cel-mir-269                             gGCAAGAcucuggcaaaacu
```

Figure 11. Candidate miRNA hairpins in *C. elegans* closely related to cel-mir-266 and cel-mir-269.

ClustalW alignment of the hairpin sequences of cel-mir-266, cel-mir-269 and the genomic hairpins 1,165,306 (chr I, 1733470..1733572 (+), *L* score = 0.030, 12th intron of F54F11.2) and 2,047,661 (chr II, 13515555..13515672 (+), *L* score = 7.2E-3, 7th intron of Y71G12B.11). The position of the mature miRNA sequences of cel-mir266 and cel-mir-269 (in lowercase) is projected on the sequences in green. Lowest two lines show again the mature miRNA sequences of cel-mir-266 (MIMAT0000325) and cel-mir-269 (MIMAT0000322), with their seed sequence in uppercase.

The data presented for the filtering protocols "Clustered" and "Similar" (Table 5) show that combined filtering on *L* score, genomic context and threshold filtering allows for compilation of a priority list of candidate miRNAs that is amenable to manual inspection and experimental verification. Apart from filtering on genomic clustering or similarity to known miRNAs, filtering on *L* score attains the largest data reduction. This shows that the *L* score was important in compilation of the priority list and demonstrates the added value of our approach for *in silico* miRNA prediction.

Discussion

We here present a new computational strategy for the *in silico* prediction of miRNA hairpins and show the applicability of the method for predicting new candidate miRNA hairpins in four viral genomes and in the genome of *C. elegans*. While using the latter as an example for model construction, the *L* score method as here optimized for *C. elegans* is well usable for other metazoan genomes. However, further improvement of performance of the method on different taxonomic groups can be achieved by constructing dedicated scoring models [see Additional file 9]. Our strategy aims at minimizing the number of false negative predictions (optimal sensitivity), rather than at minimizing the number of false positive predictions (optimal specificity), as proposed in previous studies. Focusing on sensitivity rather than specificity should help uncover new classes of miRNA molecules in biological systems. Hairpins in genomic sequences are identified with the help of an adjusted suffix-tree based method. When the performance of the hairpin prediction was benchmarked on sets of all known miRNAs hairpins from viruses and Metazoa, all 55 viral miRNAs were recovered (100%), 128 of 132 (97%) *C. elegans* miRNAs were recovered and 3,803 out of 3,902 (97.5%) metazoan miRNAs were shown to be recoverable when considered in their genomic context. Similar performance was reported for another edit distance-based hairpin identification method [26].

Four *C. elegans* miRNAs (cel-mir-262, cel-mir-260, cel-mir-272 and cel-mir-256) remained undetected due to the absence of a stringently base-pairing area in their stems. They all had extremely poor *L* scores (2.7e-41, 2.5e-27, 3.5e-20 and 3.8e-10), ranking first, third, fourth and eight among the *C. elegans* miRNA hairpins with lowest *L* scores. None of the four was found in recent high-throughput sequencing datasets, which otherwise retrieved the vast majority of known *C. elegans* miRNAs [10,16]. This suggests that these four may not be genuine miRNAs. If so, the reported performance of our hairpin prediction method is underestimated.

The biological variation and evolutionary diversity of various properties of miRNA hairpins were captured in a likelihood score *L*, based on statistics derived from accurately fitted (generally skewed normal) distributions of hairpin characteristics derived from known miRNA hairpins. In total 40 hairpin characteristics were defined and analyzed. The *L* score is a measure for a single hairpin sequence: a descriptor that captures evolutionary conservation in another species is not included. Although conservation has proven to be an extremely selective miRNA detection criterion [28], it conflicts with the aim to maximize sensitivity, because of the existence of species-specific miRNAs. The strategy was evaluated on genomic hairpins and known miRNAs from the *C. elegans* genome. Although details of analyses and results are likely to differ when applied to other genomes and other negative sets of genomic hairpins, major trends and results were shown to be similar [see Additional file 9].

Based on analyses of correlation and discriminative power, 18 hairpin characteristics were identified as most selective. Comparison of the 18 descriptor scoring model with a binary threshold filtering protocol using the same 18 descriptors (Table 3) shows that a 17% higher selectivity is achieved with the *L* score strategy. Binary decision thresholds on all or even a few descriptors can easily result in a major decrease of sensitivity. In the strategy developed here, $L < 1$ represents individual sequences that have one or more descriptors with $S < 1$, indicating that these descriptors have a relatively low probability of occurrence in miRNA hairpins because they occur in the tail(s) of their respective distribution. When a descriptor value falls outside the observed range of biological variation, the continuous likelihood score *L* allows for the compensation of an unlikely score for a single descriptor by likely scores for other descriptors. In such a case, the sequence is not *a priori* rejected as a miRNA hairpin candidate. The *L* score thus allows for more deviations from 'genuine' miRNA characteristics than binary selection (yes/no) on the basis of the same characteristics. This is an important improvement over the use of pre-defined thresholds for filtering on single or multiple descriptors published previously [19]. Assigning scores to descriptors, as opposed to binary selection on pre-defined thresholds, has also been used in previous work. For example, MIRscan [5,6] employs an heuristic score assignment to seven features and assigns weights based on the relative entropy between known miRNA hairpins and genomic hairpins. PalGrade [8] uses a statistical distribution by arbitrarily binning the ordered vector of descriptor values. SVM kernels achieve descriptor scoring and weighting as part of the SVM [35]. Novel to the approach here developed is that the score assignment is based entirely on the statistical evaluation of the variation in physical and sequence properties observed in known miRNAs and no binary selection prior to building a scoring model is used.

The *L* score is a relative measure of the likelihood that a given hairpin is a miRNA hairpin candidate. An important parameter contributing to useful *L* scores is the number of miRNAs used to derive the discriminating statistics of the individual characteristics. A minimum of about a thousand miRNA hairpins is required, but results indicate that the larger the available data set, the better the scoring model captures the variation in miRNA hairpin characteristics and performs. In the context of machine learning, the input miRNA set could be considered a training set, although in the *L* score derivation no formal 'training' is included. It is noteworthy that the evolutionary distance between the species contained in the taxonomic set for a scoring model influences *L* considerably. In general, the scoring model based on the set that is taxonomically closest to the organism being analyzed performs best. This indicates that over a wide range of characteristics, miRNA hairpins within (related) species are substantially more alike than miRNA hairpin sequences between less-related species. This should be taken into account when searching for similarity between miRNA hairpins from distantly related species. However, we observed that, for example, taxonomic sets that do not contain human miRNA hairpins can yield scoring models that accurately identify human miRNA hairpins. In addition, descriptor weighting and parameterization appeared to considerably influence the performance of a scoring model. The analyses presented allow the selection of optimal scoring models, with maximal discriminative power to distinguish true miRNA hairpins from other genomic (or random) hairpins for a selected set and a given data set. However, each set of data, for example in case of a new genome sequence, will require its own analysis to build an optimal scoring model.

Whereas the analyses started with 40 descriptors, based on extensive correlation and selectivity analyses, a subset model based on 18 descriptors was performing better than the model comprising all 40 descriptors. In view of the importance of the taxonomic composition of the set used in the model, it should be pointed out, that this may reflect a taxonomic bias for metazoan sequences that may not be valid for other taxonomic groups, such as, for example plant miRNA hairpins. In the set of 18 most informative descriptors selected, it is remarkable that three descriptors that are generally considered important are not represented: GC content and the MFE randomization descriptors P and Z. GC-content is widely used as a pre-filtering step of *in silico* miRNA prediction methods [36], but is not among the 18 descriptors used in the final scoring models. The GC-content showed the highest variability in the (skewed-normal) mean over different taxonomic sets (data not shown), reflecting the apparently large variation in GC-content found among miRNA hairpins from different species. Any scoring model that includes the fitted distribution of GC-content would disqualify hairpin structures in species containing miRNAs with relatively high or low GC-content. Also the descriptors P and Z, based on MFE randomization shown to be significantly lower for miRNAs hairpins than for randomized sequences [37,38], were not among the18 descriptors selected. Although P and Z ranked among the most selective descriptors, they were excluded because of their strong correlation [see Additional file 3] with the most selective descriptor MFEahl index, in which the MFE is adjusted for hairpin length and GC-content.

SVMs have similar aims as the *L*-score strategy and perform well in miRNA classification [35]. In assessing and comparing the performance of different methods, however, several caveats should be considered. First, methods that use evolutionary conservation perform well on conserved miRNAs [25,11] but fail to detect species-specific or fast evolving miRNAs [12]. The importance of the latter should not be underestimated. Second, the particular data set(s) on which the performance is achieved is important to the evaluation and comparison of the results from different methods. Unfortunately, there is no benchmark set of both positive and negative examples of miRNA hairpins available. Many methods are tailored on a specific organism and are likely to perform best in that exact context. Assembling a set of true negative sequences is particularly challenging: hairpins in non-coding RNAs, e.g. the set of tRNAs as used in [25], are likely to possess different and more diverse hairpin features than miRNA hairpins, whereas a set of genomic hairpins might contain *bona fide* miRNA hairpins. SVMs explicitly require negative examples for training and testing, whereas the *L* scoring method uses such a set only for benchmarking purposes. Third, data on prediction performance are not reported consistently in literature. Our method enables reporting of AUC performance as well as sensitivity and specificity values over the entire range of the ROC curve. Comparison with binary classifiers from other methods therefore requires transformation of the continuous outcome to a binary outcome by choosing an arbitrary threshold for *L* and using the associated sensitivity and specificity values as measure of the performance.

With these caveats in mind, we compared the performance of our method to three leading SVM-based methods *miPred* [35], RNAmicro [25] and miRNA SVM [26] Note that RNAmicro is based on multiple sequence alignments. Using different positive and negative datasets, these methods report the following values of sensitivity and specificity; 1) *MiPred*: 86.69% and 97.68%, using 323 human miRNAs as positive and 646 human genomic hairpins as negative set; 2) *MiPred*: 87.65% and 97.75%, using 1,918 Metazoan, non-human miRNAs as positive and 3,836 human genomic hairpins as negative set; 3) RNAmicro: 90% and 99%, using 147 Metazoan miRNA hairpin alignments as positive and 383 shuffled miRNA hairpin and tRNA alignments as negative set, and; 4) miRNA SVM: 90% and 95%, using 322 human miRNAs as positive and 3,000 random human genomic hairpins as negative set. These performances compare well to the values of 87.26% and 97.02% obtained in the 10 fold-cross validated performance of our *L* score model, using 203 Metazoan miRNA hairpins as positive and 200,000 randomly selected genomic hairpins from *C. elegans* as negative set. The performance of the *L* scoring model is most similar to that of *miPred*, which does not include sequence conservation as a parameter. An analysis of the performance of our model on sets of genomic hairpins other than those derived for the *C. elegans* genome is provided as Additional file [see Additional file 9].

A major challenge of any SVM is understanding its behavior in, for example, a biological context. While SVMs are known to produce classifiers that perform well in case of unseen data [49], SVMs are essentially black-box classifiers. This makes it difficult to judge the relative importance of individual descriptors or to translate results in biologically relevant understanding. As all parameters are embedded in the kernel function of the SVM, SVM classifiers are also difficult to adjust, although they do not require the pre-selection of parameters required for the *L* score strategy here presented. Classifier selection based on the detailed descriptor analysis presented here may improve future SVM approaches. A key advantage of the *L* score strategy over SVMs is that the contribution of individual descriptors to a scoring model can be analyzed in a straightforward way by adding or removing a descriptor or changing the parameterization or weight of descriptors. This way the scoring model becomes better tailored to the biologist's needs in a particular research environment.

Conclusion

At the laboratory bench, the criterion that is of most interest is simply how many putative hairpins should be evaluated by experimentation. With current developments in microarray analysis and high-throughput sequencing, the numbers of potential candidates that can be screened with relative ease will increase dramatically. Still large numbers of new miRNAs may be identified. Yet, for the time being, the individual laboratory would like to see as little putative candidates as possible with as high a success rate as feasible. The highest *L* score is 1, implying that the given hairpin scores are maximal for all descriptors in the model. In the 100 Mb large genome of *C. elegans*, still 3,110 hairpin loci remain that cannot be ranked further on the basis of *L*. It implies that relatively large numbers of genomic hairpins (3,110 loci from 3,526,115 hairpins, i.e. 0.09%; see Table 4) comply with all miRNA hairpin descriptors, whereas it is unlikely that they all generate mature miRNAs, given the relative small number of 132 currently known *C. elegans* miRNAs. It is likely that this situation will occur in most genomic contexts. If so, several strategies are open. The model parameters could be adjusted, so that the individual descriptor is less likely to get the maximal score. This way, the *L* score approach will convert to more traditional threshold filtering. Given that the analysis requires high sensitivity, it would however be more advantageous to incorporate more biological expert knowledge in the selection process, such as the presence of the hairpin in an intron, sequence similarity to known miRNAs, etc. When following this strategy, one has to be aware that the number of novel miRNAs that could be discovered is constrained by the filtering on genomic context. For example, in the set of 132 miRNAs in *C. elegans* (miRBase 9.0), 87 occur in singletons when clustered on distance, 56 have no other family representative and 102 are not located in an intron [10]. These numbers can thus be considered as indicative for the fraction of true miRNAs that will remain concealed when filtering on genomic context. With the filtering protocols ("Clustered" and "Similar") we show that a combination of filtering on *L* score, genomic context and threshold filtering allows for compilation of a priority list of manageable size for manual inspection and further experimentation.

In addition to good performance in comparison with other leading (SVM-based) methods and a user-defined selectivity, an additional advantage of the *L* score approach over threshold filtering and support vector machine classifiers is that the prior analysis of taxonomically defined sets and fitted distributions, correlations, and discriminative power of descriptors gives detailed insight in the behavior of a scoring model and can accommodate expert knowledge. It should therefore appeal to the experimental biologist, despite the fairly time-consuming construction of a suitable scoring model.

The scoring model proposed here is independent of the hairpin prediction step and can therefore be coupled to any *in silico* or experimental miRNA prediction method. It can facilitate the analysis of large sets of putative miRNA hairpin loci obtained in deep-sequencing efforts of small RNAs [10,14-16]. The *L* score approach can be used to rank and select interesting miRNA hairpin candidates for downstream experimental analysis in search for novel miRNAs. Moreover, our in-depth analyses of known miRNA hairpins from miRBase [44], our detailed descriptor analyses (Figure 9) and the *L* score approach here presented are likely to increase the reliability and evidence of miRBase entries and will help to further increase the biological relevance of the miRBase repository.

Methods

Sequence and annotation data

The complete set of 3,498 non-plant miRNA hairpin sequences were retrieved from the web resource miRBase version 9.0 [44]. In addition, 474 miRNA hairpin sequences from human and chimpanzee [12] and 18 from *C. elegans* [10] were obtained from the supplementary material of the respective publications. Secondary structures of the sequences were predicted using RNAfold version 1.6 [21] with the constrained folding option (-C) used to position the mature miRNA sequence(s) in the stem of the hairpin. The hairpin structure of six sequences, three from miRBase and three from the human/chimpanzee set [12], deviated considerably from the predicted characteristics of miRNA hairpins. These six sequences were therefore excluded from all subsequent analyses [see Additional file 10]. The resulting 3,984 miRNA hairpin sequences were included in this study, 3,902 from Metazoa and 82 from virus genomes. Sequence and annotation of the *C. elegans* genome (build 150) was obtained from Ensembl [50]*C. elegans* unigenes (build 28) were downloaded from NCBI UniGene [48]. Viral genome data for the Epstein-Barr virus [EMBL: AJ507799], Human cytomegalovirus [EMBL: X17403], Kaposi sarcoma-associated herpesvirus [EMBL: U75698] and Mareks disease virus [EMBL: AF243438] were obtained from EMBL [51].

