Characterizing mutagenesis in *Fusarium circinatum*

by

Sophia Johanna van Coller

Submitted in partial fulfilment of the requirements for the degree

Magister Scientiae (Microbiology)

In the faculty of Natural and Agricultural Sciences

University of Pretoria

Pretoria

(30 April 2013)

Supervisor: Prof. E.T. Steenkamp

Co-supervisors: Mrs. G Fourie

Co-supervisor: Prof. B.D. Wingfield

Declaration

I Sophia Johanna van Coller declare that the thesis/dissertation, which I hereby submit for the degree Magister Scientiae (Microbiology) at the University of Pretoria, is my own work and has not been submitted by me for a degree at this or any other tertiary institution.

SIGNATURE: _____

DATE: _____

Table of Contents

Chapter 3: Calculating the spontaneous mutation rate of *Fusarium circinatum*

List of tables

Chapter 1:

Chapter 2:

List of figures

Chapter 1:

Chapter 3:

Acknowledgements

I would like to express my appreciation for the following people and funding bodies:

My primary supervisor, Prof. Emma Steenkamp, for her help and support throughout my MSc degree. Your brilliant eye for good science and positivity around problem solving was a true blessing. I am very privileged to have had the opportunity to work with you, you are truly inspiring.

My co-supervisor Mrs. Gerda Fourie for all her help and guidance in the lab. Thank you for your willingness to always lend a hand, no matter the time of day.

My co-supervisor, Prof. Brenda Wingfield, for her support and encouragement. Thank you for the thought provoking and eye-opening discussions.

My six floor colleagues and dear friends. Vivienne Clarence, thank you for being like a second mother to me. Kershney Naidoo, to be in your presence is to have fun. Thank you for all the laughs and special moments. Quentin Santana, thank you for sharing your extraordinary computer skills. You truly made life much easier. Dr Martin Coetzee, thank you for all your help and guidance throughout my time in the lab. I have learnt much from you.

My special friends: Ronel Viljoen, Theunette Mathlener, Chantelle van Dyk, Anel Cronje, Dalinda van Wyk and Imke du Plessis. Thank you for all the love and support throughout the years. My life would have been so empty without you. Tracy Hall, thank you for all the inspiring chats, I could not have finished without them. David Read, thank your patience and understanding. Rynier Lourens, thank you for your shoulder to cry on and all the encouragement.

My amazing inspiring mother, thank you for being the person you are and for raising me the way you did. I can never imagine having a better friend than you. Thank you for always being there to pick up the pieces when things go wrong and for sharing in my joy when the world is right again.

Thank you to my stepdad for all his prayers, love and encouragement and for being a person I can look up to. My brother for his support and encouragement when I needed it.

My dearest dad for his patience, love and support. Thank you for loving me unconditionally and for always being there no matter what I needed. You are the reason I am where I am and for that I can never thank you enough.

The University of Pretoria, the National Research Foundation (NRF), the Forestry and Agricultural Biotechnology Institute (FABI) and the Tree Protection Cooperative Programme (TPCP) for the financial support for my MSc.

But most of all I thank my Heavenly Father for challenging me to be the best I can be. Thank you for the talents I have received and the determination you have blessed me with to finish this task.

"For I know the plans I have for you. They are plans for good and not for disaster, to give you a future and a hope"

Jeremiah 29:11

Preface

The introduction of mutations causes alterations in the hereditary material of all organisms. The process giving rise to these mutations, i.e., mutagenesis, is a major source of the diversity on earth and contributes hugely to the evolution of species. Spontaneous mutations are the consequence of intracellular processes (e.g., polymerase errors and reactive oxygen species) which introduces DNA lesions or damage. Although much is known about the molecular mechanisms underlying mutagenesis and DNA repair in the model organisms, exactly how these mechanisms function to drive evolution of different genes remains unknown. In this study, mutation in *Fusarium circinatum* is studied. This filamentous fungus is the causal agent of the economically devastating pine disease pitch canker and like any other pathogen, *F. circinatum* has to adapt and respond to various biotic and abiotic factors in order to survive and proliferate. Understanding how spontaneous mutagenesis function to drive the evolution and diversification of this fungus might thus aid in predicting emergence of new phenotypes e.g. strains with increase virulence.

Chapter 1 provides a review of the available literature on spontaneous mutagenesis in microbes. The review starts by discussing the sources of DNA damage that are internally produced covering the entire range from chemical mutagens to enzymatic mistakes. This is followed by a discussion of damage recognition mechanisms that lead to triggering of different DNA repair mechanisms. Depending on the type of lesion recognized an appropriate repair pathway is activated. The different repair mechanisms that exists for damaged free nucleotides, damage affecting only a single strand of DNA and lesions involving double stranded DNA breaks are considered. The interaction between the different steps of mutagenesis contributes to a genomic mutation rate. Mutation rates of prokaryotes, Archaea and eukaryotic micro-organisms are discussed respectively after which a comparison is drawn between the three. The chapter is concluded with different methods and considerations when calculating the spontaneous mutation rate.

In **Chapter 2** the general mutational patterns across the genome and genes of *F. circinatum* are studied. For this purpose single nucleotide polymorphisms (SNPs) are considered, by specifically looking at their distribution and the type of substitution that caused the SNP. The SNPs and mutational motifs characteristic of the core housekeeping and non-ribosomal peptide synthetases (NRPS) were also analysed to compare and contrast the possible mutagenic processes involved in the evolution of these two sets of genes. The last aim of the

chapter was to determine whether evolutionary stable mutational signatures of known spontaneous mutagens are present within the core and NRPS genes of *F. circinatum*.

In **Chapter 3** the spontaneous genomic mutation rate of *F. circinatum* is estimated. In order to do this, a fluctuation analysis was conducted to calculate the spontaneous reversion frequency of a nitrate non-utilizing (*nit*) mutant of *F. circinatum*. In combination with an estimate for the generation time of *F. circinatum*, the spontaneous mutation rate was estimated for the fungus under laboratory conditions. The generation time of a close relative of *F. circinatum* was used due to difficulties encountered in accurately calculating the generation time of filamentous fungi. Even though the estimate is not an accurate indication of what is happening naturally it allows for comparisons to be drawn of evolutionary rate between different species including filamentous fungi *Neurospora crassa*, *Aspergillus nidulans* and *Aspergillus fumigatus*.

Chapter 1

Mutagenesis in Microbes

1. Introduction

Genetic mutations can be defined as any event that causes a change in the genetic material of an organism, leading to the formation of new or rearranged hereditary determinants (Foster 2006; Maki 2002; Rogozin *et al.* 2003). Interest in mutations initially arose due to their capacity to cause genetic diseases. This interest triggered research that specifically focused on when, where and how these mutations happened. Mutagenesis in germ line cells was central to these studies, because they are passed on to the progeny of sexual reproducing organisms. As the understanding of mutations increased, their importance in evolutionary processes was recognized as the ultimate source of variation. Today it is widely accepted that mutation is a driving force of evolutionary processes and fundamental to our understanding of the evolution of species and their genes.

In sexually reproducing species, recombination during meiosis contributes significantly to genetic diversity by reshuffling existing variation. For asexual reproducing organisms, mutations are the major source of genetic diversity and ultimately the origin of novel species (Whitman *et al.* 1998). Additionally, many species rely on mutational events for regulating the expression of surface proteins that are functional during infection (Zhuanga *et al.* 1995). Mutations are especially crucial in pathogens, where a high rate of mutagenesis is regarded as favourable because it allows adaptation to stressful environments and different hosts (Sniegowski *et al.* 1997). Mutability is thus selected for when populations undergo a bottleneck, in this way allowing *de novo* variation generation and opening the possibility for sustained infection (Mckenzie & Rosenberg 2001).

Generally, mutations can be classified as either simple or complex (den Dunnen & Antonarakis 2000; Maki 2002; Rogozin *et al.* 2003). Simple mutations – also known as point mutations - are mutations where only one base pair is involved. Simple mutations include base pair substitutions (where one nucleotide is replaced by another) and single base pair insertions and deletions (i.e., indels). The two types of substitution mutations that may occur are transitions and transversions. A transition substitution replaces a pyrimidine (cytosine and thymine) with another pyrimidine or a purine (guanine and adenine) with another purine, thereby acting to conserve the biochemical structure of the bases. A transversion mutation on the other hand, causes a change in the biochemical structure of the bases by changing a pyrimidine to a purine or *vice versa*. Base substitution mutations have been found to occur more frequently than other point mutations and are thus of great value in molecular biology

(Maki 2002; Rogozin *et al.* 2003; Rogozin & Malyarchuk 2005; Snyder & Champness 2007) and are often used as SNPs (single nucleotide polymorphisms) to map certain phenotypic traits.

Complex mutations can affect large regions of DNA and not only a single base pair (Maki 2002; Rogozin *et al.* 2003; den Dunnen & Antonarakis 2000). Duplications, inversions and deletions are the major sources of complex mutations and are mostly regarded as artefacts of erroneous recombination (den Dunnen & Antonarakis 2000). Duplication mutations are one of the most important mechanisms contributing to genome rearrangement (Stankiewicz & Lupski 2002) and result from unequal crossing over during recombination due to misaligned homologous chromosomes. Complex deletions occur when double stranded breaks are fixed using recombination (see Section 5.3 below), while large deletions of interceding genes are caused by non-homologous recombination between two similar genes that are far apart (Gebow *et al.* 2000). Inversion mutations, typically occur when recombination occurs between two molecules in which the homology of the molecules is in opposite directions (Schmid & Roth 1983). However, some inversion mutations have been found to occur in the cell as regulated events (Johnson & Bruist 1989) and are thus not necessarily mutagenic events. Because inversion mutations are more difficult to detect than insertions and deletions, not much research have been done on this type of mutation.

Coding regions are thought to be the most vulnerable to mutations (Maki 2002). This is especially true for genes that are constitutively expressed, or vital for survival, where mutation would have more drastic effects than mutations in genes that are only expressed transitory, or whose functions are redundant (Drake *et al.* 1998; Maki 2002; Rogozin *et al.* 2003). In these regions, frame shift mutations (i.e., an insertion or deletion that causes a shift in the reading frame of the gene) can cause inaccurate amino acid incorporation downstream of the mutation. Base substitution mutations can also affect the functioning of a protein depending on which position in the codon the mutation affects. Mutations occurring in the first and second positions most commonly cause an alteration in the amino acid encoded for, whereas a mutation in the third position is tolerated due to the wobble effect (Crick 1966). The wobble effect allows for what is seen as relaxed base pairing (non Watson-Crick base pairing between two RNA nucleosides) rendering the genetic code redundant (Crick 1966; Spencer & Barral 2012), with multiple codons encoding the same amino acid. This phenomenon allows for silent mutations to occur where the mutation is not visible in the amino acid sequence. Mis-sense mutations, however, cause an amino acid alteration, the

effects of which is dependent on the chemical differences between the wild type amino acid and mutant amino acid. If the codon encoding a polar amino acid is altered to encode another polar amino acid the resulting effects on the protein might be minimal, whereas if the codon is altered to encode a non-polar amino acid the chemical properties of the protein may change significantly, altering its function. Non-sense mutations cause a change from an amino acid coding codon into a stop codon, resulting in a truncated and possibly non-functional protein (Maki 2002; Rogozin *et al.* 2003).

Mutations occurring in gene regions can thus have different effects in that it can be deleterious, neutral or advantageous (Drake *et al.* 1998; Maki 2002; Rogozin *et al.* 2003). Mutations are said to be neutral to natural selection if the resulting phenotype has no effect on the fitness of the organism harbouring the mutation. Deleterious mutations, mutations that produce phenotypes reducing the fitness of the individual, are selected against, because these individuals are easily outcompeted and inferior to the rest of the population. In contrast, advantageous mutations may be selected for as they could increase the fitness of the individual (Drake *et al.* 1998).

Certain stretches of DNA are more prone to mutations than other sequences (Maki 2002; Rogozin *et al.* 2003). This bias can be ascribed to either the nucleotide sequence of the DNA or the presence of repeat units. Sequences that are more vulnerable to mutagenesis due to the nucleotide sequence are known as intrinsic hotspots (Rogozin *et al.* 2005). Intrinsic hotspots are often the result of mutable motifs left by mutagens. A common example is CpG dinucleotides, where the C is often methylated and when the 5-methyl C is deaminated it causes C to T transitions (Rogozin *et al.* 2005). In contrast, homonucleotide runs, direct repeats, indirect repeats and microsatellite repeats undergo a higher frequency of mutational events due to the repetitive nature of the sequences and not the DNA sequence itself (Rogozin *et al.* 2005). Short insertions and deletions are often found in the repetitive regions due to misalignment between the repeats during replication leading to heterogeneity in the number of repeats.

Mutagenesis is an ongoing process which enables us to measure the rate at which mutations occur. Mutation rate can subsequently be defined as the amount of mutations that accumulate per unit of time, most often generation time (Baer *et al.* 2007). When studying mutation rate a class of mutations is typically selected e.g. substitution mutations to measure substitution rate. However, mutation rate or substitution rate seemingly differs, not only

between different organisms but also within the genome of an organism e.g. a lower mutation rate is observed in gene regions than in regions with no apparent function.

To date, no review of literature with regards to the modulation of spontaneous mutagenesis has been published. This leaves opportunities for misunderstandings with regards to sources of spontaneous mutations and their repair. In this literature review I will address the different sources of spontaneous mutations and how mutagenesis is accomplished through these sources. The next two sections will address the mechanisms of damage recognition and how the recognized damage is repaired. I will then move on to mutation rate in micro-organisms, discussing the rate of mutations in prokaryotes, archaea and eukaryotes respectively. The variability in mutation rate and the difficulties posed in its estimation will be discussed, after which I will conclude with a discussion of the different methods that can be used for estimating mutation rate across the genome. As excellent reviews regarding the process of induced mutagenesis are available (e.g., see Burney *et al.* 1999; De Flora *et al.* 1990; Sinha & Häder 2002), I will not address it here and will instead focus only the importance of spontaneously occurring mutagenesis.

2. Spontaneous mutations

The term spontaneous mutations is used to refer to mutations that arise due to internal factors (Maki 2002; Rogozin *et al.* 2003). These mainly include polymerase errors during replication and repair, recombination errors, the effects of mobile genetic elements and damage due to mutagenic metabolites produced by the cell (Figure 1)(Foster 2006; Maki 2002; Rogozin *et al.* 2003). In contrast to this, induced mutations are genetic changes that arise due to the exposure to external mutagens, such as irradiation or chemicals (Maki 2002; Rogozin *et al.* 2003). Generally speaking, any treatment that causes damage to the DNA is referred to as a mutagen. DNA damage increases the opportunity for mutations to occur when the cellular machinery attempts to fix the damage or when the damaged DNA itself causes mispairing. Spontaneous mutations are thought to occur less frequently than induced mutations (Snyder & Champness 2007).

2.1 Erroneous DNA replication and Polymerase precision

2.1.1 DNA replication

DNA replication requires a set of machinery that ensures that each nucleus in the cell is supplied with an exact copy of its parental genome. The enzymes responsible for the duplication process are known as DNA polymerases (see Section 2.1.2 below). Although these enzymes duplicate genetic material at a very high level of accuracy, the process is not faultless. For example, the incorporation of the wrong nucleotide and the incorporation of too few or too many nucleotides may occur (Pray 2008). These mistakes promote possibilities for mutations to take place.

The process by which DNA replication errors occur has been debated and changed over the years. Following the description of the double helix conformation of DNA in 1953 (Watson & Crick 1953), it was thought that errors were made during replication due to tautomeric shifts (Pray 2008). Tautomers are isomers of the different nucleotides and they form due to protons occupying different regions in the molecule (Pray 2008). Guanine and thymine have both keto (common form) and enol (rare) forms, whereas cytosine and adenine have amino (common) and imino (rare) forms. For proper binding in the Watson and Crick conformation, all the bases are required in their keto or amino forms (Pray 2008). When these bases undergo a spontaneous shift to their rare forms, it causes a change in the base pairing properties of the nucleotides involved, leading to mispairings. However, it was found that these spontaneous tautomeric shifts do not occur often enough to be responsible for all of the observed mispairings. A suitable explanation for these mispairings was proposed, implicating DNA polymerases.

2.1.2 DNA polymerase

DNA polymerases are the enzymes that are responsible for selecting and incorporating the correct, undamaged base during DNA replication. They are also responsible for the repair of damaged DNA fragments (Arana & Kunkel 2010). The duplication process is carried out at a high level of accuracy (allowing for only one mistake per 10^8-10^{10} bases), but certain factors might act to reduce the fidelity of these enzymes (see below) (Arana & Kunkel 2010; Baer *et al.* 2007; Snyder & Champness 2007). In addition to the polymerase fidelity, the mismatch repair system (MMR) contributes to the overall conformity of nucleic acid duplication and repair (Arana & Kunkel 2010).

There are seven families (A, B, C, D, X, Y and RT) of DNA polymerases found in eukaryotes. Polymerases differ with regards to substrate, nucleotide selectivity and

proofreading ability. Polymerases that play a significant role in replicating undamaged nuclear DNA belong to the B family of polymerases and are known as Pols α, δ and ε (Arana & Kunkel 2010; Kunkel & Burgers 2008; Steitz 1999). These polymerases can only replicate DNA with a normal Watson-Crick geometry; additional polymerases are required to deal with damaged DNA. The polymerases that deals with translesion synthesis (TLS) and damaged DNA are polymerases η and κ (belonging to the Y family) as well as θ and ν that belong to the A family (Andersen *et al.* 2008; Arana & Kunkel 2010; Friedberg *et al.* 2002; Prakash & Prakash 2002). The TLS polymerases all have different substrates and will act accordingly, but all of the TLS polymerases lack proofreading ability and show a decrease in nucleotide selectivity (Arana & Kunkel 2010). The only polymerases that possess intrinsic proofreading ability are Pols δ and ε, responsible for the majority of nuclear genome replication and Pol ϒ, responsible for mitochondrial genome replication. Polymerases belonging to the C (found in bacteria) and D (found in Euryarchaeota) families will not be discussed as they are not well characterized (Filee *et al.* 2002). The X family of DNA polymerases contain enzymes implicated in DNA repair. They include polymerase β that is associated with base excision repair of damaged nucleotides and the repair of abasic sites, as well as polymerase IV that is the non-homologous end joining (NHEJ) polymerase responsible for repairing double stranded breaks. The last family of polymerases, RT, are encoded for by some viruses to make DNA from their RNA genomes.

There are three main factors that contribute to DNA replication fidelity, the first of which is nucleotide sensitivity. Many of the polymerases lack exonuclease (proofreading) activity and their fidelity is thus dependent on to their ability to choose and incorporate the correct nucleotide (Arana & Kunkel 2010). During this process, the polymerase activity of the enzyme is only induced if a dNTP (deoxy-nucleotide triphosphate) of the correct Watson-Crick geometry binds to the base pair binding pocket. In contrast to this, polymerases responsible for TLS that tolerate DNA damage in order to finish DNA replication, have relaxed geometric selectivity (Arana & Kunkel 2010; Steitz 1999).

Secondly, the proofreading ability of the polymerase contributes to DNA replication fidelity by removing incorrectly inserted nucleotides. An erroneously inserted terminal nucleotide renders elongation inefficient, allowing time for the enzyme to switch from the polymerase activity to the 3′ exonuclease activity that can increase the fidelity of replication more than a hundred times (Albertson & Preston 2006; Arana & Kunkel 2010). This

subliminal quality is shared by only three polymerases, of which two, polymerase δ and ε, function in the nucleus, whereas polymerase ϒ functions in the mitochondria. The main factor influencing the proofreading ability of the polymerase is mutations in the genes encoding these polymerases. Mutations in the exonuclease active site are a common means of bypassing replication errors, but proper functioning is not limited to this active site. Mutations in the polymerase active site might enable the enzyme to extend the erroneous terminus, whereas others might act to repress the switch from polymerase to exonuclease function. In addition to mutation, target DNA sequence is implicated depending on the suitability of the sequence as a substrate e.g. repetitive sequences are poor substrates for these enzymes (Arana & Kunkel 2010). Lastly, base analogs (such as 8-oxoguanine or 8-oxoG in its *syn* conformation) will pair with a normal adenine nucleotide and escape detection (Arana & Kunkel 2010; Baer *et al.* 2007), leading to mutations in subsequent rounds of replication.

The last contributor to replication fidelity is MMR (see Section 1.1.4.2 below). This system fixes those mistakes that were overseen by the exonuclease activity of the DNA polymerase. Apart from also modulating recombination and repairing double-stranded breaks, it also triggers apoptosis in response to severe damage (Arana & Kunkel 2010). The function of MMR can increase the fidelity of DNA replication as much as a thousand times (Arana & Kunkel 2010; Hsieh 2001; Schofield & Hsieh 2003). Therefore, mutations altering the function of any one of these proteins, as well as the type and location of the error made, will have a significant influence on mutation rate (Arana & Kunkel 2010).

2.1.3 Replication Biases

Chromosomal replication encompasses a whole series of asymmetries (Rocha 2004). These asymmetries have certain important implications with regards to sequence composition and the distribution of genes, which in turn may lead to mutational biases between different replichores (an arm of the chromosome when divided by replication) and the leading and lagging strands. The main aspects causing these mutational biases are: polymerase collisions; leading and lagging strand differences; and replichore gradients.

During the normal replication cycle of a cell replication and transcription occur simultaneously, with both a DNA polymerase and an RNA polymerase loaded onto the DNA (Bremer & Dennis 1996). Both of these polymerases move in a 5′-3′ direction, causing transcripts encoded for by the lagging strand to cause head-on collisions between the two

polymerases if the replication fork arrives at a gene being transcribed. In contrast to this, transcripts of the leading strand cause co-orientated collisions. It was found that every head-on collision lead to a crash (a possible collapse of replication fork), whereas only 95% of co-orientated collisions caused crashes (Rocha 2004). In addition to this, head-on collisions were found to interfere drastically with the replication fork, causing it to stall (Deshpande & Newlon 1996). The resulting stalled fork may then need homologous recombination to get it restarted opening up possibilities for errors during replication which could lead to mutations.

Distinct differences have been found between the leading and the lagging strands that significantly influence the accuracy of replication. The lagging strand is replicated in so-called Okazaki fragments, which leaves patches of single-stranded areas on the template strand (Rocha 2004; Yudelevich *et al.* 1968). Single-stranded DNA is known to form secondary structures which may cause erroneous replication, leading to substitutions or deletions (Francino & Ochman 1997). Also, single-stranded DNA has increased chemical instability of the cytosines (Frank & Lobry 1999; Rocha 2004), with increased frequency of deamination that can lead to substitutions (a 140 times increase in cytosine deamination and an additional 4 times when the cytosine is methylated (Lutsenko and Bhagwat 1999; Rocha 2004)). In addition to such replication errors, certain substitution mutations have been found to display strand bias (Frank & Lobry 1999). Also, the efficiency of polymerases differs at replication forks (Frank & Lobry 1999; Rocha 2004), further contributing to mutational biases between the strands. Moreover, the MMR system functions more efficiently on the lagging strand due to the availability of nicks in the Okazaki fragments (Radman 1998; Rocha 2004). Together, these factors cause mutation rate variation between the two strands.

In *Escherichia coli*, it was determined that the time span between the start and termination of replication was approximately one hour (Daubin & Perrière 2003). This long replication time allowed sufficient time environmental changes to occur, leading to the creation of gradients. A simple example is that nutritionally favourable conditions that reign during the start of replication might vanish throughout the replication process, leaving unfavourable conditions at the end. This places the cell under stress at the end of replication, which may lead to mutations, and subsequent changes in the mutation rate. When looking at mutation biases, it is therefore not surprising that early studies showed substitution rates that were twice as high at the termini of chromosomes (Rocha 2004). More recent studies have found that the distance from the origin of replication accounts for about 5% of substitution

rate variation. This percentage of variation was regarded as constant with the exception of an increase in mutations in the GC poor termini, forming a less apparent gradient (Daubin & Perrière 2003; Rocha 2004). The genome compensates for this gradient by positioning the important genes close to the origin of replication and the less important genes (possible redundancy) near the termini (Daubin & Perrière 2003).

2.2 Base damage and alterations

2.2.1 Deamination

Deamination is one of the most frequently occurring DNA damages (Snyder & Champness 2007). As the word implies, deamination is the loss of amine groups. The bases adenine, cytosine and guanine are especially vulnerable to this type of damage, which might occur either spontaneously or due to intracellular chemical compounds, produced by the cell, such as nitrous acid. Thymine has no amine group and thus, cannot be deaminated.

When bases adenine, cytosine and guanine are deaminated, they become hypoxanthine, uracil and xanthine respectively - resulting in the mispairing of the bases and the incorporation of base substitutions during replication (Snyder & Champness 2007). The type of substitution that happens depends on the deaminated base. If adenine is deaminated, the resulting hypoxanthine will pair with a cytosine when the strand is replicated. Thus, a cytosine is incorporated instead of a thymine, which can result in an adenine to cytosine transversion. Similarly, if a cytosine is deaminated, the resulting uracil will pair with an adenine resulting in a cytosine to thymine transition. Lastly, a deaminated guanine, xanthine, may pair with a thymine instead of a cytosine, leading to a guanine to adenine transition. Nonetheless, all of the deaminated bases are unnatural bases, facilitating detection by repair mechanisms.

A common source of mutations is through the deamination of methylated cytosines. Most organisms have methyl transferases which transfer a methyl group to the cytosine to form a 5-methylcytosine, which plays a role in gene regulation (Gonzalgo & Jones 1997). When these methylated cytosines become deaminated, they become thymines - a base occurring naturally in DNA - and thus would not be recognized as a mistake. A special DNA glycosylase is responsible for repairing this damage. Instead of recognizing an unnatural base, this enzyme recognizes guanine-thymine mismatches and fixes the thymine (Snyder &

Champness 2007). Nonetheless, these methylated cytosines seem to be hotspots for mutations which might indicate a relatively ineffective repair system (Gonzalgo & Jones 1997).

2.2.2 Oxygen Species

All aerobic species maintain aerobic metabolic processes which unavoidably lead to the production of reactive oxygen species (ROS) (Snyder & Champness 2007). These damaging species are different forms of molecular oxygen (O_2) that are formed through the acquisition of an electron e.g. superoxide radicals, hydroxen peroxide, and hydroxyl radicals (Bayir 2005). Because of the natural occurrence of ROS in the cell, the cell has evolved different mechanisms to deal with oxidative damage (Bayir 2005; Snyder & Champness 2007). Enzymes, such as superoxide dismutases, catalases and peroxide reductases, help destroy the reactive species, whereas enzymatic glycosylases and exonucleases help repair the damage. Antioxidants, such as vitamin C, vitamin E and glutathione also help in neutralizing these molecules (Snyder & Champness 2007).

