# Microsatellite analysis of *Ceratocystis fimbriata*

by

**Melissa Claire Simpson**

Submitted in partial fulfilment of the requirements for the degree

**Magister Scientiae**

In the Faculty of Natural and Agricultural Sciences, Department of Genetics, Forestry and Agricultural Biotechnology Institute, University of Pretoria, Pretoria, South Africa

February 2012

**Supervisor:**
**Prof. Brenda D. Wingfield**

**Co-Supervisors:**
**Prof. Michael J. Wingfield**
**Dr. Martin P.A. Coetzee**
**Mr. Markus P. Wilken**

# Declaration

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

_____

Melissa Claire Simpson

February 2012

# Table of Contents

# Acknowledgments

I would like to express my sincere thanks and appreciation to the following people and institutions:

My principal supervisor, Prof. Brenda Wingfield, for her invaluable guidance, insight and contagious passion throughout my studies. She has been a true inspiration!

My co-supervisors, Prof. Mike Wingfield and Dr. Martin Coetzee, for their guidance, support and many helpful suggestions throughout the course of this study.

My co-supervisor and mentor since my second year of study, Markus Wilken, for helping me in various aspects of my studies and the many engaging and thought-provoking discussions.

My colleagues and friends at FABI, especially those that shared their knowledge and experiences with me.

The University of Pretoria, National Research Foundation (NRF) of South Africa, the South African Biosystematics Initiative (SABI), members of the Tree Protection Co-operative Programme (TPCP), and the Department of Science and Technology (DST)/NRF Centre of Excellence in Tree Health Biotechnology (CTHB) for funding and providing facilities.

My friends and family for their continual encouragement and support.

My parents for giving me the opportunity to study my passion. My Mom, my rock, for all the emotional support, encouragement and advice, especially during the stressful times. You are truly one of a kind! My Dad for his logical insight and clear head, especially in those times when I couldn't see the forest for the trees.

Mark Vrdoljak for spending many long hours and weekends assisting me in finishing my work and editing each chapter of this thesis. More importantly, for keeping my spirits up during the difficult times and for constantly reminding me of the light at the end of the tunnel!

# Preface

*Ceratocystis fimbriata* is the type species for the genus *Ceratocystis* and was first described as the causal agent of black rot in sweet potatoes. However, evidence from DNA sequence data suggests that *C. fimbriata* is in fact a species complex (*C. fimbriata sensu lato*) consisting of many morphologically similar cryptic species. Species in this complex are pathogens of important root and fruit crops and trees in the forestry industry world-wide. Population studies on some of these species have mainly relied on microsatellite markers. However, nothing is known regarding the microsatellite structure within *Ceratocystis* species or any species in the order Microascales in which *Ceratocystis* resides. The need for a more robust identification tool is also required to differentiate between species in this complex.

The first chapter of this thesis provides a review of the literature on microsatellite markers, particularly in fungi. It also discusses the history of microsatellites, mechanisms of microsatellite evolution and functional importance in selected fungal examples. In addition, isolation methodologies are compared and contrasted to newly developed techniques that include bioinformatic searches of genome sequences. Opportunities to use and develop microsatellite markers in *Ceratocystis* species is also discussed with an emphasis on the possibilities that more microsatellites markers would provide.

Microsatellites are abundant in eukaryotic genomes, and fungi are no exception. Analyses of microsatellite content in eukaryotic and fungal genomes have shown that fungi contain fewer microsatellites and that each organism shows preference for particular motifs. In Chapter 2 of this thesis, the abundance and distribution of microsatellites in the recently sequenced *C. fimbriata* genome is investigated. Comparisons to other fungi and eukaryotes show that *C. fimbriata* follows the general pattern of microsatellite structure, however it is unique in its preference for certain motifs.

The *C. fimbriata sensu lato* species complex contains morphologically indistinct species. Microsatellite markers previously developed for a population study could differentiate between some of the cryptic species based on their geographic location and host-specificity. In Chapter 3 a subset of microsatellite markers identified in gene regions in Chapter 2 are used to develop a diagnostic test to differentiate between species in the complex. Microsatellite markers that are polymorphic between species but monomorphic within species were selected for this purpose. However, not all species could be distinguished using this diagnostic test. This thesis is presented as a series of chapters in which Chapters

2 and 3 are in manuscript format. Consequently each chapter represents an independent article and repetition between these chapters has been unavoidable.

# List of Tables

## Chapter 2



## Chapter 3

# List of Figures

## Chapter 1



## Chapter 3

# Chapter 1

# Literature Review:
# Microsatellite markers with special reference to their origins and occurrence in fungi

## 1.0 Introduction

Eukaryotic genomes contain many repetitive elements (Kubis et al., 1998; Tautz and Renz, 1984). The first discovery of one such group of repetitive elements, tandem repeats, was that of a large "minisatellite" in 1980 (Wyman and White). It was subsequently found that these repetitive elements were highly polymorphic and could be used to determine familial relationships (Jeffreys et al., 1985). Another group of tandem repeat elements was also discovered during the 1980's (Hamada et al., 1982; Hotchl and Zachau, 1983; Miesfeld et al., 1981; Nishioka and Leder, 1980; Schaffner et al., 1978; Sures et al., 1978) but were of a shorter length than minisatellites and were consequently referred to as "microsatellites" (Litt and Luty, 1989).

Microsatellites were found to be highly abundant in genomes, inherited co-dominantly and hypervariable, more so than minisatellites (Litt and Luty, 1989; Tautz, 1989; Tautz and Renz, 1984; Weber and May, 1989). These ideal properties allow microsatellites to be applied to a wide range of studies, including forensic science, genome mapping, population genetics and conservation studies (Jarne and Lagoda, 1996; Selkoe and Toonen, 2006). In fungi, microsatellite markers have mostly been used to study populations with regards to their diversity and origins (Barnes et al., 2005; Breuillin et al., 2006; Cortinas et al., 2011; Kamgan Nkuekam et al., 2009; Kubisiak et al., 2007). The focus is usually on pathogenic fungi as knowledge of their population structure and movements are key to their control. For example, analyses of populations of the *Eucalyptus* pathogen *Teratosphaeria gauchensis* isolated in South America suggest that this pathogen possibly underwent a host-jump from related native species (Cortinas et al., 2011). Microsatellites have also been useful in identifying different species and strains of pathogenic fungi, especially in the medical field (Araujo et al., 2009; Botterel et al., 2001; Foulet et al., 2005; Hennequin et al., 2001).

Microsatellite markers are very popular but in the past their isolation has been difficult (Chambers and MacAvoy, 2000). Genomic libraries were required for *de novo* isolation and this proved to be time-consuming and expensive, especially in organisms with a low microsatellite density (Zane et al., 2002). With the recent advances in next generation sequencing and bioinformatic tools for database mining, the cost and speed of discovery and development of microsatellite markers has improved substantially (Abdelkrim et al., 2009; Santana et al., 2009). Comparisons of microsatellite content between different genomes can now be made and informative markers can be developed for a wide variety of studies (Demuth et al., 2007; Drury et al., 2009; Karaoglu et al., 2005).

In this review the history of microsatellites, isolation methodologies, their evolution and functional importance, particularly in fungi, are discussed. The fungal genus *Ceratocystis* will also be examined with regards to microsatellite markers and the potential applications in this genus.

## 2.0 Microsatellite markers

### 2.1. Definition of microsatellite markers

Microsatellites form part of a group of repetitive sequences called variable number of tandem repeats (VNTRs) that are found abundantly throughout the genomes of most prokaryotes and eukaryotes (Bacolla et al., 2008; Borstnik and Pumpernik, 2002; Chambers and MacAvoy, 2000; Field and Wills, 1996; Tóth et al., 2000). Minisatellites, another class of VNTRs, consist of repeat units that are larger than 10 bp and can form a repeat array of up to 30 kb in size (Chambers and MacAvoy, 2000). Microsatellites are less clearly defined and have been described as tandem repeats of 1-5 bp (Le Flèche et al., 2001), 1-6 bp (Goldstein and Pollock, 1997), 2-6 bp (Schlötterer et al., 1998) and 2-8 bp (Armour et al., 1994). Interestingly, repeat units between seven and 10 bp are not considered to be microsatellites or minisatellites because they are suggested to have a different mutational mechanism to both these VNTR classes (Chambers and MacAvoy, 2000).

Several different types of microsatellite classes can be distinguished. Six main classes (pure, interrupted pure, compound, interrupted compound, complex and interrupted complex) are recognised, based on their repeat type (Chambers and MacAvoy, 2000). A pure microsatellite is an exact copy of the repeat unit, e.g. $(TTC)_5$. Compound microsatellites are made up of two or more sets of repeat units located right next to each other, e.g. $(TTG)_4(ACT)_6$. Complex microsatellites consist of multiple microsatellites that are near to each other, e.g. $(TAAG)_3CGAC(TC)_5GA(CAT)_6$. Interrupted versions of these types of microsatellites are due to mutations, such as mismatches or small insertions or deletions which separate the microsatellite into smaller parts, e.g. $(TTC)_3AG(TTC)_2$ (Chambers and MacAvoy, 2000).

### 2.2. History and description of microsatellites

In 1980 the first minisatellite was discovered inadvertently while researchers were searching for polymorphic restriction fragment length polymorphisms (RFLPs) to use as genetic

markers (Wyman and White, 1980). After this discovery, other minisatellites were identified by chance near human myoglobin and ζ-globin genes (Goodbourn et al., 1983; Weller et al., 1984). As RFLPs are not highly polymorphic, scientists were determined to find other loci to use instead which they thought were present in the human genome (Jeffreys et al., 1985). The accidental discovery that minisatellites are highly polymorphic and inherited in a Mendelian fashion led to the development of DNA fingerprinting that became extremely useful in forensic science (Jeffreys et al., 1985).

Around the same time that minisatellites were identified, microsatellites were also discovered (Hamada et al., 1982; Hotchl and Zachau, 1983; Miesfeld et al., 1981; Nishioka and Leder, 1980; Schaffner et al., 1978; Sures et al., 1978). However, it was not until some years later that microsatellites were considered a highly significant discovery (Litt and Luty, 1989; Tautz, 1989; Weber and May, 1989). Various research groups noted the presence of simple repeats in DNA regions, for example TC repeats were found in the spacer region of the histone gene clusters (Schaffner et al., 1978), and CA repeats in the 3' flanking region of the variable κ immunoglobulin genes (Nishioka and Leder, 1980). Neither of these studies, however, could provide a function for microsatellites. The first to hint at a function came from studying TG repeats in the intergenic regions of human globin genes, and it was suggested that regions containing these repeats could be recombination hot spots for gene conversion (Miesfeld et al., 1981; Slightom et al., 1980).

Another function of dinucleotide repeats came from studying their impact on DNA structure. It had already been established that physiological conditions and DNA sequence could influence the structure of DNA *in vitro,* as tracts of the dinucleotides GC, TG and CA adopt the Z-DNA conformation (Arnott et al., 1980; Vorlickova et al., 1982; Zimmer et al., 1982). Subsequently, the genomes of human, mouse, salmon, yeast and calf were searched for TG and CG repeats that could potentially form Z-DNA (Hamada et al., 1982). Findings showed that TG repeats are highly conserved across all organisms studied, while CG was found to a lesser extent in human, mouse and salmon DNA. The change in conformation to and from Z-DNA could influence the transcriptional activity of surrounding genes, thus these dinucleotide repeats were suggested to play an important role in gene regulation under different conditions (Hamada et al., 1982). Recognition of the potential importance of these repetitive elements led to further studies to find the extent to which they are present within genomes.

Early studies on microsatellites focused only on searches for the motifs AA/TT and GT/CA (Hamada et al., 1982; Shenkin and Burdon, 1972; Shenkin and Burdon, 1974). Subsequently a larger hybridisation study was carried out to search for AA/TT, GG/CC, GT/CA, GA/CT and CAG/GTC motifs in the genomes of phylogenetically different species by hybridisation with repeat-containing probes (Tautz and Renz, 1984). The results showed that these microsatellites are present throughout eukaryotic genomes and that they are highly abundant (Tautz and Renz, 1984). Slippage and cross-over events resulting in polymorphisms were suggested to explain the presence of microsatellites in genomes (Tautz and Renz, 1984). Three studies then showed that microsatellites are highly polymorphic due to these mutational events and would be particularly useful for genotyping individuals by PCR (Litt and Luty, 1989; Tautz, 1989; Weber and May, 1989).

### 2.3. Factors influencing the evolution of microsatellites

Microsatellites originate from adjacent identical sequences, but the manner in which this occurs remains uncertain (Levinson and Gutman, 1987; Zhu et al., 2000). The first mutational mechanism proposed to explain the generation, expansion and contraction of microsatellites was DNA replication slippage (Levinson and Gutman, 1987). Other factors such as gene conversion and unequal crossing-over also play a role in the expansion and contraction of microsatellites, but the contribution of these factors is small when compared to slippage events (Ellegren, 2004; Richard and Paques, 2000). These processes, however, cannot explain the origin of short microsatellites, 15 – 20 bp in length.

Three models are currently used to describe the origin of short microsatellites. The first model poses that the accumulation of point mutations at a locus, followed by DNA slippage events, gives rise to short microsatellites (Rose and Falush, 1998). In contrast, the “continuous model” proposes that DNA slippage occurs continuously at all microsatellites loci (Noor et al., 2001; Pupko and Graur, 1998; Sokol and Williams, 2005). DNA slippage cannot account for duplications at sites where there are no previously identified tandem repeats and for this reason, a third model known as “indel slippage” was suggested (Dieringer and Schlötterer, 2003; Zhu et al., 2000). Indel slippage occurs by insertion of a copy of adjacent short sequences, thus creating a short tandem repeat (Zhu et al., 2000).

### 2.3.1. DNA slippage

DNA replication slippage was the first mutational mechanism proposed that would give rise to microsatellites (Levinson and Gutman, 1987). During DNA replication, there is a transient dissociation followed by misalignment of the elongating strand to the template strand in the microsatellite region (Fig. 1). If DNA synthesis continues on the misaligned strand, a hairpin loop will be formed due to the gain or loss of a repeat unit (Fig. 1). The mismatch repair (MMR) system, responsible for repairing errors, often recognises the hairpin loop and removes it, thereby keeping the microsatellite identical (Kelkar et al., 2008). However, if the hairpin loop is not removed, the microsatellite will gain or lose a repeat unit (Levinson and Gutman, 1987). The rate at which DNA slippage occurs is motif and length dependent, i.e. the likelihood of slippage increases as the length of the microsatellite increases (Brinkmann et al., 1998; Brohede and Ellegren, 1999; Sainudiin et al., 2004; Schlötterer and Tautz, 1992).

### 2.3.2. Generation of short microsatellites

DNA slippage can increase or decrease the length of microsatellites, but this process cannot explain the origin of short microsatellites. For this reason, three models have been proposed to account for their evolution. The first general model poses that short microsatellites arise from the accumulation of point mutations (Jarne et al., 1998). Once a threshold length has been reached (8 – 10 bp), which is minimum length of a microsatellite that would allow the stable misalignment of repeats, DNA slippage is activated (Rose and Falush, 1998).

The continuous model was subsequently proposed to explain the origin of short microsatellites as observations of microsatellite frequency did not entirely match the general model (Pupko and Graur, 1998). This model suggests that DNA slippage takes place at all microsatellite loci (Noor et al., 2001; Pupko and Graur, 1998; Sokol and Williams, 2005). Shorter microsatellites (15 – 20 bp or less), however, would have a lower rate of DNA slippage as DNA polymerase would have less sequence to “slip” on (Noor et al., 2001). Therefore, no minimum threshold length would be required for DNA slippage to take place, although the rate of DNA slippage at short and long microsatellite loci may vary (Noor et al., 2001; Pupko and Graur, 1998; Sokol and Williams, 2005).

Indel slippage is the third, and more recent, model that has been suggested as a mechanism for generating short microsatellites (Fig. 2). This type of slippage occurs randomly in the

genome and at a constant rate (Dieringer and Schlötterer, 2003; Messer and Arndt, 2007). As a result, indel slippage not only has an effect on microsatellites but also on all indels in a genome. This is in contrast to DNA slippage that only affects tandem repeats (Messer and Arndt, 2007). The observed microsatellite distribution (both long and short microsatellites) in a genome is, therefore, the result of indel slippage in conjunction with base substitutions and DNA slippage events (Dieringer and Schlötterer, 2003; Leclercq et al., 2010). This would explain the observation that long microsatellites, in contrast to short microsatellites, are over-represented in most eukaryotic genomes (Leclercq et al., 2010; Tóth et al., 2000).

The molecular mechanism of indel slippage is thought to occur through non-homologous end joining (NHEJ) repair (Messer and Arndt, 2007). NHEJ repair fixes double-stranded breaks by ligating complementary single-stranded overhangs together (Pâques and Haber, 1999). If a misalignment occurs between the overhangs, an insertion or deletion may arise (Pâques and Haber, 1999). At sites without any repeats, NHEJ may allow stable mis-pairings of non-homologous regions that would allow the break to be repaired (Leclercq et al., 2010). Overhangs can also be ligated together without any homology as only 1-4 bp micro-homologies are required for NHEJ repair (Rose and Falush, 1998). The mismatch repair system would then correct the mismatched nucleotides, resulting in a tandem duplication (Leclercq et al., 2010). NHEJ repair fills the complementary single-stranded ends before ligation, which can also result in duplications (Roth et al., 1985). However after ligation, nucleotide excision could remove tandem repeats thus resulting in deletions (Leclercq et al., 2010; Roth et al., 1985).

### 2.3.3. Mutation rate and constraints on microsatellite length

Microsatellites experience a higher nucleotide mutation rate of $10^{-6}$ to $10^{-2}$ per locus per generation when compared to the genome wide average of $10^{-8}$ per base per generation (Ellegren, 2000; Sun et al., 2009). Mutation rate among microsatellites depends on a number of factors including intrinsic features, such as the repeat motif composition, the number of tandem repeats and the total length of the microsatellite, as well as regional genomic factors (Kelkar et al., 2008). Different motif compositions form different secondary structures, each with varying stability which has an effect on the slippage events that can take place (Karthikeyan et al., 1999; Kelkar et al., 2008; Sagher et al., 1999). The mutation rate of microsatellites with the same motif composition but different number of repeats varies greatly (Kelkar et al., 2008). Mutation rate, however, increases exponentially with repeat

number of the motif and length of the microsatellite (Kelkar et al., 2008; Webster et al., 2002).

Expansion of microsatellites into very large arrays as a result of increased mutation rate due to microsatellite length has been observed (Leclercq et al., 2010). This is most likely due to point mutations and the mutational forces acting on length. Length mutations by DNA slippage can either add or remove repeat units, thus increasing or decreasing the length of the microsatellite (Schlötterer and Tautz, 1992) whereas point mutations may break the microsatellite into separate smaller tandem repeats (Ellegren, 2004). Due to differences in efficiency of the mismatch repair machinery and the proof-reading ability of DNA polymerase these mutations are not always properly corrected, which would prevent microsatellites from reaching very large sizes (Kelkar et al., 2008; Sia et al., 1997).

### 2.4. Importance of microsatellites

Microsatellites are important components of genomes and as such are also ideal as molecular markers (Jarne and Lagoda, 1996; Michael et al., 2007). Because microsatellites are present in both coding and non-coding regions and experience different selection pressures (Metzgar et al., 2000), researchers can choose the type of microsatellites that would best address a particular research question. For example, population genetic studies would require selectively neutral makers (Selkoe and Toonen, 2006), while studies attempting to differentiate between species might rather select markers within coding regions that could show differences at the species level rather than the individual level.

Microsatellites are ideal molecular markers because they are abundant, co-dominantly inherited, are highly polymorphic, easy to score and produce reproducible results (Jarne and Lagoda, 1996). They are fairly short, about 200-400 bp in length, thus they can be used to study degraded DNA, which is useful for ancient and forensic studies (Selkoe and Toonen, 2006). Microsatellites are also used extensively to study populations as well as to construct genetic maps which allows identification of quantitative trait loci (Chambers and MacAvoy, 2000). Some microsatellites can even be amplified across species as the regions flanking microsatellites are often conserved, which would allow for cross-species comparisons (Barbará et al., 2007).

2.4.1. Distribution of microsatellites within genomes

The distribution pattern of microsatellites within a genome differs between species, but some general patterns are evident. Due to the repetitive nature of these markers, the majority of microsatellites are contained within non-coding regions (Field and Wills, 1996; Morgante et al., 2002; Tóth et al., 2000). However, several studies have shown that a large amount (up to 21 %) of genes contain microsatellites (Gemayel et al., 2010; Marcotte et al., 1998).

2.4.1.1. Microsatellites in intergenic regions

The intergenic regions contain most of the microsatellites present within a genome (Morgante et al., 2002; Tóth et al., 2000). Microsatellite of all types are present but mono- and dinucleotides dominate in almost all taxa studied to date (Metzgar et al., 2000; Tóth et al., 2000). Interestingly, pentanucleotides are almost exclusively found in intergenic regions (Tóth et al., 2000; Zhang et al., 2004). Microsatellite motifs are mostly A/T-rich, however, there are a few exceptions (Tóth et al., 2000; Zhang et al., 2004). For example, the nematode *Caenorhabditis elegans* prefers C/G mononucleotides and AG dinucleotides over A/T motifs (Tóth et al., 2000). There are some similarities of microsatellite distribution in intergenic regions such as the abundance of A/T-rich motifs, however, different selection pressures experience by each organism results in different overall microsatellite content (Tóth et al., 2000).

2.4.1.2. Microsatellites within gene regions

Microsatellite distribution is markedly different between regions of a gene. Genes are comprised of exons, 3' and 5' untranslated regions (UTRs) and introns (Fig. 3). Each of these elements has a unique pattern of microsatellite abundance (Morgante et al., 2002; Tóth et al., 2000).

*3' and 5' untranslated regions*

A much higher microsatellite frequency is observed in the UTRs when compared to the rest of the genome (Morgante et al., 2002). The 3' UTRs and 5' UTRs contain more microsatellites than the other sections of the coding regions (Morgante et al., 2002; Wren et al., 2000). The most common motifs observed in 3' UTRs are tri- and tetranucleotides, while 5' UTR regions contain mostly di- and trinucleotides (Morgante et al., 2002; Wren et al.,

2000). However, 5' UTRs contain considerably more trinucleotides than 3' UTRs and show bias towards certain types of motifs (Thiel et al., 2003; Wren et al., 2000).

*Introns*

Introns, much like intergenic regions of the genome, consist mostly of mono- and dinucleotides (Metzgar et al., 2000; Tóth et al., 2000). There is a bias toward the dinucleotides AC/GT in the introns of most eukaryotes, with most plant species, nematodes and *Saccharomyces cerevisiae* being notable exceptions. In the introns of these organisms, CG/GC seems to be the most abundant dinucleotide motif (Tóth et al., 2000). Interestingly, the ACG motif has never been observed in vertebrate introns while CCG/CGG is a rare repeat in the introns of mammals, vertebrates, fungi and plants (Chambers and MacAvoy, 2000; Tóth et al., 2000). The absence of these particular trinucleotides could be explained in two ways. Firstly, the CpG dinucleotide is highly mutable, and as such will not be stable. Secondly, long tracts of CCG are key motifs for splicing, and if present in introns could interfere with recruitment of the splicing machinery to the correct location (Tóth et al., 2000).

*Exons*

Exons consist mostly of tri- and hexanucleotides but different organisms show preferences for different motifs (Metzgar et al., 2000; Tóth et al., 2000). An example of this is found in primate and rodent exons, which show an abundance of CCG and AGC, while other mammal and vertebrate exons display preference for AGC and AGG (Borstnik and Pumpernik, 2002; Tóth et al., 2000). Within avian exons, AAG and CCG are prevalent trinucleotides, while fungal exons including those of yeasts contain mostly the AAC motif (Primmer et al., 1997; Tóth et al., 2000). These biases exist as each taxon has a preference for particular codons and their corresponding amino acids (Yarus and Folley, 1985).

Tri- and hexanucleotide repeats are common in coding regions due to selection against motifs that will cause disruptive frame-shift mutations (Metzgar et al., 2000). However, other repeat motifs, such as di- and tetranucleotides, are regularly found in coding regions (Metzgar et al., 2000; Tóth et al., 2000). This is normally only possible if the total length of the microsatellite is a multiple of three in order to maintain the correct reading frame (Gibbons and Rokas, 2009; Metzgar et al., 2000). These non-triplet repeats, however, are under purifying selection in genes (Metzgar et al., 2000). One copy more or one copy less of

the repeat motif would change the length from a multiple of three thereby disrupting the reading frame and will thus be selected against (Metzgar et al., 2000).

Triplet repeats in coding regions generate homo-polymeric amino acid tracts in the resulting proteins (Kashi and King, 2006). These proteins seem to be less conserved, have a very distinctive composition when compared to the rest of the proteome and typically form part of protein classes such as transcription factors and protein kinases (Gibbons and Rokas, 2009). Homo-polymeric tracts, therefore, appear to have functions even though they are thought to be unfolded (Huntley and Golding, 2002). Some homo-polymers, such as poly-glutamine tracts, can modulate or activate transcription when bound to a DNA-binding domain (Fondon and Garner, 2004; Karlin and Burge, 1996; Radivojac et al., 2007). Although DNA composition is important in generating tandem repeats, functional and structural limitations are imposed on the tandem repeats by the type of amino acids that are tolerated in the protein (Katti et al., 2001).

The most predominant homo-polymer tracts consist of small and hydrophilic amino acids such as alanine and proline, which are often found in the 3' and 5' flanking regions of structural domains (Faux et al., 2005; van Passel and de Graaff, 2008). These amino acids could, therefore, influence the affinity and flexibility of extracellular structural proteins (Altman et al., 2003). As the repeat number of the motif within a homo-polymer tract increases, so too does the cytotoxicity of the protein as aggregates are more easily formed (Jorda and Kajava, 2010). There is thus selection pressure on these microsatellites to reduce toxic products or to decrease metabolic costs by keeping the repeat number low (van Passel and de Graaff, 2008). The best known examples of the phenotypic effect of homo-polymers are the polyglutamine tracts associated with human neurodegenerative diseases, such as Huntington's disease and Myotonic dystrophy type 2 (Gemayel et al., 2010; Kashi and King, 2006; Liquori et al., 2001).

### 2.4.2. Microsatellites generate diversity

Tandem repeats have been implicated in the rapid evolution of phenotypes because they are a source of diversity due to high mutation rates (Fondon and Garner, 2004). This was clearly demonstrated by Fondon and Garner (2004) who correlated tandem repeat variability to differences in skeletal morphology of different dog breeds. One of the regulatory genes studied, *Alx-4*, is involved in developmental gene expression. Deletion of 51 nucleotides of the tandem repeat, which codes for a proline-glutamine tract, results in reduced gene

expression of *Alx-4* dependent genes phenotypically resulting in an extra dewclaw. Another regulatory gene, *Runx-2*, contains 2 microsatellites, which code for a polyglutamine and a polyalanine tract. The ratio of glutamine to alanine repeats has a strong correlation to midface length and the degree of the dorsoventral nose bend in different breeds (Fig. 4). Even though selection over the last 150 years has resulted in distinct dog breeds and reduced genetic diversity, this study has shown that rapid evolution can still occur over a fairly short period of time in mammals due to the presence of hypervariable microsatellites.

2.4.3. Examples of functional microsatellites in fungi

*Neurospora crassa*

In *N. crassa*, the white collar-1 (WC-1) gene product regulates transcription of a component of the circadian clock (Gu et al., 2000). Circadian clocks are mechanisms that regulate biological processes to coincide with the local environmental events (Dunlap, 2006). The duration of a circadian clock cycle is about 24 hours, but dark/light and temperature cycles can influence this (Michael et al., 2007). Circadian clock function in continuous darkness requires a polyglutamine tract (tandem repeats of CAA or CAG) in the amino terminus of the WC-1 protein (Lee et al., 2003). The length of this tract is correlated to the environmental conditions such as temperature and circadian cycle length in different locations (Michael et al., 2007). Isolates of *N. crassa* from equatorial regions show a longer polyglutamine tract, associated with a shorter circadian cycle length (Michael et al., 2007). This shows that there can be selection for variation in the microsatellite based on the environmental conditions, without deleteriously affecting the functioning of the protein.

*Saccharomyces cerevisiae*

A study on the genomes of several yeast species revealed that 25% of all promoters contain tandem repeats, many of which are A/T rich (Vinces et al., 2009). These A/T rich tandem repeats are present in nucleosome-free regions which allow binding of transcription factors (Lam et al., 2008). Transcription increases as the repeat length of the microsatellite increases, but decreases once a threshold length is reached (Vinces et al., 2009). Repeat length variations, facilitated by microsatellite repeat numbers have an effect on the chromatin structure which affects transcription of a variety of genes (Vinces et al., 2009).

Microsatellites in the 3' UTR of exons can lead to the transcription of larger mRNA products that distribute unevenly in the cell. In *S. cerevisiae* CAG/CTG repeats in the *URA3* gene, which codes for orotodin-5'-phosphate decarboxylase, influence the transcription of a larger mRNA product (Fabre et al., 2002). The CAG/CTG repeats form secondary structures on which the RNA polymerase complex pauses (Fabre et al., 2002). Pausing allows the RNA polymerase complex to slip backwards and re-transcribe a section of the repeated DNA sequence (Jacques and Kolakofsky, 1991). This process causes a longer mRNA product to be transcribed that contains a longer repeated tract and forms clusters in the cytoplasm (Fabre et al., 2002).

*Candida albicans*

*Candida albicans* is an opportunistic human fungal pathogen that causes infections in mucosal surfaces (Hurley and De Louvois, 1979). It is able to obtain nutrients and colonise host tissues by using virulence factors such as proteases (Gemayel et al., 2010). The *SAP2* promoter of one of these proteases contain two microsatellites, $(GCTTT)_n$ and $(TTGAT/A)_n$, that together affect transcription (Staib et al., 2002). Although the repeat number of each microsatellite is not important, the combined length of both microsatellites determines the transcription rate (Staib et al., 2002). For example, $(GCTTT)_4(TTGAT/A)_6$ and $(GCTTT)_5(TTGAT/A)_5$ will both result in the same transcription rate as long as the combined length of the two microsatellites remains the same. The location of microsatellites influence flexibility of the DNA helix and thus variations in the combined length will affect the binding of proteins to upstream regulatory sequences, causing variation in transcriptional activity (Staib et al., 2002). A lower transcription level of *SAP2* results in reduced growth of *C. albicans* and thus less colonisation of the host tissue because there are less proteases available to break down proteins for nutrition (Staib et al., 2002)

*Podospora anserina*

In *P. anserina* the microsatellite, $(CA)_n$, found in the 5' UTR of glyceraldehyde-3-phosphate dehydrogenase affects transcription of this gene (Khashnobish et al., 1998). The UTR was placed in an expression plasmid in front of a reporter gene in order to evaluate the role of the microsatellite in gene expression. An increase in number of repeats of $(CA)_n$ resulted in an increase in reporter activity, thus confirming that this microsatellite does affect gene expression.

*Aspergillus fumigatus*

Levdansky et al. (2007) used the *A. fumigatus* genome sequence to identify VNTRs in coding regions that would contribute to pathogenesis of this human pathogen. They focused on cell-wall protein encoding genes as they are hypothesised to play a role in pathogenesis of the organism towards its host. Four putative cell-wall proteins were identified with leader sequences and GPI anchors (Levdansky et al., 2007). Three of the proteins contained minisatellites and one contained a microsatellite, which is part of a still undescribed protein that appears to play a role in the plasma membrane (Levdansky et al., 2007).

## 2.5. Isolation of microsatellites

The ability to isolate microsatellites for use as molecular markers represents the key to their application in molecular biology. Traditionally, isolating microsatellites has been a time-consuming and expensive process, especially when organisms with low microsatellite densities were considered (Dutech et al., 2007). Repeat-containing probes were used to screen genomic libraries to identify positive clones containing microsatellites (Rassmann et al., 1991). This procedure usually resulted in a less than 12% discovery rate (Zane et al., 2002). The low success rate following this strategy led to the development of a number of high-yielding and more efficient techniques based on *de novo* isolation. Although successful, several problems with this technique still exist. More recently, *in silico* discovery of microsatellites through bioinformatic analysis combined with full genome sequencing has become the gold standard in isolating microsatellites (Abdelkrim et al., 2009; Santana et al., 2009).

### 2.5.1. *De novo* isolation

*De novo* isolation of microsatellites was first developed by Tautz (1989) to find simple sequence repeats in the pilot whale, *Globicephala malaena*, for which no previous microsatellite data had been generated. This technique involves creating a genomic library by digesting DNA with two restriction enzymes and selecting short fragments (250 – 350 bp in size) that are then cloned into a vector and transformed into competent bacterial cells. Clones containing microsatellites are identified by hybridisation with GA/CT dinucleotide probes followed by sequencing of the cloned insert. Although Tautz (1989) only used one type of microsatellite probe (the dinucleotide GA/CT), other microsatellite probes can be used to find a wide variety of simple sequence repeats in an organism (Tautz and Renz,

1984). This technique allowed for microsatellites to be identified within species that have no available microsatellite sequence data. However, it can become time-consuming and expensive. Other more efficient techniques, including non-genomic library techniques and enrichment protocols, have subsequently been developed to identify microsatellites (Zane *et al.*, 2002).

2.5.1.1. Non-genomic-library techniques

Non-genomic-library techniques were developed in order to overcome the problems associated with genomic library generation (Zane et al., 2002). Two such methods use a modification of the randomly amplified polymorphic DNA (RAPD) approach to find microsatellites. This is achieved either by hybridising repeat-containing probes to RAPD profiles and subsequent cloning of positive bands (Ender et al., 1999) or by screening clones consisting of all RAPD products (Lunt et al., 1999). At the same time, other methods were developed that isolate microsatellites directly from genomic DNA by using repeat-anchored primers followed by sequencing to obtain the microsatellite and its flanking regions (Cooper et al., 1997; Lench et al., 1996). A similar technique, randomly amplified microsatellites (RAMS), was developed that amplifies PCR products using primers containing microsatellites and a degenerate 5' end (Zietkiewicz et al., 1994). RAMS has been particularly useful in generating microsatellites in fungal species (Hantula and Müller, 1997).

