Intron architecture in *Fusarium*

by

Mmatshepho Malekgale Phasha

Submitted in partial fulfillment of the requirements for the degree
Magister Scientiae (Microbiology)

In the faculty of Natural and Agricultural Sciences
University of Pretoria
Pretoria
(30 April 2012)

Supervisor: Prof. E.T. Steenkamp
Co-supervisor: Prof. B.D. Wingfield
Co-supervisor: Dr. M.P.A. Coetzee

Declaration

I Mmatshepho Malekgale Phasha 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

Declaration i
Acknowledgements iv
Preface v
Chapter 1: Literature review 1
1. Introduction 2
2. Intron Types 3
2.1 Self-splicing introns 4
2.1.1 Distribution over the TOL 4
2.1.2 Intron structure and splicing mechanisms 4
2.2 Spliceosomal introns 5
2.2.1 Distribution over the TOL 5
2.2.2 Intron structure 6
2.2.3 The Spliceosome and splicing 6
2.2.4 Mechanisms that define the region to be spliced 7
2.2.5 Genomic density and spread of Spliceosomal introns 8
2.2.6 Within-gene position and density of Spliceosomal introns 9
3. Intron Origins 12
3.1 The Introns-early theory 12
3.2 The Introns-late theory 13
4. Intron Functions 14
4.1 Alternative Splicing 15
4.1.1 The mechanisms of alternative splicing 15
4.1.2 The origin of alternative splicing 17
4.2 Intron-Mediated Enhancement 18
5. Future Prospects 19
6. References 20
7. Figures 27
Chapter 2: The architecture and distribution of introns in housekeeping genes of four *Fusarium* species 31
Abstract 32
Introduction 33
Materials and methods 35
Fusarium genomes 35

Identification of *Fusarium* HK Genes 35
Identification and annotation of introns 35
Analysis of intron architecture and distribution 36
Results 39
Identification of *Fusarium HK* genes 39
Identification and annotation of introns 39
Analysis of intron architecture and distribution 41
Discussion 46
References 53
Tables 58
Figures 60
Chapter 3: *In silico* identification and characterization of alternative splicing in the *Fusarium circinatum* genome 68
Abstract 69
Introduction 70
Materials and methods 73
Fusarium circinatum cDNA library construction and sequencing 73
Identification of genes showing alternative splicing 74
Architecture of introns in genes showing alternative splicing 75
Identification of homologues of genes with alternative splicing signals in the other *Fusarium* species 76
Identification of nonsense-mediated mRNA decay (NMD) genes 76
Results 77
Identification of genes showing alternative splicing 77
Analysis of genes showing alternative splicing 78
Identification of homologues of genes with alternative splicing signals in the other *Fusarium* species 81
Identification of NMD genes 81
Discussion 82
References 89
Tables 94
Figures 101
Summary 105
Appendix 107

Acknowledgements

I would like to acknowledge the following people and funding bodies:

- My primary supervisor Prof. Emma Steenkamp for her amazing support throughout my Masters degree and for teaching me how to improve my analytical skills and scientific reasoning.

- My co-supervisor Prof. Brenda Wingfield for her eye-opening discussions.

- My co-supervisor Dr. Martin Coetzee for his contribution into improving my writing skills and for his immediate help always.

- My lab-mates and Vivienne Clarence for their emotional and technical support throughout my Masters degree.

- My friends for their constant encouragement towards the completion of my dissertation.

- The University of Pretoria, the National Research Foundation (NRF), the Forestry and Agricultural Biotechnology Institute (FABI), the Tree Protection Cooperative Programme (TPCP), and the Mandela Rhodes foundation for the financial support that enabled my Masters research.

- My family for *everything* that they represent and the beautiful support they have given, and are still giving, me no matter the cost.

- My Wonderful Heavenly Father for all that I am and are achieving:

"For the LORD giveth wisdom, out of his mouth cometh knowledge and understanding."
Proverbs 2:6

Preface

Fusarium is a fungal genus that constitutes pathogenic and non-pathogenic species. Pathogenic species cause a variety of diseases in humans, animals and plants. The genomes of four economically important plant pathogenic *Fusarium* species have been sequenced. These include the maize pathogen *F. verticillioides*, the tomato pathogen *F. oxysporum*, *F. graminearum*, which is pathogenic to wheat and *F. circinatum* that is pathogenic to pine.

Different gene prediction software have been developed for the annotation of bacterial and eukaryotic genomes. Such software generally includes algorithms that locate the different elements of a gene, which may include, among others, exons and introns. Unlike exons, introns do not form part of mature mRNAs that are translated into proteins but have been very useful in improving the annotation of genomes. What makes introns useful in the annotation process is that fact they have *cis*-elements that define and differentiate them from exons. In addition, the position of introns in genes is highly conserved between closely related organisms such that the annotation of a newly sequenced genome can be carried out using an already annotated genome of a related organism as a reference. However, the general structure of intronic *cis*-elements is not conserved across the eukaryotic Tree of Life (TOL) and much variation exists – sometimes even among closely related organisms. Therefore, an adequate understanding of the architecture of introns in a specific group of organisms is crucial for accurate and efficient genome annotations.