Informatics and statistics

Supplemental data are available through the Additional data files and the accompanying web document *µRNALL*, which can be downloaded at http://appliedbioinformatics.wur.nl/murnall/[42]. All statistical analyses were performed using the package R [52], as integrated in python through Rpy [53].

Definition of miRNA descriptors

A set of 40 potentially discriminative features of miRNA hairpins, hereafter referred to as descriptors, was defined based on the set of 3,984 miRNA hairpins. The descriptors include both physical and sequence characteristics of miRNA hairpins [Table 1; see Additional files 1 and 2]. A subset of descriptors is given in Table 1. To take the evolutionary diversity of the descriptors into account in the statistical analyses, miRNA sequences were divided in hierarchically organized subsets based on their taxonomic relationships. Taxonomic sets that comprised at least 100 sequences were used for analysis. In total, 23 taxonomic sets were defined [see Additional file 4], including one set containing all metazoan miRNA sequences, eleven sets representing metazoan taxa and eleven species-specific sets. The virus set was not used because it contained only 82 sequences. Unless stated otherwise, all results presented in this paper use the combined taxonomic set 'Metazoa' (3,902 miRNAs).

Individual likelihood score S for each descriptor

For all 3,902 sequences, descriptor values were calculated and their distributions within each of the 23 taxonomic sets were fitted to an appropriate probability distribution. Goodness-of-fit was determined by a Chi-square test. For each distribution, the probability that the descriptor takes a value less than or equal to a specified value was calculated as the cumulative distribution function (CDF) and transformed into a likelihood distribution function (LDF). For the LDF, a default cut-off value was set at 0.05, corresponding to the 95% confidence interval of the fitted distribution of the descriptor. For each descriptor, values of the CDF above the cut-off value were transformed to the LDF likelihood score $S = 1$. Values below the cut-off were transformed to the likelihood score S = (CDF/cut-off). Table 1 and Additional file 2 list for each descriptor whether the lower tail of the CDF, upper tail or both were transformed. As a result of this transformation, each descriptor in the taxonomic set has a likelihood distribution S comprising an $S < 1$ and an $S = 1$ fraction. $S = 1$ indicates a descriptor for which characteristics of the individual sequence are in 95% of the distribution.

Likelihood score L for the combined descriptor values

To obtain a single metric for a given taxonomic set, the likelihood scores S for all descriptors were multiplied to obtain the combined likelihood score L. The ensemble of likelihood scores S for a given set of hairpin sequences is referred to as the scoring model. L is the outcome of the scoring model and functions as classifier for miRNA hairpin sequences. L ranges between 0 and 1 and represents the likelihood of a hairpin sequence to be a true miRNA hairpin given the underlying descriptors used in the scoring model. It is possible to incorporate additional expert knowledge in the scoring model by assigning a relative weight to the S score of an individual descriptor. In the default setting reported here, no difference between descriptors is made (assigned weight = 1). An L score of 1.0 for a hairpin sequence indicates that $S = 1$ for each descriptor of the set. L is only affected by descriptors with a value $S < 1$.

Correlation and discriminative power of descriptors

To prevent potential over-penalization of hairpin sequences when combining correlated descriptors, we determined the independence (orthogonality) of all descriptors in the $S < 1$ fraction by calculating Cohen's kappa [41] for each combination of descriptors. The value $\kappa = 0$ indicates that there is no more correlation between descriptors than expected by chance alone, and $\kappa = 1$ indicates that the descriptors are fully dependent. The discriminative power of a descriptor, i.e. its ability to distinguish true miRNA hairpins from non-miRNA hairpins, was calculated as the ratio of percentages of miRNA hairpins and genomic hairpins that comply with a given threshold for this descriptor. As threshold the descriptor's limiting value between *S* = 1 and *S* < 1 of the LDF was chosen (95% of the CDF). Discriminative power was calculated using known miRNA hairpins from the taxonomic set Metazoa and genomic hairpins from a set of 3,526,115 hairpins identified in *C. elegans* (see section Identification of putative miRNA hairpin structures). It is calculated with the formula for selectivity (see below), but for the sake of clarity we will here use the term 'discriminative power' for the performance of a single descriptor and the term 'selectivity' for the performance of a scoring model.

Descriptor selection and model evaluation

To select a subset of descriptors that was most informative for the combined assessment of miRNA hairpins by the *L* classifier, descriptors that either correlated with a more discriminative descriptor (κ>0.4) or that showed low discriminative power (< 1.1) were discarded from the initial set. The resulting subset was used to evaluate the impact of different settings of variables. For all models, *L* scores were calculated for 100,000 randomly selected hairpins from the *C. elegans* genome. We evaluated (1) the effect of the size of the input set, which refers to the number of miRNA hairpins in a given taxonomic set; (2) the impact of evolutionary distance between taxa; (3) the impact of different combinations of descriptors in a scoring model; (4) the effect of parameterization of descriptors and (5) the effect of weighting of descriptors.

Performance of L

The performance of the outcome classifier *L* of scoring models was measured in two ways. First, by the area under a receiver operating characteristic (ROC) curve [54]. Trapezoids were constructed as approximation of the Area Under the Curve (AUC). Unless described otherwise, ROC curves were made for the taxonomic set of metazoan miRNA hairpins (3,902) versus 200,000 randomly selected hairpins from the *C. elegans* genome. Second, sensitivity, specificity, and selectivity were calculated for each scoring model from the counts of true and false positive and negative cases (TP, FP, TN, and FN, respectively) in the following way:

Sensitivity = (100∗)TP /(TP + FN)

Specificity = (100∗)TN /(TN + FP)

Selectivity = Sensitivity /(100 - Specificity)

TP and FN were counted from taxonomic sets of known miRNA hairpins, TN and FP were determined as a fraction of genome-wide identified hairpins. Although these sets of genomic hairpins contained an unknown number of true miRNAs (so FN and TP), this number was expected to be sufficiently small to be ignored. For uniform comparison, we benchmarked selectivity at discrete values of sensitivity (95% and/or 75%). Discrete points on the ROC curve correspond to pairs of sensitivity and specificity values, and as such describe the shape of the curve.

The performance of the classifier *L* was compared with sensitivity, specificity and selectivity of threshold filtering on descriptors of miRNA and genomic hairpins. For the 18 most informative descriptors, the threshold used did represent the same cut-off value between the *S* = 1 and *S* < 1 fraction of the LDF, at 95% of the CDF at the side(s) of the distribution as listed in Table 1. This cut-off was used as a binary decision criterion: below the threshold, the (miRNA) hairpin was included; above it was rejected.

A 10-fold cross-validation was performed on 200,000 randomly chosen genomic hairpins and repeated ten times. As input set, a non-redundant variant of the taxonomic set Metazoa was constructed. This involved clustering miRNA hairpins with identical mature miRNA seed sequences and an overall hairpin sequence identity larger than 90%. All but a single representative for each cluster were then removed, yielding a subset of 2,033 sequences.

Identification of putative miRNA hairpin structures

The suffix-tree based tool VMatch [43] was used to identify small genomic hairpin structures in the genomes of *C. elegans* and four viruses, using a sliding window of 1,000 nt with an overlap of 200 nt. The latter value exceeds the length of the largest known metazoan miRNA hairpin (153 nt). Each sequence window was stored as a VMatch database (index) and its reverse complement was used as query sequence in a VMatch search for degenerate palindromic sequences, allowing GU-base pairing. Parameter settings that allowed exhaustive retrieval of known miRNA hairpins were found empirically (data not shown). Such a palindromic sequence consists of two inverse complementary sequences for the stem, at a physical distance representing the loop of a putative hairpin. Palindromes were discarded if the distance was larger than 50 nt, which represents the upper limit of loop size in the vast majority of metazoan miRNAs. Overlapping palindromes were merged if they had at most 8 non-overlapping nucleotides on either side. With these parameter settings, only a small number of known miRNA hairpins was missed. The remaining set of palindromic sequences was used for secondary structure prediction using RNA-fold version 1.6 with the constrained folding option (-C) to enforce the stem structure in the folding of the molecule [21]. All hairpin structures were filtered for five threshold values: (1) minimal hairpin length = 45 nt; (2) minimal number of base pairs in the stem = 15; (3) minimal number of paired bases in the most stringently paired window of 24 positions in the hairpin stem = 15; (4) maximum length of a bulge in the stem = 29 nt; (5) minimal ratio of the number of paired positions divided by all positions in the stem (match-ratio) = 0.45.

Grouping identified genomic hairpins into unique loci

Many of the genomic hairpins identified were overlapping or nested. Such hairpins were grouped into unique loci when the centers of their loops were less than 20 nt apart, regardless of the strand on which the hairpins were located.

Abbreviations

AUC: Area Under the (ROC) Curve; CDF: cumulative distribution function; EBV: Epstein-Barr Virus; HCMV: Human cytomegalovirus; KSHV: Kaposi sarcoma-associated herpesvirus; LDF: likelihood distribution function; MDV: Mareks disease Virus; MFE: minimal folding energy (kcal/mol); miRNA: microRNA; ROC: Receiver Operating Characteristic; RSSP: RNA secondary structure prediction; siRNA: small-interfering RNA; SN: Skewed Normal distribution; SVM: Support Vector Machine.

Authors' contributions

AvdB designed the study, performed the programming and analyses, drafted the manuscript and added all additional analyses for the revision. MF participated in the design of the study and supported the computational work. JPN and RvH contributed to the design of the study, supervised and participated in the preparation of the manuscript. All authors were involved in discussions, have read and approved the final manuscript.

Acknowledgements

We thank prof. Cajo ter Braak (Biometris, WUR) for helpful discussions regarding the statistics of scoring models and their interpretation.

Additional material

Additional file 1: Data fit of descriptors.
Results of the data fit for 40 descriptors from the taxonomic set Metazoa (3,902 miRNA hairpins).
http://www.biomedcentral.com/content/supplementary/1471-2164-10-204-S1.pdf

Additional file 2: Detailed explanation of descriptors.
A set of 40 potentially discriminative features of miRNA hairpins, referred to as descriptors, was defined and includes both physical and sequence characteristics of miRNA hairpins.
http://www.biomedcentral.com/content/supplementary/1471-2164-10-204-S2.pdf

Additional file 3: Descriptor interdependency.
Correlation among descriptors in their S<1 fractions, assessed by Cohen's kappa coefficient κ of all 780 possible pairs of descriptors, using the miRNA hairpins of the taxonomic set Metazoa.
http://www.biomedcentral.com/content/supplementary/1471-2164-10-204-S3.pdf

Additional file 4: Taxonomic sets with at least 100 miRNA sequences.
MiRNA sequences were divided in hierarchically organized subsets based on their taxonomic relationships. A total of 23 taxonomic sets comprised at least 100 sequences and were used for analysis.
http://www.biomedcentral.com/content/supplementary/1471-2164-10-204-S4.pdf

Additional file 5: Hairpins identified in Epstein-Barr Virus
Details of 23 hairpins with L score = 1.0, identified in Epstein-Barr Virus [EMBL: AJ507799] for the scoring model Metazoa.
http://www.biomedcentral.com/content/supplementary/1471-2164-10-204-S5.pdf

Additional file 6: Hairpins identified in Mareks disease Virus.
Details of 18 hairpins with L >= 0.30 identified in Mareks Disease Virus [EMBL: AF243438] for the scoring model Metazoa.
http://www.biomedcentral.com/content/supplementary/1471-2164-10-204-S6.pdf

Additional file 7: Hairpin loci in C. elegans obtained by the filtering protocol "Clustered".
Filtering on L score was combined with filtering on genomic context, a protocol referred to as "Clustered".
http://www.biomedcentral.com/content/supplementary/1471-2164-10-204-S7.pdf

Additional file 8: Hairpin loci in C. elegans obtained by the filtering protocol "Similar".
Filtering on L score was combined with filtering on a similarity threshold to known metazoan mature miRNAs, a protocol referred to as "Similar".
http://www.biomedcentral.com/content/supplementary/1471-2164-10-204-S8.pdf

Additional file 9: Performance of the scoring model Metazoa.
Analysis of the performance of the scoring model Metazoa on sets of genomic hairpins other than those derived for the C. elegans genome.
http://www.biomedcentral.com/content/supplementary/1471-2164-10-204-S9.pdf

Additional file 10: MiRNAs that excluded from the analyses.
The hairpin structure of six sequences deviated considerably from the predicted characteristics of miRNA hairpins and were excluded from all subsequent analyses.
http://www.biomedcentral.com/content/supplementary/1471-2164-10-204-S10.pdf

References

1. Bartel DP: MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004, 116(2):281-297.
2. Pfeffer S, Sewer A, Lagos-Quintana M, Sheridan R, Sander C, Grasser FA, van Dyk LF, Ho CK, Shuman S, Chien M, *et al.*: **Identification of microRNAs of the herpesvirus family.** *Nature methods* 2005, **2(4):**269-276.
3. Zeng Y, Cullen BR: **Efficient processing of primary microRNA hairpins by Drosha requires flanking nonstructured RNA sequences.** *The Journal of biological chemistry* 2005, 280(30):27595-27603.
4. Berezikov E, Plasterk RH: Camels and zebrafish, viruses and cancer: a microRNA update. *Human molecular genetics* 2005, 14(Spec No 2):R183-190.
5. Lim LP, Lau NC, Weinstein EG, Abdelhakim A, Yekta S, Rhoades MW, Burge CB, Bartel DP: **The microRNAs of Caenorhabditis elegans.** *Genes & development* 2003, **17(8):**991-1008.
6. Lagos-Quintana M, Rauhut R, Meyer J, Borkhardt A, Tuschl T: **New microRNAs from mouse and human.** *RNA (New York, NY)* 2003, **9(2):**175-179.
7. Hertel J, Lindemeyer M, Missal K, Fried C, Tanzer A, Flamm C, Hofacker IL, Stadler PF: **The expansion of the metazoan micro-RNA repertoire.** *BMC genomics* 2006, **7:**25.
8. Bentwich I, Avniel A, Karov Y, Aharonov R, Gilad S, Barad O, Barzilai A, Einat P, Einav U, Meiri E, *et al.*: **Identification of hundreds of conserved and nonconserved human microRNAs.** *Nature genetics* 2005, **37(7):**766-770.
9. Beuvink I, Kolb FA, Budach W, Garnier A, Lange J, Natt F, Dengler U, Hall J, Filipowicz W, Weiler J: **A novel microarray approach reveals new tissue-specific signatures of known and predicted mammalian microRNAs.** *Nucleic acids research* 2007, **35(7):**e52.
10. Ruby JG, Jan C, Player C, Axtell MJ, Lee W, Nusbaum C, Ge H, Bartel DP: **Large-scale sequencing reveals 21U-RNAs and additional microRNAs and endogenous siRNAs in C. elegans.** *Cell* 2006, 127(6):1193-1207.
11. Stark A, Kheradpour P, Parts L, Brennecke J, Hodges E, Hannon GJ, Kellis M: **Systematic discovery and characterization of fly microRNAs using 12 Drosophila genomes.** *Genome research* 2007, **17(12):**1865-1879.
12. Berezikov E, Thuemmler F, van Laake LW, Kondova I, Bontrop R, Cuppen E, Plasterk RH: **Diversity of microRNAs in human and chimpanzee brain.** *Nature genetics* 2006, **38(12):**1375-1377.