2.5.1.2. Enrichment protocols

Enrichment of microsatellites in genomic DNA was used to increase the number of microsatellites present within the sample before generating clone libraries so that they could more easily be isolated. One method of genomic library enrichment employed the primer extension step of PCR to produce an enriched dinucleotide library, followed by hybridisation and sequencing of positive clones to obtain microsatellites (Ostrander et al., 1992). Other enrichment protocols followed the path of selective hybridisation, a modification of the traditional *de novo* isolation method, whereby genomic libraries are enriched for di-, tri- and tetranucleotides (Armour et al., 1994; Karagyozov et al., 1993). In this approach, PCR is used to amplify the microsatellite-containing sequences after selective hybridisation to produce an enriched library.

The most popular enrichment technique that has been developed is fast isolation by amplified fragment length polymorphism (AFLP) of sequences containing repeats (FIASCO) as it is quick and inexpensive (Zane et al., 2002). This method employs the efficient

digestion-ligation reaction of AFLP which is the simultaneous digestion of DNA with *Mse*I and ligation of adaptors to the fragmented DNA. PCR is then performed to amplify all fragments that have a flanking *Mse*I site (Zane et al., 2002). A biotinylated probe containing a microsatellite, i.e. $(AC)_7$, is then hybridised to the PCR products (Zane et al., 2002). This is followed by capture of the biotinylated DNA with streptavidin beads and non-specific DNA is removed by stringent washes (Zane et al., 2002). The final product should be a library of highly-enriched PCR products that contain the *Mse*I site (Zane et al., 2002). All these methods, however, require cloning, sequencing, primer design, amplification and testing for polymorphisms (Zane et al., 2002). As a result, these methods are still relatively labour-intensive.

### 2.5.2. *In silico* isolation of microsatellites

As *de novo* isolation of microsatellites can be a long and complicated process, specialised bioinformatic tools to detect microsatellites *in silico* from sequences were developed. With the increase in data, the development of next-generation sequencing and availability of whole genome sequences, these tools have become attractive to researchers aiming at reducing time afforded on microsatellite marker development (Abdelkrim et al., 2009). Search tools, such as MSatFinder (Thurston and Field, 2005) and Tandem Repeats Finder (Benson, 1999), identify microsatellites within sequences by assessing the number of repeats or total length of the microsatellite. Different search parameters can be chosen, depending on the type of study conducted. For example, MSatFinder allows the researcher to choose the minimum number of repeats for each type of microsatellite.

#### 2.5.2.1. *In silico* searches of genome sequences

*In silico* search tools have been used to scan whole genome sequences and sequence databases for microsatellites to obtain an overview of the repeat structure in organisms (Tóth et al., 2000). Such analyses have been performed on a number of eukaryotes, including plants, birds, primates, nematodes and fungi (Borstnik and Pumpernik, 2002; Karaoglu et al., 2005; Katti et al., 2001; Lim et al., 2004; Morgante et al., 2002; Primmer et al., 1997; Tóth et al., 2000). Results from these studies show that (i) the density of microsatellites does not correlate with the size of the genome; (ii) the abundance of particular motifs varies between genomes; and (iii) that as repeat number increases, the frequency of the motif decreases (Karaoglu et al., 2005; Katti et al., 2001; Lim et al., 2004; Tóth et al., 2000). These characteristics are most likely due to differences in the organisation of genomes or

differences in mutation rate and efficiency of the mismatch repair machinery between organisms (Harr et al., 2002; Karaoglu et al., 2005; Katti et al., 2001; Tóth et al., 2000; Wierdl et al., 1997).

*In silico* searches of genomes have shown that every organism has a unique microsatellite profile (Tóth et al., 2000). However, microsatellite profiles of fungi are quite different to those of other higher eukaryotes (Karaoglu et al., 2005; Tóth et al., 2000). They have fewer microsatellites and an under-representation of long repeat motifs which could be a result of less non-coding regions present in fungi rather than just a smaller genome size (Karaoglu et al., 2005; Lim et al., 2004; Tóth et al., 2000). In fungi GC content seems to be correlated to microsatellite density, with a lower GC content correlated to a higher density (Lim et al., 2004).

2.5.2.2. *In silico* searches for the development of microsatellite markers

Search tools are useful in studies where little or no sequence data are available and quick development of microsatellite markers is required. An example of this is the use of pyrosequencing to generate millions of short reads (200-300 bp) from which microsatellites can be identified using bioinformatic tools (Abdelkrim et al. 2009). Another strategy is to first enrich genomic libraries for microsatellites (as for enrichment protocols for *de novo* isolation) before employing pyrosequencing to generate short reads (Santana et al., 2009). Search tools are then employed to obtain microsatellites from the sequence data. Both these techniques were able to rapidly identify many potential microsatellites to develop further as molecular markers from organisms for which whole genome sequences were not yet available.

New sequencing methods and bioinformatic tools can greatly increase the number of microsatellites identified in organisms for which conventional methods of *de novo* isolation have proven difficult. An example is seen in the red flour beetle, *Tribolium castaneum*, which is a pest of stored foods but also a model organism (Park et al., 1964). After many years of struggling to develop microsatellite markers, researchers were able to develop 19 markers (Pai et al., 2003). However, once the whole genome of *T. castaneum* had been sequenced, more than 12 000 microsatellites were discovered (Demuth et al., 2007). From these microsatellites, 981 primer pairs were tested of which 509 were found to be polymorphic (Demuth et al., 2007). Fifteen polymorphic microsatellite markers were chosen based on distribution within the genome and genetic variability among samples to study populations of

this model organism in the wild (Drury et al., 2009). This example demonstrates the power of whole genome sequence data to search for microsatellites when conventional methods have not been very successful in producing sufficient markers for the necessary studies.

2.5.2.3. Problems with *in silico* searches

Bioinformatic tools have the advantage of enabling researchers to quickly search through large amounts of sequence data for microsatellites. However, each tool is different and as such biases exist even when analysing the same data sets, yielding different results (Merkel and Gemmell, 2008). For example, three studies on microsatellites in the same fungus all show different results with regards to the number and types of microsatellites observed in the genomes. In *Aspergillus nidulans* 2650, 1161 and 4837 total microsatellites were reported by Lim et al. (2004), Karaoglu et al. (2005) and Li et al. (2009) respectively. This problem can be attributed to the controversy surrounding the definition of a microsatellite (i.e. 1-6 bp tandem repeats), the different microsatellite types (i.e. perfect, interrupted, complex) and the minimum number of repeats that are searched for (Merkel and Gemmell, 2008).

The minimum number of repeats that a particular search tool can seek is influenced by the parameters input by the researcher. Each software package is also limited by how small the tandem repeats are that it can detect (Merkel and Gemmell, 2008). Previously, five repeats was the preferred minimum repeat number but it has recently been shown that DNA slippage can occur on motifs consisting of just two tandem repeats, which could already be considered a microsatellite (Leclercq et al., 2010). Two suggestions have been made as to which parameters should be used: a minimum repeat length of eight bp for all microsatellites (Rose and Falush, 1998); or a minimum repeat number of nine for mononucleotides and four for di-, tri-, tetra-, penta- and hexanucleotides (Lai and Sun, 2003). At this point, there does not seem to be consensus as to which parameters for minimum length or repeat number should be used and individual research groups are currently making their own choices (Abdelkrim et al., 2009; Karaoglu et al., 2005; Santana et al., 2009). Merkel and Gemmel (2008), however, suggest that in order for consistent comparisons to be made between different studies, details on all the parameters should be given in full.

## 3.0 Opportunities to apply microsatellite markers in the fungal genus *Ceratocystis*

*Ceratocystis* is a fungal genus that consists of many pathogens and saprophytes that affect important agricultural crops and forestry plantations world-wide (Kile, 1993). These fungi are

transported by insects to new hosts where they colonise wounds (Hinds, 1972; Juzwik and French, 1983). They can cause cankers, wilting and vascular staining in trees and rot of root and fruit crops (Barnes et al., 2003; Morris et al., 1993; Ribeiro et al., 1986). DNA sequence data has increasingly shown that there are complexes of cryptic species (Fig. 5) within this genus (Barnes et al., 2003; Johnson et al., 2005; van Wyk et al., 2011; Wingfield et al., 1996).

The type species for the genus *Ceratocystis* is *C. fimbriata* which was first described in 1890 as the causal agent of black rot in sweet potatoes (Halsted). Many species have since been described as *C. fimbriata* due to their morphological similarities, but in actual fact they form part of a complex of cryptic species, *C. fimbriata sensu lato (s.l.)* that affect a wide variety of plants across the world. The first rDNA sequence data for *Ceratocystis* allowed *C. albifundus* to be described as a separate species to *C. fimbriata* (Wingfield et al., 1996). Currently there are 26 species within this complex that can be identified through morphological characteristics and phylogenies of the ITS region.

To further understand species within *Ceratocystis*, population studies have been carried out and have mainly relied on microsatellite markers (Barnes et al., 2005; Engelbrecht et al., 2007; van Wyk et al., 2006). Most of these markers have been developed for species in the *C. fimbriata sensu lato (s.l.)* complex (Barnes et al., 2001; Marin et al., 2009; Rizatto et al., 2010; Steimel et al., 2004). This is not surprising as many of the plant pathogens within *Ceratocystis* reside in the *C. fimbriata s.l.* complex (Engelbrecht and Harrington, 2005; Morris et al., 1993; Roux et al., 2004). The microsatellite markers developed from *C. fimbriata* were also able to differentiate between isolates of this species according to their geographic location and host-specialisation (Barnes et al., 2001). The only other microsatellites developed in *Ceratocystis* are from *C. polonica* which is a blue-stain fungus of conifers and belongs to the *C. coerulescens sensu lato* species complex (Marin et al., 2009). Most of the microsatellite markers are transferable to other species within the complex they were developed in, for example microsatellites developed from *C. fimbriata* and *C. cacaofunesta* could be used in a population study on *C. pirilliformis* (Barnes et al., 2005; Engelbrecht et al., 2004; Ferreira et al., 2010; Kamgan Nkuekam et al., 2009).

The development of more microsatellite markers within *Ceratocystis* would present the opportunity to perform more in-depth and robust studies on species within this genus. It has already been shown that microsatellites can differentiate between isolates of *C. fimbriata* from different geographical locations (Barnes et al., 2001). With more microsatellites, this

observation could be taken further to differentiate between the cryptic species within the complexes. The best way to develop many new microsatellite markers is by using the bioinformatics approach as a whole genome sequence can yield thousands of microsatellites (Demuth et al., 2007; Karaoglu et al., 2005). This would also provide an opportunity to analyse the microsatellite content not only within the whole genome of a *Ceratocystis* species, but also within the order Microascales, to which *Ceratocystis* belong, for which a study such as this has not yet been performed.

## 4.0 Conclusions

Microsatellites are useful markers for a variety of studies, but they are not clearly defined (Chambers and MacAvoy, 2000). There is a general consensus that each repeat motif is six bp or less in length and that they are tandemly repeated in a genome (Chambers and MacAvoy, 2000). Due to mutations, tandem repeats can be added or removed which results in polymorphisms (Schlötterer and Tautz, 1992). However, the mutational mechanisms of these insertions and deletions are still not fully understood.

A combination of point mutations, DNA slippage and indel slippage are thought to bring about the observed microsatellite distributions in genomes (Leclercq et al., 2010; Schlötterer and Tautz, 1992). These distributions are also affected by selection for particular types of microsatellite motifs and lengths in different organisms, especially in coding regions (Metzgar et al., 2000; Tóth et al., 2000). Microsatellite lengths vary in genes present among species, thus those microsatellites affecting phenotypes can be identified (Levdansky et al., 2007) and could potentially be used to better differentiate species.

Microsatellites are abundant in fungi, although not to the same extent as they are in higher eukaryotes (Tóth et al., 2000). Fungal microsatellites are also shorter, but these features cannot be accounted for by just a smaller genome size (Karaoglu et al., 2005; Lim et al., 2004). It has been suggested that the genome organisation (e.g. fungal genomes have less non-coding regions) and the efficiency of the mismatch repair system play a role in the distribution and abundance of microsatellites (Harr et al., 2002; Katti et al., 2001). As microsatellites are abundant in fungal genomes they have been popular molecular markers to use, especially in population studies (Barnes et al., 2005; Breuillin et al., 2006; Cortinas et al., 2011). Strain typing of fungi has also been achieved with microsatellite markers and species differentiation is possible in some genera (Foulet et al., 2005; Hennequin et al., 2001).

Isolating microsatellites has been time-consuming in the past as they have to be isolated *de novo* (Zane et al., 2002). Recently, whole genome sequencing projects have drastically increased and along with it the number of genomes that can be studied as well as the software that can be used to search for microsatellites *in silico* (Karaoglu et al., 2005; Lim et al., 2004). This has allowed the development of microsatellite markers to become simpler and more efficient (Abdelkrim et al., 2009; Santana et al., 2009). Many eukaryotic genomes have already been searched *in silico* for microsatellites and other interesting features, such as transposable elements (Karaoglu et al., 2005; Kim et al., 1998).

The genome of the plant pathogen *C. fimbriata sensu stricto* has recently been sequenced and the availability of this genome sequence would allow for an in depth study on microsatellite content in this fungus. Comparisons to other Ascomycete genomes could then provide insight into the evolution of this fungus and related species. With so many microsatellites that can be identified from the genome, choosing specific microsatellites for further studies will be possible. Microsatellites can then be developed that could potentially aid in differentiating the cryptic species in the *C. fimbriata s.l.* complex. The genome sequence therefore offers many more possibilities to study this plant pathogen in more detail with regards to its genome content.

## 5.0 References

Abdelkrim, J., Robertson, B. C., Stanton, J. L., Gemmell, N. J., 2009. Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. Biotechniques. 46**,** 185-192.

Altman, G. H., Diaz, F., Jakuba, C., Calabro, T., Horan, R. L., Chen, J. S., Lu, H., Richmond, J., Kaplan, D. L., 2003. Silk-based biomaterials. Biomaterials. 24**,** 401-416.

Araujo, R., Pina-Vaz, C., Rodrigues, A. G., Amorim, A., Gusmão, L., 2009. Simple and highly discriminatory microsatellite-based mulitplex PCR for *Aspergillus fumigatus* strain typing. Clinical Microbiology and Infection. 15**,** 260-266.

Armour, J. A. L., Neumann, R., Gobert, S., Jeffreys, A. J., 1994. Isolation of human simple repeats loci by hybridisation selection. Human Molecular Genetics. 3**,** 599-605.

Arnott, S., Chandrasekaran, R., Birdsall, D. L., Leslie, A. G. W., Ratliff, R. L., 1980. Left-handed DNA helices. Nature. 283**,** 743-745.

Bacolla, A., Larson, J. E., Collins, J. R., Li, J., Milosavljevic, A., Stenson, P. D., Cooper, D. N., Wells, R. D., 2008. Abundance and length of simple repeats in vertebrate

genomes are determined by their structural properties. Genome Research. 18**,** 1545-1553.

Barbará, T., Palma-Silva, C., Paggi, G. M., Bered, F., Fay, M. F., Lexer, C., 2007. Cross-species transfer of nuclear microsatellite markers: Potential and limitations. Molecular Ecology. 16**,** 3759-3767.

Barnes, I., Gaur, A., Burgess, T., Roux, J., Wingfield, B. D., Wingfield, M. J., 2001. Microsatellite markers reflect intra-specific relationships between isolates of the vascular wilt pathogen *Ceratocystis fimbriata*. Molecular Plant Pathology. 2**,** 319.

Barnes, I., Nakabonge, G., Roux, J., Wingfield, B. D., Wingfield, M. J., 2005. Comparison of populations of the wilt pathogen *Ceratocystis albifundus* in South Africa and Uganda. Plant Pathology. 54**,** 189-195.

Barnes, I., Roux, J., Wingfield, B. D., Dudzinski, M. J., Old, M. N., Wingfield, M. J., 2003. *Ceratocystis pirilliformis*, a new species from *Eucalyptus nitens* in Australia. Mycologia. 95**,** 865-871.

Benson, G., 1999. Tandem repeats finder: A program to analyse DNA sequences. Nucleic Acids Research. 27**,** 573-580.

Borstnik, B., Pumpernik, D., 2002. Tandem repeats in protein coding regions of primate genes. Genome Research. 12**,** 909-915.

Botterel, F., Desterke, C., Costa, C., Bretagne, S., 2001. Analysis of microsatellite markers of *Candida albicans* used for rapid typing. Journal of Clinical Microbiology. 39**,** 4076-4081.

Breuillin, F., Dutech, C., Robin, C., 2006. Genetic diversity of the Chestnut blight fungus *Cryphonectria parasitica* in four French populations assessed by microsatellite markers. Mycological Research. 110**,** 288-296.

Brinkmann, B., Klintschar, M., Neuhuber, F., Huhne, J., Rolf, B., 1998. Mutation rate in human microsatellites: Influence of the structure and length of the tandem repeat. American Journal of Human Genetics. 62**,** 1408-1415.

Brohede, J., Ellegren, H., 1999. Microsatellite evolution: Polarity of substitutions within repeats and neutrality of flanking sequences. Proceedings of the Royal Society of London: Biological Sciences. 266**,** 825-833.

Chambers, G. K., MacAvoy, E. S., 2000. Microsatellites: Consensus and controversy. Comparative Biochemistry and Physiology. 126**,** 455-476.

Cooper, S. J. B., Bull, C. M., Gardner, M. G., 1997. Characterization of microsatellite loci from the socially monogamous lizard *Tiliqua rugosa* using a PCR-based isolation technique. Molecular Ecology. 6**,** 793-795.

Cortinas, M. N., Barnes, I., Wingfield, B. D., Wingfield, M. J., 2011. Unexpected genetic diversity revealed in the *Eucalyptus* canker pathogen *Teratosphaeria gauchensis*. Australasian Plant Pathology. 40**,** 497-503.

Demuth, J. P., Drury, D. W., Peters, M. L., van Dyken, D., Priest, N. K., Wade, M. J., 2007. Genome-wide survey of *Tribolium castaneum* microsatellites and description of 509 polymorphic markers. Molecular Ecology Notes. 7**,** 1189-1195.

Dieringer, D., Schlötterer, C., 2003. Two distinct modes of microsatellite mutation processes: Evidence from the complete genomic sequences of nine species. Genome Research. 13**,** 2242-2251.

Drury, D. W., Siniard, A. L., Wade, M. J., 2009. Genetic differentiation among wild populations of *Tribolium castaneum* estimated using microsatellite markers. Journal of Heredity. Doi: 10.1093/jhered/esp77.

Dunlap, J. C., 2006. Proteins in the *Neurospora* circadian clockworks. Journal of Biological Chemistry. 281**,** 28489-28493.

Dutech, C., Enjalbert, J., Fournier, E., Delmotte, F., Barrès, B., Carlier, J., Tharreau, D., Giraud, T., 2007. Challenges of microsatellite isolation in fungi. Fungal Genetics and Biology. 44**,** 933-949.

Ellegren, H., 2000. Microsatellite mutations in the germline: Implications for evolutionary inference. Trends in Genetics. 16**,** 551-558.

Ellegren, H., 2004. Microsatellites: Simple sequences with complex evolution. Nature Reviews Genetics. 5**,** 435-445.

Ender, A., Schwenk, K., Stadler, T., Streit, B., Shierwater, B., 1999. RAPD identification of microsatellites in *Daphnia*. Molecular Ecology. 5**,** 437-441.

Engelbrecht, C. J. B., Harrington, T. C., 2005. Intersterility, morphology and taxonomy of *Ceratocystis fimbriata* on sweet potato, cacao and sycamore. Mycologia. 97**,** 57-69.

Engelbrecht, C. J. B., Harrington, T. C., Alfenas, A. C., Suarez, C., 2007. Genetic variations in populations of the cacao wilt pathogen, *Ceratocystis cacaofunesta*. Plant Pathology. 56**,** 923-933.

Engelbrecht, C. J. B., Harrington, T. C., Steimel, J., Capretti, P., 2004. Genetic variation in eastern North American and putatively introduced populations of *Ceratocystis fimbriata* f. *platani*. Molecular Ecology. 13**,** 2995-3005.

Fabre, E., Dujon, B., Richard, G. F., 2002. Transcription and nuclear transport of CAG/CTG trinucleotide repeats in yeast. Nucleic Acids Research. 30**,** 3540-3547.

Faux, N. G., Bottomley, S. P., Lesk, A. M., Irving, J. A., Morrison, J. R., de la Banda, M. C., Whisstock, J. C., 2005. Functional insights from the distribution and role of homopeptide repeat-containing proteins. Genome Research. 15**,** 537-551.

Ferreira, E. M., Harrington, T. C., Thorpe, D. J., Alfenas, A. C., 2010. Genetic diversity and interfertility among highly differentiated populations of *Ceratocystis fimbriata* in Brazil. Plant Pathology. 59, 721-735.

Field, D., Wills, C., 1996. Long, polymorphic microsatellites in simple organisms. Proceedings of the Royal Society of London: Biological Sciences. 263**,** 209-215.

Fondon, J. W., Garner, H. R., 2004. Molecular origins of rapid and continuous morphological evolution. Proceedings of the National Academy of Sciences of the United States of America. 101**,** 18058-18063.

Foulet, F., Nicolas, N., Eloy, O., Botterel, F., Gantier, J.-C., Costa, J.-M., Bretagne, S., 2005. Microsatellite marker analysis as a typing system for *Candida glabrata*. Journal of Clinical Microbiology. 43**,** 4574-4579.

Gemayel, R., Vinces, M. D., Legendre, M., Verstrepen, K. J., 2010. Variable number of tandem repeats accelerate evolution of coding and regulatory sequences. Annual Review of Genetics. 44**,** 445-477.

Gibbons, J. G., Rokas, A., 2009. Comparative and functional characterisation of intragenic tandem repeats in 10 *Aspergillus* genomes. Molecular Biology and Evolution. 26**,** 591-602.

Goldstein, D. B., Pollock, D. D., 1997. Launching microsatellites: A review of mutation processes and methods of phylogenetic inference. Journal of Heredity. 88**,** 335-342.

Goodbourn, S. E. Y., Higgs, D. R., Clegg, J. B., Weatherall, D. J., 1983. Molecular basis of length polymorphism in the human ζ-globin gene complex. Proceedings of the National Academy of Sciences of the United States of America. 80**,** 5022-5026.

Gu, Y. Z., Hogenesch, J. B., Bradfield, C. A., 2000. The PAS superfamily: Sensors of environmental and developmental signals. Annual Review of Pharmacology and Toxicology. 40**,** 519-561.

Hamada, H., Petrino, M. G., Kakunaga, T., 1982. A novel repeated element with Z-DNA-forming potential is widely found in evolutionarily diverse eukaryotic genomes. Proceedings of the National Academy of Sciences of the United States of America. 79**,** 6465-6469.

Hantula, J., Müller, M. M., 1997. Variation within *Gremmeniella abietina* in Finland and other countries as determined by Random Amplified Microsatellites (RAMS). Mycological Research. 101**,** 169-175.

Harr, B., Todorova, J., Schlötterer, C., 2002. Mismatch repair-driven mutational bias in *D. melanogaster*. Molecular Cell. 10**,** 199-205.

Hennequin, C., Thierry, A., Richard, G. F., Lecointre, G., Nguyen, H. V., Gaillardin, C., Dujon, B., 2001. Microsatellite typing as a new tool for identification of *Saccharomyces cerevisiae* strains. Journal of Clinical Microbiology. 39**,** 551-559.

Hinds, T. E., 1972. *Ceratocystis* canker of aspen. Phytopathology. 62**,** 213–220.

Hotchl, J., Zachau, H. G., 1983. A novel type of aberrant recombination in immunoglobulin genes and its implications for *V-J* joining mechanism. Nature. 302**,** 260-263.

Huntley, M. A., Golding, G.B., 2002. Simple sequences are rare in the Protein data bank. Proteins. 48**,** 134-140.

Hurley, R., De Louvois, J., 1979. *Candida* vaginitis. Postgraduate Medical Journal. 55**,** 645-647.

Jacques, J. P., Kolakofsky, D., 1991. Psuedo-templated transcription in prokaryotic and eukaryotic organisms. Genes & Development. 5**,** 707-713.

Jarne, P., David, P., Viard, F., 1998. Microsatellites, transposable elements and the X chromosome. Molecular and Cellular Biology. 15**,** 28-34.

Jarne, P., Lagoda, P. J. L., 1996. Microsatellites, from molecules to populations and back. Trends in Ecology and Evolution. 11**,** 424-429.

Jeffreys, A. J., Wilson, V., Thein, S. L., 1985. Hypervariable 'minisatellite' regions in human DNA. Nature. 314**,** 67-73.

Johnson, J. A., Harrington, T. C., Engelbrecht, C. J. B., 2005. Phylogeny and taxonomy of the North American clade of the *Ceratocystis fimbriata* complex. Mycologia. 97**,** 1067-1092.

Jorda, J., Kajava, A. V., 2010. Protein homorepeats: Sequences, structures, evolution and functions. Advances in Protein Chemistry and Structural Biology, Vol 79. 79**,** 59-88.

Juzwik, J., French, D. W., 1983. *Ceratocystis fagacearum* and *C. piceae* on the surfaces of free-flying and fungus-mat-inhabiting nitidulids. Phytopathology. 73**,** 1164-1168.

Kamgan Nkuekam, G., Barnes, I., Wingfield, M. J., Roux, J., 2009. Distribution and population diversity of *Ceratocystis pirilliformis* in South Africa. Mycologia. 101**,** 17-25.

Karagyozov, L., Kalcheva, I. D., Chapman, V. M., 1993. Construction of random small-insert genomic libraries highly enriched for simple sequence repeats. Nucleic Acids Research. 21**,** 3911-3912.

Karaoglu, H., Lee, C. M. Y., Meyer, W., 2005. Survey of simple sequence repeats in completed fungal genomes. Molecular Biology and Evolution. 22**,** 639-649.

Karlin, S., Burge, C., 1996. Trinucleotide repeats and long homopeptides in genes and proteins associated with nervous system disease and development. Proceedings of the National Academy of Sciences of the United States of America. 93**,** 1560-1565.

Karthikeyan, G., Chary, K. V. R., Basuthkar, J. R., 1999. Fold-back structures at the distal end influence DNA slippage at the proximal end during mononucleotide repeat expansions. Nucleic Acids Research. 27**,** 3851-3858.

Kashi, Y., King, D. G., 2006. Simple sequence repeats as advantageous mutators in evolution. Trends in Genetics. 22**,** 253-259.

Katti, M. V., Ranjekar, P. K., Gupta, V. S., 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Molecular Biology and Evolution. 18**,** 1161-1167.

Kelkar, Y. D., Tyekucheva, S., Chiaromonte, F., Makova, K. D., 2008. The genome-wide determinants of human and chimpanzee microsatellite evolution. Genome Research. 18**,** 30-38.

Khashnobish, A., Hamann, A., Osiewacz, H. D., 1998. Modulation of gene expression by $(CA)_n$ microsatellites in the filamentous Ascomycete *Podospora anserina*. Applied Microbiology and Biotechnology. 52**,** 191-195.

Kile, G. A., 1993. Plant diseases caused by species of *Ceratocystis sensu stricto* and *Chalara*, in: Wingfield, M.J., Seifert, K.A., Webber, F.J. (Eds.), *Ceratocystis* and *Ophiostoma*: Taxonomy, ecology and pathogenicity. APS Press, St. Paul, Minnesota, USA, pp. 173-183.

Kim, J. M., Vanguri, S., Boeke, J. D., Gabriel, A., Voytas, D. F., 1998. Organisation: A comprehensive survey of the retrotransposons revealed by the complete *Saccharomyces cerevisiae* genome sequence. Genome Research. 8**,** 464-478.

Kubis, S., Schmidt, T., Heslop-Harrison, J.S., 1998. Repetitive elements as a major component of plant genomes. Annals of Botany. 82, 45-55.

Kubisiak, T. L., Dutech, C., Milgroom, M. G., 2007. Fifty-three polymorphic microsatellite loci in the chesnut blight fungus, *Cryphonectria parasitica*. Molecular Ecology Notes. 7**,** 428-432.

Lai, Y., Sun, F., 2003. The relationship between microsatellite slippage mutation rate and the number of repeat units. Molecular Biology and Evolution. 20**,** 2123-2131.

Lam, F. H., Steger, D. J., O'Shea, E. K., 2008. Chromatin decouples promoter threshold from dynamic range. Nature. 453**,** 246-U16.

Le Flèche, P., Hauck, Y., Onteniente, L., Prieur, A., Denoeud, F., Ramisse, V., Sylvestre, P., Benson, G., Ramisse, F., Vergnaud, G., 2001. A tandem repeats database for bacterial genomes: Application to the genotyping of *Yersinia pestis* and *Bacillus anthracis*. BMC Microbiol. 1**,** 2.

Leclercq, S., Rivals, E., Jarne, P., 2010. DNA slippage occurs at microsatellite loci without minimal threshold length in humans: A comparative genomic approach. Genome Biology and Evolution. 2**,** 325-335.

Lee, K., Dunlap, J. C., Loros, J. J., 2003. Roles for WHITE COLLAR-1 in circadian and general photoperception in *Neurospora crassa*. Genetics. 163**,** 103-114.

Lench, N. J., Norris, A., Bailey, A., Booth, A., Markham, A. F., 1996. Vectorette PCR isolation of microsatellite repeat sequences using anchored dinucleotide repeat primers. Nucleic Acids Research. 24**,** 2190-2191.

Levdansky, E., Romano, J., Shadkchan, Y., Sharon, H., Verstrepen, K. J., Fink, G. R., Osherov, N., 2007. Coding tandem repeats generate genetic diversity in *Aspergillus fumigatus* genes. Eukaryotic Cell. 6**,** 1380-1391.

Levinson, G., Gutman, G. A., 1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. Molecular Biology and Evolution. 4**,** 203-221.

Li, C., Liu, L., Yang, J., Li, J., Su, Y., Zhang, Y., Wang, Y., Zhu, Y., 2009. Genome-wide analysis of microsatellite sequence in seven filamentous fungi. Interdisciplinary Sciences and Computer Life Sciences. 1**,** 141-150.

Lim, S., Notley-McRobb, L., Lim, M., Carter, D. A., 2004. A comparison of the nature and abundance of microsatellites in 14 fungal genomes. Fungal Genetics and Biology. 41**,** 1025-1036.

Liquori, C. L., Ricker, K., Moseley, M. L., Jacobsen, J. F., Kress, W., Naylor, S. L., Day, J. W., Ranum, L. P. W., 2001. Myotonic dystrophy type 2 caused by a CCTG expansion in intron 1 of ZNF9. Science. 293**,** 864-867.

Litt, M., Luty, J. A., 1989. A hypervariable microsatellite revealed by *in vitro* amplification of a dinucleotide repeat within the cardiac muscle actin gene. American Journal of Human Genetics. 44**,** 397-401.

Lunt, D. H., Hutchinson, W. F., Carvalho, G. R., 1999. An efficient method for PCR-based isolation of microsatellite arrays (PIMA). Molecular Ecology. 8**,** 891-893.

Marcotte, E. M., Pellegrini, M., Yeates, T. O., Eisenberg, D., 1998. A census of protein repeats. Journal of Molecular Biology. 239**,** 151-160.

Marin, M., Preisig, O., Wingfield, B. D., Wingfield, M. J., 2009. Simple sequence repeat markers reflect diversity and geogrpahic barriers in Eurasion populations of the conifer pathogen *Ceratocystis polonica*. Forest Pathology. 39**,** 249-265.

Merkel, A., Gemmell, N. J., 2008. Detecting microsatellites in genome data: Variance in definitions and bioinformatic approaches cause systematic bias. Evolutionary Bioinformatics. 4**,** 1-6.

Messer, P. W., Arndt, P. F., 2007. The majority of recent short DNA insertions in the human genome are tandem duplications. Molecular Biology and Evolution. 24**,** 1190-1197.

Metzgar, D., Bytof, J., Wills, C., 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Research. 10**,** 72-80.

Michael, T. P., Park, S., Kim, T., Booth, J., Byer, A., Sun, Q., Chory, J., Lee, K., 2007. Simple sequence repeats provide a substrate for phenotypic variation in the *Neurospora crassa* circadian clock. PLoS ONE. 2**,** e795.

Miesfeld, R., Krystal, M., Arnheim, N., 1981. A member of a new repeated sequence family which is conserved throughout eucaryotic evolution is found between the human δ and β globin genes. Nucleic Acids Research. 9**,** 5931-5947.

Morgante, M., Hanafey, M., Powell, W., 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nature Genetics. 30**,** 194-200.

Morris, M. J., Wingfield, M. J., de Beer, C., 1993. Gummosis and wilt of *Acacia mearnsii* in South Africa caused by *Ceratocystis fimbriata*. Plant Pathology. 42**,** 814-817.

Nishioka, Y., Leder, P., 1980. Organisation and complete sequence of identical embryonic and plasmacytoma κ V-region genes. The Journal of Biological Chemistry. 255**,** 3691-3694.

Noor, M. A. F., Kliman, R. M., Machado, C. A., 2001. Evolutionary history of microsatellites in the Obscura group of *Drosophila*. Molecular Biology and Evolution. 18**,** 551-556.

Ostrander, E. A., Jong, P. M., Rine, J., Duyk, G., 1992. Constructions of small-insert genomic DNA libraries highly enriched for microsatellite repeat sequences. Proceedings of the National Academy of Sciences of the United States of America. 89**,** 3419-3423.

Pai, A., Sharakhov, I. V., Braginets, O., Costa, C., Yan, G. Y., 2003. Identification of microsatellite markers in the red flour beetle, *Tribolium castaneum*. Molecular Ecology Notes. 3**,** 425-427.

Pâques, F., Haber, J. E., 1999. Multiple pathways of recombination induced by double-strand breaks in *Saccharomyces cerevisiae*. Microbiology and Molecular Biology Reviews. 349-404.

Park, T., Leslie, P. H., Mertz, D. B., 1964. Genetic strains and competition in populations of *Tribolium*. Physiological Zoology. 37**,** 97-162.

Primmer, C. R., Raudsepp, T., Chowdhary, B. P., Moller, A. R., Ellegren, H., 1997. Low frequency of microsatellites in the avian genome. Genome Research. 7**,** 471-482.