Chapter 1 provides a review of the available literature on introns, particularly Spliceosomal introns found in eukaryotic protein-coding genes. The types of introns found across the TOL, as well as the process utilized to exclude them from premature mRNA are discussed. Theories seeking to explain the origin of these DNA sequences are also reviewed. Finally the known functions of introns are considered. Throughout the Chapter differences between higher eukaryotes and lower eukaryotes are highlighted.

In **Chapter 2**, the architecture and distribution of introns in *F. verticillioides*, *F. oxysporum*, *F. graminearum* and *F. circinatum* were investigated. This was done using a set of 458 core genes that are standard to all eukaryotic genomes. Intron *cis*-elements, intron frequency per gene, intron length and intron distribution were compared between the four *Fusarium* species. The research in this Chapter also aimed to generate a set of refined consensus sequences for the various *cis*-elements to potentially allow their incorporation into genome gene prediction software to improve current and future annotations.

In **Chapter 3**, genes that undergo alternative splicing in the genome of *F. cricinatum* were analyzed. Genome and EST data for *F. cricinatum* were utilized to identify these genes. They were then analysed by comparing their *cis*-elements to those of constitutive introns within the same genes. An analysis was also performed to understand the extent to which non-sense mediated mRNA decay (NMD) could have influenced the data. This was accomplished by determining whether these fungi actually encode the known proteins involved in this process and by searching for the presence of premature stop codons in alternative transcripts, which represent the substrates for the NMD system.

Chapter 1: Literature review

1. Introduction

In 1977, two research groups working on Adenoviruses both discovered introns. The research groups, led by Sharp and Roberts, independently showed that the Adenovirus genome contains regions that do not form part of the mature mRNA. They were investigating the expression of Adenovirus late genes when they realised that the mRNAs were "mosaic molecules consisting of sequences complementary to several non-contiguous segments of the viral genome" (Sharp and Roberts, 1977). This discovery was made for the *hexon* gene, which encodes the major viral capsid. In 1978, Walter Gilbert called the regions that do not form part of the mature mRNA introns.

Introns are present in the genomes of all forms of life - Eukaryota, Bacteria, Archaea and viruses. They represent different types that have been categorized based on their structures and the mechanisms used to splice them from genes. The three main intron-types are Group I, Group II and Spliceosomal introns. Group I introns occur in the eukaryotic, bacterial and archaeal nuclear genomes, as well as in viruses and eukaryotic plastid genomes (Keating *et al*., 2000; Saldanha *et al*., 1993). Group II introns have a similar distribution except that they usually do not occur in eukaryotic nuclear genomes (Keating *et al*., 2000; Saldanha *et al*., 1993). Spliceosomal introns are exclusively found in eukaryotic nuclear genomes (Jeffares *et al*., 2006).

The origin of introns is the subject of on-going debate in molecular-evolutionary biology. Two opposing theories (*i.e.*, introns-early and introns-late) have been proposed to explain how introns originated and both are supported by experimental data. Whether introns were present before or emerged after the diversification of Life remains unclear, but what has emerged is that introns apparently have functions (Saldanha *et al*., 1993). One would expect the loss of all or most of these introns after their initial appearance in genomes if they did not render the cell some advantage. For example, Spliceosomal introns are thought to have evolved from Group II introns (Koonin, 2006) and they have been associated with alternative splicing and intron mediated enhancement. The phenomenon of alternative splicing was proposed shortly after the discovery of introns when Walter Gilbert realised that different combinations of the protein coding regions,

called exons, of the *hexon* gene could lead to the production of different mature mRNA molecules (Gilbert, 1978). These mRNA molecules would then be translated into functionally different proteins. The phenomenon of intron mediated enhancement was first discovered in 1990 when Mascarenhas *et al*. investigated the enhancing effect of two maize specific introns on the expression of bacterial chloramphenicol acetyl transferase gene. They reported stimulation of expression enhanced by the two maize introns. Since their discovery, alternative splicing and intron mediated enhancement has been observed in many eukaryotes (Conti and Izaurralde, 2005; Parra *et al*., 2011).

Some areas of intron research have received more attention than others. Moreover, more research has been done in certain organisms than in others. An area like the origin of introns has received much attention, but much remain to be researched before we would be able to reject one or more or the prevailing evolutionary hypotheses. Also, despite knowledge of the possible functions of Spliceosomal introns, the mechanisms that govern constitutive and alternative splicing of these introns are not well understood. This review serves to highlight what is known and understood about the introns of Eukaryota, specifically fungi. Therefore, fungal examples are given according to information availability and relevant comparisons are provided between so-called lower eukaryotes (e.g., fungi) and are compared to higher eukaryotes (e.g., animals).

2. Intron Types

The three main intron types (Group I, Group II and Spliceosomal introns) are categorized according to where they are found in the Tree of Life (TOL), their structure and their splicing mechanisms. Group I and Group II introns splice themselves out of genes or genomic regions. Spliceosomal introns, in contrast, require the assistance of a number of proteins for their splicing (Saldanha *et al*, 1993).