ROS may react with normal DNA bases, forming mutagenic lesions. One such an example is 7, 8-dihydro-8-oxoguanine (8-oxoG). This base appears frequently, and if the damage is not repaired, DNA polymerase pairs it with any nucleotide which could lead to mutations (Vasyunina *et al.* 2004). In fact, 8-oxoG is thought to be the major source of genomic variety due to its mutagenic capacity (Ohno *et al.* 2006). A direct relationship was found between regions with high 8-oxoG and SNPs densities (Ohno *et al.* 2006). In addition to this, it was also found to correlate with recombination frequency (Ohno *et al.* 2006), altogether contributing to genomic diversity. Enzymes responsible for the repair of this damage are expressed in all organisms and include 8-oxoG glycosylases (i.e., OGG1 and OGG2). OGG1 excises the 8-oxoG when paired with a cytosine or thymine, while OGG2 targets 8-oxoG paired with adenine (Hazra *et al.* 2001).

2.2.3 Abasic sites

Abasic sites are a common form of DNA damage that are brought about through the enzymatic cleavage and spontaneous degradation of the glycosidic bond between the nitrogen base and sugar in the nucleotide (Barsky *et al.* 2000; Boiteux & Guillet 2004). It was estimated that approximately 1000 abasic sites form spontaneously in the mammalian genome per day (Barsky *et al.* 2000). In addition to the spontaneous loss of bases, DNA damaging agents act to weaken the glycosidic bonds between bases. These damaged bases are

then recognized by enzymes, known as DNA glycosylases, which remove the damaged base, leaving a gap. These gaps are referred to as apurinic or apyrimidinic (AP) sites, depending on the removed base. If these sites are not repaired, they could lead to possible transcription and replication arrest (Barsky *et al.* 2000; Boiteux & Guillet 2004), which might then be subjected to error prone repair processes, enabling mutations to occur.

2.3 Mobile genetic elements

Mobile or transposable genetic elements are DNA segments that move freely within the genome of a cell (Galagan & Selker 2004). They commonly occur in genomes and their movement are thought to have aided in the creation of novel genes and sophisticated regulatory pathways but are often implicated in mutagenesis (Galagan & Selker 2004). These elements are regarded as mutagens because their insertion into gene regions can alter the function and expression of the gene. Multiple insertions of the same element may also create sequence stretches that are repetitive in nature, thus further increasing the vulnerability of the genome to mutagenesis. In addition, faulty DNA repair following excision of these elements creates additional opportunities for mutagenesis (Galagan & Selker 2004).

3. DNA damage recognition

In order for the genome to maintain its integrity, it has to implement mechanisms to recognize damage imposed by genotoxic stresses which could otherwise result in mutations (Figure 1). Different types of DNA damage are recognized differently, varying from the simplest enzymatic recognition to protein complexes, which in turn trigger downstream repair mechanisms. Evidently, damage sensors can trigger the same, or different, repair pathways depending on the damage recognized. In this section the recognition process, as well as the different components of repair pathways will be discussed.

Throughout the four cell cycle stages (i.e., G0, interphase (G1, S and G2) and M), there are three main checkpoints that a cell implements to ensure completion of a particular cell phase before commencing on to the next (Novak *et al.* 2002). In the first stage, interphase, there are three phases: G1; S; and G2. G1 allows for cell growth and increase in size so that it can enter the S (synthesis) phase, characterized by DNA replication. The last phase in the interphase is G2, where the cell reaches the appropriate size allowing it to enter the M phase,

dividing the cell into two daughter cells. Checkpoints exist at the end of the G1, S and G2/M phases in the cell cycle (Novak *et al.* 2002; Nyberg *et al.* 2002). Each of these checkpoints plays a distinct and important role to ensure the fidelity of replication and disbursing the replicated genomic DNAs to the daughter cells (Novak *et al.* 2002). These checkpoints seem to be similar in all eukaryotes, with some minor differences between the so-called lower and higher eukaryotes. One such difference is that in lower eukaryotes, such as fungi, the S and M phases overlap, resulting in the absence of a distinct G2 phase and subsequently a G2 checkpoint (Nyberg *et al.* 2002). In prokaryotes the cell cycle is even simpler, due to the absence of multiple chromosomes, but the underlying principles remain.

Proteins that function as part of the checkpoints allow the cell cycle to slow down and stop in order to repair damaged nucleotides (Nyberg *et al.* 2002). In addition to this, some of these proteins function either directly or indirectly in the repair process (Novak *et al.* 2002; Nyberg *et al.* 2002). For example, the RPA (replication protein A) protein complex (which binds single-stranded DNA during replication, repair and recombination) is phosphorylated by the apical kinase, ATM (see below; ataxia telangiectasia mutated) when damage is detected, thus resulting in replication arrest to allow time for the damage to be repaired (Nyberg *et al.* 2002). Some other examples of indirect involvement include: the induction of transcription factors that are responsible for the expression of repair proteins; the regulation of dNTP pools; and the participation in homologous recombination or NHEJ (Novak *et al.* 2002; Nyberg *et al.* 2002). In contrast to these indirect involvements in DNA repair, there is not much evidence for direct involvement. It was found in fission yeast that DNA damage results in the prevention of recombination by a checkpoint protein that regulates topoisomerase III (Top3), which is known to process recombination intermediates (Nyberg *et al.* 2002).

In general the DNA damage recognition process can be divided into five steps. Firstly, the sensor proteins are responsible for the interaction with the damage, whether direct or indirect (Abraham 2001). These proteins scan the genome for stalled replication forks, double-stranded or single-stranded breaks, or any additional abnormalities, although not much is known about the mechanism or molecules involved in these scanning processes. Following the sensor proteins, the apical kinases bind these abnormalities. Apical kinases are responsible for the activation of downstream phosphorylation cascades in order to amplify the damage signal. These signal relay proteins (such as scaffolding proteins, downstream kinases and regulatory proteins (Abraham 2001)) are responsible for activating the effector proteins,

such as cell cycle components, transcription factors and repair proteins (Abraham 2001; Novak *et al.* 2002; Shiloh 2001). This leads to typical responses such as cell cycle arrest, DNA repair and apoptosis.

Central to this recognition cascade are the two apical kinases ATR (ATM and Rad3-related) and ATM. These are both members of the family of phosphor-inositide-3-kinases (PIKK) and have phosphor-transferase functions, phosphorylating proteins with serine-glutamine motifs (Abraham 2001). These two proteins function similarly, although ATM is primarily triggered by double-stranded breaks and provides a rapid protective response to the damages causing these breaks (Abraham 2001; Novak *et al.* 2002; Nyberg *et al.* 2002), while ATR is triggered by single-stranded breaks that form due to stalled replication forks or due to unrepaired damage prior to replication (Abraham 2001; Novak *et al.* 2002). ATR does not appear to have a rapid response, but rather a housekeeping function fundamental for cell proliferation.

The G1 checkpoint, in lower eukaryotes, is very weak and typically delays replication for less than an hour, with the result that most of the damage remains unfixed (Nyberg *et al.* 2002). In higher eukaryotes, this checkpoint comprises of two mechanisms of delay: one acting immediately on a post translational level; and the other on a transcriptional level which takes longer to function (Novak *et al.* 2002; Nyberg *et al.* 2002). Cells that are of the right size and present in a nutritionally favourable environment pass this checkpoint, allowing genome replication, unless threats are detected that threaten the integrity of the DNA. If such threats are detected, ATM triggers BASC (Brca1 associated surveillance complex) (Novak *et al.* 2002; Nyberg *et al.* 2002) that associates with the DNA triggering repair and stalling replication. If no such threats are detected the S phase is entered.

Replication during the S phase creates opportunities for converting DNA damage into mutations through DNA polymerase errors (Novak *et al.* 2002). The S phase checkpoint monitors the cell cycle succession and decreases the rate of DNA synthesis if damage is detected. Once detected, two checkpoints act to stall progression, the intra-S checkpoint and the replication checkpoint (Novak *et al.* 2002; Nyberg *et al.* 2002). These two checkpoints, although different, share many common features. The intra-S checkpoint is triggered when a lesion stalls a replication fork, followed by lesion processing and replication resulting in a double-stranded break (Nyberg *et al.* 2002). This checkpoint acts as a backup for the replication checkpoint by preventing late origin firing and spindle elongation. The replication

checkpoint is also triggered by stalled replication forks through the functioning of, for instance, DNA helicases (Nyberg *et al.* 2002). It is speculated that the altered form of the stalled fork then triggers the damage sensor proteins, linking the checkpoints to damage repair.

Although there is no distinct G2/M checkpoint in lower eukaryotes, it is thought that this checkpoint still acts to delay transition from metaphase to anaphase (Nyberg *et al.* 2002). This is done through the inhibitory phosphorylation of a protein called Cdc2 (cell division control protein 2) (Abraham 2001; Novak *et al.* 2002; Nyberg *et al.* 2002), which is required for the transition. In addition to simply stopping the transition, the damage also activates ATM/ATR which inhibits the activator of APC (anaphase promoting complex) (Novak *et al.* 2002; Nyberg *et al.* 2002) through a signalling cascade. ATR/ATM will also gather repair proteins that will fix the damage.

4. DNA repair

DNA repair is the collection of processes that a cell employs to identity and rectify damage to the genomic DNA (Figure 1). By doing so the cell is capable of maintaining the integrity of the genomic material and thus the optimal functioning of the cellular machinery. All cells have developed multi-faceted responses to counteract the possible negative effects that can be brought about by the damage that occurred. These include direct reversal of the damage, mechanisms that fix the damage using information on complimentary or homologous DNA strands as well as mechanisms that allows replication across the damage to enable repair at a later stage.

4.1 Direct reversal

The simplest of repair mechanisms is the direct reversal of damaged nucleotides, with the reversal process being specific to the damaged that occurred (Snyder & Champness 2007). This type of repair does not require a template, since it only fixes free nucleotides. An example of such a mechanism is the repair of O^6-methyl-guanine, which is generated in low levels within the cell when certain cellular catabolites react with guanine (Rydberg & Lindahl 1982) and it mispairs with a thymine, possibly causing GC to AT transitions (Esteller *et al.* 1999). These lesions are fixed by an enzyme known as MGMT (O^6-methyl-guanine DNA

methyltransferase). As the name implies, the enzyme directly transfers the methyl group from the guanine to a cysteine in its active site, inactivating the enzyme. The methylated enzyme is then degraded by an ATP-dependant reaction (Esteller *et al.* 1999), which is thus energetically expensive and can be detrimental to a cell.

4.2 Single stranded damage

Damage to nucleotides does not always occur in the free nucleotide pools, forcing the cell to repair nucleotides incorporated into the double-stranded DNA molecule. In the case where only one of the strands is damaged, the complementary strand can act as a template. Here mechanisms act to remove the damaged base, fill the gap with the correct complementary base, and seal the nick to restore the fidelity of the double-stranded molecule. Three pathways fall into this category and will fix different types of lesions within one strand of DNA, namely base excision repair (BER), nucleotide excision repair (NER) and MMR (Figure 2).

BER is a multi-step process responsible for fixing non-bulge forming (i.e. minor disturbances to the helical DNA structure) lesions (Memisoglu & Samson 2000). These lesions include oxidative (e.g. 8-oxoG), methylation (e.g. 3-methyl adenine), deamination (hypoxanthine, uracil and xanthine but not thymidine due to deamination of 5-methyl cytosine) and spontaneous, loss of bases (Memisoglu & Samson 2000). This repair pathway can be divided into two sub-pathways: short patch repair (1 nucleotide) and long patch repair (2-10 nucleotides) (David *et al.* 2007). Both of these processes are very well studied and have been shown to function similarly as well as to be initiated by enzymes known as DNA glycosylases (David *et al.* 2007).

Short patch repair relies on DNA glycosylases to cleave the *N*-glycosidic bond between the damaged base and its sugar in the backbone of the DNA. This leaves an abasic site, also formed spontaneously, which is then processed by AP Endonuclease 1 (David *et al.* 2007; Matsumoto 1995). This enzyme leaves a 3′ hydroxyl group and a 5′ transient abasic deoxyribose phosphate (dRP) which forms the substrate for DNA polymerase β. The polymerase adds a single nucleotide to the 3′ end of the nick and subsequently removes the dRP molecule through the AP lyase activity (Matsumoto 1995). Finally, DNA ligase seals the nick, restoring the integrity of the DNA. This short patch repair takes place 80-90% of the time.

Long patch repair is the backup mechanism for the short patch repair system where certain modified bases are resistant to AP lyase and thus requires the replacement of 2-10 nucleotides (David *et al.* 2007; Matsumoto *et al.* 1994; Matsumoto 1995) including the damaged nucleotide. This sub-pathway relies on the same machinery as the short patch sub-pathway but is PCNA (proliferating cell nuclear antigen) dependant (Frosina *et al.* 1996). PCNA is a factor that acts as a processivity factor for DNA polymerase. Instead of simply adding one nucleotide to the 3′ end of the nick, the polymerase adds a string of several nucleotides to the gap, displacing the dRP as a flap (David *et al.* 2007; Matsumoto 1995). The flap is then removed by the flap endonuclease, FEN1 (Prasad *et al.* 2001), and the gap is sealed by DNA ligase. This process is the same in prokaryotes and eukaryotes with the FEN1 homolog in *Saccharomyces cerevisiae* being RAD27 (Klungland & Lindahl 1997).

NER is regarded as the most flexible repair pathway with regards to the type of lesions repaired (Batty & Wood 2000). It is required for the removal of lesions forming bulges in the double-stranded DNA structure, pyrimidine dimers, intra-strand cross links and oxidative damage (Batty & Wood 2000; Hess *et al.* 1997). NER differs between prokaryotes and eukaryotes, but only in the sense that eukaryotic NER is more complex. More than 30 proteins function in a stepwise manner from recognition, cleavage, dual incision of damaged strand, gap repair and lastly ligation; whereas only 4 proteins are involved in NER in *E.coli* (Shuck *et al.* 2008).

The NER proteins utilized by *E.coli* are UvrA, UvrB, UvrC and UvrD (also known as DNA helicaseII) (Howard-Flanders & Boyce 1966; Van Houten 1990). NER is initiated by the UvrA-B complex that continuously scans the genome and recognizes distortions. Upon recognition of such a distortion, UvrA exits the complex and is replaced by UvrC - forming a new protein complex known as UvrBC. Upon arrival, UvrB cleaves the phosphodiester bond 4 nucleotides downstream from the lesion. In addition to this, the UvrC protein cleaves the phosphodiester bond 8 nucleotides upstream of the lesion, rendering an excised oligomer consisting of 12 nucleotides. This oligomer is then removed by UvrD, the helicase, by breaking the complementary hydrogen bonds as it separates the strands. The 12 nucleotide gap is then filled by DNA polymerase I, followed by DNA ligase sealing the nick (Van Houten 1990).

The eukaryotic NER system can be separated into two sub-pathways: Global Genomic NER (GG NER) that acts in transcriptionally inactive and active sites and

Transcription Coupled NER (TC NER) that acts in transcriptionally active sites (Shuck *et al.* 2008). Both sub-pathways use the same mechanisms, except for recognition where GG NER utilizes a protein complex, while TC NER utilizes stalled transcription machinery for the recognition of lesions. Following recognition, the mechanism remains the same between the two sub-pathways. Helicases, in the near vicinity of the damage, unwind the double-stranded DNA, allowing the endonucleases to nick the damaged strand 5′ and 3′ of the lesion. Approximately 30 nucleotides (Batty & Wood 2000), including the lesion, are removed as an oligonucleotide. The gap is then filled by DNA polymerase δ or ε and the gap is sealed by DNA ligase, restoring the integrity of the DNA (Shuck *et al.* 2008).

The general mechanism of eukaryotic NER is the same as that of *E. coli*, but the logistics differ (Batty & Wood 2000). In eukaryotes, 9 major proteins have been found to play a role in NER, with different sets contributing to GG NER and TC NER because latter does not require a protein complex for lesion recognition. Proteins contributing to recognition in GG NER include the XPC-Rad23 (Xeroderma pigmentosum group C- UV excision repair protein 23) complex; RPA; and transcription factor IIH (TFIIH) (Mu *et al.* 1995; Shuck *et al.* 2008). The key component in recognition seems to be XPA, which only functions in GG NER, while TC NER uses the stalled transcription machinery as recognition to trigger downstream steps. Following recognition the XPC-Rad23 complex recruits TFIIH, which plays a role in both NER and transcription initiation. This transcription factor has 9 subunits, of which 2 have helicase activity that aid in unwinding the damaged DNA (Shuck *et al.* 2008). Once unwound, XPG and XPF-ERCC (Excision repair cross-complementing rodent repair deficiency, complementation group 1) create nicks 5′ and 3′ to the lesion, leaving a gap of approximately 24-32 nucleotides (Shivji *et al.* 1995). The gap is then filled by the joint action of replication factor C (RFC), PCNA, DNA polymerase δ or ε, DNA ligase I and RPA. This gap filling pathway, with DNA ligase I, seems to be highly dependent on the cell cycle and is only triggered in late G1 phase (Shivji *et al.* 1995; Shuck *et al.* 2008). An additional pathway was found in quiescent cells that only utilizes DNA polymerase δ, XRCC1 (X-Ray cross-complementing) and DNA ligase IIIα (Moser *et al.* 2007).

The last pathway for single-strand repair, MMR, specifically targets polymerase errors, such as mismatches and indels forming insertion/deletion loops (IDLs) (Schofield & Hsieh 2003). This pathway is also sensitive enough to identify mismatches due to the deamination of $5^{m}C$ (a cytosine nucleotide methylated at the 5^{th} atom), causing a guanine-thymine mismatch, which could possibly lead to a guanine-adenine transition. The process of

mismatch repair is similar to that of the excision repair pathways. It starts with lesion recognition, excision of the lesion and the surrounding nucleotides, followed by their replacement and completed with the replaced nucleotides' re-ligation (Hsieh 2001; Kolodner & Marsischky 1999; Li 2008; Schofield & Hsieh 2003).

Gram negative bacteria, such as *E.coli*, employ a system known as methyl directed mismatch repair (Hsieh 2001; Li 2008), where methyl groups are used to direct the repair machinery to the newly synthesized daughter strand. This pathway contributes significantly to the fidelity of replication. It is initiated by MutS (as a homodimer) binding IDLs and mismatches (except cytosine-cytosine mismatches). MutS then recruits another homodimer, MutL which then activates MutH. MutH is an endonuclease that searches the DNA for the closest GATC site to the lesion and cleaves the unmethylated strand at the hemimethylated GATC site. The niche created by the endonuclease then acts as a point of entry for the single stranded binding proteins and DNA helicase II, which then bind the DNA through MutL (Hsieh 2001; Li 2008). This pathway is capable of bidirectional repair, where the direction of repair depends on the position of the mismatch with regards to the GATC site. Depending on whether the GATC site is located upstream or downstream of the lesion, a 5′ to 3′ exonuclease (ExoI/ExoX) or a 5′ to 3′ exonuclease (RecJ or Exo VII respectively) excises the nucleotides between the nick and the lesion (Hsieh 2001). This gap is then filled by DNA polymerase III and sealed by DNA ligase.

The process of MMR appears to be highly conserved among prokaryotes and eukaryotes, yet the recognition processes differ greatly. Exactly how non model bacteria and eukaryotes go about recognizing the lesions remains uncertain. In eukaryotic organisms, the MutS homologs are two heterodimers, Msh1/Msh6 (MutSα) and Msh2/Msh3 (MutSβ) (Hsieh 2001; Kolodner & Marsischky 1999; Schofield & Hsieh 2003). The MutSα heterodimer repairs both substitutions and small IDLs, whereas MutSβ only repairs IDLs. Similar to MutS, MutL also has 2 homologs, Mlh1 and Pms1 - forming a heterodimer, while MutH has no eukaryotic homolog. In eukaryotes the endonuclease activity is taken over by the MutL homologs which have a 5′ to 3′ exonuclease activity (Schofield & Hsieh 2003). Because this exonuclease activity requires a free 3′ end, there seems to be a bias in the strands that is fixed. As with the methyl directed mismatch repair, the eukaryotic pathway also directs mismatch repair to the newly synthesized strand by utilizing the free 3′ ends of the Okazaki fragments (Hsieh 2001; Kolodner & Marsischky 1999; Schofield & Hsieh 2003).

4.3 Double-stranded breaks

Double-stranded DNA breaks form when the complementary strands of the double helix break close to each other at the same time, in such a manner that base pairing and the chromatin structure cannot keep the ends juxtaposed. This damage is considered the most dangerous type of damage in terms of mutation and cell death. These breaks are mostly brought about by mechanical stress (Jackson 2002; Lieber 2010) on chromosomes and when the DNA polymerase encounters a lesion or a single stranded binding protein. In addition to this, these breaks are formed as an intermediate during recombination (Jackson 2002), which if not controlled, can lead to inappropriate recombination. There are two main pathways responsible for the repair of these breaks: homologous recombination (HR) (Jackson 2002; Lieber 2010; Wyman *et al.* 2004); and NHEJ (Figure 3)(Hefferin & Tomkinson 2005; Jackson 2002; Lieber 2010). How the cell determines which pathway to use for fixing the damage remains unclear, but the cell cycle phase seems to play a role (Jackson 2002).

HR is the exchange of genetic material between two sequences that share significant homology. Both single-stranded gaps and double-stranded breaks initiate HR and repair using an intact homolog (Wyman *et al.* 2004). This mechanism of repair is used in both prokaryotes and eukaryotes where the processes of homology recognition and strand exchange are very similar, with the only difference being that the eukaryotic mechanism is more complex. This repair mechanism can be divided into three steps: Firstly, pre-synapsis during which the single-strand is processed to form a recombination proficient strand; secondly, synapsis when the processed strand invades the homolog in order to produce a joint molecule; and finally post-synapsis during which the repair happens and the strands that form part of the joint molecule are separated (Wyman *et al.* 2004).

In *E.coli*, pre-synapsis requires a single-stranded DNA sequence that can act as a substrate for RecA, the recombinase. Here, a free 3′ single-stranded end is formed when the RecBCD protein complex recognizes a specific DNA sequence motif (i.e., chi site) and starts processing the broken end (Kowalczykowski 2000; Smith 2001). In addition to the nucleolytic activity of this complex, it is also responsible for unwinding the DNA and loading RecA, forming a nucleoprotein filament (Wyman *et al.* 2004). The second step of HR is when the nucleoprotein filament recognizes a sequence with great homology to the broken end, and invades this intact double-stranded homologous region forming a joint structure known as a Holiday junction. Migration of the Holiday junction branch is then facilitated by RuvAB,

facilitating repair of the DNA lesion. The Holiday junction is then resolved by RuvC that binds to the RuvAB complex and dissolves the joint structure to its original partners (West 2003).

Eukaryotic organisms are expected to have proteins that have functions analogous to those mentioned above (identified in *E.coli*). But the only protein that has been found to be greatly conserved at amino acid level in eukaryotes is RecA (Wyman *et al.* 2004), with the eukaryotic homolog being Rad51 (Shinohara *et al.* 1992). The picture in eukaryotes is especially complicated due to the abundance of proteins able to perform the required functions, which is possibly why so little is known about eukaryotic proteins involved in pre-synapsis. It was proposed that the MreII complex consisting of MreII (meiotic recombination 11), Rad50 and NBS1 (Nijmegen breakage syndrome protein 1) plays a role in pre-synapsis (Cromie *et al.* 2001; Wyman *et al.* 2004) due to its affinity for single-stranded ends (Chen *et al.* 2001). Rad51 functions like RecA in the formation of a nucleoprotein filament, facilitating homolog invasion and strand exchange (Shinohara *et al.* 1992). However, strand exchange appears to require many additional proteins such as Rad54 (UV excision repair protein 54) (Sung *et al.* 2003), because Rad51 is not as efficient as RecA. Thus, even the most conserved step of the process, strand invasion, seems to differ between prokaryotes and eukaryotes. The Holiday junction was found to be resolved by a Rad51 paralog, Rad51C (Liu *et al.* 2004; Wyman *et al.* 2004). Whether this protein acts as the resolvase in itself, or activates the true resolvase, also remains unclear, but ongoing research promises some insight into this issue and HR as a whole.

HR is highly dependent on a homolog, with high sequence homology, and thus can only occur if the DNA is diploid or has already copied its DNA (Cromie *et al.* 2001). However, all organisms, even if diploid, lack the presence of a second homolog during the S phase of the cell cycle, which has forced the cell to evolve an alternative mechanism (i.e., NHEJ) to fix lesions during this phase (Lieber 2010). How the cell determines which mechanism to use is an area of active research, but it was proposed that this decision might be operational and that NHEJ is used if a homolog is not present near a break (e.g. during the G2/S phase). In contrast, the presence of a sister chromatid (homolog) close to the break favours HR (Lieber 2010).

NHEJ represents an intriguing process because of its diversity at different levels. The process seems to have evolved independently in prokaryotes and eukaryotes, although it

shares the same driving forces yielding mechanistically flexible systems in both (Hefferin & Tomkinson 2005; Lieber 2010). This thought to be linked to a fundamental requirement for substrate diversity – i.e., diverse types of damage that has to be fixed. In addition, this process is also recognized for the flexible activities of its nuclease, polymerase and ligase (Lieber 2010). NHEJ also does not convert the DNA to the original sequence, thus rendering a diversity of end results (Hefferin & Tomkinson 2005).

In lower eukaryotes, HR was found to be the most important repair pathway, thus NHEJ only occurred when HR was inactive (Hefferin & Tomkinson 2005). NHEJ functions in a stepwise manner starting with the two protein complexes - Hdf1/Hdf2 (human diploid fibroblast) and Dnl4/Lif1 (analogous to mammalian DNA ligase IV/XRCC4) - binding to the exposed end. This is followed by end bridging which is facilitated by the XRM (Rad50/MreII/Xrs2) protein complex (Chen *et al.* 2001) - with the nuclease activity of MreII facilitating alignment (Paull & Gellert 2000). Once aligned, the protein complex recruits Dnl4/Lif1, completing the bridging process. It was found that all proteins contributing to end joining do so through the interaction of Xrs2 and Lif1. Evidence for direct interaction with Hdf1/Hdf2 is still elusive, but it has been suggested that these are important in species-specific functional interactions (Chen *et al.* 2001). Once the ends are bridged, processing, gap filling (DNA pol IV) and ligation completes the repair process (Wyman *et al.* 2004).

In prokaryotes the NHEJ process is mechanistically similar with some minor changes (Wyman *et al.* 2004). Proteins showing homology to the eukaryotic end-binding proteins are encoded by single copy genes leading to the formation of homodimers, instead of heterodimers as in eukaryotes (Wyman *et al.* 2004). In *Mycobacterium tuberculosis*, MtuKu (*M. tuberculosis* Ku homolog) binds in a sequence independent manner and slides along the DNA sequence, as is the case in eukaryotes. This bacterium also has three ligases, one of which has primase (polymerase) and nuclease functions (Gong *et al.* 2004). This suggests that a single polypeptide is responsible for linking processing and ligation reactions, which is different to eukaryotes where different proteins perform separate functions.