Pupko, T., Graur, D., 1998. Evolution of microsatellites in the yeast *Saccharomyces cerevisiae*: Role of length and number of repeated units. Journal of Molecular Evolution. 48**,** 313-316.

Radivojac, P., Iakoucheva, L. M., Oldfield, C. J., Obradovic, Z., Uversky, V. N., Dunker, A. K., 2007. Intrinsic disorder and functional proteomics. Biophysical Journal. 92**,** 1439-1456.

Rassmann, K., Schlötterer, C., Tautz, D., 1991. Isolation of simple-sequence loci for use in polymerase chain reaction-based DNA fingerprinting. Electrophoresis. 12, 113-118.

Ribeiro, I. J. A., Rosetto, C. J., Sabino, J. C., Gallo, P. B., 1986. Seca da mangueira. VIII. Resistincia de porta-enxertos de mangueira ao fungo *Ceratocystis fimbriata* Ell. & Halst. Bragantia. 45**,** 317-322.

Richard, G. F., Paques, F., 2000. Mini- and microsatellite expansions: The recombination connection. Embo Reports. 1**,** 122-126.

Rizatto, S., de Araùjo Batista, C. E., Bajay, M. M., Sigrist, M. S., Ito, M. F., Monteiro, M., Cavallari, M. M., Pinheiro, J. B., Zucchi, M. I., 2010. A new set of microsatellite markers for the genetic characterisation of *Ceratocystis fimbriata*, an economically important plant pathogen. Conservation Genetics Resources. 2**,** 55-58.

Rose, O., Falush, D., 1998. A threshold size for microsatellite expansion. Molecular Biology and Evolution. 15**,** 613-615.

Roth, D. B., Porter, T. N., Wilson, J. H., 1985. Mechanisms of nonhomologous recombination in mammalian cells. Molecular and Cellular Biology. 5**,** 2599-2607.

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

Sagher, D., Hsu, A., Strauss, B., 1999. Stabilization of the intermediate in frameshift mutation. Mutation Research. 423.

Sainudiin, R., Durrett, R. T., Aquadro, C. F., Nielsen, R., 2004. Microsatellite mutation models: Insights from a comparison of humans and chimpanzees. Genetics. 168**,** 383-395.

Santana, Q. C., Coetzee, M. P. A., Steenkamp, E. T., Mlonyeni, O. X., Hammond, G. N. A., Wingfield, M. J., Wingfield, B. D., 2009. Microsatellite discovery by deep sequencing of enriched genomic libraries. Biotechniques. 46**,** 217-223.

Schaffner, W., Kunz, G., Daetwyler, H., Telford, J., Smith, H. O., Birnstiel, M. L., 1978. Genes and spacers of cloned sea urchin histone DNA analysed by sequencing. Cell. 14**,** 655-671.

Schlötterer, C., Ritter, R., Harr, B., Brem, G., 1998. High mutation rate of a long microsatellite allele in *Drosophila melanogaster* provides evidence for allele-specific mutation rates. Molecular Biology and Evolution. 15**,** 1269-1274.

Schlötterer, C., Tautz, D., 1992. Slippage synthesis of simple sequence DNA. Nucleic Acids Research. 20**,** 211-216.

Selkoe, K. A., Toonen, R. J., 2006. Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers. Ecology Letters. 9**,** 615-629.

Shenkin, A., Burdon, A. H., 1972. Deoxyadenylate-rich sequences in mammalian DNA. FEBS Letters. 22**,** 157-160.

Shenkin, A., Burdon, A. H., 1974. Deoxyadenylate-rich and deoxy-guanylate-rich regions in mammalian DNA. Journal of Molecular Biology. 85**,** 19-39.

Sia, E. A., Kokoska, R. J., Dominska, M., Greenwell, P., Petes, T. D., 1997. Microsatellite instability in yeast: Dependence on repeat unit size and DNA mismatch repair genes. Molecular and Cellular Biology. 17**,** 2851-2858.

Slightom, J. L., Blechl, A. E., Smithies, O., 1980. Human fetal $^{G}\gamma$- and $^{A}\gamma$-globin genes: Complete nucleotide sequences suggest that DNA can be exchanged between these duplicated genes. Cell. 21**,** 627-638.

Sokol, K. A., Williams, C. G., 2005. Evolution of a triplet repeat in a conifer. Genome. 48**,** 417-426.

Staib, P., Kretschmar, M., Nichterlein, T., Hof, H., Morschhäuser, J., 2002. Host versus *in vitro* signals and intrastrain allelic differences in the expression of a *Candida albicans* virulence genes. Molecular Microbiology. 44**,** 1351-1366.

Steimel, J., Engelbrecht, C. J. B., Harrington, T. C., 2004. Development and characterisation of microsatellite markers for the fungus *Ceratocystis fimbriata*. Molecular Ecology Notes. 4**,** 215-218.

Sun, J. X., Mullikin, J. C., Patterson, N., Reich, D. E., 2009. Microsatellites are molecular clocks that support accurate inferences about history. Molecular Biology and Evolution. 26**,** 1017-1027.

Sures, I., Lowry, J., Kedes, L. H., 1978. The DNA sequence of sea urchin (*S. purpuratus*) H2A, H2B, and H3 histone coding and spacer regions. Cell. 15**,** 1034-1044.

Tautz, D., 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Research. 17**,** 6463-6471.

Tautz, D., Renz, M., 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Research. 12**,** 4127-4138.

Thiel, T., Michalek, W., Varshney, R. K., Graner, A., 2003. Exploiting EST databases for the development and characterisation of gene-derived SSR-markers in barley (*Hordeum vulgare* L.). Theoretical and Applied Genetics. 106**,** 411-422.

Thurston, M. I., Field, D., 2005. Msatfinder: Detection and characterisation of microsatellites. Distributed by the authors at http://www.genomics.ceh.ac.uk/msatfinder/.

Tóth, G., Gáspári, Z., Jurka, J., 2000. Microsatellites in different eukaryotic genomes: Survey and analysis. Genome Research. 10**,** 967-981.

van Passel, M. W. J., de Graaff, L. H., 2008. Mononucleotide repeats are asymmetrically distributed in fungal genes. BMC Genomics. 9, 596.

van Wyk, M., van der Merwe, N. A., Roux, J., Wingfield, B. D., Kamgan, G. N., Wingfield, M. J., 2006. Population genetic analyses suggest that the *Eucalyptus* fungal pathogen *Ceratocystis fimbriata* has been introduced into South Africa. South African Journal of Science. 102**,** 259-263.

van Wyk, M., Wingfield, B. D., Wingfield, M. J., 2011. Four new *Ceratocystis* spp. associated with wounds on *Eucalyptus*, *Scizolobium* and *Terminalia* trees in Ecuador. Fungal Diversity. 46**,** 111-131.

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.

Vorlickova, M., Sedlacek, P., Kypr, J., Sponar, J., 1982. Conformational transitions of poly(dA-dT).poly(dA-dT) in ethanolic solutions. Nucleic Acids Research. 21**,** 6969-6979.

Weber, J. L., May, P. E., 1989. Abundant class of human DNA polymorphisms which can be typed using the polymerase chain reaction. American Journal of Human Genetics. 44**,** 388-396.

Webster, M. T., Smith, N. G. C., Ellegren, H., 2002. Microsatellite evolution inferred from human-chimpanzee genomic sequence alignments. Proceedings of the National Academy of Sciences of the United States of America. 99**,** 8748-8753.

Weller, P., Jeffreys, A. J., Wilson, V., Blanchetot, A., 1984. Organisation of the human myoglobin gene. The EMBO Journal. 3**,** 439-446.

Wierdl, M., Dominska, M., Petes, T. D., 1997. Microsatellite instability in yeast: Dependence on the length of the microsatellite. Genetics. 146**,** 769-779.

Wingfield, M. J., DeBeer, C., Visser, C., Wingfield, B. D., 1996. A new *Ceratocystis* species defined using morphological and ribosomal DNA sequence comparisons. Systematic and Applied Microbiology. 19**,** 191-202.

Wren, J. D., Forgacs, E., Fondon, J. W., Pertsemlidis, A., Cheng, S. Y., Gallardo, T., Williams, R. S., Shohet, R. V., Minna, J. D., Garner, H. R., 2000. Repeat polymorphisms within gene regions: Phenotypic and evolutionary implications. American Journal of Human Genetics. 67**,** 345-356.

Wyman, A. R., White, R., 1980. A highly polymorphic locus in human DNA. Proceedings of the National Academy of Sciences of the United States of America. 77**,** 6754-6758.

Yarus, M., Folley, L. S., 1985. Sense codons are found in specific contexts. Journal of Molecular Biology. 182**,** 529-540.

Zane, L., Bargelloni, L., Patarnello, T., 2002. Strategies for microsatellite isolation: A review. Molecular Ecology. 11**,** 1-16.

Zhang, L., Yuan, D., Yu, S., Li, Z., Cao, Y., Miao, Z., Qian, H., Tang, K., 2004. Preference of simple sequence repeats in coding and non-coding regions of *Arabidopsis thaliana*. Bioinformatics. 20**,** 1081-1086.

Zhu, Y., Strassmann, J. E., Queller, D. C., 2000. Insertions, substitutions and the origin of microsatellites. Genetic Research. 76**,** 227-236.

Zietkiewicz, E., Rafalski, A., Labuda, D., 1994. Genome fingerprinting by simple sequence repeat (SSR)-anchored polymerase chain reaction amplification. Genomics. 20**,** 178-183.

Zimmer, C., Tymen, S., Marck, C., Guschlbauer, W., 1982. Conformational transitions of poly(dA-dC).poly(dG-dT) induced by high salt or in ethanolic solution. Nucleic Acids Research. 3**,** 1081-1091.

## 6.0 Figures

**Fig. 1** DNA slippage of three tandem repeats during DNA replication. Repeat units can be deleted or duplicated as a result of transient dissociation and subsequent mispairing of single-stranded DNA (Taken from Gemayel et al. 2010).

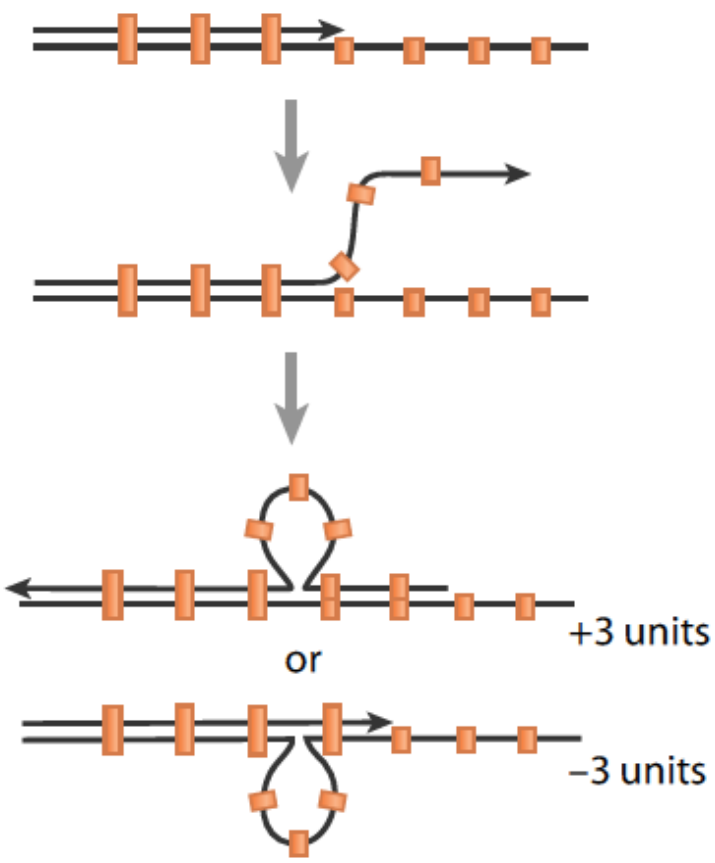


**Fig. 2** Indel slippage induced by non-homologous end joining (NHEJ) repair results in duplications or deletions. **A** After cleavage of double-stranded DNA, micro-homologies can cause misalignment of the strands during ligation and thus NHEJ occurs. This results in tandem duplications or deletions of a few bases (Pâques and Haber, 1999). **B** In the absence of microhomologies, tandem duplications can still occur during ligation due to mispairing. The resulting motif will depend on which direction the mismatch was corrected (Leclercq et al., 2010). **C** Complementary ends can directly be filled in before ligation of the double-stranded ends occur (Roth et al., 1985) which leads to duplication of the cleaved bases (Diagram taken from Leclercq et al., 2010).

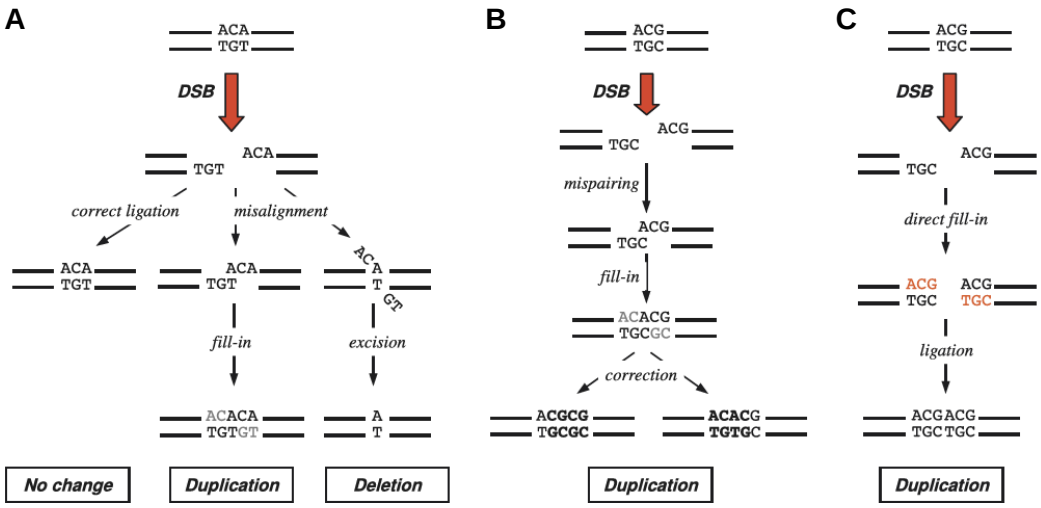


**Fig. 3** Location of tandem repeats within a gene or regulatory sequence in eukaryotic examples. Red lines indicate a tandem repeat at that particular region; parentheses show the tandem repeat motifs. # Two variable microsatellites in the *SAP2* promoter which together affect transcription in *C. albicans* (Staib et al., 2002). * An increase in the repeat number of the CA motif results in increased gene expression in *P. anserina* (Khashnobish et al., 1998). + Variation in the polyglutamine tract of WC-1 in *N. crassa* results in differences of the circadian clock cycle in isolates from different parts of the world (Michael et al., 2007). ** CCTG expansion in intron 1 of ZNF9 causes Myotonic dystrophy type 2 (Liquori et al., 2001). ## Longer mRNA products are produced by transcription slippage over the CAG/CTG repeats in *S. cerevisiae* (Fabre et al., 2002). (Diagram modified from Gemayel et al., 2010).

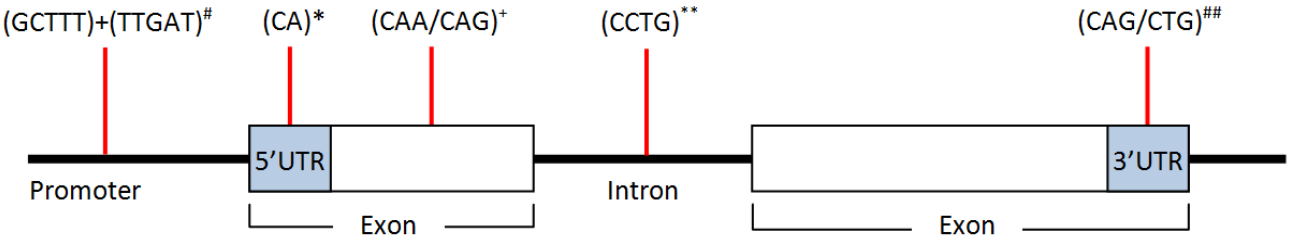


**Fig. 4** Changes in skull morphology over time of three different dog breeds that is correlated to variability in the number of tandem repeats in the *Runx-2* gene. **A** Purebred St. Bernard skulls from the year 1850 (Top), 1921 (Middle), and 1967 (Bottom). **B** Purebred bull terrier skulls from the year 1931 (Top), 1950 (Middle), and 1976 (Bottom) (24). **C** Purebred Newfoundland skulls from the year 1926 (Top), 1964 (Middle), and 1971 (Bottom). (Photo taken from Fondon and Garner, 2004).

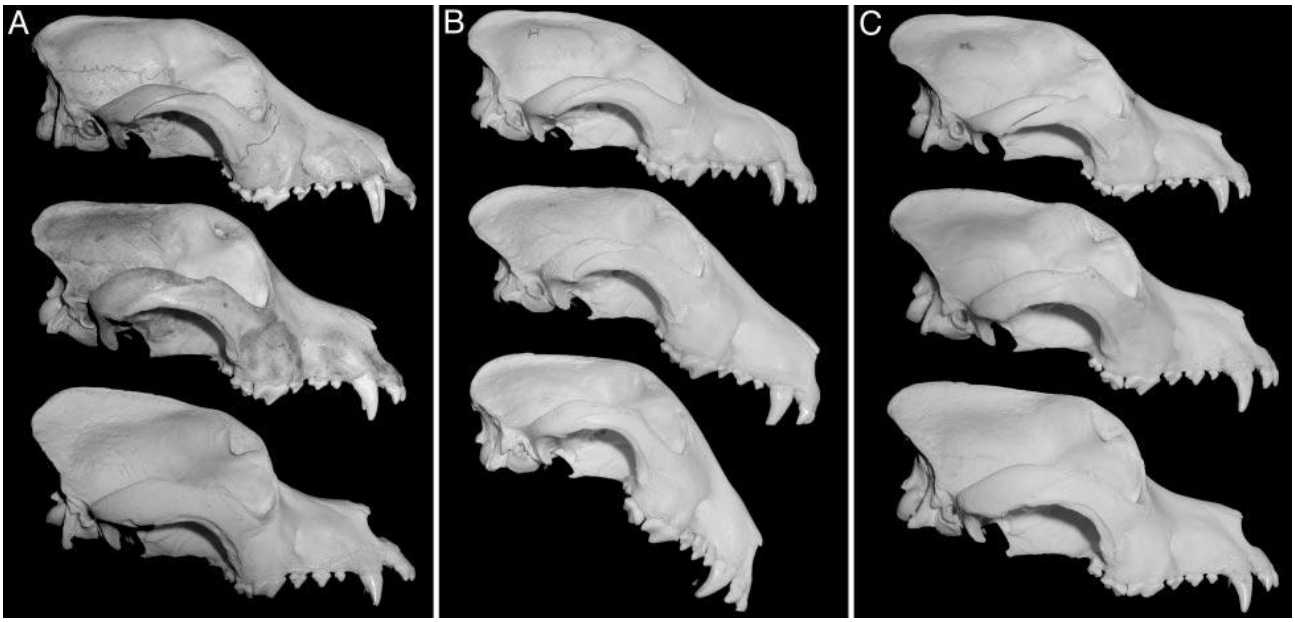
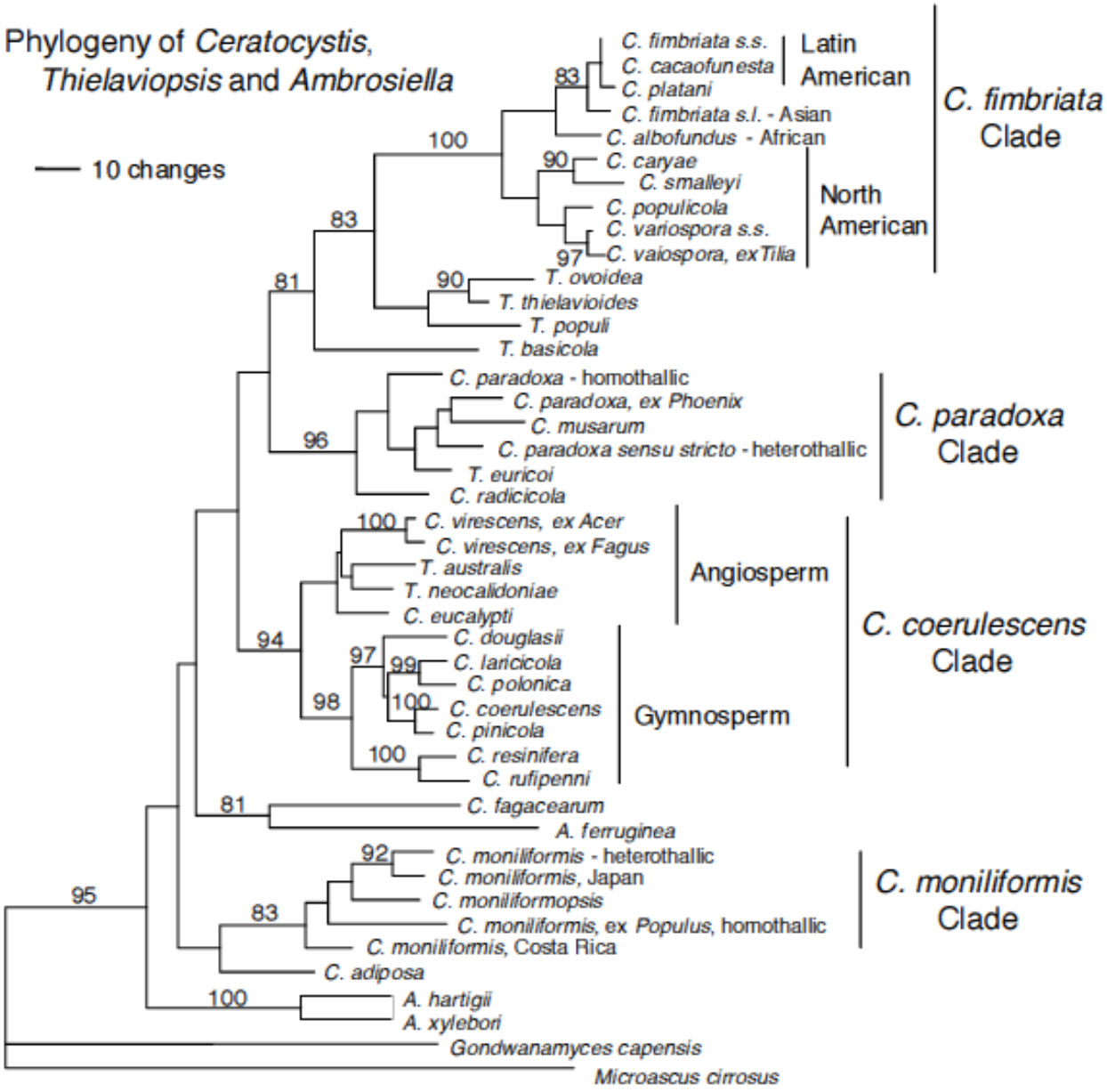


**Fig. 5** Phylogenetic tree based on DNA sequences of elongation factor 1-α for *Ceratocystis*, *Thielaviopsis* and *Ambrosiella* species. *Microascus cirrosus* is the outgroup taxon. Note the clustering of the species complexes (Phylogenetic tree taken from Harrington, 2007).



# Chapter 2

# Distribution and abundance of microsatellites in the genome of the plant pathogen *Ceratocystis fimbriata sensu stricto*

**Abstract**

*Ceratocystis fimbriata* is a fungal pathogen of important plants in the forestry and agricultural industries and represents a complex of cryptic species. Microsatellite markers have been developed for species within the complex and used to study population structure and origin of these species. Sequencing the *C. fimbriata* genome allowed a bioinformatic search of microsatellites to be carried out. The abundance and distribution of microsatellites were analysed and compared to other fungi on which similar studies have been performed. *C. fimbriata* has a medium microsatellite density even though it has a larger genome size than the other fungi it was compared to. With a GC content of almost 50%, this genome displays a more even distribution of microsatellites compared to fungi that have AT-rich genomes. The distributions of motifs within each microsatellite class are unique to *C. fimbriata*, as has been found in other eukaryotes. The coding regions consist mainly of trinucleotides that encode the amino acids alanine and glutamine. This is the first report of a microsatellite analysis performed on a whole genome sequence of a fungus within the order Microascales, to which *Ceratocystis* belongs. The microsatellites identified in this study could contribute to further population studies as well as potentially differentiating between species in this complex.

## Introduction

Species of *Ceratocystis* are insect-associated fungi that include saprophytes, plant pathogens and species that cause blue stain in cut timber (Wingfield et al., 1993). *Ceratocystis fimbriata* represents a sub-group of these fungi that are transmitted by casual insects such as nitidulid beetles and most are serious pathogens of many plants including trees and root crops (Kile, 1993). Biological and phylogenetic studies based on DNA sequence comparisons have increasingly shown that *Ceratocystis fimbriata* represents a complex of species, some of which are host specific while others appear to have broad host ranges (Barnes et al., 2003; Johnson et al., 2005; van Wyk et al., 2011).

Microsatellite markers have been useful in studies on *Ceratocystis*. They have, for example, provided insight into the population structure and origin of some of these fungi (Barnes et al., 2005; Engelbrecht et al., 2007; Engelbrecht et al., 2004; Ferreira et al., 2010; Kamgan Nkuekam et al., 2009; Ocasio-Morales et al., 2007). It has even been shown that microsatellites can differentiate between isolates from different geographical regions and hosts (Barnes et al., 2001). However, nothing is known regarding their abundance or distribution in the genomes of these fungi or even in the order Microascales that accommodates *Ceratocystis*.

Microsatellites are 1-6 bp tandem repeats that are abundant throughout eukaryotic and prokaryotic genomes (Field and Wills, 1996; Tautz and Renz, 1984). They make ideal molecular markers because they have a high level of polymorphism, are inherited in a Mendelian manner and are easy to amplify with PCR (Levinson and Gutman, 1987; Tautz, 1989). Consequently, they have been used for strain typing, genetic mapping and population structure studies in many different organisms (Field and Wills, 1996; Hennequin et al., 2001; Jarne and Lagoda, 1996). However, *de novo* isolation of microsatellites has presented a major drawback to their use in the past. Estimating the abundance of microsatellites in genomes has also presented challenges because hybridisation experiments using repeat-containing probes were difficult to perform and could sometimes be inefficient (Tautz and Renz, 1984; Zane et al., 2002). More recently it has been possible to search sequence databases and whole genome sequences to estimate microsatellite distribution and abundance and to further develop informative markers for population studies (Demuth et al., 2007; Drury et al., 2009; Richard and Dujon, 1996; Tóth et al., 2000).

Genome-wide searches for microsatellites have been performed on a number of eukaryotic and fungal genomes, including *Arabidopsis thaliana*, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae*, *Neurospora crassa* and *Fusarium graminearum* (Karaoglu et al., 2005; Katti et al., 2001; Lim et al., 2004; Tóth et al., 2000). The consensus from these studies showed that fungal genomes contain fewer microsatellites than other higher eukaryotes. Each organism also has a unique microsatellite distribution and shows a preference for certain types of microsatellite motifs (Morgante et al., 2002; Tóth et al., 2000). It is only in the last 10 years that there has been an increase in studies dealing with microsatellites and their distribution and evolution in the genomes of fungi (Karaoglu et al., 2005; Lim et al., 2004; Tóth et al., 2000). Prior to this period, most work was focused on the yeast species, *S. cerevisiae* (Field and Wills, 1998; Richard and Dujon, 1996; Sia et al., 1997; Wierdl et al., 1997).

The aim of this study was to determine the distribution and abundance of microsatellites in the genome sequence of *C. fimbriata* using a bioinformatics approach. An additional objective was to compare the microsatellite structure of the *C. fimbriata* genome with other Ascomycete genomes to determine their similarities and differences.

**Methods and Materials**

*Genome sequence and GC content*

A fully sequenced genome of *C. fimbriata* (CMW 14799) was used for analysis in this study. This genome is 47.7 Mb in size, and was assembled using Newbler (PM Wilken, Personal Communication). The sequence statistics function in CLC Genomics Workbench (CLC bio, Aarhus, Denmark) was selected to produce a table of the nucleotide content of each contig. These data were then exported to Microsoft Office Excel 2007 (Microsoft Corp., Redmond, WA, USA) and the equation GC content = (G+C)/(A+T+G+C) was used to calculate the overall GC content of all the contigs making up the genome sequence of *C. fimbriata.*

*Microsatellite discovery*

Sequence files of the assembled *C.fimbriata* genome were searched for microsatellite repeats using the web interface of MSatFinder (Thurston and Field, 2005). A regex-directed search engine was used to identify sequences containing perfect microsatellites. A perfect microsatellite is a tract consisting of exact copies of the repeat unit, e.g. $(CTA)_6$, and doesn't

contain any mismatches or interruptions. The minimum repeat number for detecting mononucleotides was 12, while a five repeat minimum was used for detecting di-, tri-, tetra-, penta- and hexanucleotides. MSatFinder generated tab-delimited files that were converted for analysis in Microsoft Office Excel 2007 (Microsoft Corp., Redmond, WA, USA). The microsatellites were sorted in Excel according to the type of motif and repeat length. The sequence lengths of each motif, the number of each type of motif and the total repeats per Mb of sequence analysed were calculated. In addition, the percentage of each type of microsatellite in the genome and within each group of microsatellite motifs was analysed.

*Microsatellites in coding regions*

The online interface of the *de novo* prediction program AUGUSTUS (Keller et al., 2011) was used to predict all genes in the genome. For this purpose the dataset from *Fusarium graminearum* was used as the reference annotated genome. The fasta output file containing the predictions for coding genes was then searched for microsatellites and analysed using MSatFinder with the same parameters as above.

*Comparisons between fungal genomes*

Comparisons of the microsatellite content in the *C. fimbriata* genome sequence were made with the Ascomycetes *Aspergillus nidulans*, *F. graminearum*, *Magnaporthe grisea*, *N. crassa*, *S. cerevisiae* and *Schizosaccharomyces pombe*. Genome size and results of microsatellite abundance in each of the fungal genomes were obtained from the study by Karaoglu et al. (2005). In addition, the GC contents of these genomes were obtained from Lim et al. (2004).

**Results**

*GC content and abundance of microsatellites*

A total of 6738 perfect microsatellites were identified from the genome of *C. fimbriata* using MSatFinder. The GC content was 48.10% and the microsatellite density was calculated as one microsatellite every 7.1 kb. Not all of the contigs contained microsatellites, whereas some of the contigs contained more than one microsatellite. In total, 1525 of 3703 (41%) contigs contained microsatellites. The contig containing the most microsatellites was 1001 bp in size, the largest contig was 134092 bp and contained 17 microsatellites while the smallest contig containing a single microsatellite was 101 bp in length.

The majority of microsatellites found in the *C. fimbriata* genome were mononucleotides, followed by dinucleotides, trinucleotides, tetranucleotides, pentanucleotides and hexanucleotides (Table 1). The most common mononucleotide motif was $T_n$, followed by $A_n$ and $(AG/GA)_n$, $(CT/TC)_n$ and $(AT/TA)_n$ were most commonly found dinucleotides. The most common trinucleotide motifs included $(ACG)_n$, $(CGT)_n$ and $(AAG)_n$, while $(TGCA)_n$ and $(ATAC)_n$ were the most abundant tetranucleotide motifs. The most abundant pentanucleotides were $(CAGCA)_n$ and $(TGTTT)_n$, and the most common hexanucleotide motif was $(TCTCTG)_n$. All types of mono-, di- and trinucleotide motifs were present but only a portion of tetra-, penta- and hexanucleotide motifs were found (for example, $(CAACAT)_n$ is not present in the genome). The trinucleotides $(CAT)_5$ and $(CAC)_5$ that have been used as DNA fingerprinting probes for a number of fungi, including species of *Ceratocystis* were found four and eight times respectively in the *C. fimbriata* genome.

*Abundance of microsatellites in coding regions*

The *C. fimbriata* genome contains 8809 genes predicted by AUGUSTUS. MSatFinder identified 739 in the predicted coding regions of the genes. The majority of microsatellites were trinucleotides, followed by dinucleotides and hexanucleotides (Table 2). The most common microsatellites present in the predicted coding regions were $(ACG)_n$, $(AAC)_n$, $(CT/TC)_n$, $(AAG)_n$ and $(CGT)_n$. The most abundant hexanucleotide motif $(CAGGCT)_n$ was found only four times. Very few mononucleotides were found in the coding regions and mainly consist of $C_n$ which was found 15 times. The only other mononucleotide found was $(G)_{12}$. Only one pentanucleotide, $(GACAG)_{18}$, was found, while no tetranucleotides were identified in the coding regions.

In the *C. fimbriata* genome, the two most abundant trinucleotide motifs, $(CAG)_n$ and $(CAA)_n$ code for the amino acid glutamine while the third most abundant trinucleotide $(GCA)_n$ codes for alanine (Table 3). The most abundant hexanucleotides $(CAGGCT)_n$, $(CAGGCA)_n$ and $(GCTCAA)_n$ code for glutamine/alanine tracts and $(CAGCAA)_n$ encodes glutamine. Other amino acids that were encoded fairly abundantly in the genome were lysine, serine and threonine.

*Longest microsatellites*

The longest mononucleotide in the *C. fimbriata* genome was $(G)_{62}$ (Table 5) and on the whole, mononucleotide motifs had mostly between 12 – 18 repeats. For dinucleotides, the

longest motifs were $(GA)_{41}$ and $(AG)_{40}$, followed by $(AC)_{33}$ and $(AC)_{32}$. The longest trinucleotides were $(AAG)_{19}$ and $(CTA)_{18}$ and the longest tetranucleotides were $(TCAC)_{15}$, $(CTGC)_{11}$ and $(TACA)_{11}$. The motif $(GACAG)_{18}$ was the longest pentanucleotide and microsatellite in the whole genome, followed by $(GACAG)_{12}$ and $(CACAG)_{12}$. The longest hexanucleotides were $(GAAAAT)_{14}$, $(TCTCTG)_{13}$ and $(TGCTGT)_{13}$. Generally, di-, tri-, tetra-, penta- and hexanucleotide motifs were repeated five to eight times in the genome. In coding regions, the longest microsatellite was the pentanucleotide $(GACAG)_{18}$ followed by the hexanucleotides $(AGAGAC)_{13}$, $(CAACAG)_{13}$, $(CAGGCT)_{11}$ and $(GACAGA)_{11}$. The longest trinucleotide was $(AAG)_{19}$, the longest dinucleotide was $(AC)_{12}$ and the longest mononucleotide was $(C)_{14}$.