13. Lindow M, Krogh A**: Computational evidence for hundreds of non-conserved plant microRNAs.** *BMC genomics* 2005, 6:119.
14. Rajagopalan R, Vaucheret H, Trejo J, Bartel DP: **A diverse and evolutionarily fluid set of microRNAs in Arabidopsis thaliana.** *Genes & development* 2006, **20(24):**3407-3425.

15. Fahlgren N, Howell MD, Kasschau KD, Chapman EJ, Sullivan CM, Cumbie JS, Givan SA, Law TF, Grant SR, Dangl JL, *et al.*: **High-throughput sequencing of Arabidopsis microRNAs: evidence for frequent birth and death of MIRNA genes.** *PLoS ONE* 2007, **2(2):**e219.
16. Zhang L, Ding L, Cheung TH, Dong MQ, Chen J, Sewell AK, Liu X, Yates JR 3rd, Han M: **Systematic identification of C. elegans miRISC proteins, miRNAs, and mRNA targets by their interactions with GW182 proteins AIN-1 and AIN-2**. *Molecular cell* 2007, 28(4):598-613.
17. Tyler DM, Okamura K, Chung WJ, Hagen JW, Berezikov E, Hannon GJ, Lai EC**: Functionally distinct regulatory RNAs generated by bidirectional transcription and processing of microRNA loci.** *Genes & development* 2008, 22(1):26-36.
18. Stark A, Bushati N, Jan CH, Kheradpour P, Hodges E, Brennecke J, Bartel DP, Cohen SM, Kellis M: **A single Hox locus in Drosophila produces functional microRNAs from opposite DNA strands.** *Genes & development* 2008, **22(1):**8-13.
19. Yoon S, De Micheli G: **Computational identification of microR-NAs and their targets.** *Birth Defects Res C Embryo Today* 2006, 78(2):118-128.
20. Lindow M, Gorodkin J: **Principles and limitations of computa-tional microRNA gene and target finding.** *DNA and cell biology* 2007, 26(5):339-351.
21. Hofacker IL: **Vienna RNA secondary structure server.** *Nucleic acids research* 2003, **31(13):**3429-3431.
22. Zuker M: **Mfold web server for nucleic acid folding and hybridization prediction.** *Nucleic acids research* 2003, 31(13):3406-3415.
23. Lai EC, Tomancak P, Williams RW, Rubin GM: **Computational identification of Drosophila microRNA genes.** *Genome biology* 2003, **4(7):**R42.
24. Sewer A, Paul N, Landgraf P, Aravin A, Pfeffer S, Brownstein MJ, Tuschl T, van Nimwegen E, Zavolan M: **Identification of clustered microRNAs using an ab initio prediction method.** *BMC bioin-formatics* 2005, **6:**267.
25. Hertel J, Stadler PF: **Hairpins in a Haystack: recognizing micro-RNA precursors in comparative genomics data.** *Bioinformatics (Oxford, England)* 2006, 22(14):e197-202.
26. Helvik SA, Snove O Jr, Saetrom P**: Reliable prediction of Drosha processing sites improves microRNA gene prediction**. *Bioinformatics (Oxford, England)* 2007, 23(2):142-149.
27. Nam JW, Shin KR, Han J, Lee Y, Kim VN, Zhang BT: **Human micro-RNA prediction through a probabilistic co-learning model of sequence and structure.** *Nucleic acids research* 2005, 33(11):3570-3581.
28. Berezikov E, Guryev V, Belt J van de, Wienholds E, Plasterk RH, Cuppen E: **Phylogenetic shadowing and computational identifica tion of human microRNA genes.** Cell 2005, 120(1):21-24.
29. Lindow M, Jacobsen A, Nygaard S, Mang Y, Krogh A: **Intragenomic matching reveals a huge potential for miRNA-mediated regulation in plants.** PLoS computational biology 2007, 3(11):e238.
30. Baskerville S, Bartel DP: **Microarray profiling of microRNAs reveals frequent coexpression with neighboring miRNAs and host genes.** RNA (New York, NY) 2005, 11(3):241-247.
31. Brennecke J, Cohen SM: **Towards a complete description of the microRNA complement of animal genomes.** Genome biology 2003, 4(9):228.
32. Llave C, Xie Z, Kasschau KD, Carrington JC: **Cleavage of Scarecrow-like mRNA targets directed by a class of Arabidopsis miRNA.** Science (New York, NY) 2002, 297(5589):2053-2056.
33. Piriyapongsa J, Marino-Ramirez L, Jordan IK: Origin and evolution of human microRNAs from transposable elements. Genetics 2007, 176(2):1323-1337.
34. Xue C, Li F, He T, Liu GP, Li Y, Zhang X: **Classification of real and pseudo microRNA precursors using local structure-sequence features and support vector machine.** BMC bioinformatics 2005, 6:310.
35. Ng KL, Mishra SK: **De novo SVM classification of precursor microRNAs from genomic pseudo hairpins using global and intrinsic folding measures.** Bioinformatics (Oxford, England) 2007, 23(11):1321-1330.
36. Grad Y, Aach J, Hayes GD, Reinhart BJ, Church GM, Ruvkun G, Kim J: **Computational and experimental identification of *C. elegans* microRNAs.** Molecular cell 2003, 11(5):1253-1263.
37. Freyhult E, Gardner PP, Moulton V: **A comparison of RNA folding measures.** BMC bioinformatics 2005, 6:241.
38. Bonnet E, Wuyts J, Rouze P, Peer Y Van de: **Evidence that micro-RNA precursors, unlike other non-coding RNAs, have lower folding free energies than random sequences.** Bioinformatics (Oxford, England) 2004, 20(17):2911-2917.

39. Ng Kwang Loong S, Mishra SK: **Unique folding of precursor microRNAs: quantitative evidence and implications for de novo identification.** RNA 2007, 13(2):170-187.
40. Azzalini A, Capitanio A: **Statistical applications of the multivariate skew normal distribution.** Journal of the Royal Statistical Society: Series B (Statistical Methodology) 1999, 61(3):579-602.
41. Cohen J: **A Coefficient of Agreement for Nominal Scales.** Educational and psychological measurement 1960, 20(1):37.
42. Applied Bioinformatics (PRI/WUR) [http://appliedbioinformatics.wur.nl/murnall/]
43. Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, Giegerich R: **REPuter: the manifold applications of repeat analysis on a genomic scale**. Nucleic acids research 2001, 29(22):4633-4642.
44. Griffiths-Jones S: **miRBase: the microRNA sequence database.** Methods in molecular biology (Clifton, NJ) 2006, 342:129-138.
45. Yao Y, Zhao Y, Xu H, Smith LP, Lawrie CH, Watson M, Nair V: **MicroRNA profile of Marek's disease virus-transformed T-cell line MSB-1: predominance of virus-encoded microRNAs.** Journal of virology 2008, 82(8):4007-4015.
46. Cai X, Schafer A, Lu S, Bilello JP, Desrosiers RC, Edwards R, Raab-Traub N, Cullen BR: **Epstein-Barr virus microRNAs are evolutionarily conserved and differentially expressed.** PLoS pathogens 2006, 2(3):e23.
47. Benson G: **Tandem repeats finder: a program to analyze DNA sequences.** Nucleic acids research 1999, 27(2):573-580.
48. Wheeler DL, Church DM, Federhen S, Lash AE, Madden TL, Pontius JU, Schuler GD, Schriml LM, Sequeira E, Tatusova TA, et al.: **Database resources of the National Center for Biotechnology.** Nucleic acids research 2003, 31(1):28-33.
49. Schölkopf B: **Support Vector Learning** R. Oldenbourg Verlag, Munich; 1997.
50. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, et al**.: Ensembl 2008.** Nucleic acids research 2008:D707-714.
51. Kulikova T, Akhtar R, Aldebert P, Althorpe N, Andersson M, Baldwin A, Bates K, Bhattacharyya S, Bower L, Browne P, et al.: **EMBL Nucleotide Sequence Database in 2006.** Nucleic acids research 2007:D16-20.
52. The R Project for Statistical Computing [http://www.Rproject.org/]
53. RPy (R from Python) [http://rpy.sourceforge.net/index.html]
54. Hanley JA, McNeil BJ: **The meaning and use of the area under a receiver operating characteristic (ROC) curve.** Radiology 1982, 143(1):29-36.

Chapter 8

General discussion

My thesis project was part of an initiative of the graduate school Experimental Plant Sciences (EPS) to stimulate bioinformatics-related research projects, within the framework of EPS 'Strategische middelen' and Plant Research International (PRI) 'Kennisbasis-financiering'. Wageningen University (WU) and Research Centre implemented the full spectrum of high throughput omics technologies in the infrastructure of its laboratories to be used in large scale research programs that focus on the biology of important agricultural species. However, the challenge is to extract and interpret the biological information from the data and to integrate this information into a comprehensive knowledge base on the functioning of cellular and organismal systems. To strengthen this effort, bioinformatic recourses are required for the handling and integration of large datasets generated with modern omics technologies. Some of the prioritized areas included plant and fungal comparative genomics (comparative studies at the genome level in areas such as gene structure, regulatory networks or metabolic pathways) and general omics data management, standardization and integration. A further prerequisite was that the development of methodology, algorithms, and software implementations must be useful for both experimentalists and bioinformaticians. Because collaborations between research groups from WU and PRI were given high priority, a twinning project `Comparative Fungal Genomics` was granted that comprised two PhDs, each based in one of the research groups.

In search for *in silico* evidence of existence of miRNAs in fungi

I started to study the occurrence and possible regulatory roles of miRNAs that were just discovered. Already in 2006 Andy Fire and Craig Mello won the Nobel Prize for their work on RNA interference (RNAi), for work mainly done on the nematode *C. elegans*, underlining its enormous impact. In chapter 7 we describe a novel and flexible approach in ranking the likeliness that a given hairpin structure represents a miRNA hairpin. We showed that our approach can assign miRNAs in *C. elegans*, *D. melanogaster* and various human viruses with great accuracy. By choosing an optimal Likeliness score *L*, known miRNAs as well as several highly likely novel miRNA candidates were assigned. Unfortunately, the potential to maturate functional miRNAs from these candidate miRNA hairpins were never tested *in vitro* or *in vivo*. This study on miRNAs was initiated for a different reason. Our hypothesis was that fungi would encode miRNAs. Some of the pathogenic fungi could even encode miRNAs that are targeted towards host defence genes (plant or mammal). Most Ascomycete fungi encode the full protein potential for one or more silencing pathways [1]; most of their genomes encode two distinct Dicer-like and a variable number of Argonaute proteins. Existence of a functional silencing pathways (including quelling and meiotic silencing by unpaired DNA) and the requirement of one of these Dicer-like proteins in these pathways had some experimental support [2],[3]. However, comparative genomics of fungal sequence data, available at that time, showed a considerable diversification of proteins known to be involved in the RNA silencing machinery [1].

The *L* score approach and the exhaustive mining of genomic hairpins applied to fungal genomes. Exhaustive sets of genomic hairpins were mined in various fungal genomes (*Neurospora crassa*, *Zymoseptoria tritici*, *Fusarium graminearum* and others) and several custom-made scoring models were employed to calculate the *L* score. Various models were taken into account to prevent the *L* score to be tailored on animal or plant miRNA hairpins, with the purpose to increase the chance to find the unexpected. The results were surprising. Hundreds of hairpins per fungal species were assigned at thresholds that yielded the majority of known metazoan hairpins. Even the most stringent filtering (*L*=1) yielded tens to a hundred of hairpins per fungal genome. Next, we tried to prioritize these hairpins by looking for conservation. Since many metazoan and plant miRNAs regulate developmental processes (for reviews see [4],[5]) by targeting highly conserved genes. As a consequence, these miRNAs are near-identical even at the complete hairpin level among distantly related species. Imposing a conservation threshold criterion on this, requiring a complete hairpin or a 21nt segment to be somewhat conserved over several species, reduced the number of miRNA candidates to zero. Analyses following the same line of thought, but applied by others in various *Aspergillus* species, yielded the same outcome, *i.e.* not a single conserved miRNA-like hairpin structure was detected [6]. Additional searches allowing for faint similarity and subsequent manual inspection of all resulting pairwise and multiple alignments yielded a few interesting candidates. Potentially the strict conservation requirement, which was shown to be successful in the prediction of plant and animal miRNA hairpins, did not apply to fungal miRNAs. Likely, developmental regulation in fast evolving single-cellular organisms as fungi is not as conserved as that in complex, multicellular organisms. Moreover, the hypothesis that a considerable pool of non-conserved, 'young' miRNAs would exist in animals and plants [7],[8] got increased scientific support.

Yet, in that particular period (~2007-2008), miRNA research started to rely on and demand more and more that published miRNAs had some degree of experimental proof. The first deep-sequencing miRNA data sets of model organisms were being published, showing that some earlier published miRNAs might be incorrect. We realized that without (large-scale) experimental support the effort to publish any of these results would be fruitless. Unfortunately, funding for these experiments were not available during my thesis project. Therefore, the difficult but sensible decision was made not to proceed this line of research during my thesis. A fundamental scientific breakthrough can only be achieved if bioinformatics, biological datasets and functional analyses are combined, preferably in an *a priori* and carefully planned manner. Precious time and resources will get lost if not all of these prerequisites are met in a scientific research project.

Validity of L score miRNA predictions

Later scientific reports have shown the value our miRNA prediction effort and provided *a posteriori* justification for the quest for miRNAs in fungi. In chapter 7, we provide lists of high-scoring genomic hairpins in three species (Additional Files 5-8). These lists show large overlap with the miRNAs known at that time, however some of these did not correspond to known miRNAs. From these, we mention a few highly compelling miRNA candidates in viruses: two in Epstein-Barr virus (EBV) and three in Mareks disease virus (MDV). By revisiting the miRBase registry [9], release 20, we noticed that the novel miRNAs ebv-mir-BART21 [10], ebv-mir-BART22 [11],[10] and mdv1-mir-M31 [12] exactly correspond to our predictions. For *C. elegans,* we elaborated on only four out of 132 known miRNA hairpins that were not among our exhaustive population of over 3.500,000 genomic hairpins. We speculated that these (cel-mir-262, cel-mir-260, cel-mir-272 and cel-mir-256) might not be genuine miRNAs. Current miRBase entries provide accumulated Reads Per Million (RPM) data obtained from various next-generation sequencing (NGS) experiments. Cel-mir-256 and cel-mir-272 have yet to be detected in any experiment, and cel-mir-260 and 262 are only supported by very low RPM values (12.2 and 18.3, respectively). This is in sharp contrast to typical expression levels of the earliest described, non-constitutively expressed miRNAs lin-4 (29.300 RPM) and let-7 (6.510 RPM), and even to the median (41 RPM) of 223 currently known miRNAs. Overall, these results underline the strengths of our predictions using the *L*-score classifier.

Discovery of miRNA-like small RNAs and cross-kingdom RNAi in plant-pathogenic fungi

The hypothesis that miRNAs also occur in fungi was later demonstrated in *N. crassa* [13]. It was experimentally shown that 21nt miRNA-like small RNAs (milRNAs) were cleaved from endogenous RNA precursor hairpins. Evidence was provided that these milRNAs regulate gene expression in *Neurospora* by downregulation of (protein-coding) transcripts that contain imperfectly complementary target sites, just like reported for animal miRNAs [14]. The results uncover several pathways for small RNA production in filamentous fungi, and show that the milRNA-maturation pathway is indeed slightly different from those reported in animals and plants [14],[15], which could explain why several distinct types of RNA hairpins were substrates for Dicer. These differences let to the naming of a miRNA-like pathway in fungi, indicating that further research was required to fully understand this pathway.