4.4 Translesion Synthesis (TLS)

It is a common occurrence that some DNA damaging agents lead to more mutations than could be expected based only on the types of lesions formed (Snyder & Champness 2007). This was found true for all organisms, from bacteria to humans, implying that certain DNA repair pathways are error prone (Goodman 2002). Early studies on *E. coli* suggested

that this error prone pathway is encoded for by the SOS genes and was named the TLS pathway. Because of this pathway's tendency to incorporate errors, it is only used as a last resort when it is essential for the survival of the cell to continue replication despite the damage. TLS is thus only activated when the damage is too severe to be repaired in a timely fashion by other pathways.

Since its discovery, much research has been conducted with regards to the functioning of TLS. It was found that UmuC (DNA polymerase V subunit UmuC) and UmuD (DNA polymerase V subunit UmuD) proteins allow the cell to tolerate damage but, in the process, increases mutagenesis in *E.coli* (Friedberg *et al.* 2002; Lehmann 2006). These proteins function as heterotrimers ($UmuD_2C$) (Lehmann 2006) that forms upon expression of the UmuDC operon, following extensive damage to the DNA. The *umuC* gene encodes a polymerase (Friedberg *et al.* 2002; Lehmann 2006) with a larger active domain than normally found, allowing for the incorporation of nucleotides opposite severe DNA damage. This polymerase, however, is inactive in the heterotrimer but binds to DNA polymerase III. This leads to autocleavage of UmuD in the presence of accumulated RecA nucleoprotein filaments (Snyder & Champness 2007). This step acts as a checkpoint, allowing other non-error prone repair pathways to act prior to TLS activation. The cleavage of UmuD activates UmuC, allowing replication to proceed past the lesion by incorporating random nucleotides, which could lead to mutations.

In eukaryotes it was found that the *rad6* and *rad18* (Bailly *et al.* 1994) genes are involved in TLS. Rad6 occurs in a complex with Rad18, a DNA binding protein, which together are responsible for binding to a stalled polymerase at a lesion as to allow replication. Furthermore, the pathway involves three additional proteins, REV1 (protein reversionless 3), Rev3 and Rev7, although the function of the latter is not essential for TLS (Johnson *et al.* 1994). Rev3 and Rev7 form the subunits of DNA polymerase ζ, the translesion polymerase, while Rev1 is a deoxycytidyl transferase responsible for inserting a dCMP opposite an abasic site. For activation of the translesion polymerase, eukaryotes employ the conserved protein ubiquitin (Ub) that binds the amino group in lysine instead of the RecA homolog Rad51 (Andersen *et al.* 2008). A commonly found Ub-like protein, small Ub like modifier (SUMO), was found to regulate protein function through conjugation (Prakash & Prakash 2002). In eukaryotes the polymerase switch is regulated through the ubiquitination of PCNA, DNA polymerase processivity factor. PCNA, in the form of a homotrimer, encircles the DNA and aids in the assembly of DNA synthesis proteins and the unwinding of the double strand. It

was found that PCNA is only ubiquitinated following DNA damage, allowing the specialized polymerase to function (Andersen *et al.* 2008; Prakash & Prakash 2002).

5. Mutation rate modulation

Mutation rate is defined as the amount of mutations that will arise within the genome of a cell during its lifetime (Snyder & Champness 2007; Drake *et al.* 1998). Mutation rate should not be confused with the number of mutations that accumulate at varying rates within the genome and at different genomic regions. For example, mutations in so-called "junk DNA" accumulate at a high rate because they apparently do not affect the functioning of the organism and are thus a consequence of lower levels purifying selection. Terms that are commonly used in spontaneous mutation rate studies are summarized in Table 1.

Natural selection has a significant effect on mutation rate, but exactly how it affects mutation rate is subject to debate (Sniegowski *et al.* 2000). In general it has been proposed that there are four control points whereby natural selection can control mutation rate: DNA replication; mutagen exposure; DNA repair; and mutational buffering. How the first three control points may influence mutation has been discussed in Sections 2-4. Mechanisms acting to buffer the effects of mutations on the fitness of the organism will indirectly affect mutation rates (Baer *et al.* 2007). Here redundant elements in DNA (introns and repetitive DNA) or proteins (redundant amino acids) are thought to function as buffers by aiding the retention of the functional DNA/protein conformation (Conrad 1985). Examples of such mechanisms include chaperones, such as Hsp90 (heat shock protein 90) (Chiosis *et al.* 2004), and repeat-induced point mutation (RIP) pathways such as that found in *Neurospora crassa*. The RIP pathway detects DNA duplications prior to meiosis and interrupts them with G:C to A:T mutations (Singer *et al*. 1995).

6. Mutation rates in microbes

Cellular organisms such as microbes have a lower genomic mutation rate than elements with RNA as their genetic material (Drake 1991). This can mainly be ascribed to the highly error prone RNA polymerase that is responsible for the replication of the RNA genomes. Not

only is the genomic mutation rate (μ_g) of microbes lower, but their mutation rate per base pair (μ_b) is lower (Table 2) (Drake 1991).

There are clear patterns that can be observed with regards to the μ_b, μ_g and genome size (G) (Table 2). In general μ_g remains relatively constant when comparing the prokaryotes, and even *S. cerevisiae*, *S. pombe* and *N. crassa*. Drake (1998) recorded this value as conspicuously robust compared to other constants in evolutionary processes. Drake (1998) also predicted that μ_g will remain constant even in organisms exposed to hypermutagenic environments. This prediction was addressed by Grogan *et al.* in 2001, who showed that μ_g had a value of 0.0018 for the archaeon, *Sulfolobus acidocaldarius*, which is indeed close the average for non extremophiles (0.0034). This is in contrast to μ_b and genome size (G) that vary over 4 orders of magnitude. The difference in μ_b is compensated for by G seeing as $\mu_g = \mu_b / G$. A high μ_b will thus be cancelled out by a large G and yield a lower μ_g value.

6.1 Mutation rate studies in prokaryotes

The *in vivo* role of mutation rates in the adaptation of prokaryotes to their ever changing and stressful environments was studied by Giraud *et al.* in 2001. Here it was found that mutator bacteria (bacteria bearing alleles that increase mutations at other loci) start to dominate in a short period of time during colonization of the mouse gut. In order to determine whether an increase in the mutation rate resulted in rapid adaptation, the authors conducted competition studies between the mutator bacteria and wild types, where it was found that the mutator had an advantage over the wild type. The advantage of such a mutator is greatly dependent on its capacity to induce advantageous mutations which are then rapidly fixed in the population, following which, the advantage conferred by the mutator phenotype seems to vanish. It became clear in that study that the success of the mutator phenotype was not hindered by the deleterious mutations generated, as long as adaptive mutations arose. The authors came to the conclusion that the heterogeneity of the environment possibly favours flexibility of mutation rate (Giraud *et al.* 2001).

An increase in mutation rate is linked to the growth conditions of a prokaryotic population. The scarcity of nutrients challenges microbes to maintain a steady growth rate, and most respond by entering a state known as stationary phase (Navarro Llorens *et al.* 2010). In 1988 a paper was published by Cairns *et al.* challenging the idea of Luria and Delbrük in 1943 who suggested that mutations arise spontaneously independent of selection pressures. It was proposed that mutations can be directed to certain sites by certain

environmental stimuli, forming adaptive mutations (Navarro Llorens *et al.* 2010). Stationary phase mutations are adaptive mutations that are brought about when the cells are exposed to non-lethal stresses, increasing evolutionary processes during times of stress (McKenzie and Rosenberg 2001). The most common mechanism used to increase the mutation rate during these stress conditions is by repressing the methyl directed mismatch repair system (Navarro Llorens *et al.* 2010). Ultimately the role of stress-induced mutations is controversial, seeing as some studies (Roth *et al.* 2006) showed that the increased mutation rate during stress conditions would have a negative impact on the fitness of the population over a long period of time (Navarro Llorens *et al.* 2010).

In a study conducted by Gerrish and Garcia-Lerma in 2003, the prokaryotic mutation rate was considered from a medicinal point of view by studying treatment efficacy. The main factor determining the efficacy of a treatment was the fitness of the microbial population that persevered despite treatment application (Gerrish & García-Lerma 2003). Generally speaking the efficacy of drug treatment is thought to be inversely proportional to the fitness of the population. In turn, the fitness of the infecting microbe is influenced by its genomic mutation rate in two ways. The more obvious of the two is that with an increase in the mutation rate the amount of pre-existing mutants will increase. The second way in which it impacts, and which is also thought to dominate when mutation rates are high, is the increase in harmful mutations. These harmful mutations can then render the persevering microbes less viable or non-viable. Since the only way for fitness recovery is through either the reversion of harmful mutations or the acquisition of mutations counteracting the harmful effects, drug treatment has been found to be most effective when the mutation rate of the pathogen is either extremely high or low (Gerrish & García-Lerma 2003).

6.2 Mutation rate studies in Archaea

By 1997, mutation rate studies had only been conducted for various bacteriophages (phage M13, phage λ, phage T2, phage T4) and *E. coli*, *S. cerevisiae* and *N. crassa* (Grogan *et al.* 2001). Jacobs and Grogan addressed the information gap for Archaea by studying the mutation rate of *S. acidocaldarius*. The results indicated that the mutation rate of the phenotype studied was 3.37×10^{-7} per cell division i.e. 3.37×10^{-7} mutations per cell division renders a resistant phenotype). The data was then used to determine the genomic mutation rate that was calculated to be 0.0018 mutations per genome per replication. Surprisingly, when this value was comparable to that of the other DNA based microbes - whose average

mutation rate was calculated to be 0.0034 (Jacobs & Grogan 1997). These results were surprising because they showed no variation between organisms that grow at ambient temperatures and high temperatures, where higher mutagenesis is expected.

6.3 Mutation rate studies in eukaryotes

Tian *et al.* 2008 conducted a study in which the relationship between indels (insertions and deletions) and diversity was studied. Here it was found that as the distance to the nearest indel increased, and the indel interval length increased, the diversity or divergence (D) decreased. The data was shown to support the indel associated mutation hypothesis (Longman-Jacobsen *et al.* 2003; Petrov 2002) that predicted an increase in substitutions close to an indel. It was concluded that indel-associated mutations occur throughout the kingdom Eukarya, stressing the importance of indels in genome evolution (Tian *et al.* 2008).

It was recorded by Fox *et al.* in 2008 that eukaryotic species such as insects and all characterized mammals showed a regional bias with regards to the mutation rate in different genomic regions. This observation was made by studying the mutation rate in gene synonymous sites. In strong contrast to this, a uniform mutation rate was observed in all fungi, except for the genus *Candida* (Fox *et al.* 2008). The genome wide mutation rate of *S. cerevisiae* has been calculated as 0.0031. For calculations in *N. crassa*, two genomic regions were studied, the *ad-3AB* (purple adenine genes) and *mtr* (methyltryptophan resistance) genes. Here μ_{bp} was found to be $4.47^{10\text{-}11}$ and $9.96^{10\text{-}11}$ respectively and μ_g was calculated as 0.00187 and 0.00417 respectively (Drake 1991). In conclusion Drake estimated the average genomic mutation rate for all DNA based microbes to be 0.003 (Drake 1991).

7. Variation between mutation rates

Neutral mutation rate (mutations that are not selected for or against) was thought to be uniform for a long time (Baer *et al.* 2007; Fox *et al.* 2008). However, it has been discovered that mutation rate not only varies from one species to the next, but it even varies within a genome (Zeyl & DeVisser 2001; Baer *et al.* 2007; Fox *et al.* 2008). Heterogeneity within the genome has been confirmed in a variety of species such as humans, chimps, dogs and cows, whereas a uniform mutation rate was found among *Saccharomyces* species (*S. cerevisiae*, *S. paradoxus*, *S. bayanus* and *S. mikatae*) (Fox *et al.* 2008). It has previously been found that the

frequency of gene inactivation due to mutations can also vary greatly and is dependent on factors such as the organism, gene size and the sensitivity of the gene to base pair substitutions (Drake 1999).

Mitochondrial DNA (mtDNA) is thought to have a higher mutation rate than that of the nuclear genome (Brown *et al.* 1979). It is this characteristic that makes mtDNA a rapidly evolving sequence, and thus a target for phylogenetic studies (Baer *et al.* 2007). The higher substitution rate is currently ascribed to a higher mtDNA mutation rate and is not due to relaxed selection on this organellar genome, as was shown in a study comparing the nuclear and mitochondrial genomes of *Caenorhabditis elegans* (Denver *et al.* 2000; Denver *et al.* 2004). In contrast, it was found that the nuclear and organellar silent mutation rates are in the same order of magnitude in non-metazoan eukaryotes (Lynch *et al.* 2006). The situation in plants is also different as it was found that the nuclear substitution rate was lower than that of organellar substitution rates (Palmer & Herbon 1988). The fact that metazoa apparently lack certain mitochondrial activities that occur in non-metazoa might form the foundation for a possible explanation for these differences. An example of such an activity is MutS mediated repair systems (Baer *et al.* 2007).

When one considers heterogeneity within the genome, little is known concerning the observed regional biases. Because most previous studies focussed on mammalian genomes, mutation rate studies in a range of other species might aid in elucidating genomic mutational properties (Fox *et al.* 2008). Currently there are a variety of possible explanations for heterogeneity, including base composition, gene density, local recombination rate, gene expression patterns and chromatin structure (Baer *et al.* 2007; Fox *et al.* 2008). For example, in a study conducted by Lichtenauer-Kaligis *et al.* (1993) it was found that the substitution rate of a reporter gene, integrated into the human genome, varied up to 60 times depending on the site of integration.

Fox *et al.* (2008) used 27 fungal species to study heterogeneity in the fungal kingdom. It was concluded that most fungi had uniform mutation rates because regions with high or low mutation rates could not be found (Fox *et al.* 2008). Heterogeneity could only be detected in 5 of the 27 fungal species (*Candida albicans, C. dubliniensis, C. tropicalis, N. crassa* and *C. globosum*). These species were found to encode ribosomal and metabolic genes with mutation rates that deviated from the global mutation rate. These genes are known to be under selection for codon usage bias, which could be a possible explanation for the deviation.

To confirm the obtained results, a neighbouring gene analysis was done which supported the results, showing that fungal species generally do not exhibit the regional mutation rate heterogeneity as found in mammalian species (Fox *et al.* 2008).

8. Estimating mutation rate

Methods for determining the genomic mutation rate in different organisms have to take into account the diversity of both mutation types and pathways responsible for mutagenesis (Drake 1991). Five criteria should be satisfied by such methods: (i) the mutational target should be large enough as to be representative sample of the genome; (ii) if a growth rate difference is observed between the wild type and mutant, it should be accounted for in downstream calculations; (iii) if the expression of the generated mutations suffers from phenotypic lag, it is to be quantified and mathematically nullified; (iv) the system should be explored to such an extent that all limits and artefacts of the system are known; and lastly (v) it is required that the mutational spectrum (mutation pattern) is described at a molecular level as to ensure that all of the kinds of mutations are represented and to determine what fraction of mutations escapes detection (Drake 1991).

There are two methods that can be used to calculate mutation rate – mutant accumulation assays and fluctuation analysis (Balin & Cascalho 2010; Drake 1991; Foster 2006; Baer *et al.* 2007). Both of these methods rely on the Lea-Coulson Model for expansion of mutant clones, which in turn relies on the following assumptions: exponential growth of cells; mutational events occur independently of previous mutation events; the chance for mutation is equal in each generation; the mutant fraction is small; reversion mutations are negligible; no mutants go undetected; mutants only arise prior to selection; the initial cell number is negligible compared to the final number; and lastly, the probability of the mutation per generation remains constant throughout culture growth (Foster 2006).

Mutant accumulation assays regards the mutation rate as the increase in mutant fraction over time (Foster 2006). As time passes, the mutant fraction increases due to the proliferation of pre-existing mutants as well as the generation of new mutants. Something that is important to remember when performing these assays is the population size, the population must be large enough as to limit or eliminate the domination of chance mutations. For batch

cultures, the population size should be sufficiently large to ensure that the average number of mutations per culture is higher than one (Foster 2006). Although this method is conceptually simple, it poses the problem that one cannot accurately determine the accumulation of new mutants, because clones are produced from existing mutants (Balin & Cascalho 2010; Foster 2006). In order to overcome this obstacle, a large population with a small amount of mutants needs to be generated e.g. testing and selecting a population with a low mutant fraction or purging pre-existing mutants (Foster 2006).

Fluctuation analysis was originally developed by Luria and Delbruck to determine whether bacteria acquired resistance through spontaneous mutations or as a result of selective pressure (Luria & Delbruck 1943; Balin & Cascalho 2010). A problem encountered with this analysis was that mutants produced clonal offspring which then increase the apparent mutation rate if mutants were to be counted directly. This problem was addressed by determining the average number mutations that occur per culture. This was done by plotting the frequency of cultures lacking resistant mutants, to the equation that describes the zero order of the Poisson process (Balin & Cascalho 2010).

The Poisson process, also known as the Poisson distribution, is a probability distribution which expresses the probability of a number of events occurring in a set amount of time, given the events occur at a known average and an event occurs independently of the previous one (Balin & Cascalho 2010). This distribution is expressed as follows:

$$f(\mathrm{k};\lambda) = \frac{\lambda \mathrm{e}^{-\lambda}}{k!}$$

Where: f (k, λ) represents the probability that there are exactly k occurrences

e = the base of the natural algorithm

k = the amount occurrences of an event

λ = expected amount of occurrences during the set time

Balin and Cascalho conducted a study in 2010 in order to study the feasibility of determining the mutation rate using DNA sequence information. Here a gene of interest was sequenced, and the average amount of mutations was calculated using the number of non-mutated positions according to the Poisson distribution. Assumptions made when this equation was derived are that the probability of each base mutating in the sequence was

equal, and the probability of a base mutating twice was negligible. It was concluded that the mutation rate of a gene can be calculated using direct sequencing methods independent of reporter genes (Balin & Cascalho 2010).

Determining whether the target organism grows as free cells or as mycelia is an important consideration when calculating the mutation rate of that organism. (Baracho & Baracho 2003). The aforementioned methods for calculating spontaneous mutation rates can only be applied to microbes that grow as single cells (Alvarez-Perez *et al.* 2010; Foster 2006). Mycelial organisms complicate spontaneous mutation rate calculations, leaving the results incomplete (Baracho & Baracho 2003), because it is impossible to determine whether the mutation occurred during spore formation of previous nuclear divisions. As for unicellular organisms, ideally one would determine the mutation probability per nucleus and per generation, but the only feasible way of obtaining nucleus samples for filamentous organisms is by collecting spores (for calculating the proportion of mutants) which allows for miscalculations that are based on the lack of knowledge about whether the mutations in the spores occurred during spore formation or prior nuclear divisions (Alvarez-Perez *et al.* 2010; Baracho & Baracho 2003). However, Baracho and Baracho (2003) demonstrated that the Kolmogorov-Smirnov test can be applied to link the mean number of nuclei that underwent mutation to the mean number of mutant conidia facilitating the quantification of the mutation rate per nucleus per generation (Baracho & Baracho 2003).

In the study conducted by Alvarez, Perez *et al.* (2010) a fluctuation analysis was done to illustrate that rare, random spontaneous mutations were responsible for the transformation of elastase$^-$ (EA^-) *Aspergillus fumigatus* strains to elastase$^+$ (EA^+) strains. In addition to this, the frequency at which the EA conferring mutation occurred was determined using an alternative approach. Instead of measuring colony size and counting conidia per colony, the DNA duplication rate (DDR) was determined. This was done using a nuclear stain, 4′, 6-diamidino-2-phenylindole or DAPI. This stain fluoresces, and thus allows for nuclear quantification by measuring fluorescence. The following formula was then applied to calculate the rate at which spontaneous mutations converted EA^- strains to EA^+ strains:

$$\mu = -ln\frac{\mathrm{Po}}{(\mathrm{Fo} - \mathrm{Ft})}$$

Where P_o = the proportion of non-mutant cultures

F_o= fluorescence at the beginning of the experiment

F_t = fluorescence at the end of the experiment

By applying this, the authors determined the mutation rate to be 6.17 x 10^{-8} mutants per nuclear division under lab conditions. It was stressed that the mutation rate will most likely deviate from this estimation under natural conditions (Alvarez-Perez *et al.* 2010).

9. Conclusions

Mutagenesis is a well known, deeply studied occurrence in model organisms. Research based on the functional processes of mutagens as well as repair machinery has given us the necessary tools for understanding the underlying principles of mutagenesis. Variation in the process itself, as well as its regulation undoubtedly occurs between different organisms. Therefore different mutagenic sources will fulfil different roles in different organisms, depending on the exposure and repair pathways available.

The process of spontaneous mutagenesis is well regulated process in all cells. Although extensive research has been done on how these mutagens act to cause substitution mutations, the context of these mutations seem to be less well understood. The dependence on certain sequences for recognition and repair renders the fixation of spontaneous mutations non-random. This non-random process is less well studied, with studies mainly limited to humans and certain bacterial genes. The inability to predict where mutations are most likely to take place, in evolutionary important regions (genes), leads to an incomplete understanding of the evolution of the genes.

A variety of selective forces act to form a level of tolerance to mutations at different regions of the genome. It is likely that mutations in genes with a vital function will not be tolerated at the same level as those genes with a redundant function. One will therefore expect different patterns of mutagenesis in the two groups of genes due to two different levels of selection. Firstly, selection has acted in over long periods of time to shape the genome e.g. gene position. Important genes, with housekeeping functions, are positioned close to the origin of replication in bacteria as to allow high fidelity. In addition to this, certain pathways

of repair are directly linked to the level of gene transcription, further increasing the fidelity of duplication of constitutively expressed genes. On the second level of selection one finds housekeeping genes to be under purifying selection, eliminating individuals with a variety that causes a reduction in fitness. This is in contrast to species specific genes that may be under diversifying selection.

Genomic mutation rate is a good quantitative measure in estimating mutagenesis across the genome. Comparisons between genomic mutation rates of different species can elucidate the rate of evolution of the species. In addition to this comparison, intra-genomic comparisons to that of the genomic mutation rate will illustrate the different levels of tolerance for mutations. The lack of mutation rate data not only hinders the understanding of species evolution but also genome evolution and gene evolution within a species.

The research conducted in this study aims to address the gaps in the literature. The study of context of substitution mutations will form the basis for understanding which spontaneous mutagens play significant roles in the evolution of the species. The role of different mutagens differs between regions of the genome, contributing to the variation in mutability of different genomic regions and their evolution. In addition to addressing mutation variation within the genome the estimation of genomic mutation rate will allow for cross species comparisons. Furthermore, the availability of mutation rate data across species and lineages might aid in comprehending the evolution of the species.

10.References

Abraham RT, 2001. Cell cycle checkpoint signaling through the ATM and ATR kinases. *Genes and Development* **15**: 2177–2196.

Albertson T, Preston B, 2006. DNA Replication Fidelity: Proofreading in Trans. *Current biology*, **16**: 206–209.

Alvarez-Perez S, Blanco JL, Lopez-Rodas V, Flores-moya A, Costas E, Garc ME, 2010. Elastase activity in *Aspergillus fumigatus* can arise by random , spontaneous mutations. *International Journal of Evolutionary Biology*: doi:10.4061/2010/602457

Andersen PL, Xu F, Xiao W, 2008. Eukaryotic DNA damage tolerance and translesion synthesis through covalent modifications of PCNA. *Cell Research* **18**: 162–173.

Arana ME, Kunkel TA, 2010. Mutator phenotypes due to DNA replication infidelity. *Seminars in Cancer Biology* **20**: 304–311.

Auerbach C, 1959. Spontaneous mutations in dry spores of *Neurospora crassa. Zeitschrift für Vererbungslene* **90**: 335-346.

Baer CF, Miyamoto MM, Denver DR, 2007. Mutation rate variation in multicellular eukaryotes : causes and consequences. *Nature* **8**: 619–632.

Bailly V, Lamb J, Sung P, Prakash S, Prakash L, 1994. Specific complex formation between yeast RAD6 and RAD18 proteins: a potential mechanism for targeting RAD6 ubiquitin-conjugating activity to DNA damage sites. *Genes & Development* **8**: 811–820.

Balin SJ, Cascalho M, 2010. The rate of mutation of a single gene. *Nucleic Acids Research* **38**: 1575–1582.

Baracho MS, Baracho IR, 2003. An analysis of the spontaneous mutation rate measurement in filamentous fungi. *Genetics and Molecular Biology* **26**: 83–87.

Barsky D, Foloppe N, Ahmadia S, Wilson DM, Mackerell AD, 2000. New insights into the structure of abasic DNA from molecular dynamics simulations. *Nucleic Acids Research* **28**: 2613–2626.

Batty DP, Wood RD, 2000. Damage recognition in nucleotide excision repair of DNA. *Gene*, **24**: 193–204.

Bayir H, 2005. Reactive oxygen species. *Critical Care Medicine* **33**: 498–501.

Boiteux S, Guillet M, 2004. Abasic sites in DNA : repair and biological consequences in *Saccharomyces cerevisiae*. *DNA Repair* **3**: 1–12.

Bremer H, Dennis PP, 1996. Modulation of chemical composition and other parameters of the cell by growth rate. *Escherichia coli and Salmonella: cellular and molecular biology* ***2***: 1553-1569.

Brown WM, George M, Wilson AC, 1979. Rapid evolution of animal mitochondrial DNA. *Proceedings of the National Academy of Science* **76**:1967–1971.

Burney S, Caulfield JL, Niles JC, Wishnok JS, Tannenbaum SR, 1999. The chemistry of DNA damage from nitric oxide and peroxynitrite. *Mutation Research* **424**: 37–49.

Cairns J, Overbauch J, Miller S, 1988. The origin of mutants. *Nature* **335**: 142-145.

Chen L, Trujillo K, Ramos W, Sung P, 2001. Promotion of Dnl4-catalyzed DNA end-joining by the Rad50/Mre11/Xrs2 and Hdf1/Hdf2 complexes. *Molecular Cell* **8**: 1105–1115.

Chiosis G, Vilenchik M, Kim J, Solit D, 2004. Hsp90 : the vulnerable chaperone. *Drug Discovery Today* **9**: 881–888.

Crick FHC, 1966. The genetic code- yesterday, today, and tomorrow. *Old Spring Harbour Symposium on Quantitative Biology* **31**: 3-9.

Conrad M, 1985. The mutation buffering concept of biomolecular structure. *Journal of Biosciences and Bioengineering* **8**: 669–679.

Cromie GA, Connelly JC, Leach DR, 2001. Recombination at double-strand breaks and DNA ends: conserved mechanisms from phage to humans. *Molecular cell* **8**: 1163–74.

Daubin V, Perrière G, 2003. G+C3 structuring along the genome: a common feature in prokaryotes. *Molecular biology and evolution* **20**: 471–83.

David SS, O'Shea VL, Kundu S, 2007. Base-excision repair of oxidative DNA damage. *Nature*, **447**: 941–50.

