Mating locus structure in *Ceratocystis moniliformis* and a gene duplication in *Ceratocystis* species

T. A. GODLONTON

Mating locus structure in *Ceratocystis moniliformis* and a gene duplication in *Ceratocystis* species

by

Tracy Alison Godlonton

Submitted in partial fulfilment of the requirements for the degree

MAGISTER SCIENTIAE

In the faculty of Natural & Agricultural Science

University of Pretoria

Pretoria

28 July 2014

Study Leaders:

Prof. Brenda D. Wingfield

Mr. P. Markus Wilken

Prof. Michael J. Wingfield

Declaration

I, Tracy Alison Godlonton, hereby declare that the dissertation, which I hereby submit for the degree *Magister Scientiae* at the University of Pretoria, is my own work and has not been submitted by me for a degree at this or any other tertiary institution.

Tracy A. Godlonton

28 July 2014

TABLE OF CONTENTS

ACKNOWLEDGEMENTS

I would like to express my heartfelt thanks and appreciation to the following:

Most importantly to my Heavenly Father Who has so graciously guided me. May He receive all the glory.

My supervisors, Prof. Brenda Wingfield for her support, guidance and remarkable mentorship not only academically but personally too. To Prof. Mike Wingfield for his guidance and support. To Mr Markus Wilken for his continued support on a practical level and for challenging me to broaden my thinking.

My colleagues and friends at FABI, especially Anandi Reitmann, Phia van Coller and Kershney Naidoo for their support, encouragement and help during this degree.

The University of Pretoria (UP), members of the Tree Protection Co-operative Programme (TPCP), the DST-NRF Centre of Excellence (CoE) in Tree Health Biotechnology and the National Research Foundation (NRF) for financial assistance.

My amazing parents for their prayers, encouragement and continued support over the past 7 years.

My dear late Grandfather, Hamilton who always expressed such interest in fungi.

My incredible husband, Christopher, for your love, encouragement, motivation, understanding and support throughout this journey.

PREFACE

The genomics era is advancing rapidly as new technologies are developed to uncover a plethora of information regarding species biology. Whole genome sequencing can be used to directly investigate the genetic make up of a species with increasing accuracy. In the Fungal Kingdom, this technology is providing invaluable insights into pathogenicity, mating systems and species evolution, to name a few. The genus, *Ceratocystis* has a number of species that are prominent in the forestry industry as pathogens and saprophytes. With the availability of a number of genomic sequences belonging to the *Ceratocystis* genus, an opportunity presented itself for gene discovery and interspecies comparisons.

The aim of this study was firstly to increase the number of genome sequences in the *Ceratocystis* genus with that of *C. moniliformis. S*econdly, to investigate the mating type system of *C. moniliformis* and finally to compare a repertoire of genes between species spanning the genus to gain insights into pathogenicity and species evolution. The thesis is comprised of three independent chapters.

The first chapter delivers a historical review of genomics and the progression of the technologies in this field with special reference to fungal research. The ability of genome studies to examine individual isolates raises the question as to whether model organisms have been rendered redundant as a tool for biological research. The advantages and disadvantages of each of these research avenues provided a platform for a discussion regarding the relevance of both. The impact of genomics on microbial research in terms of new technologies and applications clearly depict the importance of the genomics era.

Chapter 2 deals with the sequencing of a saprophytic fungus belonging to the *Ceratocystis* genus, *C. moniliformis.* Subsequently, the assembled genome was exploited to determine the mating type structure and gene organisation of this species using bioinformatic tools. *Ceratocystis moniliformis* is a homothallic fungus which would typically contain the genes from both the opposite mating type idiomorphs. Another species belonging to this genus, *C. fimbriata* undergoes mating type switching and has been thoroughly characterised. With the availability of the assembled genome, the mating type locus in *C. moniliformis* was identified and compared to that of *C. fimbriata* and those of other homothallic species.

In chapter 3, the repertoire of *β-fructofuranosidase* genes present in 17 genomes spanning the *Ceratocystis* genus was investigated. A genomic comparison was performed on 17 genomes belonging to the *Ceratocystis* genus. The β-fructofuranosidase proteins are implicated in carbohydrate metabolism and have been reported to play a possible role in pathogenicity. An interest in these genes led to an investigation of their presence in all 17 genomes. Phylogenetic analysis was used to provide insights into the involvement of duplications in the evolution of species in this genus.

The genome sequence of *C. moniliformis* has added to the collection of sequenced genomes in the *Ceratocystis* genus and affords opportunities for thorough species comparisons in the future. Together, this research adds to the knowledge of mating in fungi as well as population biology and species evolution in *Ceratocystis.*

CHAPTER 1

The Impact of Genomics on Fungal Research

1. Introduction

Genomics encompasses the determination of the complete DNA sequence of an organism in combination with bioinformatics for the assembly of the genome, identification of genes and genome analyses (Kuo *et al.* 2014). Although the first genome (bacteriophage MS2 RNA) was sequenced in 1976 (Fiers *et al.*), it wasn't until the sequencing of the *Saccharomyces cerevisiae* genome in 1996 (Goffeau *et al.*) that the field of mycology truly entered the genomics era (Foster *et al.* 2006). Ultimately, it was the completed draft human genome (Lander *et al.* 2001) that launched genomics into the international scientific arena. This resulted in the ability to unravel biological phenomena from genome data and to rapidly develop associated technologies.

Substantial work on the genomes of different yeast species laid the foundation for genomics of fungal species and higher Eukaryotes. Since the publication of the genome of the baker's yeast, *Saccharomyces cerevisiae*, in 1996 (Goffeau *et al.*), the number of filamentous fungal genomes that have been sequenced has increased rapidly with this fungus serving as a model for other fungi. *Encephalitozoon cuniculi,* an atypical fungus that is an obligate animal parasite (Katinka *et al.* 2001) and another yeast, *Schizosaccharomyces pombe* (Wood *et al.* 2002) were the second and third fungal genomes to be sequenced.

The field of fungal genomics gained momentum in November 2000 when Gerry Fink of the Broad Institute (then the Whitehead Institute) established the Fungal Genomes Initiative (FGI) (Foster *et al.* 2006). The aim of the FGI was to propose genome sequencing for a broad range of fungal species. These species were selected based on their potential to provide insights into their ecological, evolutionary, medical and basic biological characteristics (Birren *et al*. 2002; 2003a; 3003b; 2004). The first filamentous fungus for which the genome was sequenced was red bread mould fungus, *Neurospora crassa* (Galagan *et al.* 2003). By 2006, on-going as well as completed fungal genome projects amounted to 115 (Foster *et al.* 2006). These included projects run by the FGI and The Institute for Genome Research (TIGR) among other public as well as private initiatives. To date, there are 1903 fungal genome projects listed on Gold Genomes (http://www.genomesonline.org/), while the Joint Genome Institute (JGI) site reports 1568 fungal whole genome projects (http://genome.jgi.doe.gov/programs/fungi/fungal-projects.jsf) and all data were retrieved on 12/04/2014.

The numerous fungal genomes available to the scientific community have provided attractive opportunities for comparative genomics of different species promoting an improved understanding of their genomic structure and function. Comparative genomics is a valuable tool to better understand the genes involved in host specialisation, evolution, development and pathogenicity. Applications of comparative genomics have helped to determine the role of gene losses and gains as well as horizontal gene transfer in the evolution of prokaryotes (Koonin and Wolf 2008) and some filamentous fungi (Andersen *et al.* 2011).

2. Model Organisms

2.1 *Saccharomyces cerevisiae:* the platform for eukaryotic genetic research

Saccharomyces cerevisiae has provided a foundation for fungal genome studies. The genome of this model yeast was sequenced via a worldwide collaboration of some 600 scientists (Goffeau *et al.* 1996) and it gave rise to a substantial body of research in subsequent years (Figure 1). One of the first steps towards understanding the data emerging from the genome sequence was to functionally characterise novel genes. A research network, the European Functional Analysis Network (EUROFAN) (Oliver 1996) undertook the analysis of these genes along with other teams from the United States, Canada and Japan (Goffeau *et al.* 1996). The knowledge gained from this vast project established a remarkable foundation for future research, especially in eukaryotic and fungal genomics.

To determine the function of every gene in an organism is an arduous task, though the knowledge of where and when a specific gene is expressed may provide evidence for a reasonable hypothesis (DeRisi *et al.* 1997). Researchers from Stanford in the USA developed a transcriptome microarray slide for *S. cerevisiae* in order to measure gene expression profiles (Schena *et al.* 1995; Shalon *et al.* 1996). The genome provided almost all of the *S. cerevisiae* gene sequences that were printed as targets in a high-density array on a glass microscope slide. The experimental procedure began with the extraction of cellular mRNA which was then easily converted to cDNA by reverse transcription. These sequence data were hybridised to the microarray chip and the fluorescence of different colours could then indicate increased and decreased expression levels. This technique allowed for a rapid method to investigate gene expression differences on a genomic scale to aid in the identification of gene functions. For this reason, microarrays have become a popular tool for the elucidation of gene function (Nowrousian *et al.* 2005).

Three years after the development of the microarray for *S. cerevisiae*, one third of the yeast open reading frames (ORF) were still described as proteins of unknown function (Mewes *et al.* 1997; Uetz *et al.* 2000). In an attempt to rectify this shortcoming and to further reveal protein-protein interactions, a study was conducted using a yeast two-hybrid assay on predicted yeast ORFs (Uetz *et al.* 2000). This assay involves the establishment of a library of yeast ORF's that are each hybridised with a GAL4 activation domain. The roughly 6000 predicted proteins are each fused to a GAL4 DNA binding domain. Both the activation domain and the DNA binding domain are required for a particular downstream gene to be expressed. If the yeast proteins attached to the respective domains interact, the domains are brought together resulting in the expression of the reporter gene. In this way, interacting yeast genes can be identified. This breakthrough allowed for uncharacterised proteins to be assigned an appropriate biological function based on their interactions with other known proteins, or at least link proteins that function together in a cell (Uetz *et al.* 2000).

In 2001, a revolutionary proteome chip was developed for global analysis of protein activities (Zhu *et al*. 2001). In this study, 5800 ORFs were cloned (80% of all the *S. cerevisiae* ORFs) and the proteins over-expressed. These proteins were printed onto slides to generate a yeast proteome micro-array. The chip was utilised to identify calmodulin-binding proteins and characterise a common potential binding motif for these proteins. This technology provided a means to screen interactions between proteins on the chip, and other proteins or phospholipids. Furthermore it allowed for the elucidation of post-translational modifications (Zhu *et al.* 2001).

A further advance in terms of understanding S. *cerevisiae* gene function was the gene knock-out project which produced deletions of 96% of the organism's open reading frames (Giaever *et al.* 2002). Each of the genes was completely deleted and replaced with a "deletion cassette module". This represents a form of "reverse genetics" that is possible only where a complete genome sequence is available. Here, the gene is identified first and subsequently the phenotype analysed or linked to the gene; as opposed to the formidable task of generating random mutants and then attempting to identify the gene responsible for the observed phenotype (Giaever *et al.* 2002). These knock-out libraries provided a means by which to allocate gene functions based on the resultant phenotype of the yeast (Foster *et al.* 2006).

Yeasts and particularly *S. cerevisiae* became interesting to scientists as early as the 1800's, largely because of their role in in alcoholic fermentation (Barnett 2003). Studies on *S. cerevisiae* contributed substantially to the fields of microbiology and biochemistry, and this has continued to be true after the genome of this fungus was sequenced. The wealth of data and the advanced techniques that have emerged from studies of this organism have led to it being referred to as the "biotechnology workhorse" (Hofmann *et al.* 2003). While its features make it an opportune resource for use in genetics research, their unique genomes which are aneuploid, polyploid or hybrids, hinders the transfer of all the acquired knowledge and techniques to other species (Hofmann *et al.* 2003).

2.2 The model organism debate

For many decades, research has focussed very strongly on the genetics and physiology of various model organisms across the Kingdoms of Life. Model organisms are defined as species that have been the target of extensive research made available to the scientific community and have well established techniques for genomic manipulation (Fields and Johnston 2005). These organisms serve as representatives for other related species, and it is assumed that the knowledge gained will be transferable across related species. However, with the advent of technology to sequence and analyse genomes from non-model species, the question arises as to whether this will decrease our dependence on model organisms.

During the mid-1900's, scientists sought to unravel the mystery of how cells functioned at a molecular level using a reductionist approach. This was accomplished by focussing on central molecular mechanisms such as transcription, protein synthesis and replication. Bacteria and bacteriophages, the simplest forms of life were used for this purpose (Hunter 2008). As the fields of biology and technology expanded, research extended to more complex organisms and systems. This in turn required the use of higher organisms as model species including insects (*Drosophila melanogaster*), plants (*Arabidopsis thaliana*) (Hunter 2008) and fungi (*Saccharomyces cerevisiae*).

There were three main reasons why a reliance on model organisms for future research might no longer have been reasonable. Firstly, it was noted that the most basic universal questions have been answered, even if only partially. These questions relate to processes, including DNA repair, replication and signalling pathways that are relatively conserved throughout the Tree of Life (Fields and Johnston 2005). Secondly, the resources and tools which rendered the simple organisms as attractive subjects can now be applied to more complex organisms (Fields and Johnston 2005). Finally, with the advent of genomic studies on multiple, non-model organisms, there is an expanding resource of genes and proteins available and research can be directed at these specific non-model organisms. Researchers no longer need to base their views regarding the genetics of a particular test subject on what is known for a model organism. Rather, a genome sequence can reveal the exact gene sequences that exist in the organism of interest and these can then be directly investigated and described.

Having a genome sequence available for an organism has been compared to having a dictionary with lists of words, however lacking the associated word descriptions (Foster *et al.* 2006). In order for a newly sequenced genome to be interpreted at a biological level, it is necessary to assign gene functions and understand organisational implications of the genome. Substantial funding has been invested to accomplish whole genome sequencing for a number of model organisms (Fields and Johnston 2005). Several databases, such as The BioKnowledge Library (http://www.proteome.com), have been compiled that contain sets of data describing genes, primarily from model organisms (Costanzo *et al.* 2001). Information available includes the biochemical function, structure and protein domains, regulation, genetic interactions, localisation and mutant phenotype among other relevant data. Scientists who aim to explain biochemical pathways or annotate uncharacterised genes in new organisms, can access these databases and search for gene orthologs in model organisms (Costanzo *et al.* 2001). The data can then be extrapolated to the genome of the organism being studied. In this way research conducted on model organisms remains invaluable in the genomics era.

The use of popular model organisms will continue in the years to come for several reasons. As technology advances, deeper insights and methods for observing the basic biological processes will develop (Fields and Johnston 2005). Model organisms will offer the best opportunities to achieve this due to their well characterised biology. Furthermore, researchers who identify a pathogen or disease implicated gene in humans or plants can use applicable model organisms to ethically study interactions and functions of these genes thoroughly. For example, the puffer fish, *Fugu rubripes* is an ideal model organism to study human genes because genomic analysis has revealed that three quarters of predicted human proteins matched genes in the puffer fish genome (Aparicio *et al.* 2002).

Model organisms will be increasingly used for direct investigation of medical problems in humans (Fields and Johnston 2005) and pathogenicity in agriculture. Important genes and pathways implicated in the regulation of aging have been elucidated by studying model organisms (Guarente and Kenyon 2000). The advantage of studying this process in yeast is that direct mutants can be identified and the phenotypic implications described. These results can then offer insights into such processes in humans. Another example includes the investigation of pathogenic pathways in fungi. Comparative genomics (Tettelin *et al.* 2008; Ellegren 2008) and bioinformatic systems biology (Alves *et al.* 2008; Karathia *et al.* 2011) can be applied to the genomes of *S. cerevisiae* and the related pathogenic species, *Candida albicans* to identify genetic anomalies that may be linked to pathogenicity (Fields and Johnston 2005).

Model organisms such as the yeast *S. cerevisiae* and the fruit fly, *Drosophila melanogaster*, are ideal for genetic studies because it is relatively easy to manipulate their gene sequences. These organisms have short generation spans (Hedges 2002), which allows for many generations to be analysed. In addition, model organisms such as these are also not complicated by ethical issues that arise with target organisms such as humans (Karathia *et al.* 2011).

Model organisms will remain the "testing ground" for the development of new technologies (Fields and Johnston 2005). Development of reliable techniques for experimental research is invaluable to the scientific community. The substantial amount of work performed on model organisms to optimise laboratory techniques has allowed for rapid advancement in genetic research. This is evident in the gene knock-out system developed in *S. cerevisiae* (Giaever *et al.* 2002). Other examples include a protocol for the study of pre-mRNA splicing developed on budding yeast (Meyer and Vilardell 2009) and the establishment of a Real-Time PCR-based approach to quantify bacterial RNA targets in water using *Salmonella* as a model organism (Fey *et al.* 2004). Likewise, the yeast-two-hybrid system developed in yeast has been used extensively for protein-protein interaction studies (Uetz *et al.* 2000; Yu *et al.* 2014).

Biological complexity is incompletely understood and this is a topic of substantial current interest. By combining the already well described pathways in model organisms, researchers will advance toward a more quantitative understanding of life processes (Fields and Johnston 2005). This will ultimately lead to the desired elucidation of a molecular interactions network. In order to accomplish this goal, a combination of genomic data sets, multiple technologies, bioinformatic specialists and genetic experimentalists will be needed. Thus genomic studies of non-model organisms need to be incorporated into studies with model organisms to achieve this goal. This evidence serves to demonstrate the continued value of model organisms for research in the genomics era.

3. Impact of Genomics on Research

Full genome sequencing has had a substantial impact on many aspects of biological research. The task of sequencing a genome carries with it various computational challenges, including the need to develop physical and genetic mapping algorithms, a requirement to work with large scale sequence assemblies, data processing, databases and annotation. The sheer volume of data that needs to be processed illustrates the problems that arise in terms of computing capacity. Before the availability of whole genome sequences, the identification and mapping of genes required the use of many laborious methods. However, with a full genome as a starting point, new techniques can be developed and used as alternatives.

Pathologists argue that novel therapies could be developed based on a deeper understanding of the fundamental biological processes that govern diseases (Fischer 2005). This holds true across all aspects of biology, including reproduction and evolution. Prior to genomic studies, biological issues were addressed using a reductionist approach limited to one or two molecular levels. Developments in high-throughput sequencing technologies have made increasingly quantitative views of the complexity of an organism's biology possible. These technologies monitor various molecular levels of a cell and are often referred to as the "-omics technologies" (Fischer 2005) (Figure 2).

By applying high throughput technologies, it is possible to undertake a systematic investigation of biological systems and to gain a more complete view of the species biology. These great advances are, however, hampered by the biological interpretations of the data that involve and rely heavily on computational data integration systems, predictive models and pipelines (Fischer 2005). The field of computational biology has had to develop rapidly to deal with the accumulation of large collections of data. The two study areas most impacted by the exponentially increasing number of available genomes are bioinformatics and comparative genomics. These are discussed briefly in the following section.

3.1 Bioinformatics

Bioinformatics involves the computational management and processing of genetic data and it has become one of the most valuable fields of modern science (Ouzounis and Valencia 2003). This field has needed to keep up with the large volumes of data emerging from genome sequencing. All the sequence, structural and functional data has led to the need for optimal storage as well as computational tools for analysis and organisation (Alves *et al.* 2008).

Computer simulations and mathematical algorithms used for molecular biology research were first published in the 1950s. This technology allowed for experimental results to be understood in an efficient progression for biological mechanisms (Alves *et al.* 2008). In the 1970's many key developments were made in the progress of bioinformatics with such studies as the first sequence alignment algorithms (Gibbs and McIntyre 1970; Needleman and Wunsch 1970); primary structure of proteins (Krzywicki and Slonimski 1967) and the popular use of molecular data in evolutionary studies (Jukes 1969).

By the end of the 1970's, it became necessary to develop computer archives for protein sequences and structures to be stored and distributed (Dayhoff 1978; Bernstein *et al*. 1977). This is also a trend that would grow substantially (Ouzounis and Valencia 2003). As massive volumes of data accumulated, the capabilities of computer tools and capacity of genomic high throughput (HTP) techniques increased drastically. This led to a shift from reductionism towards an integrative approach (Alves *et al.* 2008). The reductionist approach can be explained as dividing the problem into individual components that are easier to work with (Ahn *et al.* 2006) and thus only focusing on one facet or separate aspects of the complex process. Alternatively, the integrative systems approach is based on the notion that studying individual elements of a system does not necessarily reflect the functioning of the entire system (Ahn *et al.* 2006).

The influx of genomic data and the desire for an integrative approach to dealing with them, has led to the development of databases and pipelines for storage, analysis and predictions of genomic sequences. An example of such a pipeline is the Gene Ontology (GO) project (Gene Ontology, C. 2004). This Consortium aims to establish firstly, a set of ontologies to describe critical concepts in molecular biology and secondly, the software that can be publically accessed for the annotation of genomic sequence. The overall purpose is to achieve worldwide conservation of biological terminology enabling accurate and common interpretation of data for all researchers. Similarly, a pipeline such as MAKER (Cantarel *et al.* 2008) functions to set up a database and annotate eukaryotic genomes in a simplistic manner. This is beneficial for small scientific communities that lack bioinformatics expertise. For a systems approach, a database such as Kyoto Encyclopedia of Genes and Genomes (KEGG) is favourable. The database elucidates gene functions based on networks of molecules and diagrams of known biochemical pathways are also stored (Ogata *et al.* 1999).

These online programs and pipelines make genomic sequence data readily available for research groups. Further development of systems biology based pipelines will allow for a more thorough, integrated understanding of the manner in which biological pathways interact to result in a particular function. Having the means by which to analyse vast amounts of genomic data, an example being the 1000 fungal genome project (http://1000.fungalgenomes.org/), may also lead to a continual increase in genomic sequencing efforts. An advantage is that a collection of multiple genomes spanning many species allows for an even deeper investigation of genetics.

3.2 Comparative Genomics

The growing volumes of research performed on genomes of model and other organisms and bioinformatics has led to the development of the field of comparative genomics. Here comparisons are drawn between genomes of various species (Foster *et al.* 2006). These comparisons can be focussed on the complete gene sets, organisation of genes or specific gene clusters, depending on the research question. Due to the substantial number of genomic sequences available, as predicted by Hofmann *et al*. (2003), comparative genomics has become a major tool for gene identification.

3.2.1 Application in Gene Discovery

Comparative genomics draws its strength from synteny, the conservation of gene order as well as the DNA sequence of genes in closely related organisms (Foster *et al.* 2006). This can be exploited in gene discovery, when searching for particular genes in a poorly characterised genome. The gene can be located in an annotated genome and subsequently, the equivalent genomic region in the related genome of interest investigated. Prior to the use of genome comparisons for gene discovery, degenerate primers designed based on characterised genes as well as hybridisation techniques were pivotal to determine the presence and sequence of a gene in a genome.

An example of the prior technique is provided by the work of Glass *et al.* (1988) for the elucidation of mating type genes in the model species, *Neurospora crassa.* Glass *et al.* (1998) cloned the A mating type gene by exploiting its proximity to a selectable marker. The cloned sequence was then used in hybridisation assays as a probe to search for common DNA that flanks mating type genes in opposite mating isolates. Their aim was to determine the sequence of the “a” mating type gene, the alternate gene present in an individual that has to cross with an A mating type individual to reproduce. This method can be time consuming and inefficient. Furthermore, sequences may be too divergent for adequate hybridisation and researchers may forfeit the ability to view the complete gene sequence.