*Comparisons between fungal genomes*

When compared to other Ascomycetes, the *C. fimbriata* genome is larger, shares an almost 50% GC content with some of the other Ascomycetes and has a microsatellite density of one microsatellite every 7.1 kb (Table 4). The most abundant microsatellites in *C. fimbriata* are mononucleotides followed by di-, tri-, tetra-, penta- and hexanucleotides, which is similar to some other Ascomycetes (Table 5). The longest microsatellites in *C. fimbriata* are shorter than the longest microsatellites found in other Ascomycetes (Table 6).

**Discussion**

This study is the first to characterise microsatellites in *C. fimbriata* and a member of the order Microascales. Results from this study showed that microsatellites are fairly abundant in the genome of *C. fimbriata* and compare well with analyses from other Ascomycetes. It is important to recognise that search tools using different algorithms can sometimes produce significantly different results with the same data set (Merkel and Gemmell, 2008). However, results in this study support those of previous studies showing that genome size does not correlate with microsatellite density (Karaoglu et al., 2005; Lim et al., 2004; Tóth et al., 2000). We also found the frequency of a particular motif decreases with an increase in the repeat number of that motif, as was reported for other fungi and eukaryotes (Harr et al., 2002; Karaoglu et al., 2005; Katti et al., 2001; Tóth et al., 2000; Wierdl et al., 1997). Fungal genomes contain fewer microsatellites than other eukaryotic genomes; in this regard, *C. fimbriata* was no exception (Karaoglu et al., 2005; Lim et al., 2004). This is most likely due to the presence of fewer non-coding regions in fungal genomes, rather than a smaller genome size (Karaoglu et al., 2005).

The genome of *C. fimbriata* is larger than those of many other Ascomycetes that have been considered for whole genome microsatellite analyses (Karaoglu et al., 2005; Lim et al., 2004). However, it has a medium microsatellite density in comparison to the other fungi. This phenomenon is not without precedent as other studies have shown that genome size does not correlate with microsatellite density (Karaoglu et al., 2005; Lim et al., 2004; Tóth et al., 2000). For example, in the Ascomycetes *N. crassa* and *F. graminearum*, the former fungus has a genome size of 38 Mb and a density of one microsatellite every 2.7 kb, whereas the latter has a genome size of 36.1 Mb has a density of one microsatellite every 12.5 kb (Karaoglu et al., 2005).

Different factors could contribute to the variation in microsatellite density observed in the different fungal species above. It could be as a result of differences in genome organisation and efficiency of the mismatch repair machinery, and other factors that control variation in the abundance of microsatellites (Karaoglu et al., 2005; Tóth et al., 2000). Another factor that can affect microsatellite density is GC content. Lim et al. (2004) showed that fungal genomes with a higher abundance of microsatellites have a lower GC content, whereas those with a GC content of 50% show a more equal distribution of microsatellites. This is clearly seen with the yeasts, *S. cerevisiae* and *Sch. Pombe*, which have a low GC content and a high microsatellite density, with mononucleotides being the most abundant microsatellite class (Karaoglu et al., 2005; Lim et al., 2004). The *C. fimbriata* genome demonstrates this characteristic with an almost 50% GC content and a more equal microsatellite distribution.

With regards to microsatellite distribution, every genome is unique and this is also true for particular motifs within each microsatellite class. In the *C. fimbriata* genome, the most common mononucleotide and microsatellite is T, followed by A. In other fungal genomes a similar pattern is seen, where A and T mononucleotides dominate (Karaoglu et al., 2005; Lim et al., 2004). The three most common dinucleotides in other fungal genomes are AT/TA, AG/GA and CT/TC (Karaoglu et al., 2005; Lim et al., 2004), which is also the case in the *C. fimbriata* genome. The repeats CG/GC are uncommon in most fungal genomes (Karaoglu et al., 2005; Lim et al., 2004), and this is also reflected in the *C. fimbriata* genome. Trinucleotides represent the third most abundant microsatellite class in the *C. fimbriata* genome. Similarly this motif is abundant in other fungal genomes, such as *A. nidulans* and *F. graminearum* (Karaoglu et al., 2005; Lim et al., 2004). Tetra-, penta- and hexanucleotides are the least abundant microsatellites in the *C. fimbriata* genome. As with other fungal genomes, there is a preference for different motifs in each of these classes of microsatellites

(Karaoglu et al., 2005; Lim et al., 2004). For example, *M. grisea* contains mostly CG-rich trinucleotides which is probably a result of its slightly higher than 50% GC content, and *S. cerevisiae* and *Sch. pombe* contain AT-rich trinucleotides, which coincides with their AT-rich genome (Karaoglu et al., 2005; Lim et al., 2004).

Tri- and hexanucleotides are expected to be the most abundant microsatellites in coding regions as they would not change the reading frame of the coding region (Metzgar et al., 2000). Similarly, other motifs whose total length is a multiple of three are also expected to be present (Gibbons and Rokas, 2009; Metzgar et al., 2000). It was, therefore, surprising that no tetranucleotides were found in the coding regions of this genome and that the longest mononucleotide, $C_{14}$, has a length of 14 bp and was found four times. The single pentanucleotide, $GACAG_{18}$, found in the coding regions of the genome has a length that is a multiple of three and therefore is tolerated in the reading frame. Generally, pentanucleotides are found almost exclusively in non-coding regions of eukaryotes (Toth et al., 2000). Most of the microsatellites (81%) in coding regions have lengths that are a multiple of three, which correlates well to the most abundant microsatellite class - the trinucleotides. However, the second most abundant class, dinucleotides, did not occur in lengths a multiple of three. These microsatellites still seem to be stable but any increase or decrease in the number of repeats may not be tolerated as changes in the resulting protein could adversely affect the growth and development of the fungus.

Tri- and hexanucleotides within coding regions are likely to code for amino acids and often produce homopolymeric tracts in the resulting proteins (Kashi and King, 2006). In the *C. fimbriata* genome, the most abundant trinucleotides encoded glutamine and alanine, followed by lysine, threonine and serine. The most abundant hexanucleotides encoded glutamine or glutamine/alanine tracts. Trinucleotides within coding regions of higher eukaryotes often encode the amino acids serine, asparagine, glutamine, proline and threonine (Katti et al., 2001; Tóth et al., 2000). In yeast genomes the most abundant trinucleotides encode glutamine, aspartic acid, asparagine and glutamic acid (Katti et al., 2001; Malpertuy et al., 2003), while other fungi show preferences for asparagine, serine, threonine and lysine (Li et al., 2009; Tóth et al., 2000). The trinucleotides encoding amino acids in *C. fimbriata* are thus similar to those previously found for fungi and other eukaryotes.

The trinucleotide fingerprinting probes $(CAT)_5$ and $(CAC)_5$ have been used previously to distinguish between different fungal species. In the fungal genus *Ceratocystis*, species from different geographical locations as well as those with different reproductive strategies could

be differentiated from one another using this method (DeScenzo and Harrington, 1994; Engelbrecht et al., 2007; Harrington et al., 1998). These particular motifs were, in this study, not abundant in the genome of *C. fimbriata*. Analyses of the $(CAT)_5$ DNA fingerprinting gels of *C. cacaofunesta*, *C. virescens* and *C. eucalypti* in previous studies showed that there are more than 10 bands present for $(CAT)_5$ and $(CAC)_5$ (Engelbrecht et al., 2007; Harrington et al., 1998), which is more than expected from the number of motifs that are found in the *C. fimbriata* genome. This could be due to the probe binding less specifically than expected and might have included mismatches in the microsatellite, e.g. CATCAGCATCATCAT. As $(CAT)_5$ and $(CAC)_5$ are rare in the genome they can be used successfully as diagnostic markers.

The *C. fimbriata* genome is larger than other Ascomycete genomes for which similar studies on microsatellites have been carried out; however this genome is smaller compared to other Ascomycetes in general, e.g. some strains of *Fusarium oxysporum* (Ma et al., 2010). The longest microsatellites in this fungus are shorter than the longest microsatellites in other Ascomycetes. This could mean that microsatellites in the *C. fimbriata* genome are imperfect and are therefore no longer uninterrupted tracts. Generally, microsatellites in fungi are short, which is also seen in *C. fimbriata* with most motifs repeated five to eight times. Interestingly, the longest mononucleotide consists of a G motif, which is in contrast to other Ascomycetes, such as *F. graminearum* and *A. nidulans* where the T motif is the longest mononucleotide. The trinucleotide $(AAG)_{19}$ is the longest microsatellite in coding regions, which is expected, and is only found once within the genome. As observed in other genome studies, the occurrence of a particular motif decreases as the repeat number of that motif increases (Karaoglu et al., 2005; Lim et al., 2004; Tóth et al., 2000).

*Conclusions*

Searching whole genome sequences for microsatellites is not only beneficial for analysing abundance and distribution but the microsatellites identified can be used further in studies on the evolution of microsatellites, genome organisation as well as for the development of molecular markers. Microsatellites have already been used successfully for population studies within the species complex *Ceratocystis fimbriata sensu lato* (Barnes et al., 2001; Engelbrecht et al., 2007). Some of these microsatellites were able to distinguish between different isolates based on their location and specific host (Barnes et al., 2001). This demonstrates that microsatellite markers potentially have the power to differentiate between the morphologically similar species in this complex. We found that the *C. fimbriata* genome

contains over 6000 perfect microsatellites and some of these will likely be useful for species recognition and for more robust population genetic studies.

**References**

Barnes, I., Gaur, A., Burgess, T., Roux, J., Wingfield, B. D., Wingfield, M. J., 2001. Microsatellite markers reflect intra-specific relationships between isolates of the vascular wilt pathogen *Ceratocystis fimbriata*. Molecular Plant Pathology. 2**,** 319.

Barnes, I., Nakabonge, G., Roux, J., Wingfield, B. D., Wingfield, M. J., 2005. Comparison of populations of the wilt pathogen *Ceratocystis albifundus* in South Africa and Uganda. Plant Pathology. 54**,** 189-195.

Barnes, I., Roux, J., Wingfield, B. D., Dudzinski, M. J., Old, M. N., Wingfield, M. J., 2003. *Ceratocystis pirilliformis*, a new species from *Eucalyptus nitens* in Australia. Mycologia. 95**,** 865-871.

Demuth, J. P., Drury, D. W., Peters, M. L., van Dyken, D., Priest, N. K., Wade, M. J., 2007. Genome-wide survey of *Tribolium castaneum* microsatellites and description of 509 polymorphic markers. Molecular Ecology Notes. 7**,** 1189-1195.

DeScenzo, R. A., Harrington, T. C., 1994. Use of $(CAT)_5$ as a DNA fingerpringting probe for fungi. Phytopathology. 84**,** 534-540.

Drury, D. W., Siniard, A. L., Wade, M. J., 2009. Genetic differentiation among wild populations of *Tribolium castaneum* estimated using microsatellite markers. Journal of Heredity. Doi: 10.1093/jhered/esp77.

Engelbrecht, C. J. B., Harrington, T. C., Alfenas, A. C., Suarez, C., 2007. Genetic variations in populations of the cacao wilt pathogen, *Ceratocystis cacaofunesta*. Plant Pathology. 56**,** 923-933.

Engelbrecht, C. J. B., Harrington, T. C., Steimel, J., Capretti, P., 2004. Genetic variation in eastern North American and putatively introduced populations of *Ceratocystis fimbriata f. platani*. Molecular Ecology. 13**,** 2995-3005.

Ferreira, E. M., Harrington, T. C., Thorpe, D. J., Alfenas, A. C., 2010. Genetic diversity and interfertility among highly differentiated populations of *Ceratocystis fimbriata* in Brazil. Plant Pathology. 59, 721-735.

Field, D., Wills, C., 1996. Long, polymorphic microsatellites in simple organisms. Proceedings of the Royal Society of London: Biological Sciences. 263**,** 209-215.

Field, D., Wills, C., 1998. Abundant microsatellite polymorphism in *Saccharomyces cerevisiae*, and the different distributions of microsatellites in eight prokaryotes and *S. cerevisiae*, result from strong mutation pressures and a variety of selective forces.

Proceedings of the National Academy of Sciences of the United States of America. 95**,** 1647-1652.

Gibbons, J. G., Rokas, A., 2009. Comparative and functional characterisation of intragenic tandem repeats in 10 *Aspergillus* genomes. Molecular Biology and Evolution. 26**,** 591-602.

Harr, B., Todorova, J., Schlötterer, C., 2002. Mismatch repair-driven mutational bias in *D. melanogaster*. Molecular Cell. 10**,** 199-205.

Harrington, T. C., Steimel, J., Kile, G., 1998. Genetic variation in three *Ceratocystis* species with outcrossing, selfing and asexual reproductive strategies. European Journal of Forest Pathology. 28**,** 217-226.

Hennequin, C., Thierry, A., Richard, G. F., Lecointre, G., Nguyen, H. V., Gaillardin, C., Dujon, B., 2001. Microsatellite typing as a new tool for identification of *Saccharomyces cerevisiae* strains. Journal of Clinical Microbiology. 39**,** 551-559.

Jarne, P., Lagoda, P. J. L., 1996. Microsatellites, from molecules to populations and back. Trends in Ecology & Evolution. 11**,** 424-429.

Johnson, J. A., Harrington, T. C., Engelbrecht, C. J. B., 2005. Phylogeny and taxonomy of the North American clade of the *Ceratocystis fimbriata* complex. Mycologia. 97**,** 1067-1092.

Kamgan Nkuekam, G., Barnes, I., Wingfield, M. J., Roux, J., 2009. Distribution and population diversity of *Ceratocystis pirilliformis* in South Africa. Mycologia. 101**,** 17-25.

Karaoglu, H., Lee, C. M. Y., Meyer, W., 2005. Survey of simple sequence repeats in completed fungal genomes. Molecular Biology and Evolution. 22**,** 639-649.

Kashi, Y., King, D. G., 2006. Simple sequence repeats as advantageous mutators in evolution. Trends in Genetics. 22**,** 253-259.

Katti, M. V., Ranjekar, P. K., Gupta, V. S., 2001. Differential distribution of simple sequence repeats in eukaryotic genome sequences. Molecular Biology and Evolution. 18**,** 1161-1167.

Keller, O., Kollmar, M., Stanke, M., Waak, S., 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. In print.

Kile, G. A., Plant diseases caused by species of *Ceratocystis sensu stricto* and *Chalara*. In: M. J. Wingfield, et al., Eds.), *Ceratocystis* and *Ophiostoma*: Taxonomy, ecology and pathogenicity. APS Press, St. Paul, Minnesota, USA, 1993, pp. 173-183.

Levinson, G., Gutman, G. A., 1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. Molecular Biology and Evolution. 4**,** 203-221.

Li, C., Liu, L., Yang, J., Li, J., Su, Y., Zhang, Y., Wang, Y., Zhu, Y., 2009. Genome-wide analysis of microsatellite sequence in seven filamentous fungi. Interdisciplinary Sciences and Computer Life Sciences. 1**,** 141-150.

Lim, S., Notley-McRobb, L., Lim, M., Carter, D. A., 2004. A comparison of the nature and abundance of microsatellites in 14 fungal genomes. Fungal Genetics and Biology. 41**,** 1025-1036.

Ma, L., van der Does, H.C., Borkovich, K.A., Coleman, J.J., Daboussi, M., Di Pietro, A., et al., 2010. Nature. 464, 367-373.

Malpertuy, A., Dujon, B., Richard, G., 2003. Analysis of microsatellites in 13 hemiascomycetous yeast species: Mechanisms involved in genome dynamics. Journal of Molecular Evolution. 56**,** 730-741.

Merkel, A., Gemmell, N. J., 2008. Detecting microsatellites in genome data: Variance in definitions and bioinformatic approaches cause systematic bias. Evolutionary Bioinformatics. 4**,** 1-6.

Metzgar, D., Bytof, J., Wills, C., 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Research. 10**,** 72-80.

Morgante, M., Hanafey, M., Powell, W., 2002. Microsatellites are preferentially associated with nonrepetitive DNA in plant genomes. Nature Genetics. 30**,** 194-200.

Ocasio-Morales, R. G., Tsopelas, P., Harrington, T. C., 2007. Origin of *Ceratocystis platani* on native *Platanus orientalis* in Greece and its impact on natural forests. Plant Disease. 91**,** 901-904.

Richard, G. F., Dujon, B., 1996. Distribution and variability of trinucleotide repeats in the genome of the yeast *Saccharomyces cerevisiae*. Gene. 174**,** 165-174.

Sia, E. A., Kokoska, R. J., Dominska, M., Greenwell, P., Petes, T. D., 1997. Microsatellite instability in yeast: Dependence on repeat unit size and DNA mismatch repair genes. Molecular and Cellular Biology. 17**,** 2851-2858.

Tautz, D., 1989. Hypervariability of simple sequences as a general source for polymorphic DNA markers. Nucleic Acids Research. 17**,** 6463-6471.

Tautz, D., Renz, M., 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Research. 12**,** 4127-4138.

Thurston, M. I., Field, D., Msatfinder: detection and characterisation of microsatellites. CEH Oxford, Mansfield Road, Oxford OX1 3SR, 2005, pp. Distributed by the authors at http://www.genomics.ceh.ac.uk/msatfinder/.

Tóth, G., Gáspári, Z., Jurka, J., 2000. Microsatellites in different eukaryotic genomes: Survey and analysis. Genome Research. 10**,** 967-981.

van Wyk, M., Wingfield, B. D., Wingfield, M. J., 2011. Four new *Ceratocystis* spp. associated with wounds on *Eucalyptus*, *Scizolobium* and *Terminalia* trees in Ecuador. Fungal Diversity. 46**,** 111-131.

Wierdl, M., Dominska, M., Petes, T. D., 1997. Microsatellite instability in yeast: Dependence on the length of the microsatellite. Genetics. 146**,** 769-779.

Wingfield, M. J., Seifert, K. A., Webber, J. A., 1993. *Ceratocystis* and *Ophiostoma*: Taxonomy, Ecology and Pathogenicity. St Paul, MN, USA: APS Press.

Zane, L., Bargelloni, L., Patarnello, T., 2002. Strategies for microsatellite isolation: A review. Molecular Ecology. 11**,** 1-16.

**Table 1** The most abundant microsatellite motifs, their density and total number in the *C. fimbriata* genome

| Microsatellite class | Total motifs (% of all microsatellites) | Total/Mb | Abundant motifs | Number of motifs | % of microsatellite class | % of all microsatellites |
|---|---|---|---|---|---|---|
| Mononucleotides | 2995 (44.45%) | 63 | T | 823 | 27.48 | 12.21 |
| | | | A | 770 | 25.71 | 11.43 |
| | | | G | 713 | 23.81 | 10.58 |
| | | | C | 689 | 23.01 | 10.23 |
| Dinucleotides | 2340 (34.73%) | 49 | AG/GA | 565 | 24.14 | 8.39 |
| | | | CT/TC | 521 | 22.26 | 7.73 |
| | | | AT/TA | 474 | 20.26 | 7.03 |
| | | | AC/CA | 390 | 16.67 | 5.79 |
| | | | GT/TG | 362 | 15.47 | 5.37 |
| | | | CG/GC | 28 | 1.20 | 0.42 |
| Trinucleotides | 984 (14.60%) | 21 | ACG/AGC/CAG/CGA/GAC/GCA | 167 | 16.97 | 2.48 |
| | | | CGT/CTG/GCT/GTC/TCG/TGC | 125 | 12.7 | 1.86 |
| | | | AAG/AGA/GAA | 93 | 9.45 | 1.38 |
| | | | CTT/TCT/TTC | 85 | 8.64 | 1.26 |
| Tetranucleotides | 208 (3.09%) | 4 | TGTA | 9 | 4.33 | 0.13 |
| | | | ATAC | 8 | 3.85 | 0.12 |
| | | | CCAG | 6 | 2.88 | 0.09 |
| | | | TACA | 6 | 2.88 | 0.09 |
| | | | AGAC | 5 | 2.40 | 0.07 |
| | | | GCTG | 5 | 2.40 | 0.07 |
| Pentanucleotides | 110 (1.63%) | 2 | CAGCA | 3 | 2.73 | 0.04 |
| | | | TGTTT | 3 | 2.73 | 0.04 |
| Hexanucleotides | 101 (1.50%) | 2 | TCTCTG | 3 | 2.97 | 0.04 |
| Total of all microsatellites | 6738 (100%) | 141 | - | - | - | - |

**Table 2** Total number of microsatellites and most abundant motifs in coding regions of the *C. fimbriata* genome

| Microsatellite class | Total motifs (% of all microsatellites) | Abundant motifs | Number of motifs | % of microsatellite class | % of all microsatellites |
|---|---|---|---|---|---|
| Mononucleotides | 16 (2.17%) | C | 15 | 93.75 | 2.03 |
| | | G | 1 | 6.25 | 0.14 |
| Dinucleotides | 154 (20.64%) | CT/TC | 50 | 32.47 | 6.77 |
| | | AC/CA | 43 | 27.92 | 5.82 |
| | | AG/GA | 36 | 23.38 | 4.67 |
| | | CG/GC | 14 | 9.09 | 1.89 |
| | | GT/TG | 9 | 5.84 | 1.22 |
| | | AT/TA | 2 | 1.30 | 0.27 |
| Trinucleotides | 511 (69.15%) | ACG/AGC/CAG/CGA/GAC/GCA | 181 | 35.42 | 24.49 |
| | | AAC/ACA/CAA | 89 | 17.42 | 12.04 |
| | | AAG/AGA/GAA | 47 | 9.20 | 6.36 |
| | | CTG/GCT/TCG/TGC | 46 | 9.00 | 6.22 |
| Tetranucleotides | 0 | - | - | - | - |
| Pentanucleotides | 1 (0.13%) | GACAG | 1 | 100.00 | 0.13 |
| Hexanucleotides | 57 (7.71%) | CAGGCT | 4 | 7.02 | 0.55 |
| | | CAGCAA | 3 | 5.26 | 0.41 |
| | | CAGGCA | 3 | 5.26 | 0.41 |
| | | GCTCAA | 3 | 5.26 | 0.41 |
| Total of all microsatellites | 739 (100%) | - | - | - | - |

**Table 3** Amino acids that are coded for by trinucleotides and hexanucleotides within coding regions of the *C. fimbriata* genome

| Microsatellite class | Motif | Number of motifs | Amino acid |
|---|---|---|---|
| Trinucleotides | AAC | 12 | Asparagine |
| | ACA | 11 | Threonine |
| | ACG | 11 | Threonine |
| | AGA | 6 | Arginine |
| | AGC | 20 | Serine |
| | AGG | 24 | Lysine |
| | CAA | 66 | Glutamine |
| | CAG | 85 | Glutamine |
| | CGA | 9 | Arginine |
| | CTG | 13 | Leucine |
| | GAA | 17 | Glutamic Acid |
| | GAC | 14 | Aspartic acid |
| | GCA | 42 | Alanine |
| | GCT | 18 | Alanine |
| | TCG | 1 | Serine |
| | TGC | 14 | Cystine |
| Hexanucleotides | CAGGCT | 4 | Glutamine/Alanine |
| | CAGCAA | 3 | Glutamine |
| | CAGGCA | 3 | Glutamine/Alanine |
| | GCTCAA | 3 | Glutamine/Alanine |

**Table 4** Comparison of the genome size, GC content, microsatellite density and total microsatellites in the genomes of various Ascomycetes

| Organism | Sequence analysed (Mb) | GC content | Microsatellite density | Total microsatellites |
|---|---|---|---|---|
| *A. nidulans*[1] | 30.1 | 50.3[2] | 1/12.5 kb | 2410 |
| *F. graminearum*[1] | 36.1 | 48.3[2] | 1/12.5 kb | 2896 |
| *M. grisea*[1] | 37.9 | 51.6[2] | 1/3.3 kb | 11642 |
| *N. crassa*[1] | 38.0 | 49.9[2] | 1/2.7 kb | 14319 |
| *S. cerevisiae*[1] | 14.2 | 37.7[2] | 1/3.9 kb | 3618 |
| *Sch. pombe*[1] | 13.1 | 36.0[2] | 1/4.4 kb | 3232 |
| *C. fimbriata*[3] | 47.7 | 48.1[3] | 1/7.1 kb | 6739 |

[1]Results from study by Karaoglu et al. (2005), [2]GC content from Lim et al. (2004), [3]Results from this study

**Table 5** Comparison of the distribution of microsatellites in the genomes of various Ascomycetes

| Organism | Mono (%) | Di (%) | Tri (%) | Tetra (%) | Penta (%) | Hexa (%) |
|---|---|---|---|---|---|---|
| *A. nidulans*[1] | 51.83 | 31.24 | 13.49 | 1.49 | 1.16 | 0.79 |
| *F. graminearum*[1] | 38.29 | 35.70 | 19.16 | 2.94 | 2.52 | 1.38 |
| *M. grisea*[1] | 69.08 | 14.86 | 13.51 | 1.88 | 0.28 | 0.39 |
| *N. crassa*[1] | 41.50 | 22.40 | 28.52 | 5.29 | 1.34 | 0.94 |
| *S. cerevisiae*[1] | 65.31 | 22.58 | 10.95 | 0.39 | 0.25 | 0.53 |
| *Sch. pombe*[1] | 72.15 | 20.48 | 6.06 | 0.65 | 0.56 | 0.09 |
| *C. fimbriata*[2] | 44.45 | 34.73 | 14.60 | 3.09 | 1.63 | 1.50 |

[1]Results from study by Karaoglu et al. (2005), [2]Results from this study.

**Table 6** Comparison of the longest motifs in each different group of microsatellite motifs in the genomes of various Ascomycetes

| Organism | Mono | Di | Tri | Tetra | Penta | Hexa |
|---|---|---|---|---|---|---|
| *A. nidulans*[1] | $(T)_{94}$ | $(GT)_{36}$ | $(TGA)_{31}$ | $(AAAT)_{13}$ | $(AAACG)_{14}$ | $(TTAGGG)_{22}$ |
| *F. graminearum*[1] | $(T)_{41}$ | $(CT)_{28}$ | $(GAA)_{46}$ | $(CTTT)_{13}$ | $(GTATG)_{18}$ | $(TGAAGA)_{22}$ |
| *M. grisea*[1] | $(T)_{59}$ | $(GA)_{92}$ | $(TGG)_{37}$ | $(TACC)_{48}$ | $(GGCAA)_{29}$ | $(GCCTGA)_{58}$ |
| *N. crassa*[1] | $(T)_{89}$ | $(TC)_{78}$ | $(TTA)_{93}$ | $(AGGA)_{51}$ | $(AAGGA)_{32}$ | $(AGGGTT)_{28}$ |
| *S. cerevisiae*[1] | $(T)_{42}$ | $(GA)_{32}$ | $(TAT)_{36}$ | $(AAAT)_{13}$ | $(GATGA)_{7}$ | $(TGTTTT)_{8}$ |
| *Sch. pombe*[1] | $(T)_{39}$ | $(TG)_{19}$ | $(CAA)_{28}$ | $(TAAA)_{7}$ | $(TATTT)_{9}$ | $(ATTATC)_{6}$ |
| *C. fimbriata*[2] | $(G)_{62}$ | $(GA)_{41}$ | $(AAG)_{19}$ | $(TCAC)_{15}$ | $(GACAG)_{18}$ | $(GAAAAT)_{14}$ |

[1]Results from study by Karaoglu et al. (2005), [2]Results from this study.

# Chapter 3

# A diagnostic test using microsatellite markers to differentiate between cryptic species in the *Ceratocystis fimbriata sensu lato* complex

**Abstract**

*Ceratocystis fimbriata sensu lato* represents a complex of cryptic and commonly plant pathogenic species that are morphologically similar. To date, 26 species within this complex have been described using morphological characteristics, intersterility tests and phylogenetics. Population studies using microsatellites markers have been informative in defining population diversity and the origins of some of these species. Of the 47 published microsatellite markers, 35 could be mapped onto the *C. fimbriata* genome sequence and 25 onto the *C. albifundus* genome sequence. Thirty-six microsatellites, identified in putative genes from the *C. fimbriata* genome sequence in the previous chapter of this thesis, were selected to develop a diagnostic test to differentiate between the cryptic species in the complex. Up to five isolates of each species in the complex were tested with a subset of 10 newly developed polymorphic microsatellite markers. The absence of amplicons at some of the loci was diagnostic and a few species could be identified solely on this basis, including the outgroup species *C. virescens*. Genescan analysis was used to assign allele sizes for each locus. By determining consensus allele sizes from multiple isolates of a species, 11 species in the complex could be delimited successfully. However, the results were ambiguous for 6 species that gave mixed results and for 9 species for which only one or two isolates were tested. Consensus sizes could not be confidently defined for these species and further work will be needed to provide a fully functional diagnostic test for *Ceratocystis* species in the *C. fimbriata s.l.* complex.

## Introduction

*Ceratocystis fimbriata* was first described as the causal agent of sweet potato rot in 1890 (Halsted). Since then many fungi have been identified as representing this species and infecting a wide variety of plants of agricultural and economic importance around the world, including coffee (Pontis, 1951), poplar (Gremmen and de Kam, 1977), *Acacia* species (Morris et al., 1993) and *Eucalyptus* species (Roux et al., 2004; Roux et al., 2000). Phylogenetic inference based on DNA sequence data has led to the recognition that *C. fimbriata sensu lato* represents a complex of cryptic species, some of which might be host-specific (Barnes et al., 2003; Engelbrecht and Harrington, 2005; Johnson et al., 2005; van Wyk et al., 2009). The first species in the complex to be described after *C. fimbriata* was *C. albifundus* where it causes a serious wilt disease on non-native *Acacia mearnsii* in South Africa (Wingfield et al., 1996). Since these first two descriptions, 24 more species have been named based on studies using DNA sequence comparisons, intersterility tests and molecular markers (Baker-Engelbrecht and Harrington, 2005; Johnson et al., 2005).

Molecular markers have proved useful for various studies considering the population structure and origin of species in the *C. fimbriata s.l*. complex. Microsatellite markers developed from *C. cacaofunesta* were used to genotype invasive strains and to consider the population structure of the fungus in Latin America and *C. platani* in North America and Greece (Engelbrecht et al., 2007; Engelbrecht et al., 2004; Ocasio-Morales et al., 2007; Steimel et al., 2004). Recently, twenty microsatellite markers were developed to differentiate between mango-associated isolates of *C. fimbriata s.l.* in Brazil (Rizatto et al., 2010). From a South African perspective, microsatellite markers developed for an undescribed species in the *C. fimbriata s.l.* species complex were used to study the population structure of *C. albifundus* and *C. pirilliformis* (Barnes et al., 2001; Kamgan Nkuekam et al., 2009). These microsatellite markers could also differentiate between some of the cryptic species in the complex according to their geographic location and host-specificity and showed congruence with phylogenetic trees using the ITS region (Barnes et al. 2001).

Microsatellite markers are 1-6 bp tandem repeats, e.g. $(CAG)_5$, that are found throughout the genomes of most prokaryotes and eukaryotes (Field and Wills, 1996; Tautz and Renz, 1984). They are mainly present in non-coding regions but can also be found in coding regions (Jarne and Lagoda, 1996). Microsatellites are highly polymorphic and experience a high mutation rate of $10^{-6}$ to $10^{-2}$ per locus per generation (Ellegren, 2000). These length polymorphisms are due to DNA replication slippage and indel slippage that insert or delete

repeat units, and give rise to many different alleles at a single locus (Dieringer and Schlötterer, 2003; Schlötterer and Tautz, 1992). The polymorphic nature of microsatellites, along with their reproducibility, ease of use and co-dominant inheritance makes them ideal molecular markers (Levinson and Gutman, 1987; Tautz and Renz, 1984).

Isolation of microsatellites was traditionally achieved through *de novo* isolations whereby genomic libraries were screened with probes or PCR followed by cloning and sequencing (Zane et al., 2002). This is, however, time-consuming and expensive especially in organisms with a low microsatellite density (Selkoe and Toonen, 2006). The recent advances in DNA sequencing technology and bioinformatic software, such as MSatFinder (Thurston and Field, 2005), has made isolating microsatellites much simpler and more efficient (Abdelkrim et al., 2009; Santana et al., 2009). Bioinformatic searches generate data on each microsatellite in the genome, which can then be used for studying the microsatellite structure or developing microsatellite markers for further studies in population genetics or genetic mapping (Karaoglu et al., 2005; Santana et al., 2009). For example, microsatellite markers have always been difficult to identify in the red flour beetle *Tribolium castaneum* using conventional methods and as such only 19 were developed (Pai et al., 2003), whereas bioinformatic analysis of the whole genome sequence yielded over 12000 microsatellites from which 891 were tested and 509 were found polymorphic (Demuth et al., 2007). Fifteen of these polymorphic microsatellites markers were then used in a population study on wild red flour beetles (Drury et al., 2009).

The aim of this study was firstly to map the previously published microsatellite markers onto the genome of *C. fimbriata* and thus to understand their potential taxonomic value for the complex as well as to determine if any were closely linked. In addition we developed additional microsatellite markers from the genome sequence for a diagnostic test to differentiate between the cryptic species in the *C. fimbriata s.l.* complex.

**Methods and Materials**

*Genome sequences*

The genome sequences of *C. fimbriata* (CMW 14799) and *C. moniliformis* (CMW 10134) were sequenced using 454 sequencing technology (Roche Diagnostics, Mannheim, Germany) and both were assembled using Newbler. The size of the *C. fimbriata* genome was 47.7 Mb and had 11x coverage, while *C. moniliformis* had a size of 31 Mb and 20x

coverage. The *C. albifundus* (CMW 17274) genome was sequenced using Illumina sequencing technology (Illumina, Inc., California, USA). The raw read sequences were then assembled using Velvet version 1.0.19 (Zerbino and Birney, 2008). The *C. albifundus* genome had a size of 28 Mb and 58x coverage.