De Flora S, Bagnasco M, Serra D, Zanacchi P, 1990. Genotoxicity of chromium compounds. A review. *Mutation Research* **238**: 99–172.

Den Dunnen JT, Antonarakis SE, 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. *Human mutation* **15**: 7–12.

Denver DR, Denver DR, Morris K, Lynch M, Vassilieva LL, Thomas WK, 2000. High direct estimate of the mutation rate in the mitochondrial genome of *Caenorhabditis elegans*. *Science* **28**: 2342–2344.

Denver DR, Morris K, Lynch M, Thomas WK, 2004. High mutation rate and predominance of insertions in the *Caenorhabditis elegans* nuclear genome. *Nature* **430**: 679–682.

Deshpande AM, Newlon CS, 1996. DNA replication fork pause sites dependent on transcription. *Science* **272**: 1030–3.

Drake JW, 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proceedings of the National Academy of Science* **88**: 7160–7164.

Drake JW, Charlesworth B, Charlesworth D, Crow JF, 1998. Rates of Spontaneous Mutation. *Genetics* **148**: 1667–1686.

Drake JW, 1999. The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. *Annals of the New York Academy of Sciences* **18**: 100-107.

Drake JW, 2009. Avoiding dangerous missense : thermophiles display especially low mutation rates. *PLoS Genetics* **5**: 1-11.

Esteller M, Hamilton S, Burger P, Baylin S, 1999. Inactivation of the DNA repair gene O6-methylguanine-DNA methyltransferase by promoter hypermethylation is a common event in primary human neoplasia. *Cancer Research* **59**: 793–797.

Filee J, Forterre P, Sen-lin T, Laurent J, 2002. Evolution of DNA polymerase families : evidences for multiple gene exchange between cellular and viral proteins. *Journal of Molecular Evolution* **54**: 763–773.

Foster PL, 2006. Methods for Determining Spontaneous Mutation Rates. *Methods in Enzymology* **409**: 1–16.

Fox A, Tuch B, Chuang JH, 2008. Measuring the prevalence of regional mutation rates: an analysis of silent substitutions in mammals, fungi, and insects. *BMC Evolutionary Biology* **13**: 1–13.

Francino M, Ochman H, 1997. Strand asymmetries in DNA evolution. *Trends in Genetics* **13**: 240–246.

Frank A, Lobry JR, 1999. Asymmetric substitution patterns: a review of possible underlying mutational or selective mechanisms. *Gene* **238**: 65–77.

Friedberg EC, Wagner R, Radman M, 2002. Specialized DNA polymerases, cellular survival, and the genesis of mutations. *Science* **296**: 1627–30.

Frosina G, Fortini P, Rossi O, Carrozzino F, Raspaglio G, Cox LS, Lane DP, 1996. Two pathways for base excision repair in mammalian cells. *The Journal of biological chemistry*, *271*(16), 9573–8.

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

Gebow D, Miselis N, Liber HL, Gebow DAN, Miselis N, 2000. Homologous and nonhomologous recombination resulting indeletion : effects of p53 status , microhomology , and repetitive DNA length and orientation. *Molecular and Cellular Biology* **20**: 4028–4035.

Gerrish PJ, García-Lerma JG, 2003. Mutation rate and the efficacy of antimicrobial drug treatment. *Mutation rate* **3**: 28–32.

Giraud A, Matic I, Tenaillon O, Clara A, Radman M, Fons M, 2001. Costs and benefits of high mutation rates : adaptive evolution of bacteria in the mouse gut. *Science* **291**: 2606–2608.

Gong C, Martins A, Bongiorno P, Glickman M, Shuman S, 2004. Biochemical and genetic analysis of the four DNA ligases of mycobacteria. *The Journal of Biological Chemistry* **279**: 20594–606.

Gonzalgo ML, Jones PA, 1997. Mutagenic and epigenetic effects of DNA methylation. *Mutation Research* **386**: 107–18.

Goodman MF, 2002. Error-prone repair DNA polymerases in prokaryotes and eukaryotes. *Annual Reviews in Biochemistry* **71**: 17–50.

Grogan DW, Carver GT, Drake JW, 2001. Genetic fidelity under harsh conditions : Analysis of spontaneous mutation in the thermoacidophilic archaeon *Sulfolobus acidocaldarius*, *Proceedings of the National Academy of Science* **98**: 7928–7933.

Hazra T, Hill J, Izumi T, 2001. Multiple DNA glycosylases for repair of 8-oxoguanine and their potential in Vivo functions. *Progress in Nucleic Acid Research and Molecular Biology* **68**: 193–205.

Hefferin ML, Tomkinson AE, 2005. Mechanism of DNA double-strand break repair by non-homologous end joining. *DNA Repair* **4**: 639–648.

Hess M, Schwitter U, Petretta M, 1997. Bipartite substrate discrimination by human nucleotide excision repair. *Proceedings of the National Academy of Science* **94**: 6664–6669.

Howard-Flanders P, Boyce R, 1966. Three loci in *Escherichia coli* K-12 that control the excision of pyrimidine dimers and certain other mutagen products from DNA. *Genetics* **53**: 1119–1136.

Hsieh P, 2001. Molecular mechanisms of DNA mismatch repair. *Mutation Research* **486**: 71–87.

Jackson SP, 2002. Sensing and repairing DNA double-strand breaks. *Carcinogenesis* **23**: 687–696.

Jacobs KL, Grogan DW, 1997. Rates of Spontaneous Mutation in an Archaeon from Geothermal Environments, *Journal of Bacteriology* **179**: 3298–3303.

Johnson RC, Bruist MF, 1989. Intermediates in Hin-mediated DNA inversion: a role for Fis and the recombinational enhancer in the strand exchange reaction. *The EMBO journal* **8**: 1581–90.

Johnson RE, Prakash S, Prakash L, 1994. Yeast DNA repair protein RAD5 that promotes instability of simple repetitive sequences is a DNA-dependent ATPase. *The Journal of Biological Chemistry* **269**: 28259–28262.

Klungland A, Lindahl T, 1997. Second pathway for completion of human DNA base excision-repair: reconstitution with purified proteins and requirement for DNase IV (FEN1). *The EMBO journal* **16**: 3341–8.

Kolodner RD, Marsischky GT, 1999. Eukaryotic DNA mismatch repair. *Current Opinion in Genetics & Development* **9**: 89–96.

Kowalczykowski SC, 2000. Initiation of genetic recombination and recombination-dependent replication. *Trends in Biochemical Sciences* **25**: 156–65.

Kunkel TA, Burgers PM, 2008. Dividing the workload at a eukaryotic replication fork. *Trends in Cell Biology* **18**: 521–527.

Lehmann AR, 2006. Clubbing together on clamps: The key to translesion synthesis. *DNA Repair*, **5**: 404–407.

Li G, 2008. Mechanisms and functions of DNA mismatch repair. *Cell Research* **18**: 85–98.

Lichtenauer-Kaligis EGR, van der Velde- van Dijke I, den Duik H, van de Putte P, Giphart Gasslet M, Tasseron-da JG, 1993. Genomic position influences spontaneous mutagenesis of an integrated retroviral vector containing the hprt cDNA as target for mutagenesis. *Human Molecular Genetics* **2**: 173-182.

Lieber M, 2010. The mechanism of double-strand break repair by the nonhomologous DNA end joining pathway. *Annual Reviews in Biochemistry* **79**: 181–211.

Liu Y, Masson J, Shah R, O'Regan P, 2004. RAD51C is required for Holliday junction processing in mammalian cells. *Science* **303**: 243–246.

Longman-Jacobsen N, Williamson JF, Dawkins RL, Gaudieri S, 2003. In polymorphic genomic regions indels cluster with nucleotide polymorphism : quantum genomics. *Gene* **312**: 257–261.

Luria S, Delbruck M, 1943. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**: 491–511.

Lutsenko E, Bhagwat AS, 1999. Principal causes of hot spots for cytosine to thymine mutations at sites of cytosine methylation in growing cells. A model, its experimental support and implications. *Mutation Research* **437**: 11–20.

Lynch M, Lynch M, Koskella B, Schaack S, 2006. Mutation pressure and the evolution of organelle genomic architecture. *Science* **311**: 1727–1730.

Maki H, 2002. Origins of Spontaneous Mutations : Specificity and Directionality of Base-Substitution, Frameshift, and Sequence-Substitution Mutageneses. *Annual Reviews in Genetics* **36**: 279–303.

Matsumoto Y, Kim K, Bogenhagen, DF, 1994. Proliferating cell nuclear antigen-dependent abasic site repair in *Xenopus laevis* oocytes: an alternative pathway of base excision DNA repair. *Molecular and Cellular Biology* **14**: 6187–6197.

Matsumoto Y, 1995. Excision of deoxyribose phosphate residues by DNA polymerase beta during DNA repair. *Science* **269**: 699–702.

Mckenzie GJ, Rosenberg SM, 2001. Adaptive mutations , mutator DNA polymerases and genetic change strategies of pathogens. *Current Opinion in Microbiology* **4:** 586–594.

Memisoglu A, Samson L, 2000. Base excision repair in yeast and mammals. *Mutation Research* **451**: 39–51.

Moser J, Kool H, Giakzidis I, Caldecott K, Mullenders LHF, Fousteri MI, 2007. Sealing of chromosomal DNA nicks during nucleotide excision repair requires XRCC1 and DNA ligase III alpha in a cell-cycle-specific manner. *Molecular Cell* **27**: 311–23.

Mu D, Park C, Matsunaga T, Hsu DS, Reardon JT, Sancar A, 1995 Reconstruction of human DNA repair excision nuclease in a highly defined system. *Journal of Biological Chemistry* **270**: 2415-2418.

Navarro Llorens JM, Tormo A, Martinez-Garcia E, 2010. Stationary phase in gram-negative bacteria. *FEMS Microbiology Reviews* **34**: 476–495.

Novak B, Sible J, Tyson J, 2002. Checkpoints in the Cell Cycle. *Encyclopedia of Life Sciences:* 1–8.

Nyberg KA, Michelson RJ, Putnam CW, Weinert TA, 2002. Toward Maintaining the Genome : DNA Damage and Replication Checkpoints. *Annual Reviews in Genetics* **36**: 617–656.

Ohno M, Miura T, Furuichi M, Tominaga Y, Tsuchimoto D, Sakumi K, Nakabeppu Y, 2006. A genome-wide distribution of 8-oxoguanine correlates with the preferred regions for recombination and single nucleotide polymorphism in the human genome. *Genome Research* **16**: 567–75.

Palmer JD, Herbon LA, 1988. Plant mitochondrial DNA evolves rapidly in structure , but slowly in sequence. *Journal of Molecular Evolution* **28**: 87–97.

Paull TT, Gellert M, 2000. A mechanistic basis for Mre11-directed DNA joining at microhomologies. *Proceedings of the National Academy of Sciences* **97**: 6409–6414.

Petrov DA, 2002. DNA loss and evolution of genome size in *Drosophila*. *Genetics* **115**: 81–91.

Prakash S, Prakash L, 2002. Translesion DNA synthesis in eukaryotes : A one- or two-polymerase affair. *Genes and Development* **16**: 1872–1883.

Prasad R, Lavrik OI, Kim SJ, Kedar P, Yang XP, Vande Berg BJ, Wilson SH, 2001. DNA polymerase beta -mediated long patch base excision repair. Poly(ADP-ribose)polymerase-1 stimulates strand displacement DNA synthesis. *The Journal of Biological Chemistry* **276**: 32411–32414.

Pray L, 2008. DNA replication and causes of mutation. *Nature Education* **1**: 1-3.

Radman M, 1998. DNA replication: One strand may be more equal. *Proceedings of the National Academy of Science* **95**: 9718–9719.

Rocha EPC, 2004. The replication-related organization of bacterial genomes. *Microbiology* **150**: 1609–1627.

Rogozin IB, Babenko VN, Milanesi L, Pavlov YI, 2003. Computational analysis of mutation spectra. *Briefings in Bioinformatics* **4**: 210–227.

Rogozin IB, Malyarchuk BA, Pavlov YI, Milanesi L, 2005. From context-dependence of mutations to molecular mechanisms of mutagenesis. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing* 10: 409–420.

Roth J, Kugelberg E, Reams A, 2006. Origin of mutations under selection: the adaptive mutation controversy. *Annual Reviews* **60**: 477–501.

Rydberg B, Lindahl T, 1982. Nonenzymatic methylation of DNA by the intracellular methyl group donor S-adenosyl-L-methionine. *The EMBO Journal* **1**: 211–216.

Schmid MB, Roth JR, 1983. Genetic methods for analysis and manipulation of inversion mutations in bacteria. *Genetics* **105**: 517–537.

Schofield MJ, Hsieh P, 2003. DNA mismatch repair: molecular mechanisms and biological function. *Annual Reviews in Microbiology* **57**: 579–608.

Shiloh Y, 2001. ATM and ATR: networking cellular responses to DNA damage. *Current Opinion in Genetics & Development* **11**: 71–77.

Shinohara A, Ogawa H, Ogawa T, 1992. Rad51 protein involved in repair and recombination in *S. cerevisiae* is a RecA-like protein. *Cell* **69**: 457–470.

Shivji M, Podust V, Huebscher U, 1995. Nucleotide excision repair DNA synthesis by DNA polymerase. epsilon. in the presence of PCNA, RFC, and RPA. *Biochemistry* **34**: 5011–5017.

Shuck SC, Short EA, Turchi JJ, 2008. Eukaryotic nucleotide excision repair : from understand- ing mechanisms to influencing biology. *Cell Research* **18**: 64–72.

Singer MJ, Marcotte BA, Selker EU, 1995. DNA methylation associated with repeat-induced point mutation in *Neurospora crassa*. *Moleculae and Cellular Biology* **15**: 5585-5597.

Sinha RP, Häder D, 2002. UV-induced DNA damage and repair : a review. *Photochemical and Photobiological Sciences* **1**: 225–236.

Smith GR, 2001. Homologous recombination near and far from DNA breaks: alternative roles and contrasting views. *Annual Review of Genetics* **35**: 243–274.

Sniegowski PD, Gerrish PJ, Lenski RE, 1997. Evolution of high mutation rates in experimental populations of *E . coli*. *Nature* **387**: 703–705.

Sniegowski P, Gerrish P, Johnson T, 2000. The evolution of mutation rates: separating causes from consequences. *Bioessays* **22**: 1057–1066.

Snyder L, Champness W, 2007 Molecular genetics of bacteria (third edition): Chapter 11: *DNA repair and mutagenesis*, 459-497. ASM Press. Washington DC.

Spencer PS, Barral JM, 2012. Genetic code redundancy and its influence on the encoded polypeptides. *Computational and Structural Biotechnology Journal* **1**: 1-8.

Stankiewicz P, Lupski JR, 2002. Genome architecture, rearrangements and genomic disorders. *Trends in Genetics* **18**: 74–82.

Steitz TA, 1999. DNA polymerases: structural diversity and common mechanisms. *The Journal of Biological Chemistry* **274**: 17395–8.

Sung P, Krejci L, Van Komen S, Sehorn MG, 2003. Rad51 recombinase and recombination mediators. *The Journal of Biological Chemistry* **278**: 42729–42732.

Tian D, Wang Q, Zhang P, Araki H, Yang S, Kreitman M, Nagylaki T, 2008. Single-nucleotide mutation rate increases close to insertions/deletions in eukaryotes. *Nature* **455**: 105-110.

Van Houten B, 1990. Nucleotide excision repair in *Escherichia coli*. *Microbiological Reviews* **54**: 18–51.

Vasyunina EA, Rogozin IB, Sinitsina OI, Plaksina AS, Rotskaya U N, 2004. Theoretical and experimental study of mutations by 8-oxoGuanine. *Computational Structural and Functional Genomics*: 200–203.

Watson JD, Crick F, 1953. The molecular structure of nucleic acids: a structure for deoxyribose nucleic acid. *Nature* **17**: 737-738.

West SC, 2003. Molecular views of recombination proteins and their control. *Nature Reviews. Molecular Cell Biology* **4**: 435–445.

Whitman WB, Coleman DC, Wiebe WJ, 1998. Perspective Prokaryotes : The unseen majority, *Proceedings of the National Academy of Science* **95**: 6578–6583.

Wyman C, Ristic D, Kanaar R, 2004. Homologous recombination-mediated double-strand break repair. *DNA Repair* **3**: 827–833.

Yudelevich A, Ginsberg B, Hurwitz J, 1968. Discontinuous Synthesis of DNA during Replication. *Proceedings of the National Academy of Science* **61**: 1129–1136.

Zeyl C, DeVisser J, 2001. Estimates of the rate and distribution of fitness effects of spontaneous mutation in *Saccharomyces cerevisiae*. *Genetics* **157**: 53–61.

Zhuanga WZ, Sugimotoa C, Kubota S, Onoe S, Onuma M, 1995. Antigenic alteration in major piroplasm surface proteins of *Theileria sergenti* during infection. *Veterinary Parasitology*, **60**, 191–198.

Table 1. Common terms used in mutation rate studies.

Term	Definition
M	Number of mutations per culture
μ	Mutation rate (probability of mutation per cell per division or generation)
N	Number of cells
N_0	Cell number in inoculums
N_t	Final cell number in culture
R	Observed number of mutants
F	Mutant fraction (frequency)
V	Volume of culture
C	Number of cultures in experiment
P_o	Proportion of cultures without mutants
Z	Dilution factor/ fraction of culture plated

Table 2. Mutation rate in well studied DNA based micro-organisms.

Organism	μ_b (mutations/bp/genome replication)	μ_g (mutations/genome/genome replication)	**Reference**
Bacteriophage M13	7.5×10^{-7}	0.0048	Drake 1991
Bacteriophage λ	6.6×10^{-8}	0.0032	Drake 1991
Bacteriophage T_2	2.7×10^{-8}	0.0043	Drake 1991
Bacteriophage T_4	2.8×10^{-8}	0.0047	Drake 1991
Escherichia coli	7.9×10^{-10}	0.0037	Drake 1991
Saccharomyces cerevisiae	2.9×10^{-10}	0.0037	Drake 1991
Schizosaccharomyces pombe	3.2×10^{-10}	0.0044	Drake 2009
Sulfobolus acidocaldarius	3.4×10^{-10}	0.0018	Jacobs & Grogan 1997
Neurospora crassa	6.6×10^{-11}	0.0028	Auerbach 1956
Aspergillus nidulans	2.7×10^{-13}	0.0000082	Baracho & Baracho 2003
Aspergillus fumigates	2×10^{-15}	0.0000000617	Alvarez-Perez *et al.* 2010

Figure 1. Summary of the processes occurring in spontaneous mutagenesis.

Depicted at the beginning is the starting point of how DNA can be spontaneously damaged. This is followed by recognition and repair mechanisms. Finally if the DNA is not repaired prior to the next round of replication, then a mutation occurs.

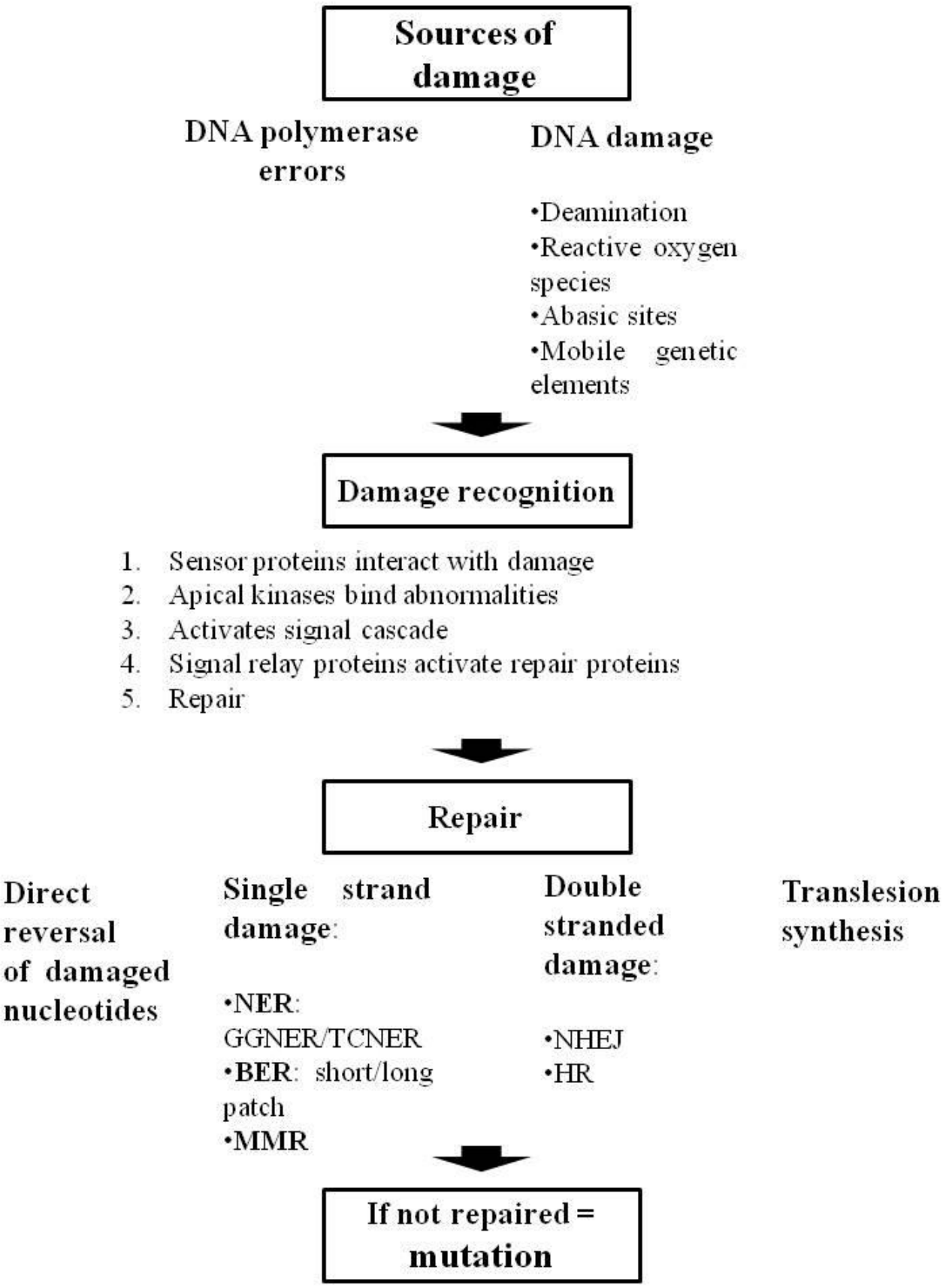


Figure 2. Diagrammatic illustration of the different pathways contributing to the repair of only one strand of DNA damage.

Base excision repair (BER) can repair both single nucleotides (short patch) and longer stretches of DNA consisting of 2-10 nucleotides (long patch repair). The processes described in steps 1 to 4 are used by both prokaryotes and eukaryotes, where the eukaryotic homolog of FEN1 is represented in red. The nucleotide excision repair (NER) system consists of four steps in both prokaryotes and eukaryotes. When a homolog is present in eukaryotes, its protein name is represented in red. Should the same protein is used in both prokaryotes and eukaryotes, it is depicted in black. The four steps of mismatch repair (MMR) shown above are present in both prokaryotes and eukaryotes. The eukaryotic homologs are represented in red, and when the same protein is used in both prokaryotes and eukaryotes, these are shown in black.

Single Stranded Breaks Repair

BER

Short patch repair

1. DNA glycosylase recognition and cleavage
2. AP endonuclease 1 processes cleaved site
3. DNA pol β fills gap
4. DNA ligase seals nick

Long patch repair

1. DNA glycosylase recognition and cleavage
2. AP endonuclease 1 processes cleaved site
3. DNA pol β adds flap of nucleotides
4. Flap removed by FEN1/ RAD27

NER

1.Recognition:
UvrA-B/XPC-RAD23-RPA-TEIIFH

2.Remove damage:
UvrB-C-/ XPG-ERCC

3.Gap filled by:
DNA pol I/ RCF-PCNA-DNA pol δ

4.Nick sealed by:
DNA ligaseI

MMR

1.Recognition:
MutS/Msh1-Msh6 and Msh2-Msh3

2.Remove damage:
MutH/ Mlh1-Pms1

3.Gap filled by:
DNA pol III

4.Nick sealed by:
DNA ligase

Figure 3. An illustration of repair mechanisms acting on damaged DNA which causes double stranded breaks.

Homologous recombination (HR) acts to repair DNA damaged strands in sequences with high homology. Non-homologous end joining (NHEJ) does not need another DNA sequence with high homology to facilitate repair of the damaged DNA. The three steps of HR are present in both prokaryotes and eukaryotes. The eukaryotic homologs are represented in red. The process of NHEJ differs between prokaryotes and eukaryotes as is depicted.

Double Stranded Breaks Repair

HR

1. **Pre-synapsis**
   RecBCD/MreII-Rad50-NBSI

2. **Synapsis**
   RecA/Rad51-Rad54

3. **Post synapsis**
   RuvAB and RuvC/ Rad51C

NHEJ

Prokaryotes:

1. **Recognition**
   MtuKu

2. **End processing/gap filling/sealing nick**
   3 ligases

Eukaryotes:

1. **Recognition**
   Hdf1-Hdf2 and Dnl4-Lif1

2. **End Processing**
   XRM, Xrs2 and Lif1

3. **Gap filling**
   DNA pol IV

4. **Sealing the nick**
   DNA ligase

Chapter 2

Unique mutational motifs within *Fusarium circinatum* core and NRPS genes

Abstract

Mutagenesis is an important driving force in all evolutionary processes. Different pathways of mutagenesis and DNA repair are known to target specific sequences, leaving notable and evolutionary stable motifs that are not randomly distributed across genomes. The presence of these motifs, however, has only been examined in the genomes of a few model eukaryote species. The aim of this study was to identify and characterize the mutational motifs in the genome of the filamentous fungus *Fusarium circinatum*. The study had two objectives, first to determine whether the exons, introns and intergenic regions of the core housekeeping and the non-ribosomal peptide synthetase (NRPS) genes harbour evolutionarily stable motifs that were identified in previous studies, and secondly to determine if the 5′ and 3′ flanking bases influence the occurrence of specific point mutations. Chi squared analyses were employed to determine if single nucleotide polymorphisms (SNPs) identified in the core and NRPS genes of *F. circinatum* are associated with specific flanking bases to form new mutational motifs. Results showed that none of the known evolutionarily stable motifs are present in the core genes suggesting that these motifs do not contribute to the diversity of the genes. However, a motif that is characteristic of the action of the translesion synthesis polymerase is present in the NRPS genes. Furthermore, flanking bases were found to exert a significant influence on the substitutions flanking certain SNPs in the core and NRPS genes. These new and previously unidentified mutational motifs might indicate that the pathways causing the stable motifs may function differently in *F. circinatum* to those observed in model organisms or that these are the result of a previously unrecognised mutagenic pathway.