The ability to compare genomic data allows for an *in silico* BLAST search of syntenic sites in genomes to identify and characterise genes. This was demonstrated in the genome comparison of necrotrophic plant pathogenic fungi, *Sclerotinia sclerotiorum* and *Botrytis cinerea* (Amselem *et al.* 2011). These broad host-range pathogens are close relatives, however they differ in reproductive behaviour. This dissimilarity was explained by comparing the genes present at the mating type loci of each fungus. Whole idiomorph characterisation was possible from the assembled genome sequences. Differences in mating behaviour were therefore explained based on the variation in gene content at the mating type loci.

The *in silico* comparative approach illustrated above is largely beneficial for studying organisms that differ in specific characteristics to elucidate genes of particular interest. Prior to the era of genomics, specific candidate genes were targeted or mutants generated to elucidate genes responsible for specific lifestyles such as parasitism (Foster *et al.* 2006). Now, genomes of related species that possess alternate ecological characteristics can rather be compared computationally (Foster *et al.* 2006). Genes or suites of genes can then be directly identified and tested.

The term "reverse-ecology" (Li *et al.* 2008) has been applied to the research approach where genomic discrepancies are identified and then functionally described as opposed to a focus on a specific trait and then searching for the gene targets. Reverse-ecology harnesses the power of comparative genomics to identify target genes of natural selection by analysing genetic diversity within and between species populations (Ellison *et al.* 2011; Akey *et al.* 2004; Begun *et al.* 2007). This can, in turn, uncover the genes that control ecological and evolutionary traits (Li *et al.* 2008).

Local adaptations to temperature were elucidated by the identification of high divergence regions between *N. crassa* populations using reverse-ecology. This unbiased whole-genome approach identifies differences in the genome and thereafter, the function or implication of such divergence is deduced. Subsequent growth assays showed differences in temperature fitness and BLAST analyses of the genes involved indicated associations with cold response and circadian clocks (Ellison *et al.* 2011). In this way, the ecologically adaptive phenotypes were assigned to regions of interest identified by comparative genomics.

Another example of reverse-genetics can be found in studies of the model organism *S. cerevisiae*. The combination of reverse-genetics and the tools and data available for a model organism have been exploited in order to study fundamental traits in biology. The genome of *S. cerevisiae* was analysed for tandem repeats (short stretches of repetitive DNA) (Vinces *et al.* 2009). It was demonstrated that tandem repeats occupy nucleosome free genomic regions, and these were consistently embedded at promoter regions, thus affecting the expression of genes. This evidence made it possible for Vinces *et al.* (2009) to speculate that tandem repeats affect the local chromatin structure and thereby play a role in the "evolutionary tuning" of gene expression. By first analysing the genome, avenues emerged to explore core mechanisms of gene regulation leading to evolutionary diversity.

3.2.2 Application in Microbial Studies

While a “reverse” approach has been used on other systems such as mice and trees (Turner *et al.* 2010; Harr 2006) it may be most useful in microbial research (Ellison *et al.* 2011). Due to the fact that a nearly complete reference genome is needed for a reverse-genetics method of study (Storz and Wheat 2010), the concern arises as to whether it is worth studying non-model organisms, that are currently lacking genome sequences, in this manner (Ellison *et al.* 2011). But the generally smaller genome size and lower repeat density compared with macrobes, has made it possible to generate reliable *de novo* assemblies of microbial genomes using low-cost short reads. With the advanced sequencing technologies available, it is thus plausible to generate a reference genome from a single individual and to compare it with many other individuals re-sequenced at a lower coverage (Ellison *et al.* 2011).

The advantage of comparing genomes of species differing in life-style and habitat is the ability to thoroughly identify notable genetic differences *in silico* and then focus on these areas. Such a comparative study was conducted on the genomes of two primary mammalian pathogens, *Coccidioides immitis* and *Coccidioides posadasiia,* a non-pathogen, *Uncinocarpus reesii* and that of a distantly related plant pathogen, *Histoplasma capsulatum*. Analyses revealed expansions and contractions of gene family size associated with a host/substrate shift from plants to animals (Sharpton *et al.* 2009). Coverage of the genomes ranged from 14.5X down to 5X. By considering fungi with different ecological associations, for example pathogen vs non-pathogen and plant pathogen vs animal pathogen, differences can be targeted and interrogated for associations with particular phenotypes.

Comparative genomics has revolutionised the manner in which research is performed. Rather than randomly searching for the genes responsible for specific traits, it is possible to initially compare genomes that differ with regards to a trait and identify genetic discrepancies between the genomes (Ellison *et al.* 2011). The genetic data can thus be used as a starting point and worked backward to explain the phenotypic result and thereby often increasing the efficiency of research.

4. Conclusions

The field of genomics has dramatically enhanced the way in which we study the biology of organisms. Fungal genomics emerged after the sequencing of *S. cerevisiae* over 15 years ago and it is a field that has continued to advance exponentially since that time. Numerous model organisms have been well characterised at the genome level and these provide a basis to study related organisms. Debates as to whether genomic studies of individual organisms will render model organisms redundant tend to favour the value of model organisms.

Genomics has shifted biological research from a disparate focus of DNA only or single gene analyses to an integrative approach. It now is possible to identify all aspects of a particular biological pathway by using, to name a few, genomics and transcriptomics along with molecular techniques to elucidate all the factors that contribute to the pathway as a system. Comparative genomics is used along with "reverse-ecology" to identify specific areas to target, eliminating the tedious task of searching for a gene that may not even be relevant.

The fungal genus *Ceratocystis* that provides the focus of this dissertation comprises a number of economically relevant plant pathogens and interesting saprophytes. *Ceratocystis moniliformis* is vectored by nitidulid beetles to tree wounds and is generally considered as a saprophyte or weak pathogen (van Wyk *et al.* 2004; Roux *et al.* 2004; Tarigan *et al.* 2010). This is unlike the type species of the genus, *Ceratocystis fimbriata* which is a very serious plant pathogen known to cause black rot on sweet potato (Baker *et al.* 2003; Barnes *et al.* 2003; van Wyk *et al.* 2007). Another difference between *C. fimbriata* and *C. moniliformis* is the lack of mating type switching in the latter species (Witthuhn *et al.* 2000; Harrington 2007). Apart from the lack of mating type switching, nothing is known about the mating system of this fungus. With the recently sequenced genome of *C. fimbriata* available, obtaining the genomic sequence of *C. moniliformis* would allow us to answer a number of questions relating to this fungus using a comparative approach. The power of comparative genomics can in this case be harnessed in order to describe the mating type system and uncover interesting gene duplications in the genus as a whole.

7. References

Ahn, A.C., Tewari, M., Poon, C-S., Phillips, R.S. (2006). The limits of reductionism in medicine: Could systems biology offer an alternative? *PLoS Med*. **3**, e208.

Akey, J.M., Eberle, M.A., Rieder, M.J., Carlson, C.S., Shriver, M.D., Nickerson, D.A., Kruglyak, L. (2004). Population history and natural selection shape patterns of genetic variation in 132 genes. *PLoS Biol*. **2**, e286.

Alves, R., Vilaprinyo, E., Sorribas, A. (2008). Integrating bioinformatics and computational biology: Perspectives and possibilities for *in silico* network reconstruction in molecular systems biology. *Current Bioinformatics*. **3**, 98-129.

Amselem, J., Cuomo, C.A., van Kan, J.A.L., Viaud, M., Benito, E.P., Couloux, A., Coutinho, P.M., de Vries, R.P., Dyer, P.S., Fillinger, S. (2011). Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. *PLoS genetics*. **7**, e1002230.

Andersen, M. R., Salazar, M. P., Schaap, P. J., van de Vondervoort, P. J. I., Culley, D., Thykaer, J., Frisvad, J. C., Nielsen, K. F., Albang, R., Albermann, K., Berka, R. M., Braus, G. H., Braus-Stromeyer, S. A., Corrochano, L. M., Dai, Z., van Dijck, P. W. M., Hofmann, G., Lasure, L. L., Magnuson, J. K., Menke, H., Meijer, M., Meijer, S. L., Nielsen, J. B., Nielsen, M. L., van Ooyen, A. J. J., Pel, H. J., Poulsen, L., Samson, R. A., Stam, H., Tsang, A., van den Brink, J. M., Atkins, A., Aerts, A., Shapiro, H., Pangilinan, J., Salamov, A., Lou, Y., Lindquist, E., Lucas, S., Grimwood, J., Grigoriev, I. V., Kubicek, C. P., Martinez, D., van Peij, N. I. N. M. E., Roubos, J. A., Nielsen, J., and Baker, S. E. (2011). Comparative genomics of citric-acid-producing *Aspergillus niger* ATCC 1015 versus enzyme-producing CBS 513.88. *Genome Research* **21,** 885-897.

Aparicio, S., Chapman, J., Stupka, E., Putnam, N., Chia, J-M., Dehal, P., Christoffels, A., Rash, S., Hoon, S., Smit, A., Gelpke, M.D.S., Roach, J., Oh, T., Ho, I.Y., Wong, M., Detter, C., Verhoef, F., Predki, P., Tay, A., Lucas, S., Richardson, P., Smith, S.F., Clark, M.S., Edwards, Y.J.K., Doggett, N., Zharkikh, A., Tavtigian, S.V., Pruss, D., Barnstead, M., Evans, C., Baden, H., Powell, J., Glusman, G., Rowen, L., Hood, L., Tan, Y.H., Elgar, G., Hawkins, T., Venkatesh, B., Rokhsar, D., Brenner, S. (2002). Whole-genome shotgun assembly and analysis of the genome of *Fugu rubripes*. *Science (New York, N.Y.)*. **297**, 1301-1310.

Baker, C. J., Harrington, T. C., Krauss, U., and Alfenas, A. C. (2003). Genetic variability and host specialization in the Latin American Clade of *Ceratocystis fimbriata*. *Phytopathology* **93,** 1274-1284.

Barnes, I., Roux, J., Wingfield, B. D., O'Neill, M., and Wingfield, M. J. (2003). *Ceratocystis fimbriata* infecting *Eucalyptus grandis* in Uruguay. *Australasian Plant Pathology* **32,** 361-366.

Barnett, J.A. (2003). Beginnings of microbiology and biochemistry: the contribution of yeast research. *Microbiology*. **149**, 557-567.

Begun, D.J., Holloway, A.K., Stevens, K., Hillier, L.D.W., Poh, Y.P., Hahn, M.W., Nista, P.M., Jones, C.D., Kern, A.D., Dewey, C.N. (2007). Population genomics: whole-genome analysis of polymorphism and divergence in *Drosophila simulans*. *PLoS biology*. **5**, e310.

Bernstein, F.C., Koetzle, T.F., Williams, G.J.B., Meyer Jr, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M. (1977). The protein data bank: A computer-based archival file for macromolecular structures. *Journal of Molecular Biology*. **112**, 535-542.

Birren, B., Fink, G., Lander, E.S. (2002). Fungal Genome Initiative, White paper developed by the Fungal Research Community. Whitehead Institute/MIT Center for Genome Research, Cambridge, MA 02141, USA. http://www.broad.mit.edu/annotation/fungi/fgi/ history.html.

Birren, B. (2003). Fungal Genome Initiative, A white paper for fungal comparative genomics. Whitehead Institute/MIT Center for Genome Research, Cambridge, MA 02141, USA. http://www.broad.mit.edu/annotation/fungi/fgi/history.html.

Birren, B., Fink, G., Lander, E.S. (2003). Fungal Genome Initiative, White paper for fungal comparative genomics. Whitehead Institute/MIT Center for Genome Research, Cambridge, MA 02141, USA. http://www.broad.mit.edu/annotation/fungi/fgi/history.html.

Birren, B. (2004). Fungal Genome Initiative, A White paper for fungal genomics. The Broad Institute of Harvard and MIT, Cambridge, MA 02141 USA. http://www.broad.mit.edu/annotation/fungi/fgi/history.html.

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

Costanzo, M.C., Crawford, M.E., Hirschman, J.E., Kranz, J.E., Olsen, P., Robertson, L.S., Skrzypek, M.S., Braun, B.R., Hopkins, K.L., Kondu, P., Lengieza, C., Lew-Smith, J.E., Tillberg, M., Garrels, J.I. (2001). YPD™, PombePD™ and WormPD™: model organism volumes of the BioKnowledge™ Library, an integrated resource for protein information. *Nucleic Acids Research*. **29**, 75-79.

Costanzo, M., Baryshnikova, A., Bellay, J., Kim, Y., Spear, E. D., Sevier, C. S., Ding, H., Koh, J. L. Y., Toufighi, K., Mostafavi, S., Prinz, J., St. Onge, R. P., VanderSluis, B., Makhnevych, T., Vizeacoumar, F. J., Alizadeh, S., Bahr, S., Brost, R. L., Chen, Y., Cokol, M., Deshpande, R., Li, Z., Lin, Z.-Y., Liang, W., Marback, M., Paw, J., San Luis, B.-J., Shuteriqi, E., Tong, A. H. Y., van Dyk, N., Wallace, I. M., Whitney, J. A., Weirauch, M. T., Zhong, G., Zhu, H., Houry, W. A., Brudno, M., Ragibizadeh, S., Papp, B. z., PÃ¡l, C., Roth, F. P., Giaever, G., Nislow, C., Troyanskaya, O. G., Bussey, H., Bader, G. D., Gingras, A.-C., Morris, Q. D., Kim, P. M., Kaiser, C. A., Myers, C. L., Andrews, B. J., and Boone, C. (2010). The genetic landscape of a cell. *Science* **327,** 425-431.

Dayhoff, M.O. (1978). Atlas of protein sequence and structure, Vol. 4, Suppl. 3. National Biomedical Research Foundation,Washington, D.C., U.S.A.

DeRisi, J.L., Iyer, V.R., Brown, P.O. (1997) Exploring the metabolic and genetic control of gene expression on a genomic scale. *Science* **278,** 680-686.

Dixon, S. J., Costanzo, M., Baryshnikova, A., Andrews, B., and Boone, C. (2009). Systematic mapping of genetic interaction networks. *Annual Review of Genetics* **43,** 601-625.

Ellegren, H. (2008). Comparative genomics and the study of evolution by natural selection. *Molecular Ecology*. **17**, 4586-4596.

Ellison, C.E., Hall, C., Kowbel, D., Welch, J., Brem, R.B., Glass, N.L., Taylor, J.W. (2011). Population genomics and local adaptation in wild isolates of a model microbial eukaryote. *Proceedings of the National Academy of Sciences*. **108**, 2831-2836.

Fey, A., Eichler, S., Flavier, Sb., Christen, R., Höfle, M.G., Guzmán, C.A. (2004). Establishment of a real-time PCR-based approach for accurate quantification of bacterial RNA targets in water, using *Salmonella* as a model organism. *Applied and Environmental Microbiology*. **70**, 3618-3623.

Fields, S., and Johnston, M. (2005). Whither model organism research? *Science* **307,** 1885-1886.

Fiers, W., Contreras, R., Duerinck, F., Haegeman, G., Iserentant, D., Merregaert, J., Min Jou, W., Molemans, F., Raeymaekers, A., and Van den Berghe, A. (1976). Complete nucleotide sequence of bacteriophage MS2 RNA: primary and secondary structure of the replicase gene. *Nature* **260,** 500-507.

Fischer, H. P., and El-Gewely, M. R. (2005). Towards quantitative biology: Integration of biological information to elucidate disease pathways and to guide drug discovery. *In* "Biotechnology Annual Review", Vol. Volume 11, pp. 1-68. Elsevier.

Foster, S.J., Monahan, B.J., Bradshaw, R.E. (2006). Genomics of the filamentous fungi - moving from the shadow of the bakers yeast. *Mycologist*. **20**, 10-14.

Galagan, J. E., Calvo, S. E., Borkovich, K. A., Selker, E. U., Read, N. D., Jaffe, D., FitzHugh, W., Ma, L.-J., Smirnov, S., Purcell, S., Rehman, B., Elkins, T., Engels, R., Wang, S., Nielsen, C. B., Butler, J., Endrizzi, M., Qui, D., Ianakiev, P., Bell-Pedersen, D., Nelson, M. A., Werner-Washburne, M., Selitrennikoff, C. P., Kinsey, J. A., Braun, E. L., Zelter, A., Schulte, U., Kothe, G. O., Jedd, G., Mewes, W., Staben, C., Marcotte, E., Greenberg, D., Roy, A., Foley, K., Naylor, J., Stange-Thomann, N., Barrett, R., Gnerre, S., Kamal, M., Kamvysselis, M., Mauceli, E., Bielke, C., Rudd, S., Frishman, D., Krystofova, S., Rasmussen, C., Metzenberg, R. L., Perkins, D. D., Kroken, S., Cogoni, C., Macino, G., Catcheside, D., Li, W., Pratt, R. J., Osmani, S. A., DeSouza, C. P. C., Glass, L., Orbach, M. J., Berglund, J. A., Voelker, R., Yarden, O., Plamann, M., Seiler, S., Dunlap, J., Radford, A., Aramayo, R., Natvig, D. O., Alex, L. A., Mannhaupt, G., Ebbole, D. J., Freitag, M., Paulsen, I., Sachs, M. S., Lander, E. S., Nusbaum, C., and Birren, B. (2003). The genome sequence of the filamentous fungus *Neurospora crassa*. *Nature* **422,** 859-868.

Gene Ontology, C. (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Research* **32**(suppl 1)**,** D258-D261.

Giaever, G., Chu, A. M., Ni, L., Connelly, C., Riles, L., Veronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., Andre, B., Arkin, A. P., Astromoff, A., El Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K.-D., Flaherty, P., Foury, F., Garfinkel, D. J., Gerstein, M., Gotte, D., Guldener, U., Hegemann, J. H., Hempel, S., Herman, Z., Jaramillo, D. F., Kelly, D. E., Kelly, S. L., Kotter, P., LaBonte, D., Lamb, D. C., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S. L., Revuelta, J. L., Roberts, C. J., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D. D., Sookhai-Mahadeo, S., Storms, R. K., Strathern, J. N., Valle, G., Voet, M., Volckaert, G., Wang, C.-y., Ward, T. R., Wilhelmy, J., Winzeler, E. A., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J. D., Snyder, M., Philippsen, P., Davis, R. W., and Johnston, M. (2002). Functional profiling of the *Saccharomyces cerevisiae* genome. *Nature* **418,** 387-391.

Gibbs, A. J., and McIntyre, G. A. (1970). The diagram, a method for comparing sequences. *European Journal of Biochemistry* **16,** 1-11.

Glass, N. L., Vollmer, S. J., Staben, C., Grotelueschen, J., Metzenberg, R. L., and Yanofsky, C. (1988). DNAs of the two mating-type alleles of *Neurospora crassa* are highly dissimilar. *Science* **241,** 570 - 573.

Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., and Oliver, S. G. (1996). Life with 6000 genes. *Science* **274,** 546-567.

Guarente, L., and Kenyon, C. (2000). Genetic pathways that regulate ageing in model organisms. *Nature.* **408**, 255-262.

Harr, B. (2006). Genomic islands of differentiation between house mouse subspecies. *Genome Research*. **16**, 730-737.

Harrington, T. C. (2007). The genus *Ceratocystis*: Where does the oak wilt fungus fit. In *Proceedings of the National Oak Wilt Symposium* (pp. 21-35).

Hedges, S.B. (2002). The origin and evolution of model organisms. *Nature Reviews Genetics*. **3**, 838.

Hofmann, G., McIntyre, M., and Nielsen, J. (2003). Fungal genomics beyond *Saccharomyces cerevisiae*? *Current Opinion in Biotechnology* **14**(2)**,** 226-231.

Hunter, P. (2008). The paradox of model organisms. *EMBO Rep*. **9**, 717-720.

Jukes, T.H. (1969). Recent advances in studies of evolutionary relationships between proteins and nucleic acids. *Space Life Sciences*. **1**, 469-490.

Karathia, H., Vilaprinyo, E., Sorribas, A., Alves, R. (2011). *Saccharomyces cerevisiae* as a model organism: a comparative study. *PLoS One*. **6**, e16015.

Katinka, M. D., Duprat, S., Cornillot, E., Metenier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., Delbac, F., El Alaoui, H., Peyret, P., Saurin, W., Gouy, M., Weissenbach, J., and Vivares, C. P. (2001). Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature* **414,** 450-453.

Koonin, E. V., and Wolf, Y. I. (2008). Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. *Nucleic Acids Research* **36,** 6688-6719.

Krzywicki, A., and Slonimski, P. P. (1967). Formal analysis of protein sequences: I. Specific long-range constraints in pair associations of amino acids. *Journal of Theoretical Biology* **17,** 136-158.

Kuo, A., Bushnell, B., Grigoriev, I. V., and Francis, M. M. (2014). Chapter one - Fungal genomics: sequencing and annotation. *In* "Advances in Botanical Research", Vol. Volume 70, pp. 1-52. Academic Press.

Lander, E.S., *et al.* International Human Genome Sequencing Consortium* (2001). Initial sequencing and analysis of the human genome. *Nature*. **409**, 860.

Li, Y.F., Costello, J.C., Holloway, A.K., Hahn, M.W. (2008). "Reverse ecology" and the power of population genomics. *Evolution*. **62**, 2984-2994.

Mewes, H. W., Albermann, K., Heumann, K., Liebl, S., and Pfeiffer, F. (1997). MIPS: a database for protein sequences, homology data and yeast genome information. *Nucleic Acids Research* **25,** 28-30.

Meyer, M., and Vilardell, J. (2009). The quest for a message: budding yeast, a model organism to study the control of pre-mRNA splicing. *Briefings in Functional Genomics & Proteomics* **8,** 60-67.

Needleman, S. B., and Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. *Journal of Molecular Biology* **48,** 443-453.

Nowrousian, M., Ringelberg, C., Dunlap, J.C., Loros, J.J., Kück, U. (2005). Cross-species microarray hybridization to identify developmentally regulated genes in the filamentous fungus *Sordaria macrospora*. *Molecular Genetics and Genomics*. **273**, 137-149.

Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., Kanehisa, M. (1999). KEGG: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*. **27**, 29-34.

Oliver, S. (1996). A network approach to the systematic analysis of yeast gene function. *Trends in Genetics*. **12**, 241-242.

Ouzounis, C. A., and Valencia, A. (2003). Early bioinformatics: the birth of a discipline- a personal view. *Bioinformatics*. **19**, 2176-2190.

Roux, J., Van Wyk, M., Hatting, H., and Wingfield, M. J. (2004). *Ceratocystis* species infecting stem wounds on *Eucalyptus grandis* in South Africa. *Plant Pathology* **53,** 414-421.

Schena, M., Shalon, D., Davis, R.W., Brown, P.O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science*. **270**, 467-470.

Shalon, D., Smith, S.J., Brown, P.O. (1996). A DNA microarray system for analyzing complex DNA samples using two-color fluorescent probe hybridization. *Genome Research*. **6**, 639-645.

Sharpton, T. J., Stajich, J. E., Rounsley, S. D., Gardner, M. J., Wortman, J. R., Jordar, V. S., Maiti, R., Kodira, C. D., Neafsey, D. E., Zeng, Q., Hung, C.-Y., McMahan, C., Muszewska, A., Grynberg, M., Mandel, M. A., Kellner, E. M., Barker, B. M., Galgiani, J. N., Orbach, M. J., Kirkland, T. N., Cole, G. T., Henn, M. R., Birren, B. W., and Taylor, J. W. (2009). Comparative genomic analyses of the human fungal pathogens *Coccidioides* and their relatives. *Genome Research* **19,** 1722-1731.