Existence of an extensive, diverse and tissue-specific population of small RNAs in fungi was shown by NGS sequencing [13],[16],[17],[18],[19] of which only a small proportion probably resemble milRNAs. Remarkably, in a recent study in *B. cinerea* it was shown that fungal-encoded, processed small RNAs selectively silence host immunity genes [19]. These small RNAs hijack the host RNAi machinery by binding to Argonaute 1; “thus, this fungal pathogen transfers ‘virulent’ small RNA effectors into host plant cells to suppress host immunity and achieve infection” [19]. The loci which produce these small RNAs are LTR retrotransposons and do not correspond to milRNA loci. Such cross-kingdom RNAi was already demonstrated to exist in mammalian virus-encoded miRNAs ([20]; for review see [21]). Yet, it is an exciting finding for pathology, which opens up a new level of manipulation in fungus-host interactions. The obvious question that remains is how the fungus manages to transfer small RNAs into plants.

Comparative genomics of *Dothideomycete* genomes

In the last decade numerous fungal genomes were sequenced and most of them became publicly available. Focus In our laboratory is on the *Ascomycetes*, and more specifically the class of *Dothiodeomycetes*. As part of this thesis, the genome sequencing, annotation and comparative analyses of *Cladosporium fulvum* and *Dothiostroma septosporum* was performed in a joint effort with an affiliated group, and by commitment of all colleagues and co-authors. The availability of those two genome sequences in particular, fortified by a steady growing array of (closely) related genomes, opened up the possibility of various comparative genomics analyses. In the following part of the discussion, the most striking observations are further discussed.

Increased level of pseudogenization in Cladosporium fulvum

Pseudogenization of effector genes of a plant pathogenic fungus have been reported on. Examples are the avirulence (AVR) genes of *C. fulvum* [22] and *Leptosphaeria maculans* [23] which show allelic variation by acquiring mutations resulting in frame shifts or in-frame stops, which effectively results in a loss of function mutation of the conserved AVR. In the studied fungi in chapter 3 and 4, or fungal genomes in general, pseudogenization was expected to occur at very low frequencies but this was so far never quantified. The exception to the rule of a low pseudogenization frequency is the observation of abundant gene loss in the genomes of the ectomycorrhizal symbiont *Tuber melanosporum* [24] and the obligate biotrophic pathogen

Blumeria graminis [25]. In *T. melanosporum*, the loss of gene is explained primarily by pseudogenization. In *B. graminis*, it was stated that the genes were completely lost from its genome [25]. However, TBLASTX analyses using annotated fungal proteins yielded significant and abundant similarity all throughout its genome (data not shown), indicating that gradual pseudogenization is a better explanation and in accordance with the observation in *T. melanosporum*. In both species, the loss of genes was attributed to transposon activity [25],[24], and simultaneously this proliferation of transposable elements resulted in an expansion of their genome sizes to around 120-125MB. The loss of numerous individual genes up to complete pathways and an biotrophic lifestyle seems to go hand in hand. It might represent an evolutionary dead end of an organism that becomes increasingly specialised but at the same time fully dependent on its host. Therefore it is difficult to quantify the total number or the pseudogenized fraction of the (ancestral) gene catalogue of these fungi. Based on the estimation of a typical fungal gene catalogue (12,000) and the number of predicted proteins in *B. graminis* (5,854 [25]) and *T. melanosporum* (~7,500 [24]), 38 to 51% of all genes could have become pseudogenes.

Of the species that were studied in chapter 4, up to 372 representing 5% of the annotated genes of *C. fulvum* might represent pseudogenes. Additionally, hundreds of not yet annotated gene loci were found to be significantly similar to predicted fungal proteins, yet only when tolerating in-frame stop codons and frame shift mutations. This similarity was not limited to single or partial domain hits, but spanned the full query protein sequence. In most cases a HMMER model could serve as query, based on a multiple protein sequence alignment of commonly observed fungal proteins. This observation, combined with a likely somewhat overestimated number of genes in the annotated gene catalogue (based on ongoing RNA-seq analyses; data not shown), indicates that the stated 5% likely represents a two-fold underestimation. We showed that for a small minority of these mutated genes the description gene truncation might be more justified than their classification as pseudogene. For mutations that are not dramatic to protein continuity, the encoded protein might still have complete, residual or a new function. Without functional characterization, the impact of a mutation in regard to preservation, loss or change of function is difficult to indicate.

Comparison of the observed extend of pseudogenization among the six studied *Ascomycetes* revealed a relation to reproductive mode. A higher degree of pseudogenization was observed for species that reproduce preferably asexually compared to those that are known to reproduce both sexual and asexual. It might be possible that *C. fulvum*, the species with by far the highest number of pseudogenes, is at the onset of becoming an obligate biotroph like *B. graminis*. The additional striking resemblance between *C. fulvum* and these two species is the expanded genome size due to massive retrotransposon proliferation. Although the legacy of its putative ancestral hemi-biotrophic lifestyle is still encoded on its genome, many of these genes are either pseudogenized or not expressed during host colonization (see chapter 2).

Introner- and Introner-Like Elements could represent the missing link in intron dynamics

Chapter 5 and 6 describe the discovery and characterization of introner-like elements (ILE) in several Dothideomycete fungi. Multiplication of ILEs was shown to be responsible for massive and on an evolutionary timescale recent gain of introns. ILEs were shown to be the main contributor to intron gains. In addition, inter- and intra-species comparison of ILEs showed a considerable degree of analogy by sequence similarity and predicted secondary structure, which is maintained by a manifold of compensatory mutations. On the contrary, individual copies of ILEs rapidly diverge in sequence, and tend to lose their particularly stable secondary structure during this process. ILEs received their name because of their resemblance to introner elements (IE) which were discovered in the green alga *Micromonas pusilla* [26],[27]. IEs were recognized to be responsible for the massive and apparently sudden invasion of new introns in this species. Because genomes of closely related species were not available, no further comparative analyses was performed in that species.

The origin of introns remains a great mystery. Although the genomes of virtually all eukaryotes are intron-rich, little evidence for intron gain could be found [28]. One of the most accepted hypotheses explains the unambiguous and abundant presence of introns in eukaryotes by a sudden and massive invasion as early as the last eukaryotic common ancestor (LECA) [29]. Thus the LECA genome was full of introns and from this ancestor the radiation of eukaryotic life initiated. Some would have lost introns at a fast pace (e.g. *Saccharomyces cerevisiae* has few introns) or at a slow pace , like all vertebrate genomes which are nowadays still rich in introns. This hypothesis would imply that intron gain would be very rare. Successive analyses in various taxonomic clades kept this hypothesis on introns alive [28],[30], but it would ultimately result in the disappearance of spliceosomal introns. However, some lineages such as fungi have experienced more balanced rates of intron gains and losses [31],[32]. Recently intron gains have also been described in in multicellular animals like the micro-crustacean *Daphnia pulex* (water flea) [33] and the tunicade (sea-squirts) *Oikopleura dioica* [34]. These observations suggest that intron gains does still occur in present day genomes.

Although the mechanism of IE and ILE multiplication is unknown, it caused intron gains at a so far unprecedented scale in recent evolutionary history and in two unrelated branches of the eukaryotic tree of life. Hundreds of (gained) introns in different genes in several *Dothideomycete* species could be attributed to the multiplication of ILEs. In *Zymoseptoria tritici*, and a subset of insertions could be dated to the last 20,000 to 2,000 years. Moreover, evolutionary distance could be correlated to sequence divergence of ILEs, which shows that ILE sequences degenerate within 100ky to become indistinguishable from regular spliceosomal introns assuming an average mutation rate. This would indicate that, assuming that ILE multiplication is been ongoing for longer than 100ky, many more intron gains could be identified in ILE-containing *Dothideomycetes*. This is exactly what we observed, as shown in Figure 1.

This figure shows the paradox of apparent decrease of ILE-accounted contribution to intron gain at increasing branch length. ILE's account for up to 90% of all most recent intron gains in various *Zymoseptoria* species, which decreases to 40-60% in species that lack a recent branching speciation event (*D. septosporum*, *C. fulvum* and *Septoria passerinii*) and finally drops to less than 10% in the common ancestors.

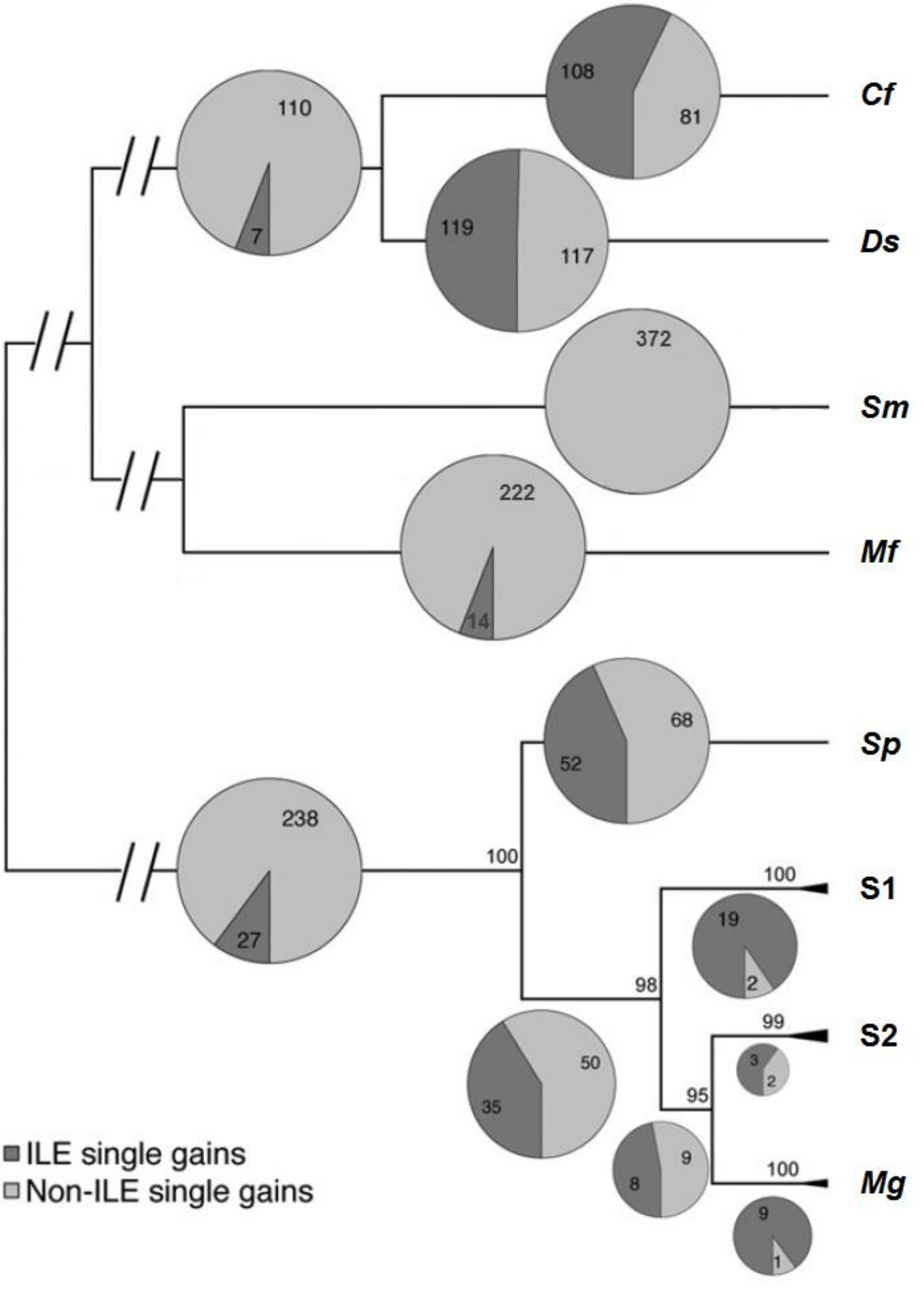


Figure 1: Contribution of Introner-like Elements to Single Intron Gains

Contribution of ILE-multiplication to intron gain plotted on the species tree of Cladosporium fulvum (Cf), Dothistroma septosporum (Ds), Septoria musiva (Sm), Mycosphaerella fijiensis (Mf) and recently speciated Zymoseptoria species: Mycosphaerella graminicola (Mg, renamed Zymoseptoria tritici, sister species S1, sister species S2 and Septoria passerinii. Adapted from Figure 5c in Chapter 5; bootstrap-supported branch lengths are only informative for the Zymoseptoria species; topology of the Dothideomycete species tree according to Figure 2a of Chapter 2. For M. fijiensis and S. musiva, only half of the total number of intron loci could be inspected due to the requirement of an additional outgroup node (see Table S3, Chapter 5). Most informative ,however, is the proportion of ILE-multiplication to intron gain.

At the same time, the overall rate of intron gains since the radiation of the *Capnodiales* species does not seem to vary. This pattern is also observed for *M. fijiensis*. For *Septoria musiva*, no recently active ILE cluster was detected, supported by absence of several near-identical copies. Yet, an overwhelming number of hits was retrieved by a HMM search of ILE sequences in its intronome (see Table S2, Chapter 5), which is consistent with the fast sequence degeneration of individual ILE copies.

This implies that, as exemplified for *C. fulvum* and *D. septosporum*, the total number of ILE-intron gains since the divergence from the *Zymoseptoria* branch could represent nearly four times the number that is now recognized as such (538 and 445 ILEs, respectively) in their genomes. This brings the total number of ILE-mediated gained introns at around 2,000 which is at the same order of magnitude as observed in *M. pusilla*. A major difference remains the apparent timescale during which those thousands of multiplications occurred. In these fungi, the estimated timescale would be one million year, whereas in *M. pusilla* lack of sequence divergence among individual IE copies suggest a very narrow and recent window of time. However, this suggestion is based on the pattern of fast sequence degeneracy of ILEs in fungi. Future analyses of yet to be discovered closely related *M. pusilla* species might give further insight.

What is the origin and the molecular mechanism behind the mobility of Introners and introner-like elements?

An hypothesis for the intron-richness of current eukaryotic species is a sudden invasion of introns in the LECA. Several mechanism were initially proposed to explain the acquisition of introns [35]. Some of these have some experimental support to occur in nowadays genomes, but none of these are supported by an observed frequency that can explain the current abundance of introns in eukaryotes. Therefore, it was proposed that the mechanism of the ancestral intron invasion would be different from that in the nowadays genomes [35]. The observed multiplication frequency of IEs in *M. pusilla* and ILEs in *Dothideomyectes* makes the hypothesis of a massive and sudden intron invasion of the LECA plausible. Simultaneously it would imply that mechanistically there is no difference between ancestral and nowadays intron gain. Because of the high frequency, ILE and IE multiplication likely involves a mechanism different from that proposed for intron gain in earlier reports [35]. The models that currently received most support involve intron transposition [36],[35],[27] and spliceosomal retrohoming [37]. Both involve reverse splicing and reverse transcription, but the first model assumes spicing into mRNA and the latter directly into DNA. In the case of ILEs, it is tempting to speculate that their secondary structures might significantly contribute to the multiplication mechanism. The predicted stable secondary structure of ILEs seems to be under selection pressure as can be concluded from the multiple compensatory mutations observed in ILEs. Noteworthy here is that IEs have comparable stable predicted secondary structures, but only on their reverse complementary strand compared to their normal orientation in genes (data not shown). From the results presented above and in Chapter 6, multiplication of ILEs in fungi and IEs in *M. pusilla* is certainly the main mechanism of intron gain in these species. If large-scale intron invasion is indeed limited to IEs and ILEs, absence of these elements might explain why only few genomes have experienced recent (massive) intron gains. But, the question how some genomes acquired Introners remains.