*Screening the Ceratocystis genomes for published microsatellites*

Fasta files of the microsatellite sequences for the *C. fimbriata s.l.* species complex published by Barnes et al. (2001), Steimel et al. (2004) and Rizatto et al. (2010) were obtained from Genbank (Accession numbers in Table 1). Using CLC Genomics Workbench (CLC bio, Aarhus, Denmark), local BLAST searches of the microsatellite sequences against the *C. fimbriata*, *C. albifundus* and *C. moniliformis* genome sequences were performed. Primer sequences for each microsatellite were obtained from the relevant studies (Barnes et al., 2001; Rizatto et al., 2010; Steimel et al., 2004). The binding sites of the primers were then identified using the primer function in CLC Genomics Workbench. If the microsatellite could still not be identified, a local BLAST search was performed of all the raw reads against the microsatellite sequences.

Contigs from the *C. fimbriata* genome containing the microsatellite sequences were then analysed using the online interface of AUGUSTUS (Keller et al., 2011) with the *Fusarium graminearum* genome sequence as the reference annotated genome to determine where the predicted genes were located. The microsatellite locations were then compared to the putative gene locations to determine their presence within coding regions, introns or non-coding regions. Proteins of the predicted genes that contain microsatellites were then identified by analysing their amino acid sequences using BLASTp (Altschul et al., 1990).

*Fungal isolates and DNA isolation*

Ninety isolates representing 26 species in the *C. fimbriata s.l.* complex and one species from the *C. coerulescens* complex, *C. virescens* (Table 2) were grown on 2% (v/w) malt extract agar (MEA, Biolab, Midrand, South Africa) supplemented with 100 mg/L streptomycin sulphate salt (SIGMA, Steinheim, Germany) and 100 mg/L thymine hydrochloride (SIGMA, Steinheim, Germany) for two weeks at 25°C. Hyphal tips were isolated and grown on 2% MEA. DNA extraction was performed as previously described by van Wyk et al. (2006).

*Microsatellite discovery and primer design*

In Chapter 2 of this thesis, microsatellites were discovered in putative coding regions in the *C. fimbriata* genome. This did not include introns of genes, thus some genes were screened for the presence of microsatellites within introns. Microsatellite motifs, excluding mononucleotides, of five repeats or more were targeted for the design of primers. Where the microsatellite was less than 50 bp from the either end of the contig, it was not analysed further. Using Primer3 (Rozen and Skaletsky, 2000), primers were designed for forty microsatellite loci (Table 3). These primers were then tested on the genome sequence in CLC Bio Genomics Workbench to determine whether they would result in the amplification of a single fragment. Primers for 36 microsatellite loci were synthesized at Inqaba Biotec (Pretoria, South Africa) and then tested on four isolates representing different species in the *C. fimbriata s.l.* complex; *C. cacaofunesta* (CMW 26375), *C. manginecans* (CMW 13851), *C. platani* (CMW 1896) and *C. polyconidia* (CMW 23818).

*PCR amplification*

The PCR reactions were prepared using 20-50 ng of DNA in a 25 µl reaction containing one unit of Mytaq polymerase (Bioline Ltd, London, United Kingdom), 5x reaction buffer (consisting of five mM dNTPs and 15 mM $MgCl_2$) and 10 mM of each primer. These reactions were performed on an Eppendorf thermocycler (Eppendorf, Hamburg, Germany). The first denaturation step was carried out at 95°C for one minute, followed by 35 cycles of 95°C for 15 s, 50°C for 15s and 72°C for 10 s, with a final elongation step of 72°C for seven minutes. The PCR products were then visualised on a 2% (v/w) Agarose gel under UV light.

*Sequencing and analysis*

The PCR products were purified with the DNA Clean and Concentrator™ Kit (Zymo Research Corporation, California, USA). Sequencing was performed in 10 µl reactions using either the forward or reverse primers for each microsatellite and a Big Dye cycle sequencing Kit v 3.1 (Perkin-Elmer, Warrington, UK) following the manufacturer's instructions. Sequencing PCR reactions were purified using the ZR DNA Sequencing Clean-Up™ Kit (Zymo Research Corporation, California, USA). An ABI PRISM 3300 Genetic Analyser was then used to generate the sequences and the chromatograms that were produced were analysed using Bioedit (Hall, 1999). The sequences were aligned with MAFFT (Katoh and Toh, 2008).

*Genescan analysis*

Ten microsatellite loci that showed polymorphisms between the four species were selected for screening using Genescan (Table 3). The forward primer from each pair was resynthesized and fluorescently labelled with PET, 6-FAM, NED or VIC (Applied Biosystems, California, USA). PCR was then performed as above to amplify fragments in all isolates used in this study (Table 2) but with annealing temperatures ranging from 42°C to 50°C and a final elongation step of 45 minutes. Four of the PCR products were then combined, each to a dilution of 1:100, according to their amplicon size and the type of fluorescent dye. One µL of the mix was combined with 0.18 µL Genescan-600 Liz internal size standard (Applied Biosystems, California, USA) and 10 µL formamide. These mixes were then separated on a 36 cm capillary with POP$^{TM}$ polymer on an ABI Prism 3100 Genetic Analyser. The program Peak Scanner Software v1.0 (Applied Biosystems, California, USA) was used to analyse the fragment sizes.

*Diagnostic test*

A diagnostic test was developed using the presence or absence of PCR products on a 2% (v/w) agarose gel, and fragment sizes of the amplicons analysed using Genescan analysis. Consensus fragment sizes for each microsatellite locus for each species were then determined in order to distinguish between the species in the complex.

*Transposon analysis*

The program Censor was used to identify transposons against the Repbase database (Kohany et al., 2006). All the microsatellites, those previously published and those developed in this study were searched for transposons that spanned or were within one bp of the microsatellite motif. The microsatellite sequences as well as the contigs on which they were found in the *C. fimbriata* genome were analysed with the online version of Censor.

**Results**

*Mapping published microsatellites onto the Ceratocystis genomes*

Thirty-five of the 47 published microsatellites could be placed onto the *C. fimbriata* genome (Table 1), most of which were present on different contigs. The full sequences of four

microsatellites (CfCAA80, CfCAG15, Cfim16 and Cfim18) could not be determined in the genome as they fall at the ends of the contigs. Ten of these microsatellites were present in putative gene regions (Table 4), however, some had a low coverage and/or a high E–value. In *C. albifundus*, 25 microsatellites could mostly be mapped onto different contigs in the genome (Table 1) and two microsatellites (CfCAA80 and Cfim20) were located at the ends of contigs. None of the published microsatellites could be identified in the *C. moniliformis* genome.

The presence and locations of nine microsatellite sequences could not be determined in the genomes of *C. fimbriata* and *C. albifundus*. Six microsatellites (Cfim01, Cfim02, Cfim05, Cfim08, Cfim13 and Cfim14) had similar motifs, i.e. the motif $(AC)_{14}$ was present in three of the microsatellites and the rest consist of various $(AC)_n$ or $(TG)_n$ repeats, as well as similar flanking regions (Fig.1). These six similar microsatellite sequences all aligned to the same position in both *C. fimbriata* and *C. albifundus*. The other three microsatellites (Cfim03, Cfim06 and Cfim07), with $(AC)_n$ motifs, also aligned to the same position although their flanking regions were not as similar. Analyses of the raw reads failed to resolve these microsatellites further as there was little or no coverage of the flanking regions, and the raw reads that did cover parts of the microsatellite sequence had low quality scores.

Microsatellites CF17/18 and CF23/24, with $(CA)_{15}$ and $(TG)_{15}$ motifs respectively, aligned well to the same position in the *C. fimbriata* and *C. albifundus* genomes. Both their primers had 100% similarity within this region in *C. fimbriata*, while CF17 did not have a binding site in this region for *C. albifundus*. These two microsatellite sequences appeared to be the same sequence as they aligned well to one another (Fig. 2). Similarly, the microsatellites Cfim16 and Cfim18, with motifs $(AC)_{21}$ and $(TG)_{16}$ respectively, both aligned to the same position in the *C. fimbriata* and *C. albifundus* genomes (Fig. 3). There were, however, differences between the sequences especially at the ends of the flanking regions, and the number of repeats of the microsatellite itself. Only the reverse primers had similarity to these regions in the *C. fimbriata* and *C. albifundus* genomes, and in the *C. fimbriata* genome this microsatellite is at the end of the contig. Another two microsatellites, Cfim09 and Cfim17, had the same motif $(AC)_9$, and aligned well to one another except for the end flanking regions of the sequences (Fig. 4). They also both aligned to the same position in the *C. fimbriata* and *C. albifundus* genomes. However, the alignment of Cfim17 to the contig was better than that of Cfim09. Only the reverse primers had some similarity to this region. Locations of all these microsatellites could not be determined further even after analysing the raw reads.

*Development of microsatellites from the C. fimbriata genome*

Thirty-six microsatellites within gene regions (Table 5) were chosen as genes are expected to be present across species, but would show differences between species while being similar within a species. Some of the microsatellites were chosen from different coding regions present on the same contig in order to test whether they would still produce polymorphisms between species. Proteins would be slightly different between species even if they are close to each other in the genome as both the species and different regions of the genome would experience different selection pressures (Metzgar et al., 2000). Most of the putative microsatellites were trinucleotides, with some dinucleotides, tetranucleotides, hexanucleotides and one octanucleotide. In the *C. albifundus* genome three (CF_CTCTCTGT5, CF_CG5 and CF_TCC7) of the 36 microsatellites could be not be identified and none could be identified in the *C. moniliformis* genome.

After amplification of the 36 microsatellites, one (CF_GTT6) was discarded as it produced two fragments in all the isolates tested. The remainder of the microsatellite primers produced single amplicons. Nine microsatellites showed no polymorphisms in any of the isolates tested. The rest had at least two alleles with six showing different alleles in four species, including *C. fimbriata.* Most of the loci present on the same contigs had different allele complements between the four different species tested but some loci were monomorphic. Single nucleotide polymorphisms were identified at nine loci (Fig. 5), some of which disrupted the microsatellite motifs but in essence retained the same repeat number. From the 26 polymorphic microsatellites, 10 loci were chosen for further analysis using Genescan; one contained a SNP (CF_CAGAAG5) but still showed differences in the microsatellite repeat number.

*Genescan Analysis*

The 10 polymorphic primers were tested on isolates representing different species within the *C. fimbriata s.l.* species complex and an outgroup species, *C. virescens* (Table 6). Some primer pairs failed to amplify loci in all the isolates representing a species and this was considered an allele. The 10 polymorphic markers produced a total of 141 alleles with a size range of 123 bp to 360 bp. The smallest number of alleles per locus is eight (CF_CAAG5) and the largest is 20 (CF_GCT11). The locus with the lowest allele diversity is found in the microsatellite designed within an intron of a putative gene. Consensus allele sizes were then

determined for each species by analysing fragment sizes for each isolate (Table 7). Some species had conflicting allele sizes and thus a consensus size could not be assigned.

In this study we wished to find loci that were fixed within a species; a locus that is monomorphic in a species. In addition, the locus should be variable between species. The 10 polymorphic microsatellite markers could therefore be used to design a diagnostic test for species identification. Twelve species (*C. cacaofunesta*, *C. colombiana*, *C. caryae*, *C. curvata*, *C. diversiconidia*, *C. ecuadoriana*, *C. fimbriata*, *C. fimbriatomima*, *C. larium*, *C. manginecans*, *C. platani* and *C. tsitsikammensis*) in the *C. fimbriata s.l.* complex had consistent allele sizes at each of the microsatellite loci in most isolates. Some isolates of a species produced different allele sizes at some of the microsatellite loci (Table 6), e.g. the locus CF_CAA/CAG81 had four allele sizes for five isolates of *C. pirilliformis*. Isolates of *C. albifundus*, *C. cacaofunesta*, *C. neglecta*, C*. papillata*, and *C. variospora* also displayed mixed results at some of the microsatellite loci.

*Diagnostic test*

A diagnostic test was developed for the *C. fimbriata s.l.* complex (See Fig. 6 for a flow chart on how to differentiate between the species). Banding patterns from the microsatellite PCRs were in some cases diagnostic, as PCR products for some loci could not be amplified from all isolates of a species. The outgroup species *C. virescens* produced a distinctive banding pattern compared to species in the *C. fimbriata s.l.* complex when subjected to PCR of all the microsatellite loci and could therefore be identified on this basis. It displayed double bands at two loci (CF_GCT11 and CF_CTCTCTGT5) and no amplicons at three loci (CF_CAA/CAG80, CF_CAA/CAG81 and CF_CAA/CAG24), with the rest of the loci showing mixed results in all the isolates tested. Other species that could be identified based on the banding patterns were *C. acaciivora*, *C. albifundus, C. caryae*, *C. polyconidia*, C*. populicola*, *C. smalleyi*, *C. tanganyicensis* and C*. variospora*. However, only one isolate of *C. polyconidia*, C*. populicola*, and *C. smalleyi* were tested; thus the result is tentative.

The rest of the species all resulted in amplicons of similar size based on agarose gel electrophoresis. Genescan analysis was therefore required to identify those species further by assigning allele sizes. In *C. colombiana*, *C. fimbriata*, *C. fimbriatomima*, *C. larium*, *C. manginecans*, *C. neglecta* and *C. platani* some of the primer pairs failed to amplify the loci. However, the remaining isolates (out of four or five isolates) were tested in each species and they all produced the same fragment sizes, therefore a fairly confident consensus size could

be obtained. For the species *C.atrox*, C*. caryae*, *C. obpyriformis*, *C. papillata*, *C. pirilliformis*, and *C. zombamontana* amplification in only some of the loci along with mixed results in other loci and/or low number of isolates tested implies that confident consensus sizes could not be properly assigned for every locus.

*Transposon analysis*

Out of the 47 previously published microsatellites, 16 did not contain any transposable elements and nine could not be mapped onto the *C. fimbriata*, thus they could not be tested for the presence of transposons. The transposable elements that could be identified consisted of one DNA transposon, six long terminal repeat (LTR) retrotransposons, five non-LTR retrotransposons and one endogenous virus (Table 8). One of the transposable elements identified was one bp away from the microsatellite and was included (Cfim20 with Chapaev3-1_MR). For the 36 microsatellite markers developed in this study, transposable elements could not be found in the vicinity of 22 microsatellite motifs. For the 14 remaining motifs, one endogenous virus, five non-LTR retrotransposons, five LTR retrotransposons and three DNA transposons could be identified (Table 9).

**Discussion**

The aims of this study were to screen the *C. fimbriata*, *C. albifundus* and *C. moniliformis* genomes for published microsatellites and then to develop microsatellites in gene regions using the *C. fimbriata* genome sequence to distinguish between cryptic species in the *C. fimbriata s.l.* species complex. The published microsatellite markers were developed from *C. cacaofunesta* (Steimel et al., 2004), *C. fimbriata* from Mango trees in Brazil (Rizatto et al., 2010) and *C. fimbriata* from various geographical locations (Barnes et al., 2001). Most of these markers could be mapped onto the *C. fimbriata* genome; however, there were difficulties in mapping some of the microsatellites identified by Rizatto et al. (2010). This could possibly be due to the fact that the specific regions of the genome where these microsatellites are located have not been sufficiently sequenced. Barnes et al. (2001) showed that the microsatellites markers that they developed could be used to distinguish between samples of *C. fimbriata* isolated from different geographical regions. This observation led to the idea that microsatellites in gene regions could possibly differentiate between the cryptic species in the *C. fimbriata s.l.* complex.

In this study, we have taken this observation further and developed a tool to identify species in a more robust manner than is possible using only morphological characteristics. There were, however, some problems with differentiating between some of the cryptic species in the complex. Some isolates representing a species presented mixed results, thus definitive allele sizes could not be assigned. In other cases, a large enough sample size could not be tested for some of the species and therefore consensus allele sizes could not be obtained with confidence.

*Mapping the published microsatellites onto the Ceratocystis genomes*

The *C. fimbriata* and *C. albifundus* genomes did not contain all of the published microsatellites. This could mean that the microsatellites are either not there or that these genomes have not been completely sequenced. No published microsatellites could be identified in the *C. moniliformis* genome, which confirms that it is distinctly different to species in the *C. fimbriata s.l.* complex and forms part of a different complex, the *C. moniliformis s.l.* complex. The 10 published microsatellites in putative genes could pose a problem for studies on populations or genetic mapping as neutral markers are required for such studies which are normally present in non-coding regions (Selkoe and Toonen, 2006). However, these 10 microsatellites are present within genes and are less likely to be polymorphic within a species as they could be under selection. The low coverage and high E-value of some of the genes indicates that their identities are unknown in fungi and may be unique to *C. fimbriata* or that the gene prediction program was incorrect in assigning these regions as genes.

Three of the microsatellites found in genes and 10 in non-coding regions were associated with transposons, most of which are retrotransposons. Retrotransposons make copies of themselves and then transfer to another location within the genome. It is not uncommon to find microsatellites associated with transposons (Akagi et al., 2001; Coates et al., 2011; Temnykh et al., 2001). The microsatellites would therefore have the potential to move in the genome but could also be duplicated. This would result in more than one locus and could complicate analyses for population studies and genetic mapping when using these microsatellite markers.

Primers developed for microsatellites in non-coding regions are less likely to be transferable to related species as these regions experience less selection than coding regions and are thus more likely to undergo mutations (Barbará et al., 2007). This has been seen in

population studies where some of the microsatellites developed in one species are monomorphic or the primers do not work in another species within the *C. fimbriata s.l.* complex (Barnes et al., 2005; Engelbrecht et al., 2007; Engelbrecht et al., 2004; Ferreira et al., 2010; Kamgan Nkuekam et al., 2009; Ocasio-Morales et al., 2007). For example, microsatellite primers developed for *C. fimbriata* were not all transferable to *C. albifundus* and *C. pirilliformis* (Barnes et al., 2005; Kamgan Nkuekam et al., 2009). However, mapping the published microsatellites onto the *C. albifundus* genome showed that most are present but some of the primer binding sites contain mismatches which explains why the microsatellite primers were not transferable to *C. albifundus*.

Some of the published microsatellites could be linked. Microsatellites CF17/18 and CF23/24 share the same motif (although in the reverse complement), flanking regions and align to the same position in the *C. fimbriata* and *C. albifundus* genomes. Using these two markers together is therefore not recommended as they would result in the same data. The other microsatellite markers that are similar were all developed by Rizatto et al. (2010). These 13 microsatellite markers are problematic as they have similar, if not the same, motifs (AC, CA or TG) and flanking regions, and align to the same position in the *C. fimbriata* and *C. albifundus* genomes. Results from our analyses and the fact that other microsatellites are available therefore lead us to conclude that continued use of these markers is not desirable.

*Development of microsatellite markers from the C. fimbriata genome*

Microsatellite markers are usually designed in non-coding regions to address questions on populations, such as population diversity, of a single species (Selkoe and Toonen, 2006). However, in order to differentiate between species using microsatellite markers, they should ideally be under selection and thus gene regions of the genome would be preferable. In this study, microsatellite markers were developed to differentiate between the different species in the *C. fimbriata s.l.* species complex. Microsatellites were chosen from gene regions as they are expected to be more similar within a species but potentially different between species. However, it is expected that some would be the same between species as the genes could be conserved and thus the microsatellite as well. This would be especially true for non-triplet repeats, i.e. dinucleotides, where there may be selection in keeping the microsatellite length constant so as not to disrupt the reading frame (Metzgar et al., 2000).

The proteins within which the monomorphic microsatellites reside either have no similarity to any previously identified protein or are similar to hypothetical proteins, mostly belonging to

other fungi. Some of these proteins could play a role in functions that require the protein to be conserved. Two of the monomorphic microsatellites found in this study were also associated with transposons, particularly DNA transposons. DNA transposons can relocate to a different part of the genome or they can be copied to another location (Kapitonov and Jurka, 2001; Kidwell and Lisch, 1997; Pritham et al., 2007). The fact that the microsatellites appear to be monomorphic could mean that they are required for a function of the transposon and are therefore conserved.

Microsatellites in genes that are polymorphic between species are more likely able to differentiate species. The 25 polymorphic microsatellites identified in this study consisted mostly of trinucleotides but also included dinucleotides, hexanucleotides, a tetranucleotide and an octanucleotide. One dinucleotide and the tetranucleotide are present within introns while the rest are found in coding regions. The tri- and hexanucleotides showed the most polymorphisms, probably because they are triplet repeats and the loss or gain of a repeat unit would not disrupt the reading frame (Metzgar et al., 2000). Also, constraints might not be placed on the number of repeated amino acids in a homo-polymer tract as the protein function may not necessarily be influenced by a change in repeat number.

The polymorphic microsatellites in this study were contained within genes encoding proteins that are mostly similar to fungal proteins, including from some pathogens such as *Verticillium albo-atrum*, *Glomerella graminicola* and *Arthroderma gypseum* (Hastie, 1962; Politis, 1975; Stockdale, 1964). Two proteins, disulphide isomerase and acid phosphatase, are potentially involved in metabolism and three other proteins, sulphate permease, Ras GTPase and a major facilitator superfamily transporter, in membrane transport. However, the low coverage and high e-values of some of these proteins indicates that they may not actually have this function in *C. fimbriata*. It has already been established that different repeat numbers of microsatellite motifs within some cell-wall proteins can cause differences in pathogenicity between strains (Levdansky et al., 2007). These microsatellites could potentially influence pathogenicity in these fungi and in *C. fimbriata* in which they have been identified.

Eleven microsatellites were associated with transposable elements, mostly retrotransposons, which shows that microsatellites are not only mobile in non-coding regions but also mobile in gene regions. This feature could thus have potential value in species diagnostics and is supported by the fact that some microsatellite loci (CF_CTCTCTGT5 and CF_CAA/CAG24) did not amplify in some species. However, it could also be detrimental as these

microsatellites might be found in more than one location as a consequence of the mobility of these elements.

Markers other than microsatellites also show polymorphisms and can also be diagnostic for species differentiation, for example single nucleotide polymorphisms (SNPs). Some of the microsatellites tested in this study showed SNPs not only in the flanking regions but also in the microsatellite motifs themselves. SNPs present in the microsatellite motifs disrupt the microsatellite but the overall length of the microsatellite is maintained and would be indistinguishable if only fragment sizes are analysed. These SNPs could help to further differentiate between species in the *C. fimbriata s.l.* complex if they are found to be present in all isolates of a single species. Further analysis will be needed to verify that the SNPs identified are present within more than one isolate of each species.

Microsatellites in non-coding regions (including introns) are expected to be more polymorphic as they are under neutral selection (Selkoe and Toonen, 2006). It was interesting that the microsatellite found in the intron produced the smallest number of alleles when tested on all the species in the complex. This result suggests that some introns could be under selection, especially if splicing occurs in the intron of that gene. If the microsatellite forms part of the splicing recognition site, this would disrupt splicing of the transcripts and could cause the resulting protein not to be produced (Hastings and Krainer, 2001). However, it is possible that the microsatellite might not actually reside in an intron as the *de novo* gene prediction program AUGUSTUS may not have set the intron-exon boundaries correctly for this genus.

*Development of a diagnostic test*

The diagnostic test developed in this study produced some interesting results in which the identity of the species did not entirely match the allele complement at the 10 microsatellite loci. All 10 loci were amplified in the three isolates of *C. cacaofunesta*, however, isolate CMW 26375 which is the ex-type of the species, had different allele sizes to the other two isolates at five loci. All three isolates were isolated from *Theobroma cacao* in different South American countries. *Ceratocystis cacaofunesta* can be identified using this set of 10 microsatellites; however, this would exclude the ex-type isolate CMW 26375 as it is significantly different from the other two isolates. Thus, the ex-type of *C. cacaofunesta* could represent a different species. In order to confirm this, more isolates would need to be tested.

Another species that presents a conflict in terms of its taxonomy is *C. neglecta*. This species displays two different "types" from analysis of allele sizes at the 10 loci even though all isolates originated from *Eucalyptus* trees in Colombia. Isolates CMW 11284 and CMW 11285 share alleles at four loci that are different to the alleles that are shared by isolates CMW 17808 and CMW 18194. Interestingly, isolate CMW 11284 has the same allele complement as one isolate representing *C. platani* (CMW 1896). It also has almost the same complement as two other isolates, CMW 23450 and CMW 23918, of *C. platani* with allele sizes differing at one locus for each of these two isolates. As all alleles are shared between *C. neglecta* isolate CMW 11284 and *C. platani* isolate CMW 1896 it is possible that they represent the same species. However, it is also possible that they have the same allele complement by chance due to homoplasy (Estoup et al., 2002) and more microsatellite loci would be required to clarify this further.

It has been suggested that *C. manginecans* is a hybrid species and that the one parent is *C. acaciivora* while the other is unknown. This hypothesis originated from the observation that *C. manginecans* is clonal and shares an ITS sequence with *C. acaciivora* (Al-Adawi et al., unpublished). The results of this study support this hypothesis as *C. manginecans* shares the same alleles at nine of the 10 microsatellite loci with *C. acaiivora* isolates. Although the results for *C. acaiivora* are not entirely complete, alleles of at least one isolate are shared at each locus with *C. manginecans.* None of the other species in the complex have the same allele as *C. manginecans* at the CF_GCT11 locus. Further studies on a wider range of *C. fimbriata s.l.* isolates and species would be required to identify the unknown parent.

Developing an effective diagnostic technique was hampered by lack of isolates and variable success in the amplification of microsatellite regions. For the species *C. atrox*, *C. caryae*, C. *polychroma*, *C. polyconidia*, *C. populicola*, *C.obpyriformis*, *C. smalleyi*, *C. tanganyicensis,* and *C. zombamontana*, only one or two isolates were tested, therefore consensus allele sizes could not be assigned with confidence. The diagnostic test could be used to tentatively identify these species but more isolates should be tested in order to develop a robust diagnostic technique. Although three to five isolates were tested for *C. acaciivora*, *C. albifundus*, *C. papillata*, *C. pirilliformis, C. variospora* and *C. virescens,* mixed results were obtained for many of the loci, i.e. some loci did not amplify in some of the isolates and some allele sizes differed between some of the isolates representing a species. Thus these species cannot be identified properly using this set of ten microsatellites at this point in time. *C. virescens,* however, produced two amplicons at two different loci. This banding pattern

can be used to differentiate this species (and other related species in the *C. coerulescens* complex) from species in the *C. fimbriata s.l.* complex.

Due to some of the microsatellite markers used in this study showing variable results between isolates within some species, this presents a problem for differentiating effectively between species. A potential solution lies in other studies on species identification in fungi which have made use of barcoding regions, PCR-RFLPs and SNPs. The current unofficial barcoding gene for fungi is the ITS region is routinely used to identify *Ceratocystis* species (Nilsson et al., 2008; van Wyk et al., 2011) and thus would not provide further resolution to the species in the complex. The mitochondrial barcoding gene cytochrome *c* oxidase 1 is, however, a candidate for use in species differentiation as this gene shows variation among closely related species of *Penicillium* (Seifert et al., 2007). PCR-RFLPs have already been used to differentiate between some species of *Ceratocystis* successfully (Witthuhn et al., 1999). As the *C. fimbriata* and *C. albifundus* genomes are available they could be searched for restriction sites within specific regions to develop this technique further specifically for the *C. fimbriata s.l.* complex. SNPs provide another option to species differentiation; however, just using SNPs alone to differentiate species could require many loci (Travanti et al., 2005). A combination of these methodologies in conjunction with microsatellites would improve the diagnostic test and provide a robust method of species differentiation within the *C. fimbriata* species complex.

*Conclusions*

The diagnostic test developed in this study can be used to differentiate many of the species in the *C. fimbriata s.l.* complex. There are, however, some missing data points and conflicting allele sizes that should be reanalysed. Sequencing of all loci for each isolate should be carried out as there are likely SNPs that could improve this diagnostic test further. Testing a number of isolates from different geographic regions would be useful to ensure that variation between isolates is accounted for and could provide insight into whether the environment plays a role in microsatellite variation between isolates of a single species. Higher confidence of consensus allele sizes for each species would also be obtained with a larger sample of isolates. In cases where there seem to be more than one “type” present within a described species, more isolates should be tested. The previously published microsatellites that could be found within the *C. fimbriata* genome and contained within genes should also be tested, both with Genescan analysis and sequencing. As it has already been shown that some of the published microsatellites could differentiate between

different lineages of *C. fimbriata s.l.* isolates (Barnes et al., 2001), it is likely that some of published microsatellites would also be useful as diagnostic markers. Once these are complete, this diagnostic test would help to more easily identify species within the *C. fimbriata s.l.* complex and possibly aid in the description of new species.

**References**

Abdelkrim, J., Robertson, B. C., Stanton, J. L., Gemmell, N. J., 2009. Fast, cost-effective development of species-specific microsatellite markers by genomic sequencing. Biotechniques. 46**,** 185-192.

Akagi, H., Yokozeki, Y., Inagaki, A., Mori, K., Fujimura, T., 2001. Micron, a microsatellite-targeting transposable element in the rice genome. Molecular Genetics and Genomics. 266**,** 471-480.

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.

Barbará, T., Palma-Silva, C., Paggi, G. M., Bered, F., Fay, M. F., Lexer, C., 2007. Cross-species transfer of nuclear microsatellite markers: Potential and limitations. Molecular Ecology. 16**,** 3759-3767.

Barnes, I., Gaur, A., Burgess, T., Roux, J., Wingfield, B. D., Wingfield, M. J., 2001. Microsatellite markers reflect intra-specific relationships between isolates of the vascular wilt pathogen *Ceratocystis fimbriata*. Molecular Plant Pathology. 2**,** 319.

Barnes, I., Nakabonge, G., Roux, J., Wingfield, B. D., Wingfield, M. J., 2005. Comparison of populations of the wilt pathogen *Ceratocystis albifundus* in South Africa and Uganda. Plant Pathology. 54**,** 189-195.

Barnes, I., Roux, J., Wingfield, B. D., Dudzinski, M. J., Old, M. N., Wingfield, M. J., 2003. *Ceratocystis pirilliformis*, a new species from *Eucalyptus nitens* in Australia. Mycologia. 95**,** 865-871.

Coates, B. S., Kroemer, J. A., Sumeford, D. V., Hellmich, R. L., 2011. A novel class of miniture inverted repeat transposable elements (MITEs) that contain hitchhiking $(GTCY)_n$ microsatellites. Insect Molecular Biology. 20**,** 15-27.

Demuth, J. P., Drury, D. W., Peters, M. L., van Dyken, D., Priest, N. K., Wade, M. J., 2007. Genome-wide survey of *Tribolium castaneum* microsatellites and description of 509 polymorphic markers. Molecular Ecology Notes. 7**,** 1189-1195.

Dieringer, D., Schlötterer, C., 2003. Two distinct modes of microsatellite mutation processes: Evidence from the complete genomic sequences of nine species. Genome Research. 13**,** 2242-2251.

Drury, D. W., Siniard, A. L., Wade, M. J., 2009. Genetic differentiation among wild populations of *Tribolium castaneum* estimated using microsatellite markers. Journal of Heredity. Doi: 10.1093/jhered/esp77.

Ellegren, H., 2000. Microsatellite mutations in the germline: Implications for evolutionary inference. Trends in Genetics. 16**,** 551-558.

Engelbrecht, C. J., Harrington, T. C., 2005. Intersterility, morphology and taxonomy of *Ceratocystis fimbriata* on sweet potato, cacao and sycamore. Mycologia. 97**,** 57-69.

Engelbrecht, C. J. B., Harrington, T. C., Alfenas, A. C., Suarez, C., 2007. Genetic variations in populations of the cacao wilt pathogen, *Ceratocystis cacaofunesta*. Plant Pathology. 56**,** 923-933.

Engelbrecht, C. J. B., Harrington, T. C., Steimel, J., Capretti, P., 2004. Genetic variation in eastern North American and putatively introduced populations of *Ceratocystis fimbriata* f. *platani*. Molecular Ecology. 13**,** 2995-3005.

Estoup, A., Jarne, P., Corunet, J., 2002. Homoplasy and mutation model at microsatellite loci and their consequences for population genetics. Molecular Ecology. 11, 1591-1604.

Ferreira, E. M., Harrington, T. C., Thorpe, D. J., Alfenas, A. C., 2010. Genetic diversity and interfertility among highly differentiated populations of *Ceratocystis fimbriata* in Brazil. Plant Pathology. 59, 721-735.

Field, D., Wills, C., 1996. Long, polymorphic microsatellites in simple organisms. Proceedings of the Royal Society of London: Biological Sciences. 263**,** 209-215.

Gremmen, J., de Kam, M., 1977. *Ceratocystis fimbriata*, a fungus associated with poplar canker in Poland. European Journal of Forest Pathology. 7**,** 44–47.

Hall, T. A., 1999. Bioedit: A user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. Nucleic Acid Symposium Series. 41**,** 95-98.

Halsted, B. D., 1890. Some fungous diseases of the sweet potato. The black rot. Bulletin of the New Jersey Agricultural College Experiment Station. 76**,** 7-14.

Hastie, A. C., 1962. Genetic recombination in the Hop-wilt fungus *Verticilium albo-atrum*. Journal of General Microbiology. 27**,** 373-382.

Hastings, M. L., Krainer, A. R., 2001. Pre-mRNA splicing in the new millenium. Current Opinion in Cell Biology. 13**,** 302-309.

Jarne, P., Lagoda, P. J. L., 1996. Microsatellites, from molecules to populations and back. Trends in Ecology and Evolution. 11**,** 424-429.

Johnson, J. A., Harrington, T. C., Engelbrecht, C. J. B., 2005. Phylogeny and taxonomy of the North American clade of the *Ceratocystis fimbriata* complex. Mycologia. 97**,** 1067-1092.

Kamgan Nkuekam, G., Barnes, I., Wingfield, M. J., Roux, J., 2009. Distribution and population diversity of *Ceratocystis pirilliformis* in South Africa. Mycologia. 101**,** 17-25.

Kapitonov, V. V., Jurka, J., 2001. Rolling-circle transposons in eukaryotes. Proceedings of the National Academy of Sciences of the United States of America. 98**,** 8714-8719.

Karaoglu, H., Lee, C. M. Y., Meyer, W., 2005. Survey of simple sequence repeats in completed fungal genomes. Molecular Biology and Evolution. 22**,** 639-649.

Katoh, K., Toh, H., 2008. Recent developments in the MAFFT sequence alignment program. Briefings in Bioinformatics. 9, 286-298.