Storz, J. F., and Wheat, C. W. (2010). Integrating evolutionary and functional approaches to infer adaptation at specific loci. *Evolution* **64,** 2489-2509.

Tarigan, M., Van Wyk, M., Roux, J., Tjahjono, B., and Wingfield, M. J. Three new *Ceratocystis* spp. in the *Ceratocystis moniliformis* complex from wounds on *Acacia mangium* and *A. crassicarpa*. *Mycoscience* **51,** 53-67.

Tettelin, H., Riley, D., Cattuto, C., Medini, D. (2008). Comparative genomics: the bacterial pan-genome. *Current Opinion in Microbiology*. **11**, 472-477.

Turner, T.L., Bourne, E.C., Von Wettberg, E.J., Hu, T.T., Nuzhdin, S.V. (2010). Population resequencing reveals local adaptation of *Arabidopsis lyrata* to serpentine soils. *Nature Genetics*. **42**, 260-263.

Uetz, P., Giot, L., Cagney, G., Mansfield, T. A., Judson, R. S., Knight, J. R., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. M. (2000). A comprehensive analysis of protein-protein interactions in *Saccharomyces cerevisiae*. *Nature* **403,** 623-627.

van Wyk, M., Roux, J., Barnes, I., Wingfield, B. D., Chhetri, D. B., Kirisits, T., and Wingfield, M. J. (2004). *Ceratocystis bhutanensis* sp. nov., associated with the bark beetle *Ips schmutzenhoferi* on *Picea spinulosa* in Bhutan. *Studies in Mycology* **50,** 365-379.

van Wyk, M., Al Adawi, A. O., Khan, I. A., Deadman, M. L., Al Jahwari, A. A., Wingfield, B. D., Ploetz, R., and Wingfield, M. J. (2007). *Ceratocystis manginecans* sp. nov., causal agent of a destructive mango wilt disease in Oman and Pakistan. *Fungal Diversity* **27,** 213-230.

Vinces, M.D., Legendre, M., Caldara, M., Hagihara, M., Verstrepen, K.J. (2009). Unstable tandem repeats in promoters confer transcriptional evolvability. *Science*. **324**, 1213-1216.

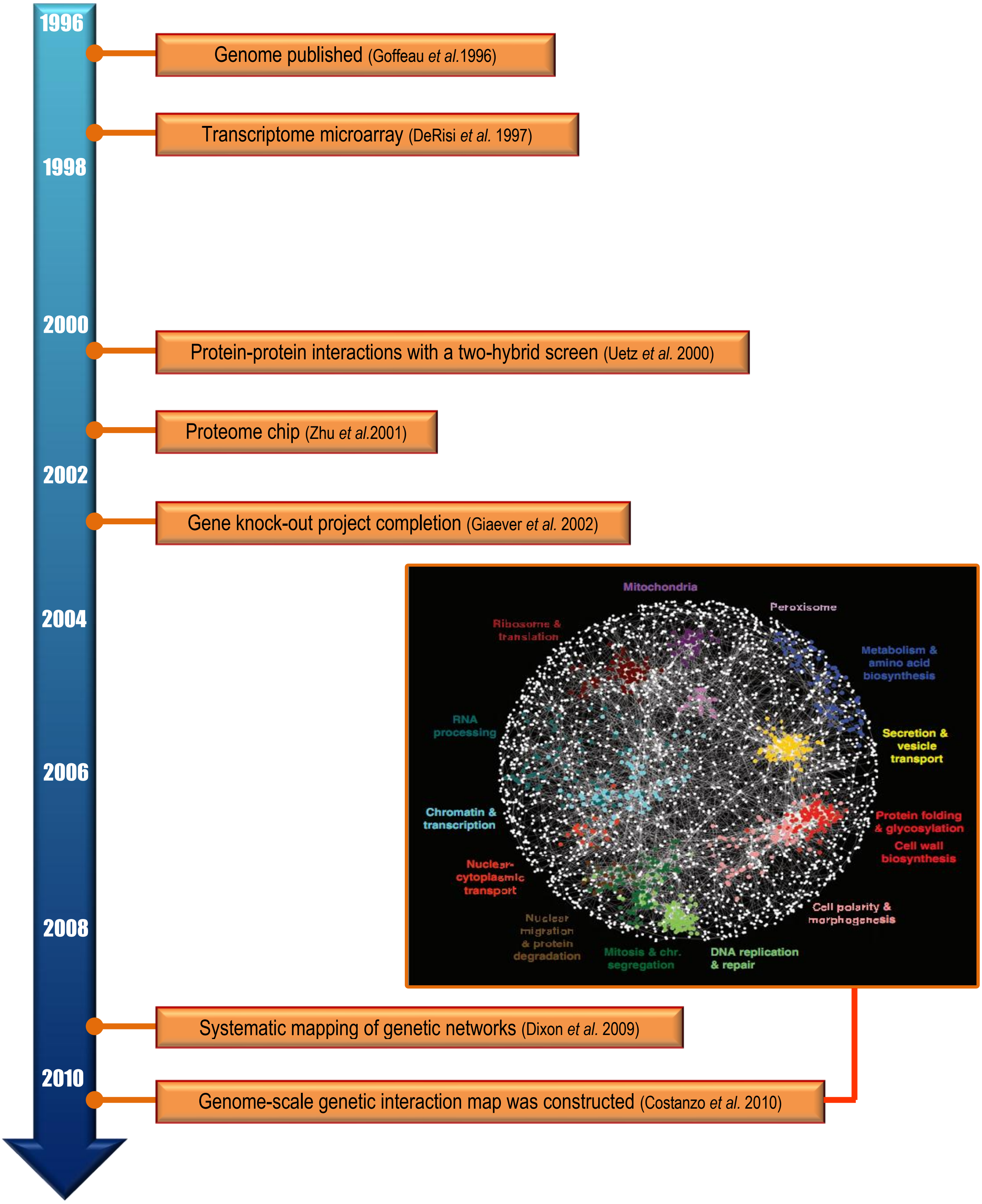
Witthuhn, R. C., Harrington, T. C., Wingfield, B. D., Steimel, J. P., and Wingfield, M. J. (2000). Deletion of the *MAT-2* mating-type gene during uni-directional mating-type switching in *Ceratocystis*. *Current Genetics* **38,** 48-52.

Wood, V., Gwilliam, R., Rajandream, M. A., Lyne, M., Lyne, R., Stewart, A., Sgouros, J., Peat, N., Hayles, J., Baker, S., Basham, D., Bowman, S., Brooks, K., Brown, D., Brown, S., Chillingworth, T., Churcher, C., Collins, M., Connor, R., Cronin, A., Davis, P., Feltwell, T., Fraser, A., Gentles, S., Goble, A., Hamlin, N., Harris, D., Hidalgo, J., Hodgson, G., Holroyd, S., Hornsby, T., Howarth, S., Huckle, E. J., Hunt, S., Jagels, K., James, K., Jones, L., Jones, M., Leather, S., McDonald, S., McLean, J., Mooney, P., Moule, S., Mungall, K., Murphy, L., Niblett, D., Odell, C., Oliver, K., O'Neil, S., Pearson, D., Quail, M. A., Rabbinowitsch, E., Rutherford, K., Rutter, S., Saunders, D., Seeger, K., Sharp, S., Skelton, J., Simmonds, M., Squares, R., Squares, S., Stevens, K., Taylor, K., Taylor, R. G., Tivey, A., Walsh, S., Warren, T., Whitehead, S., Woodward, J., Volckaert, G., Aert, R., Robben, J., Grymonprez, B., Weltjens, I., Vanstreels, E., Rieger, M., Schäfer, M., Müller-Auer, S., Gabel, C., Fuchs, M., Fritzc, C., Holzer, E., Moestl, D., Hilbert, H., Borzym, K., Langer, I., Beck, A., Lehrach, H., Reinhardt, R., Pohl, T. M., Eger, P., Zimmermann, W., Wedler, H., Wambutt, R., Purnelle, B., Goffeau, A., Cadieu, E., Dréano, S., Gloux, S., Lelaure, V., Mottier, S., Galibert, F., Aves, S. J., Xiang, Z., Hunt, C., Moore, K., Hurst, S. M., Lucas, M., Rochet, M., Gaillardin, C., Tallada, V. A., Garzon, A., Thode, G., Daga, R. R., Cruzado, L., Jimenez, J., Sánchez, M., del Rey, F., Benito, J., Domínguez, A., Revuelta, J. L., Moreno, S., Armstrong, J., Forsburg, S. L., Cerutti, L., Lowe, T., McCombie, W. R., Paulsen, I., Potashkin, J., Shpakovski, G. V., Ussery, D., Barrell, B. G., and Nurse, P. (2002). The genome sequence of *Schizosaccharomyces pombe*. *Nature* **415,** 871-880.

Yu, Z., Ge, Y., Xie, L., Zhang, T., Huang, L., Zhao, X., Liu, J., and Huang, G. (2014). Using a yeast two-hybrid system to identify FTCD as a new regulator for HIF-1α in HepG2 cells. *Cellular Signalling* **26,** 1560-1566.

Zhu, H., Bilgin, M., Bangham, R., Hall, D., Casamayor, A., Bertone, P., Lan, N., Jansen, R., Bidlingmaier, S., Houfek, T., Mitchell, T., Miller, P., Dean, R. A., Gerstein, M., and Snyder, M. (2001). Global analysis of protein activities using proteome chips. *Science* **293,** 2101-2105.

Figure 1. A timeline of major research accomplishments in *Saccharomyces cerevisiae* since the genome publication (Adjusted from Hofmann *et al.* 2003). The Genome-scale genetic interaction map connects genes with similar genetic interaction profiles (Taken from Costanzo *et al.* 2010).



Genome published (Goffeau *et al.*1996)

Transcriptome microarray (DeRisi *et al.* 1997)

Protein-protein interactions with a two-hybrid screen (Uetz *et al.* 2000)

Proteome chip (Zhu *et al.*2001)

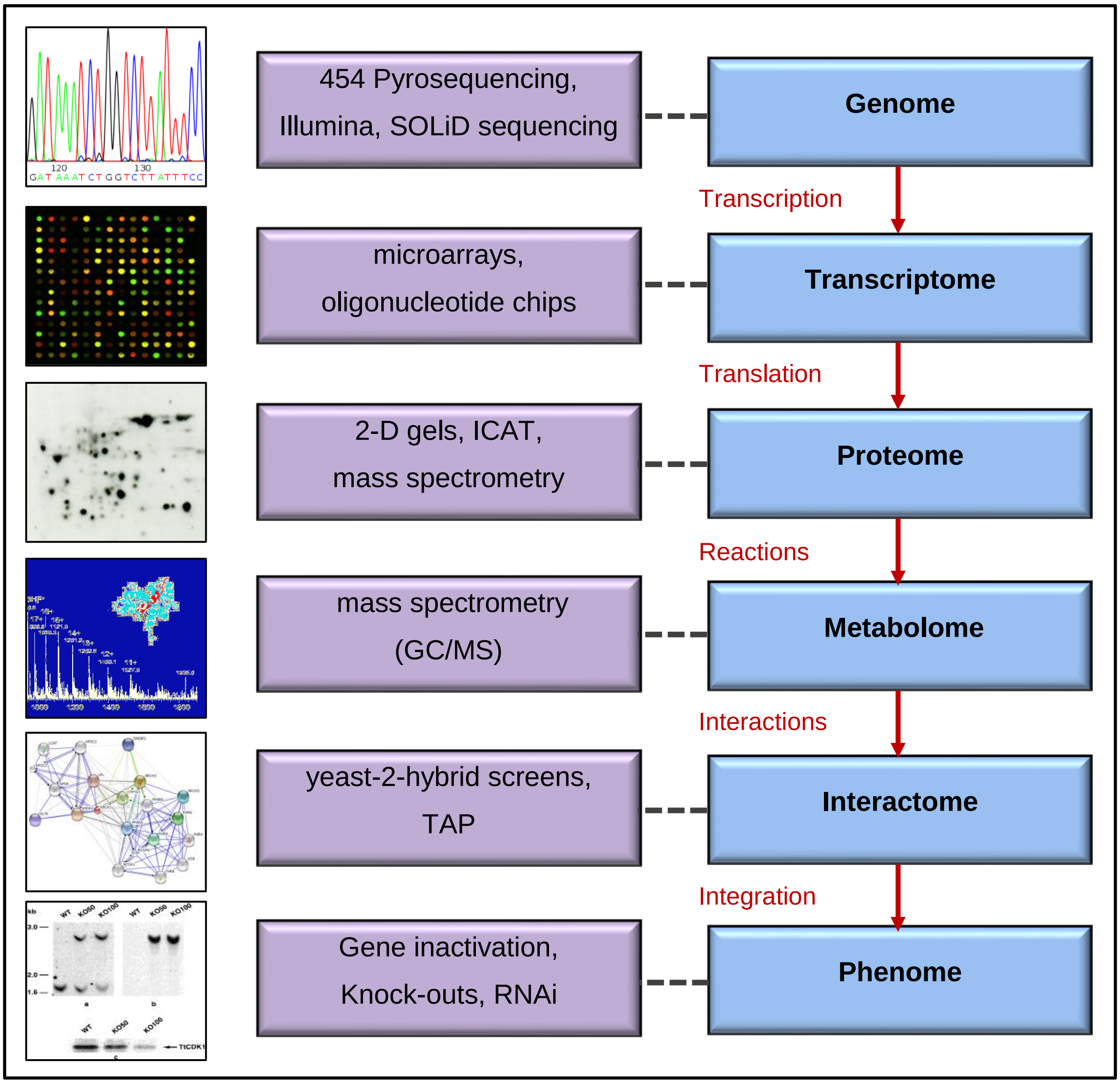
Gene knock-out project completion (Giaever *et al.* 2002)

Systematic mapping of genetic networks (Dixon *et al.* 2009)

Genome-scale genetic interaction map was constructed (Costanzo *et al.* 2010)

1996
1998
2000
2002
2004
2006
2008
2010
Mitochondria
Peroxisome
Ribosome & translation
Metabolism & amino acid biosynthesis
RNA processing
Secretion & vesicle transport
Chromatin & transcription
Protein folding & glycosylation
Cell wall biosynthesis
Nuclear-cytoplasmic transport
Cell polarity & morphogenesis
Nuclear migration & protein degradation
Mitosis & chr. segregation
DNA replication & repair

Figure 2. A schematic representation of the experimental data acquisition process in biological sciences. The basic molecular levels of a cell and the biological processes link these levels are depicted in the blue boxes. The grey boxes indicate examples of ‘‘-omics’’ technologies that can study the respective biological levels quantitatively (adapted from Fischer 2005).



454 Pyrosequencing, Illumina, SOLiD sequencing — Genome

Transcription

microarrays, oligonucleotide chips — Transcriptome

Translation

2-D gels, ICAT, mass spectrometry — Proteome

Reactions

mass spectrometry (GC/MS) — Metabolome

Interactions

yeast-2-hybrid screens, TAP — Interactome

Integration

Gene inactivation, Knock-outs, RNAi — Phenome

CHAPTER 2

Novel organisation of the Mating type locus in *Ceratocystis moniliformis*

ABSTRACT

Sexual reproduction in Ascomycetes is controlled by the mating type (*MAT*) locus which includes the *MAT1-1* and *MAT1-2* idiomorphs. Individuals with a heterothallic mating system have a single *MAT* idiomorph and sexually reproduce with individuals having the opposite idiomorph, while homothallic species carry both versions of the idiomorphs in their genome and can complete the sexual cycle without an opposite mating partner. *Ceratocystis moniliformis* (Ascomycete) is a common non-pathogenic wound-infecting fungus found on trees in the tropics and sub-tropics. The fungus produces sexual structures without out-crossing, indicating a homothallic mating strategy and thus, single strains of the fungus should contain both the *MAT1-1* and *MAT 1-2* idiomorphs. The genome of *C. moniliformis* was sequenced using a combination of 454 pyrosequencing and Illumina technology. The genome was assembled and mined for mating type genes and 28 other genes known to be involved in the mating process. Interestingly, results showed that the *MAT1-2-1* gene was the only *MAT* gene present in the genome. These results were unexpected, as isolates from a homothallic species would typically contain both *MAT* idiomorphs. Furthermore, analyses showed the pheromone precursors were also absent. A homothallic species that harbours only one mating type idiomorph has been identified in *Neurospora*, but this is the first example of this phenomenon where only the *MAT1-2-1* gene is present. Thus we suggest a novel mechanism of sexual reproduction in *C. moniliformis.*

INTRODUCTION

Regulation of sexual reproduction is one of the key processes in the life cycle of fungi (Coppin *et al.* 1997; Kronstad and Staben 1997). The genes regulating these functions are situated at the mating type (*MAT*) locus that shares a highly conserved organisation and structure in most filamentous Ascomycetes. The MAT proteins function as transcription factors that regulate the process of mating by controlling sexual development and coordinating the expression of downstream, mating type specific genes (Coppin *et al.* 1997; Shiu and Glass 2000; Jones and Bennett 2011). The pheromone signalling and response pathway, meiotic and recombination processes and pathways controlling fruiting body development are also crucial parts of sexual reproduction (Galagan *et al.* 2005). The cloning of the mating locus in *Saccharomyces cerevisiae* (Astell *et al.* 1981) represented an important step towards understanding mating in fungi. The first filamentous Ascomycete for which the mating type loci were cloned was *Neurospora crassa* (Glass 1988), now considered a model for the study of *MAT* genes.

Mating type genes are located at a single locus (*MAT1*) in the Ascomycete fungal genome (Coppin *et al.* 1997; Turgeon and Yoder 2000). In many fungal species, this locus is flanked by the DNA lyase (*APN*) and *SLA2* genes (Debuchy and Turgeon 2006; Waalwijk *et al.* 2004; Butler *et al.* 2004). However, some species have only one of these genes flanking the *MAT* locus (Martin *et al.* 2010). Although a role for *APN* and *SLA2* in reproduction has not been found, their position flanking the *MAT* idiomorph could prove useful in identifying MAT information.

Most fungi have one of two mating systems; either homothallism or heterothallism. A single homothallic fungal isolate can produce viable ascospores without fusing with a mate, and this is achieved by harbouring all the genes for reproduction at a single locus (Nelson 1996; Coppin *et al.* 1997; Pöggeler 2001). Heterothallic species require an opposite mating partner in order to give rise to sexual progeny. In these fungi, each partner has one version of the mating type alleles, referred to as *MAT1-1* and *MAT1-2* (Nelson 1996; Pöggeler 2001). The alleles are termed idiomorphs because they are unrelated in sequence or structure, yet occupy the same locus on homologous chromosomes (Metzenberg and Glass 1990). While the genetic content of the idiomorphs can differ between species, two key genes are typically always present. These are the *MAT1-1-1* gene in the *MAT1-1* idiomorph, and the *MAT1-2-1* present at the *MAT1-2* idiomorph.

In heterothallic fungi, opposite isolates are distinguished by the presence of distinct genes. *MAT1-2* isolates contain the hallmark *MAT1-2-1* gene, with an HMG (High motility group) DNA binding domain in the protein (Turgeon and Yoder 2000; Coppin *et al*. 1997). This gene is generally the only one present in the idiomorph, although a number of other genes such as the *MAT1-2-2* (Kanamori *et al.* 2007)*, MAT1-2-3* (Martin *et al.* 2011) and *MAT1-2-4* (Mandel *et al.* 2007) have been found, but they appear to be restricted to certain fungal genera. *MAT1-1* isolates are characterised by the presence of the *MAT1-1-1* gene (Turgeon and Yoder 2000). This protein contains an α-box DNA binding motif and is homologous to *MATα1* of *Saccharomyces cerevisiae.* Other genes which may also be found at this idiomorph are the *MAT1-1-2* and *MAT1-1-3* genes. These encode proteins with an alpha-like domain and an HMG box, respectively. A number of homothallic *Neurospora* species have, however been reported to contain only Type A genes (homologous to *MAT1-1-1)* (Glass *et al.* 1988; Glass and Smith 1994). To the best of our knowledge, there are no examples of a homothallic fungus containing only the *MAT1-2* genes.

Ceratocystis spp*.* are Ascomycetes for which the mating systems have not been intensively studied but species with both heterothallic and homothallic mating are known. Those with homothallic mating such as the well-known group of plant pathogens, *C. fimbriata* and its cryptic relatives (Baker *et al.* 2003) are particularly interesting as they undergo uni-directional mating type switching (Harrington and McNew 1998; Witthuhn *et al.* 2000). In this case, individuals generated from single ascospores can give rise to both self-fertile and self-sterile progeny. The self-fertile individuals contain both MAT1-1 and MAT1-2 mating type sequences. The self-sterile individuals are unable to revert back to homothallism (Webster and Butler 1967; Harrington and McNew 1997) and there is evidence that the *MAT1-2-1* gene is no longer present in these isolates (Witthuhn *et al.* 2000; Wilken *et al.* 2014). Not all *Ceratocystis* species undergo mating type switching and species such as *C. moniliformis* Hedgcock and *C. omanensis* Al-Subhi, M. J. Wingf., M. Van Wyk & Deadman are strictly homothallic (Harrington 2007; Al-Subhi *et al.* 2006) while *C. fagacearum* (Bretz) Hunt and *C. eucalypti* Z. Q. Yuan & Kile are heterothallic (Harrington 2007; Juzwik *et al.* 2008, Kile *et al.* 1996).

Ceratocystis moniliformis sensu stricto (s.s.) is a non-pathogenic species and it defines a phylogenetically related clade of species that should logically reside in a discrete genus (Wingfield *et al.* 2013). Although the mating system of *C. moniliformis* and at least one of its close relatives, *C. omanensis* are known to be defined by homothallism (Al-Subhi *et al.* 2006), it is not known whether other species in the *C. moniliformis* complex all share this system. The objectives of this study were to identify the *MAT* genes in the genome of an isolate of *C. moniliformis* and to elucidate the structure and organisation of the *MAT* idiomorph and its flanking regions. These results were also compared with those for other homothallic fungi. An ancillary objective was to investigate the presence of mating related genes in addition to the *MAT* genes.

MATERIALS AND METHODS

Strains and DNA isolation

A single isolate of *C. moniliformis* s.s. (CMW 10134) collected from wounds on *Eucalyptus* in Mpumulanga, South Africa in 2002 was selected for this study. This isolate was chosen because it produces sexual structures. Hyphal tips were isolated from single germinating ascospores to generate cultures that represented only a single genotype was present. The Internal transcribed spacer regions (ITS) of the ribosomal DNA operon were sequenced to confirm the identity of the isolate. For this verification step, the primer set ITS1F and ITS4 (Gardes and Bruns 1993) was used.

To isolate DNA, single spore cultures were grown on 2% Malt extract agar (MEA: 20g malt extract [Biolab, Merck], 20g agar [Biolab, Merck], 1L dH2O) for 6-8 days at 25°C. Mycelium was scraped from the surface of cultures and the DNA extracted using a phenol chloroform extraction method as described by Barnes *et al.* (2001). The isolate used in this study is maintained in the culture collection (CMW) of the Forestry and Agricultural Biotechnology Institute, University of Pretoria.