Introduction

Evolutionary processes are driven by mutation (Baer *et al*. 2007), making mutagenesis important for our understanding of how fungi diversify and species evolve. Spontaneous mutations arise through a variety of intracellular processes that primarily involve replication errors associated with DNA lesions. These are caused by tautomerisation of DNA bases, spontaneous deamination of methylated cytosines, spontaneous depurination of nucleotides and intracellular mutagenic metabolites such as reactive oxygen species (Rogozin *et al*. 2003; Rogozin *et al*. 2005). Even though the causes of mutation are well studied, the importance of different mutagenic pathways in the evolution of different gene regions is less well understood.

The advent of next generation sequencing has revolutionised the study of mutation. Previously, data on mutagenesis and mutation spectra was limited by standard molecular mutation detection techniques, such as restriction fragment length polymorphisms (RFLPs) (Botstein *et al*. 1980; Dryja *et al*. 1991) and amplified fragment length polymorphisms (AFLPs) (Vos *et al*. 1995; Miyashita *et al*. 1999). For example, studies on mammals and *Drosophila* revealed that spontaneous mutations associated with single nucleotide substitution affect only a small proportion of a genome and was estimated to arise at an average rate of 10^{-9} substitution mutations per replication (Petrov & Hartl 1999). However, analysis with all of these traditional methods is limited to only certain "scorable" regions of the genome (i.e., those regions that are represented by RFLP or AFLP fragments). The application of next generation sequencing technologies thus allow for in-depth and genome-wide analyses of mutation (Smith *et al*. 2008), because information about mutation can now be assessed directly through the analysis of single nucleotide polymorphisms (SNPs).

Substitution mutations can be classified as either transitions (i.e., exchange of a purine for a purine or a pyrimidine for a pyrimidine) or transversions (i.e., exchange of a purine for a pyrimidine or vice versa) (Maki 2002; Rogozin *et al*. 2005). Transitions generally occur at a higher frequency than transversions due to their capacity to conserve the biochemical structure of the nucleotides and the chemical properties of base pairing (Wakeley 1996). Through this conservation, a bias towards A-G and C-T substitutions is generated. This bias generally varies across the genome and among specific genes, and is dependent on the selection pressure exerted on the region, the GC content and codon usage bias (Rosenberg *et al*. 2003; Wakeley 1996). In gene regions, transition mutations are less likely to cause amino

acid substitutions (due to the "wobble" or non-Watson-Crick pairing of bases during translation) and ultimately act to conserve the properties of the protein, while transversions often cause non-synonymous substitutions (Rosenberg *et al*. 2003; Wakely 1996).

Mutability differs along and among nucleotide sequences, with some sequences being more prone to spontaneous mutation than others, which leads to the creation of so-called "mutable sequence motifs" or mutational motifs (Rogozin *et al*. 2003; Rogozin & Pavlov 2003; Rogozin *et al*. 2005). The appearance of these motifs or hotspots is usually associated with specific repair pathways and/or exposure to certain endogenous mutagens (Rogozin & Pavlov 2003). The target motifs for many endogenous mutagens, their repair pathways and the sequence motifs resulting from the repair have been identified (Rogozin & Pavlov 2003). Because of the wide conservation of their action (Rogozin *et al*. 2005), these mutagens and repair pathways can lead to the emergence of evolutionarily stable sequence motifs (Rogozin *et al*. 2005). These may vary in certain lineages and species, because of intrinsic differences in the efficiency of the DNA repair pathways, as well as the occurrence of unique mutagens and repair pathways in specific taxa (Maki 2002; Rogozin *et al*. 2003; Rogozin *et al*. 2005). Nevertheless, very little is known about the mutagens and repair pathways associated with specific species as most previous research on mutagenesis and DNA repair have focussed on model bacteria (Lenhart *et al*. 2012) and eukaryotes such as *Neurospora* (Inoue 2011) and mammals (Nouspikel 2009).

Knowledge regarding mutation hotspots is fundamental to our understanding of mutation and the DNA repair pathways that act to counter their appearance. The discovery of mutational motifs is based on the identification of SNPs that are dependent on the surrounding nucleotide sequence (sequence context) where the 5′ or 3′ neighbours or both, affect the SNP. Previous studies have indicated that this sequence context effect is most easily studied over short distances thus resulting in di- or trinucleotide motifs (Rogozin *et al*. 2003). A common example is that of Sn-1 alkylating agents where significantly more mutations are associated with the RG motif (where G is the mutable base and R represents A or G) than the YG motif (where Y represents C or T) (Rogozin & Pavlov 2003). There are, however, also examples where the sequence context of a mutable base occurs over much longer distances, e.g., the mutation rate of a specific base in the *supF* suppressor transfer RNA gene is affected by a base that occurs 80 bases away (Canella & Seidman 2000). The majority of research regarding the sequence context effects of mutations has been done in bacteria and humans (Krawczak *et al*. 1998; Rogozin *et al*. 1992; Shapiro *et al*. 2012).

Almost all of these previous studies concluded that the processes determining context dependent mutation and mutation rate are complex, highlighting that large gaps still exist in our understanding of mutational footprints and processes in non-model organisms and organelles.

In this study we considered mutation in the filamentous fungus *Fusarium circinatum* (Nirenberg & O'Donnell emend. Britz, Coutinho, Wingfield & Marasas), which is the causal agent of the debilitating disease, pitch canker, on various *Pinus* species (Wingfield *et al*. 2008). Like most members of the *Gibberella fujikuroi* species complex, this fungus is an economically important pathogen (Kvas *et al*. 2009). To account for the possible effects that different evolutionary forces might have on mutation, we focussed on two set of genes – the core genes and the nonribosomal peptide synthetase (NRPS, see below) genes. The core genes include those common to eukaryotes and that encode most of the cellular housekeeping functions (Parra *et al*. 2007). The NRPS genes, on the other hand, produce secondary products that can impart one or more competitive advantage to the cell, but that are non-essential for housekeeping functions and survival (Bushley & Turgeon 2010). The NRPS genes are thus thought to evolve primarily through diversifying selection, while those in the core set probably experience purifying selection, suggesting that the mutational effects will be markedly different for the two types of genes (Amoutzias *et al*. 2008).

Unlike the normal translational machinery, NRPSs can produce peptide or protein products without using mRNA and ribosomes (Bushley & Turgeon 2010; Stack *et al*. 2007). The genes encoding these proteins generally consist of different modules that contain different domains, with a minimum module consist of an adenylation domain, a thiolation or peptidyl carrier protein domain and a condensation domain (Bushley & Turgeon 2010; Stack *et al*. 2007). The steps involved in NRPS proteins synthesis are as follows: i) the substrate (e.g., a single amino acid) enters through the adenylation domain where it is recognized and activated using ATP, ii) attachment of the activated substrate to a phosphopantetheine cofactor in the peptidyl carrier protein, and iii) formation of peptide bonds between adjacent substrates within the condensation domain (Bushley & Turgeon 2010; Stack *et al*. 2007). The presence of additional or accessory domains can allow for the modification (e.g., epimerization domain which converts L amino acids to their D conformation) of the peptide product before release (Bushley & Turgeon 2010). The most commonly recognised products of NRPSs are mycotoxins such as the AM toxin of *Alternaria alternata* (Johnson *et al*. 2000), penicillin in *Aspergillus nidulans* (Stack *et al*. 2007) and fusarin in *F. verticillioides* and *F.*

graminearum (Stack *et al*. 2007). Other NRPS products include proteins involved in fungal reproduction, pathogen development, cell surface properties and stress control (Bushley & Turgeon 2010; Stack *et al*. 2007).

The overall goal of this study was to improve our understanding of spontaneous mutagenic processes in *F. circinatum* by investigating the distribution and frequency of SNPs occurring in specific sequence contexts. To achieve this goal, our specific aims were four-fold. Firstly, we characterised SNPs across the genome of *F. circinatum* by considering the distribution and density of different substitution types. Secondly, we characterized the conserved motifs (i.e., potentially representing mutable motifs or remnants motifs that were generated by DNA repair pathways) by testing for significant associations between specific substitutions and their respective 5′ and 3′ neighbouring nucleotides. Thirdly, the conserved mutational motifs occurring in the core and NRPS genes were compared with one another and with those that are known from the model organism to improve our understanding of how these genes evolve and to determine whether the known mutagens and repair pathways leave the same recognizable motifs in *F. circinatum*. The results of this study will thus not only shed light on how mutation potentially drive diversification of NRPS and core genes in *F. circinatum*, but also on how the mutagenic processes and repair systems result in genetic diversity in this fungus.

Materials and Methods

Genome sequences

Whole genome sequence data for two *F. circinatum* isolates, FSP34 and KS17, was used in this study. Isolate FSP34 originated from diseased *P. radiata* trees in California (United States of America), while KS17 was isolated from a diseased *P. radiata* seedling in a commercial pine seedling nursery in the Western Cape Province of South Africa. The genome sequence for strain FSP34 was available from a previous study (Wingfield *et al*. 2012).

The KS 17 isolate was prepared and DNA was extracted by Mr. QC Santana. The following protocol was followed: single spore cultures were grown on half strength potato dextrose agar (Merck Biolab, Wadeville, Gauteng) for one week at 25°C. Mycelial plugs from the leading

edges of the cultures were then used to inoculate half strength potato dextrose broth (Merck Biolab, Wadeville, Gauteng). These were then incubated at 25°C for five days in an orbital shaker (100rpm). Mycelia were harvested via filtration using Whatman filter paper and freeze dried. The freeze dried material was then ground to a powder in the presence of liquid nitrogen and stored at -80°C.

Genomic DNA was extracted as previously described by Steenkamp *et al* 1999. DNA quality and quantity was determined using a Nanodrop spectrophotometer ND-1000 (Nanodrop Technologies, Rockland USA) and 10µg of DNA was pooled together. To check for possible contamination, the Internal Transcribed Spacer (ITS) and translation elongation factor 1 α (TEF-1α) was amplified as previously described (O'Donnell *et al*. 2000) and compared to both Genbank and the Fusarium ID databases (http://isolate.fusariumdb.org/index.php).

Genome sequencing was performed at Seqomics (Domaszek, Hungary) using the Applied Biosystems SOLiD v4 platform. Mate pair libraries with an insert size of 1000bp were prepared by the company. The library was sequenced using a full slide on the SOLiD v4 platform with a read length of 50bp on either end of the mate pair.

Identifying core and NRPS gene sets

The set of core genes for *F. circinatum* was available from a previous study (Phasha *et al*. 2012). Briefly, the construction of this dataset utilized the genes encoding housekeeping functions in *Saccharomyces cerevisiae* and presumably all other eukaryotes (Holt & Yandell 2011), which were retrieved from the Core Eukaryotic Genes Mapping Approach (Σ-cegma; Parra *et al*. 2007) (http://korflab.ucdavis.edu/Datasets/cegma/index.html) database and used in BLASTp searches against the *Fusarium* Comparative Database (Broad Institute). This set of core gene sequences identified in *F. verticillioides*, *F. graminearum* and *F. oxysporum* were then used to identify the nucleotide sequences of the core genes in *F. circinatum* through BLASTn searches. The NRPS and their accessory genes were identified in the FSP34 genome using the online tool AntiSMASH (www.antismash.secondarymetabolites.org), which uses profile hidden Markov models to identify gene clusters encoding secondary metabolites (Medema *et al*. 2011).

SNP datasets

In order to facilitate SNP identification, the sequence reads of the isolate KS17 were mapped to the genome of FSP34 using CLCBio Genomics workbench, version 4.9 (CLC bio A/S). SNPs were identified in regions with a coverage of more than 100 reads of which 65% or more supported the variant. The SNP distribution was studied at two different levels – across all regions of the genome and across putative genes. For the SNPs in the core and NRPS gene sets, their specific location in genes (introns, exons and codon position) and between genes were also determined.

Testing for associations between SNPs and their 5′ and 3′ neighbouring bases

For the core and NRPS gene sets, the different types of substitutions and their neighbouring bases were assessed to determine the possible association between specific SNPs and their neighbouring bases. For this purpose, χ^2 squared analyses were used for number of comparisons. For the analyses involving the NRPS genes, the occurrence of mutational motifs were also investigated in terms of the domains and modules potentially harbouring them.

To test for associations between bases at the 5′ neighbouring position of SNPs within exons, the dataset generated for this analysis was divided into transition and transversion substitutions, after which transitions and transversions were further subdivided into the four types if transitions (ts: C->T, G->A, T->C, A->G) and six types of tranversions (tv: C->A, C->G, A->C, A->T, G->C, G->T, T->A and T->G). The occurrence of each of the four possible nucleotides at the 5′ position was recorded as the observed values for the χ^2 analysis i.e. *M*(CX), *M*(GX), *M*(TX) and *M*(AX) where X is the SNP and *M* is the total amount of mutations that occur at nucleotide combinations. The number of times the dinucleotide (CX, GX, TX and AX) occurs within the region, were recorded as *N*(CX), *N*(GX), *N*(TX) and *N*(AX). The equation below was used to determine the expected amount of every substitution to occur within each of the dinucleotides i.e. *E*(CX), *E*(GX), *E*(TX) and *E*(AX) where I represents A/C/G/T (Rogozin & Pavlov 2003).

$$E(IX) = \frac{N(IX)\times(M(AX)+ M(CX)+ M(GX)+ M(TX))}{N(AX)+N(CX)+N(GX)+N(TX)}$$

A standard χ^2 test was used to test Ho: *M*(AX): *M*(CX): *M*(GX): *M*(TX) = 1:1:1:1. The *P*-value was determined with three degrees of freedom (df). The test was repeated for XA, XC, XG and XT (3′ neighbouring bases) to test Ho: *M*(XA): *M*(XC): *M*(XG): *M*(XT) in the exons of the core genes. The dinucleotide combinations that occur more frequently than the rest at

significance ($P < 0.05$) were recorded. A similar approach was done to examine the bases at the 3′ neighbouring position of SNPs within exons. NRPS genes, the dinucleotides occurring in different domains and modules were also compared between modules.

For the analysis of trinucleotide mutational motifs, the 5′ and 3′ data was combined and χ^2 analyses were conducted to test Ho: AXA: AXC: AXG: AXT: CXA: CXC: CXG: CXT: GXA: GXC: GXG: GXT: TXA: TXC: TXG: TXT = 1:1:1:1:1:1:1:1:1:1:1:1:1:1:1:1. The observed values were quantified as *M*(AXA), *M*(AXC), *M*(AXG), *M*(AXT), *M*(CXA), *M*(CXC), *M*(CXG), *M*(CXT), *M*(GXA), *M*(GXC), *M*(GXG), *M*(GXT), *M*(TXA), *M*(TXC), *M*(TXG) and *M*(TXT). The *N*-values and subsequent *E*-values were determined in the same manner as for the dinucleotides. A df of 15 was used to determine the *P*-value. The trinucleotide combinations that occur more frequently than the rest at a significance level above ($P<0.05$) were recorded.

Known mutational motifs

The significant associations were further examined in order to determine whether they correspond to any of the mutational motifs of five known endogenous mutagens, namely Sn1-type alkylating agents, translesion synthesis polymerase, 8-oxoGuanine, cytidine deaminase and repeat induced point (RIP) mutations are evident (Rogozin *et al*. 2003; Rogozin *et al*. 2005). Sn1- type alkylating agents were found to target the dinucleotide RG (R=A/G), commonly causing G->A transitions (Rogozin *et al*. 2003). TLS polymerase targets WA motifs (W=A/T) (Rogozin *et al*. 2003). A common product of reactive oxygen species is 8-oxoGuanine which targets AA motifs, causing A->C transversions (Vasyunina *et al*. 2004). Cytidine deaminase targets GG dinucleotides (Rogozin *et al*. 2003) and lastly RIP causes C->T transitions at CA targets in repeat sequences larger than 400bp (Clutterbuck 2011).

Results

SNPs datasets

Comparisons of the genomes of *F. circinatum* isolates KS17 (493 979 732 mate pair reads) and FSP34 revealed a total of 132 386 SNPs with an average density of 3.0 SNPs/kb (1 SNP every 325bp) (Table 1). From the 132 386 SNPs, most were transition substitutions (94 524 SNPs), with a ts/tv ratio that is significantly deviated from the expected 2ts:tv ratio

(χ^2=86327.88, df=1, *P*<0.0001). When considering the difference in occurrence between the different types of transitions and transversions, the χ^2 squared tests again showed significant deviation from the expected 1:1:1:1 and 1:1:1:1:1:1:1:1 ratios, indicating biases toward specific transition (χ^2=58.34, df=3, *P*<0.0001) and transversion substitutions (χ^2=265.87, df=7, *P*<0.0001) (Table 1). Of the possible transitions, C->T substitutions are generally favoured. Of the possible transversions, G->C, C->A and T->G substitutions occur at roughly equal frequencies, but at higher frequencies than other transversions (Table 2).

The genome of *F. circinatum* (FSP34) contains 14 975 annotated putative genes, spanning a total of 21 683 768 bp (approximately 5% of the genome) (Wingfield *et al*. 2012). The putative genes were found to contain 47 148 (35.6%) SNPs, with an average density of 2.2 SNPs/kb (1 SNP every 455 bp). The putative gene regions showed a general bias towards transitions (χ^2=32889.05, df=1, *P*<0.0001), as well as C->T transitions (χ^2=53.33, df=3, *P*<0.0001) and C-> A transversions (χ^2=336.33, df=7, *P*<0.0001) (Table 1).

A set of 370 *F. circinatum* core genes were used in this study. These genes spanned 329 contigs and 1.3 Mb. Comparison of the sequences for the two isolates (FSP34 and KS17) revealed 846 SNPs (202 in the exons, 137 in the introns and 507 in the intergenic regions) at an overall density of 0.64 SNPs/kb (i.e., 0.38 SNPs/kb in the exons, 1.65 SNPs/kb in the introns and 0.72 SNPs/kb in the intergenic regions) (Table 1). As found across the genome, the intergenic regions showed a significant transition bias (χ^2=267.166, df=1, *P*<0.0001), and a significant deviation from the expected 1:1 ratio between the different types of transitions (χ^2=32.125, df=3, *P*<0.0001) and transversions (χ^2=20.067, df=7, *P*=0.0054) (Table 1). In the core gene dataset significantly more C->T transitions were detected, whereas C->A transversions occurred at the highest frequency (Table 2). When assessing the exons and introns of the core genes no bias between the different transitions (χ^2=2.64, df=3, *P*=4505; χ^2=1.184, df=3, *P*=0.7569) and transversions (χ^2=12.963, df=7, *P*=0.073; χ^2=7.564, df=7, *P*=0.3726) were found, although a bias towards transitions was apparent (χ^2=148,324, df=2, *P*<0.0001; χ^2=89.96 df=1, *P*<0.0001) (Table 1). This absence of a bias between the different transitions and transversions in the core exons and introns, however, could be ascribed to the small sample size. Mutations within the core genes were observed mostly at the third codon position (75/202 mutations).

Analysis of the genome sequence for FSP34 with the online tool antiSMASH allowed identification of 438 putative secondary metabolite genes. Of these, 84 genes were identified

as putative NRPS and NRPS accessory genes, which were located on 10 contigs that together span 409 588 bp of sequence. Within this set of 84 genes, a total of 1 175 SNPs were identified. Of these, 714 SNPs were located in the NRPS coding regions, 34 were in introns and 403 were located in intergenic regions. The average density of SNPs within this dataset was 2.87 SNPs/kb, with an average of 2.85 for the exons and intergenic regions and 3.34 for the introns (Table 1). In addition, we found a significant deviation from the expected 2ts:tv ratio (χ^2=771.889, df=1, P<0.0001), although there was no apparent bias between the different types of transitions (χ^2=1.382, df=3, P=0.7098) (Table 1) (note, this is probably due to the small sample size, because artificial increase of the proportions of substitutions observed in generated significantly different ratios). With regards to the transversion substitutions, C->A transversions occurred most frequently whereas C->G transversions occurred least frequently (χ=26.15, df=7, P=0.0005) (Table 2). SNPs within the NRPS genes were found to mainly affect the first codon position (263/738 mutations).

Associations between SNPs and their 5′ and 3′ neighbouring bases

For both datasets, clear associations between certain substitutions and their neighbouring bases (i.e., 3′ alone, 5′ alone or both), were observed (Tables 3-7). A total of thirteen conserved motifs were identified within the core gene dataset. Of these seven motifs were identified in exons (three dinucleotide and four trinucleotide) (Table 3), two motifs in introns (one dinucleotide and one trinucleotide) (Table 4) and four in the examined intergenic regions (three dinucleotide and one trinucleotide) (Table 5). For the NRPS dataset, a total of 29 conserved motifs were identified. Of these 17 were detected in exons (seven dinucleotide and 10 trinucleotide) (Table 6) and 12 in the intergenic regions (eight dinucleotide and four trinucleotide) (Table 7). Due to the small number of SNPs in the introns of the NRPS genes, statistical analyses for detecting association were not done.

The specific domains encoded by the nine NRPS genes were examined for the occurrence of 29 conserved motifs detected in the NRPS dataset (Tables 6 and 7). All but two of the nine NRPS genes contained the three domains to form a minimal module, while an additional epimerization domain was present in two of the NRPS genes (Table 8). Of the 242 SNPs that occurred within the nine NRPS genes, 54 occurred within the adenylation domain, 31 in the condensation domain and only six in the peptidyl carrier domain (Table 8). However, a 1:1 relationship was detected between the proportion of SNPs and the proportion of the gene covered by domains across all nine of the genes. The percentage SNPs detected within the

identified conserved motifs varied from 14.3% to 40% (Table 8). However, all of the identified conserved motifs, including the motif of the TLS polymerase (see below), occur randomly throughout the nine genes. The TT motif is especially evident within the second adenylation domain of one of the NRPS genes, with mutations occurring at five of the 61 TT dinucleotides within the second adenylation domain. This is in contrast to the first adenylation domain that has no mutations occurring in the possible 62 TT dinucleotides. In addition to the adenylation domain the second condensation domain also seems more prone to mutations in the TT motif than the first condensation domain (Table 9). The peptidyl carrier domains contained only six of the 242 SNPs, all of which are transitions and only one mutation that is present within a conserved motif.

Known mutational motifs

All the significant associations between specific SNPs and their neighbouring bases were compared to five of the known mutational motifs. None of these were evident in the core gene dataset. The only known mutational motif detected in the NRPS dataset was that of the TLS polymerase with its WA (W=A/T) target that was apparent in the TA mutational motifs. This motif was found in the NRPS exons (χ^2=30.145, df=3, P<0.0001) and the intergenic regions associated with the NRPS and accessory genes (χ^2=24.16281, df=3, P<0.0001). A homolog of the *rev7* gene of *Neurospora crassa*, which is expected to contribute to TLS (Murakumo, 2002), was found (FCIRC 14775) when the amino acid sequence of the *N. crassa* gene was used in a tblastp search against a protein database constructed from the genome sequence of the FSP34 *F. circinatum* isolate (Wingfield *et al*. 2012).

Discussion

Between the two *F. circinatum* genomes compared, SNPs were estimated to occur at an average density of 1 SNP/333bp, which is markedly different from what has been demonstrated for other eukaryotes such as humans and yeast. The genomes of two humans contain on average 1 SNP/1000 bp of sequence (Zhao *et al*. 2003), while those of some *S. cerevisiae* strains can harbour as many as one SNP every 37 bp of sequence (Borneman *et al*. 2011). The unusually high SNP density observed in this yeast is thought to be due to the domestication process and the artificial selection associated with it (Borneman *et al*. 2011). In a SNP study of the common cereal pathogen, *F. graminearum*, a density range of 0-17.5

SNPs/kb was estimated (Cuomo *et al*. 2007). This study demonstrated that most of the detected SNPs occurred in discrete genomic regions or were located in telomeric regions. Although the lack of complete genome sequence information did not allow for the estimation of the genome wide SNP density in *F. graminearum*, a similar distribution pattern would suggest that *F. circinatum* in general harbours more SNPs. This is because *F. circinatum* has twelve chromosomes (De Vos *et al*. 2007) as opposed to the four chromosomes of *F. graminearum*. Future studies should thus investigate whether the SNP density distribution in *F. circinatum* follow a similar bias towards telomeric regions.

A bias towards transitions was evident throughout the genome of *F. circinatum*. Previous studies have shown that transition bias is a universal phenomenon that occurs in mitochondrial genomes, as well as the nuclear genomes of most organisms, from bacteriophage to primates (Wakeley 1996). This substitution bias is a consequence of the intrinsic properties of DNA that cause a difference in mutagenesis between the 4 bases (Fu *et al*. 2011; Wakeley 1996). However, Keller *et al*. (2007) showed that the transition bias in grasshopper is due to the high mutability of methylated cytosines, causing the original cytosine to adopt the structure of a thymine leading to C transitions. When the methylated cytosine sites were excluded from their datasets they showed that the bias disappears. In our study, however, methylated cytosines did not contribute significantly to the transition bias observed, because the exclusion of all C transitions still resulted in a deviation from the expected 2Ti:Tv ratio (χ^2=16 4141, df=1, P<0.0001).

The ratios in which the different types of transition and transversion substitutions occurred in *F. circinatum* deviated significantly from what was expected when considering the amount of transition and transversion types. In general, C transitions and G transitions were more common than the rest of the substitutions (Table 2). The high frequency of C transitions could possibly be remnants of RIP in addition to spontaneous deamination of methylated cytosines. RIP was first identified in *Neurospora crassa* (Cambareri *et al*. 1991) and was suggested to be a defence mechanism against transposable elements. This defence is accomplished by introducing C transitions in repeat regions throughout the genome, disabling the movement of transposable elements that threatens normal gene function (Clutterbuck 2011). The high frequency of G transitions could be associated with inter-base double proton transfer, which is a spontaneous chemical process that leads to tautomerisation that occurs more frequently within GC base pairs (Fu *et al*. 2011). The G->C, C->A and T->G transversions dominated throughout the *F. circinatum* genome where they occur at roughly

equal frequencies. Different mutations clearly do not occur at random throughout the genome.

Consistent with earlier studies, a number of SNPs in *F. circinatum* were significantly associated with their neighbouring bases resulting in dinucleotide and trinucleotides motifs that potentially represent mutation hotspots (Rogozin *et al*. 2003; Rogozin *et al*. 2005; Rogozin & Pavlov 2003). The occurrence of these motifs in the introns, exons and intergenic regions of the core genes, as well as, the exons and intergenic regions of the NRPS genes suggest that they might play an important role in determining the base substitution rate across the genome. It would suggest that regions with more mutational motifs will have a higher substitution rate than those with fewer motifs, which potentially have important implications when conducting phylogenetic analyses (Rogozin & Pavlov 2003). It has thus been suggested that data on mutational motifs should be included in phylogenetic analyses, which, even though it threatens to increases the complexity of the underlying substitution models, will improve phylogenetic analyses. Alternatively, substitution models compensating for substitution variation amongst sequences should be used or the mutational motifs can be excluded from phylogenetic analyses (Rogozin & Pavlov 2003).