Keller, O., Kollmar, M., Stanke, M., Waak, S., 2011. A novel hybrid gene prediction method employing protein multiple sequence alignments. Bioinformatics. In print.

Kidwell, M. G., Lisch, D., 1997. Transposable elements as sources of variation in animals and plants. Proceedings of the National Academy of Sciences of the United States of America. 94**,** 7704-7711.

Kohany, O., Gentles, A. J., Hankus, L., Jurka, J., 2006. Annotation, submission and screening of repetitive elements in Repbase: RepbaseSubmitter and Censor. BMC Bioinformatics. 7, 474.

Levdansky, E., Romano, J., Shadkchan, Y., Sharon, H., Verstrepen, K. J., Fink, G. R., Osherov, N., 2007. Coding tandem repeats generate genetic diversity in *Aspergillus fumigatus* genes. Eukaryotic Cell. 6**,** 1380-1391.

Levinson, G., Gutman, G. A., 1987. Slipped-strand mispairing: A major mechanism for DNA sequence evolution. Molecular Biology and Evolution. 4**,** 203-221.

Metzgar, D., Bytof, J., Wills, C., 2000. Selection against frameshift mutations limits microsatellite expansion in coding DNA. Genome Research. 10**,** 72-80.

Morris, M. J., Wingfield, M. J., de Beer, C., 1993. Gummosis and wilt of *Acacia mearnsii* in South Africa caused by *Ceratocystis fimbriata*. Plant Pathology. 42**,** 814-817.

Nilsson, R.H., Kristiansson, E., Ryberg, M., Hallenberg, N., Larsson, K., 2008. Intraspecific *ITS* variability in the Kingdom *Fungi* as expressed in the International Sequence Databases and its implications for molecular species identification. Evolutionary Bioinformatics. 4, 193-201.

Ocasio-Morales, R. G., Tsopelas, P., Harrington, T. C., 2007. Origin of *Ceratocystis platani* on native *Platanus orientalis* in Greece and its impact on natural forests. Plant Disease. 91**,** 901-904.

Pai, A., Sharakhov, I.V., Braginets, O., Costa, C., Yan, G.Y., 2003. Identification of microsatellite markers in the red four beetle, *Tribolium castaneum*. Molecular Ecology Notes. 3, 425-427.

Politis, D. J., 1975. The identity and perfect state of *Colletotrichum graminicola*. Mycologia. 67**,** 56-62.

Pontis, R. E., 1951. A canker disease of the coffee tree in Colombia and Venezuela. Phytopathology. 41**,** 179-184.

Pritham, E. J., Putliwala, T., Feschotte, C., 2007. *Mavericks*, a novel class of giant transposable elements widespread in eukaryotes and related to DNA viruses. Gene. 390**,** 3-17.

Rizatto, S., de Araùjo Batista, C. E., Bajay, M. M., Sigrist, M. S., Ito, M. F., Monteiro, M., Cavallari, M. M., Pinheiro, J. B., Zucchi, M. I., 2010. A new set of microsatellite markers for the genetic characterisation of *Ceratocystis fimbriata*, an economically important plant pathogen. Conservation Genetics Resources. 2**,** 55-58.

Rozen, S., Skaletsky, H.J., 2000. Primer3 on the WWW for general users and for biologist programmers, in: Krawetz, S., Misener, S. (Eds.), Bioinformatic methods and protocols: Methods in Molecular biology. Humana Press, Totowa, New Jersey, pp. 365-386.

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

Roux, J., Wingfield, M. J., Bouillet, J. P., Wingfield, B. D., Alfenas, A. C., 2000. A serious new wilt disease of *Eucalyptus* caused by *Ceratocystis fimbriata* in Central Africa. Forest Pathology. 30**,** 175-184.

Santana, Q. C., Coetzee, M. P. A., Steenkamp, E. T., Mlonyeni, O. X., Hammond, G. N. A., Wingfield, M. J., Wingfield, B. D., 2009. Microsatellite discovery by deep sequencing of enriched genomic libraries. Biotechniques. 46**,** 217-223.

Schlötterer, C., Tautz, D., 1992. Slippage synthesis of simple sequence DNA. Nucleic Acids Research. 20**,** 211-216.

Seifert, K.A., Samson, R.A., deWaard, J.R., Houbraken, J., Lévesque, C.A., Moncalvo, J., Louis-Seize, G., Hebert, P.D.N., 2007. Prospects for fungus identification using *CO1* DNA barcodes, with *Penicillium* as a test case. Proceedings of the National Academy of Science of the United States of America. 104, 3901-3906.

Selkoe, K. A., Toonen, R. J., 2006. Microsatellites for ecologists: A practical guide to using and evaluating microsatellite markers. Ecology Letters. 9**,** 615-629.

Steimel, J., Engelbrecht, C. J. B., Harrington, T. C., 2004. Development and characterisation of microsatellite markers for the fungus *Ceratocystis fimbriata*. Molecular Ecology Notes. 4**,** 215-218.

Stockdale, P. M., 1964. The *Microsporum gypseum* complex (*Nannizzia incurvata* Stockd., *N. gypsea* (Nann.) comb. nov., *N. fulva* sp. nov.). Medical Mycology. 3**,** 114-126.

Tautz, D., Renz, M., 1984. Simple sequences are ubiquitous repetitive components of eukaryotic genomes. Nucleic Acids Research. 12**,** 4127-4138.

Temnykh, S., DeClerck, G., Lukashova, A., Lipovich, L., Cartinhour, S., McCouch, S., 2001. Computational and experimental analysis of microsatellites in rice (*Oryza sativa* L.): Frequency, length variation, transposon associations and genetic marker potential. Genome Research. 11**,** 1441-1452.

Thurston, M. I., Field, D., 2005. Msatfinder: Detection and characterisation of microsatellites. Distributed by the authors at http://www.genomics.ceh.ac.uk/msatfinder/.

Travanti, A., Davidson, A.D., Fordyce, M.J., Gow, N.A.R., Maiden, M.C.J., Odds, F.C., 2005. Population structure and properties of *Candida albicans*, as determined by multilocus sequence typing. Journal of Clinical Microbiology. 43, 5601-5613.

van Wyk, M., van der Merwe, N. A., Roux, J., Wingfield, B. D., Kamgan, G. N., Wingfield, M. J., 2006. Population genetic analyses suggest that the *Eucalyptus* fungal pathogen *Ceratocystis fimbriata* has been introduced into South Africa. South African Journal of Science. 102**,** 259-263.

van Wyk, M., Wingfield, B. D., Clegg, P. A., Wingfield, M. J., 2009. *Ceratocystis larium* sp nov., a new species from *Styrax benzoin* wounds associated with incense harvesting in Indonesia. Persoonia. 22**,** 75-82.

Wingfield, M. J., DeBeer, C., Visser, C., Wingfield, B. D., 1996. A new *Ceratocystis* species defined using morphological and ribosomal DNA sequence comparisons. Systematic and Applied Microbiology. 19**,** 191-202.

Witthuhn, R.C., Wingfield, B.D., Wingfield, M.J., Harrington, T.C., 1999. PCR-based identification and phylogeny of species of *Ceratocystis sensu stricto*. Mycological Research. 103, 743-749.

Zane, L., Bargelloni, L., Patarnello, T., 2002. Strategies for microsatellite isolation: A review. Molecular Ecology. 11**,** 1-16.

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

**Table 1** Presence of published microsatellite motifs in the *C. fimbriata* and *C. albifundus* genome sequences and those that are transferable to other species based on population studies

| Microsatellite name | Motif in published sequence | In the genome sequence of *C. fimbriata* or *C. albifundus* | Transferable to other species based on population studies | In gene? | GenBank accession number |
|---|---|---|---|---|---|
| AG1/2[1] | $(T)_7C(T)_2CGC(T)_4(CTTT)_2GC(T)_4C(T)_3C(T)_2G(T)_4(CTT)_2$ | *C. fimbriata* and *C. albifundus* | *C. pirilliformis*[5] | NO | AY055016 |
| AG7/8[1] | $(TC)_{21}(TTC)_2$ | *C. fimbriata* and *C. albifundus* | *C. albifundus*[4] and *C. pirilliformis*[5] | NO | AY055017 |
| AG15/16[1] | Regions rich in A interrupted by C and G | *C. fimbriata* and *C. albifundus* | *C. albifundus*[4] | NO | AY055018 |
| AG17/18[1] | $(T)_5(C)_2(CT)_2T(CTT)_6(T)_2(C)_3TC(T)_3$ | *C. fimbriata* and *C. albifundus* | *C. albifundus*[4] | YES | AY055019 |
| CF5/6[1] | $(TGC)_{11}$ | *C. fimbriata* | *C. albifundus*[4] | NO | AY055020 |
| CF11/12[1] | $CA(AC)_7GC(AC)_2(N)x(G)_8$ | *C. fimbriata* and *C. albifundus* | *C. pirilliformis*[5] | NO | AY055021 |
| CF13/14[1] | $(T)_5(N)x(A)_7(N)x(C)11(N)x(AGCAC)_5$ | *C. fimbriata* | *C. pirilliformis*[5] | NO | AY055022 |
| CF15/16[1] | $(CT)_5(N)x(CT)_3(N)x(CT)_3$ sequence rich in T | *C. fimbriata* and *C. albifundus* | *C. albifundus*[4] and *C. pirilliformis*[5] | NO | AY055023 |
| CF17/18[1] | $(CA)_{15}$ sequence rich in GT and T | *C. fimbriata* and *C. albifundus* | *C. albifundus*[4] and *C. pirilliformis*[5] | NO | AY055024 |
| CF21/22[1] | $(T)_8(N)x(T)_6(N)x(C)_2(T)_3C(CT)_2(CCTT)_2C(T)_3C(T)_2C(T)_4$ | *C. fimbriata* and *C. albifundus* | *C. albifundus*[4] and *C. pirilliformis*[5] | NO | AY055025 |
| CF23/24[1] | $TGCA(TG)_{15}$ | *C. fimbriata* and *C. albifundus* | *C. albifundus*[4] and *C. pirilliformis*[5] | NO | AY055026 |
| CfAAG8[2] | $(AAG)_{11}$ | *C. fimbriata* | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], *C. pirilliformis*[5] and *C. platani*[7,8] | NO | AY494859 |
| CfAAG9[2] | $(CAG)_2+(CAG)_7+(AAG)_7$ | *C. fimbriata* and *C. albifundus* | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], *C. pirilliformis*[5] and *C. platani*[7,8] | NO | AY494860 |
| CfCAA9[2] | $(CAA)4(CAG)_2 + (CAG)_2 + (CAG)_6 + (CAG)_4 + (CAG)_2(CAA)_{20}(CAG)_5$ | *C. fimbriata* and *C. albifundus* | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], *C. pirilliformis*[5] and *C. platani*[7,8] | YES | AY494861 |
| CfCAA10[2] | $(CAA)_4(CAG)_6$ | *C. fimbriata* and *C. albifundus* | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9,] and *C. platani*[7,8] | YES | AY494862 |
| CfCAA15[2] | $(CAA)_6(CAG)_8(CAA)_2 + (CAA)_2 + (CAA)_2(CAG)_3(CAA)_2$ | *C. fimbriata* and *C. albifundus* | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], *C. pirilliformis*[5] and *C. platani*[7,8] | NO | AY494863 |

**Table 1 (continued)** Presence of published microsatellite motifs in the *C. fimbriata* and *C. albifundus* genome sequences and those that are transferable to other species based on population studies

| Microsatellite name | Motif in published sequence | In the genome sequence of *C. fimbriata* or *C. albifundus* | Transferable to other species based on population studies | In gene? | GenBank accession number |
|---|---|---|---|---|---|
| CfCAA38[2] | $(CAG/CAA)_{47}$ | *C. fimbriata* and *C. albifundus* | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], *C. pirilliformis*[5] and *C. platani*[7,8] | YES | AY494864 |
| CfCAA80[2] | $(CAG/CAA)_{43} + (CAG)_3 + (CAG)_2 + (CAG)_2 + (CAA)_3$ | *C. fimbriata* (end of contig) and *C. albifundus* (end of contig) | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], and *C. platani*[7,8] | YES | AY494865 |
| CfCAG5[2] | $(CAG/CAA)_{12}$ | *C. fimbriata* | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], *C. pirilliformis*[5] and *C. platani*[7,8] | YES | AY494870 |
| CfCAG15[2] | $(CAG)_4(CAA)_2(CAG)_4 + (CAG)_2 + (CAG)_3(CAA)_6(CAG)_9 + (CAG)_2(CAA)_{10} + (CAG)_4$ | *C. fimbriata* (end of contig) and *C. albifundus* | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], *C. pirilliformis*[5] and *C. platani*[7,8] | NO | AY494871 |
| CfCAG900[2] | $(CAG)_4 + (CAG)_2$ | *C. fimbriata* and *C. albifundus* | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], *C. pirilliformis*[5] and *C. platani*[7,8] | NO | AY494873 |
| CfCAT1[2] | $(CAT)_{11} + (CAT)_2$ | *C. fimbriata* and *C. albifundus* | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], *C. pirilliformis*[5] and *C. platani*[7,8] | NO | AY494866 |
| CfCAT3K[2] | $(CAT)_6 + (CAT)_2$ | *C. fimbriata* | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], and *C. platani*[7,8] | NO | AY494867 |
| CfCAT9X2 | $(CAT)_6$ | *C. fimbriata* | *C. cacaofunesta*[6] *and C. platani*[7,8] | NO | AY494868 |
| CfCAT12002 | $(CAT)_7$ | *C. fimbriata and C. albifundus* | *C. cacaofunesta*[6], *C. fimbriata Brazil*[9], *and C. platani*[7,8] | NO | AY494869 |
| CfGACA602 | $(GACW)_4 + (CACAGCA)_4$ | *C. fimbriata and C. albifundus* | *C. cacaofunesta*[6], *C. fimbriata Brazil*[9], *and C. platani*[7,8] | NO | AY494874 |
| CfGACA6502 | $(TG)_4 + (CA)_2(GACA)_4 + (CCT)_2$ | *C. fimbriata and C. albifundus* | *C. cacaofunesta*[6], *C. fimbriata Brazil*[9], *C. pirilliformis*[5] *and C. platani*[7,8] | YES | AY494875 |
| C. fim 01[3] | $(AC)_{14}$ | *C. fimbriata* and C. albifundus** | - | - | GF100701 |
| C. fim 02[3] | $(TG)_{18}$ | *C. fimbriata* and C. albifundus** | - | - | GF100702 |
| C. fim 03[3] | $(AC)_9$ | *C. fimbriata# and C. albifundus#* | - | - | GF100703 |
| C. fim 04[3] | $(GT)_8$ | *C. fimbriata and C. albifundus* | - | NO | GF100704 |
| C. fim 05[3] | $(AC)_{14}$ | *C. fimbriata* and C. albifundus** | - | - | GF100705 |

**Table 1 (continued)** Presence of published microsatellite motifs in the *C. fimbriata* and *C. albifundus* genome sequences and those that are transferable to other species based on population studies

| Microsatellite name | Motif in published sequence | In the genome sequence of *C. fimbriata* or *C. albifundus* | Transferable to other species based on population studies | In gene? | GenBank accession number |
|---|---|---|---|---|---|
| C. fim 06[3] | $(TG)_{11}$ | *C. fimbriata*# and *C. albifundus*# | - | - | GF100706 |
| C. fim 07[3] | $(AC)_{13}$ | *C. fimbriata*# and *C. albifundus*# | - | - | GF100707 |
| C. fim 08[3] | $(CA)_{15}$ | *C. fimbriata** and *C. albifundus** | - | - | GF100708 |
| C. fim 09[3] | $(AC)_{17}$ | *C. fimbriata*º and *C. albifundus*º | - | - | GF100709 |
| C. fim 10[3] | $(CACT)_{6}$ | *C. fimbriata* | - | NO | GF100710 |
| C. fim 11[3] | $(CA)_{9}(CT)_{7}$ | *C. fimbriata* | - | YES | GF100711 |
| C. fim 13[3] | $(AC)_{14}$ | *C. fimbriata** and *C. albifundus** | - | - | GF100713 |
| C. fim 14[3] | $(AC)_{11}$ | *C. fimbriata** and *C. albifundus** | - | - | GF100714 |
| C. fim 15[3] | $(CA)_{14}$ | *C. fimbriata* | - | NO | GF100715 |
| C. fim 16[3] | $(AC)_{21}$ | *C. fimbriata (end of contig)*^ and *C. albifundus*^ | - | ? | GF100716 |
| C. fim 17[3] | $(AC)_{17}$ | *C. fimbriata*º and *C. albifundus*º | - | YES | GF100717 |
| C. fim 18[3] | $(TG)_{16}$ | *C. fimbriata (end of contig)*^ and *C. albifundus*^ | - | ? | GF100718 |
| C. fim 19[3] | $(AC)_{13}$ | *C. fimbriata* and *C. albifundus* | - | YES | GF100719 |
| C. fim 20[3] | $(CT)_{17}$ | *C. fimbriata* and *C. albifundus (end of contig)* | - | NO | GF100720 |
| C. fim 21[3] | $(TG)_{10}$ | *C. fimbriata* | - | NO | GF100721 |

[1]Barnes et al. (2001), [2]Steimel et al. (2004), [3]Rizatto et al. (2010), [4]Barnes et al. (2005), [5]Kamgan Nkuekam et al. (2009), [6]Engelbrecht et al. (2007), [7]Engelbrecht et al. (2004), [8]Ocasio-Morales et al. (2007), [9]Fereirra et al. (2010). *, #, º and ^ indicate sequences that align to the same position in each genome.

**Table 2** Isolates of *Ceratocystis* species used in this study

| Species | Isolate number[1] | Alternative number | Host | Geographical origin | Collector |
|---|---|---|---|---|---|
| *C. acaciivora* | CMW 22595 | | *Acacia mangium* | Indonesia | M. Tarigan |
| *C. acaciivora* | CMW 22621 | | *A. mangium* | Indonesia | M. Tarigan |
| *C. acaciivora* | CMW 22562 | | *A. mangium* | Indonesia | M. Tarigan |
| *C. acaciivora* | CMW 22563 | | *A. mangium* | Indonesia | M. Tarigan |
| *C. albifundus* | CMW 4068 | | *Acacia mearnsii* | South Africa | J. Roux |
| *C. albifundus* | CMW 4090 | | *A. mearnsii* | South Africa | J. Roux |
| *C. albifundus* | CMW 5329 | CBS119681 | *A. mearnsii* | Uganda | J. Roux |
| *C. albifundus* | CMW 14159 | | *Burkea africana* | South Africa | J. Roux & L. Labuschagne |
| *C. albifundus* | CMW 15760 | | *A. mearnsii* | Uganda | J. Roux |
| *C. albifundus* | CMW 17274 | | *Faurea saligna* | South Africa | J. Roux |
| *C. atrox* | CMW 19385 | CBS120518, PREM 59012 | *Eucalyptus grandis* | Australia | M.J. Wingfield |
| *C. atrox* | CMW 19389 | CBS120225 | *E. grandis* | Australia | M.J. Wingfield |
| *C. cacaofunesta* | CMW 14809 | CBS115169 | *Theobroma cacao* | Ecuador | P.W.C. Crous |
| *C. cacaofunesta* | CMW 15051 | CBS152.62 | *T. cacao* | Costa Rica | A.J. Hansen |
| *C. cacaofunesta* | CMW 26375 | CBS115172 | *T. cacao* | Brazil | T.C. Harrington |
| *C. caryae* | CMW 14793 | CBS114716 | *Carya cordiformis* | USA | P.W.C. Crous |
| *C. caryae* | CMW 14808 | CBS115168 | *Carya ovata* | USA | P.W.C. Crous |
| *C. colombiana* | CMW 5751 | CBS121792 | *Coffee arabica* | Colombia | M.J. Wingfield |
| *C. colombiana* | CMW 5761 | CBS121791 | *C. arabica* | Colombia | M.J. Wingfield |
| *C. colombiana* | CMW 9565 | CBS121790 | Soil in coffee plantation | Colombia | B. Castro |
| *C. colombiana* | CMW 11280 | | *Schizolobium parahybum* | Colombia | Unknown |
| *C. curvata* | CMW 22433 | CBS122513 | *Eucalyptus deglupta* | Ecuador | M.J. Wingfield & M. van Wyk |
| *C. curvata* | CMW 22435 | CBS122604 | *E. deglupta* | Ecuador | M.J. Wingfield & M. van Wyk |
| *C. curvata* | CMW 22442 | CBS122603 | *E. deglupta* | Ecuador | M.J. Wingfield & M. van Wyk |
| *C. diversiconidia* | CMW 22445 | CBS123013 | *Terminalia ivorensis* | Ecuador | M.J. Wingfield & M. van Wyk |
| *C. diversiconidia* | CMW 22446 | | *T. ivorensis* | Ecuador | M.J. Wingfield & M. van Wyk |
| *C. diversiconidia* | CMW 22448 | CBS122605 | *T. ivorensis* | Ecuador | M.J. Wingfield & M. van Wyk |
| *C. ecuadoriana* | CMW 22092 | CBS124020 | *E. deglupta* | Ecuador | M.J. Wingfield & M. van Wyk |
| *C. ecuadoriana* | CMW 22093 | CBS 124021 | *E. deglupta* | Ecuador | M.J. Wingfield & M. van Wyk |
| *C. ecuadoriana* | CMW 22097 | CBS 124022 | *E. deglupta* | Ecuador | M.J. Wingfield & M. van Wyk |
| *C. ecuadoriana* | CMW 22405 | | *E. deglupta* | Ecuador | M.J. Wingfield & M. van Wyk |

**Table 2 (continued)** Isolates of *Ceratocystis* species used in this study

| Species | Isolate number[1] | Alternative number | Host | Geographical origin | Collector |
|---|---|---|---|---|---|
| *C. fimbriata* | CMW 1547 | CBS123010 | *Ipomoea batatas* | Papa New Guinea | E.H.C.M. Kenzie & F.M. Quinn |
| *C. fimbriata* | CMW 14799 | CBS114723 | *I. batatas* | USA | P.W.C. Crous |
| *C. fimbriata* | CMW 15049 | CBS141/37 | *I. batatas* | USA | C.F. Andrus |
| *C. fimbriatomima* | CMW 24174 | CBS121786 | *Eucalyptus* sp. | Venezuela | M.J. Wingfield |
| *C. fimbriatomima* | CMW 24176 | CBS121787 | *Eucalyptus* sp. | Venezuela | M.J. Wingfield |
| *C. fimbriatomima* | CMW 24377 | | *Eucalyptus* sp. | Venezuela | M.J. Wingfield & M. van Wyk |
| *C. fimbriatomima* | CMW 24378 | | *Eucalyptus* sp. | Venezuela | M.J. Wingfield & M. van Wyk |
| *C. fimbriatomima* | CMW 24379 | | *Eucalyptus* sp. | Venezuela | M.J. Wingfield & M. van Wyk |
| *C. larium* | CMW 25434 | CBS122512 | *Styrax benzoin* | Indonesia | M.J. Wingfield & M. van Wyk |
| *C. larium* | CMW 25435 | CBS122606 | *S. benzoin* | Indonesia | M.J. Wingfield & M. van Wyk |
| *C. larium* | CMW 25436 | CBS122607 | *S. benzoin* | Indonesia | M.J. Wingfield & M. van Wyk |
| *C. larium* | CMW 25437 | | *S. benzoin* | Indonesia | M.J. Wingfield & M. van Wyk |
| *C. manginecans* | CMW 13851 | CBS121659, PREM59612 | *Mangifera indica* | Oman | M. Deadman |
| *C. manginecans* | CMW 13852 | CBS121660, PREM59613 | *Hypocryphalus mangifera* | Oman | M. Deadman |
| *C. manginecans* | CMW 15314 | | *M. indica* | Oman | A. Al-Adawi |
| *C. manginecans* | CMW 15317 | | *H. mangifera* | Oman | A. Al-Adawi |
| *C. manginecans* | CMW 23634 | CBS121661, PREM59614 | *H. mangifera* | Pakistan | A. Al-Adawi |
| *C. moniliformis** | CMW 10134 | CBS118127 | *E. grandis* | South Africa | M. van Wyk |
| *C. neglecta* | CMW 11284 | CBS121349 | *E. grandis* | Colombia | C. Rodas |
| *C. neglecta* | CMW 11285 | | *E. grandis* | Colombia | C. Rodas |
| *C. neglecta* | CMW 17808 | CBS121789 | *E. grandis* | Colombia | C. Rodas |
| *C. neglecta* | CMW 18194 | | *E. grandis* | Colombia | C. Rodas |
| *C. obpyriformis* | CMW 23806 | CBS122609 | *A. mearnsii* | South Africa | R.N. Heath |
| *C. obpyriformis* | CMW 23807 | CBS122608 | *A. mearnsii* | South Africa | R.N. Heath |
| *C. papillata* | CMW 8850 | CBS121794 | *Tangelo mineola* | Colombia | B. Castro |
| *C. papillata* | CMW 8856 | CBS121793 | *Citrus limon* | Colombia | B. Castro |
| *C. papillata* | CMW 8857 | CBS123009 | *Annona muricata* | Colombia | B. Castro |
| *C. papillata* | CMW 28662 | CBS121795 | *S. parahybum* | Colombia | B. Castro |
| *C. pirilliformis* | CMW 6579 | CBS118128 | *Eucalyptus nitens* | Australia | M.J. Wingfield |
| *C. pirilliformis* | CMW 6583 | CBS118596 | *E. nitens* | Australia | M.J. Wingfield |

**Table 2 (continued)** Isolates of *Ceratocystis* species used in this study

| Species | Isolate number[1] | Alternative number | Host | Geographical origin | Collector |
|---|---|---|---|---|---|
| *C. pirilliformis* | CMW 12671 | | *E. grandis* | South Africa | H. Hatting & J. Roux |
| *C. pirilliformis* | CMW 16511 | | *Eucalyptus* sp. | South Africa | G. Kamgan Nkuekam |
| *C. pirilliformis* | CMW 28200 | | *Eucalyptus* logs | South Africa | G. Kamgan Nkuekam |
| *C. platani* | CMW 1896 | | *Platanus* sp. | Switzerland | O. Petrini |
| *C. platani* | CMW 14802 | CBS115162 | *Platanus occidentalis* | USA | P.W.C. Crous |
| *C. platani* | CMW 23450 | | *Platanus orientalis* | Greece | P. Tspopelas |
| *C. platani* | CMW 23918 | | *P. occidentalis* | Greece | M.J. Wingfield |
| *C. platani* | CMW 26380 | CBS115162, C1317 | *P. occidentalis* | USA | T.C. Harrington |
| *C. polychroma* | CMW 11424 | CBS115778, PREM57818 | *Syzygium aromaticum* | Indonesia | E.C.Y Liew & M.J. Wingfield |
| *C. polychroma* | CMW 14281 | | *S. aromaticum* | Indonesia | E.C.Y. Liew |
| *C. polyconidia* | CMW 23818 | CBS122290 | *A. mearnsii* | South Africa | R.N. Heath |
| *C. populicola* | CMW 14789 | CBS119.78 | *Populus* sp. | Poland | J. Gremmen |
| *C. smalleyi* | CMW 14800 | CBS114724, C684 | *C. cordiformis* | USA | G. Smalley |
| *C. tanganyicensis* | CMW 15992 | CBS122293 | *A. mearnsii* | Tanzania | R.N. Heath |
| *C. tanganyicensis* | CMW 15999 | CBS122294 | *A. mearnsii* | Tanzania | R.N. Heath |
| *C. tsitsikammensis* | CMW 13982 | | *Rapanea melanophloeos* | South Africa | J. Roux |
| *C. tsitsikammensis* | CMW 14275 | | *R. melanophloeos* | South Africa | G. Kamgan Nkuekam |
| *C. tsitsikammensis* | CMW 14276 | PREM59424 | *R. melanophloeos* | South Africa | G. Kamgan Nkuekam |
| *C. tsitsikammensis* | CMW 14280 | | *Ocotea bullata* | South Africa | G. Kamgan Nkuekam |
| *C. variospora* | CMW 20935 | CBS114715, C1843 | *Quercus alba* | USA | J. Johnson |
| *C. variospora* | CMW 26384 | CBS773.73 | *Quercus ellipsoidalis* | USA | R. Campbell |
| *C. variospora* | CMW 26386 | CBS114714, C1846 | *Quercus robur* | USA | J. Johnson |
| *C. virescens*[#] | CMW 3225 | C254 | *Acer saccharum* | USA | D. Houston |
| *C. virescens*[#] | CMW 11160 | C252 | *A. saccharum* | USA | D. Houston |
| *C. virescens*[#] | CMW 11164 | CBS123166, C69 | *Fagus americana* | USA | D. Houston |
| *C. virescens*[#] | CMW 17335 | C525 | *A. saccharum* | Unknown | Unknown |
| *C. virescens*[#] | CMW 17339 | C261 | *A. saccharum* | USA | D. Houston |
| *C. zombamontana* | CMW 15235 | CBS122297 | *E. grandis* | Malawi | R.N. Heath |
| *C. zombamontana* | CMW 15236 | CBS122296 | *E. grandis* | Malawi | R.N. Heath |

[1]Culture collection of the Forestry and Agricultural Biotechnology Institute (FABI), University of Pretoria, Pretoria, 0002, South Africa. * *C. moniliformis* forms part of the *C. moniliformis sensu lato* species complex. [#] *C. virescens* forms part of the *C. coerulescens sensu lato* species complex

**Table 3** Primers designed in this study to amplify microsatellites within predicted genes

| Locus | Contig | Microsatellite motif | Primer | Primer sequences | Fluorescent label | Actual size in *C. fimbriata* genome (bp) | Result |
|---|---|---|---|---|---|---|---|
| CF_CAA/CAG80 | 3789 | $(CAA/CAG)_8$ | CAA/CAG80F | 5'- catcagctgctcctgtcgta -3' | PET | 227 | Polymorphic |
| | | | CAA/CAG80R | 5'- aggcgggtagtcggagtaat -3' | - | | |
| CF_CAAG5 | 3933 | $(CAAG)_5$ | CAAG5F | 5'- cccatctgcttttcttcctg -3' | VIC | 198 | Polymorphic |
| | | | CAAG5R | 5'- gggttgtgcgtagaggatgt -3' | - | | |
| CF_GTT50 | 17 | $(GTT)_5$ | GTT50F | 5'- cagcgagcaaaaatcaaaca -3' | 6-FAM | 235 | Polymorphic |
| | | | GTT50R | 5'- tgtctcaggcgcaaatacag -3' | - | | |
| CF_GAT5 | 211 | $(GAT)_5$ | GAT5F | 5'- gtttttgacgcacgtagagc -3' | PET | 240 | Polymorphic |
| | | | GAT5R | 5'- tggtcgtatagcgacgtgag -3' | - | | |
| CF_CAGAAG5 | 351 | $(CAGAAG)_5$ | CAGAAG5F | 5'- gggagtgggtatgagtgtgg -3' | VIC | 216 | Polymorphic |
| | | | CAGAAG5R | 5'- gctgctgctgctagttcaga -3' | - | | |
| CF_GCT11 | 351 | $(GCT)_{11}$ | GCT11F | 5'- gaatgctggagcttgggtag -3' | 6-FAM | 235 | Polymorphic |
| | | | GCT11R | 5'- ggcgatgaacatggagagat -3' | - | | |
| CF_CTCTCTGT5 | 351 | $(CTCTCTGT)_5$ | CTCTCTGTF | 5'- tgaatgctgtgggagatgaa -3' | NED | 237 | Polymorphic |
| | | | CTCTCTGTR | 5'- aacatgactgtcgggaggag -3' | - | | |
| CF_CAA/CAG81 | 920 | $(CAA/CAG)_8$ | CAA/CAG81F | 5'- ccatggaccccatcaactac -3' | PET | 226 | Polymorphic |
| | | | CAA/CAG81R | 5' - gcagccaaagttccaaagac -3' | - | | |
| CF_CCG5 | 1506 | $(CCG)_5$ | CCG5F | 5'- agcggatggaacaaacagac -3' | NED | 192 | Polymorphic |
| | | | CCG5R | 5'- aggtctccgagggacgtatt -3' | - | | |
| CF_CAA/CAG24 | 2130 | $(CAA/CAG)_{24}$ | CAA/CAG24F | 5'- acccacagtcacagcatttg -3' | NED | 175 | Polymorphic |
| | | | CAA/CAG24R | 5'- gagccttgctgagtgtgtga -3' | - | | |
| CF_CT7 | 4768 | $(CT)_7$ | CT7F | 5'- cccatcgatcttcacacaaa -3' | - | 184 | Polymorphic |
| | | | CT7R | 5'- attggactttccgacaccag -3' | - | | |
| CF_TC60 | 3711 | $(TC)_6$ | TC60F | 5'- tcagaggatttgatgcgtga -3' | - | 169 | Polymorphic |
| | | | TC60R | 5'- ggtctgggatggtgatgaat -3' | - | | |
| CF_CTT6 | 3711 | $(CTT)_6$ | CTT6F | 5'- gaacaggccgagtttgagag -3' | - | 225 | Polymorphic |
| | | | CTT6R | 5'- accgccttcatcaagatcag -3' | - | | |
| CF_TG5 | 3711 | $(TG)_5$ | TG5F | 5'- cgtctttacgaggaccaagc -3' | - | 226 | Monomorphic |
| | | | TG5R | 5'- ctaaggcaaaacgcctcaag -3' | - | | |