Errors in the gene catalogue of fungal species

Any biological data set that heavily relies on *in silico* prediction will contain errors. However, the frequency, type and gravity of errors will depend on many different factors. The application of a specific tool or method – or the omission of it – could be a source for a particular systematic error to occur more frequently. Having insight in type and frequency of expected errors is of crucial importance for planning and performing experiments, regardless if they are done *in vivo*, *in vitro* or *in silico*.

During the analyses on ILEs in six *Dothideomycetes* (chapter 5), we discovered that many of the introns recognized as ILEs likely had incorrectly predicted boundaries. Sensitive HMMER models of multiple DNA sequence alignments of ILEs we used to align spurious introns and by manual inspection the boundaries of many ILEs could be corrected. Such tedious work and its results on intron-exon structures are rarely reported in the main text of a publication; at best in supplementary data it is mentioned that `many genes/sites/introns were manually curated`. In case of the manual curation of ILEs, hundreds of ILE sequences that showed aberrancies to their expected size were checked. In the different species 6-39 ILEs were corrected (5-10% of all ILEs), meaning that their currently annotated boundaries were incorrect. Some introns were completely mis-annotated; often two introns would be fused leading to kb-sized long introns that linked two distinct genes in a false gene fusion. In extreme cases, the wrongly predicted intron camouflaged a third, intermediary positioned, gene that was overlooked by the gene prediction software. In total 52 ILEs (3%) were mapped to genomic positions lacking a gene model. Convincing similarity obtained by TBLASTX against proteins from other fungi at these loci indicate that these loci likely correspond to missing (parts of) genes. In addition, 7 to 55 introns per species were removed to yield the further studied datasets of 45-538 ILEs. These introns, around 10% of all ILEs per species, are certainly ILEs, but their exact 5’ (donor) or 3’ (acceptor) boundary could not be verified yet. Recent inspection of these ILEs by using RNA-Seq data of *C. fulvum* and *Z. tritici* (kindly provided by Eva Stukenbrock) confirmed that indeed the majority of these boundaries were incorrect (data not shown). This drilled down list of all errors encountered in a simple dataset of only 1750 introns underlines the apparent high error rate in the predicted introns of these species.

In chapter 6, we identified putative events of intron duplication in 24 fungal genomes. To achieve this, stringent filtering was required to remove introns that are themselves embedded in a duplicated sequence (see chapter 5, Supplemental Experimental Procedures). In those cases, the intron was most likely co-duplicated along with its adjacent sequence. By far the largest number of discarded introns were identified as being embedded in a longer, highly repetitive sequence context (data not shown). This presumably reflect a similar type of error as was shown in the introduction: parts of transposons that were erroneously predicted as protein-coding genes.

The overall poor quality of gene catalogues of fungi was quantified in chapter 3, where a substantial part of the gene catalogues of six different Ascomycete species were analysed. The method used, called Alignment-Based Fungal Gene Prediction (ABFGP) predicted models identical to current annotated genes in 57-77% of all cases. On the contrary, 22-41% of all gene models had proposed revisions. Corrected for the accurately defined error-rate of ABFGP itself, this account for thousands of gene models containing errors in a complete fungal gene catalogue. A significant decrease in average intron length was observed for the revised catalogues versus the current ones. This average length decrease was primarily caused by removal of large introns (>604nt) that were part of falsely fused genes and splitting of medium sized introns (79-604nt) into two small introns. Overall, this result in sharpening of the peak of small introns (47-78nt) in the length distribution of the intronome (compare Figure 1 and 2 in Chapter 1). The same trend, a decrease in mean intron length upon improvement of gene models, was also observed by others in the re-annotation of *F. graminearum* [38]. Generalized comparisons of the gene catalogues of the initially sequenced Ascomycete fungi yielded the observation that `Introns are typically short in fungi, averaging between 80 and 150 bp in many ascomycetes` [39]. Such generalisations likely contributed to incorrect assumptions in gene prediction efforts.

An additional outcome of the results obtained by ABFGP is the quantification of sequence errors in coding regions and the discovery of a considerable number of pseudogenes in various fungi (chapter 4). Existing (*ab initio*) gene prediction approaches focus on predicted genes, but are not capable of recognizing pseudogenes. For a genome with a high number of pseudogenes this will result in a high error rate.

The genome sequences of some of the fungi analysed were supported by low sequence coverage (4x to 8x coverage, based on traditional Sanger-sequencing). Many genes that were recognized (in chapter 3) to contain inframe stops and frameshift mutations were shown by comparison to NGS resequencing data to actually represent sequence errors. Recently, a resequenced NGS assembly (50x) update of the *B. cinerea* strain B05.10 reference assembly (4x) was described [40]. The comparison of both assemblies yielded 19,917 wrong base calls and 11,258 short indels (personal communication Martijn Staats), indicating a sequence error rate of at least 0.8 bases/kb. A typical fungal Sanger-sequenced genome with 4x to 7x coverage will contain numerous sequence errors in genic areas [41], and will result in inaccurate *de novo* gene predictions at these loci.

Comparative genomics studies on data sets with high error rates

In all chapters (2-6) that direct or indirectly dealt with gene catalogues of fungi, a considerable number of errors in these gene catalogues was encountered. There are serious pitfalls of working with large-scale (gen)omics data that mainly rely on predictions. Even the simplest of all meta-datasets describing a genome sequence, namely its gene catalogue, will have an intrinsically high error rate.

For most of the best studied model organisms, integration of data obtained from various expression data sets and research community efforts to manually improve gene models will have reduced the number of errors to an acceptable level by consecutive rounds of updated gene annotations. This is useful and appreciated by biologists that will work on functional analysis of genes. This is best exemplified by the TAIR annotation of Arabidopsis (TAIR10, 24/10/2013), which has seen its 6th major update since the first whole-genome TIGR 4.0 release (June 2003). This is an ideal scenario that is rarely achieved for fungi, except for a few of the best studied species (e.g. *S. cerevisae* [42], *S. pombe* [43]). For some model organisms, the annotations are so detailed and accurate that they describe characterized non-coding RNAs (tRNA, rRNAs, snoRNAs, various types of silencing small RNAs, etc) and differentiate between genes and pseudogenes. Most fungal annotations were never revisited since their original annotation and/or publication was released. The majority of annotations have not gone beyond version 1 [44],[45]. There are several reasons for this. First, semi-manual (re-)annotation of a gene catalogue is time- and labour-consuming, and thus unrealistic for small scientific communities working on one or several of these fungi. Second, projecting gene-model evidence from related fungi is far from straightforward. Besides the risk of propagating an error in the gene model of an informant protein, the evolutionary distances between fungi are often large leading to poor sequence conservation. Besides these practical limitations, there is also a reason of science policy: a study describing an improved gene catalogue of a particular species as the principal result is not particularly appreciated by scientific journals and funding agencies. Even research groups with in-house improved – potentially RNA-seq guided – annotations do not undertake the effort to publish their results in peer-reviewed journals for reasons mentioned above.

An often used argument, since the availability of RNA-sequencing technology, is that inclusion of RNA-seq data will solve the problem of low-quality gene predictions. This is likely too optimistic, at least for the timeframe in which this improvement will take place. First, current software dealing with RNA-seq data is often aimed at identifying expression level variation (of characterized and correct transcript catalogues), and not *per se* structural annotation of transcripts. Although some tools do perform the second task (e.g the popular tool Cufflinks [46]), they perform generally well on the genomes of eukaryotes that are relatively gene-poor due to large intergenic distances. However, when intergenic distances are short, as observed in fungal and oomycete genomes, reference-based assembly of mapped RNA-seq data is highly prone to predict erroneously merged models. RNA-seq data will, however, indisputably reduce the error-rate in novel published gene catalogues. On the contrary, the genomes of earlier published model fungi that nowadays serve as text-book examples, often still have their initial, erroneous gene catalogues. But often their gene catalogues are required to be included in comparative genomics studies as representatives of their clade, in favour of their lesser known, but putatively better annotated, more recently sequenced and annotated phylogenetically related fungal genomes.

This preference for generation of novel, yet error-prone data over improving existing data is largely imposed by the demands from society. Our society is in constant search for new, unexpected solutions for the unanswered global problems as pollution, depletion of natural resources and disease of crop, livestock and humans. Genomic sequencing and analyses without being over-accurate has shown its value to discover the unexpected. As best examples are the restriction enzymes and polymerase genes derived from extremophile bacteria, which paved the road for molecular biological research. A contemporary example is the interest in the lignin-degradation capability of mainly white-rot fungi that can bring commercial biofuel-production to maturity.

In many research groups that study the interactions between plants and bacteria/fungi research is usually aimed at delivering *in planta* experimental evidence on the role of one or a few genes involved in virulence or avirulence. For plant pathologists small errors in gene catalogues are not disturbing, because the gene model(s) will be manually checked and verified by single-gene resequencing. However, exploiting bioinformatics in these research disciplines is increasing.

Experimental data acquired in one species needs to be quickly and reliably transferred to another, and comparative genomics studies are more frequently undertaken. With the increase in efficiency and decrease of costs, such systems biology studies are being initiated in the discipline of phytopathology. The ultimate goal is to explain pathogenicity and virulence of a pathogen: to understand the molecular basis and evolution of pathosystems up to the complex interactions between a complete microbiome with host plants. Thus, studies have evolved from single gene-for-gene interactions into multiple gene-for gene interactions between (pathogenic) microbes and their hosts It is important to understand mechanisms of adaptation of pro- and eukaryotic microbes to their host plants in time and space in order to predict durability of plant resistance genes. For those types of studies, good quality of the starting data is of crucial importance.

As soon as multiple datasets are integrated, individual error rates will accumulate in a higher level error rate. For comparative genomics studies that critically depend on accurate gene annotations, we must reluctantly and as a community admit that already this first layer of meta-data on a fungal genome sequence, namely its structural gene annotations, does not often meet the required quality. Proteins that are functionally characterized are often from non-fungi or the hemi-ascomycetes like the yeasts *S. cerevisae* or *S. pombe*. The ability to confidently recognize functional orthologues at these evolutionary distances is limited to a subset of highly conserved genes (e.g. CEGMA [47]). However, when less conserved or evolutionary more dynamic genes or pathways of genes are studied, proteins are poorly conserved at the sequence level and their occurrence throughout the fungal kingdom varies significantly. In many comparative genomics studies, predicted gene catalogues are used as-is, without any quality control. And in cases where it is done, how and which quality-improving improvements were undertaken exactly is often poorly documented. Due to strict formats and quota imposed by peer-reviewed journals, typical phrasing included in materials and methods comprises \`when appropriate, gene models were corrected manually\` or \`genes A-Z were corrected\`. Current emphasis in research on NGS projects (as the 1000 Fungal Genomes Project) will yield an explosion on fungal genomic data. \`Science is a quickly moving front\` is often stated by colleagues. Advances in bioinformatics will in time reduce the error level in many omics data sets to workable levels. The timeframe in which this will take place will vary per type of error, and the perception of this timeframe might vary between individual scientists. It requires a change in attitude within research communities to either make an effort to correct errors in datasets and use these in meaningful analyses, or to abstain from analyses that will lead to inconclusive or flawed interpretation due to errors in the input data.

References

1. Nakayashiki H, Kadotani N, Mayama S: **Evolution and diversification of RNA silencing proteins in fungi**. *Journal of molecular evolution* 2006, **63**(1):127-135.
2. Catalanotto C, Pallotta M, ReFalo P, Sachs MS, Vayssie L, Macino G, Cogoni C: **Redundancy of the two dicer genes in transgene-induced posttranscriptional gene silencing in Neurospora crassa**. *Molecular and cellular biology* 2004, **24**(6):2536-2545.
3. Kadotani N, Nakayashiki H, Tosa Y, Mayama S: One of the two Dicer-like proteins in the filamentous fungi Magnaporthe oryzae genome is responsible for hairpin RNA-triggered RNA silencing and related small interfering RNA accumulation. *The Journal of biological chemistry* 2004, 279(43):44467-44474.
4. Bartel DP: MicroRNAs: genomics, biogenesis, mechanism, and function. *Cell* 2004, 116(2):281-297.
5. Krol J, Loedige I, Filipowicz W: **The widespread regulation of microRNA biogenesis, function and decay**. *Nature reviews Genetics* 2010, **11**(9):597-610.
6. McGuire AM, Galagan JE: **Conserved secondary structures in Aspergillus**. *PloS one* 2008, **3**(7):e2812.
7. Zhang B, Pan X, Cannon CH, Cobb GP, Anderson TA: **Conservation and divergence of plant microRNA genes**. *The Plant journal : for cell and molecular biology* 2006, **46**(2):243-259.
8. Jones-Rhoades MW: **Conservation and divergence in plant microRNAs**. *Plant molecular biology* 2012, **80**(1):3-16.
9. Kozomara A, Griffiths-Jones S: **miRBase: integrating microRNA annotation and deep-sequencing data**. *Nucleic acids research* 2011, **39**(Database issue):D152-157.
10. Zhu JY, Pfuhl T, Motsch N, Barth S, Nicholls J, Grasser F, Meister G: **Identification of novel Epstein-Barr virus microRNA genes from nasopharyngeal carcinomas**. *Journal of virology* 2009, **83**(7):3333-3341.
11. Cosmopoulos K, Pegtel M, Hawkins J, Moffett H, Novina C, Middeldorp J, Thorley-Lawson DA: **Comprehensive profiling of Epstein-Barr virus microRNAs in nasopharyngeal carcinoma**. *Journal of virology* 2009, **83**(5):2357-2367.
12. Burnside J, Ouyang M, Anderson A, Bernberg E, Lu C, Meyers BC, Green PJ, Markis M, Isaacs G, Huang E *et al*: **Deep sequencing of chicken microRNAs**. *BMC genomics* 2008, **9**:185.
13. Lee HC, Li L, Gu W, Xue Z, Crosthwaite SK, Pertsemlidis A, Lewis ZA, Freitag M, Selker EU, Mello CC *et al*: **Diverse pathways generate microRNA-like RNAs and Dicer-independent small interfering RNAs in fungi**. *Molecular cell* 2010, **38**(6):803-814.
14. Lee HC, Chang SS, Choudhary S, Aalto AP, Maiti M, Bamford DH, Liu Y: **qiRNA is a new type of small interfering RNA induced by DNA damage**. *Nature* 2009, **459**(7244):274-277.
15. Yang Q, Li L, Xue Z, Ye Q, Zhang L, Li S, Liu Y: Transcription of the major neurospora crassa microRNA-like small RNAs relies on RNA polymerase III. *PLoS genetics* 2013, 9(1):e1003227.
16. Nicolas FE, Moxon S, de Haro JP, Calo S, Grigoriev IV, Torres-Martinez S, Moulton V, Ruiz-Vazquez RM, Dalmay T: Endogenous short RNAs generated by Dicer 2 and RNA-dependent RNA polymerase 1 regulate mRNAs in the basal fungus Mucor circinelloides. *Nucleic acids research* 2010, 38(16):5535-5541.
17. Nunes CC, Gowda M, Sailsbery J, Xue M, Chen F, Brown DE, Oh Y, Mitchell TK, Dean RA: **Diverse and tissue-enriched small RNAs in the plant pathogenic fungus, Magnaporthe oryzae**. *BMC genomics* 2011, **12**:288.