DNA sequencing and assembly

Genome sequencing was performed by Inqaba Biotech (Pretoria, South Africa) on a Roche GS-FLX sequencer with the Titanium plate upgrade. The resulting sequences were assembled using the Newbler software package into 649 contigs with an average of 20x coverage. The contigs were divided into 12 files and subjected to the online version of the MAKER genome annotation pipeline (Cantarel *et al.* 2008; http://www.yandell-lab.org/software/mwas.html). Another round of sequencing was undertaken using the same fungal isolate on the Genome Analyzer IIx platform (Illumina) at the University of North Carolina, Chapel Hill High-Throughput Sequencing Facility. The 39.5 million reads were assembled into 778 contigs using the CLC Genomics workbench 5.5.1 (CLC Bio, Aarhus, Denmark). The N50 value for the assembly was 153 651. The CEGMA (Core Eukaryotic Genes Mapping Approach) pipeline (Parra *et al.* 2007) was used to measure the completeness of the assembly based on the number of a set of core eukaryotic genes present in the genome as complete or partial orthologs.

Gene discovery and BLAST analyses

The respective *MAT* gene sequences from *C. fimbriata* (Wilken *et al.* 2014) as well as from other Ascomycetes, *Gibberella zeae* and *Fusarium oxysporum* (Table S1) were used in local tBLASTn and BLASTn analyses on CLC Main Workbench 5.5 in attempt to detect their homologs in *C. moniliformis*. Positive hits were designated as those with an *E* value lower than 6e-08 (Martin *et al.* 2010). Sequences that produced these hits were subjected to protein predictions using the online tools AUGUSTUS (Stanke *et al.* 2004) and FGENESH (Salamov and Solovyev 2000). BLAST (Altschul *et al.* 1990) searches were done against the NCBI/GenBank protein databases to confirm the identity of the predicted genes.

Mating type gene and protein comparisons

MAT1-2-1 protein sequences from other Ascomycete fungi were downloaded from the NCBI database for comparison with the predicted *C. moniliformis* protein (Table 1). These sequences were aligned using the ClustalX2 alignment program (Larkin *et al.* 2007) and imported into CLC Main Workbench v. 5.5. Gene and protein comparisons of the *MAT1-2-1* gene were performed to detect gaps, percentage identity, distance and differences.

Comparisons of the *MAT* idiomorphs

Detailed idiomorph structures were obtained for homothallic species, *C. fimbriata* (Wilken *et al.* 2014); *F. graminearum* (Yun *et al.* 2000; Waalwijk *et al.* 2004) which was downloaded from the website of the Broad Institute (http://www.broadinstitute.org/); *Sclerotinia sclerotiorum* (Amselem *et al.* 2011); and *Cochliobolus luttrellii* (Yun *et al.* 1999). These idiomorphs were used for organisational and structural comparisons with the *C. moniliformis* *MAT* locus.

Identification of other mating related genes

Protein sequences for other genes implicated in the mating process were obtained from the NCBI database (Table S1). These genes were identified in the genome of *C. moniliformis* using the same bioinformatic approach as that described for *MAT* gene discovery.

RESULTS

Strain of *C. moniliformis*

Cultures of *C. moniliformis* that grew from single hyphal tip isolations produced perithecia and viable ascospores within 10-15 days of growth. As is common in many Ascomycetes, the cultures lost the ability to sporulate over time and after successive rounds of sub-culturing (Hanlin 1985). While the transfer of ascospores is suggested to prevent this (Hanlin 1985), *C. moniliformis* continued to grow as white cultures lacking perithecia and spores after this method of sub-culturing. Interestingly, cultures which had ceased sporulating, were maintained at 25°C and had undergone periodic sub-culturing nine times, spontaneously resumed the production of perithecia and ascospores. This culture was used for DNA extraction. The sequenced ITS region verified the identity of the strain as *C. moniliformis* strain 10134 (Accession number FJ151422.1).

Genome sequencing and assembly

The estimated genome size in the sequenced isolate of *C. moniliformis* for both the Newbler (454 pyrosequencing data) and CLC Genomics (Illumina data) assemblies were 25.2 Mb. This was less than the expected size based on other Ascomycete fungi such as *C. fimbriata*, 29.4 Mb (Wilken *et al.* 2013); *C. manginecans*, 31.7 Mb (van der Nest *et al.* 2014); *Fusarium graminearum,* 36.1 Mb (Cuomo *et al.* 2007); *Aspergillus oryzae,* 36.7 Mb (Payne *et al.* 2006); and *Grosmannia clavigera,* 29.8 Mb (DiGuistini *et al.* 2011). The CEGMA pipeline showed that 92.7% and 95.6% of conserved eukaryotic genes (CEGs) were present as complete or partial transcripts, respectively. An average of 1.10 – 1.13 genes per CEG was detected. Together, this suggests that the genome assembly had a high coverage.

Gene discovery

A BLASTn using the *C. fimbriata MAT1-2-1* gene sequence failed to identify any significant hits in the *C. moniliformis* genome. A tBLASTn using the protein sequence produced a significant hit and the *C. moniliformis MAT1-2-1* gene was then identified and described by subjecting the hit sequence to AUGUSTUS protein predictions and verified by BLAST match against NCBI non-redundant (NR) proteins. The gene was 938 bp in length, with two introns (60, 53 bp) and is predicted to encode a 274 aa protein with a characteristic HMG motif. The protein had homology to *Metarhizium acridum* MAT1-2-1 (Gao *et al.* 2011) with 39% identity and 74% coverage, *Ceratocystis eucalypti* MAT2-1, partial protein (Witthuhn *et al.* 2000) with 56% identity and 39% coverage and *Ophiocordyceps sinensis* MAT1-2-1 (Zhang *et al.* 2009) with 37% identity and 69% coverage.

MAT1-2-1 gene comparisons in *Ceratocystis*

The *MAT1-2-1* gene of *C. moniliformis* was substantially smaller than that of *C. fimbriata* (Table 2). A pairwise comparison of the genes revealed 31% identity, 948 nucleotide differences, 552 gaps and 429 identical bases where the sequence alignment overlapped. A protein pairwise comparison showed 19.52% identity, 400 differences, 262 gaps and 97 identical amino acids where the sequence alignment overlapped. The putative protein product of the *MAT1-2-1* open reading frame (ORF) exhibited similarity to MAT1-2 proteins of a number of Ascomycetes listed in Table 1 (Figure 1), however a notable degree of sequence divergence is evident between species (Figure 1). Alignment of the *MAT1-2-1* genes of *C. moniliformis, C. omanensis, C adiposa* and *C. bhutanensis* showed a high degree of sequence similarity, these species all belong to the *C. moniliformis* species complex (Figure 2).

Structural organisation of the *MAT* idiomorph

To determine whether any other mating related genes were present around the *MAT1-2-1* gene, the full contig of 118 035 bp in length was subjected to AUGUSTUS analysis. An additional 27 genes were thus identified. The SLA2 gene was found on the 5' end of the *MAT1-2-1* gene, typically found flanking *MAT* genes, followed by an *APC* gene (Martin *et al.* 2010) (Figure 3A) which was also present in *C. fimbriata* (Wilken *et al.* 2014) (Figure 4). There was a gene with no homology or known motifs followed by a gene which encodes a hypothetical protein (Figure 3A) on the 3' end of the *MAT1-2-1* gene. The remaining 23 genes on the contig were not found to be associated with other described mating type idiomorphs.

Genome wide searches failed to identify any additional mating specific genes. But a BLAST search located the *DNA lyase* gene on a separate contig to the *MAT1-2-1* gene and was adjacent to a *cytochrome C oxidase* gene (Figure 3B), but no *MAT* genes are present on this contig.

Comparisons of the Homothallic *MAT* idiomorph structure

The *MAT* idiomorph of *C. moniliformis* contained one mating type gene, *MAT1-2-1* and was flanked by an *SLA2* gene on its 5' end with 4.5 kb of sequence between them. No other mating type genes were present on the contig (Figure 4). We compared this organisation and structure with those in other known homothallic fungi including *C. fimbriata*, *F. graminearum, S. sclerotiorum* and *Cochliobolus luttrellii*. The composition and structure of *F. graminearum* and *S. sclerotiorum* adhere to the accepted “model structure” of homothallic species, having the two characteristic genes, *MAT1-1-1* and *MAT1-2-1* present at the idiomorph along with the flanking genes, *DNA Lyase* and *SLA2* (Figure 4). *Ceratocystis fimbriata* also harbours the expected *MAT1-1-1* and *MAT1-2-1* genes, and while the flanking genes are present at the idiomorph, they do not flank the whole set of *MAT* genes. *Cochliobolus luttrellii* exhibited a fusion of the *MAT1-1-1* and *MAT1-2-1* genes.

Presence of other mating related genes

BLAST analysis using the *MAT* genes *MAT1-1-2* and *MAT1-1-3* from both *C. fimbriata* and *G. zeae,* failed to identify any homologs in *C. moniliformis.* Of the 26 proteins involved in the pheromone signalling pathway, 22 were present (Table S1). No homologs were found for the pheromone precursor genes, MFα1, MFα2 (*ppgA*) and MFa1, MFa2 (*ppgB*). A search for four genes involved in fruiting body development in Ascomycetes yielded versions of these genes in the genome of *C. moniliformis* (Table S1).

DISCUSSION

Genome sequencing of *C. moniliformis* and the availability of genomic data for other fungal genomes made it possible to characterise the *MAT* idiomorph in this fungus, and compare this with those in other Ascomycetes. The fact that cultures generated from single sexual spores (ascospores) gave rise to cultures with sexual structures having viable spores confirmed this fungus has a homothallic mating system. It was thus expected that the *C. moniliformis MAT* idiomorph would contain at the very least, the two characteristic genes of opposite mating type, *MAT1-1-1* and *MAT1-2-1,* similar to that of other homothallic species (Nelson 1996; Pöggeler 2001). However, on a genetic level the *C. moniliformis MAT* locus is atypical in that it only has the *MAT1-2-1* gene present, and no *MAT1-1-1* gene.

Because homothallic fungi are known to contain both mating type idiomorphs, the absence of the *MAT1-1-1* gene raises the question whether the analyses in this study might have failed to identify the MAT1-1 associated gene. This is unlikely because the *MAT1-1-1* sequence that failed to identify the ortholog in *C. moniliformis* was successful in the identification of the ortholog in other species of the *C. moniliformis* species complex [data not shown]. A PCR could be done in future studies to verify the absence of this gene. The CGEMA analysis showed that the genome sequence contains over 90% of the core genes and thus has a high degree of completion. In addition, the *MAT1-2-1* gene was identified in the expected position adjacent to the *SLA* gene, similar to *C. fimbriata*. No ORFs with homology to *MAT1-1-1, MAT1-1-2* or *MAT1-1-3* genes were identified in the vicinity of the *MAT1-2-1* gene in *C. moniliformis*. To the best of our knowledge, this has never been observed in any other homothallic species. The only atypical mating type structure of a homothallic species has been reported in the genus, *Neurospora* where only MAT1-1 sequence and no MAT1-2 sequence was present (Glass *et al.* 1990)*.*

We considered the *MAT* idiomorph structures of other homothallic species to determine whether a similar scenario exists within the Ascomycetes. The absence of the *MAT1-1-1* gene in *C. moniliformis* (in the class Sordariomycetes of the Pezizomycotina sub-phylum) is unique to the group of characterised homothallic Ascomycetes. The *MAT* idiomorph of *F. graminearum* (Hypocreales; Yun *et al.* 2000) corresponded with the general trend in homothallic idiomorph structure having genes homologous to both heterothallic mating types, *MAT1-1-1* and *MAT1-2-1* (*Glass et al*. 1990; Glass and Smith 1994; Pöggeler *et al*. 1997; Yun *et al*. 1999) (Figure 4). *S. sclerotiorum* (order Helotiales in the class Leotiomycetes of the Pezizomycotina sub-phylum) belongs to the class most closely related to Sordariomycetes (Spatafora *et al.* 2006) and is also similar to most homothallic fungi characterized to date (Amselem *et al.* 2011).

We investigated a number of phenomena from literature that could be responsible for the abnormal mating type structure in *C. moniliformis.* Evidence which may lend credence to a single *MAT* gene being sufficient to confer self-fertility in a fungus was found in the *Neurospora* species. The heterothallic *N. crassa* represents the model *MAT* locus and many species studied to date conform to this model (Arie *et al.* 1999). Opposite mating type isolates are designated A and a, which are homologous to MAT1-1 and MAT1-2 respectively. Hybridization assays using *MATA* and *MATa* probes identified one typical homothallic species which had both *MATA* and *MATa* sequence (*N. terricola*), and four atypical homothallic species with only *MATA* gene information (Glass *et al.* 1990). These species were *N. africana, N. dodgei, N. galapagosensis* and *N. lineolata*. Although unlike *C. moniliformis,* these species lack the MAT1-2 sequence, this could illustrate a similar mating mechanism in that a homothallic species lacks one of the two mating type idiomorphs in the genome.

Sequence analysis showed no indication of a fusion of the opposite mating type genes in the genome of *C. moniliformis.* This scenario has been observed in a related order, Pleosporales in *Cochliobolus luttrellii* (Yun *et al.*1999) (Figure 4). A study on *C. luttrelli* suggested that MAT alone is sufficient to change the reproductive life style in *Cochliobolus* (Yun *et al.*1999) by the expression of a fused *MAT* gene from a homothallic species in a *MAT*-deletion strain of heterothallic *C. heterostrophus* (Yun *et al.* 1999; reviewed by Pöggeler 2001). The fused genes gave rise to homothallism. A fusion of genes however, is not responsible for the homothallic nature of *C. moniliformis.*

The *MAT* idiomorph of *C. moniliformis* is highly dissimilar to those of other *Ceratocystis* species. The MAT1-2-1 protein showed only 19.52% similarity to that of *C. fimbriata.* This observation supports the argument that there is extensive divergence in the *MAT1-2-1* gene sequence between related species complexes (Turgeon 1998). The *MAT* idiomorph in *Fusarium sacchari* is highly dissimilar from others in the complex (Martin *et al.* 2011) and this species represents an unusual example of homothallic reproduction. Early studies suggested that this species might be cross-fertile, being able to mate with several unique *Fusarium* mating populations (Kuhlman, 1982). However, a study by Britz *et al*. (1999) demonstrated that the progeny produced in such crosses were uniparental, and not a result of mating. This result implies that *F. sacchari* is capable of homothallic reproduction but has a heterothallic composition with only MAT1-1 sequence. In this case *MAT1-1-1* has the ability to confer independent reproduction and it may be possible that *MAT1-2-1* is capable of the same function.

The close association of *DNA lyase* and *SLA2* to the *MAT* locus has been observed in many Ascomycetes, including *Microsporidia* and *Ascomycota* (Debuchy and Turgeon 2006, Butler *et al.* 2004) and this was tested in the present study. The *SLA2* gene was located adjacent to the *MAT1-2-1* gene in *C. moniliformis,* however the other common flank, *DNA lyase* (Butler *et al.* 2004)*,* was located on a separate contig. Although complete synteny is not seen between the *Microsporidia* and *Ascomycota* mating types, the presence of the orthologous *DNA lyase* gene in the vicinity of the mating type locus is noteworthy, supporting a common origin (Martin *et al.* 2010). In most cases these genes are immediately adjacent to the *MAT* genes such as in *F. graminearum* and *S. sclerotiorum* (Figure 4). There are however some exceptions to this. The putative *MAT1-2* mating type locus of *Encephalitozoon cuniculi (Microsporidia*) reveals the presence of a *DNA lyase* homolog 7 kb away from the *MAT* locus (Katinka *et al.* 2001; Martin *et al.* 2010). The separation of the *DNA lyase* gene from the *MAT1-2-1* gene in *C. moniliformis* lends support to the idea that the *MAT* idiomorph contains only the *MAT1-2* sequence.

In most homothallic species, the genes from both idiomorphs are closely linked or fused, being located on the same chromosome (Yun *et al*. 1999; Pöggeler, 1999). This association has been the basis of suggestions that a transcriptional interaction between these genes might take place (reviewed by Metzenberg and Glass, 1990). A possibility remains that for *C. moniliformis,* the *MAT1-1-1* gene is unlinked to *MAT1-2-1* and is sufficiently divergent and/or truncated to evade identification by bioinformatic analysis. *Cochliobolus cymbopogonis* provides an example of unlinked *MAT1-1-1* and *MAT1-2-1* genes in a homothallic species (Yun *et al*. 1999). Neither PCR nor gel blot analysis showed evidence for linkage within 30 kb between the *MAT1-1-1* and *MAT1-2-1* genes. Three other homothallic species of *Cochliobolus,* however, contain closely linked and even fused *MAT1-1-1* and *MAT1-2-1* genes (Yun *et al*. 1999). Evidence for a similarly “split” idiomorph in *C. moniliformis* could also be attributed to the fact that the adjacent genes, *DNA lyase* and *Cytochrome C oxidase* are not present on the same contig as *MAT1-2-1* and *SLA2* as would be expected (Figure 3).

The possibility exists that *MAT1-1* may have been dispensable in *C. moniliformis.* The question as to whether both opposite mating type proteins are essential for homothallic reproduction has produced conflicting answers. The *MAT* genes of the heterothallic *Neurospora* species are suggested to be under positive (diversifying) selection (Wik *et al.* 2008). In contrast the *MAT* genes of homothallic species appeared to have diverged under a lack of selective constraint, with some containing premature stop codons (Wik *et al.* 2008). This line of evidence suggests that *MAT* genes may, therefore, be dispensable in homothallic *Neurospora* spp. Knock-out studies using the homothallic *F. graminearum* provides information that contradicts this proposal. Strains in which either the *MAT1-1-1* or *MAT1-2-1* genes were knocked out allowed reproduction only in a heterothallic manner, therefore completely removing the homothallic behaviour (Lee *et al*. 2003). This shows that, at least in *Fusarium* the *MAT* genes are not dispensable (Martin *et al.* 2011). A gene knock-out study on a homothallic *Ceratocystis* species would provide insight regarding the dispensability of *MAT* genes in this genus.

The fact that *C. moniliformis* is apparently missing several key *MAT* genes resulted in our hypothesis that other genes known to be associated with mating might also not be present. In order to interrogate this question further, we investigated the presence of 24 genes involved in the pheromone signalling pathway and four genes involved in fruiting body development (Table S1). BLAST analyses identified the presence of all of these genes except for those that encode the pheromone precursors MFα1 and MFα2, as well as MFa1 and MFa2 (Table S1). It has been shown that pheromones are essential for successful mating in heterothallic species (Bender and Sprague-Jr 1989; Bobrowicz *et al*. 2002; Mayrhofer *et al*. 2006). However, in homothallic species that can reproduce sexually through self-fertilization, it is suggested that mating pheromones might be dispensable to varying degrees (Mayrhofer and Pöggeler 2005; Lee *et al*. 2008). *Ceratocystis moniliformis* is homothallic and it was thus not surprising to find that pheromones might be absent. Furthermore, mutant analysis in *Neurospora* suggests that the product of the *mtA* idiomorph, which is equivalent to *MAT1-1*, controls functions such as pheromone production, reception and/or transduction of the pheromone signal (Robertson *et al.* 1998). If this is the case, it may be possible that both the pheromone precursors and *MAT1-1* genes are redundant in homothallic species and have thus all been lost in *C. moniliformis*.

CONCLUSIONS

Attempts to understand *MAT* organization in diverse homothallic Ascomycetes is still in its infancy (discussed in Yun *et al*. 1999; Pöggeler, 1999). The rapid diversification of *MAT* genes hampers their identification in uncharacterised species even further. To the best of our knowledge, an idiomorph harbouring only the *MAT1-2-1* gene has never been reported from a homothallic fungus. The finding that *C. moniliformis* which is phenotypically homothallic but lacks the pheromone precursors along with the *MAT1-1* genes, suggests that this species undergoes reproduction in a manner not yet understood. Detailed sequence and structural analyses of the mating type loci in other homothallic relatives of *C.*

moniliformis should provide interesting insights into the evolution of sexual reproduction systems in these fungi and their relatives.

REFERENCES

Al-Subhi, A. M., Al-Adawi, A.O., Van Wyk, M., Deadman, M. L., Wingfield M. J. (2006). *Ceratocystis omanensis*, a new species from diseased mango trees in Oman. *Mycological Research*. **110**, 237-245.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology*. **215**, 403-410.

Amselem, J., Cuomo, C. A., van Kan, J. A. L., Viaud, M., Benito, E. P., Couloux, A., Coutinho, P. M., de Vries, R. P., Dyer, P. S., Fillinger, S., Fournier, E., Gout, L., Hahn, M., Kohn, L., Lapalu, N., Plummer, K. M., Pradier, J.-M., Quévillon, E., Sharon, A., Simon, A., ten Have, A., Tudzynski, B., Tudzynski, P., Wincker, P., Andrew, M., Anthouard, V. r., Beever, R. E., Beffa, R., Benoit, I., Bouzid, O., Brault, B., Chen, Z., Choquer, M., Collémare, J. r., Cotton, P., Danchin, E. G., Da Silva, C., Gautier, A. l., Giraud, C., Giraud, T., Gonzalez, C., Grossetete, S., Güldener, U., Henrissat, B., Howlett, B. J., Kodira, C., Kretschmer, M., Lappartient, A., Leroch, M., Levis, C., Mauceli, E., Neuvéglise, C. c., Oeser, B., Pearson, M., Poulain, J., Poussereau, N., Quesneville, H., Rascle, C., Schumacher, J., Ségurens, B. a., Sexton, A., Silva, E., Sirven, C., Soanes, D. M., Talbot, N. J., Templeton, M., Yandava, C., Yarden, O., Zeng, Q., Rollins, J. A., Lebrun, M.-H., and Dickman, M. Genomic analysis of the necrotrophic fungal pathogens *Sclerotinia sclerotiorum* and *Botrytis cinerea*. *PLoS Genet* **7,** e1002230.

Arie, T., Yoshida, T., Shimizu, T., Kawabe, M., Yoneyama, K., Yamaguchi, I. (1999). Assessment of *Gibberella fujikuroi* mating type by PCR. *Mycoscience*. **40**, 311-314.

Astell, C. R., Ahlstrom-Jonasson, L., Smith, M., Tatchell, K., Nasmyth, K. A., Hall, B. D. (1981). The Sequence of the DNAs coding for the mating-type loci of *Saccharomyces cerevisiae*. *Cell*, **27**, 15-23.

Baker, C. J., Harrington, T. C., Krauss, U., and Alfenas, A. C. (2003). Genetic variability and host specialization in the Latin American clade of *Ceratocystis fimbriata*. *Phytopathology* **93,** 1274-1284.

Barnes, I., Roux, J., Coetzee, M. P. A., Wingfield, M. J. (2001). Characterization of *Seiridium* spp. associated with cypress canker based on β-tubulin and histone sequences. Plant Disease **85**: 317-321.

Bender, A., Sprague, G.F. (1989). Pheromones and pheromone receptors are the primary determinants of mating specificity in the yeast *Saccharomyces cerevisiae*. *Genetics*. **121**, 463-476.

Bobrowicz, P., Pawlak, R., Correa, A., Bell-Pedersen, D., Ebbole, D.J. (2002). The *Neurospora crassa* pheromone precursor genes are regulated by the mating type locus and the circadian clock. *Molecular Microbiology*. **45**, 795-804.

Britz, H., Coutinho, T.A., Wingfield, M.J., Marasas, W.F.O., Gordon, T.R., Leslie, J.F. (1999). *Fusarium subglutinans* f. sp.*pini* represents a distinct mating population in the *Gibberella fujikuroi* species complex. *Applied and Environmental Microbiology*. **65**, 1198-1201.

Butler, G., Kenny, C., Fagan, A.s., Kurischko, C., Gaillardin, C., Wolfe, K.H. (2004). Evolution of the MAT locus and its Ho endonuclease in yeast species. *Proceedings of the National Academy of Sciences of the United States of America*. **101**, 1632-1637.