**Table 3 (continued)** Primers designed in this study to amplify microsatellites within predicted genes

| Locus | Contig | Microsatellite motif | Primer | Primer sequences | Fluorescent label | Actual size in *C. fimbriata* genome (bp) | Result |
|---|---|---|---|---|---|---|---|
| CF_CAG7 | 3789 | $(CAG)_7$ | CAG7F | 5'- tgcttgaagaaatggtgcag -3' | - | 206 | Polymorphic |
| | | | CAG7R | 5'- ccacatggctacacgatcaa -3' | - | | |
| CF_CG5 | 3858 | $(CG)_5$ | CG5F | 5'- gaaacgacagacagcacagc -3' | - | 205 | Monomorphic |
| | | | CG5R | 5'- gttccgccttttttctttcc -3' | - | | |
| CF_TCC7 | 3868 | $(TCC)_7$ | TCC7F | 5'- agttggatacgaggcaatgg -3' | - | 160 | Polymorphic |
| | | | TCC7R | 5'- gtcccttcatcgccagaata -3' | - | | |
| CF_TCC5 | 11 | $(TCC)_5$ | TCC5F | 5'- agccagactcctcctcatca -3' | - | 240 | Monomorphic |
| | | | TCC5R | 5'- ctgccgctgagctagaagtt -3' | - | | |
| CF_GTT6 | 351 | $(GTT)_6$ | GTT6F | 5'- tacgacccttgctgtgactg -3' | - | 181 | Double band |
| | | | GTT6R | 5'- ggggctgctatagggacttc -3' | - | | |
| CF_GCA5 | 543 | $(GCA)_5$ | GCA5F | 5'- ccttcttcgcaaaggacttg -3' | - | 397 | Polymorphic |
| | | | GCA5R | 5'- tcacaatgaagtcgctcctg -3' | - | | |
| CF_ACA5 | 550 | $(ACA)_5$ | ACA5F | 5'- cagcagcagcagtagcaaag -3' | - | 151 | Polymorphic |
| | | | ACA5R | 5'- gtggtggtggctagatcgtt -3' | - | | |
| CF_TC10 | 920 | $(TC)_{10}$ | TC10F | 5'- ggccttttcccattcttttt -3' | - | 218 | Polymorphic |
| | | | TC10R | 5'- gaagagcaagaggcaggaga -3' | - | | |
| CF_CAA/CAG14 | 920 | $(CAA/CAC/CAG)_{14}$ +$(TCC)_5$ | CAA/CAG14F | 5'- atgaacaccagcagcaacac -3' | - | 356 | Polymorphic |
| | | | CAA/CAG14R | 5'- attggggctgatacaagctg -3' | - | | |
| CF_GAA6 | 969 | $(GAA)_6$ | GAA6F | 5'- ccaccatctccaacccatac -3' | - | 231 | Monomorphic |
| | | | GAA6R | 5'- attgccgatgctaccaagac -3' | - | | |
| CF_TTC5 | 969 | $(TTC)_5$ | TTC5F | 5' - ccaaccttcgtggttgagtt -3' | - | 199 | Monomorphic |
| | | | TCC5R | 5'- aacgagggatcaaagcaaga -3' | - | | |
| CF_CCT6 | 969 | $(CCT)_6$ | CCT6F | 5'- cctccacaatcacctccatt -3' | - | 228 | Monomorphic |
| | | | CCT6R | 5'- gcgactgtgggttttgtttt -3' | - | | |
| CF_GAT7 | 978 | $(GAT)_7$ | GAT7F | 5' - aatctcacccgtttgattgg -3' | - | 227 | Polymorphic |
| | | | GAT7R | 5'- atgctgcgagggatcattag -3' | - | | |
| CF_GCTCCC6 | 980 | $(GCTCCC)_6$ + $(GTG)_8$ | GCTCCC6F | 5'- tccgaacgaagattgaggac -3' | - | 293 | Polymorphic |
| | | | GCTCCC6R | 5'- ccccctcatagccaatatca -3' | - | | |
| CF_CAA/CAG11 | 2125 | $(CAA/CAG)_{11}$ | CAA/CAG11F | 5'- cggcgtcttagaccactctc -3' | - | 229 | Polymorphic |
| | | | CAA/CAG11R | 5'- ggaagacattgccgaggata -3' | | | |

**Table 3 (continued)** Primers designed in this study to amplify microsatellites within predicted genes

| Locus | Contig | Microsatellite motif | Primer | Primer sequences | Fluorescent label | Actual size in *C. fimbriata* genome (bp) | Result |
|---|---|---|---|---|---|---|---|
| CF_TCG6 | 2125 | $(TGC)_6$ | TCG6F | 5'- agaggtggtggcatcacagt -3' | - | 200 | Monomorphic |
| | | | TCG6R | 5'- tgttctacgaggccctgact -3' | - | | |
| CF_TC61 | 2136 | $(TC)_6$ | TC61F | 5'- ttcaagacccacctcaacct - 3' | - | 160 | Monomorphic |
| | | | TC61R | 5'- ccagaaaccattgctcgatt -3' | - | | |
| CF_GTT51 | 2364 | $(GTT)_5$ | GTT51F | 5'- taacacgcggactccctatc -3' | - | 156 | Polymorphic |
| | | | GTT51R | 5'- cagtggacgctgctgataaa -3' | - | | |
| CF_GGC7 | 2567 | $(GGC)_7$ | GGC7F | 5'- cggcacatccttctaacaca -3' | - | 246 | Polymorphic |
| | | | GGC7R | 5'- gctgtttccccctcttcttc -3' | - | | |
| CF_GAC5 | 2227 | $(GAC)_5$ | GAC5F | 5'- ggtacaagcacattgcctca -3' | - | 239 | Polymorphic |
| | | | GAC5R | 5'- gggcttagcagccttcttct -3' | - | | |
| CF_GA5 | 13 | $(GA)_5$ | GA5F | 5'- agcagcaaccagccactatt -3' | - | 190 | Monomorphic |
| | | | GA5R | 5'- ggttgctgacagagggtgtt -3' | - | | |
| CF_CGG5 | 152 | $(CGG)_5$ | CGG5F | 5'- ccataaggagcttgcgtagg -3' | - | 242 | Polymorphic |
| | | | CGG5R | 5'- atgcctccacacgactatcc -3' | - | | |

**Table 4** Published microsatellite sequences identified in the *C. fimbriata* genome that were identified within predicted proteins

| Locus | Motif in published sequence | Motif in *C. fimbriata* genome | Region | Predicted protein | E-value | Coverage |
|---|---|---|---|---|---|---|
| AG17/18[1] | $(T)_5(C)_2(CT)_2T(CTT)_6(T)_2(C)_3TC(T)_3$[1] | $(T)_5(C)(CTT)_4(T)_2(C)_3TC(T)_3$ | Coding | Fungal hypothetical protein | 6.00E-18 | 95% |
| CfCAG5[2] | $(CAG/CAA)_{12}$[2] | $(CGA/CAA)_9$ | Coding | Fungal hypothetical protein | 2.00E-135 | 49% |
| CfCAA9[2] | $(CAA)_4(CAG)_2 + (CAG)_2 + (CAG)_6 + (CAG)_4 + (CAG)_2(CAA)_{20}(CAG)_5$ | $(CGA)_7(CAA)_{12}(CGA)_5 + (CAA) + (CGA)_4 + (CAA)_3 + (CGC)(CAC)_3(CGA)$ | Coding | Fungal-specific transcription factor | 0 | 71% |
| CfCAA38[2] | $(CAG/CAA)_{47}$[2] | $(CGA/CAA)_{15}$ | Coding | CCR4-NOT complex component | 0 | 100% |
| CfCAA80[2] | $(CAG/CAA)_{43} + (CAG)_3 + (CAG)_2 + (CAG)_2 + (CAA)_3$ | (end of contig) … $(CAA)_3$ | Intron | GH22395 (*Drosophila grimshawi*) | 0.13 | 52% |
| CfGACA650[2] | $(TG)_4 + (CA)_2(GACA)_4 + (CCT)_2$ | $(TG)_4 + (CA)_2(GACA)_3 + (CA)_{11}$ | Coding | Hypothetical protein (*Drosophila*) | 3.1 | 55% |
| CfCAA10[2] | $(CAA)_4(CAG)_6$[2] | $(CAA)_5(CAG)_7$ | Coding | Protein kinase A catalytic subunit | 2.00E-177 | 49% |
| Cfim11[3] | $(CA)_9(CT)_7$ | $(CA)_9(CT)_7$ | Coding | Patatin family protein (*Spirochaeta thermophila*) | 4.8 | 48% |
| Cfim17[3] | $(AC)_{17}$ | $(AC)_{11}$ | Coding | SPRY-containing protein (*Glomerella graminicola*) | 2.00E-17 | 100% |
| Cfim19[3] | $(AC)_{13}$ | $(AC)_{12}$ | Coding | Subtilase (*Glomerella graminicola*) | 1.00E-25 | 84% |

[1]Barnes et al. (2001), [2]Steimel et al. (2004), [3]Rizatto et al. (2010)

**Table 5** Microsatellite markers developed in this study that were detected within putative proteins (Loci highlighted in grey are those used for Genescan analysis).

| Locus | Region | Predicted protein | Coverage | E-value |
|---|---|---|---|---|
| CF_CAA/CAG80 | Coding | Hypothetical protein (*Verticillium albo-atrum*) | 100% | 6.00E-13 |
| CF_CAAG5 | Intron | Hypothetical protein (*Parabacteroides johnsonii*) | 85% | 3.50 |
| CF_GTT50 | Coding | SCF E3 ubiquitin ligase complex F-box protein (*Verticillium albo-atrum*) | 79% | 0.13 |
| CF_GAT5 | Coding | Hypothetical protein (*Verticillium albo-atrum*) | 94% | 7.00E-27 |
| CF_CAGAAG5 | Coding | No similarity | - | - |
| CF_GCT11 | Coding | No similarity | - | - |
| CF_CTCTCTGT5 | Coding | Hypothetical protein (*Magnaporthe oryzae*) | 56% | 8.00E-05 |
| CF_CAA/CAG81 | Coding | No similarity | - | - |
| CF_CCG5 | Coding | Snf7 family protein (*Glomerella graminicola*) | 100% | 7.00E-32 |
| CF_CAA/CAG24 | Coding | No similarity | - | - |
| CF_CT7 | Intron | Ras GTPase Rab11 (*Neosartorya fischeri*) | 100% | 3.00E-21 |
| CF_TC60 | Coding | No similarity | - | - |
| CF_CTT6 | Coding | Hypothetical protein (*Podospora anserina*) | 69% | 5.00E-05 |
| CF_TG5 | Coding | No similarity | - | - |
| CF_CAG7 | Coding | Major facilitator superfamily transporter (*Ferroplasma acidarmanus*) | 48% | 3.70 |
| CF_CG5 | Coding | ORF1ab polyprotein (*Bat coronavirus* 1B) | 80% | 2.40 |
| CF_TCC7 | Coding | Disulfide isomerase (*Erythrobacter litoralis*) | 75% | 5.00 |
| CF_TCC5 | Coding | Hypothetical protein (*Candida albicans*) | 85% | 0.33 |
| CF_GTT6 | Coding | No similarity | - | - |
| CF_GCA5 | Coding | Acid phosphatase (*Metarhizium acridum*) | 60% | 5.00E-04 |
| CF_ACA5 | Coding | No similarity | - | - |
| CF_TC10 | Coding | No similarity | - | - |
| CF_CAA/CAG14 | Coding | No similarity | - | - |
| CF_GAA6 | Coding | Hypothetical protein (*Podospora anserina*) | 100% | 3.00E-23 |
| CF_TTC5 | Intron | Hypothetical protein (*Podospora anserina*) | 100% | 3.00E-23 |
| CF_CCT6 | Coding | No similarity | - | - |
| CF_GAT7 | Coding | No similarity | - | - |
| CF_GCTCCC6 | Coding | No similarity | - | - |
| CF_CAA/CAG11 | Coding | No similarity | - | - |
| CF_TCG6 | Coding | No similarity | - | - |
| CF_TC61 | Coding | Hypothetical protein (*Podospora anserina*) | 93% | 7.00E-11 |
| CF_GTT51 | Coding | Sulfate permease 2 (*Arthroderma gypseum*) | 94% | 1.00E-11 |
| CF_GGC7 | Coding | Hypothetical protein (*Giberella zeae*) | 46% | 1.00E-04 |
| CF_GAC5 | Coding | Hypothetical protein (*Nectria haematococca*) | 94% | 9.00E-27 |
| CF_GA5 | Coding | Patched protein (*Drosophila melanogaster*) | 93% | 0.83 |
| CF_CGG5 | Coding | Hypothetical protein (*Verticillium albo-atrum*) | 80% | 5.00E-22 |

**Table 6** Genescan analysis of microsatellite loci fragment sizes for each isolate of *Ceratocystis* used in this study. Grey boxes indicate conflicting allele sizes within a species, x indicates no PCR product and 0 indicates double bands.

| Species | CMW | CF_CAA/ CAG80 | CF_CAAG5 | CF_GTT50 | CF_GAT5 | CF_CAG AAG5 | CF_GCT 11 | CF_CTCTC TGT5 | CF_CAA/ CAG81 | CF_CCG5 | CF_CAA/ CAG24 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *C. acaciivora* | 22562 | x | x | x | 236.7 | 231.8 | x | x | 221.4 | x | 173.9 |
| *C. acaciivora* | 22563 | 227.9 | 196.7 | x | 236.8 | x | x | x | 221.5 | x | 173.5 |
| *C. acaciivora* | 22595 | 227.9 | 189 | x | 236.9 | x | x | 236.9 | 221.4 | 190.2 | 173.7 |
| *C. acaciivora* | 22621 | 227.9 | x | 233.5 | 236.8 | 220 | x | 237.1 | 221.5 | 190.3 | 167.7 |
| *C. albifundus* | 15760 | x | 192.9 | 248.3 | 254 | 207.4 | 238 | x | 215.6 | 189.3 | 152.6 |
| *C. albifundus* | 4068 | x | x | 254.3 | 248.3 | 196.4 | x | x | 218.8 | 189.1 | 156.1 |
| *C. albifundus* | 5329 | x | x | 245.2 | x | 231.7 | 237.6 | x | 218.7 | 189.6 | 153 |
| *C. albifundus* | 14159 | x | x | x | x | 196 | x | 190.3 | x | 189.3 | x |
| *C. albifundus* | 4090 | x | x | 254.1 | x | 196.8 | x | x | 218.7 | 189.2 | 152.5 |
| *C. atrox* | 19389 | 242.5 | x | 247.9 | x | 208.6 | 189.1 | 194.5 | 218.3 | 181.2 | 123.2 |
| *C. atrox* | 19385 | 246 | x | 248.4 | x | 208.7 | 201.4 | 153.3 | 218.4 | 181.1 | 122.8 |
| *C. cacaofunesta* | 14809 | 222.1 | 205.1 | 230.1 | 236.8 | 196.8 | 237.4 | 194.3 | 224.3 | 190.6 | 159 |
| *C. cacaofunesta* | 26375 | 222.1 | 195.6 | 233.1 | 232.7 | 208.2 | 238.2 | 194 | 230.2 | 196.3 | 161.9 |
| *C. cacaofunesta* | 15051 | 222 | 197.4 | 230.1 | 228.4 | 196.5 | 237.7 | 194.1 | 224.3 | 190 | 158.9 |
| *C. caryae* | 14793 | x | x | 223.9 | x | 200 | 348.3 | x | 250.4 | x | 156.5 |
| *C. caryae* | 14808 | x | x | 223.7 | 245.9 | x | 304.8 | x | 256.4 | x | x |
| *C. colombiana* | 5751 | 239.5 | 196.7 | 236.8 | 239.8 | 244.5 | 234.9 | 267.9 | 232.9 | x | 182.8 |
| *C. colombiana* | 11280 | 239.6 | 197.6 | 236.9 | 239.8 | 226 | 234.4 | 307.9 | 232.8 | 187.4 | 182.8 |
| *C. colombiana* | 9565 | 248.6 | 197.6 | 236.5 | 239.8 | 231.6 | x | 275.9 | 232.9 | 187.4 | 182.7 |
| *C. colombiana* | 5761 | 239.7 | 197.1 | 236.5 | 240 | 244.4 | 234.9 | 267.8 | 233 | 187 | 182.7 |
| *C. curvata* | 22433 | 225.1 | 192.6 | 233.1 | 246.1 | 214.4 | 256.2 | 240.7 | 221.8 | 190.2 | 167.7 |
| *C. curvata* | 22435 | 225.1 | 193.1 | 233.2 | 246.1 | 213.4 | 256 | 241.1 | 221.5 | 190.4 | 168 |
| *C. curvata* | 22442 | 225 | 193.4 | 232.9 | 246 | 213.8 | 256.3 | 241 | 226.3 | 190.3 | 168 |
| *C. diversiconidia* | 22446 | 222.3 | 185 | 242.1 | 236.8 | 239.1 | 219.8 | 179.6 | 215.4 | 190.8 | 156 |
| *C. diversiconidia* | 22445 | 222.2 | 184.7 | 242.1 | 236.9 | 238.7 | 219.2 | 180 | 215.2 | 189.8 | 155.7 |
| *C. diversiconidia* | 22448 | 222.2 | 184.9 | 242 | 236.9 | 239.1 | 219.4 | 179.9 | 215.1 | 190.3 | 155.7 |

**Table 6 (continued)** Genescan analysis of microsatellite loci fragment sizes for each isolate of *Ceratocystis* used in this study. Grey boxes indicate conflicting allele sizes within a species, x indicates no PCR product and 0 indicates double bands.

| Species | CMW | CF_CAA/ CAG80 | CF_CAAG5 | CF_GTT50 | CF_GAT5 | CF_CAG AAG5 | CF_GCT 11 | CF_CTCTC TGT5 | CF_CAA/ CAG81 | CF_CCG5 | CF_CAA/ CAG24 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *C. ecuadoriana* | 22405 | 230.9 | 192.8 | 236.5 | 240 | 205.8 | 244.4 | 185.9 | 218.3 | 190 | 165.1 |
| *C. ecuadoriana* | 22092 | 239.8 | 192.9 | 236.8 | 248.3 | 206 | 244.3 | 185.9 | 218.4 | 190.2 | 164.5 |
| *C. ecuadoriana* | 22097 | 230.8 | 193.6 | 236.2 | 249.3 | 205.9 | 244 | 185.9 | 218.3 | 190.3 | 165 |
| *C. ecuadoriana* | 22093 | 230.8 | 193.2 | 235.9 | 249.3 | 205.4 | 244.2 | 186 | 218.3 | 190 | 164.4 |
| *C. fimbriata* | 1547 | 225 | 193.2 | 233.3 | 239.9 | 214 | 231.1 | 234.7 | 224.3 | 187.5 | 168 |
| *C. fimbriata* | 14799 | 225 | 193.4 | 233.2 | 240.6 | 214.2 | 231.6 | 234.6 | 224.4 | 187.4 | 167.9 |
| *C. fimbriata* | 15049 | 225.1 | 193 | 232.9 | 239.6 | 214.4 | 231.6 | 234.6 | x | 187.1 | 167.6 |
| *C. fimbriatomima* | 24377 | 222 | 200.9 | 233.3 | 240 | 208.3 | 237.3 | 192.4 | 224.3 | 190.1 | 170.9 |
| *C. fimbriatomima* | 24174 | 221.9 | 200.7 | 233.3 | 239.9 | 208.5 | 237.2 | x | 209.6 | 190 | 170.9 |
| *C. fimbriatomima* | 24176 | 222 | 200.7 | 233.7 | x | 208.3 | 237.3 | 192.3 | 224.3 | 190.2 | 170.5 |
| *C. fimbriatomima* | 24378 | 222.1 | 200.8 | 232.9 | 240 | 208.5 | 237 | 192.3 | 224.2 | 190.3 | 170.6 |
| *C. fimbriatomima* | 24379 | 221.7 | 200.9 | 233.5 | x | x | 237.3 | x | 224.3 | 190.5 | 171.2 |
| *C. larium* | 25435 | 242.5 | x | 233.5 | 234.7 | 208.7 | 189.1 | x | 217.5 | 203.7 | 149.6 |
| *C. larium* | 25434 | 242.5 | 191.8 | 233.3 | 233.8 | 208.5 | 188.9 | 194.6 | 217.7 | 203.5 | x |
| *C. larium* | 25437 | 242.5 | 192.8 | 233.3 | 234.5 | x | 188.9 | 194.5 | 218.7 | 203.6 | 149.7 |
| *C. larium* | 25436 | 242.5 | 192.9 | 233.2 | 234.4 | 208.8 | 188.9 | x | x | 203.6 | 150.3 |
| *C. manginecans* | 23634 | 227.9 | 197 | 233.2 | 236.6 | 220.2 | 240.5 | 236.9 | 221.3 | 190.1 | 173.4 |
| *C. manginecans* | 13852 | 227.7 | x | 233.7 | 236.3 | 219.9 | 240.4 | 236.9 | 221.4 | 190.3 | 173.6 |
| *C. manginecans* | 15317 | 228 | 196.6 | 233.6 | 236.5 | 220.2 | 240.4 | 236.8 | 221.4 | 190.2 | 173.5 |
| *C. manginecans* | 15314 | 228 | 197.1 | 233.2 | 236.5 | 220.3 | 240.6 | 236.9 | 221.3 | 190.2 | 173.8 |
| *C. manginecans* | 13851 | 227.9 | 197.1 | 233.8 | 236.8 | 220.2 | 240.4 | 236.9 | 221.3 | 190.6 | 174 |
| *C. neglecta* | 11285 | x | 185.2 | 239.3 | 236.4 | 203.5 | 246.9 | 208.3 | 227.1 | 190.5 | 161.8 |
| *C. neglecta* | 11284 | 230.8 | 185.2 | 239.3 | 236.7 | 219.9 | 247.2 | 202.6 | 227.1 | 190.4 | 161.9 |
| *C. neglecta* | 18194 | 227.8 | 192.9 | 236.3 | x | 214.5 | 243.8 | 186.4 | 215.5 | 190 | 165 |
| *C. neglecta* | 17808 | 230.8 | 193.2 | 236.1 | 248.2 | 214.3 | 243.8 | 186 | 215.6 | 190.2 | 162.1 |
| *C. obpyriformis* | 23806 | 227.9 | 193.2 | 286.9 | 228.4 | 196.8 | 232.2 | 153.3 | 212.5 | 180.9 | 122.3 |
| *C. obpyriformis* | 23807 | x | x | x | 228.1 | 196.3 | x | x | 212.4 | 180.9 | 230 |

**Table 6 (continued)** Genescan analysis of microsatellite loci fragment sizes for each isolate of *Ceratocystis* used in this study. Grey boxes indicate conflicting allele sizes within a species, x indicates no PCR product and 0 indicates double bands.

| Species | CMW | CF_CAA/ CAG80 | CF_CAAG5 | CF_GTT50 | CF_GAT5 | CF_CAG AAG5 | CF_GCT 11 | CF_CTCTC TGT5 | CF_CAA/ CAG81 | CF_CCG5 | CF_CAA/ CAG24 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *C. papillata* | 8856 | 230.9 | 196.9 | 235.9 | 239.5 | x | 222.4 | 179.9 | 218.4 | x | 171.1 |
| *C. papillata* | 8850 | x | x | x | 227.9 | x | 237.4 | x | x | x | x |
| *C. papillata* | 28662 | 239.4 | 197.1 | 235.8 | 239.6 | x | x | x | x | 199.2 | 170.2 |
| *C. papillata* | 8857 | x | x | x | 228 | 237.7 | x | x | x | x | x |
| *C. pirilliformis* | 28200 | x | 185.4 | 280.3 | x | x | 233 | x | 218.7 | x | 173.7 |
| *C. pirilliformis* | 6583 | 233.9 | x | 280.1 | 228.1 | 202.5 | 232.4 | 153 | 209.6 | 181.4 | 122.3 |
| *C. pirilliformis* | 12671 | 233.8 | x | 274.6 | 231.2 | 202.7 | x | 153.4 | 209.6 | 180.9 | 122.5 |
| *C. pirilliformis* | 16511 | 233.9 | 197.2 | 274.6 | 231.4 | 202.8 | 232 | 153.5 | 251.3 | 181.1 | 122.9 |
| *C. pirilliformis* | 6579 | x | x | 248.3 | 228.4 | 202.5 | x | 153.4 | 248.3 | 181.1 | 122.8 |
| *C. platani* | 1896 | 230.7 | 185.7 | 238.9 | 236.8 | 219.8 | 247.6 | 202.5 | 227.2 | 190 | 162.1 |
| *C. platani* | 26380 | 233.7 | 184.9 | 239.8 | 236.7 | x | 247.3 | 202.5 | 227.1 | 190 | 161.8 |
| *C. platani* | 23450 | 230.8 | 285.3 | 239.6 | 236.6 | 219.9 | 246.9 | 202.3 | 227.2 | 189.9 | 161.8 |
| *C. platani* | 14802 | 233.7 | 185.4 | 239.7 | 236.6 | 220.4 | 247 | 202.6 | 227.2 | 190.1 | 161.6 |
| *C. platani* | 23918 | 230.8 | 185.7 | 239.4 | 236.3 | 220.2 | 247.1 | x | 227.2 | 190.3 | 161.5 |
| *C. polychroma* | 14281 | x | x | 251.3 | x | 199.1 | x | x | 218.2 | 181.9 | 167.5 |
| *C. polychroma* | 11424 | 227.7 | x | 251.4 | x | x | 201.5 | 152.2 | 215.3 | 181.9 | 167.8 |
| *C. polyconidia* | 23818 | 225 | 196.9 | 233.5 | 236.8 | 214 | 271 | 226.7 | 221.5 | 189.9 | x |
| *C. populicola* | 14789 | 224.8 | x | 223.2 | x | 257.3 | 256.6 | x | 223.6 | 207.3 | 132.8 |
| *C. smalleyi* | 14800 | 242.9 | x | 224.1 | 227.9 | x | 360.3 | x | 216.7 | 181 | 141.2 |
| *C. tanganyicensis* | 15992 | 222.3 | 192.8 | x | 231.2 | 178.6 | 183.2 | 156.1 | 218.8 | 193.5 | 152.6 |
| *C. tanganyicensis* | 15999 | 221.9 | 192.1 | x | x | 178.3 | x | 156.2 | x | x | 152.9 |
| *C. tsitsikammensis* | 14275 | 224.9 | 185.2 | 245.3 | 245.9 | 208.9 | 192.5 | 171.4 | 230.7 | 187.4 | 173.6 |
| *C. tsitsikammensis* | 13982 | 224.8 | 185 | 245.1 | 246 | 208.9 | 232 | 171.8 | 218.8 | 187.3 | 173.8 |
| *C. tsitsikammensis* | 14280 | 225 | 185.4 | 245.3 | 246 | 208.8 | 192.3 | 171.5 | 218.4 | 187.3 | 173.6 |
| *C. tsitsikammensis* | 14276 | 224.8 | 185 | 245.4 | 245.6 | 209 | 192.5 | 171.4 | 218.6 | 187.4 | 173.2 |
| *C. variospora* | 20935 | 222.3 | 146.4 | 208.2 | 268.9 | 218.4 | 384.5 | x | 227.2 | 211.5 | 161.9 |
| *C. variospora* | 26384 | x | x | x | 228.2 | 241.4 | 231 | x | x | x | x |
| *C. variospora* | 26386 | 222.3 | x | x | 268.6 | x | x | x | 228.7 | 211.4 | 164.7 |

**Table 6 (continued)** Genescan analysis of microsatellite loci fragment sizes for each isolate of *Ceratocystis* used in this study. Grey boxes indicate conflicting allele sizes within a species, x indicates no PCR product and 0 indicates double bands.

| Species | CMW | CF_CAA/ CAG80 | CF_CAAG5 | CF_GTT50 | CF_GAT5 | CF_CAG AAG5 | CF_GCT 11 | CF_CTCTC TGT5 | CF_CAA/ CAG81 | CF_CCG5 | CF_CAA/ CAG24 |
|---|---|---|---|---|---|---|---|---|---|---|---|
| *C. virescens* | 3225 | x | x | 233.3 | 228.2 | x | 0 | 0 | x | x | x |
| *C. virescens* | 11160 | x | 196.9 | x | x | 185.3 | 0 | 0 | x | x | x |
| *C. virescens* | 17335 | x | x | x | 228.3 | x | 0 | 0 | x | 219.8 | x |
| *C. virescens* | 17339 | x | x | 233.1 | x | x | 0 | 0 | x | 219.6 | x |
| *C. virescens* | 11164 | x | 197 | 233.2 | 228.3 | x | 0 | 0 | x | 219.6 | x |
| *C. zombamontana* | 15235 | 233.9 | 192.8 | 251.4 | 231.5 | 196.6 | 231.5 | 153.2 | 209.7 | 181.4 | 122.3 |
| *C. zombamontana* | 15236 | 233.8 | 197.3 | 250.9 | 231.1 | 196.6 | x | x | 253.8 | 190 | 123 |

**Table 7** Consensus allele sizes for each species of *Ceratocystis* used in this study. Alleles in **bold** indicate that all isolates within the species have the same allele size, alleles not in bold indicate that most of isolates within the species have the same allele, ? indicates that there is no consensus as to the allele size in that species and x indicates that no PCR products were amplified in that species.

| Species | CF_CAA /CAG80 | CF_CAAG5 | CF_GTT50 | CF_GAT5 | CF_CAG AAG5 | CF_GCT 11 | CF_CTCTC TGT5 | CF_CAA/ CAG81 | CF_CCG5 | CF_CAA/ CAG24 |
|---|---|---|---|---|---|---|---|---|---|---|
| *C. acaciivora* | 228 | ? | ? | **236** | ? | **x** | 237 | **221** | 190 | 173 |
| *C. albifundus* | **x** | ? | 254 | ? | 196 | 238 | ? | 218 | **190** | 153 |
| *C. atrox* | ? | **x** | **248** | **x** | **208** | ? | ? | **218** | **181** | **123** |
| *C. cacaofunesta* | **222** | ? | 230 | ? | 196 | **238** | **194** | 224 | 190 | 159 |
| *C. caryae* | **x** | **x** | **224** | ? | ? | ? | **x** | ? | **x** | ? |
| *C. colombiana* | 239 | **197** | **236** | **240** | 244 | 235 | 267 | **233** | 187 | **182** |
| *C. curvata* | **225** | **193** | **233** | **246** | **214** | **256** | **241** | 221 | **190** | **168** |
| *C. diversiconidia* | **222** | **185** | **242** | **236** | **239** | **219** | **180** | **215** | **190** | **156** |
| *C. ecuadoriana* | **230** | **193** | **236** | 249 | **205** | **244** | **186** | **218** | **190** | **165** |
| *C. fimbriata* | **225** | **193** | **233** | **240** | **214** | **231** | **234** | 224 | **187** | **168** |
| *C. fimbriatomima* | **222** | **201** | **233** | 240 | 208 | **237** | 192 | 224 | **190** | **171** |
| *C. larium* | **242** | **?** | **233** | **234** | 208 | **189** | ? | 218 | **203** | **150** |
| *C. manginecans* | **228** | 197 | **233** | **236** | **220** | **240** | **237** | **221** | **190** | **173** |
| *C. neglecta* | ? | **185/193** | **236/239** | ? | ? | **244/247** | ? | **227/215** | **190** | ? |
| *C. obpyriformis* | ? | ? | ? | **228** | **196** | ? | ? | **212** | **181** | ? |
| *C. papillata* | ? | 197 | 236 | 228/240 | ? | ? | ? | ? | ? | 171 |
| *C. pirilliformis* | 233 | ? | 274/280 | 228/231 | 203 | 233 | 153 | 209 | 181 | 123 |
| *C. platani* | **230/233** | 185 | **239** | **236** | 220 | **247** | 202 | **227** | **190** | **162** |
| *C. polychroma* | ? | **x** | **251** | **x** | ? | ? | ? | ? | **181** | **168** |
| *C. polyconidia* | 225 | 197 | 233 | 236 | 241 | 271 | 226 | 221 | 190 | **x** |
| *C. populicola* | 225 | **x** | 223 | **x** | 257 | 256 | **x** | 224 | 207 | 132 |
| *C. smalleyi* | 242 | **x** | 224 | 228 | **x** | 360 | **x** | 216 | 181 | 141 |
| *C. tanganyicensis* | **222** | **193** | **x** | ? | **178** | ? | **156** | ? | ? | **152** |
| *C. tsitsikammensis* | **225** | **185** | **245** | **246** | **209** | 192 | **171** | 218 | **187** | **173** |
| *C. variospora* | **222** | **x** | 233 | ? | ? | ? | **x** | 228 | **211** | ? |
| *C. virescens* | **x** | 197 | 233 | 228 | ? | ? | **x** | **x** | 220 | **x** |
| *C. zombamontana* | **233** | ? | **251** | **231** | **196** | ? | ? | ? | ? | **123** |

**Table 8** Analysis of transposons that span or are right next to microsatellite motifs in the published microsatellite sequences compared to the *C. fimbriata* genome, also showing which species the microsatellites can be transferred to

| Locus | Transposon name | Transposon type | Similarity (%) | Distance from microsatellite | | Transferable to other species based on population studies |
|---|---|---|---|---|---|---|
| | | | | 5' end | 3' end | |
| AG1/2[1] | Copia-53_MLP-I | LTR retrotransposon | 75 | 7 bp outside | 57 bp outside | *C. pirilliformis*[5] |
| AG7/8[1] | Gypsy-128_ZM-LTR | LTR retrotransposon | 74 | 77 bp outside | 7 bp outside | *C. albifundus*[4] and *C. pirilliformis*[5] |
| AG15/16[1] | RTAg4 | Non-LTR retrotransposon | 80 | 286 bp inside | 57 bp inside | *C. albifundus*[4] |
| CF11/12[1] | Gypsy-4_SI-I | LTR retrotransposon | 77 | 8 bp inside | 29 bp inside | *C. pirilliformis*[5] |
| CF13/14[1] | CATCH2LTR_DR | LTR retrotransposon | 79 | 36 bp inside | 169 bp inside | *C. pirilliformis*[5] |
| CF15/16[1] | HERVIP10FH | Endogenous retrovirus | 74 | 101 bp outside and 1 bp inside | - | *C. albifundus*[4] and *C. pirilliformis*[5] |
| CfAAG9[2] | Ag-Jock-1 | Non-LTR retrotransposon | 67 | 32 bp outside | 0 bp | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], *C. pirilliformis*[5] and *C. platani*[7,8] |
| CfCAA10[2] | Ag-Jock-1 | Non-LTR retrotransposon | 70 | 211 bp outside and 84 bp inside | - | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9,] and *C. platani*[7,8] |
| CfCAA15[2] | Ag-Jock-1 | Non-LTR retrotransposon | 72 | 48 bp outside and 163 bp inside | - | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], *C. pirilliformis*[5] and *C. platani*[7,8] |
| CfCAA38[2] | Gypsy-22-SB-1 | LTR retrotransposon | 82 | 36 bp outside and 39 bp inside | - | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], *C. pirilliformis*[5] and *C. platani*[7,8] |
| CfCAG5[2] | RTEX-1_CR | Non-LTR retrotransposon | 70 | 3 bp inside | 107 bp outside | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], *C. pirilliformis*[5] and *C. platani*[7,8] |
| CfCAG900[2] | Copia-3_CCO-I | LTR retrotransposon | 88 | 1 bp outside | 11 bp outside | *C. cacaofunesta*[6], *C. fimbriata* Brazil[9], *C. pirilliformis*[5] and *C. platani*[7,8] |
| C. fim 20[3] | Chapaev3-1_MR | DNA transposon | 85 | 1 bp outside | - | - |

[1]Barnes et al. (2001), [2]Steimel et al. (2004), [3]Rizatto et al. (2010), [4]Barnes et al. (2005),[5]Kamgan Nkuekam et al. (2009),[6]Engelbrecht et al. (2007), [7]Engelbrecht et al. (2004), [8]Ocasio-Morales et al. (2007), [9]Fereirra et al. (2010). LTR = long terminal repeat.