18. Zhou J, Fu Y, Xie J, Li B, Jiang D, Li G, Cheng J: Identification of microRNA-like RNAs in a plant pathogenic fungus Sclerotinia sclerotiorum by high-throughput sequencing. *Molecular genetics and genomics : MGG* 2012, 287(4):275-282.
19. Weiberg A, Wang M, Lin FM, Zhao H, Zhang Z, Kaloshian I, Huang HD, Jin H: **Fungal small RNAs suppress plant immunity by hijacking host RNA interference pathways**. *Science* 2013, **342**(6154):118-123.
20. Pfeffer S, Zavolan M, Grasser FA, Chien M, Russo JJ, Ju J, John B, Enright AJ, Marks D, Sander C *et al*: **Identification of virus-encoded microRNAs**. *Science* 2004, **304**(5671):734-736.
21. Kincaid RP, Sullivan CS: **Virus-encoded microRNAs: an overview and a look to the future**. *PLoS pathogens* 2012, **8**(12):e1003018.
22. Stergiopoulos I, De Kock MJ, Lindhout P, De Wit PJ: Allelic variation in the effector genes of the tomato pathogen Cladosporium fulvum reveals different modes of adaptive evolution. *Molecular plant-microbe interactions : MPMI* 2007, 20(10):1271-1283.

23. 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.
24. Martin F, Kohler A, Murat C, Balestrini R, Coutinho PM, Jaillon O, Montanini B, Morin E, Noel B, Percudani R *et al*: **Perigord black truffle genome uncovers evolutionary origins and mechanisms of symbiosis**. *Nature* 2010, **464**(7291):1033-1038.
25. Spanu PD, Abbott JC, Amselem J, Burgis TA, Soanes DM, Stuber K, Ver Loren van Themaat E, Brown JK, Butcher SA, Gurr SJ *et al*: **Genome expansion and gene loss in powdery mildew fungi reveal tradeoffs in extreme parasitism**. *Science* 2010, **330**(6010):1543-1546.
26. Worden AZ, Lee JH, Mock T, Rouze P, Simmons MP, Aerts AL, Allen AE, Cuvelier ML, Derelle E, Everett MV *et al*: **Green evolution and dynamic adaptations revealed by genomes of the marine picoeukaryotes Micromonas**. *Science* 2009, **324**(5924):268-272.
27. Verhelst B, Van de Peer Y, Rouze P: The complex intron landscape and massive intron invasion in a picoeukaryote provides insights into intron evolution. *Genome biology and evolution* 2013, 5(12):2393-2401.
28. Roy SW, Gilbert W: **Rates of intron loss and gain: implications for early eukaryotic evolution**. *Proceedings of the National Academy of Sciences of the United States of America* 2005, **102**(16):5773-5778.
29. Koonin EV: **The origin of introns and their role in eukaryogenesis: a compromise solution to the introns-early versus introns-late debate?** *Biology direct* 2006, 1:22.
30. Csuros M, Rogozin IB, Koonin EV: **A detailed history of intron-rich eukaryotic ancestors inferred from a global survey of 100 complete genomes.** *PLoS computational biology* 2011, 7(9):e1002150.
31. Nielsen CB, Friedman B, Birren B, Burge CB, Galagan JE: **Patterns of intron gain and loss in fungi**. *PLoS biology* 2004, **2**(12):e422.
32. Zhang LY, Yang YF, Niu DK: **Evaluation of models of the mechanisms underlying intron loss and gain in Aspergillus fungi**. *Journal of molecular evolution* 2010, 71(5-6):364-373.
33. Li W, Tucker AE, Sung W, Thomas WK, Lynch M: **Extensive, recent intron gains in Daphnia populations**. *Science* 2009, **326**(5957):1260-1262.
34. Denoeud F, Henriet S, Mungpakdee S, Aury JM, Da Silva C, Brinkmann H, Mikhaleva J, Olsen LC, Jubin C, Canestro C *et al*: **Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate**. *Science* 2010, **330**(6009):1381-1385.
35. Yenerall P, Zhou L: Identifying the mechanisms of intron gain: progress and trends. *Biology direct* 2012, 7:29.
36. Torriani SF, Stukenbrock EH, Brunner PC, McDonald BA, Croll D: **Evidence for extensive recent intron transposition in closely related fungi**. *Curr Biol* 2011, **21**(23):2017-2022.
37. Roy SW, Irimia M: **Mystery of intron gain: new data and new models**. *Trends in genetics : TIG* 2009, **25**(2):67-73.
38. Wong P, Walter M, Lee W, Mannhaupt G, Munsterkotter M, Mewes HW, Adam G, Guldener U: **FGDB: revisiting the genome annotation of the plant pathogen Fusarium graminearum**. *Nucleic acids research* 2011, **39**(Database issue):D637-639.
39. Galagan JE, Henn MR, Ma LJ, Cuomo CA, Birren B: **Genomics of the fungal kingdom: insights into eukaryotic biology**. *Genome research* 2005, **15**(12):1620-1631.
40. Staats M, van Kan JA: **Genome update of Botrytis cinerea strains B05.10 and T4**. *Eukaryotic cell* 2012, **11**(11):1413-1414.
41. Weber JL, Myers EW: **Human whole-genome shotgun sequencing**. *Genome research* 1997, **7**(5):401-409.
42. SGD: Saccharomyces genome database [http://www.yeastgenome.org]
43. pombase: the scientific resource for fission yeast [http://www.pombase.org/]
44. **FGI: Fungal Genome Initiative** [http://www.broadinstitute.org/scientific-community/science/projects/fungal-genome-initiative/fungal-genome-initiative]
45. **FGP: Fungal Genomics Program** [http://genome.jgi.doe.gov/programs/fungi/index.jsf]
46. Roberts A, Pimentel H, Trapnell C, Pachter L: **Identification of novel transcripts in annotated genomes using RNA-Seq**. *Bioinformatics* 2011, **27**(17):2325-2329.
47. Parra G, Bradnam K, Korf I: CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics* 2007, 23(9):1061-1067.

Summary

Fungi are a diverse group of eukaryotic micro-organisms particularly suited for comparative genomics analyses. Fungi are important to industry, fundamental science and many of them are notorious pathogens of crops, thereby endangering global food supply. Dozens of fungi have been sequenced in the last decade and with the advances of the next generation sequencing, thousands of new genome sequences will become available in coming years. In this thesis I have used bioinformatics tools to study different biological and evolutionary processes in various genomes with a focus on the genomes of the *Dothideomycete* fungi *Cladosporium fulvum*, *Dothistroma septosporum* and *Zymoseptoria tritici*.

Chapter 1 introduces the scientific disciplines of mycology and bioinformatics from a historical perspective. It exemplifies a typical whole-genome sequence analysis of a fungal genome, and focusses in particular on structural gene annotation and detection of transposable elements. In addition it shortly reviews the microRNA pathway as known in animal and plants in the context of the putative existence of similar yet subtle different small RNA pathways in other branches of the eukaryotic tree of life.

Chapter 2 addresses the novel sequenced genomes of the closely related *Dothideomycete* plant pathogenic fungi *Cladosporium fulvum* and *Dothistroma septosporu*m. Remarkably, it revealed occurrence of a surprisingly high similarity at the protein level combined with striking differences at the DNA level, gene repertoire and gene expression. Most noticeably, the genome of *C. fulvum* appears to be at least twice as large, which is solely attributable to a much larger content in repetitive sequences.

Chapter 3 describes a novel alignment-based fungal gene prediction method (ABFGP) that is particularly suitable for plastic genomes like those of fungi. It shows excellent performance benchmarked on a dataset of 7,000 unigene-supported gene models from ten different fungi. Applicability of the method was shown by revisiting the annotations of *C. fulvum* and *D. septosporum* and of various other fungal genomes from the first-generation sequencing era. Thousands of gene models were revised in each of the gene catalogues, indeed revealing a correlation to the quality of the genome assembly, and to sequencing strategies used in the sequencing centres, highlighting different types of errors in different annotation pipelines.

Chapter 4 focusses on the unexpected high number of gene models that were identified by ABFGP that align nicely to informant genes, but only upon toleration of frame shifts and in-frame stop-codons. These discordances could represent sequence errors (SEs) and/or disruptive mutations (DMs) that caused these truncated and erroneous gene models. We revisited the same fungal gene catalogues as in chapter 3, confirmed SEs by resequencing and successively removed those, yielding a high-confidence and large dataset of nearly 1,000 pseudogenes caused by DMs. This dataset of fungal pseudogenes, containing genes listed as bona fide genes in current gene catalogues, does not correspond to various observations previously done on fungal pseudogenes. Moreover, the degree of pseudogenization showing up to a ten-fold variation for the lowest versus the highest affected species, is generally higher in species that reproduce asexually compared to those that in addition reproduce sexually.

Chapter 5 describes explorative genomics and comparative genomics analyses revealing the presence of introner-like elements (ILEs) in various *Dothideomycete* fungi including *Zymoseptoria tritici* in which they had not identified yet, although its genome sequence is already publicly available for several years. ILEs combine hallmark intron properties with the apparent capability of multiplying themselves as repetitive sequence. ILEs strongly associate with events of intron gain, thereby delivering in silico proof of their mobility. Phylogenetic analyses at the intra- and inter-species level showed that most ILEs are related and likely share common ancestry.

Chapter 6 provides additional evidence that ILE multiplication strongly dominates over other types of intron duplication in fungi. The observed high rate of ILE multiplication followed by rapid sequence degeneration led us to hypothesize that multiplication of ILEs has been the major cause and mechanism of intron gain in fungi, and we speculate that this could be generalized to all eukaryotes.

Chapter 7 describes a new strategy for miRNA hairpin prediction using statistical distributions of observed biological variation of properties (descriptors) of known miRNA hairpins. We show that the method outperforms miRNA prediction by previous, conventional methods that usually apply threshold filtering. Using this method, several novel candidate miRNAs were assigned in the genomes of *Caenorhabditis elegans* and two human viruses. Although this chapter is not applied on fungi, the study does provide a flexible method to find evidence for existence of a putative miRNA-like pathway in fungi.

Chapter 8 provides a general discussion on the advent of bioinformatics in mycological research and its implications. It highlights the necessity of *a priori* planning and integration of functional analysis and bioinformatics in order to achieve scientific excellence, and describes possible scenarios for the near future of fungal (comparative) genomics research. Moreover, it discusses the intrinsic error rate in large-scale, automatically inferred datasets and the implications of using and comparing those.

Samenvatting

Schimmels zijn een heterogene groep van eukaryotische micro-organismen die zich bijzonder goed lenen voor het doen van vergelijkend DNA-onderzoek. Schimmels zijn belangrijk voor de industrie en fundamenteel wetenschappelijk onderzoek; vele soorten zijn beruchte ziekteverwekkers van voedingsgewassen en bedreigen de wereldvoedselvoorziening. Van vele tientallen schimmels zijn de genomische DNA sequenties inmiddels bekend, en door de beschikbaarheid van `next-generation` DNA sequencing methodes zullen er naar verwachting op korte termijn duizenden nieuwe genomen beschikbaar komen. In dit proefschrift heb ik door middel van bio-informatica verschillende biologische en evolutionaire processen in verscheidene schimmels bestudeerd en vergeleken, daarbij gebruik makend van genomisch DNA van een aantal schimmels uit de klasse der Dothideomyceten waaronder *Cladosporium fulvum*, *Dothistroma septosporum* en *Zymoseptoria tritici*.

In **Hoofdstuk 1** introduceer ik de mycologie en de bio-informatica vanuit een historisch perspectief. Er wordt een voorbeeld gegeven van de analyse van een compleet schimmelgenoom, waarbij met name gekeken is naar structurele gen voorspelling en detectie van repetitieve DNA-sequenties. Daarnaast wordt een kort overzicht gegeven van de microRNA-pathway in dieren en planten, met het oogpunt om eventuele overeenkomsten en verschillen in de microRNA-pathway op te sporen Dit heeft als doel om gericht te kunnen zoeken naar vergelijkbare pathways in andere takken van de eukaryotische boom van het leven.

Hoofdstuk 2 is gewijd aan de nieuw gesequencete genomen van de nauw verwante, plantpathogene schimmels *C. fulvum* en *D. septosporum* uit de klasse der Dothideomyceten. In de genomen van deze twee schimmels worden twee tegenstrijdige verschijnselen waargenomen, namelijk een onverwacht grote overeenkomst op het niveau van eiwit-sequentie en opvallende verschillen op DNA-niveau, waaronder verschillen in genrepertoire en genexpressie. Meest opvallend is dat het genoom van *C. fulvum* minstens twee keer zo groot is als dat van *D. septosporum*, wat te verklaren is op grond van een veel grotere hoeveelheid repetitief DNA in het genoom van eerstgenoemde schimmel.

Hoofdstuk 3 beschrijft een nieuwe genvoorspellingsmethode (ABFGP: alignment-based fungal gene prediction) die bijzonder geschikt is voor dynamische genomen zoals die van schimmels. De methode voldoet uitstekend in een vergelijkende prestatietest op een dataset van 7.000 door geobserveerde expressie bewezen genmodellen uit tien verschillende schimmels. Toepasbaarheid van de methode wordt aangetoond door de gen-annotatie in *C. fulvum*, *D. septosporum* en enkele andere schimmelgenomen die nog op klassieke wijze waren gesequenced opnieuw onder de loep te nemen. Duizenden genmodellen werden gereviseerd per gencatalogus. Er werden, zoals verwacht, correlaties gevonden tussen de kwaliteit van de genoomassemblage en de DNA sequencing strategieën zoals toegepast in verschillende centra voor DNA sequencing. Er is ook een correlatie tussen de verschillende typen van waargenomen fouten en de gebruikte strategieën voor gen-annotatie.

Hoofdstuk 4 beschrijft het onverwacht grote aantal door ABFGP voorspelde genmodellen dat in vergelijking met homologe informant genen frame-shifts en/of stopcodons in het open leesraam heeft. Deze onregelmatigheden kunnen sequentie-fouten (SEs) en/of disruptieve mutaties (DMs) zijn, die beide resulteren in verkorte of foutieve genmodellen. We bestudeerden hier weer dezelfde gencatalogi als beschreven in **hoofstuk 3**, we bevestigden SEs door opnieuw te sequencen en verwijderden deze vervolgens. Hierdoor verkregen we een opgeschoonde dataset van genmodellen met hoge betrouwbaarheid met bijna 1000 door DMs veroorzaakte pseudogenen. Observaties aan deze dataset van pseudogenen in schimmels, enkel bestaande uit genen die als *bonafide* te boek staan in huidige versies van hun gencatalogi, komen niet overeen met verschillende eerder gedane observaties aan pseudogenen in schimmelgenomen. Daarnaast vertoont de mate van pseudogenisatie in een zestal schimmelgenomen een variatie van een factor tien tussen het laagste en hoogste niveau. De trend tekent zich af dat er meer pseudogenen lijken te zitten in soorten die aseksueel reproduceren dan in soorten die daarnaast ook seksueel reproduceren.