Different non-overlapping sets of potential mutable motifs were identified in the two sets *F. circinatum* sequences with the NRPS dataset harbouring more than twice as many of these motifs as the core gene dataset. Although the difference in the number of these motifs between the two sets might be due to the differences in the numbers of SNPs detected, the large number of NRPS mutational motifs identified might be linked to the evolutionary forces acting on these genes. At a functional level, the products encoded by these two gene sets are hugely different and it follows that the selection pressures acting on them will allow for the emergence of dissimilar mutational patterns (Amoutzias *et al*. 2008). In addition, the difference in the mutational patterns observed in the two datasets could be ascribed to intrinsic differences in these genes regarding accessibility to mutagens and repair proteins (Reeves & Adair 2005). In fact, DNA repair is known to be influenced by the general transcriptional level, chromatin structure, DNA sequence context, and the type of lesion that was formed (Rogozin *et al*. 2003). Because core genes encode basic cellular functions and are commonly constitutively expressed (Eisenberg & Levanon 2003), they are not only more exposed to mutagens but also more accessible to repair machinery (Russev & Boulikas 1992). Core genes are thus likely to be repaired more efficiently than other tightly packed, less expressed genes or regions in or near heterochromatin. This accessibility may explain why

conserved mutational motifs were not observed in the exons, introns and intergenic regions of the core genes. The idea that the absence in overlap between motifs in the core and NPRS genes might indicate the contribution of novel mutagenic pathways to either group of genes can also not be discounted. Whatever the case may be, these motifs might be crucial for our understanding how genes evolve in *F. circinatum*.

With the exception of TLS polymerase in the NPRS dataset, none of the known mutational motifs were observed in *F. circinatum*. This could illustrate that none of the causes of mutational motifs studied here contribute to diversification of *F. circinatum*. However, another possibility is that, for example, the mutagenic pathways function differently in *F. circinatum* and its close relatives due to e.g. repair mechanisms or enzymes that were not present in the model organism where the motif was identified. The possibility arises that some of the newly identified motifs in this study might represent known mutagenic pathways which leave a different signature in *F. circinatum* due to the repair mechanisms or enzymes present in this fungus. It is thus important to explore mutation in the close relatives of *F. circinatum* to see if similar patterns arise. Furthermore, it will be important to link known mutagenic pathways to our newly discovered mutational motifs.

Our results suggest a role of the TLS polymerase in the diversification of the NRPS and their accessory genes. TLS polymerase is the enzyme responsible for the error prone DNA repair pathway, translesion synthesis (Arana & Kunkel 2010; Friedberg *et al*. 2002). The mutagenicity of TLS polymerases depends on a lesion bypass event, which is influenced by the DNA binding properties of the polymerase (Arana & Kunkel 2010). The presence of a homolog of the TLS polymerase encoding gene further support the hypothesis that the TLS contributes to the diversity of NRPS genes..

An association could not be detected between the occurrence of specific SNPs in mutational motifs and NRPS domains. SNPs occur randomly throughout the domains and between domains regardless of mutational motifs. This is in contrast to what was found in human immunoglobulin genes, where it was discovered that the chromatin structure of different domains influences the occurrence of SNPs in different motifs (Bachl *et al*. 1997). The only difference detected in mutational motifs in the present study was between different modules of the same NRPS encoding gene. The domains of the first module seem to be better conserved than the same domains of the second module. Multimodular NRPSs are thought to occur in majority in Euascomycetes and it is suggested that they have evolved through

duplication of modules followed by selective loss of certain domains of modules (Bushley & Turgeon 2010). This would explain why more mutations are observed within one module than the next. The small amount of transitions occurring within the peptidyl carrier protein domain in this study may solely be due to the fact that the domain is approximately five times smaller than any of the other domains. In addition, the conservation of the long loop of the peptidyl carrier protein is essential for protein phosphatase recognition and subsequent protein release (Weber *et al*. 2000), further explaining the predominance of transitions that generally act to conserve amino acid sequence.

It is evident from the results of this study that the mutability of certain dinucleotide combinations in the NRPS and their accessory gene regions differs. The TA dinucleotide combination was present in five of the 12 identified dinucleotide mutational associations and six out of the 13 identified trinucleotide mutational motifs. These results are congruent with other studies that showed a high mutability of the TA combination in memory B cells of humans (Smith *et al*. 1996). It was suggested that this type of mutation is the result of a variety of mismatches due to alternative biochemical conformations of the bases e.g. $T_{enol}G$, TG wobble and $C_{imino}A$ mismatches, and predictions regarding mutable trinucleotides could be made. For *F. circinatum*, however, no trinucleotides motif occurred dominantly, although two of the trinucleotides motifs found to be highly mutable in the B cells, TAC and GTA, were also apparent in the NRPS and accessory genes of *F. circinatum*. This may suggest that the target sequences of mutation mechanisms between fungi and humans are similar due to the universal intrinsic properties of DNA.

Similar to TAC and GTA motifs, most of the mutable trinucleotide combinations and all of the mutable dinucleotide combinations were found to consist of alternating purines and pyrimidines instead of, for example, a pyrimidine tract. Due to the repetitive nature of the tracts they are hotspots for mutagenesis. These homopolymeric tracts are prone to mutations in the form of insertions and deletions (indels) that form due to misalignment during replication (Ma *et al*. 2012). Although the tracts are prone to indels, it was shown that e.g. a polyadenine tract allows for the formation of additional hydrogen bridges through an increased propeller twist (Li *et al*. 1998), making them less likely to mutate. Furthermore, as was the case in the current study, homopolymeric tracts are not commonly found in coding regions of genes due to their instability (Denver *et al*. 2004).

SNPs were distributed differently between codon positions in the two gene sets examined, again indicating that mutation affects these gene differently. Due to the degeneracy of the genetic code, most mutations usually occur at the third position and results in synonymous codon changes (O'Neill & Ryan 2001; Taylor & Coates 1989), while limited mutations at the first codon position generally ensures conservation of the chemical properties of the amino acid (Taylor & Coates 1989). In this study, it was found that mutations mainly affected the third codon position in the core genes whereas the first position was affected predominantly in the NRPS and accessory genes. Because of the fact that core genes probably undergoes significant levels of purifying selection, the more frequent occurrence of SNPs in the third position was not unexpected as it allows for conservation of the amino acids encoded. The fact that there were more SNPs in the first codon positions of the NRPs genes was consistent with the expectations. Genes encoding secondary metabolites such as toxins have been found to be under positive selection (Stewart *et al*. 2011), thus favouring non-synonymous mutations yet conserving the chemical properties of the toxin protein. Our findings thus suggest that the NRPS gene diversity in *F. circinatum* is increased by the occurrence of mutations in first codon positions without greatly altering the chemical properties of the encoding products.

The purpose of this study was to provide a better understanding of mutagenesis, specifically substitutions, within the core genes and NRPS genes of *F. circinatum*. This promises to expand our current knowledge with regards to the evolutionary forces driving diversification and evolution of the different gene groups under different selection pressures. Our results showed that the 5′ and 3′ neighbouring bases influence the occurrence of a specific substitution in both the core genes and the NRPS and accessory genes. Furthermore, the identified mutational motifs differ between the core genes and NRPS and accessory genes. This can indicate that either different mutagenic pathways act in on the two different gene groups or that DNA repair machinery function differently between the two groups. Lastly, the presence of the TLS mutational motif in the NRPS and accessory genes could point to an important source of diversification of secondary metabolite encoding genes. The molecular mechanisms involved in spontaneous mutagenesis as a process are complex and influenced by a variety of factors including DNA packaging, functioning of the mutagen and DNA repair mechanisms available (Rogozin & Pavlov 2003). Until more studies are done on mutational motifs across the genomes of both closely related species as well as other non related species it will remain unknown as to whether the observed motifs are unique or a

common occurrence. Once this is known the role of each factor contributing to spontaneous mutagenesis in the formation of these motifs can be studied. Future work should aim to include more isolates of a species as to identify hotspots in different genes or groups of genes. Identification of possible mutagens targeting the identified hotspots is also crucial for understanding forces driving diversification of different groups of genes within the genome of *F. circinatum.*

References

Amoutzias GD, Van De Peer Y, Mossialos D, 2008. Evolution and taxonomic distribution of nonribosomal peptide and polyketide synthases. *Future Microbiology* **3**: 361–370.

Arana ME, Kunkel TA, 2010. Mutator phenotypes due to DNA replication infidelity. *Seminars in Cancer Biology* **20**: 304–311.

Bachl J, Steinberg C, Wabl M, 1997. Critical test of hot spot motifs for immunoglobulin hypermutation. *European Journal of Immunology* **27**: 3398–3403.

Baer CF, Miyamoto MM, Denver DR, 2007. Mutation rate variation in multicellular eukaryotes : causes and consequences. *Nature* **8**: 619–632.

Borneman AR, Desany BA, Riches D, Affourtit JP, Forgan AH, Pretorius IS, Egholm M, *et al*, 2011. Whole-genome comparison reveals novel genetic elements that characterize the genome of industrial strains of Saccharomyces cerevisiae. *PLoS Genetics* **7**: 1-10.

Botstein D, White RL, Skolnick M, Davis RW, 1980. Construction of a genetic linkage map in man using restriction fragment length polymorphisms. *The American Journal of Human Genetics* **32**: 314–331.

Bushley KE, Turgeon BG, 2010. Phylogenomics reveals subfamilies of fungal nonribosomal peptide synthetases and their evolutionary relationships. *BMC Evolutionary Biology* **10**: 26–49.

Cambareri EB, Singer MJ, Selker EU, 1991. Recurrence of repeat-induced point mutation (RIP) in *Neurospora crassa*. *Genetics* **127**: 699–710.

Canella KA, Seidman MM, 2000. Mutation spectra in supF: approaches to elucidating sequence context effects. *Mutation Research* **450**: 61–73.

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

Cuomo C, Guldener U, Xu J, Trail F, et al, 2007. The Fusarium graminearum genome reveals a link between localized polymorphism and pathogen specialization. *Science* **317**: 1400–1402.

De Vos L, Myburg A, Wingfield MJ, Desjardins A, Gordon T, Wingfield BD,2007. Complete genetic linkage maps from an interspecific cross between *Fusarium circinatum* and *Fusarium subglutinans*. *Fungal Genetics and Biology* **44**: 701–714.

Denver DR, Morris K, Kewalramani A, Harris KE, Chow A, Estes S, Lynch M, et al, 2004. Abundance , distribution , and mutation rates of homopolymeric nucleotide runs in the genome of *Caenorhabditis elegans*. *Journal of Molecular Evolution* **58**: 584–595.

Dryja TP, Hahn LB, Cowley GS, McGee TL, Berson EL, 1991. Mutation spectrum of the rhodopsin gene among patients with autosomal dominant retinits pigmentosa. *Proceedings of the National Academy of Sciences* **88**: 9370-9374.

Eisenberg E, Levanon EY, 2003. Human housekeeping genes are compact. *Trends in Genetics* **19**: 362–365.

Friedberg EC, Wagner R, Radman M, 2002. Specialized DNA polymerases, cellular survival, and the genesis of mutations. *Science* **296**: 1627–1630.

Fu LY, Wang GZ, Ma BG, Zhang HY, 2011. Exploring the common molecular basis for the universal DNA mutation bias: revival of Löwdin mutation model. *Biochemical and Biophysical Research Communications* **409**: 367–371.

Holt C, Yandell M, 2011. MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC Bioinformatics* **12**: 491–505.

Inoue H, 2011. Exploring the processes of DNA repair and homologous integration in *Neurospora*. *Mutation Research-Reviews in Mutation Research* **728**: 1–11.

Johnson RD, Johnson L, Itoh Y, Kodama M, Otani H, Kohmoto K, 2000. Cloning and characterization of a cyclic peptide synthetase gene from *Alternaria alternata* apple pathotype whose product is involved in AM-toxin synthesis and pathogenicity. *Molecular Plant-Microbe Interactions* **13**: 742–753.

Keller I, Bensasson D, Nichols RA, 2007 Transition-transversion bias is not universal: a counter example from grasshopper pseudogenes. *PLOS Genetics* **3**: 185-191.

Krawczak M, Ball EV, Cooper DN, 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. *The American Journal of Human Genetics* **63**: 474–488.

Kvas M, Marasas W, Wingfield BD, Wingfield, MJ, Steenkamp ET, 2009. Diversity and evolution of *Fusarium* species in the *Gibberella fujikuroi* complex. *Fungal Diversity*: 1–21.

Lenhart JS, Schroeder JW, Walsh BW, Simmons LA, 2012. DNA repair and genome maintenance in *Bacillus subtilis*. *Microbiology and Molecular Biology Reviews* **76**: 530–564.

Li T, Jin Y, Vershon AK, Wolberger C, 1998. Crystal structure of the MATa1 / MATα2 homeodomain heterodimer in complex with DNA containing an A-tract. *Nucleic Acids Research* **26**: 5707–5718.

Ma X, Rogacheva MV, Nishant KT, Zanders S, Bustamante CD, Alani E, 2012. Report mutation hot spots in yeast caused by long-range clustering of homopolymeric sequences. *Cell Reports* **1**: 36–42.

Maki H, 2002. Origins of spontaneous mutations : specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. *Annual Reviews in Genetics* **36**: 279–303.

Medema MH, Blin K, Cimermancic P, Jager VD, Zakrzewski P, Fischbach MA, Weber T, et al, 2011. antiSMASH : rapid identification , annotation and analysis of secondary metabolite biosynthesis gene clusters in bacterial and fungal genome sequences. *Nucleic Acids Research* **39**: 339–346.

Miyashita NT, Kawabe A, Innan H, 1999. DNA variation in the wild plant *Arabidopsis thaliana* revealed by amplified fragment length polymorphism analysis. *Genetics* **152**: 1723-1731.

Murakumo Y, 2002. The property of DNA polymerase ζ: REV7 is a putative protein involved in translesion DNA synthesis and cell cycle control. *Mutation Research* **510**: 37–44.

Nirenberg H, O'Donnell K, 1998. New *Fusarium* species and combinations within the Gibberella fujikuroi species complex. *Mycologia* **90**(3): 434–458.

Nouspikel T, 2009. So DNA repair really is that important? *Cellular and Molecular Life Sciences* **66**: 965–967.

O'Donnell K, Nirenberg HI, Cigelnik E, 2000. A Multigene phylogeny of the *Gibberella fujikuroi* species complex : Detection of additional phylogenetically distinct species. *Mycoscience* **41**: 61–78.

O'Neill M, Ryan C, 2001. Grammatical Evolution. *IEEE Transactions on Evolutionary Computation* **5**: 349–358.

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

Petrov D, Hartl D, 1999. Patterns of nucleotide substitution in *Drosophila* and mammalian genomes. *Proceedings of the National Academy of Science* **96**: 1475–1479.

Phasha MM, Steenkamp ET, Wingfield BD, Coetzee MPA, 2012. MSc Thesis, University of Pretoria, Intron architecture in *Fusarium*: Chapter 2, The architecture and distribution of introns in housekeeping genes of four *Fusarium* species: 31-67.

Reeves R, Adair JE, 2005. Role of high mobility group (HMG) chromatin proteins in DNA repair. *DNA Repair* **4**: 926–938.

Rogozin IB, Babenko VN, Milanesi L, Pavlov YI, 2003. Computational analysis of mutation spectra. *Briefings in Bioinformatics* **4**: 210–227.

Rogozin IB, Pavlov YI, 2003. Theoretical analysis of mutation hotspots and their DNA sequence context specificity. *Mutation Research* **544**: 65–85.

Rogozin IB, Malyarchuk BA, Pavlov YI, Milanesi L, 2005. From context-dependence of mutations to molecular mechanisms of mutagenesis. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*: 409–420.

Rosenberg MS, Subramanian S, Kumar S, 2003. Patterns of Transitional Mutation Biases Within and Among Mammalian Genomes. *Molecular Biology and Evolution* **20**: 988–993.

Russev G, Boulikas T, 1992. Repair of transcriptionally active and inactive genes during S and G2 phases of the cell cycle. *European Journal of Biochemistry* **204**: 267–272.

Shapiro GS, Aviszus K, Ikle D, Lawrence J, 2012. Predicting Regional Mutability in Antibody V Genes Based Solely on Di- and Trinucleotide Sequence Composition. *Journal of Immunology* **163**: 259-268.

Smith DR, Quinlan AR, Peckham HE, Smith DR, Quinlan AR, Peckham HE, Makowsky K, et al, 2008. Rapid whole-genome mutational profiling using next-generation sequencing technologies. *Genome Research* **18**: 1638–1642.

Smith DS, Creadon G, Jena PK, Portanova JP, Kotzin BL, Wysocki LJ, 1996. Di-and trinucleotide target preferences of somatic mutagenesis in normal and autoreactive B cells. *The Journal of Immunology*: 2642–2652.

Stack D, Neville C, Doyle S, 2007. Nonribosomal peptide synthesis in *Aspergillus fumigatus* and other fungi. *Microbiology* **153**: 1297–1306.

Stewart JE, Kawabe M, Abdo Z, Arie T, Peever TL, 2011. Contrasting codon usage patterns and purifying selection at the mating locus in putatively asexual *Alternaria* fungal species. *PloS One* **6**: 1-8.

Taylor FJ, Coates D, 1989. The code within the codons. *BioSystems* **22**: 177–187.

Vasyunina EA, Rogozin IB, Sinitsina OI, Plaksina AS, Rotskaya UN, 2004. Theoretical and experimental study of mutations by 8-oxoGuanine. *Computational Structural and Functional Genomics*: 200–203.

Vos P, Hogers R, Bleeker M, Reijans M, van de Lee T, Frijters A, Pot J, et al, 1995. AFLP : a new technique for DNA fingerprinting. *Nucleic Acids Research* **23**: 4407–4414.

Wakeley J, 1996. The excess of transitions among nucleotide substitutions: new methods of estimating transition bias underscore its significance. *Tree* **11**: 158–162.

Weber T, Baumgartner R, Renner C, Marahiel MA, Holak TA, 2000. Solution structure of PCP , a prototype for the peptidyl carrier domains of modular peptide synthetases. *Structure* **8**: 407–418.

Wingfield BD, Steenkamp ET, Santana Q, Coetzee MP, Bam S, *et al*, 2012. First fungal genome sequence from Africa : A preliminary analysis. *South African Journal of Science* **108**: 1–9.

Wingfield MJ, Hammerbacher A, Ganley RJ, Steenkamp ET, Gordon TR., Wingfield BD, Coutinho TA, 2008. Pitch canker caused by Fusarium circinatum – a growing threat to pine plantations and forests worldwide. *Australasian Plant Pathology* **37**: 319-334.

Zhao Z, Fu Y, Hewett-Emmett D, Boerwinkle E, 2003. Investigating single nucleotide polymorphism (SNP) density in the human genome and its implications for molecular evolution. *Gene* **312**: 207–213.

Table 1. The number of SNPs, SNP density and their distribution, relative to what is expected, over different regions of the *F. circinatum* genome.

Genome region[a]	Number of SNPs	SNP density (SNPs/k)	2Ti:Tv[b] (yes/no)	Ti:Ti:Ti:Ti[c] (yes/no)	Tv:Tv:Tv:Tv:Tv:Tv:Tv:Tv[d] (yes/no)
Whole genome	132 386	3	N	N	N
ORFs	47 148	2.2	N	N	N
Core exons	202	0.38	N	Y	Y
Core introns	137	1.65	N	Y	Y
Core intergenic regions	507	0.72	N	N	N
Core gene regions (exons, introns and intergenic regions)	846	0.46	Y	Y	Y
Core genes (exons and introns)	339	0.55	Y	N	N
NRPS exons	738	2.85	N	Y	N
NRPS introns	34	3.34	N	Y	Y
NRPS intergenic regions	403	2.86	N	Y	Y
NRPS gene regions (exons, introns and intergenic regions)	1175	2.87	Y	N	Y
NRPS genes (exons and introns)	772	2.87	Y	Y	Y

[a] Genome region refers to the different regions of the genome studied individually.

[b] 2Ti:TV refers to the expected 2 transitions: transversions ratio, this ratio is either observed (yes) or not observed (no).

[c] Ti:Ti:Ti:Ti refers to the expected 1:1:1:1 ratio between the different types of transitions, this ratio is either observed (yes) or not observed (no).

[d] Tv:Tv:Tv:Tv:Tv:Tv:Tv:Tv refers to the expected 1:1:1:1:1:1:1:1 ratios between the different types of transversions, this ratio is either observed (yes) or not observed (no).

Table 2. Frequency in percentage of the transitions and transversions throughout the respective regions of the *F. circinatum* genome.

Substitution	Genome wide	ORFs	Core gene region[a]	Core genes[b]	Core gene exons	Core gene introns	Core gene intergenic regions	NRPS gene regions[c]	NRPS genes[d]	NRPS exons	NRPS introns	NRPS intergenic
Transitions	**71.40**	**71.18**	**69.60**	**73.10**	**74.30**	**71.50**	**67.70**	**71.94**	**74.87**	**74.79**	**76.47**	**66.50**
G -> A	18.20	17.90	19.20	18.90	18.80	19.00	19.50	17.29	17.91	18.49	5.88	16.13
A -> G	17.50	17.00	11.08	15.00	14.80	15.30	8.30	17.90	19.12	18.91	23.53	15.63
T -> C	17.10	17.40	19.20	21.20	21.80	21.20	17.90	17.55	17.51	17.51	17.65	17.62
C -> T	18.70	18.90	20.30	18.00	18.80	16.80	21.90	19.29	20.32	19.89	29.41	17.12
Transversions	**28.60**	**28.60**	**30.40**	**27.40**	**25.70**	**28.50**	**32.30**	**28.06**	**25.13**	**25.21**	**23.53**	**33.50**
A -> C	3.90	4.00	4.80	5.00	3.00	6.60	5.10	4.08	4.14	4.06	5.88	3.97
A -> T	3.30	2.90	3.50	3.20	2.50	4.40	3.70	2.78	2.67	2.80	0.00	2.98
G -> T	3.00	4.20	5.30	5.30	5.90	4.40	5.30	4.34	3.74	3.78	2.94	5.46
G -> C	4.00	3.20	2.80	2.70	3.50	1.50	3.00	3.04	3.07	3.08	2.94	2.98
C -> A	4.00	4.30	4.60	2.50	2.50	2.90	5.90	4.78	4.55	4.76	0.00	5.21
C -> G	3.20	3.40	1.90	2.10	2.50	1.50	1.80	1.48	0.13	0.00	2.94	3.97
T -> A	3.20	2.80	3.90	2.10	1.00	3.60	5.10	3.39	3.48	3.36	5.88	3.23
T -> G	4.00	4.00	3.50	5.00	6.00	3.60	2.60	4.17	3.34	3.36	2.94	5.71

[a] Core gene region includes core gene exons, introns and intergenic regions.

[b] Core genes include core gene exons and introns.

[c] NRPS gene region includes NRPS exons, introns and intergenic regions.

[d] NRPS genes include NRPS exons and introns.

Table 3. Results of the tests for associations between specific substitutions and neighbouring base in the exons of the *F. circinatum* core genes.

Nucleotide affected[a]	Type of mutation[b]	Neighbour	Association with neighbour (Yes/No)[c]	Chi squared value[d]	Degrees of freedom[e]	*P* value[f]	Motif[g]
A	Ti	5′	N	7.3	3	>0.05	
		3′	N	1.18	3	>0.05	
		Both	Y	30.15	15	<0.02	CAG
	Tv	5′	N	0.29	3	>0.05	
		3′	N	1.18	3	>0.05	
		Both	N	6.99	15	>0.05	
	All	5'	N	5.24	3	>0.05	
		3'	N	0.62	3	>0.05	
		Both	N	22.69	15	>0.05	
C	Ti	5'	N	0.39	3	>0.05	
		3'	N	7.32	3	>0.05	
		Both	N	22.06	15	>0.05	
	Tv	5'	Y	10.83	3	<0.02	CC
		3'	N	7.22	3	>0.05	
		Both	Y	29.51	15	<0.02	CCG
	All	5'	N	2.83	3	>0.05	
		3'	N	7.18	3	>0.05	
		Both	N	21.38	15	>0.05	
G	Ti	5'	N	3.92	3	>0.05	
		3'	Y	25.52	3	<0.05	TGT
		Both	N	1.86	15	>0.05	
	Tv	5'	N	4.73	3	>0.05	
		3'	Y	26.02	3	<0.05	GT
		Both	Y	9.39	15	<0.05	CGG
	All	5'	N	5.31	3	>0.05	
		3'	Y	26.12	3	<0.05	TGT
		Both	N	7.66	15	>0.05	
T	Ti	5'	N	2.66	3	>0.05	
		3'	Y	11.89	3	<0.01	TA
		Both	N	22.18	15	>0.05	
	Tv	5'	N	2.53	3	>0.05	
		3'	N	1.29	3	>0.05	
		Both	N	7.81	15	>0.05	
	All	5'	N	1.89	3	>0.05	
		3'	N	3.06	3	>0.05	
		Both	N	21.82	15	>0.05	

[a] Nucleotide affected refers to nucleotide at SNP position on the reference sequence.

[b] Type of substitution tested for significant association with neighbouring bases, transitions (Ti), transversions (Tv) or all substitutions (both).

[c] A significant association between a certain substitution and its neighbour is observed when the χ^2 squared analysis yields a $P<0.5$.

[d] Chi squared value obtained from the analysis.

[e] Degrees of freedom (df) used in χ^2 squared analysis, df=n-1 where n=number of categories.

[f] $P<0.5$ indicates an occurrence of a SNP flanked by a specific neighbour more often than expected.

[g] Resulting motifs due to associations between SNPs and their neighbouring bases, the mutated base is underlined.

Table 4. Results of the tests for associations between specific substitutions and neighbouring base in the introns of the *F. circinatum* core genes.

Nucleotide affected	Type of mutation	Neighbour	Association with neighbour (Yes/No)	Chi squared value	Degrees of freedom	*P* value	Motif
A	Ti	5'	N	2.13	3	>0.05	
		3'	N	4022	3	>0.05	
		Both	N	15.61	15	>0.05	
	Tv	5'	N	4.5	3	>0.05	
		3'	N	1.76	3	>0.05	
		Both	N	11.47	15	>0.05	
	All	5'	N	3.12	3	>0.05	
		3'	N	3.76	3	>0.05	
		Both	N	13.74	15	>0.05	
C	Ti	5'	N	1.39	3	>0.05	
		3'	N	1.93	3	>0.05	
		Both	N	15.38	15	>0.05	
	Tv	5'	N	0.33	3	>0.05	
		3'	N	2.12	3	>0.05	
		Both	N	7.78	15	>0.05	
	All	5'	N	0.73	3	>0.05	
		3'	N	2.99	3	>0.05	
		Both	N	14.29	15	>0.05	
G	Ti	5'	N	3.5	3	>0.05	
		3'	N	1.71	3	>0.05	
		Both	N	22.32	15	>0.05	
	Tv	5'	N	0.49	3	>0.05	
		3'	N	0.46	3	>0.05	
		Both	N	10.34	15	>0.05	
	All	5'	N	2.25	3	>0.05	
		3'	N	0.69	3	>0.05	
		Both	N	17.68	15	>0.05	
T	Ti	5'	N	2.37	3	>0.05	
		3'	N	1.59	3	>0.05	
		Both	N	20.82	15	>0.05	
	Tv	5'	Y	8.71	3	<0.05	AT
		3'	N	1.97	3	>0.05	
		Both	N	18	15	>0.05	
	All	5'	N	7.11	3	>0.05	
		3'	N	1.92	3	>0.05	
		Both	Y	29.29	15	<0.02	ATA

[a] see the footnotes of Table 3 for description of the various entries.