**Table 9** Analysis of transposons that span or are right next to microsatellite motifs developed in this study (Loci highlighted in grey are those used for Genescan analysis).

| Locus | Transposon name | Transposon type | Similarity (%) | Distance from microsatellite | |
|---|---|---|---|---|---|
| | | | | 5' end | 3' end |
| CF_CAGAAG5 | Ag-Jock-1 | Non-LTR retrotransposon | 72 | 19 bp outside | 5 bp inside |
| CF_CTCTCTGT5 | RAL_Rn_1 | Endogenous retrovirus | 69 | 16 bp outside | 75 bp outside |
| CF_CAA/CAG24 | TART_DV | Non-LTR retrotransposon | 79 | 0 bp | 143 bp outside |
| CF_CTT6 | Gypsy-46_PIT-LTR | LTR retrotransposon | 71 | 31 bp inside | 35 bp outside |
| CF_TCC5 | EnSpm-17_Sbi | DNA transposon | 76 | 27 bp inside | 13 bp outside |
| CF_GCA5 | Copia-7_ES-I | LTR retrotransposon | 84 | 9 bp outside | 24 bp outside |
| CF_TC10 | EnSpm-6_ZM | DNA transposon | 81 | 5 bp outside | 31 bp outside |
| CF_CAA/CAG14 | TART_DV | Non-LTR retrotransposon | 77 | 35 bp outside | 82 bp inside |
| | R21-4_PI | Non-LTR retrotransposon | 67 | 116 bp inside | 68 bp outside |
| CF_TTC5 | Polinton-4_NV | DNA transposon | 80 | 29 bp inside | 76 bp outside |
| CF_CCT6 | Gypsy-50_PIT-1 | LTR retrotransposon | 64 | 184 bp outside | 52 bp outside |
| CF_GAT7 | COPI2_I | LTR retrotransposon | 76 | 33 bp outside | 1 bp outside |
| CF_GCTCCC6 | Gypsy-2_Dfa-I | LTR retrotransposon | 80 | 87 bp outside | 45 bp inside |
| CF_CAA/CAG11 | TART_DV | Non-LTR retrotransposon | 82 | 2 bp outside | 64 bp outside |
| CF_GAC5 | Harbinger-1_MLP | DNA transposon | 91 | 6 bp outside | 15 bp outside |

**Fig. 1** Alignment of microsatellites Cfim01, Cfim02, Cfim05, Cfim08, Cfim13 and Cfim14 to show the similarities between these sequences. The microsatellite motif $(AC)_n$ is shown in the block.

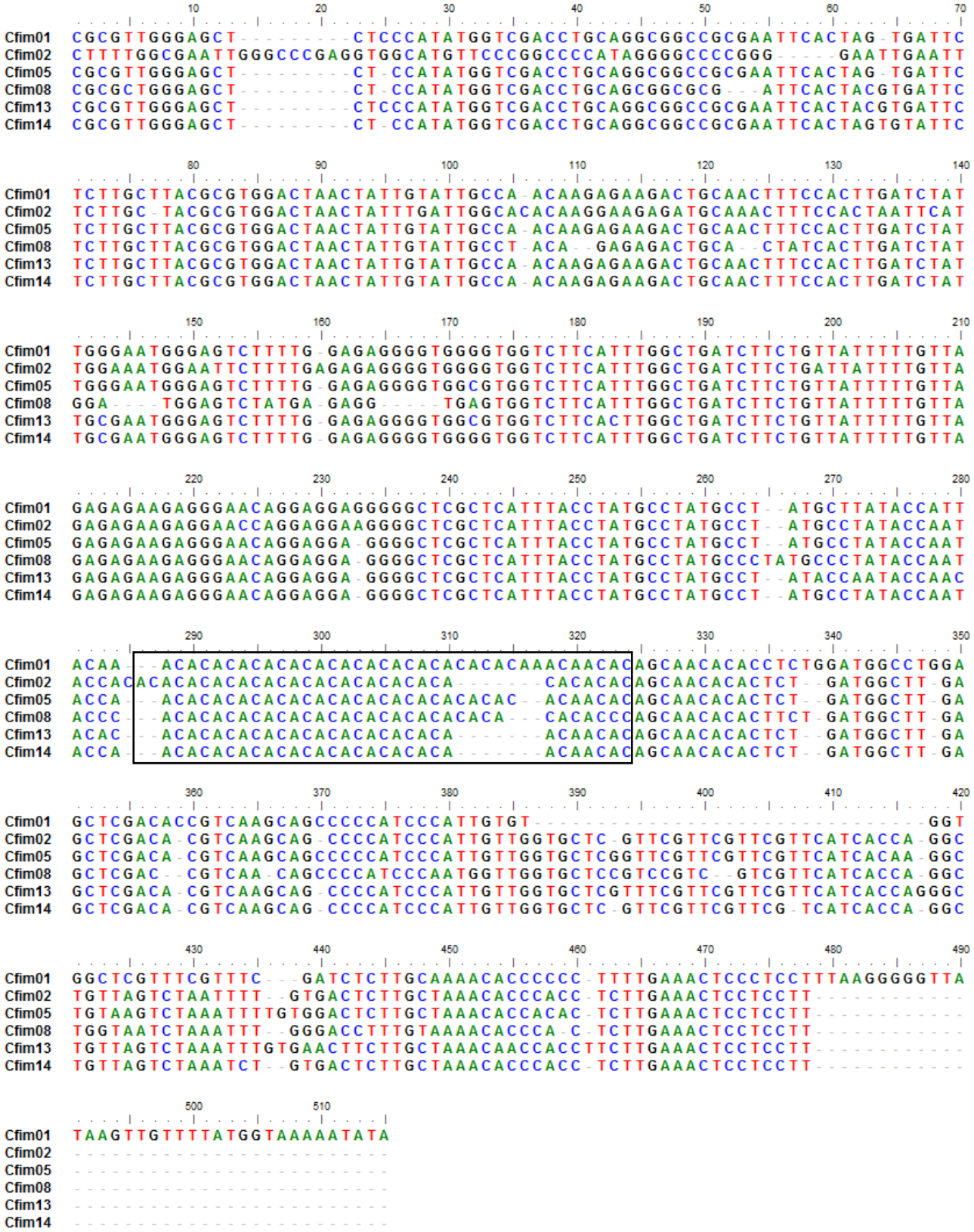


**Fig. 2** Alignment of microsatellites CF17/18 and CF23/24 to show the similarities between these sequences. The microsatellite motif $(TG)_{15}$ is shown in the block.

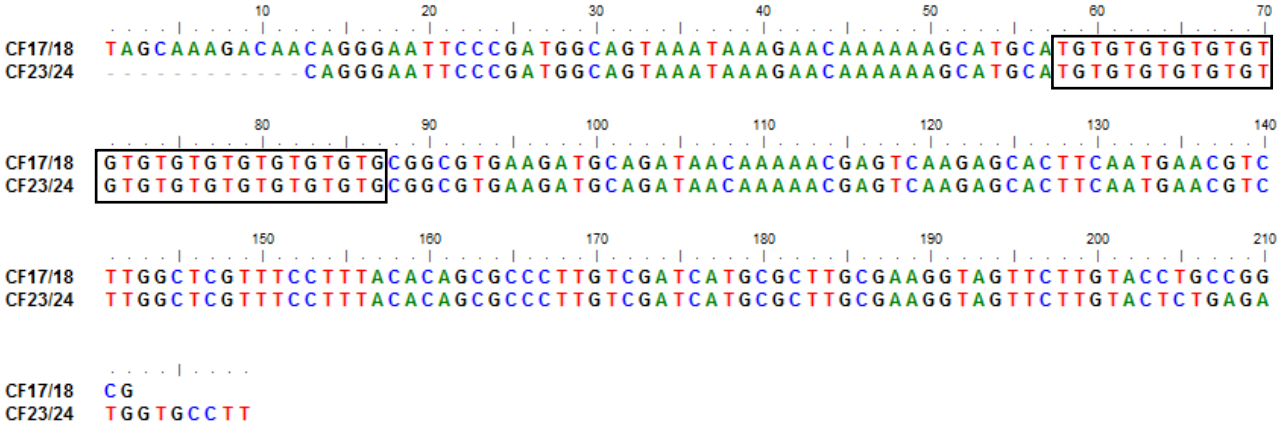


**Fig. 3** Alignment of microsatellites Cfim16 and Cfim18 to show the similarities between these sequences. The microsatellite motifs $(AC)_{21}$ and $(AC/TG)_{16}$, belonging to Cfim16 and Cfim18 respectively, are shown in the block.

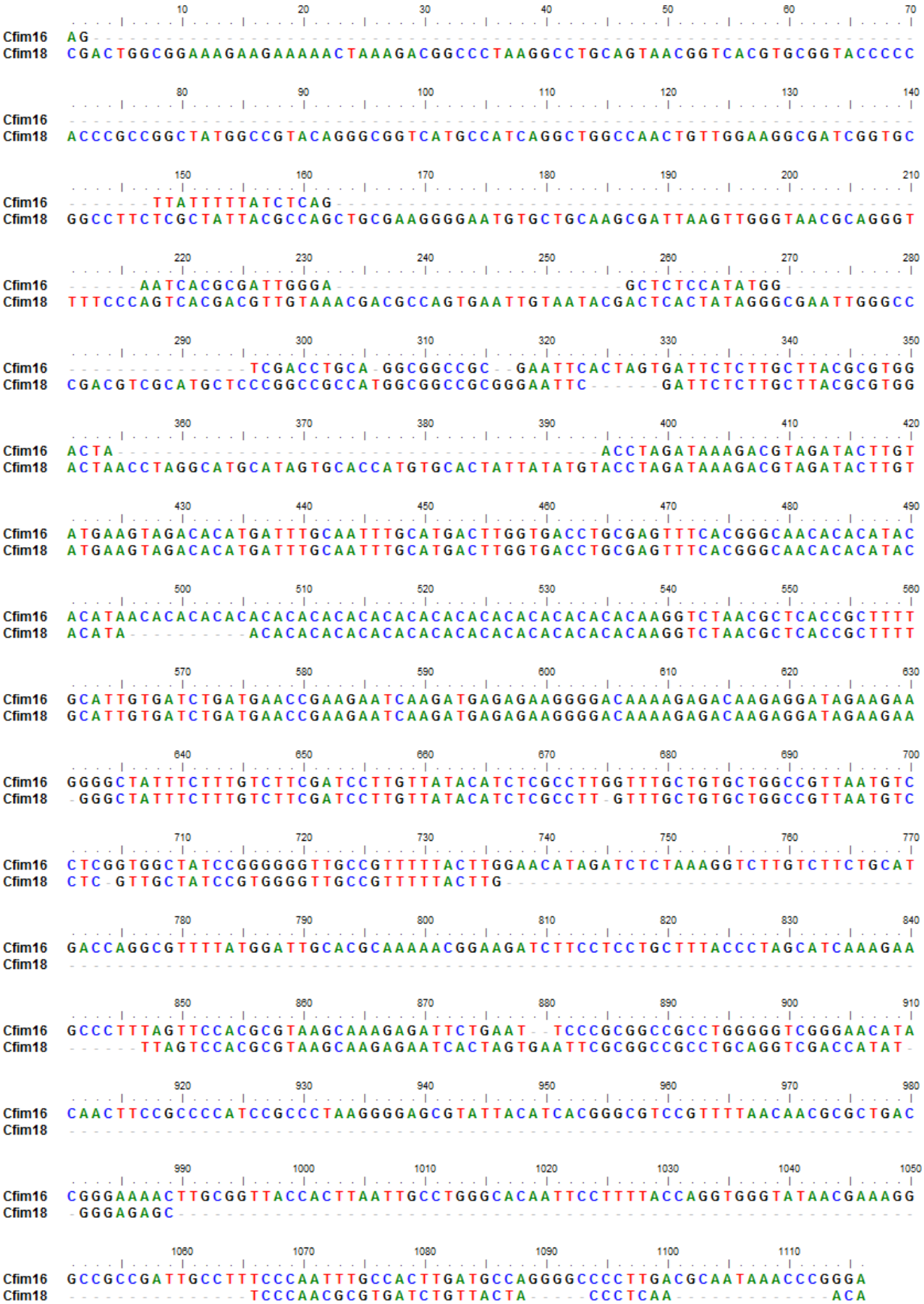


**Fig. 4** Alignment of microsatellites Cfim09 and Cfim17 to show the similarities between these sequences. The microsatellite motif $(AC)_{17}$ is shown in the block.

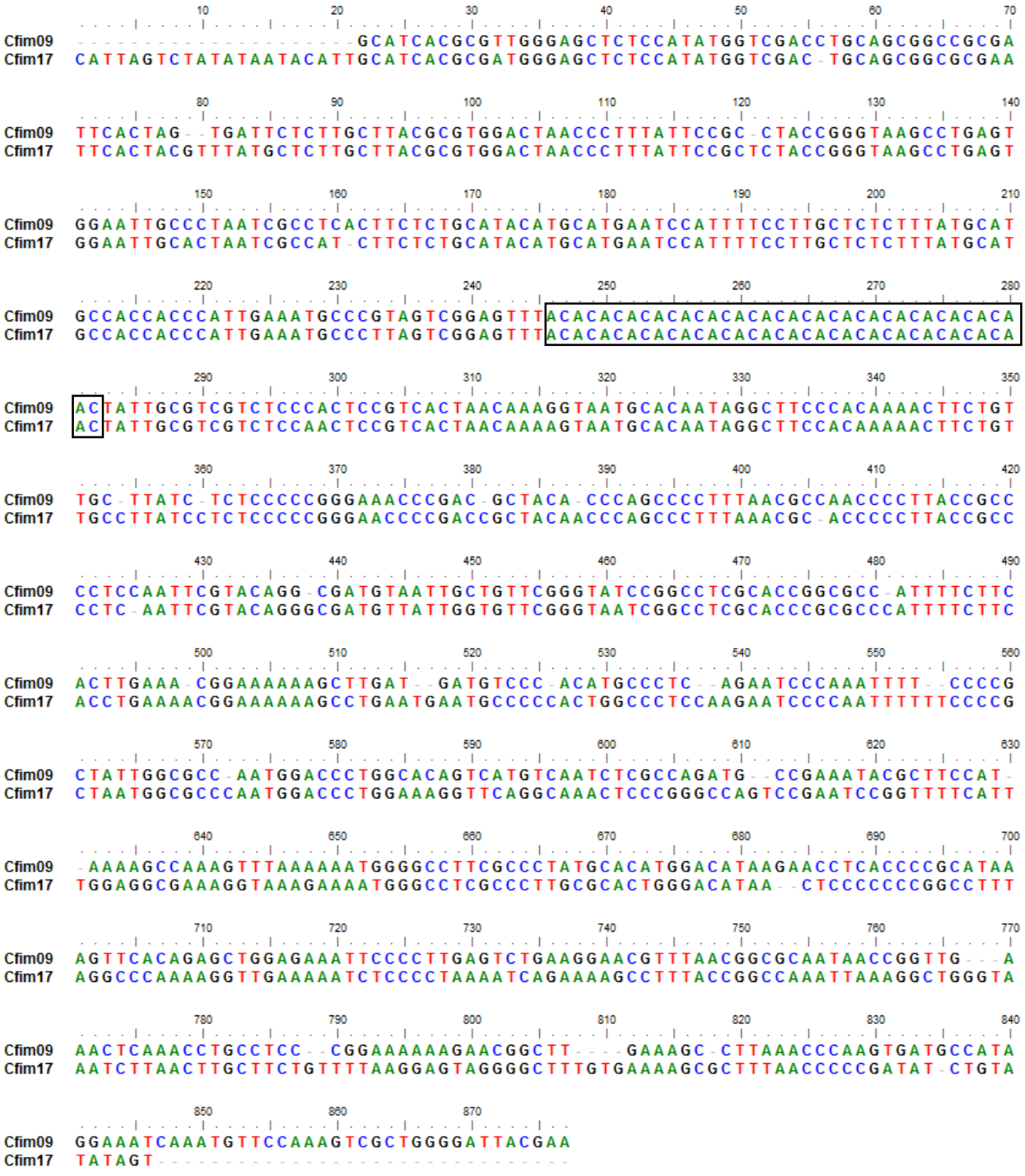


**Fig. 5** SNPs identified in the microsatellite markers developed in this study. The microsatellite motif is shown in the block and SNPs are highlighted in grey. **A** depicts four SNPs in the locus CF_GCTCCC6, three within the microsatellite GTG motif and one SNP in the flanking region. **B** depicts the microsatellite locus CF_CTT6 that contains six SNPs in the flanking regions.

**Fig. 5A**

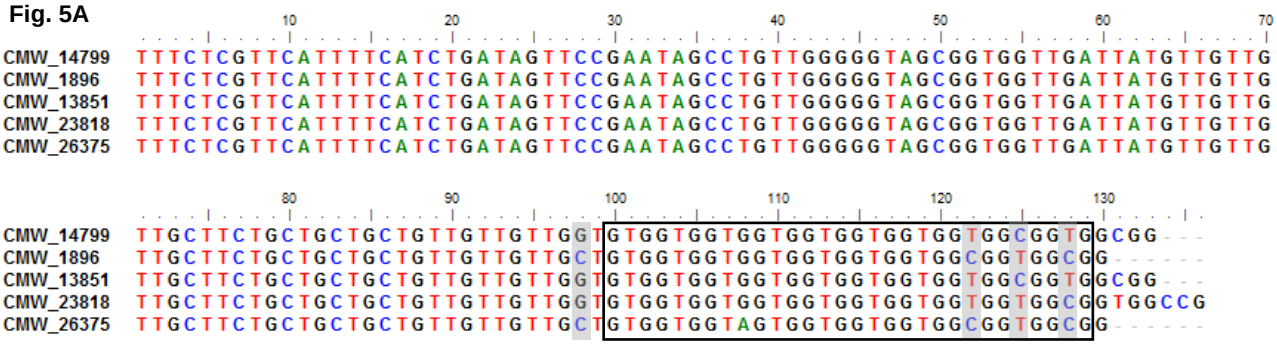



**Fig. 5B**

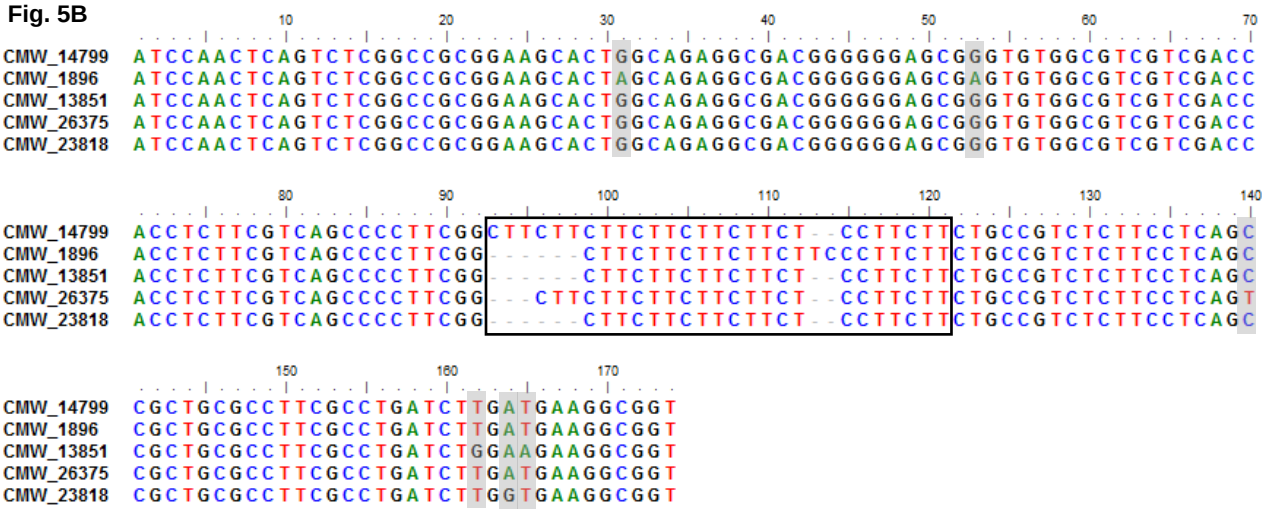
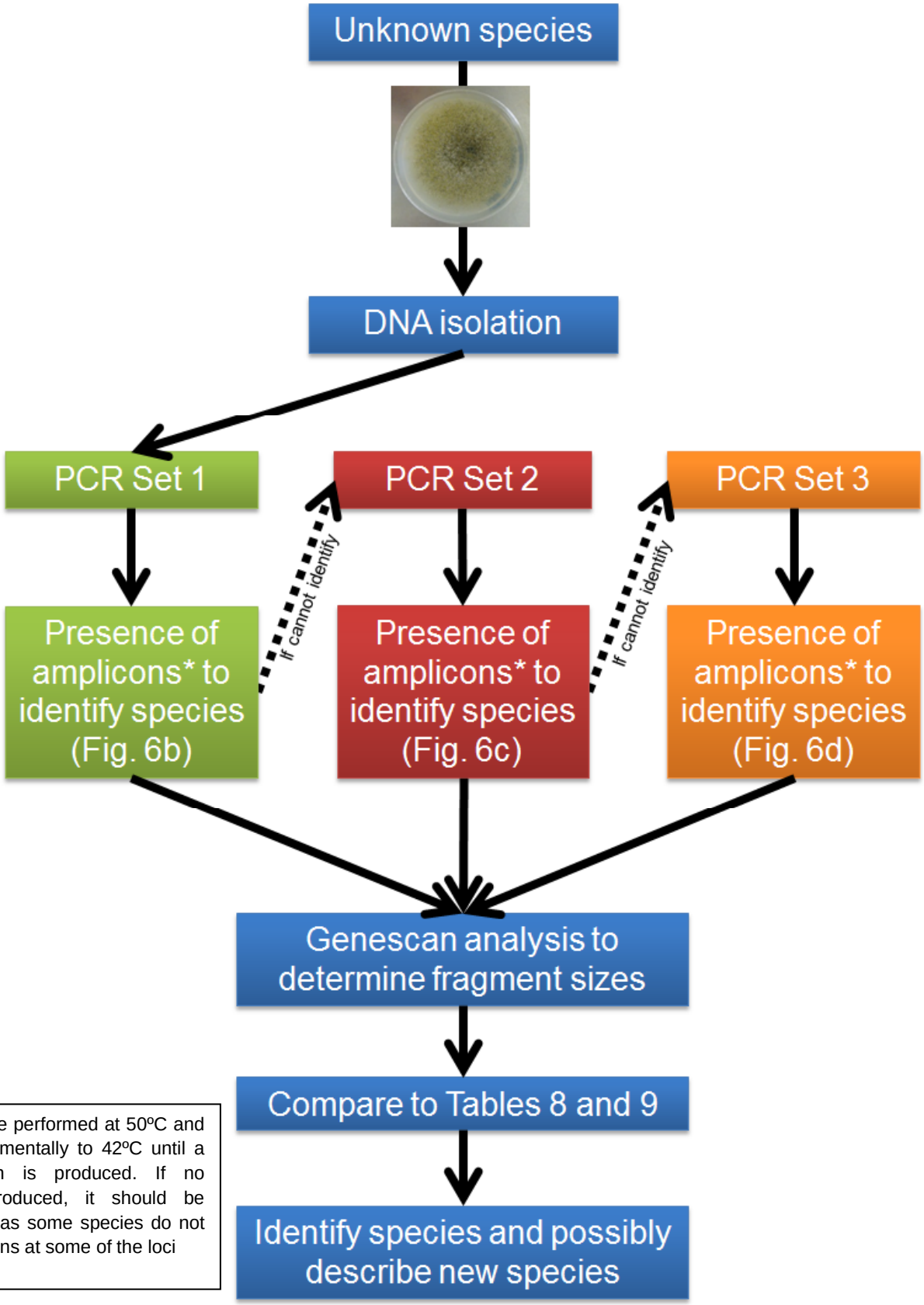


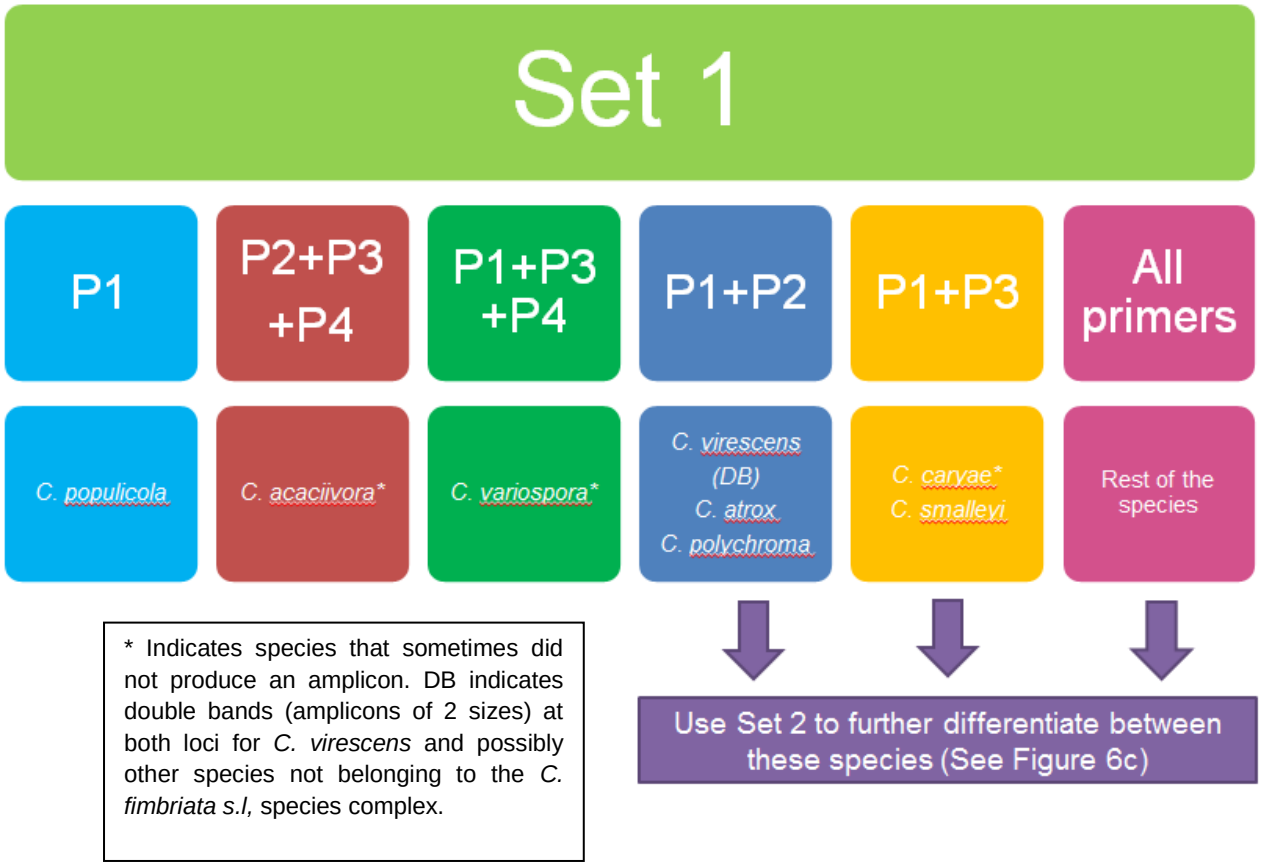
**Fig. 6** The diagnostic test to identify species using the microsatellite loci developed in this study. **A** Flow diagram of the overall diagnostic test. **B** Flow diagram of Set 1 primer combinations used to differentiate species based on the presence of amplicons for each locus. P1 = CF_GCT11, P2 = CF_CTCTCTGT5, P3 = CF_GAT5 and P4 = CF_CAAG9. **C** Flow diagram of Set 2 primer combinations used to differentiate species based on the presence of amplicons for each locus. P5 = CF_GTT50, P6 = CF_CAGAAG5, P7 = CF_CAA/CAG80 and P8 = CF_CAA/CAG24. **D** Flow diagram of Set 3 primer combinations used to differentiate species based on the presence of amplicons for each locus. P9 = CF_CGG5 and P10 = CF_CAA/CAG81.

**Fig. 6A**



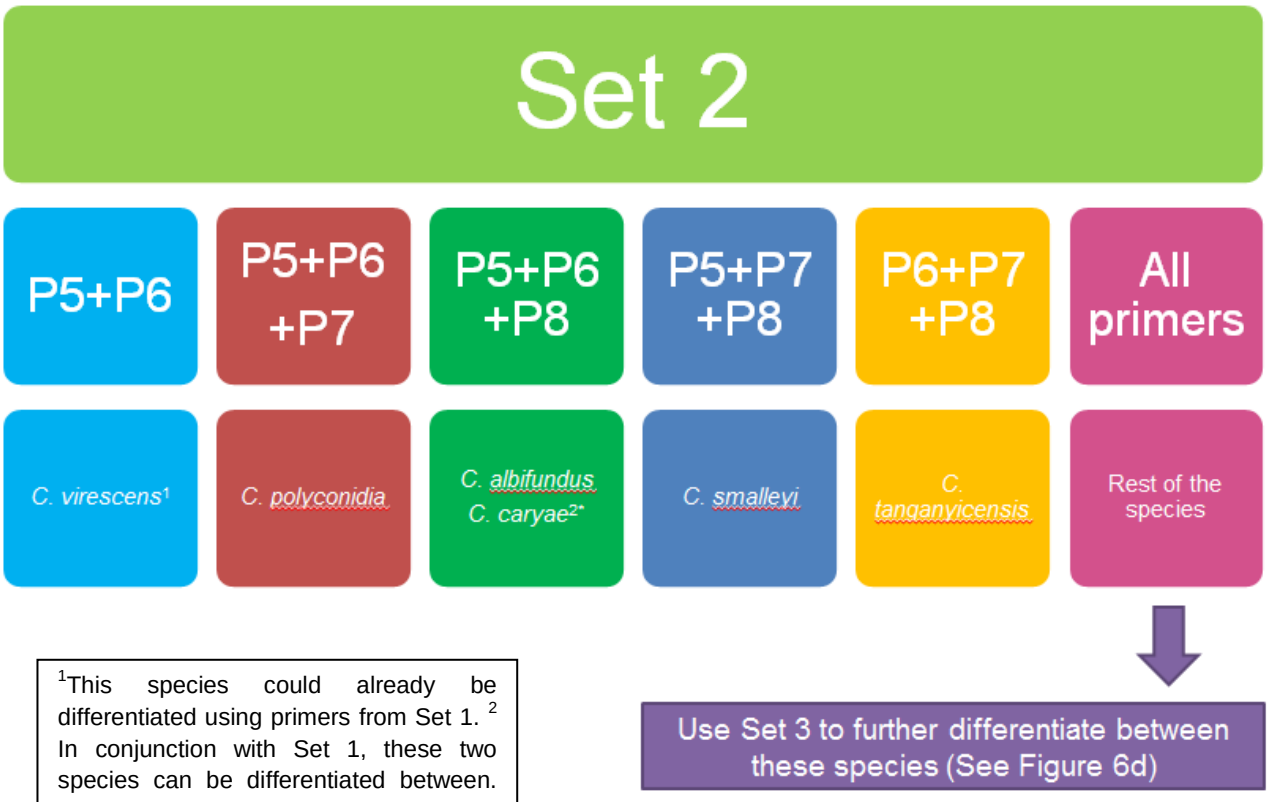

*PCRs should be performed at 50ºC and decreased incrementally to 42ºC until a single amplicon is produced. If no amplicon is produced, it should be scored as such as some species do not produce amplicons at some of the loci

**Fig. 6B**




* Indicates species that sometimes did not produce an amplicon. DB indicates double bands (amplicons of 2 sizes) at both loci for *C. virescens* and possibly other species not belonging to the *C. fimbriata s.l,* species complex.

**Fig. 6C**




[1]This species could already be differentiated using primers from Set 1. [2] In conjunction with Set 1, these two species can be differentiated between. *Species that sometimes did not produce an amplicon.

**Fig. 6D**


[1]These species could already be differentiated using primers from Set 1 and Set 2.

# Summary

*Ceratocystis fimbriata* represents a complex of cryptic species which are pathogens of important agricultural crops and forestry plantations. Population studies of these species have been carried out with microsatellite markers, some of which could differentiate between some of the cryptic species in the complex. A review of microsatellites with regards to their history, past and present isolation methodologies, mechanisms of evolution and functional importance, particularly within fungi is presented. The lack of knowledge regarding microsatellite structure in this group of fungi was highlighted and the need for a more robust identification tool to identify the cryptic species in the complex is also emphasised.

Sequencing of the *C. fimbriata* genome provided an opportunity to study the microsatellite distribution and abundance more in depth. The *C. fimbriata* genome has a medium microsatellite density although it has a larger genome than that of other Ascomycetes. It also compares well with regards to the general trend of microsatellite structure in Ascomycetes and has a unique preference for particular motifs. As the *C. fimbriata* genome contains a significant number of microsatellites, specific microsatellites were identified and a diagnostic test was successfully developed that differentiates between most of the cryptic species in the complex. This study provides an important foundation for future studies on microsatellites in *Ceratocystis* and provides a proof of concept for the use of microsatellite regions in differentiating between sibling species in this genus.