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

Coppin, E., Debuchy, R., Arnaise, S., Picard, M. (1997). Mating types and sexual development in filamentous ascomycetes. *Microbiology and Molecular Biology Reviews*, **61**, 411-428.

Cuomo, C. A., Güldener, U., Xu, J.-R., Trail, F., Turgeon, B. G., Di Pietro, A., Walton, J. D., Ma, L.-J., Baker, S. E., Rep, M., Adam, G., Antoniw, J., Baldwin, T., Calvo, S., Chang, Y.-L., DeCaprio, D., Gale, L. R., Gnerre, S., Goswami, R. S., Hammond-Kosack, K., Harris, L. J., Hilburn, K., Kennell, J. C., Kroken, S., Magnuson, J. K., Mannhaupt, G., Mauceli, E., Mewes, H.-W., Mitterbauer, R., Muehlbauer, G., Münsterkötter, M., Nelson, D., O'Donnell, K., Ouellet, T. r. s., Qi, W., Quesneville, H., Roncero, M. I. G., Seong, K.-Y., Tetko, I. V., Urban, M., Waalwijk, C., Ward, T. J., Yao, J., Birren, B. W., and Kistler, H. C. (2007). The *Fusarium graminearum* genome reveals a link between localized polymorphism and pathogen specialization. *Science* **317,** 1400-1402.

Debuchy, R., Turgeon, B.G. (2006). Mating-type structure, evolution, and function in Euascomycetes. *The Mycota I: Growth, Differentiation and Sexuality*, 293 - 323.

DiGuistini, S., Wang, Y., Liao, N.Y., Taylor, G., Tanguay, P., Feau, N., Henrissat, B., Chan, S.K., Hesse-Orce, U., Alamouti, S.M., Tsui, C.K.M., Docking, R.T., Levasseur, A., Haridas, S., Robertson, G., Birol, I., Holt, R.A., Marra, M.A., Hamelin, R.C., Hirst, M., Jones, S.J.M., Bohlmann, Jr. , Breuil, C. (2011). Genome and transcriptome analyses of the mountain pine beetle-fungal symbiont *Grosmannia clavigera*, a lodgepole pine pathogen. *Proceedings of the National Academy of Sciences*. **108**, 2504-2509.

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

Gao, Q., Jin, K., Ying, S-H., Zhang, Y., Xiao, G., Shang, Y., Duan, Z., Hu, X., Xie, X-Q., Zhou, G., Peng, G., Luo, Z., Huang, W., Wang, B., Fang, W., Wang, S., Zhong, Y., Ma, L-J., St. Leger, R.J., Zhao, G-P., Pei, Y., Feng, M-G., Xia, Y., Wang, C. (2011). Genome sequencing and comparative transcriptomics of the model entomopathogenic fungi *Metarhizium anisopliae* and *M. acridum*. *PLoS Genet*. **7**, e1001264.

Gardes, M., Bruns, T.D. (1993). ITS primers with enhanced specificity for basidiomycetes - application to the identification of mycorrhizae and rusts. *Molecular Ecology*. **2**, 113-118.

Glass, N.L., Metzenberg, R.L., Raju, N.B. (1990). Homothallic *Sordariaceae* from nature: The absence of strains containing only the a mating type sequence. *Experimental Mycology*. **14**, 274-289.

Glass, N.L., Smith, M.L. (1994). Structure and function of a mating-type gene from the homothallic species *Neurospora africana*. *Mol Gen Genet*. **244**, 401 - 409.

Glass, N.L., Vollmer, S.J., Staben, C., Grotelueschen, J., Metzenberg, R.L., Yanofsky, C. (1988). DNAs of the two mating-type alleles of *Neurospora crassa* are highly dissimilar. *Science*. **241**, 570 - 573.

Hanlin, R.T. (1985). The pedagogical ascomycete. *Mycologia*. **77**, 1-10.

Harrington, T. C. (2007). The genus *Ceratocystis*: Where does the oak wilt fungus fit. In *Proceedings of the National Oak Wilt Symposium* (pp. 21-35).

Harrington, T. C., and McNew, D. L. (1997). Self-fertility and uni-directional mating-type switching in *Ceratocystis coerulescens*, a filamentous ascomycete. *Current Genetics* **32,** 52-59.

Harrington, T.C., and McNew, D.L. (1998). Partial interfertility among the *Ceratocystis* species on conifers. *Fungal Genetics and Biology*. **25**, 44-53.

Jones, S.K., Bennett, R.J. (2011). Fungal mating pheromones: choreographing the dating game. *Fungal genetics and biology : FG & B*. **48**, 668-676.

Juzwik, J., Harrington, T.C., MacDonald, W.L., Appel, D.N. (2008). The origin of *Ceratocystis fagacearum*, the oak wilt fungus. *Annu. Rev. Phytopathol.* **46**, 13-26.

Kanamori, M., Kato, H., Yasuda, N., Koizumi, S., Peever, TL., Kamakura, T., Teraoka, T., Arie, T. (2007). Novel mating type-dependent transcripts at the mating type locus in *Magnaporthe oryzae*. *Gene*. **403**, 6-17.

Katinka, M.D., Duprat, S., Cornillot, E., Metenier, G., Thomarat, F., Prensier, G., Barbe, V., Peyretaillade, E., Brottier, P., Wincker, P., Delbac, F., El Alaoui, H., Peyret, P., Saurin, W., Gouy, M., Weissenbach, J., Vivares, C.P. (2001). Genome sequence and gene compaction of the eukaryote parasite *Encephalitozoon cuniculi*. *Nature*. **414**, 450-453.

Kile, G.A., Harrington, T.C., Yuan, Z.Q., Dudzinski, M.J., Old, K.M. (1996). *Ceratocystis eucalypti* sp. nov., a vascular stain fungus from eucalypts in Australia. *Mycological Research*. **100**, 571-579.

Kronstad, J. W., and Staben, C. (1997). Mating type in filamentous fungi. *Annual Review of Genetics* **31,** 245-276.

Kuhlman, E.G. (1982). Varieties of *Gibberella fujikuroi* with anamorphs in *Fusarium* section *Liseola*. *Mycologia*. **74**, 759-768.

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

Lee, J., Lee, T., Lee, Y.W., Yun, S.H., Turgeon, B.G. (2003). Shifting fungal reproductive mode by manipulation of mating type genes: obligatory heterothallism of *Gibberella zeae*. *Mol Microbiol*. **50**, 145 - 152.

Lee, J., Leslie, J.F., Bowden, R.L. (2008). Expression and function of sex pheromones and receptors in the homothallic ascomycete *Gibberella zeae*. *Eukaryotic Cell*. **7**, 1211-1221.

Mandel, M.A., Barker, B.M., Kroken, S., Rounsley, S.D., Orbach, M.J. (2007). Genomic and population analyses of the mating type loci in *Coccidioides* species reveal evidence for sexual reproduction and gene acquisition. *Eukaryotic Cell*. **6**, 1189-1199.

Martin, S.H., Wingfield, B.D., Wingfield, M.J., Steenkamp, E.T. (2011) Causes and consequences of variability in peptide mating pheromones of ascomycete fungi. *Molecular Biology and Evolution*. **28**, 1987-2003.

Martin, T., Lu, S-W., van Tilbeurgh, H., Ripoll, D.R., Dixelius, C., Turgeon, B.G., Debuchy, R. (2010). Tracing the origin of the fungal α1 domain places its ancestor in the HMG-Box superfamily: Implication for fungal mating-type evolution. *PLoS One*. **5**, e15199.

Mayrhofer, S., Pöggeler, S. (2005). Functional characterization of an α-factor-like *Sordaria macrospora* peptide pheromone and analysis of its interaction with its cognate receptor in *Saccharomyces cerevisiae*. *Eukaryotic Cell*. **4**, 661-672.

Mayrhofer, S., Weber, J.M., Pöggeler, S. (2006). Pheromones and pheromone receptors are required for proper sexual development in the homothallic ascomycete *Sordaria macrospora*. *Genetics*. **172**, 1521-1533.

Metzenberg, R.L., Glass, N.L. (1990). Mating type and mating strategies in *Neurospora*. *Bioessays*. **12**, 53 - 59.

Nelson, M.A. (1996). Mating systems in Ascomycetes: a romp in the sac. *Trends in Genetics*, **12**, 69-74.

Parra, G., Bradnam, K., Korf, I. (2007). CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*. **23**, 1061-1067.

Payne, G.A., Nierman, W.C., Wortman, J.R., Pritchard, B.L., Brown, D., Dean, R.A., Bhatnagar, D., Cleveland, T.E., Machida, M., Yu, J. (2006). Whole genome comparison of *Aspergillus flavus* and *A. oryzae*. *Medical Mycology*. **44**, 9-11.

Pöggeler, S., Risch, S., Kück, U., Osiewacz, H.D. (1997). Mating-type genes from the homothallic fungus *Sordaria macrospora* are functionally expressed in a heterothallic ascomycete. *Genetics*. **147**, 567-580.

Pöggeler, S. (1999). Phylogenetic relationships between mating-type sequences from homothallic and heterothallic ascomycetes. *Curr Genet*. **36**, 222 - 231.

Pöggeler, S. (2001). Mating-type genes for classical strain improvements of ascomycetes. *Applied Microbiology and Biotechnology*. **56**, 589-601.

Robertson, S.J., Bond, D.J., Read, N.D. (1998). Homothallism and heterothallism in *Sordaria brevicollis*. *Mycol Res*. **102**, 1215 - 1223.

Salamov, A.A., Solovyev, V.V. (2000). Ab initio gene finding in *Drosophila* genomic DNA. *Genome Research*. **10**, 516-522.

Shiu, P.K.T., Glass, N.L. (2000). Cell and nuclear recognition mechanisms mediated by mating type in filamentous ascomycetes. *Current Opinion in Microbiology*. **3**, 183-188.

Spatafora, J.W., Sung, G-H., Johnson, D., Hesse, C., O'Rourke, B., Serdani, M., Spotts, R., Lutzoni, Fo., Hofstetter, V., Miadlikowska, J., Reeb, V., Gueidan, C., Fraker, E., Lumbsch, T., Lücking, R., Schmitt, I., Hosaka, K., Aptroot, A., Roux, C., Miller, A.N., Geiser, D.M., Hafellner, J., Hestmark, G., Arnold, A.E., Büdel, B., Rauhut, A., Hewitt, D., Untereiner, W.A., Cole, M.S., Scheidegger, C., Schultz, M., Sipman, H., Schoch, C.L. (2006). A five-gene phylogeny of Pezizomycotina. *Mycologia*. **98**, 1018-1028.

Stanke, M., Morgenstern, B. (2004). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research*. **33**, W465-W467.

Turgeon, B.G. (1998). Application of mating type gene technology to problems in fungal biology. *Annu Rev Phytopathol*. **36**, 115 - 137.

Turgeon, B.G., Yoder, O.C. (2000) Proposed nomenclature for mating type genes of filamentous ascomycetes. *Fungal Genetics and Biology*. **31**, 1-5.

van der Nest, M. A., Bihon, W., De Vos, L., Naidoo, K., Roodt, D., Rubagotti, E., Slippers, B., Steenkamp, E. T., Wilken, P. M., and Wilson, A. (2014). Draft genome sequences of

Diplodia sapinea, *Ceratocystis manginecans*, and *Ceratocystis moniliformis*. *IMA Fungus* **5,** 135-140.

Waalwijk, C., van der Lee, T., de Vries, I., Hesselink, T., Arts, J., Kema, G. (2004). Synteny in toxigenic *Fusarium* species: The fumonisin gene cluster and the mating type region as examples. *European Journal of Plant Pathology*. **110**, 533-544.

Webster, R.K., Butler, E.E. (1967). The origin of self-sterile, cross-fertile strains and culture sterility in *Ceratocystis Fimbriata*. *Mycologia*. **59**, 212-221.

Wik, L., Karlsson, M., Johannesson, H. (2008). The evolutionary trajectory of the mating-type (mat) genes in *Neurospora* relates to reproductive behavior of taxa. *BMC Evolutionary Biology*. **8**, 109.

Wilken, P. M., Steenkamp, E. T., Wingfield, M. J., De Beer, Z. W., and Wingfield, B. D. (2013). Draft nuclear genome sequence for the plant pathogen, *Ceratocystis fimbriata*. *IMA Fungus* **4,** 357-358.

Wilken, P. M., Steenkamp, E. T., Wingfield, M. J., de Beer, Z. W., and Wingfield, B. D. (2014). DNA loss at the *Ceratocystis fimbriata* mating locus results in self-sterility. *PLoS One.* **9,** e92180.

Wingfield, B.D., Van Wyk, M., Roos, H., Wingfield, M.J., 2013. *Ceratocystis*: Emerging evidence for discrete generic boundries. In: Seifert, K.A., De Beer, W., Wingfield, M.J. (Eds.), Ophiostomatoid Fungi: Expanding Frontiers. CBS-KNAW Fungal Biodiversity Centre, AD Utrecht, The Netherlands.

Witthuhn, R.C., Harrington, T.C., Wingfield, B.D., Steimel, J.P., Wingfield, M.J. (2000). Deletion of the *MAT-2* mating-type gene during uni-directional mating-type switching in *Ceratocystis*. *Curr. Genet.* **38**, 48-52.

Yun, S.H., Berbee, M.L., Yoder, O.C., Turgeon, B.G. (1999). Evolution of the fungal self-fertile reproductive life style from self-sterile ancestors. *Proc Natl Acad Sci USA*. **96**, 5592 - 5597.

Yun, S.H., Arie, T., Kaneko, I., Yoder, O.C., Turgeon, B.G. (2000). Molecular organization of mating type loci in heterothallic, homothallic, and asexual *Gibberella/Fusarium* species. *Fungal Genet Biol*. **31**, 7 - 20.

Zhang, Y., Xu, L., Zhang, S., Liu, X., An, Z., Wang, M., Guo, Y. (2009). Genetic diversity of *Ophiocordyceps sinensis*, a medicinal fungus endemic to the Tibetan Plateau: Implications for its evolution and conservation. *BMC Evolutionary Biology*. **9**, 290.

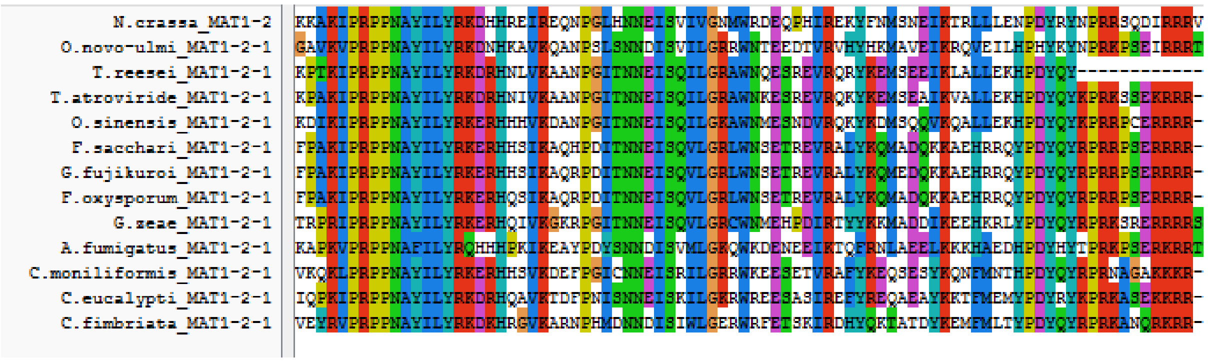
Table 1. Species names and accession numbers for the MAT1-2-1 proteins of various Ascomycetes discussed in this study.

Species name	**Accession number**
Ceratocystis eucalypti	AAF00498.1
Ophiocordyceps sinensis	ACV60385.1
Trichoderma reesei	ADB28880.1
Trichoderma atroviride IMI 206040	EHK42953.1
Fusarium sacchari	BAE94382.1
Gibberella fujikuroi	AAC71056.1
Fusarium oxysporum* f. sp*. radicis-lycopersici	BAA28611.1
Gibberella zeae	AAG42810.1
Neurospora crassa	AAA33598.2
Ophiostoma novo-ulmi	AAX83067.1
Aspergillus fumigatus Af293	XP_754989.2
Ceratocystis fimbriata	Wilken *et al.* 2014

Table 2. A comparison of the *MAT1-2-1* gene and protein sequences between *C. moniliformis* and *C. fimbriata.*

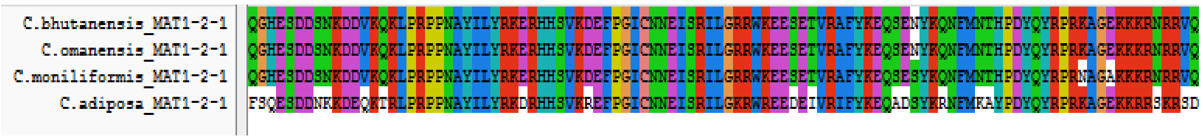
	C. moniliformis	***C. fimbriata***
Number of amino acids	274	458
Number of base pairs	937	1670
Number of introns	2	3

Figure 1. Alignment of the MAT1-2-1 proteins of *C. moniliformis* and other Ascomycetes. The figure only includes a conserved portion of the MAT1-2-1 amino acid sequence containing the HMG-box domain. Species and accession numbers are listed in Table 1.



```
      N.crassa_MAT1-2  KKAKIPRPPNAYILYRKDHHREIREQNPGLHNNEISVIVGNMWRDEQPHIREKYFNMSNEIKTRLLLENPDYRYNPRRSQDIRRRV
  O.novo-ulmi_MAT1-2-1  GAVKVPRPPNAYILYRKDNHKAVKQANPSLSNNDISVILGRRWNTEEDTVRVHYHKMAVEIKRQVEILHPHYKYNPRKPSEIRRRT
     T.reesei_MAT1-2-1  KPTKIPRPPNAYILYRKDRHNLVKAANPGITNNEISQILGRAWNQESREVRQRYKEMSEEIKLALLEKHPDYQY------------
 T.atroviride_MAT1-2-1  KPAKIPRPPNAYILYRKDRHNIVKAANPGITNNEISQILGRAWNKESREVRQKYKEMSEAIKVALLEKHPDYQYKPRKSSEKRRR-
   O.sinensis_MAT1-2-1  KDIKIPRPPNAYILYRKERHHHVKDANPGITNNEISQILGKAWNMESNDVRQKYKDMSQQVKQALLEKHPDYQYKPRRPCERRRR-
   F.sacchari_MAT1-2-1  FPAKIPRPPNAYILYRKERHHSIKAQHPDITNNEISQVLGRLWNSETREVRALYKQMADQKKAEHRRQYPDYQYRPRRPSERRRR-
   G.fujikuroi_MAT1-2-1  FPAKIPRPPNAYILYRKERHHSIKAQRPDITNNEISQVLGRLWNSETREVRALYKQMEDQKKAEHRRQYPDYQYRPRRPSERRRR-
  F.oxysporum_MAT1-2-1  FPAKIPRPPNAYILYRKERHQSIKAQRPDITNNEISQVLGRLWNSETREVRALYKQMADQKKAEHRRQYPDYQYRPRRPSERRRR-
       G.zeae_MAT1-2-1  TRPRIPRPPNAYILYRKERHQIVKGKRPGITNNEISQVLGRCWNMEHPDIRTYYKKMADDIKEEHKRLYPDYQYRPRKSRERRRRS
  A.fumigatus_MAT1-2-1  KAPKVPRPPNAFILYRQHHHPKIKEAYPDYSNNDISVMLGKQWKDENEEIKTQFRNLAEELKKKHAEDHPDYHYTPRKPSERKRRT
C.moniliformis_MAT1-2-1  VKQKLPRPPNAYILYRKERHHSVKDEFPGICNNEISRILGRRWKEESETVRAFYKEQSESYKQNFMNTHPDYQYRPRNAGAKKKR-
  C.eucalypti_MAT1-2-1  IQPKIPRPPNAYILYRKDRHQAVKTDFPNISNNEISKILGKRWREESASIREFYREQAEAYKKTFMEMYPDYRYKPRKASEKKRR-
  C.fimbriata_MAT1-2-1  VEYRVPRPPNAYILYRKDKHRGVKARNPHMDNNDISIWLGERWRFETSKIRDHYQKIATDYKEMFMLTYPDYQYRPRKANQRKKR-
```

Figure 2. Alignment of *C. bhutanensis; C. omanensis C. moniliformis* and *C. adiposa* MAT1-2-1 putative protein sequences. These 4 species belong to the proposed *C. moniliformis s.l.* species complex.



```
C.bhutanensis_MAT1-2-1  QGHESDDSNKDDVKQKLPRPPNAYILYRKERHHSVKDEFPGICNNEISRILGRRWKEESETVRAFYKEQSENYKQNFMNTHPDYQYRPRKAGEKKKRNRRVQ
  C.omanensis_MAT1-2-1  QGHESDDSNKDDVKQKLPRPPNAYILYRKERHHSVKDEFPGICNNEISRILGRRWKEESETVRAFYKEQSENYKQNFMNTHPDYQYRPRKAGEKKKRNRRVQ
C.moniliformis_MAT1-2-1 QGHESDDSNKDDVKQKLPRPPNAYILYRKERHHSVKDEFPGICNNEISRILGRRWKEESETVRAFYKEQSESYKQNFMNTHPDYQYRPRNAGAKKKRNRRVQ
    C.adiposa_MAT1-2-1  FSQESDDNKKDEQKTRLPRPPNAYILYRKDRHHSVKREFPGICNNEISRILGKRWREEDEIVRIFYKEQADSYKRNFMKAYPDYQYRPRKAGEKKRRSKRSD
```

Figure 3. A) The structure of the *MAT* idiomorph of *C. moniliformis* on contig 73. B) Organisation of contig 53 which contains *DNA lyase*, a gene commonly associated with the *MAT* idiomorph. Boxes represent genes with the direction of transcription indicated by arrows. Gene names are indicated above each arrow. Hyp indicates a hypothetical protein.