Table 5. Results of the tests for associations between specific substitutions and neighbouring base in the intergenic regions associated with the exons of the *F. circinatum* core genes.

Nucleotide affected	Type of mutation	Neighbour	Association with neighbour (Yes/No)	Chi squared value	Degrees of freedom	*P* value	Motif
A	Ti	5'	N	3.01	3	>0.05	
		3'	N	1.96	3	>0.05	
		Both	N	24.34	15	>0.05	
	Tv	5'	N	1.08	3	>0.05	
		3'	N	1.99	3	>0.05	
		Both	N	12.05	15	>0.05	
	All	5'	N	3.31	3	>0.05	
		3'	N	2.49	3	>0.05	
		Both	N	11.86	15	>0.05	
C	Ti	5'	N	0.72	3	>0.05	
		3'	Y	13.19	3	<0.01	CG
		Both	N	16.61	15	>0.05	
	Tv	5'	N	5.31	3	>0.05	
		3'	N	2.04	3	>0.05	
		Both	N	18.02	15	>0.05	
	All	5'	N	0.34	3	>0.05	
		3'	Y	10.72	3	<0.02	CG
		Both	N	11.94	15	>0.05	
G	Ti	5'	N	5.07	3	>0.05	
		3'	N	5.67	3	>0.05	
		Both	N	24.04	15	>0.05	
	Tv	5'	N	0.78	3	>0.05	
		3'	N	5.42	3	>0.05	
		Both	N	12.74	15	>0.05	
	All	5'	N	5.1	3	>0.05	
		3'	N	8.07	3	>0.05	
		Both	Y	27.2	15	<0.05	CGT
T	Ti	5'	N	2.56	3	>0.05	
		3'	N	6.56	3	>0.05	
		Both	N	12.22	15	>0.05	
	Tv	5'	N	3.33	3	>0.05	
		3'	N	1.96	3	>0.05	
		Both	N	13.55	15	>0.05	
	All	5'	N	2.5	3	>0.05	
		3'	N	2.97	3	>0.05	
		Both	N	15.12	15	>0.05	

[a] see the footnotes of Table 3 for description of the various entries.

Table 6. Results of the tests for associations between specific substitutions and neighbouring base in the exons of the *F. circinatum* NRPS genes.

Nucleotide affected	Type of mutation	Neighbour	Association with neighbour (Yes/No)	Chi squared value	Degrees of freedom	*P* value	Motif
A	Ti	5'	Y	30.15	3	<0.01	TA
		3'	N	2.01	3	>0.05	
		Both	Y	35.79	15	<0.01	TAT
	Tv	5'	Y	10.81	3	<0.02	
		3'	N	4.07	3	>0.05	
		Both	Y	34.59	15	<0.01	GAC
	All	5'	Y	39.31	3	<0.0001	TA
		3'	N	4.96	3	>0.05	
		Both	Y	55.76	15	<0.0001	TAA
C	Ti	5'	N	5.18	3	>0.05	
		3'	N	6.72	3	>0.05	
		Both	N	28.64	15	>0.05	
	Tv	5'	N	2.48	3	>0.05	
		3'	N	0.73	3	>0.05	
		Both	N	12.56	15	>0.05	
	All	5'	N	4.37	3	>0.05	
		3'	N	7.16	3	>0.05	
		Both	N	30.09	15	>0.05	
G	Ti	5'	N	3.39	3	>0.05	
		3'	Y	17.94	3	<0.001	GT
		Both	Y	27.18	15	<0.05	CGT
	Tv	5'	N	5.76	3	>0.05	
		3'	N	0.82	3	>0.05	
		Both	N	24.502	15	>0.05	
	All	5'	N	4.29	3	>0.05	
		3'	Y	12.73	3	<0.01	GT
		Both	Y	27.62	15	<0.05	CGG
T	Ti	5'	N	7.09	3	>0.05	
		3'	N	6.01	3	>0.05	
		Both	Y	34.72	15	<0.05	GTA
	Tv	5'	Y	8.23	3	<0.05	TT
		3'	N	5.99	3	>0.05	
		Both	Y	35.134	15	<0.05	TTT
	All	5'	Y	10.99	3	<0.02	TT
		3'	N	4.19	3	>0.05	
		Both	Y	36.37	15	<0.05	TTA

[a] see the footnotes of Table 3 for description of the various entries.

Table 7. Results of the tests for associations between specific substitutions and neighbouring base in the intergenic regions associated with the exons of the *F. circinatum* core genes.

Nucleotide affected	Type of mutation	Neighbour	Association with neighbour (Yes/No)	Chi squared value	Degrees of freedom	*P* value	Motif
A	Ti	5'	Y	24.16	3	<0.0001	TA
		3'	N	2.94	3	>0.05	
		Both	Y	35.01	15	<0.01	TAC
	Tv	5'	N	5.71	3	>0.05	
		3'	N	1.96	3	>0.05	
		Both	N	15.8	15	>0.05	
	All	5'	Y	28.13	3	<0.0001	TA
		3'	N	1.32	3	>0.05	
		Both	Y	33.55	15	<0.01	TAA
C	Ti	5'	Y	10.34	3	<0.02	AC
		3'	N	6.49	3	>0.05	
		Both	N	18.13	15	>0.05	
	Tv	5'	Y	9.54	3	<0.05	AC
		3'	N	4.77	3	>0.05	
		Both	N	21.43	15	>0.05	
	All	5'	Y	14.76	3	<0.001	AC
		3'	Y	9.55	3	<0.05	CA
		Both	Y	27.46	15	<0.05	ACA
G	Ti	5'	N	2.89	3	>0.05	
		3'	N	2.13	3	>0.05	
		Both	N	9.27	15	>0.05	
	Tv	5'	N	0.52	3	>0.05	
		3'	N	5.17	3	>0.05	
		Both	N	16.05	15	>0.05	
	All	5'	N	3	3	>0.05	
		3'	N	5.41	3	>0.05	
		Both	N	15.47	15	>0.05	
T	Ti	5'	N	4.62	3	>0.05	
		3'	Y	12.35	3	<0.02	TA
		Both	Y	26.74	15	<0.05	CTA
	Tv	5'	N	0.99	3	>0.05	
		3'	N	1.94	3	>0.05	
		Both	N	14.87	15	>0.05	
	All	5'	N	2.52	3	>0.05	
		3'	Y	13.31	3	<0.01	TA
		Both	N	16.44	15	>0.05	

[a] see the footnotes of Table 3 for description of the various entries.

Table 8. SNPs in the different domains of the 9 NRPS genes of *F. circinatum* examined in this study.

FSP34 Contig[a]	Total number of SNPs[b]	NRPS domains[c]	SNPs in domain
51	5	C_1	1
		PCP_1	0
		A_1	2
		C_2	0
		PCP_2	0
		A_2	0
		C_3	0
383	9	A	1
		PCP	1
		E	0
		C	1
869	3	A_1	0
		A_2	0
		PCP_1	0
		C_1	1
		PCP_2	0
		C_2	0
		A_3	0
		C_3	0
		PCP_3	0
1064	38	C	8
		A	1
		PCP	2
1228	10	C	5
		A	1
1385	7	C	2
		PCP	1
		A	2
2244	11	A	1
		PCP	0
		C	1
2636	95	C_1	0
		A_1	0
		PCP_1	0
		E_1	2
		C_2	18
		A_2	29
		PCP_2	2
		E_2	19
2669	17	A	17

[a] The FSP34 contig number on which the NRPS gene occurs.

[b] The total amount of SNPs within the protein coding sequence, sequence encoding domains as well as regions between domains.

[c] The different domains forming part of the NRPS. A = adenylation domain, PCP = peptidyl carrier domain, C = condensation domain and E = epimerization domain. For the multimodular enzyme encoding genes A1 = the adenylation domain of the first module and A2 = the second modules adenylation domain.

Table 9. Mutated TT dinucleotides in NRPS gene on contig 2636

Part of gene[a]	**Number of TT dinucleotides**[b]	**Number of TT mutated**[c]
IGR_1	20	0
C_1	56	0
IGR_2	26	0
A_1	62	0
IGR_3	11	0
PCP_1	6	0
IGR_4	1	0
E_1	60	0
IGR_5	25	0
C_2	39	2
IGR_6	26	0
A_2	61	5
IGR_7	12	0
PCP_2	10	0
IGR_8	1	0
A_3	50	0
IGR_9	0	0

[a] The region of the gene studied where IGR 1-9 are regions occurring between domains, A = adenylation domain, PCP = peptidyl carrier domain, C = condensation domain and E = epimerization domain. For the multimodular enzyme encoding genes A1 = the adenylation domain of the first module and A2 = the second modules adenylation domain.

[b] The total amount of TT dinucleotides occurring within the specific region.

[c] The amount of TT dinucleotides that harboured a SNP.

Chapter 3

Calculating the spontaneous mutation rate of *Fusarium circinatum*

Abstract

Mutation is a fundamental component of evolution and speciation. However, other than for the model organisms, almost no information is available about mutation and the rate at which it occurs. The aim of this study was to estimate the spontaneous mutation rate in the filamentous fungus *Fusarium circinatum*. For this purpose a fluctuation analysis was conducted to determine the spontaneous reversion rate of a *nit3 F. circinatum* mutant. From these results the mutation rate of this fungus was determined to be 3.7 x 10^{-6} mutations/genome/generation. The mutation rate of *F. circinatum*, in terms of number of mutations/bp/generation is similar to what was calculated for *Aspergillus nidulans*. Both of these organisms have a higher mutation rate than what was estimated for *A. fumigatus*. This difference could be due to the choice of genes/markers that were used for the fluctuation analysis. Although this estimate gives an idea of how the spontaneous mutation rate of *F. circinatum* compares to other fungi, it is most likely not an accurate estimate of what is happening under natural conditions. This is due to the difficulties encountered when calculating the generation time of filamentous fungi as well as the way in which *F. circinatum* produces conidia. Future work should include identifying a set of genes that is representative of the whole genome and can be used to estimate mutation rates of other *Fusarium spp.* and strains within a species. Furthermore the optimization and perfection of current molecular methods will aid in accurate calculation of generation time of filamentous fungi. Problems encountered due to the conidiation process will require the development of a technique to trace the exact timing of a mutation.

Introduction

Spontaneous mutation and the rate at which it occurs are fundamental to all the biological diversity on earth. Mutation is a nucleotide sequence alteration that occurs due to DNA damage or lesions that have not been repaired sufficiently (Rogozin *et al.* 2003). These lesions can be a consequence of the normal internal processes of the cell (e.g., DNA mismatches due to polymerase errors during replication) causing so-called spontaneous mutations, or they can be caused by external mutagens such as UV irradiation leading to induced mutations (Maki 2002; Foster 2006). The type of DNA alteration that occurs can generally be classified as substitution mutations (where one nucleotide is exchanged for another) or indels (insertions and deletions of a nucleotide), although larger stretches of DNA can also be affected, e.g., recombination can cause large insertions, deletions, transpositions and inversions (Maki 2002; Rogozin *et al*. 2003). The rate at which spontaneous mutations occur is known as the spontaneous mutation rate and is most often measured as the number of mutations that occurred within the genome in one cellular generation (i.e., the number of mutations that occur during genome duplication) (Baer *et al.* 2007).

Knowledge regarding mutation and mutation rate is crucial for our understanding of evolutionary processes. For example, mutation drives the evolutionary processes involved in the diversification of pathogens enabling them to adapt and proliferate and to cause disease (Rogozin *et al*. 2003; Foster 2006). The evolutionary rate of an organism is thus linked to its mutation rate. In the fungal kingdom limited mutation rate studies have been conducted, leaving mutation rate data incomplete and subsequently limiting our understanding of how evolutionary rates of species compares within the fungal kingdom. Furthermore, there seems to be variation in mutation rate between different taxa, as well as within a taxon. The variation is thought to be associated with different factors including environmental stresses and available DNA repair machinery (Baer *et al*. 2007). However, the exact reason for variation in mutation rate and thus evolutionary rate between and within a species or groups of species remains unknown.

Mutation rate is typically determined using one of two widely used methods, mutant accumulation and fluctuation analyses (Foster 2006). Mutant accumulation is based on the emergence of new mutants and proliferation of existing mutants in an exponentially growing culture. The mutation rate is the mutation fraction increase over a certain period of time.

Although seemingly simple, a major limitation of mutant accumulation analysis is that mutations occur within a population before it reaches exponential phase (Foster 2006). This is in contrast to mutation rate calculations using fluctuation analyses where parallel cultures are grown and subsequently plated out on selective media to differentiate between the mutants and wild type (Foster 2006). Fluctuation analysis was first introduced in 1943 by Luria and Delbrück to show that mutations arise in the absence of selection pressures. They showed that bacterial cultures generated bacteriophage resistance prior to the addition of the bacteriophages. Mathematical methods complementing fluctuation analyses to calculate spontaneous mutation rates have been developed and improved since 1943 to maximize the precision of mutation rate calculations (Foster 2006).

A number of markers can typically be employed for determining mutation rate. These include DNA-based markers such as single nucleotide polymorphisms, short tandem repeats (microsatellites) and haplotypes where a number of DNA markers are combined, as well as phenotypic mutants (Morin *et al*. 2004). In most cases, including this study, phenotypic mutants are used as markers and a change in phenotype is regarded as a single mutation (Luria & Delbruck 1943). This might lead to an underestimation of the spontaneous mutation rate as more than one DNA change might give rise to the altered phenotype (Stewart *et al*. 1990). Spontaneous mutation rate estimates may also be influenced by the physical genomic location of the marker. For example, markers in euchromatin and heterochromatin will yield different mutation rates due to the influence of DNA packaging on the ease of access of the DNA repair machinery proteins to the DNA. Next generation sequencing has further opened up the opportunities to study mutagenesis at a molecular level and to trace mutations from one generation to the next.

Mutation rate data within the fungal kingdom is limited and mostly restricted to yeast and some filamentous species, while most work has been done on animals. It was found in mammals and insects that local substitution patterns differ at silent sites (loci that are putatively under neutral selection) across the genome. The results of a study by Fox *et al*. (2008) demonstrated that mutation rates across the genomes of 27 fungal species were uniform except for three *Candida* species in which non-uniform within-genome mutation rates were found. The exact reason for this apparently widespread intragenomic mutation rate uniformity is unknown but it was hypothesised that replication timing plays a role. According to this view, replication timing is regarded as mutagenic (i.e., regions of the genome that are replicated last are more prone to mutations (Fox *et al*. 2008) and in species with non-uniform

within-genome mutation rates, replication timing is static (Fox *et al.* 2008). An experiment on replication timing in *Candida* and a fungal species showing within-genome mutation rate heterogeneity is required to further investigate the matter (Fox *et al.* 2008).

In this study, mutation and mutation rate in the filamentous ascomycete *Fusarium circinatum* (Nirenberg & O'Donnell emend. Britz, Coutinho, Wingfield & Marasas) will be considered. *F. circinatum* is the causal agent of pitch canker of *Pinus* species (Wingfield *et al*. 2008) and regarded as the most important pathogen in South African commercial *Pinus* nurseries where it is annually responsible for huge losses (Mitchell *et al* 2011; Wingfield *et al*. 2008). Although the genome of this pathogen has been published recently (Wingfield *et al*. 2012), the methodologies associated with whole-genome based quantification of spontaneous mutation rate are still in their infancy. This study therefore employed genetic mutants, specifically nitrogen-non-utilizing (*nit*) mutants, to determine the spontaneous mutation rate of *F. circinatum*. The simple procedure for differentiating *nit* and wild-type phenotypes, which in turn allows for easy identification of spontaneous revertants, makes the use of this marker suitable for fluctuation analyses.

Nitrogen is vital for growth (Amaar & Moore 1998; Jargeat *et al*. 2003) and can be obtained from a variety sources including ammonia, glutamine and glutamate, as well as nitrate (Amaar & Moore 1998). When only nitrate is available the nitrate transporter is responsible for the uptake thereof after which it is converted to nitrite by the enzyme nitrate reductase subsequently leading to its conversion to ammonia by nitrite reductase (Amaar & Moore 1998). The functioning of these enzymes are regulated by other factors such as the product of the *nit3* gene in *Fusarium* (Leslie *et al*. 2007), which is thought to be a transcription factor that acts as a positive regulator involved in the induction of nitrate metabolism (Johnstone *et al*. 1990). In *Fusarium*, *nit* mutants are observed when cells are grown in the presence of the chlorate (Leslie *et al*. 2006), which is an alternate substrate for nitrate reductase, leading to the reduction of chlorate to chlorite. This process is toxic because chlorate occupies the active site of the enzyme rendering it unable to reduce nitrate to nitrite, the consequence is that only individuals with a *nit* mutation survive in the presence of chlorate (Leslie *et al*. 2006). Mutant isolates where the *nit3* gene product functions abnormally will be distinguishable by their inability to utilize either nitrate or nitrite as sole nitrogen sources (Leslie *et al*. 2006).

The aim of this study was to estimate the mutation rate of *F. circinatum*. To achieve this aim a fluctuation analysis was conducted to study the spontaneous reversion of a *nit3* mutantion of this fungus. Addressing the gap in mutation rate data within the fungal kingdom will give insight into how evolutionary rates of different species compares. Furthermore it will help address the question whether environmental stresses in general increases mutation rates or if a specific factor in the habitat induces mutagenesis to cause variation in spontaneous mutation rate among and within a species.

Materials and Methods

1. Fungal isolate and generation of *nit* mutants

The isolate used in this study is FSP34. It was isolated from *Pinus* spp. in California (USA) by TR Gordon (De Vos *et al*. 2007). A hybrid genetic linkage map based on amplified fragment length polymorphisms (De Vos *et al*. 2007) as well as a fully annotated genome sequence is available for FSP34 (Wingfield *et al*. 2012). The isolate was cultured on half strength potato dextrose agar (PDA) (Merck Biolab, Wadeville, Gauteng) in the dark at 25°C. After a week of incubation, FSP34 was inoculated onto minimal medium (MM) amended with 33 mM $KClO_3$ and 1.6% (w/v) L-asparagine (Merck Biolab). These cultures were then incubated at 25°C for 10 days or until the culture started to sector, which is indicative of the emergence of chlorate resistant mutants. Mycelia from fast growing sectors were plated onto MM supplemented with 23.5 mM $NaNO_3$ (referred to as nitrate medium) followed by incubation at 25°C for 7 days. This allowed identification and disposal of chlorate resistant, nitrate utilizing (*crn*) mutants.

The MM was prepared from basal medium (BM) which contained 30% (w/v), sucrose (Merck Biolab), 4.15 mM $MgSO_4.7H_2O$ (Merck Biolab), 6.71 mM KCl (Saarchem (Pty) Ltd., Krugersdorp, Gauteng), 0.036 mM FeSO4.7H2O (Merck Biolab), 20% (w/v) agar (Merck Biolab) and 0.2% (v/v) trace element solution. This solution included 26 µM citric acid, 19 µM $ZnSO_4.6H_2O$, 0.5 µM $Fe(NH_4)_2(SO_4)_2.6H_2O$, 1.6 µM $CuSO_4.5H_2O$, 0.3 µM $MnSO_4$, 0.8 µM H_3BO_3 and 0.2 µM $NaMoO_4.2H_2O$ (Correll *et al*. 1987).

Cultures, showing mutant growth (absence of aerial mycelia) on nitrate media after 10 days were phenotyped based on their ability to utilize nitrite, hypoxanthine and ammonium as exclusive nitrogen sources. For this purpose nitrite medium containing MM with 7.25 mM

$NaNO_2$ (Merck Biolab), hypoxanthine medium containing MM with 1.47 mM hypoxanthine (Sigma-Aldrich, Aston Manor, South Africa) and ammonium medium containing MM with 5.43 mM NH_4-tartrate (Sigma-Aldrich) were used. Again, plates were incubated at 25°C for 10 days and mutant growth was recorded. Mutants that were able to utilize ammonium and hypoxanthine as sole nitrogen sources allowing profuse growth, but not on nitrate and nitrite (showing mutant growth with no aerial mycelia) were classified as *nit3* mutants (Klittich & Leslie 1989). The ability of the isolates to utilize nitrate as a sole nitrogen source has already been determined when the chlorate resistant sectors were plated out on nitrate medium.

2. Fluctuation analysis

A *nit3* mutant isolate was used in the analyses to determine phenotype reversion rate. The isolate was cultured on half strength PDA for 10 days at 25°C to allow spore formation. A spore suspension was then made by washing the plate with sterile water (Adcock Ingram, Midrand, Gauteng) and filtering the suspension through cheese cloth. The spore concentration was determined by direct counting using a haemocytometer (Marienfeld, Lasec, Cape Town). To determine the optimum number of germinating spores per plate for easy phenotyping, a dilution series (undiluted up to a 10^{-10} dilution) was prepared for the spore suspension, which was plated onto nitrite containing MM. The latter was also used to calculate the germination efficiency, which involved inoculating 100 µl of the spore suspension onto a plate with nitrite containing MM and counting the number of germinated spores following incubation (72 hours at 25 °C). A total of 15 microfuge tubes were prepared with a spore concentration of 1500 spores/ml basal broth (BB, *i.e.*, BM without agar). These were incubated for 24 hours at 25 °C in the dark without shaking.

Following incubation the suspensions in the 15 microfuge tubes were centrifuged (16, 000 x g, 6 min) after which the supernatant was discarded. The pellet was washed by adding 1ml of 1 x PBS (phosphate buffered saline) (Sigma-Aldrich) and centrifugation (16, 000 x g, 6 min). The supernatant was discarded again and the pellet was resuspended in 25µl PBS. The washed spore suspension was then plated out on 145 x 20 mm petri dishes (Concorde plastics, Longdale, Johannesburg) with nitrite containing BM to allow identification of revertants through the observation of profuse fungal growth. This process (Figure 1) was repeated twice to yield 3 technical repeats.

3. Mutation rate calculation

Results from the fluctuation analysis were used to determine P_o (the fraction of cultures that remained mutants). P_o was used to determine *m*, which is the number of mutations (Foster 2006):

$$m = - \ln P_o \quad (1)$$

The P_o method of mutation rate calculation is restricted by biological factors such as the germination efficiency. These restrictions can lead to an erroneously low *m* value. This was compensated for by taking into consideration the germination efficiency (*z*) when calculating the actual *m* value (m_{act}) using the observed *m* value (m_{obs}) (Foster 2006):

$$m_{act} = m_{obs} \frac{z-1}{z \ln (z)} \quad (2)$$

The generation time of *F. circinatum* was assumed to be similar to that of "*F. moniliforme*" (Seifert et al 2003), which has been calculated in a previous study (Macris & Kokke 1977). This generation time was used to calculate the number of generations that passed in the 24 h (*n*) of incubation to estimate the amount of genomes after 24 h (N_t) with *No* being the starting number of genomes:

$$N_t = N_o (2^n) \quad (3)$$

The mutation rate was then calculated using Drake's formula (Drake 1991):

$$\mu = \frac{m}{Nt} \quad (4)$$

4. DNA extraction and sequencing of the *nit3* gene

Eight randomly selected revertant progeny, 6 randomly selected mutant progeny, the *nit3* mutant parent and the FSP34 isolate were sub-cultured on half strength PDA and incubated for 7 days at 25°C. DNA was extracted using a modification of the CTAB protocol as used by (Steenkamp *et al*. 1999). Briefly, mycelia were harvested directly from the surface of the growth media and homogenized in extraction buffer using ceramic beads. The resulting homogenate was incubated at 60°C for 3 h and centrifuged at room temperature for 30 min (16,000 x g). This was followed by a series of phenol : chloroform (1:1, v/v) extractions until the interphase between the aqueous and organic phases was clear. A final chloroform extraction was performed to eliminate residual phenol. Nucleic acid precipitation

was achieved by adding 0.1 volume of 3M sodium acetate (pH 6) and 1 volume of 100% ethanol and overnight incubation at -20°C. The DNA was harvested by centrifugation (16,000 x g, 30 min), washed with 70% ethanol and resuspended in 25 µl sterile distilled water. The integrity of the extracted DNA was confirmed using 2% (w/v) gel electrophoresis at 100V/cm and Gel Red dye (Biotium, Hayward, CA, USA) for visualization of the samples under UV. The DNA was quantified, using a Nanodrop spectrophotometer ND-1000 (Nanodrop Technologies, Rockland USA), and subsequently diluted to 50 ng/µl

To design primers for amplifying the *nit3* gene in *F. circinatum*, the *Fusarium* Comparative Database (Broad Institute) was used to find the sequence of this gene in *F. verticillioides, F. oxysporum* and *F. graminearum*. These sequences were used in a local BLAST search against the genome sequence of *F. circinatum* to identify the *nit3* gene of this fungus. Two primers, Primer Forward (5'-TCCCTCGCTCATTATATTCT-3') and Primer Reverse (5'-TCACCGTTCTCACTATTCCA-3'), were designed using CLC-bio genomics workbench version 4.9 (CLC bio A/S) to allow specific PCR amplification of the *F. circinatum nit3* gene. These 50 µl reactions were conducted in Super-Therm DNA polymerase reaction buffer (International Healthcare Distributors, Sandton, Gauteng) containing 1 mM dNTPs (25mM of each dNTP), 2.5 mM $MgCl_2$, 0.25 µM of each primer, 100 ng DNA template and 2.5 units of Super-Therm DNA polymerase (International Healthcare Distributors). PCR conditions comprised of an initial denaturation step of 1 min at 95°C followed by 30 cycles denaturation at 94°C (30 s), annealing at 58°C (30 s) and elongation at 72°C (1 min). Extension was completed at 72°C (5 min) on the T100™ thermal cycler (Biorad, Berkeley, California).

PCR products were purified using Promega MSB® SpinPCRapace clean-up kit (Southern Cross Biotechnology, Clarmont, Western Cape). The purified products were sequenced in both directions on an ABI PRISM 3130 using an ABI PRISM™ Dye Terminator Cycle Sequencing Ready Reaction Kit (Applied Biosystems, Fairlands, Gauteng). CLCBio Genomics workbench, version 4.9 was used to assemble the forward and reverse sequences. This software was also used compare the sequences of the revertants to those of the *nit3* mutant parent and FSP34.