Hoofdstuk 5 beschrijft exploratieve en vergelijkende genomische analyses die de aanwezigheid van Introner-Like Elements (ILEs) in verscheidene schimmels uit de klasse der Dothideomyceten aantonen. Daaronder ook *Z. tritici*, waarin ILEs nog niet beschreven waren, hoewel de DNA-sequentie van deze schimmel reeds enkele jaren publiek beschikbaar is. ILEs combineren tekstboek-eigenschappen van intronen met het klaarblijkelijke vermogen zich te kunnen multipliceren als repetitieve sequenties. ILEs zijn sterk geassocieerd met nieuw verworven intronen, en leveren *in silico* bewijs voor hun mobiliteit. Fylogenetische analyses van ILEs binnen en buiten verschillende soorten bewijzen dat de meeste ILEs verwant zijn en waarschijnlijk allemaal een gemeenschappelijke voorouder delen.

Hoofdstuk 6 levert additioneel bewijs, dat de multiplicatie van ILEs sterk domineert over mogelijke andere vormen van intronduplicatie in schimmels. De waargenomen grote multiplicatie-frequentie gevolgd door snelle sequentie-degeneratie heeft ons tot de hypothese gebracht, dat ILE-multiplicatie de belangrijkste bijdrage levert aan en voorziet in een mechanisme voor het ontstaan van nieuwe intronen in schimmels. Wij achten het aannemelijk dat deze observatie gegeneraliseerd kan worden naar alle eukaryoten.

Hoofdstuk 7 beschrijft een nieuwe strategie om microRNA hairpins te voorspellen op basis van geobserveerde biologische variatie in statistische distributies van eigenschappen (descriptoren) van bekende microRNA hairpins. We laten zien, dat onze methode betere resultaten geeft dan de eerder beschreven conventionele methodes die op de gebruikelijke wijze filteren op gekozen grenswaardes. Met deze methode konden enkele nieuwe kandidaat-microRNAs gevonden worden in het genomisch DNA van *Caenorhabditis elegans* en twee humane virussen. Alhoewel dit hoofdstuk niet toegespitst is op schimmels, verschaft deze studie een flexibele methode om bewijs te vinden in schimmels voor het mogelijk bestaan van een soortgelijke microRNA-pathway.

In **Hoofdstuk 8** bediscussieer ik de bijdrage van bio-informatica aan mycologisch onderzoek en de implicaties daarvan. Ik benadruk de noodzakelijkheid van *a priori* planning en integratie van functionele analyse en bio-informatica om wetenschappelijke excellentie te kunnen bereiken, en ik beschrijf mogelijke scenario's voor (vergelijkend) onderzoek aan genomisch DNA in schimmels. Daarnaast worden de intrinsieke foutenmarges in grootschalige, automatisch gegenereerde datasets en de implicaties van het gebruik en vergelijken daarvan besproken.

Dankwoord - Acknowledgement - Danksagung

Zonder Pierre was dit dankwoord er nooit geweest. Simpelweg omdat mijn promotie nooit zou zijn afgerond. Voor het bieden van deze tweede kans ben ik je erg dankbaar. In vele, soms moeizame discussies heb ik veel van je geleerd, niet in de laatste plaats over mezelf. Ik hoop dat je, net als ik, met voldoening terug kan kijken op wat we de afgelopen jaren samen hebben weten te bereiken. De vrijheid en het vertrouwen die je geeft aan mij en aan iedereen die je onder je hoede neemt, ook als de ingeslagen weg niet per se jouw eerste keus is, tekenen jouw bijzondere persoonlijkheid. Ik gun het je van harte, dat je de komende jaren je wetenschappelijke carrière weet af te bouwen met hetzelfde enthousiasme als waarmee je hem hebt opgebouwd.

Dear Jérôme, where would I have been without you? Last year you were officially assigned as my co-promotor, a role that you in practice had fulfilled since december 2011. You had a positive and strongly convergent influence on my research, where divergence is both my talent and pitfall. Your calmness and confidence were of crucial added value during emotional discussions, and contributed to achieving consensus between Pierre and me. Time will learn if we will manage to successfully finish our current project of epic proportions. I wish you all the best in your future career.

Edouard en Martijn, bedankt dat jullie mijn paranimfen wilden zijn. Edouard, enorm bedankt voor de onder grote tijdsdruk geboden hulp bij het onderzoek naar de Introner-Like Elements. Ook voor je bijdrage aan hoofdstuk drie ben ik je erkentelijk. Ik vind het achteraf jammer dat we zo weinig hebben samengewerkt toen we nog directe collega’s waren.

Martijn, al sinds 2002 hebben wij de perfecte LAT-relatie. Tot twee maal toe waren we collega’s bij Wageningen UR, en zonder enige druk en net zoals het uitkomt gingen we samen trainen, bier drinken, op trainingsstage en bergvakantie. Na jaren volharding bleek zelfs La Città een simpel spelletje te zijn: burgers komen en gaan en uiteindelijk wint Martijn. Aan de gehele, van samenstelling wisselende groep immer weer verliezende La-Città-spelers wil ik graag twee dingen kwijt. Ten eerste bedankt voor vele onvergetelijke uren, avonden en dagen. Vooral voor diegenen van het eerste uur (Martijn, Olaf, Mart, Ralph!), omdat deze mannenavond aanvankelijk was opgezet om mij uit mijn toenmalige isolement en vermoeidheid te helpen. En ook dank aan de tweede generatie (Mark, Co, Allard): blijven oefenen!

Jan, het was een leerzame tijd om jouw kamergenoot te mogen zijn. Ik heb jouw parate moleculair-biologische en mycologische kennis wel eens vergeleken met `fungipedia` en daarom ook vaak geraadpleegd. Daarnaast wil ik je bedanken voor het helpenstructureren en corrigeren van enkele van mijn hoofdstukken, in de eerste plaats de introductie en discussie van dit proefschrift.

Jeroen, ook jij hebt geleden onder mijn PhD. Jij houdt niet van het woordje als, maar hoe anders zou het gelopen kunnen hebben als het eens een keer wél iets meer mee zou hebben gezeten? Je was het ook niet altijd met me eens, maar toch bleef je begrip opbrengen voor mijn volharding om toch op twee paarden te blijven wedden. Bedankt voor de carte blanche en voor het op alle fronten begeleiden, steunen en coachen van 2005 tot en met 2011, en ach, af en toe bij vlagen nog wel eens. Hopelijk kunnen we snel weer samen de loopschoenen aandoen: de winter is weer bijna voorbij!

Na mentale inspanning moet er ook ruimte zijn voor … mis! fysieke inspanning! Tijdens, na en voor de werkdag hebben velen samen met mij gelopen, gefietst, gebeuld en gedobberd: fysiek tegenwicht ter mentale ontspanning! Hoe vaak hebben jullie mij niet moeten opvrolijken als het met mijn PhD even tegen zat? Waarschijnlijk incompleet (excuses), en geordend op het totaal aantal meegelopen kilometers, mijn dank hiervoor aan Jeroen, Alex, Martijn, Wobbe, Jan R.† (wat had ik graag nu nog met je gelopen), Jan K., Ronald van L., Mark R., Johan, Neel, Esther, Ronnie, Aalt-Jan en Carl. Sorry Carl, you are definitely last in this series!

I thank all my (former) colleagues of the Cladosporium group. Thank you Bilal, Carl, Henriek and Mansoor for your fruitful collaborations on various projects of which many resulted in meanwhile published or recently accepted manuscripts. Carl, I really enjoyed our short but intense boosts of collaboration leading to the in silico assignment of Avr5. Your time-consuming in planta experiments immediately showed that we were right: I love it when a plan comes together ;-). Harrold, bedankt voor de introductie in de vele facetten van het voor mij onbekende fytopathologische onderzoekswereld. Joost, Ronnie, Luigi and Michael, thanks for many fruitful sparring-sessions on bioinformatics issues. Robbin, ik vond het erg leuk om je te begeleiden in jouw BSc-thesis. Zo’n talentvolle student begeleiden was een makkie! Sorry dat het er (nog) niet van is gekomen ons onderzoek in een publicatie te verwerken.

Roeland, bij dezen hartelijk bedankt voor het begeleiden en het mij bijna letterlijk naar mijn eerste publicatie slepen. Jouw inzet om het combineren van promoveren met topsport voor mij mogelijk te maken, betekende automatisch moeilijkere randvoorwaarden voor jou in het begeleiden van mijn promotietraject.

Mark, jou valt de twijfelachtige eer te beurt dat ik überhaupt begonnen ben aan een PhD. Jouw enthousiasme stak mij tijdens een MSc-afstudeervak aan, waardoor mijn fascinatie voor het doen van bio-informatica-onderzoek ontstond. Bedankt voor alle hulp bij het wegwijs maken in de wondere wereld van de bio-informatica. Erwin, bedankt – en onbegrijpelijk - dat je het zo lang met mij als kamergenoot hebt weten uit te houden. Die mierenplaag was inderdaad een beetje mijn schuld. Ik vond het verrassend leuk om met iemand met een compleet andere persoonlijkheid en kijk op het leven samen te werken. En ik heb veel opgestoken van jouw eenvoudige maar verrassend doeltreffende manier van oplossen van programmeerproblemen.

Jan, Bas, en Henri, bedankt voor alle hulp met het oplossen, maar vooral voorkomen van haperende hardware en software. Niet alleen jullie, maar ook de andere (oud)collega’s van Applied Bioinformatics (Sandra, Aalt-Jan, Judith, Marleen, Elio, Paul, Sander) wil ik hartelijk danken voor alle hulp en vooral de welgemeende interesse en morele steun gedurende de ups en downs bij zowel onderzoek als topsport. Gabino, thanks for offering the not self-evident support at the technical level during my time at Phytopathology.

Voor verschillende bijdragen in het begin van mijn PhD-traject wil ik ook graag Jan-Peter Nap, Gert Kema, Theo van der Lee, Cees Waalwijk en Teun Boekhout bedanken. Theo, leuk dat je wilde opponeren tijdens mijn verdediging. Jan-Peter, ik denk dan toch een soort bedankt voor jouw hulp en aanwijzingen bij mijn eerste pennenvruchten. Misschien dat jij en Pierre hierover een keer samen van gedachten kunnen wisselen.

Bijzonder veel dank ook aan iedereen die in de loop der jaren een luisterend oor bood dan wel moest bieden aan mijn klagen en twijfels gedurende mijn PhD-traject. Ik denk hierbij vooral aan Jeroen, Martijn, Corien en Gerke, Amely, Hannie, Mischa, Ellie en Leo, Pim, mijn ouders en niet in de laatste plaats Elise. Soms was stilzwijgend luisteren al voldoende, vaak echter kreeg ik suggesties of kleine lichtpuntjes aangereikt, waarmee ik weer vooruit kon. Martien, jij ook bedankt hiervoor, en natuurlijk voor vele onvergetelijke bergvakanties!

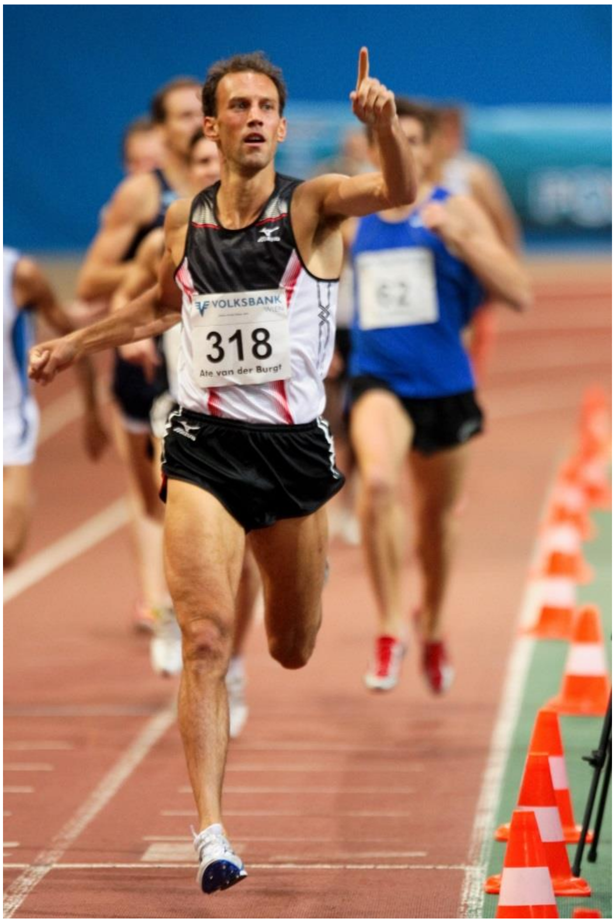
Meine Herren Strahler-Kollegen Mischa und Monsieur Président Stéphane, für was kann oder soll ich ihnen nicht danken? Hervorheben möchte ich insbesondere Ihr großartiges Verständnis. Leider war es mir im letzten Jahr nicht möglich, meine Beiträge zu unserer gemeinsamen mineralogischen Karriere in dem von mir gewünschten Umfang zu leisten. Glücklicherweise konnten Sie in dieser Zeit mehrfach meinen „Arsch“ retten, auch wenn Sie selbst unter Zeitdruck standen oder schwierigen persönlichen Umständen zu meistern hatten. Meinen herzlichen Dank für Euren riesigen Einsatz. Auch Herr Strahler Frank, Danke für viele unvergessliche Erlebnisse und gemeinsame Träume. Zusätzlich, sind auch die letzten Worte dieser Danksagung an Euch allen gerichtet. In der Zukunft werde ich demütig meine Aufgaben als Packesel, Spürhund und Küchenknecht während unserer gemeinsamen Streifzüge erfüllen.

Ook wil ik graag mijn ouders en schoonouders Agnes, Marius, Ellie en Leo bedanken. Mama, bedankt voor de noodzakelijke correcties aan de Engelse en Nederlandse teksten van mijn thesis. Had ik maar beter mijn best gedaan om te leren spellen en schrijven, dat zou mijn promotietraject aanzienlijk hebben kunnen verkorten. Jullie hebben allemaal bijgesprongen om de situatie voor Elise en Imane makkelijker te maken, als ik weer eens geen tijd had om huishoudelijke of administratieve klussen te doen. Leo en Ellie, bedankt voor het meermaals op Imane (en Elise) passen als ik weer eens de hort op moest om mineralen te gaan zoeken. Eigenlijk heeft de hele familie Van der Burgt me een klein beetje uit de wind gehouden, namelijk door mij nog steeds niet als vrijwilliger aan te wijzen voor het organiseren van de familiedag. Bij dezen kan en wil ik daar voor 2015 niet meer onderuit komen, en daarom wijs ik mijzelf vrijwillig aan als vrijwilliger.

Elise, dank je voor bijna 10 jaar geduld, meeleven en omdenken. Twee vervlochten carrières succesvol laten verlopen was te veel hooi op mijn vork, en alles wat ervan afviel of bleef liggen heb jij voor me opgeknapt. Je hebt daardoor lang je eigen plannen opzij moeten schuiven, en daar ben ik je enorm dankbaar voor.

Elise, bij dezen beloof ik plechtig, dat ik nooit meer zo onnadenkend zal zijn om aan een promotieonderzoek te beginnen. Ik zal voortaan andere stomme beslissingen nemen ...

Curriculum vitae



Ate van der Burgt werd op 17 maart 1978 geboren in Amsterdam. In 1996 behaalde hij het vwo-diploma aan het Mencia de Mendoza Lyceum in Breda. In datzelfde jaar begon hij met de studie Bioprocestechnologie aan Wageningen Universiteit. In zijn eerste afstudeervak, bij de vakgroep proceskunde, bestudeerde hij de metabole flux in de glycolyse van Lactoccocus lactis. Eind 2002 liep hij stage bij de afdeling Bio-informatica van Plant Research International, onderdeel van Wageningen UR, waarbij hij ondersteunende software voor de analyse van eiwit-massaspectrometriedata ontwikkelde. Bij ditzelfde onderzoeksinstituut deed Ate een MSc-afstudeeronderzoek waarbij hij micro-RNA's bestudeerde en voorspelde in genomen van chloroplasten. Al tijdens zijn MSc combineerde Ate studie met atletiek op nationaal niveau (midden-lange afstand). Een tweede uit de hand gelopen hobby is het zoeken naar Alpenmineralen in het Binntal, Zwitserland. Hierover publiceert hij regelmatig in populairwetenschappelijke tijdschriften.