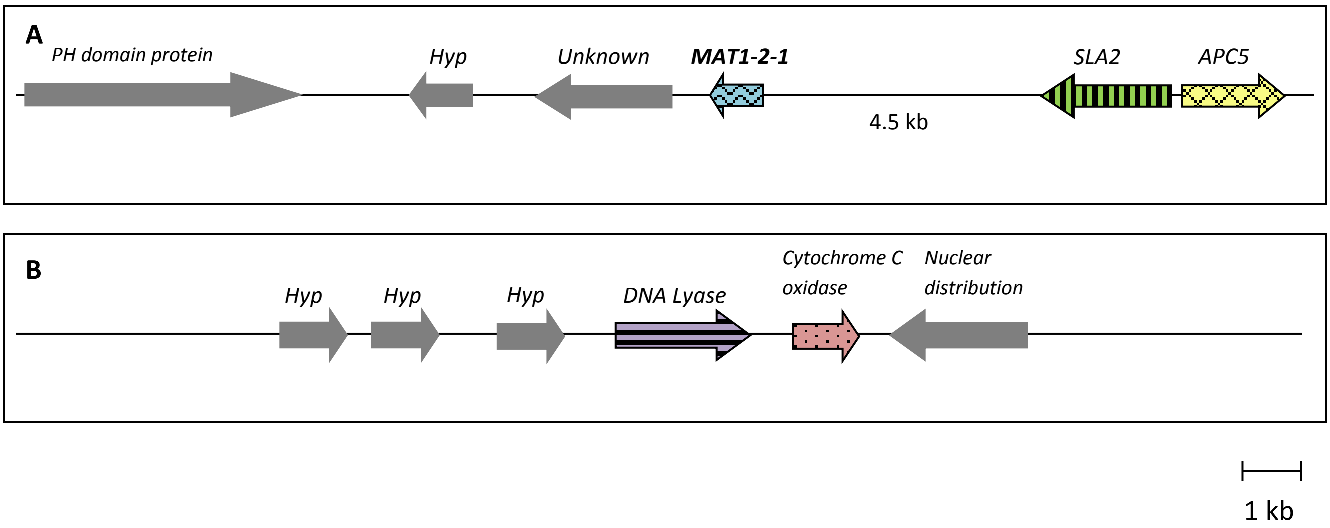
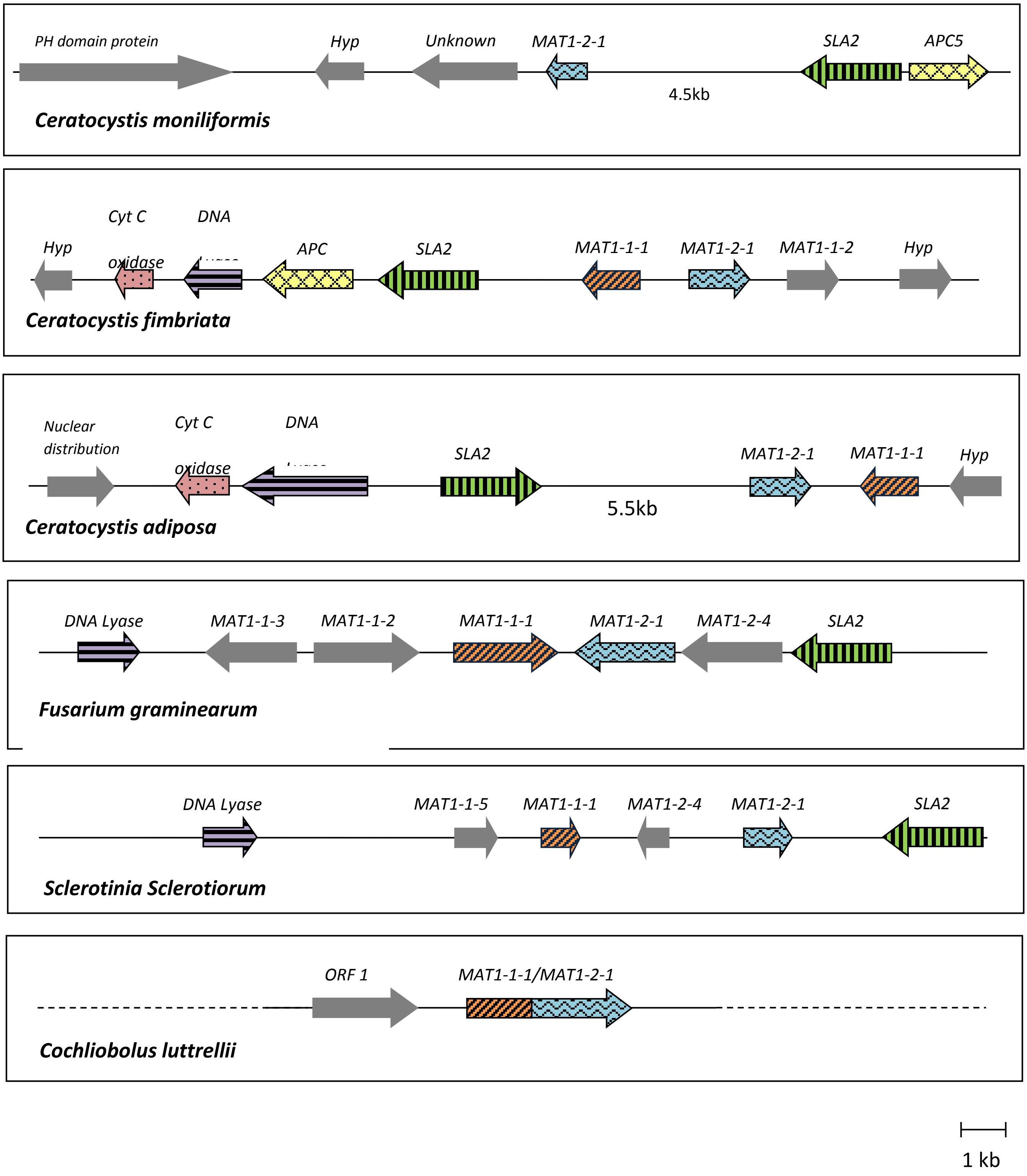

A
PH domain protein
Hyp
Unknown
MAT1-2-1
4.5 kb
SLA2
APC5
B
Hyp
Hyp
Hyp
DNA Lyase
Cytochrome C oxidase
Nuclear distribution
1 kb

Figure 4. Microsynteny around the *MAT* genes in different Ascomycete species. Boxes with colours and patterns represent genes with the direction of transcription indicated by arrows. Gene names are indicated above the gene model. Hyp indicates a hypothetical protein. Dashed lines indicate unknown sequence.


PH domain protein
Hyp
Unknown
MAT1-2-1
SLA2
APC5
4.5kb
Ceratocystis moniliformis
Cyt C
DNA
Hyp
oxidase
APC
SLA2
MAT1-1-1
MAT1-2-1
MAT1-1-2
Hyp
Ceratocystis fimbriata
Nuclear
distribution
Cyt C
DNA
oxidase
SLA2
MAT1-2-1
MAT1-1-1
Hyp
5.5kb
Ceratocystis adiposa
DNA Lyase
MAT1-1-3
MAT1-1-2
MAT1-1-1
MAT1-2-1
MAT1-2-4
SLA2
Fusarium graminearum
DNA Lyase
MAT1-1-5
MAT1-1-1
MAT1-2-4
MAT1-2-1
SLA2
Sclerotinia Sclerotiorum
ORF 1
MAT1-1-1/MAT1-2-1
Cochliobolus luttrellii
1 kb

Supplementary Data

Table S1. Genes implicated in sexual reproduction in fungi that were investigated for their presence in *C. moniliformis*. The genes were grouped according to their role in sexual reproduction: Core mating genes, the pheromone signalling pathway and fruiting body development. The last two columns indicate the BLAST methods used to screen for these genes in the *C. moniliformis* genome, and whether the gene was identified by at least one of these methods.

Gene	Function	Species	Reference	BLAST method	Identified
		CORE MATING GENES			
MAT1-1-1	Mating type (alpha-box domain transcriptional activator)	*C.fimbriata*	Wilken *et al.* 2014	Blastn Blastp	No
		Gibberella zeae	Yun *et al.* 2000	tBlastn tBlastx	No
		Fusarium oxysporum	Yun *et al.* 2000	tBlastn tBlastx	No
MAT1-1-2	Mating type Transcription Factor, amphipathic helix	*C.fimbriata*	Wilken *et al.* 2014	Blastn Blastp	No
		Gibberella zeae	Yun *et al.* 2000	tBlastn tBlastx	No
MAT1-1-3	Mating type Transcription Factor, HMG box	*C.fimbriata*	Wilken *et al.* 2014	Blastn Blastp	No
		Gibberella zeae	Yun *et al.* 2000	tBlastn tBlastx	No
MAT1-2-1	Mating type (HMG-box transcriptional activator)	*C.fimbriata*	Wilken *et al.* 2014	Blastn tBlastn	Yes
		PHEROMONE SIGNALLING PATHWAY			
ppgA (MFα1, MFα2)	Pheromone precursor (alpha-factor like)	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	No
ppgB	Pheromone	*Aspergillus*	Galagan *et*	tBlastn	No

(MFa1, MFa2)	Precursor (a-factor like)	*nidulans*	*al.* 2005		
KEX1	Carboxypeptidase alpha-factor processing	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
KEX2	Endoprotease for alpha-factor processing	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
STE13	Dipeptidyl aminopeptidase alpha-factor processing	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
STE23	Dipeptidyl aminopeptidase for a-factor processing	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
STE2 (preB)	Pheromone receptor (for alpha-factor like pheromone)	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
STE3 (preA)	Pheromone receptor (for a-factor like pheromone)	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
fadA (GPA1)	Alpha subunit G protein	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
sfaD (STE4)	Beta subunit G protein	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
STE5	Scaffold protein	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
RAM1/ STE16	CAAX-farnesyltransferase beta subunit; a-factor modification	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
RAM2	CAAX-farnesyltransferase alpha subunit	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
RCE1	CAAX prenyl protease a-factor C-terminal processing	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
STE14	CAAX prenyl cysteine	*Aspergillus*	Galagan *et*	tBlastn	Yes

	carboxymethyltransferase	*nidulans*	*al.* 2005		
STE24	CAAX prenyl protease	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
STE6	ATP-dependent afflux pump for a-factor like pheromone	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
STE18	Gamma-subunit G protein	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
STE20	Serine/threonine protein kinase MKKKK	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
STE11 (steC)	Serine/threonine protein kinase MKKK	*Aspergillus nidulans*	Galagan *et al.* 2005	BlastP (against Maker predicted protein database)	Yes
STE7	Serine/threonine protein kinase MKK	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
mpkB (FUS3)	Mitogen activated protein kinase	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
SteA (Ste12)	Transcriptional activator, Homeodomain DNA binding	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
STE50	Pheromone adaptation feedback response	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
FRUITING BODY DEVELOPMENT					
Nc asd-1	Ascus development; rhamnogalacuronase B	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
Nc asd-4	Ascus development; GATA-Zn finger TF	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes
Um rum1	Repressor of b mating type regulated genes	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes

Um umc1	MADS-box TF, modulator of pheromone-inducible gene expression	*Aspergillus nidulans*	Galagan *et al.* 2005	tBlastn	Yes

CHAPTER 3

Three independent duplications of the *β-fructofuranosidase* gene in *Ceratocystis*

ABSTRACT

Gene duplications and deletions play a prominent role in the evolution of species and there are a number of mechanisms that could cause these effects. *β-fructofuranosidase* is a glycoside hydrolase family 32 gene that is involved in the hydrolysis of sucrose by plants and fungi. We determined the copy number of *β-fructofuranosidase* genes in each genome from several species in the fungal genus *Ceratocystis*, known to include both plant pathogens and saprophytes. Pathogenic species were found to have two copies of the *β-fructofuranosidase* gene while all but two non-pathogenic species had only one copy of this gene. Some species had a gene that resembled a retrotransposon in the vicinity of the *β-fructofuranosidase* gene. Two possible models are proposed to describe the evolution of the *β-fructofuranosidase* gene in *Ceratocystis.* The first model suggests that the ancestral genome contained a single copy of the *β-fructofuranosidase* gene and the second model proposes two copies in the ancestral genome. Phylogenetic analysis supported the first model and suggests that three independent duplication events have occurred in the evolution of *Ceratocystis* spp.

INTRODUCTION

The fungal genus *Ceratocystis* includes species that are important plant pathogens as well as others that are saprophytes (Wingfield *et al.* 1993; 2013). At the generic level, the taxonomy of these fungi does not reflect their ecology or phylogentic relationships (Wingfield *et al.* 2013). Those species that are pathogens predominately reside in the *C. fimbriata senslu lato* complex (van Wyk *et al.* 2004a). The *C. moniliformis s.l.* complex (van Wyk *et al.* 2006) includes exclusively saprophytic species, while those in the *C. coerulescens s.l.* complex (Harrington *et al.* 1996) include associates of conifer-infesting bark beetles and agents of sap stain. These different ecological groups of species in *Ceratocystis* provide a superb model to compare genes and gene complexes amongst them.

Relatively little is known regarding the evolution of genes in *Ceratocystis* and which genetic traits cause pathogenicity. Gene loss and duplication events play an important role in the evolution of species. Gene duplications supply raw genetic material for biological evolution (Zhang 2003) while gene losses are responsible for differing genetic repertoires between species (Krylov *et al*. 2003). The prevalence of gene duplications is high in all the domains of life including the Archaea, Bacteria and Eukarya (Zhang 2003). In the genome of the eukaryote *Saccharomyces cerevisiae*, 30% (1858) of the 6241 genes are duplicated, while 17% (284) of the 1709 genes from the bacterium *Haemophilus influenzae* are duplicates (Rubin *et al.* 2000). These duplications drive the evolution of species by providing a platform for novel functions or increased copy number of proteins in the cells.

A duplication event can occur from unequal crossover; chromosomal duplication or retrotransposition, with each of these events resulting in a distinct genetic outcome (Zhang 2003). Unequal crossing over produces a tandem duplication of DNA with the duplicated region adjacent to the original sequence. Chromosomal duplication most often occurs due to a lack of disjunction among daughter chromosomes after replication and will usually not result in a tandem repeat (Zhang 2003). Retrotransposition describes the process whereby mRNA is retrotranscribed to complementary DNA (cDNA) and then randomly inserted into the genome (Xiao *et al.* 2008; Zhang 2003). This process results in the loss of introns and regulatory sequences from the genome and is in contrast to both unequal crossover and chromosomal duplication, which results in the retention of introns. Due to the loss of the regulatory sequences such as promoters, the newly inserted retrotranscribed gene becomes a pseudogene, unless it is inserted downstream of a new promoter. This movement throughout the genome could result in a pseudogene being positioned away from the original copy, unlike the linked nature of genes produced via other modes of duplication (Zhang 2003).

The fate of a duplicated gene is largely dependent on its biological function (Krylov *et al.* 2003). If multiple gene copies confer no advantages to the organism, one of the copies will become functionally redundant. A duplicated gene can be advantageous if the extra amounts of a protein or RNA product are beneficial to the organism (Zhang 2003) Also, duplications provide genetic material for processes such as mutation, drift and selection to produce genes with novel or specialised functions. This provides a means by which a genome can be plastic allowing for adaptation to changing environments, species divergence and species specific traits. Gene duplications have been suggested to play a role in the evolution of gene networks with regard to establishing refined gene networks (Wagner 1994).

A study by Parrent *et al.* (2009) showed that pathogenic filamentous Ascomycota had a higher copy number of genes involved in carbohydrate metabolism which may aid pathogens in host infection (Wapinski *et al.* 2007). The *β-fructofuranosidase* gene codes for an invertase which catalyses the hydrolysis of sucrose into fructose and glucose (Kotwal and Shankar 2009). This gene belongs to the glycosyl hydrolase family 32 (GH32) of the sequence-based classification of carbohydrate-active enzymes (Cantarel *et al.* 2009). A number of plant pathogenic fungi are also known to utilize GH32 enzymes to access plant-derived sucrose (Parrent *et al.* 2009). A recent study by van Wyk *et. al.* (2013) showed that the saphrophytic species *C. moniliformis s.l.* had only a single copy of this gene. In this study we exploited the availability of several *Ceratocystis* genomes to assess the copy variation of the *β-fructofuranosidase* gene. Using a bioinformatics approach, the structure of the loci containing the *β-fructofuranosidase* gene in each species was identified and these were compared between species complexes.

The genus *Ceratocystis* was first established in the late 1800's to be associated with black rot on sweet potato and has since expanded into a number of clades with numerous species belong to each (Wingfield *et al.* 2013). The *C. fimbriata* species complex includes *C. acaciivora*, *C. albifundus, C. fimbriata, C. manginecans, C. populicola* and *C. smalleyi* that are used in this study (among others). Species belonging to this complex are characteristically pathogenic on woody angiosperms or herbaceous root crops (Wingfield *et al.* 2013). The *C. moniliformis* complex is comprised of *C. bhutanensis, C. omanensis, C. moniliformis* and *C. savannae* that are considered in this study (among others). Species belonging to this complex are considered to be saprophytes except for *C. bhutanensis* that is associated with a bark beetle and is a weak pathogen (Wingfield *et al.* 2013). The *C. coerulescens* species complex includes *C. laricicola* and *C. virescens* (as well as others that were not used in this study). This complex includes both saprophytes that cause blue stain and pathogens that are associated with bark beetles on conifers (Wingfield *et al.* 2013). Two other smaller groupings have been noted, the first includes *C. paradoxa* and *C. radicicola* and the second includes *C. adiposa* and *C. fagacearum*. The former group are pathogens on monocotyledonous hosts and the latter are saprophytes (Wingfield *et al.* 2013).

MATERIALS AND METHODS

Isolates and Genomic data

All 17 isolates used in this study are maintained in the culture collection (CMW) of the Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, South Africa. These isolates are also duplicated in the culture collection of the Centraal Bureau voor Schimmelcultures (CBS), Utrecht, Netherlands. The genome of *Ceratocystis moniliformis* (CMW 10134) was sequenced as described in an earlier study (Chapter 2). A number of other *Ceratocystis* genomes were available including that of *C. fimbriata*$_T$ (Wilken *et al.* 2013), *C. acaciivora* (CMW 22563), *C. albifundus* (CMW 17620), *C. adiposa* (CMW 2573)*,* *C. bhutanensis* (CMW 8217), *C. deciphens* (CMW 30855), *C. fagacearum* (CMW 2656), *C. laricicola* (CMW 20928), *C. manginecans* (CMW 17570) (van der Nest *et al.* 2014), C. *omanensis* (CMW 11056), *C. paradoxa* (CMW 1546), *C. populicola* (CMW 14789), *C. radicicola* (CMW 1032), *C. savannae* (CMW 17300), *C. smalleyi* (CMW 14800) and *C. virescens* (CMW 17339). These isolates were sequenced on a Genome Analyzer IIx Illumina platform at the Chapel Hill High-Throughput Sequencing Facility, University of North Carolina. Each genome was *de novo* assembled using the CLC Genomics workbench 5.5.1 (CLC Bio, Aarhus, Denmark) with default settings. These species have been classified into four distinct complexes (*C. fimbriata s.l., C. moniliformis s.l., C. coerulescens s.l.* and *C. paradoxa* complexes) and two species, *C. adiposa* and *C. fagacearum*, that have not yet been assigned to a species complex. These fungi represent a suite of different ecological adaptations including those that are pathogens and others that are saprophytes or agents of blue stain (Wingfield *et al.* 2013) (Table 1).

Gene discovery and locus structure

The *C. fimbriata β-fructofuranosidase* genes were identified using a local tBLASTn analysis on the CLC Genomics workbench. For this search, the *Fusarium oxysporum β-fructofuranosidase* protein sequence (NCBI accession number ENH69192) was used as the query sequence, and any regions identified were subjected to gene and protein prediction using FGENESH (Salamov and Solovyev 2000) and AUGUSTUS (Stanke *et al.* 2004). After identification, the *C. fimbriata* β-fructofuranosidase protein was used to identify the gene(s) in the genomes of all other *Ceratocystis* spp. used in this study. Whole contigs that contained the *β-fructofuranosidase* gene(s) from each species were subjected to gene and protein predictions using FGENESH and AUGUSTUS to determine the locus structure. A partial region of the *Fusarium oxysporum* 28S large subunit (LSU) rRNA (JN938909.1) was downloaded from the database of the National Center for Biotechnology Information (NCBI; http://www.ncbi.nlm.nih.gov) and used in BLASTn analysis to identify the corresponding sequence in each of the 17 *Ceratocystis* genomes to construct a species phylogeny.

Gene identification and gene comparisons

All protein predictions were characterised by BLASTp analysis against the NCBI database to verify the presence of the *β-fructofuranosidase* genes and to identify each of the putative genes. This made it possible to construct the regions flanking the *β-fructofuranosidase* gene(s).The putative protein sequences of all *β-fructofuranosidase* genes were analysed for conserved domains using the online program InterProScan (v. 4.8) (http://www.ebi.ac.uk/Tools/pfa/iprscan/). The protein and nucleotide sequences were aligned using the ClustalW algorithm in the MEGA (v. 5) software package (Tamura *et al.* 2011). The alignments were subjected to the CLC Main Workbench to determine percentage identity, distance and differences between the sequences.

Contig Extensions

Reference mapping was used to increase the contig length to match that of the largest *β-fructofuranosidase* containing contig identified in each species complex. Raw sequence reads from the species with a smaller contig size were reference assembled against the complete contig from another species belonging to the same complex using the CLC Genomics workbench 5.5.1.

Phylogenetic analyses

The individual LSU sequences for all the isolates were aligned for each species using the online version of MAFFT v. 6 (Katoh *et al.* 2002). The most appropriate models for a maximum likelihood phylogenetic analysis were determined by using jModelTest (Posada 2008). The Maximum likelihood phylogenetic estimates were generated with PhyML 3.0 (Guindon *et al.* 2009) using the values indicated by jModelTest and 1000 bootstrap replicates. The *Fusarium circinatum* (Wingfield *et al.* 2012) *β-fructofuranosidase* protein and LSU gene regions were used as outgroups for the protein and LSU phylogenies. Nucleotide and protein sequences for the *Ceratocystis β-fructofuranosidase* genes were also used for phylogenetic analyses as described above, but Prottest 3.2 (Darriba *et al.* 2011) was used to determine the most appropriate model for protein sequence analysis. The *Aspergillus niger β-fructofuranosidase* gene (XM_001396534.2) was used as the outgroup for the *β-fructofuranosidase* coding sequence phylogeny. An LSU phylogeny was used rather than an ITS phylogeny as it resulted in higher bootstrap values.

RESULTS

Gene copy number and sequence comparisons

Two positive tBLASTn hits were obtained when searching the genome of *C. fimbriata* for the *β-fructofuranosidase* gene. Protein prediction and blast analysis against the NCBI database verified that these were both *β-fructofuranosidase* genes adjacent to one another and on the same contig. The predicted protein sequence from *C. fimbriata* was then used to do a tBLASTn analysis against the other *Ceratocystis* genomes (Table 2).

All six sequenced *Ceratocystis* species in the *C. fimbriata* complex contained two copies of the *β-fructofuranosidase* gene adjacent to each other and the two genes were present in the same orientation (Figure 1). The *β-fructofuranosidase* genes in this species complex always contained two introns and the predicted gene and protein sizes were reasonably consistent with differences no larger than 3 nucleotides and 1 amino acid (Table 2).

All but one of the sequenced species in the *C. moniliformis* complex (Table 2) harboured a single copy of the *β-fructofuranosidase* gene. The exception in this case was *C. bhutanensis* (Figure 1) that had two genes that matched to the *β-fructofuranosidase* gene from different genome contigs (designated a and b in Figures 2 and 3). One of these was the only gene present on the contig. It was not possible to link the two genes or contigs due to the repetitive sequences that flanked the genes and a lack of sequence homology with other sequenced species in the *C. moniliformis* complex. No introns were present in any of the *β-fructofuranosidase* genes and the predicted gene and protein sizes were identical in all five species (Table 2).

All four species residing in the *C. paradoxa* complex and the two isolates not yet assigned to a complex contained a single copy of the *β-fructofuranosidase* gene (Figure 1). Each gene contained a single intron and in contrast to species sequenced in the other complexes, the predicted gene and protein sizes differed for each of the species, ranging from 1832 to 1872 nucleotides and 611 to 620 amino acids.

Of the two species in the *C. coerulescens* complex, *C. laricicola* contained a single *β-fructofuranosidase* gene while *C. virescens* showed the presence of two gene copies on different contigs (Figure 1). Attempts to link these two genes failed due to a lack of sequence homology with other species. Both species contained a single intron in each of the *β-fructofuranosidase* genes and there was conservation of predicted gene and protein sizes between the two species.