Results

1. Generation of a *nit3* mutant

A total of 20 chlorate resistant isolates were plated out on MM. Of these 12 of the 20 cultures were identified as *crn* mutants and discarded, because even though they were resistant to chlorate they showed wild type growth on the MM. Of the 8 remaining isolates phenotyped further, one was able to utilize ammonium and hypoxanthine as sole nitrogen sources but unable to utilize nitrite and nitrate exclusively. This isolate therefore represented a *nit3* mutant.

2. Fluctuation analysis

Abundant spores were produced by the *nit3* mutant and an optimal dilution had to be identified to allow efficient differentiation between wild-type and mutant growth. In this study, the growth of 10-15 germinating spores per plate could be differentiated comfortably. Different spore concentrations were plated out to determine the germination efficiency of the spores to control for non-viable spores. Based on the growth at these spore densities, the germination efficiency was estimated at 10%. The germination efficiency was determined by dividing the amount of germinated spores by the number of spores that was plated out. With a 10% germination efficiency a total of 100-150 spores in the microfuge tubes will yield 10-15 germinating spores per plate, which can be differentiated comfortably.

Following the appropriate wash and incubation steps of the germinating spores in the 15 microfuge tubes a total of 100 isolates were observed with 91 showing mutant growth and 9 showing wild-type growth (growing profusely, forming aerial mycelia). For the first repeat experiment a total of 133 isolates were examined of which 122 showed mutant and 11 wild-type growth. In the third repeat, a total of 109 isolates were examined of which 100 showed mutant growth and 9 displayed the wild-type phenotype (Figure 1).

3. Mutation rate calculation

The mutation frequency (proportion of isolates that reverted) obtained from each of the repeats was calculated by dividing the number of revertants by the total number of colonies. The mutation frequencies for the three experiments were respectively 0.09, 0.083 and 0.083, with an average of 0.085. This yielded a value of 0.915 for P_o. Based on this value for P_o and the observed plating efficiency, m_{act} was calculated as 0.41.

It was estimated that "*F. moniliforme*" had a generation time of approximately 3.91 and 3.15 in 2% and 4% carob sugar medium (Macris & Kokke, 1977). In this study an average of these two generation times, 3.53 h, was considered to reflect the generation time

of *F. circinatum* in broth MM medium with 3% sugar. It was therefore assumed that a new genome of *F. circinatum* emerged every 3.53 h, thus replicated 6.818 times during the 24 hours of incubation, giving rise to a total of 111430 genomes (equation (3)). This allowed estimation of the mutation rate of *F. circinatum* as 3.7 x 10^{-6} mutations/genome/generation (equation (4)).

4. Sequence of the *nit3* gene

Sequence analysis used to determine whether the observed mutant and wild-type phenotypes were in any way linked to nucleotide substitutions in the *nit3* gene of the various isolates. However, in 810 base pair (bp) fragment examined in the mutant parent, 8 randomly selected revertant progeny, 6 randomly chosen mutant progeny and the FSP34 isolate, no sequence differences were observed.

Discussion

Based on the results of this study, a mutation rate of 3.7 x 10^{-6} mutations/genome/generation was estimated for *F. circinatum*. Compared to the mutation rates estimated for *A. nidulans* and *A. fumigatus*, this value for *F. circinatum* is in the same order of magnitude as *A. nidulans* (Baracho & Baracho 2003) and 2 orders of magnitude higher than that of *A. fumigatus* (Alvarez-Perez *et al*. 2010). The mutation rate estimated for *F. circinatum* is also 3 orders of magnitude lower than that estimated for *Neurospora crassa* (Auerbach 1959). However, when genome size is taken into consideration and mutation rate is expressed as mutations/bp/generation, the value for *F. circinatum* is an order of magnitude lower than that of *A. nidulans* with its 31 Mb genome (Brody & Carbon 1989)), but still 2 orders of magnitude higher than the mutation rate of *A. fumigatus* with a genome size of 28-30 Mb (Denning *et al*. 2003) and 3 orders of magnitude lower than what was estimated for *Neurospora crassa* (Drake 2009).

A possible reason for difference between the mutation rates of *F. circinatum* and other filamentous fungi could be linked to the genes or markers that were used for these estimations. Elastase activity was used to estimate the mutation rate of *A. fumigatus* (Alvarez-Perez *et al*. 2010) whereas functionality of the *nit3* gene was used in this study. Elastase activity is important in pathogenesis (Alvarez-Perez *et al*. 2010), while *nit3* plays a fundamental role in nitrate assimilation, which is crucial for growth and survival (Amaar &

Moore 1998). For the *N. crassa* mutation rate estimation, so-called recessive lethals recovered from heterokaryons were used (Auerbach 1959), although it has been suggested that this approach could have lead to erroneously high mutation rate estimations for *N. crassa* (Auerbach 1959; Drake 2009). Nevertheless, because the evolutionary forces acting on these markers are markedly different (Amaar & Moore 1998; Winstanley *et al*. 2009), the rates at which mutations accumulate in each are probably drastically different.

Mutant frequency refers to the fraction of the population that acquired a certain mutation. Mutant frequency is often used instead of mutation rate to study the diversification of organisms (Couce & Blazquez). Although still informative the utilization of mutant frequencies does not allow for predictions of mutation acquisition per generation. Determining the mutant frequency is the first step in conducting a fluctuation analysis and was calculated as 0.02 ($1 - P_o$) for this specific study. This is similar to what was found in a study on *Escherichia coli* where the spontaneous generation of auxotrophic mutants were studied. Here it was determined that a fraction close to 0.02 of wild type colonies grown on nutrient agar mutated to become an auxotroph (Goldstein & Smoot 1955). Similarly in a study by Alvarez-Perez the rate at which elastase activity arises spontaneously in *A. fumigatus* was studied. Of the population only 2% was found to gain elastase activity spontaneously thus the mutant frequency is 0.02. The mutant frequency estimated in this study thus is similar to what has been found in previous studies.

Two important factors could potentially have influenced the *F. circinatum* mutation rate determined in this study. Firstly, the experimental conditions used do not reflect conditions naturally encountered by this fungus. In its natural habitat *F. circinatum* will experience variations in, for example, nutrient availability and temperature, all of which will induce stress which in turn will increase the spontaneous mutation rate by inducing mutagenesis and recombination (Alvarez-Perez *et al*. 2010; Lamb *et al*. 2008). In fact, stress in the form of irradiation, temperature and drought have been shown to increase the mutation frequencies of *Penicillium lanosum* and *A. niger* 4- and 6-fold, respectively (Lamb *et al*. 2008). Secondly, the growth characters of *F. circinatum*, which involves different forms of conidiation (i.e., the production of microconidia and macroconidia) (Subramanian 1971), could potentially obscure correlations between the number of mutants observed and the actual number of mutations that occurred. For example, multiple conidia are typically formed from a single conidiogenous cell (Van Wyk *et al*. 1991) and the nuclei of macroconidia are formed by multiple divisions of the nucleus in the developing macroconidia (van Wyk *et al*. 1988).

Thus, if the mutation occurred in the conidiogenous cell, more mutants will be observed than when the mutation occurred in one of the progeny spores. Therefore, both these factors could potentially have caused the mutation rate determined to be significantly over- or underestimated.

Another factor potentially affecting the accuracy of the *F. circinatum* mutation rate determined here is the specific generation time-value that was employed for calculating it. Here, the generation time for *F. circinatum* was assumed to be similar to that of a related *Fusarium* species (i.e., "*F. moniliforme*") determined previously (Macris & Kokke 1977). This is because estimations of generation time of fungi are generally difficult due to their non-homogenous or thread-like growth (Meletiadis *et al*. 2001). Consequently, the spectrometry-based methods typically used for quantifying generation time of unicellular organisms, generally produce erroneous results due the clumping of fungal mycelia (Joubert *et al*. 2010). Although alternative methods for determining generation time in fungi are available (e.g., using colony expansion rates and microscopic measurements) these are usually tedious and time consuming and not suitable for large and complex studies involving liquid cultures. Furthermore, it is unclear how *F. circinatum* is related to the "*F. moniliforme*" strain originally used to study growth rate (Macris & Kokke 1977). This is because the obsolete taxon "*F. moniliforme*" refers to numerous distinct species (Seifert *et al*. 2003), which reduces the feasibility of extrapolating the growth characteristics of the fungus used in this initial study to *F. circinatum*. Nevertheless, generation time has a significant influence on mutation rate estimations. For example, with "*Fusarium moniliforme*" the mutation rate of an isolate with a generation time of 3.15 h would be almost three times less than an isolate with a generation time of 3.53 h. Future research should thus seek to determine more accurately the generation time of *F. circinatum* under the same conditions used to perform fluctuation analyses.

No differences between the *nit3* gene sequences of FSP34, the mutant parent, the revertants and the mutant progeny were observed. However, the normal functioning of the *nit3* gene product can be affected by mutations in regions other than the *nit3* coding region. In *Aspergillus*, for example the phenotype of *nit3* mutants correlates to that of *nirA* mutants (Klittich & Leslie 1989). In this fungus, the *nirA* homolog of *nit3* is a positive regulator that binds DNA with zinc finger domains (Johnstone *et al*. 1990), and activates transcription allowing mediation of nitrate and nitrite metabolism (Burger *et al*. 1991a). In turn, the *nirA* gene is regulated by the gene product of *areA* (Johnstone *et al*. 1990) and a putative

regulatory region located upstream of the *nirA* coding region (Burger *et al*. 1991b). Therefore, mutations in these regions required for normal functioning of the *nirA* gene product might alter normal functioning of the *nirA* gene product and could possibly explain why no mutations could be found in the *nit3* gene of the *nit3* mutants. An alternative explanation involves the receptor binding sites for the *nit3* transcription factor in the promotors of the regulated genes (Tudzynski *et al*. 1996). A mutation in one or more of these binding sites would also affect the functioning of the nit3 gene product leading to the mutant phenotype.

Mutation rates of higher eukaryotes, measured as mutations/bp/genome duplication, is much lower (on average four orders of magnitude) than what was estimated for *F. circinatum* in this study. The mutation rates of multicellular eukaryotes such as *Caenorhabditis elegans*, *Drosophila melanogaster*, *Mus musculus* and *Homo sapiens* are generally in the same order of magnitude (Drake 1999). However, mutation rate comparison between multicellular and less complex or unicellular eukaryotes are difficult, because mutation rates of multicellular eukaryotes are commonly measured per sexual generation and not cell division (Drake 1999, Guo *et al.* 2004). Also the majority of the genome of a typical multicellular eukaryote has no apparent function compared to those of less complex eukaryotes and if these non-coding regions are included in or used for mutation rate analyses, overestimates might be obtained for multicellular eukaryotes. This is because lower levels of purifying selection will act on these regions that are much more abundant in multicellular eukaryotes (Drake 1999). Similar differences in the evolutionary forces acting on different regions of the genome probably also account for the order of magnitude differences in mutation rate that have been determined for *C. elegans* using multiple gene specific markers (Baer 2007). This is probably also true for the estimates for human where mutation rate estimated using sequences of twenty Mendelian disease loci yielded a value two orders of magnitude higher (Kondrashov, 2002) than what has been recorded by Drake (1999). Overall however, mutation rate of multicellular eukaryotes appears much lower than that of filamentous (6.17×10^{-8} - 8.2×10^{-6} mutations/genome/generation) (Drake 2009) and unicellular (0.0037 - 0.0044 mutaitons/genome/generation) (Drake 2009) fungi.

Mutation rate data is important for many reasons and will allow many future studies addressing many questions. Mutation rate data on different genome regions will give insight into how different regions of the genome evolve. Similarly mutation rate between different species can be compared as to study how the evolutionary rate (driven by mutations) of the

species compare. Furthermore, the mutation rate of different organelles within the eukaryotic cell can be studied separately to study their diversification. This opens up opportunities to study how mutation rates differ between organelles in fungi. It is known that the mutation rate of mitochondrial genome in animal cells is much higher (Lynch *et al.* 2006) than nuclear genome whereas the opposite seems to be true for plant cells (Wolfe *et al*. 1987). Furthermore, by knowing the rate at which mutations arise and diversification occurs in different genomic regions one can predict the probability that a certain phenotype will arise which will represent valuable information for epidemiological studies.

This study aimed to estimate the mutation rate of *F. circinatum*. Even though the calculated value is most probably not an accurate estimation of what is happening under natural conditions, it gives an idea of how the mutation rate of *F. circinatum* compares to that of other species. Our results show that the spontaneous mutation rate of *F. circinatum* per base pair is comparable to that of *A. nidulans*. As the type of gene used in the fluctuation analysis will influence the mutation rate value the difference between the mutation rates of *F. circinatum* and *A. fumigatus* might be due to the choice of genes. Future work will be aimed at the use of different genes to estimate the mutation rate of a species from a representative set of genes or even whole genome comparisons that can be repeated within and between species.

References

Alvarez-Perez S, Blanco JL, Lopez-Rodas V, Flores-moya A, Costas E, Garc ME, 2010. Elastase Activity in *Aspergillus fumigatus* can arise by random , spontaneous mutations. *International Journal of Evolutionary Biology*: doi: 10.4061/2010/602457.

Amaar YG, Moore MM, 1998. Mapping of the nitrate-assimilation gene cluster (crnA-niiA-niaD) and characterization of the nitrite reductase gene (niiA) in the opportunistic fungal pathogen *Aspergillus fumigatus*. *Current Genetics* **33**: 206–215.

Auerbach C, 1959. Spontaneous mutations in dry spores of *Neurospora crassa*. *Zeitschrift fur Vererbungslehre* **90**: 335–346.

Baer CF, Miyamoto MM, Denver DR, 2007. Mutation rate variation in multicellular eukaryotes : causes and consequences. *Nature* **8**, 619–632.

Baracho MS, Baracho IR, 2003. An analysis of the spontaneous mutation rate measurement in filamentous fungi. *Genetics and Molecular Biology* **26**: 83–87.

Bowden R, Leslie J, 1992. Nitrate-nonutilizing mutants of *Gibberella zeae* (*Fusarium graminearum*) and their use in determining vegetative compatibility. *Experimental Mycology* **16**: 308–315.

Brody H, Carbon J, 1989. Electrophoretic karyotype of *Aspergillus nidulans*. *Proceedings of the National Academy of Science 86*: 6260–6263.

Burger G, Strauss J, Scazzocchio C, Lang BF, 1991. *nirA* , the pathway-specific regulatory gene of nitrate assimilation in *Aspergillus nidulans* , encodes a putative GAL4-type zinc finger protein and contains four introns in highly conserved regions. *Molecular and Cellular Biology* **11**: 5746–5755.

Burger G, Tilburn J, Scazzocchio C, 1991. Molecular cloning and functional characterization of the pathway- specific regulatory gene *nirA* , which controls nitrate assimilation in *Aspergillus nidulans*. *Molecular and Cellular Biology* **11**: 795–802.

Couce A, Blazquez J, 2011. Estimating mutation rates in low-replication experiments. *Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis* **714**: 26-32.

De Vos L, Myburg A, Wingfield M, Desjardins A, Gordon T, Wingfield B, 2007. Complete genetic linkage maps from an interspecific cross between *Fusarium circinatum* and *Fusarium subglutinans*. *Fungal Genetics and Biology* **44**: 701–714.

Denning DW, Anderson MJ, Turner G, Latgé J, Bennett JW, 2003. Sequencing the *Aspergillus fumigatus* genome. *The Lancet Infectious Diseases* **2**: 251–253.

Drake JW, 1991. A constant rate of spontaneous mutation in DNA-based microbes. *Proceedings of the National Academy of Science* **88**: 7160–7164.

Drake JW, 1999. The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. *Annals of the New York Academy of Sciences* **18**: 100-107.

Drake JW, 2009. Avoiding dangerous missense : thermophiles display especially low mutation rates. *PLoS Genetics* **5**: 1–11.

Foster PL, 2006. Methods for determining spontaneous mutation rates. *Methods in Enzymology* **409**: 1–16.

Fox A, Tuch B, 2008. Measuring the prevalence of regional mutation rates: an analysis of silent substitutions in mammals, fungi, and insects. *BMC Evolutionary Biology* **13**: 1–13.

Goldstein A, Smoot JS, 1955. A strain of *Escherichia coli* with an unusually high rate of auxotrophic mutation. *Journal of Bacteriology* **70**: 588-593.

Guo HH, Choe J, Loeb LA, 2004. Protein tolerance to random amino acid change. *Proceedings of the National Academy of Science* **101**: 9205–9210.

Jargeat P, Rekangalt D, Verner MC, Gilles G, Debaud JC, Marmeisse R, Fraissinet-Tachet L, *et al*., 2003. Characterisation and expression analysis of a nitrate transporter and nitrite reductase genes, two members of a gene cluster for nitrate assimilation from the symbiotic basidiomycete *Hebeloma cylindrosporum*. *Current Genetics* **43**: 199–205.

Johnstone I, McCebe P, Greaves P, Gurr S, Cole G, Brow M, Unkles S, 1990. Isolation and characterisation of the *crnA-niiA-niaD* gene cluster for nitrate assimilation in *Aspergillus nidulans*. *Gene* **90**: 181–192.

Joubert A, Calmes B, Berruyer R, Pihet M, Bouchara J, Simoneau P, Guillemette T, 2010. Laser nephelometry applied in an automated microplate system to study filamentous fungus growth. *BioTechniques* **48**: 399–404.

Klittich C, Leslie JF, 1989. Chlorate-resistant, nitrate-utilizing (*crn*) mutants of *Fusarium moniliforme* (*Gibberella fujikuroi*). *Journal of General Microbiology* **135**: 721–727.

Kondrashov AS, 2002. Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Human Mutation* **27**: 12–27.

Lamb BC, Mandaokar S, Bahsoun B, Grishkan I, Nevo E, 2008. Differences in spontaneous mutation frequencies as a function of environmental stress in soil fungi at “‘ Evolution Canyon ,’” Israel. *Proceedings of the National Academy of Science* **105**: 5792–5796.

Leslie JF, Summerell BA, Bullock S, 2006 Vegetative Compatibility Groups. In: *The Fusarium Laboratory Manual*: 31-43. Blackwell Publishing, New Jersey, USA.

Luria S, Delbruck M, 1943. Mutations of bacteria from virus sensitivity to virus resistance. *Genetics* **28**: 491–511.

Lynch, M., Lynch, M., Koskella, B., & Schaack, S, 2006. Mutation pressure and the evolution of organelle genomic architecture. *Science*, **311**, 1727–1730

Macris B, Kokke R, 1977. Kinetics of growth and chemical composition of *Fusarium moniliforme* cultivated on carob aqueous extract for microbial protein production. *European Journal of Applied Microbiology* **4**: 93–99.

Maki H, 2002. Origins of spontaneous mutations : specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. *Annual Reviews in Genetics* **36**: 279–303.

Meletiadis J, Meis JFGM, Mouton JW, Verweij PE, 2001. Analysis of growth characteristics of filamentous fungi in different nutrient media. *Journal of Clinical Microbiology* **39**: 478–484.

Mitchell RG, Steenkamp ET, Coutinho TA, Wingfield MJ, 2011. The pitch canker fungus, *Fusarium circinatum*: implications for South African forestry. *Southern Forests* **73**: 1-13.

Morin PA, Gordon L, Wayne RK, 2004. SNPs in ecology, evolution and conservation. *Trends in Ecology and Evolution* **19**: 208-216.

Rogozin IB, Babenko VN, Milanesi L, Pavlov YI, 2003. Computational analysis of mutation spectra. *Briefings in Bioinformatics* **4**: 210–227.

Seifert KA, Aoki T, Baayen RP, Brayford D, Burgess LW, *et al*., 2003. The name *Fusarium moniliforme* should no longer be used. *Mycological Research* **107**: 643-644.

Subramanian CV, 1971. The phialide. In: *Taxonomy of the Fungi Imperfecti* (ed. W. B. Kendrick): 92-115. Toronto: University of Toronto Press.

Steenkamp E, Wingfield B, Coutinho T, Wingfield M, Marasas WFO, 1999. Differentiation of *Fusarium subglutinans* f . sp . pini by histone gene sequence data. *Applied and Environmental Microbiology* **65**: 3401–3406.

Stewart FM, Gordon DM, Levin BR, 1990. Fluctuation analysis: the probability distribution of the number of mutants under different conditions. *Genetics* **124**: 175-185.

Tudzynski B, Mende K, Weltring K, Kinghorn J, Unkles S, 1996. The *Gibberella fujikuroi niaD* gene encoding nitrate reductase: isolation, sequence, homologous transformation and electrophoretic karyotype location. *Microbiology* **142**: 533–539.

van Wyk P, Venter E, Wingfield M, Marasas W, 1988. Nuclear division and septation in macroconidia of *Fusarium crookwellense*. *South African Journal of Botany* **54**: 118–122.

van Wyk P, Wingfield M, Marasas WFO, Bosman J, 1991. Development of microconidia in *Fusarium* section Sporotrichiella. *Mycological Research* **95**: 284–289.

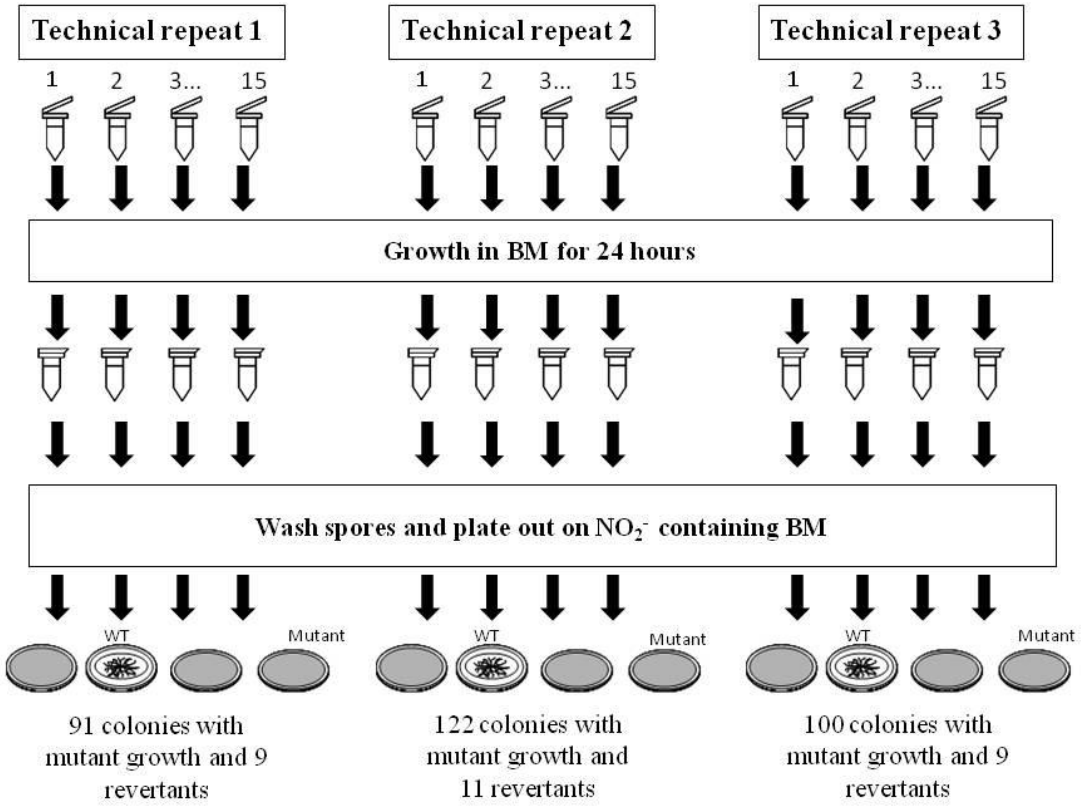
Wingfield BD, Steenkamp ET, Santana Q, Coetzee MP, Bam S, *et al*,. 2012. First fungal genome sequence from Africa : A preliminary analysis. *South African Journal of Science* **108**: 1–9.

Wingfield MJ, Hammerbacher A, Ganley RJ, Steenkamp ET, Gordon TR, Wingfield BD, Coutinho TA, 2008. Pitch canker caused by *Fusarium circinatum* - a growing threat to pine plantations and forests worldwide. *Australasian Plant Pathology* **37**: 319–334.

Winstanley C, Langille MGI, Gothergill JL, Kukavica-Ibrulj I, Paradis-Bleau C, *et al*., 2009. Newly introduced genomic prophage islands are critical determinants of *in vivo* competitiveness in the Liverpool epidemic strain of *Pseudomonas aeruginosa*. *Genome Research* **19**: 12-23.

Wolfe KH, Li WH, Sharp PM, 1987. Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. *Proceedings of the National Academy of Science* **84**: 9054-9058.

Figure 1. Diagrammatic representation of the experimental design (modified from Alverez-Perez *et al*. 2010) and results of this study. A total of 15 microfuge tubes with spore suspensions (N_0 = 1500 spores/ml media) were prepared for each repeat. The inoculum (N_0) was grown under non-selective conditions (24h) in BM after which the spores (N_t) were harvested and cultured on selective media (nitrite containing BM). The number of wild type and mutant colonies was counted. Progeny showing mutant growth (the grey plates) produced sparse growth with no aerial mycelia because the isolates have a *nit3* mutation making them incapable of utilizing nitrite as sole nitrogen source. Progeny representing spontaneous revertants displayed the wild type phenotype with profuse growth with aerial mycelia due their ability to utilize nitrite as a sole nitrogen source.



Summary

Spontaneous mutagenesis can be divided into three main steps: the introduction of DNA damage and lesions, damage recognition and DNA repair. All sources of spontaneous mutagenesis originate from within the cell itself, e.g., polymerase errors cause DNA mismatches and reactive oxygen species alter the chemical composition of DNA bases. The combined effects of all these processes influence spontaneous genomic mutation rates, which are thought to be a characteristic of individual species and/or groups of species. Although much is known about different mutagens and how they cause mutations the sequence context of these mutations are less well understood. The results of this MSc study on mutation in the filamentous fungus *Fusarium circinatum* showed that the 5′ and 3′ neighbouring bases of a single nucleotide polymorphism can significantly influence the type of substitution that occurred leading to the formation of mutational motifs. This was the case for both sets of genes examined (core housekeeping and non-ribosomal protein synthetase genes), whose evolution is known to differ. The fact that none of the identified motifs are shared between the two sets of genes could indicate that the cellular mutagens and/or repair machinery function differently for the two gene groups. Furthermore, none of the mutable motifs that have been identified for the well-known mutagens in model organisms could be detected in the fungus, which suggests that mutagens and/or DNA repair mechanisms of this fungus are unique. Although limited information is available for non-model eukaryotes, an estimate for the rate at which mutations arise across the genome of *F. circinatum* could be a good starting point for comparisons of its evolutionary rate to those of its close relatives. This was accomplished using a fluctuation analysis involving nitrate non-utilizing mutation reversion. Although mutation rate determined in this study is probably not precisely accurate, it represents a good starting point for future comparative studies on the evolutionary rate of *Fusarium* species. As a whole this study laid the foundation for a better understanding of spontaneous mutagenesis at specific sites in certain groups of genes as well as across the genome of the economically important plant pathogen *F. circinatum*.