September 2004 startte hij zijn PhD-onderzoek 'Comparative Fungal Genomics' bij Plant Research International. In de jaren die volgden combineerde hij zijn PhD-onderzoek met atletiek en behaalde hij, onder supervisie van Jeroen Zeinstra, aansluiting bij de Europese subtop op de 1500m. Dit omlijstte hij met enkele nationale titels en deelname aan meerdere EK´s en Europacup. Voor een gedeelte van die periode werd Ate professioneel ontzien en financieel ondersteund door Wageningen UR. Op 19-honderdste van een seconde miste Ate in 2007 deelname aan de Wereldkampioenschappen in Japan. Echter, met een tijd van 3.38'19" op de 1500m staat hij momenteel op de 11e plek van de Nationale Ranglijst allertijden.

In oktober 2010 pakte hij, op uitnodiging van prof. Dr. Ir. Pierre de Wit, zijn promotie fulltime op bij de vakgroep Fytopathologie van Wageningen Universiteit. Een flink aantal onderzoeksonderwerpen werd onderhanden genomen en zijn beschreven in dit proefschrift, met als terugkerende rode draad de schimmel Cladosporium fulvum en enkele nauw verwante Dothideomyceten. Zijn meest aansprekende publicaties betreffen de ontdekking van Introner-Like Elements in schimmels.

List of publications

Collemare J, Beenen HG, de Vries RP, Crous P, de Wit PJGM, **van der Burgt A**: Identification of Introner-Like Elements (ILEs) in closely related fungal species reveals high intron dynamics. Submitted.

Mesarich CH, Griffiths SA, **van der Burgt A**, Ökmen B,1 Beenen HG, Etalo DW, Joosten MHAJ, and de Wit PJGM: Transcriptome Sequencing Uncovers the *Avr5* Avirulence Gene of the Tomato Leaf Mould Pathogen *Cladosporium fulvum*. Mol Plant Microbe Interact. 2014 Mar 28. [Epub ahead of print].

Ökmen B, Collemare J, Griffiths S, **van der Burgt A**, Cox R, and de Wit PJGM: Functional analysis of the conserved transcriptional regulator CfWor1 in *Cladosporium fulvum* reveals diverse roles in the virulence of plant pathogenic fungi. Mol Microbiol. 2014, 92(1):10-27.

van der Burgt A, Severing E, Collemare J, de Wit PJGM: Automated alignment-based curation of gene models in filamentous fungi. BMC Bioinformatics 2014, 15(1):19.

van der Burgt A, Karimi Jashni M, Bahkali AH, de Wit PJGM: Pseudogenization in pathogenic fungi with different host plants and lifestyles might reflect their evolutionary past. Molecular Plant Pathology 2014, 15(2):133-144.

Collemare J, **van der Burgt A**, de Wit PJGM: At the origin of spliceosomal introns: Is multiplication of introner-like elements the main mechanism of intron gain in fungi? Communicative & Integrative biology 2013, 6(2):e23147.

van der Burgt A, Severing E, de Wit PJ, Collemare J: Birth of new spliceosomal introns in fungi by multiplication of introner-like elements. Current biology 2012, 22(13):1260-1265.

de Wit PJGM, **van der Burgt A**, Okmen B, Stergiopoulos I, Abd-Elsalam KA, Aerts AL, Bahkali AH, Beenen HG, Chettri P, Cox MP, Datema E, de Vries RP, Dhillon B, Ganley AR, Griffiths SA, Guo Y, Hamelin RC, Henrissat B, Kabir MS, Jashni MK, Kema G, Klaubauf S, Lapidus A, Levasseur A, Lindquist E, Mehrabi R, Ohm RA, Owen TJ, Salamov A, Schwelm A, Schijlen E, Sun H, van den Burg HA, van Ham RC, Zhang S, Goodwin SB, Grigoriev IV, Collemare J, Bradshaw RE: The genomes of the fungal plant pathogens *Cladosporium fulvum* and *Dothistroma septosporum* reveal adaptation to different hosts and lifestyles but also signatures of common ancestry. PLoS Genetics 2012, 8(11):e1003088.

Goodwin SB, M'Barek S B, Dhillon B, Wittenberg AH, Crane CF, Hane JK, Foster AJ, van der Lee TA, Grimwood J, Aerts A, Antoniw J, Bailey A, Bluhm B, Bowler J, Bristow J, **van der Burgt A**, Canto-Canche B, Churchill AC, Conde-Ferraez L, Cools HJ, Coutinho PM, Csukai M, Dehal P, De Wit PJGM, Donzelli B, van de Geest HC, van Ham RC, Hammond-Kosack KE, Henrissat B, Kilian A, Kobayashi AK, Koopmann E, Kourmpetis Y, Kuzniar A, Lindquist E, Lombard V, Maliepaard C, Martins N, Mehrabi R, Nap JP, Ponomarenko A, Rudd JJ, Salamov A, Schmutz J, Schouten HJ, Shapiro H, Stergiopoulos I, Torriani SF, Tu H, de Vries RP, Waalwijk C, Ware SB, Wiebenga A, Zwiers LH, Oliver RP, Grigoriev IV, Kema GH: Finished genome of the fungal wheat pathogen Mycosphaerella graminicola reveals dispensome structure, chromosome plasticity, and stealth pathogenesis. PLoS Genetics 2011, 7(6):e1002070.

van der Burgt A, Fiers MW, Nap JP, van Ham RC: *In silico* miRNA prediction in metazoan genomes: balancing between sensitivity and specificity. BMC Genomics 2009, 10:204.

Kema GH, van der Lee TA, Mendes O, Verstappen EC, Lankhorst RK, Sandbrink H, **van der Burgt A**, Zwiers LH, Csukai M, Waalwijk C: Large-scale gene discovery in the septoria tritici blotch fungus Mycosphaerella graminicola with a focus on *in planta* expression. Molecular Plant-Microbe Interactions : Molecular Plant-Microbe Interactions 2008, 21(9):1249-1260.

Fiers MW, **van der Burgt A**, Datema E, de Groot JC, van Ham RC: High-throughput bioinformatics with the Cyrille2 pipeline system. BMC Bioinformatics 2008, 9:96.

Kourmpetis YA, **van der Burgt A**, Bink MC, Ter Braak CJ, van Ham RC: The use of multiple hierarchically independent gene ontology terms in gene function prediction and genome annotation. In Silico Biology 2007, 7(6):575-582.

Hoefnagel MH, **van der Burgt A**, Martens DE, Hugenholtz J, Snoep JL: Time dependent responses of glycolytic intermediates in a detailed glycolytic model of *Lactococcus lactis* during glucose run-out experiments. Molecular Biology Reports 2002, 29(1-2):157-161.

Publications on the mineralogy of the Binntal area, Wallis, Switzerland

Cuchet S, Crumbach M, **van der Burgt A**, Brugger J: Die Region Binntal – Alpe Devero : Schauplatz einer Spektakulären Trennung. In preparation

Cuchet S, Crumbach M, **van der Burgt A**: I Ein auf unstetiger Kornvergrösserung basierendes Modell zur Entstehung der SEE-mineralisierten Quarz-Feldspat-Knauern in der Region Binntal, CH – Alpe Veglia, I. In preparation

Cuchet S, Crumbach M, **van der Burgt A**: Das Binntal enthüllt ein grosses Geheimnis: Entdeckung eines „neuen“ SEE-Mineralisationstyp im Gebiet Binntal, VS, CH, - Alpe Veglia, I, mit 5 neuen Mineralien für die Schweiz : Beta-Fergusonit-(Y), Genthelvin, Gramaccioliit-(Y), Hingganit-(Y) und Polykras-(Y). Schweizer Strahler Mai 2014, in press

Crumbach M, Cuchet S, **van der Burgt A**: Auf der Spur der vielleicht grössten Seltene-Erdelemente-Mineralisation der Alpen. Schweizer Strahler 2014, (1):8-13

Van der Burgt A, Cuchet S: Gasparit und Chernovit vom Chummibort, Binntal. Lapis 2006, 31(4) 35

Cuchet S, **van der Burgt A**, Meisser N: Chummibort, eine neue Fundstelle für Arsenmineralien im Binntal / Le Chummibort, une nouvelle localité à arsénites et arséniates du Binntal. Schweizer Strahler 2005, (2):19-29

Van der Burgt A, Cuchet S: Neu: Gadolinit, Aeschynit und Synchisit vom Fleschsee, Binntal (CH). Lapis 2005, 30(4): 19-29

Van der Burgt A, Cuchet S: Neufunde von Goyazit im Binntal, Wallis (Schweiz). Lapis 2002, 27(2): 18-19

Van der Burgt A, Cuchet S: Entdeckung von Torbernit am Gischigletscher, Binntal (VS). Schweizer Strahler 2000, 12(2): 85-87

De Wit FC, **van der Burgt A**: De Gischigletscher (Binntal, Wallis, Zwitserland). Over strahlen, kluften en de geologie & mineralogie. GEA 1996, 29(4):109-124

De Wit FC, **van der Burgt A**: Turtschi, een dolomiet-ontsluiting in het Binntal, Wallis, Zwitserland. GEA 1993, 26(4):137-139

Education certificate

Education Statement of the Graduate School

Experimental Plant Sciences

Issued to: **I.A. (Ate) van der Burgt**
Date: **12 May 2014**
Group: **Laboratory of Phytopathology, Wageningen University & Research Centre**

1) Start-up phase	*date*
► **First presentation of your project**	
Fungal Comparative Genomics	Oct 22, 2004
► **Writing or rewriting a project proposal**	
Project proposal Fungal Comparative Genomics	Dec 10, 2004
► **Writing a review or book chapter**	
MSc courses	
► **Laboratory use of isotopes**	
Subtotal Start-up Phase	*7.5 credits**

2) Scientific Exposure	*date*
► **EPS PhD student days**	
EPS PhD Student Day, Radboud University Nijmegen	Jun 02, 2005
EPS PhD Student Day, Wageningen University	Sep 13, 2007
► **EPS theme symposia**	
EPS theme 3 symposium 'Metabolism and Adaptation', Wageningen	Nov 06, 2007
EPS theme 4 symposium 'Genome plasticity', Wageningen	Dec 12, 2008
EPS theme 4 symposium 'Genome plasticity', Nijmegen	Dec 11, 2009
► **NWO Lunteren days and other National Platforms**	
Netherlands Conference on BioInformatics 2004 "Images of life", Groningen	Oct 04, 2004
ALW meeting Experimental Plant Sciences, Lunteren	Apr 04-05, 2005
NBIC: Netherlands Bioinformatics Conference, Ede	Apr 24, 2006
ALW meeting Experimental Plant Sciences, Lunteren	Apr 04-05, 2011
► **Seminars (series), workshops and symposia**	
Fungal Pathogenicity symposium, CBS, Utrecht (whole day)	Jul 08, 2005
Evolutionary Bioinformatics symposium, Utrecht (whole day)	Nov 16, 2006
EPS Flying seminar James C. Carrington:	Mar 26, 2007
Invited seminar Regine Kahmann: Effectors of the plant-pathogen *Ustilago maydis*	Oct 29, 2010
EPS/Plantum symposium 'Intraspecific pathogen variation'	Jan 22, 2013
Invited seminar Howard Judelson: Molecular Insights into Spore Biology and Metabolism of *Phytophtora infestans*, the Potato Blight Pathogen	May 07, 2013
Invited seminar Rays Jiang: Integrative genomics of destructive pathogens from oomycetes to malaria parasites	May 07, 2013

	Invited seminar Brian Staskawicz:	May 21, 2013
	Invited seminar Eric Schranz: Whole genome duplications as drivers of evolutionary innovations and radiations?	Nov 21, 2013
►	**Seminar plus**	
	James Carrington: EPS Flying seminar	Mar 26, 2007
►	**International symposia and congresses**	
	Benelux Bioinformatics Conference, Gent, Belgium	Apr 14-15, 2005
	Comp. Genomics of Euk. Micro-Organisms, Spain	Nov 12-17, 2005
	M. gramminicola annotation jamboree, JGI / Walnut Creek (CA, USA)	Jun 07-09, 2006
	Comp. Genomics of Euk. Micro-Organisms, Spain	Oct 20-25, 2007
	Benelux Bioinformatics Conference, Leuven, Belgium	Nov 12-13, 2007
	Benelux Bioinformatics Conference, Maastricht, Netherlands	Dec 15-16, 2008
►	**Presentations**	
	Oral: Benelux Bioinformatics Conference 2006, Wageningen	Oct 17-18, 2006
	Oral: 1th joint WUR-Marburg meeting: Annotated gene structure assessment and improvement	Oct 2010
	Oral: 2th joint WUR-Marburg meeting: Introner-like Elements in Fungi	Jan 2012
	Oral: ECFG 11 Dothideomycete Sattelite Meeting, Marburg: Introner-like Elements in Fungi	Apr 03, 2012
	Oral: ALW Platform Molecular Genetics Annual Meeting, Lunteren: Introner-like Elements in Fungi	Oct 04-05, 2012
	Oral: Benelux Bioinformatics Conference 2012, Nijmegen	Dec 10-11, 2012
►	**IAB interview**	
	Meeting with a member of the International Advisory Board	Sep 14, 2007
►	**Excursions**	
	Joint Lab meeting Regine Kahmann group (Max-Planx Institut, Marburg), Wageningen	Oct 2010, 2 days
	Joint Lab meeting Regine Kahmann group (Max-Planx Institut, Marburg), Marburg	Jan 2012, 2 days
	Subtotal Scientific Exposure	*18.5 credits**
3) In-Depth Studies		<u>date</u>
►	**EPS courses or other PhD courses (higly recommended)**	
	Postgraduate course 'Molecular Phylogenies: Reconstruction and Interpretation'	Oct 17-21, 2005
	Postgraduate course 'Multivariate Analysis'	Apr 19-27, 2006
►	**Journal club**	
	Literature discussion Lab.of Bioinformatics / Applied Bioinformatics	2004-2008
►	**Individual research training**	
	Subtotal In-Depth Studies	*6.0 credits**
4) Personal development		<u>date</u>
►	**Skill training courses**	
	WGS course 'Scientific Writing'	Feb-Apr 2006
	WGS course 'Career Orientation'	May-Jun 2009
	EPS ExPectationS Day: Scientific Integrity and Dealing with Supervisors	Mar 28, 2014
►	**Organisation of PhD students day, course or conference**	
►	**Memberschip of Board, Committee or PhD council**	
	Subtotal Personal Development	*3.6 credits**
TOTAL NUMBER OF CREDIT POINTS*		**35.6**

Herewith the Graduate School declares that the PhD candidate has complied with the educational requirements set by the Educational Committee of EPS which comprises of a minimum total of 30 ECTS credits

** A credit represents a normative study load of 28 hours of study.*

This research was conducted at the Laboratory of Phytopathology of Wageningen University and at Applied Bioinformatics, Plant Research International,part of Wageningen UR.

Printed by: Wöhrmann Print Service, Zutphen, NL