Searches on the *β-fructofuranosidase* putative proteins identified a number of domains, including the glycoside hydrolase family 32 domain and glucanases. Other than *C. bhutanensis* gene b (Figure 1) all sequenced species in the *C. moniliformis* complex harboured a transmembrane region*.* All species that were sequenced in the *C. fimbriata* complex contained this region in the upstream gene copies while the downstream copies lacked a transmembrane region (Figure 3). Species sequenced in the *C. coerulescens* complex, *C. paradoxa* complex and the two species not assigned to a complex did not contain a transmembrane region in their copies of the *β-fructofuranosidase* gene (Figure 3).

Locus structure

The *β-fructofuranosidase*-containing contigs in *C. fimbriata, C. manginecans* and *C. acaciivora* all harboured a number of additional genes (Figure 1), while the *C. albifundus* contig had only the two *β-fructofuranosidase* genes. The remaining species in the *C. fimbriata* complex considered in this study, *C. smalleyi* and *C. populicola* contained each of the *β-fructofuranosidase* genes on a separate contig in the initial analyses, but the reference mapping approach allowed these to be assembled into a single continuous sequence (Figure 1). In a similar way, reference mapping was used to increase the contig length and attempt to produce a single contig for all species relative to the contigs of *C. fimbriata*, *C. moniliformis* and *C. paradoxa,* respective to each of the species complexes considered. However, due to nucleotide repeats that confounded assemblies and a high degree of sequence difference in intergenic regions between species, it was not possible to increase the contig lengths in other species.

The genome sequence of only *C. moniliformis* provided a clear view of the flanking genes for species in the *C. moniliformis* complex (Figure 1). Other than *C. bhutanensis*, the contigs from all the species sequenced in this complex resembled the same locus structure, illustrated by matching downstream genes linked to the *β-fructofuranosidase* gene*. Ceratocystis bhutanensis* had the upstream genes linked to one of the two predicted *β-fructofuranosidase* genes, while the second gene copy was alone on a separate contig (Figure 1).

The flanking genes in sequenced species in the *C. coerulescens* complex, *C. paradoxa* complex and the two species that do not belong to a complex resembled those of the species in the *C. moniliformis* complex (Figure 1). However, *C. fagacearum, C. adiposa* and *C. radicicola* contained a small HSK3 domain-encoding protein between the *β-fructofuranosidase* gene and its downstream flanking regions (Figure 1), which was absent in the other species. *Ceratocystis fimbriata, C. manginecans* and *C. acaciivora* also contained this sequence. BLAST analysis of the HSK3 domain encoding protein identified it with a putative retrotransposon protein Ty1-copia subclass (ABA98084.1; the rice chromosome 11 and 12 Sequencing Consortia, 2005) with an e-value of 5e-11. InterProScan analysis indicated the presence of an RVT_2 (reverse transcriptase RNA-dependent DNA polymerase) and a GAG-POL-related retrotransposon domain.

Phylogenetic relationships

The phylogeny constructed using the LSU sequences from each genome showed that the *C. moniliformis* complex is more closely related to species residing in the *C. paradoxa* and the *C. coerulescens* complexes and the other two species without a defined complex, although the bootstrap value is low, 60% (Figure 4). A similar phylogeny emerged for the protein sequences (Figure 2) and coding nucleotide sequences (Figure 3) of all the *β-fructofuranosidase* sequences present in these 17 species.

The protein phylogeny revealed a clear grouping together of all the upstream genes in the species belonging to the *C. fimbriata* complex and a grouping together of all the downstream genes, even though the bootstrap value was very low at the point of divergence (Figure 2). The protein sequences for species in the *C. coerulescens, C. paradoxa* and the two species lacking a defined complex grouped closer to those in the *C. moniliformis* complex than to those in the *C. fimbriata* species complex. The bootstrap values at each divergence point separating two distinct copies of a gene from the same species were low in all 3 cases (Figure 2). The coding sequence phylogeny however, grouped the sequences for species in the *C. coerulescens* complex and the *C. paradoxa* complex closer to those of the *C. fimbriata* complex. In contrast, this phylogeny provided high bootstrap values at the divergence points for the two copies of the *β-fructofuranosidase* genes in the same species (Figure 3).

DISCUSSION

The availability of genomic sequence data for a number of *Ceratocystis* species made it possible in this study to determine the sequences and copy number of *β-fructofuranosidase* genes for a relatively large assemblage of these fungi representing different ecological groups. It was thus possible to trace the evolutionary history of this gene for species of *Ceratocystis*. Species sequenced from the *C. fimbriata* species complex were shown to have two copies of the gene while fungi that were sequenced in the other four complexes (*C. coerulescens*, *C. paradoxa, C. moniliformis* and the two species in an undefined complex), harboured only one copy of the gene. The only exceptions to this pattern were found in *C. virescens* and *C. bhutanensis* (residing in the *C. coerulescens* and *C. moniliformis* complexes, respectively), that contained two copies of the *β-fructofuranosidase* gene.

A phylogeny (Figure 2) produced using the β-fructofuranosidase protein sequences showed that within the *C. fimbriata* complex, the upstream *β-fructofuranosidase* genes cluster together and the downstream genes cluster together across all the species in the complex, with the exception of the upstream copies of *C. smalleyi* and *C. populicola*. In addition, the two gene copies had different sizes but were conserved in all the species. This provides a strong indication that this locus was present in the ancestral genome for the *C. fimbriata* complex and that it has been maintained through the evolution of new species in the complex (Figure 4). These genes are also more closely related to each other than to those in other species complexes. This suggests that there has been a duplication event in the ancestor of these species complexes and that the genes have diverged in sequence since the duplication event as complexes evolved.

The two copies of the gene in *C. bhutanensis* are slightly different, one being grouped closer to the gene sequences of other species in the complex than to its duplicate but both copies the same in size (Figure 2 and 3). This could suggest an independent duplication event with the accumulation of mutations over time which has caused sequence divergence. The two genes in *C. virescens* cluster together in the phylogeny and are very similar in sequence and size. This could suggest that a more recent independent duplication may have occurred. Because we were unable to link the genes in both species, we are not certain if these genes are adjacent to one another or if the duplicated genes are unlinked in the genomes. Further studies will be required to determine if both genes are expressed or if one is a pseudogene. However, the duplicated genes are very similar to each other in each species.

Results of this study led to the formulation of two possible models to describe the evolution of *β-fructofuranosidase* genes in *Ceratocystis*. A diagrammatic depiction of these models is presented in Figure 5. In the first of these models (Figure 5A) it is proposed that the ancestral genome harboured only one copy of the gene. It is apparent that the *C. moniliformis* complex evolved from this ancestral state and a single copy of the gene was maintained. An independent duplication may have occurred in this complex because *C. bhutanensis* diverged from the other species and contains two gene copies. Similarly, one copy of the *β-fructofuranosidase* gene was maintained in the genomes of species in the *C. coerulescens* complex, *C. paradoxa* complex and species in the unresolved complex. It is possible that a more recent independent duplication event occurred in the *C. virescens* genome. There was some indication that a translocation event occurred at the diversion of the *C. fimbriata* complex as the genes have completely different flanking regions to the other two complexes. The presence of retrotransposon-like genes in some of the species suggests that retrotransposition may have been the source of this alternate locus. A duplication event must also have occurred at this point because all the genomes in this complex have two copies of the gene.

The second model presents a situation where the ancestral genome contained two copies of the *β-fructofuranosidase* gene (Figure 5B). In this case the *C. fimbriata* complex would have evolved from the ancestral genome and maintained two copies of this gene at the same locus. Species in the *C. moniliformis, C. paradoxa*, *C. coerulescens* complexes and the other two species not assigned to a complex would all have diverged from the ancestral genome after a translocation event. This is evident from the fact that these complexes harboured the gene at the same locus, but a different locus to those in the *C. fimbriata* complex. Simultaneously a loss of one of the *β-fructofuranosidase* genes took place. The gene sequence in each complex is significantly different from the gene sequences in other complexes (Figure 2 and 3). Thus, the gene would have evolved independently as each complex diverged from each other. Here, the retrotransposition would have been responsible for the translocation due to its presence in some species at these loci. Support for this model lies in the lack of introns in the genes of species in the *C. moniliformis* complex and the reduced number of introns in the complexes defined by *C. paradoxa* and *C. coerulescens* complexes and those not yet assigned to a complex. Retrotransposition results in the loss of introns as the sequence is generated from mRNA (Zhang 2003). In terms of parsimony, this model would also be supported as it involves only two gene duplication events instead of three as suggested in the first model.

The protein and gene phylogenies (Figures 2 and 3) showed a distinct grouping of duplicate genes within the same species complex. This would support the proposed Model A showing three independent duplication events occurred (Figure 2). In this case, it would suggest that retrotransposition was responsible for the change in location of the *β-fructofuranosidase* genes between the *C. fimbriata* complex and the other complexes. Short direct repeats on both sides of the genes in *C. manginecans* support the view that the *β-fructofuranosidase* gene could have been translocated from the ancestral locus to a new genomic region through the mobilisation of the transposon (Zhang 2003).

It seems unlikely that duplication did occur as a result of retrotransposition. The genes in the *C. fimbriata* complex with two gene copies contained two introns and those in the *C. moniliformis* complex and the remaining three complexes that had a single gene copy and had no introns and one intron, respectively. This would be in contrast to the scenario that retrotransposition was involved in a duplication event as the introns would be lost after a duplication. No loss of introns occurred in species that contained two copies of the gene. Rather, the tandem order of genes in the *C. fimbriata* complex suggests unequal crossover as the origin of duplication. There is evidence to support a gain or loss of introns in duplicated genes (Lin *et al.* 2006), although intron loss is more common than intron gain. Repeat regions that might verify that unequal crossover had occurred were not convincingly identified. Further sequencing analysis is needed to produce contigs with higher sequence coverage to accurately search for repeat sequences.

Previous studies on *β-fructofuranosidase* genes have revealed that there has been gene family expansion in pathogenic filamentous Ascomycota (Parrent *et al.* 2009). These genes are responsible for acquiring sucrose from plant tissues that have a high sucrose concentration, including the sink tissues and phloem cells. This results in augmented fungal growth in plants which would be a disadvantage to them (Parrent *et al.* 2009). Wapinski *et al.* (2007) suggest that genes involved in stress responses and carbohydrate metabolism, as is the case in *β-fructofuranosidase*, undergo duplications more rapidly than other genes such as those involved in growth. It is thus not surprising that there are two copies of these genes in members of the *C. fimbriata* complex as these species include some very aggressive pathogens. In contrast, species in the *C. moniliformis* complex are either saprophytes or weak pathogens and mostly contain a single gene copy.

It is interesting that *C. bhutanensis* is the only species in the *C. moniliformis* complex that had two copies of the *β-fructofuranosidase* genes. *Ceratocystis bhutanensis,* unlike other species in the complex, is associated with a bark beetle and causes necrotic lesions on the host trees (van Wyk *et al.* 2004b; Kirisits *et al.* 2013). It is thus not surprising that this fungus had a second copy of the gene. *Ceratocystis virescens* also had two copies of the gene in contrast to other species in the *C. coerulescens* complex. It is amongst the few species in its complex that are found on hardwoods (Witthuhn *et al.* 1998) and is a pathogen on maple trees. These characteristics support the increased *β-fructofuranosidase* gene copy number in this fungus. Expression analysis of the *β-fructofuranosidase* genes would be required to verify whether both genes are active in *C. fimbriata* species as well as in *C. bhutanensis* and *C. virescens*. This would determine whether the duplicate gene is a pseudogene or if both are fully functional genes.

CONCLUSIONS

Although we don't have a clear understanding of how this gene has evolved, the phylogenies show that gene duplicates are more closely related to the genes from other species in the same species complex than to genes from other species complexes. This suggests a duplication event and not gene loss. Thus we support the proposed model A (Figure 5). The ancestral genome contained one copy of the *β-fructofuranosidase* gene and three independent duplication events have occurred subsequently. We propose that a retrotransposition event may have been responsible for the change in locus between the *C. fimbriata* complex species and the other two species complexes, but played no role in the duplication events. Further sequence analysis and research into the expression of these genes would provide a deeper understanding of the *β-fructofuranosidase* gene evolution in *Ceratocystis.*

REFERENCES

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of Molecular Biology* **215,** 403-410.

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Research* **37**(suppl 1)**,** D233-D238.

Darriba, D., Taboada, G. L., Doallo, R. N., and Posada, D. (2011). ProtTest 3: fast selection of best-fit models of protein evolution. *Bioinformatics* **27,** 1164-1165.

Guindon, S. P., Delsuc, F. D. R., Dufayard, J.-F. O., and Gascuel, O. (2009). Estimating maximum likelihood phylogenies with PhyML. *In* "Bioinformatics for DNA Sequence Analysis", Vol. 537, pp. 113-137. Humana Press.

Harrington, T. C., Steimel, J. P., Wingfield, M. J., and Kile, G. A. (1996). Isozyme variation and species delimitation in the *Ceratocystis coerulescens* complex. *Mycologia* **88,** 104-113.

Katoh, K., Misawa, K., Kuma, K. I., and Miyata, T. (2002). MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Research* **30,** 3059-3066.

Kirisits, T., Konrad, H., Wingfield, M. J., Chhetri, D. B. (2013). *Ophiostomatoid fungi* associated with the Eastern Himalayan spruce bark beetle, *Ips schmutzenhoferi* , in Bhutan and their pathogenicity to *Picea spinulosa* and *Pinus wallichiana*. In: *The Ophiostomatoid Fungi: Expanding Frontiers*. Seifert KA, de Beer ZW, Wingfield MJ. (eds). CBS-KNAW Fungal Biodiversity Centre: Utrecht, The Netherlands. pp 99-112.

Kotwal, S. M., and Shankar, V. (2009). Immobilized invertase. *Biotechnology Advances* **27,** 311-322.

Krylov, D. M., Wolf, Y. I., Rogozin, I. B., and Koonin, E. V. (2003). Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution. *Genome Research* **13,** 2229-2235.

Lin, H., Zhu, W., Silva, J., Gu, X., and Buell, C. R. (2006). Intron gain and loss in segmentally duplicated genes in rice. *Genome Biology* **7,** R41.

Parrent, J., James, T., Vasaitis, R., and Taylor, A. (2009). Friend or foe? Evolutionary history of glycoside hydrolase family 32 genes encoding for sucrolytic activity in fungi and its implications for plant-fungal symbioses. *BMC Evolutionary Biology* **9,** 148.

Posada, D. (2008). jModelTest: Phylogenetic model averaging. *Molecular Biology and Evolution* **25,** 1253-1256.

Rubin, G. M., Yandell, M. D., Wortman, J. R., Gabor, G. L., Miklos, Nelson, C. R., Hariharan, I. K., Fortini, M. E., Li, P. W., Apweiler, R., Fleischmann, W., Cherry, J. M., Henikoff, S., Skupski, M. P., Misra, S., Ashburner, M., Birney, E., Boguski, M. S., Brody, T., Brokstein, P., Celniker, S. E., Chervitz, S. A., Coates, D., Cravchik, A., Gabrielian, A., Galle, R. F., Gelbart, W. M., George, R. A., Goldstein, L. S. B., Gong, F., Guan, P., Harris, N. L., Hay, B. A., Hoskins, R. A., Li, J., Li, Z., Hynes, R. O., Jones, S. J. M., Kuehl, P. M., Lemaitre, B., Littleton, J. T., Morrison, D. K., Mungall, C., O'Farrell, P. H., Pickeral, O. K., Shue, C., Vosshall, L. B., Zhang, J., Zhao, Q., Zheng, X. H., Zhong, F., Zhong, W., Gibbs, R., Venter, J. C., Adams, M. D., and Lewis, S. (2000). Comparative genomics of the eukaryotes. *Science* **287,** 2204-2215.

Salamov, A. A., and Solovyev, V. V. (2000). *Ab initio* gene finding in *Drosophila* genomic DNA. *Genome Research* **10,** 516-522.

Stanke, M., and Morgenstern, B. (2004). AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. *Nucleic Acids Research* **33**(suppl 2)**,** W465-W467.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M., and Kumar, S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular Biology and Evolution* **28,** 2731-2739.

van der Nest, M. A., Bihon, W., De Vos, L., Naidoo, K., Roodt, D., Rubagotti, E., Slippers, B., Steenkamp, E. T., Wilken, P. M., and Wilson, A. (2014). Draft genome sequences of *Diplodia sapinea*, *Ceratocystis manginecans*, and *Ceratocystis moniliformis*. *IMA Fungus* **5,** 135-140.

van Wyk, M., Roux, J., Barnes, I., Wingfield, B. D., Liew, E. C. Y., Assa, B., Summerell, B. A., and Wingfield, M. J. (2004). *Ceratocystis polychroma* sp. nov., a new species from *Syzygium aromaticum* in Sulawesi. *Stud Mycol* **50,** 273-282.

van Wyk, M., Roux, J., Barnes, I., Wingfield, B. D., Chhetri, D. B., Kirisits, T., and Wingfield, M. J. (2004). *Ceratocystis bhutanensis* sp. nov., associated with the bark beetle *Ips schmutzenhoferi* on *Picea spinulosa* in Bhutan. *Studies in Mycology* **50,** 365-379.

van Wyk, M., Roux, J., Barnes, I., Wingfield, B. D., and Wingfield, M. J. (2006). Molecular phylogeny of the *Ceratocystis moniliformis* complex and description of *C. tribiliformis* sp*. nov*. *Fungal Diversity* **21,** 181-201.

van Wyk, N., Trollope, K., Steenkamp, E., Wingfield, B., and Volschenk, H. (2013). Identification of the gene for beta-fructofuranosidase from *Ceratocystis moniliformis* CMW 10134 and characterization of the enzyme expressed in *Saccharomyces cerevisiae*. *BMC Biotechnology* **13,** 100.

Wagner, A. (1994). Evolution of gene networks by gene duplications: A mathematical model and its implications on genome organization. *Proceedings of the National Academy of Sciences of the United States of America* **91,** 4387-4391.

Wapinski, I., Pfeffer, A., Friedman, N., and Regev, A. (2007). Natural history and evolutionary principles of gene duplication in fungi. *Nature* **449,** 54-61.

Wilken, P. M., Steenkamp, E. T., Wingfield, M. J., De Beer, Z. W., and Wingfield, B. D. (2013). Draft nuclear genome sequence for the plant pathogen, *Ceratocystis fimbriata*. *IMA Fungus* **4,** 357-358.

Wingfield, B. D., Steenkamp, E. T., Santana, Q. C., Coetzee, M. P. A., Bam, S., Barnes, I., Beukes, C. W., Chan, W. Y., Vos, L. d., and Fourie, G. (2012). First fungal genome sequence from Africa: A preliminary analysis. *South African Journal of Science* **108**.

Wingfield, B. D., van Wyk, M., Roos, H., Wingfield, M. J. (2013). *Ceratocystis* : emerging evidence for discrete generic boundaries**.** In: *The Ophiostomatoid Fungi: Expanding Frontiers*. Seifert KA, de Beer ZW, Wingfield MJ. (eds). CBS-KNAW Fungal Biodiversity Centre: Utrecht, The Netherlands. pp 57-64.

Witthuhn, R. C., Wingfield, B. D., Wingfield, M. J., Wolfaardt, M., and Harrington, T. C. (1998). Monophyly of the conifer dpecies in the *Ceratocystis coerulescens* complex based on DNA sequence data. *Mycologia* **90,** 96-101.

Xiao, H., Jiang, N., Schaffner, E., Stockinger, E. J., and van der Knaap, E. (2008). A retrotransposon-mediated gene duplication underlies morphological variation of tomato fruit. *Science* **319,** 1527-1530.

Zhang, J. (2003). Evolution by gene duplication: an update. *Trends in Ecology & Evolution* **18,** 292-298.

Table 1. Characterisation of the 17 *Ceratocystis* species considered in this study (Wingfield *et al.* 2013).

Species name	**Species Complex**	**Ecological adaptation**
C. manginecans	*C. fimbriata s.l.*	Pathogen
C. fimbriata	*C. fimbriata s.l.*	Pathogen
C. acaciivora	*C. fimbriata s.l.*	Pathogen
C. albifundus	*C. fimbriata s.l.*	Pathogen
C. smalleyi	*C. fimbriata s.l.*	Pathogen
C. populicola	*C. fimbriata s.l.*	Pathogen
C. virescens	*C. coerulescens s.l.*	Pathogen
C. laricicola	*C. coerulescens s.l.*	Saprophyte (blue stain)
C. radicicola	*C. paradoxa s.l.*	Pathogen on monocotyledonous hosts
C. paradoxa	*C. paradoxa s.l.*	Pathogen on monocotyledonous hosts
C. fagacearum	Not assigned to a complex	Saprophyte
C. adiposa	Not assigned to a complex	Saprophyte
C. moniliformis	*C. moniliformis s.l.*	Saprophyte
C. savannae	*C. moniliformis s.l.*	Saprophyte
C. omanensis	*C. moniliformis s.l.*	Saprophyte
C. deciphens	*C. moniliformis s.l.*	Saprophyte
C. bhutanensis	*C. moniliformis s.l.*	Weak Pathogen

Table 2. Summary of the characteristics of the *β-fructofuranosidase* genes present in the *Ceratocystis* species.

***Species**	**No. of *β-fructofuranosidase* genes**	**No. of introns present in the gene(s)**	**No. of base pairs in the coding sequence**	**No. of Amino acids in the predicted protein**
C. manginecans	2	2	1884 and 1878	627 and 625
C. fimbriata	2	2	1884 and 1878	627 and 625
C. acaciivora	2	2	1884 and 1878	627 and 625
C. albifundus	2	2	1884 and 1878	627 and 625
C. smalleyi	2	2	1881 and 1878	626 and 625
C. populicola	2	2	1881 and 1878	626 and 625
C. virescens	2	1	1872 and 1872	623 and 623
C. laricicola	1	1	1872	623
C. radicicola	1	1	1863	620
C. paradoxa	1	1	1854	617
C. fagacearum	1	1	1845	614
C. adiposa	1	1	1836	611
C. moniliformis	1	0	1848	615
C. savannae	1	0	1848	615
C. omanensis	1	0	1848	615
C. deciphens	1	0	1848	615
C. bhutanensis	2	0	1848 and 1848	615 and 615

*The complexes are distinguished by different colours, C. *fimbriata* is green, *C. coerulescens* is purple, *C. paradoxa* is orange, the species that have not been assigned to a complex are burgundy and *C. moniliformis* is blue.

Figure 1. Locus structure of the *β-fructofuranosidase* genes in *Ceratocystis* species. The solid arrows indicate the *β-fructofuranosidase* genes with duplicates distinguished by numbers or letters. In the *C. fimbriata* complex, upstream gene copies are identified by the number, 1 and downstream copies by the number, 2. The direction of the arrow indicates gene orientation. Solid lines represent linkage on a contig while broken lines portray different contigs.

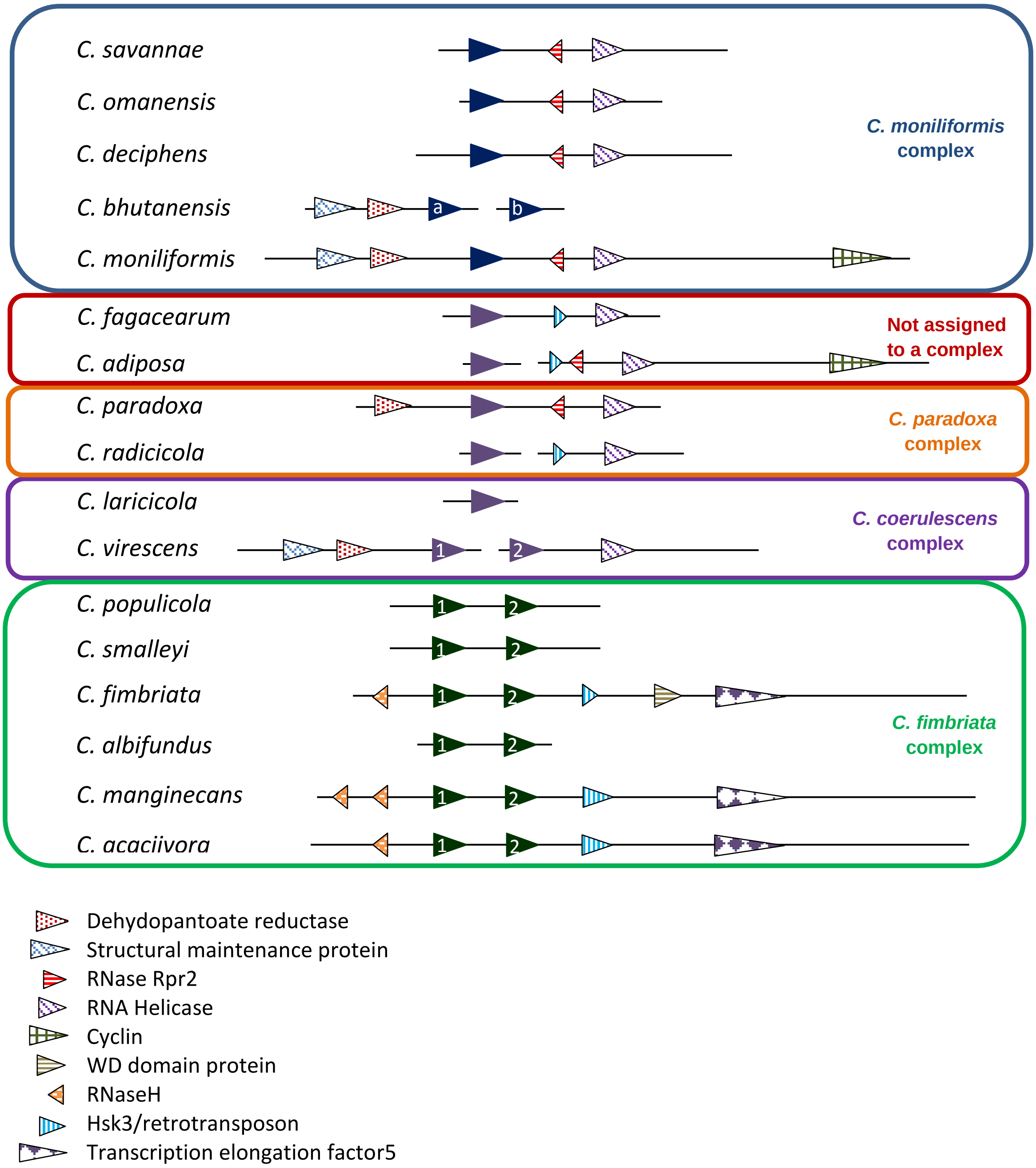

C. savannae
C. omanensis
C. deciphens
C. bhutanensis
a
b
C. moniliformis
C. moniliformis complex
C. fagacearum
C. adiposa
Not assigned to a complex
C. paradoxa
C. radicicola
C. paradoxa complex
C. laricicola
C. virescens
1
2
C. coerulescens complex
C. populicola
1
2
C. smalleyi
1
2
C. fimbriata
1
2
C. albifundus
1
2
C. manginecans
1
2
C. acaciivora
1
2
C. fimbriata complex
Dehydopantoate reductase
Structural maintenance protein
RNase Rpr2
RNA Helicase
Cyclin
WD domain protein
RNaseH
Hsk3/retrotransposon
Transcription elongation factor5

Figure 2. Maximum likelihood phylogeny of the predicted *β-fructofuranosidase* proteins of 17 *Ceratocystis* species. *Fusarium circinatum* represents the out-group taxon. Bootstrap values are indicated at the branch nodes. The number preceding the name of the species indicates the position of the gene in cases where two copies of the gene are present in the genome. a and b differentiate the two genes in *C. bhutanensis.* The brackets show the grouping of the duplicate genes in the *C. fimbriata* complex. Points of individual duplication events are indicated by red arrows.

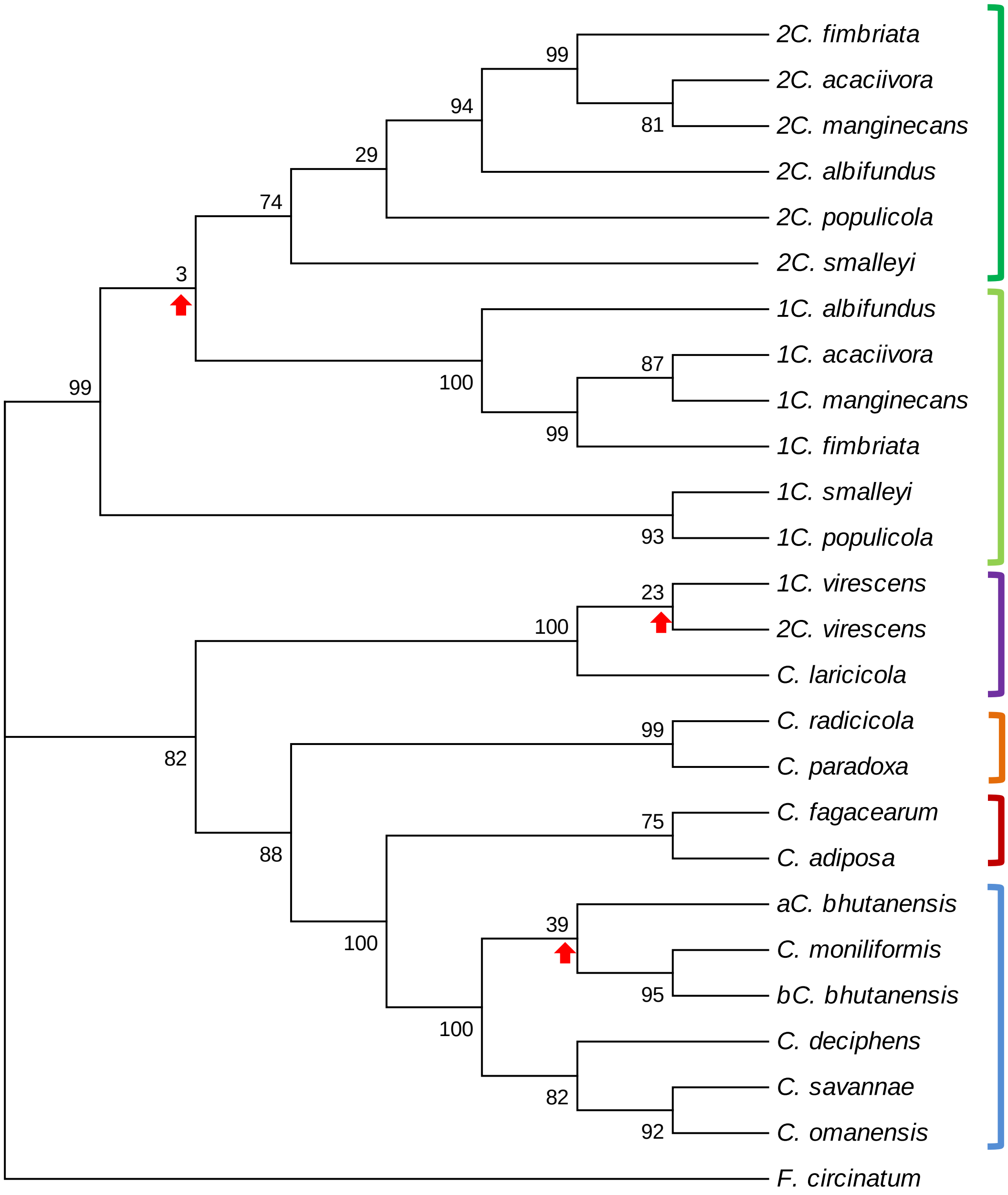

99
2C. fimbriata
2C. acaciivora
81
2C. manginecans
94
2C. albifundus
29
2C. populicola
74
2C. smalleyi
3
1C. albifundus
87
1C. acaciivora
100
1C. manginecans
99
1C. fimbriata
99
1C. smalleyi
93
1C. populicola
23
1C. virescens
100
2C. virescens
C. laricicola
99
C. radicicola
82
C. paradoxa
75
C. fagacearum
88
C. adiposa
aC. bhutanensis
39
C. moniliformis
100
95
bC. bhutanensis
100
C. deciphens
C. savannae
82
92
C. omanensis
F. circinatum

Figure 3. Maximum likelihood phylogeny of the predicted *β-fructofuranosidase* coding sequences of 17 *Ceratocystis* species. *Aspergillus niger* represents the out-group taxon. Bootstrap values are indicated at the branch nodes. The number preceding the name of the species indicates the position of the gene in cases where two copies of the gene are present in the genome. a and b differentiate the two genes in *C. bhutanensis.* The orange 4 point star indicates genes which lack a transmembrane region.

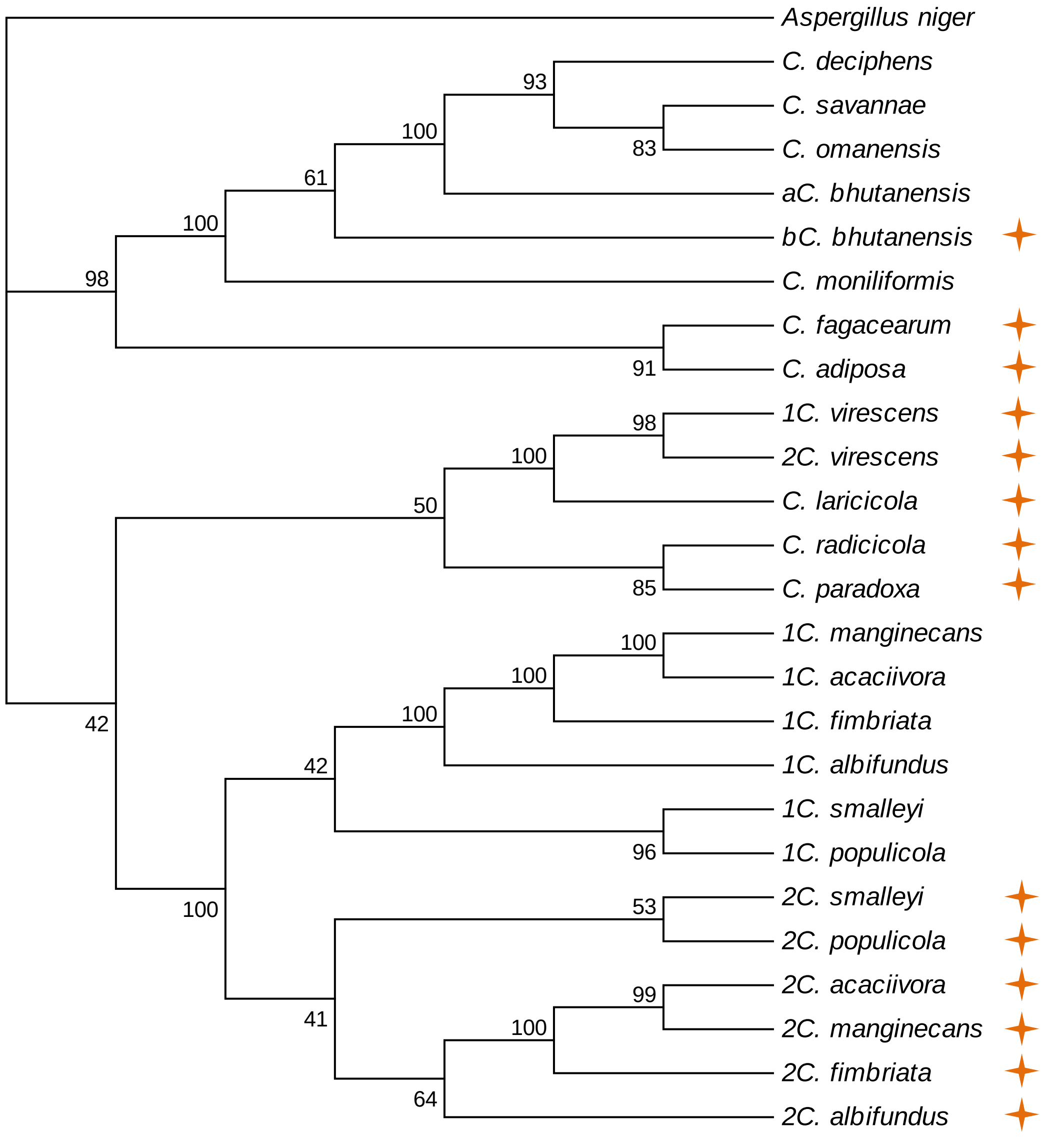

Aspergillus niger
C. deciphens
C. savannae
C. omanensis
aC. bhutanensis
bC. bhutanensis
C. moniliformis
C. fagacearum
C. adiposa
1C. virescens
2C. virescens
C. laricicola
C. radicicola
C. paradoxa
1C. manginecans
1C. acaciivora
1C. fimbriata
1C. albifundus
1C. smalleyi
1C. populicola
2C. smalleyi
2C. populicola
2C. acaciivora
2C. manginecans
2C. fimbriata
2C. albifundus
93
83
100
61
100
98
91
98
100
50
85
100
100
100
42
42
96
53
99
100
100
41
64

Figure 4. Maximum likelihood phylogeny of the complete LSU sequence of 17 *Ceratocystis* species isolates. *Fusarium circinatum* represents the out-group taxon. Bootstrap values are indicated at the branch nodes. Brackets identify species complexes within the genus: blue, the *C. moniliformis* complex; burgundy, species not yet assigned to a complex; orange, the *C. paradoxa* complex; purple, the *C. coerulescens* complex; and green, the C. *fimbriata* complex. The black dot indicates the ancestral genome of the *C. fimbriata* complex which harbours two copies of the *β-fructofuranosidase* gene.

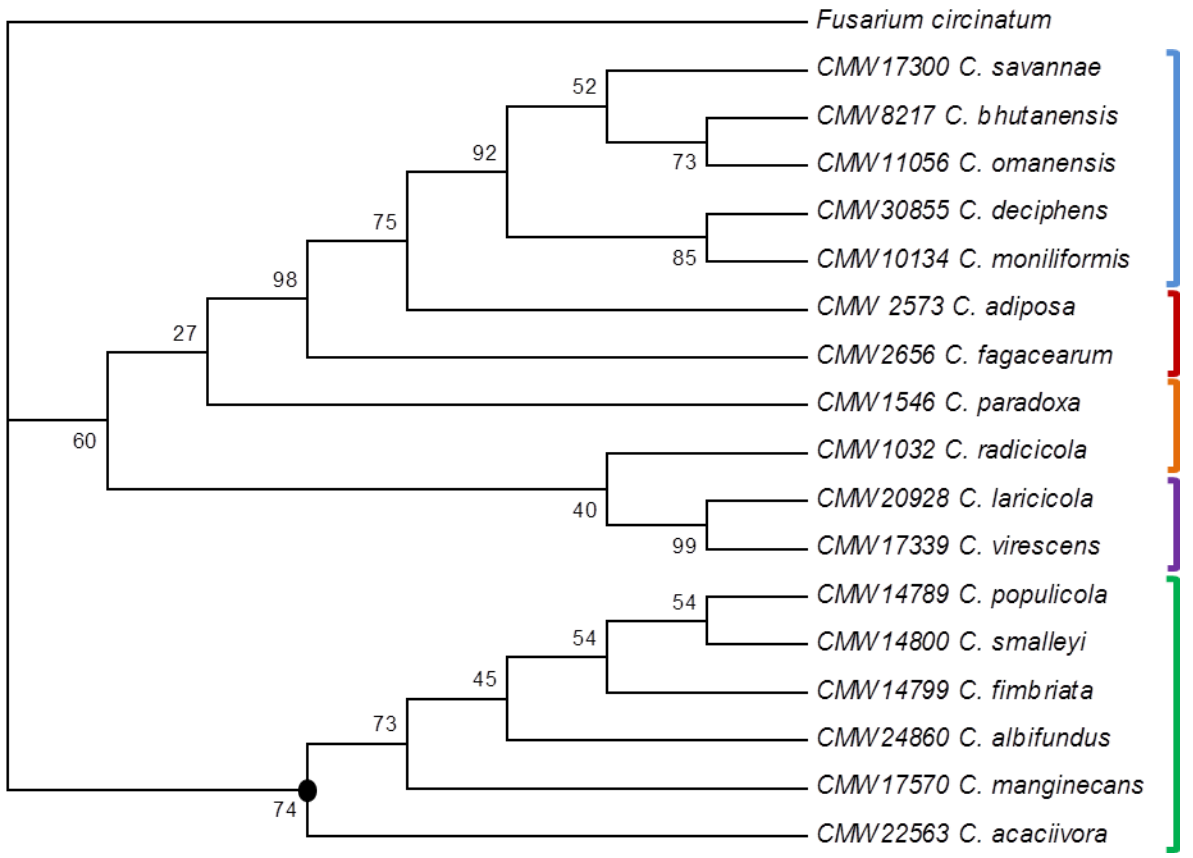
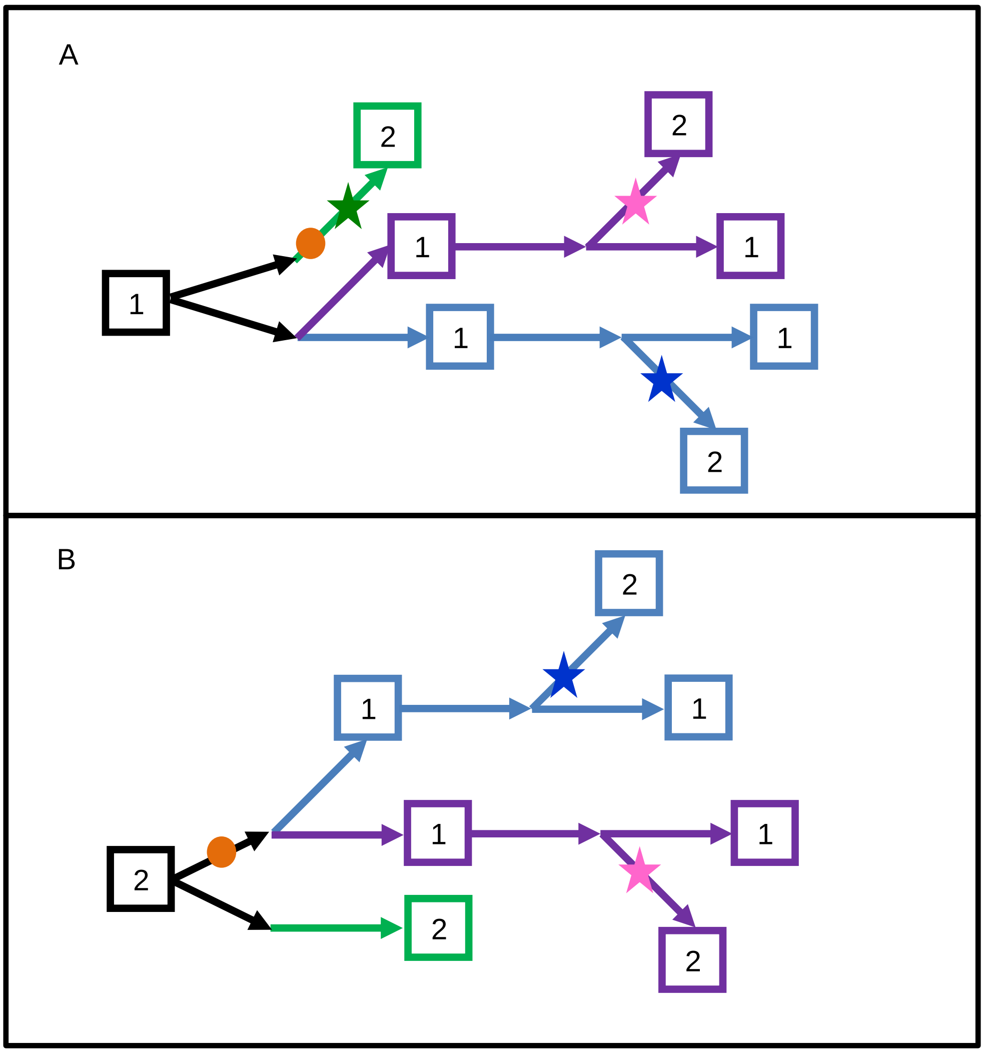

Fusarium circinatum
CMW17300 C. savannae
CMW8217 C. bhutanensis
CMW11056 C. omanensis
CMW30855 C. deciphens
CMW10134 C. moniliformis
CMW 2573 C. adiposa
CMW2656 C. fagacearum
CMW1546 C. paradoxa
CMW1032 C. radicicola
CMW20928 C. laricicola
CMW17339 C. virescens
CMW14789 C. populicola
CMW14800 C. smalleyi
CMW14799 C. fimbriata
CMW24860 C. albifundus
CMW17570 C. manginecans
CMW22563 C. acaciivora
52
73
92
85
75
98
27
60
40
99
54
54
45
73
74

Figure 5. Proposed timeline of models. **A** represents the model if the most ancestral species had a single *β-fructofuranosidase* gene. **B** represents the model if the ancestral species had two *β-fructofuranosidase* genes. The orange dot represents the possible retrotransposition event. Stars represent possible independent duplication events. The black boxes and arrows indicate the ancestral genome; Green boxes and arrows represent the *C. fimbriata* complex; purple represents the *C. coerulescens,* C. *paradoxa* and the two species not yet assigned to a complex (combined for simplicity) and blue represents the *C. moniliformis* complex. The numbers inside the boxes indicate the number of *β-fructofuranosidase* genes present in the genome.


A
2
2
1
1
1
1
1
1
2
B
2
1
1
1
2
1
1
2
2

SUMMARY

The field of genomics is rapidly advancing as new technologies develop. Bioinformatics and genomic comparisons are revolutionary tools used to understand fungal biology. In this study, the genome of a saprophyte, *C. moniliformis* was sequenced and used to increase the knowledge available on mating and evolution in species belonging to the *Ceratocystis* genus. The genome was exploited to determine the mating type structure of the fungus. It is known to reproduce using a homothallic mating strategy and so it was hypothesised that the genome would contain both mating type idiomorphs reported necessary for selfing. Interestingly, the genetic structure was unlike any known homothallic mating type structures within the *Ceratocystis* genus and other Ascomycetes. *Ceratocystis moniliformis* harboured only one mating type gene, *MAT1-2-1*, usually typical of a heterothallic isolate. This was the first report of a homothallic Ascomycete that contains only *MAT1-2* mating type sequence.

A number of *Ceratocystis* species are pathogens on economically important trees, including Eucalyptus and Pine*. β-fructofuranosidase* genes are suspected to have a possible influence on the pathogenicity of a fungus on plants. With the availability of 17 genomes belonging to the *Ceratocystis* genus, including both pathogens and saprophytes, *β-fructofuranosidase* genes were compared between species. A distinct difference was apparent between saprophytes and pathogens which contain one and two gene copies, respectively. Phylogenetic analysis suggested that three independent duplication events occurred and that the ancestral genome contained a single copy of the *β-fructofuranosidase* gene. These results provide novel insights into species evolution in *Ceratocystis* and increase the knowledge regarding genetic differences between pathogens and saprophytes.