2.1 Self-splicing introns

2.1.1 Distribution over the TOL

Group I and Group II introns are widespread among the different organisms in the TOL. Self-splicing was first discovered in the rRNA intron of *Tetrahymena thermophila* (Kruger *et al*., 1982), and following this discovery many Group I introns have been found in diverse eukaryotes and bacteria (Cannone *et al*., 2002). About 2300 Group I introns had been sequenced from these organisms by 2008 (Vicens*et al*., 2008). Group II introns commonly occur in the plastids of eukaryotes and are abundant in bacteria and archaea (Keating *et al*., 2000). They are also found in the organellar genes of yeasts and plants (Haugen *et al*., 2005). Group II introns are thought to have entered the Eukaryota through the early bacterial endosymbiois events from which cellular organelles evolved (Rest and Mindell, 2003).

2.1.2 Intron structure and splicing mechanisms

Group I introns have a conserved core made up of 10 paired segments (P1-P10) organized in three domains (Figure 1) (Vicens *et al*., 2008). These core domains are stabilized by peripheral elements forming long range tertiary interactions resulting in the promotion of splicing (Vicens *et al*., 2008). In some Group I introns this is dependent on protein co-factors . Two examples of such co-factors are the tyrosyl-tRNA synthetase (CYT-18) in *Neurospora crassa*, and cytochrome b pre-mRNA processing protein 2 (CBP) in yeast mitochondria. These co-factors interact with the peripheral elements of the Group I introns causing the RNA to fold into an active catalytic structure (Vicens *et al*., 2008).

Group I introns employ a splicing mechanism that involves two transesterification reactions (Saldanha *et al*., 1993). The folding of the intron brings its 5′ and 3′ splice site and external Guanosine into proximity (Saldanha *et al*., 1993). During splicing, the external Guanosine attacks the 5′ splice site with its 3′ OH and displaces the 5′ exon (Saldanha *et al*., 1993). The 5′ exon with its free 3′ OH then reacts with the 3′ splice site thereby displacing the 3′ exon. Finally, the 5′ and 3′ exons are ligated together to form an uninterrupted coding sequence (Saldanha *et al*., 1993). In *T. thermophila* it has been shown that the formation and cleavage of

phosphodiester bonds are dependent on Mg^{2+} (Saldanha *et al*., 1993; Michel *et al*., 1990; Yarus, 1993).

Group II introns are made up of six domains (Saldanha *et al*., 1993; Figure 1). Among them, domain 5 (D5) is the most conserved and the most important for splicing. D5 resembles and behaves like U6, a conserved Spliceosomal RNA involved in splicing during nuclear pre-mRNA splicing (see below). D5 is therefore a vital link to a possible common ancestor between Group II introns and the spiceosomal machinery. (Keating *et al*., 2009)

The splicing mechanism of Group II introns is similar to that of Group I introns and also resembles that of Spliceosomal introns. Splicing of Group II introns is initiated by folding, when an internal Adenosine attacks the Guanosine in the 5′ splice site with its 2′ OH, which results in the formation of a phosphodiester bond (Saldanha *et al*., 1993). This has been coined the lariat structure and is similar to that formed during the splicing of Spliceosomal introns. Similar to the splicing of Group I introns, this reaction displaces the 5′ exon and then the free 3′ OH of the 5′ exon attacks the 3′ splice site and displaces it (Saldanha *et al*., 1993). The final step involves the ligation of the two displaced exons (Saldanha *et al*., 1993).

2.2 Spliceosomal introns

2.2.1 Distribution over the TOL

Spliceosomal introns occur in the genomes of all members of the Eukaryota but are not found in those of Bacteria and Archaea (Jeffares *et al*., 2006). These introns interrupt the coding regions of nuclear genes (Nguyen *et al*., 2006). In general, the proteins and intron signatures involved in the splicing of Spliceosomal introns, and the different mechanisms employed to carry out this process are well defined. However, the origin of these introns is currently vague. A better understanding of the evolution of such introns is essential in order to interpret the evolution of eukaryotic genomes due to the fact that introns and genomes evolution is connected (Nguyen *et al*., 2006) e.g., introns have contributed to the size expansion of eukaryotic genomes.

2.2.2 Intron structure

For an intron to be classified as spliceosomal, there are certain elements it must harbour (Ast, 2004). In the search for characteristics that define a spliceosomal intron, only four sequence motifs or signatures have been found. These include the 5′ and 3′ intron-exon junctions or the so-called the 5′ and 3′ splice sites, the branch-point-containing branch site (usually found at least 10-20 nucleotides from the 3′ splice site), and the polypyrimidine tract (usually found between the branch site and the 3′ splice site). The 5′ and 3′ splice sites contain sequences of varying lengths located at the exon-intron junctions.

Genome-wide sequence comparisons combined with expressed sequence tag (EST) data has allowed for the characterization of the intron elements in five representative fungi (Kupfer *et al*., 2004). In addition to being shorter than those of higher eukaryotes, approximately 98% of fungal introns examined by Kupfer *et al.* (2004) belonged to the common splice site class (5′GU…..AG3′) (Kupfer *et al*., 2004; Figure 2). The polypyrimidine tracts of these introns were predominantly located between the 5′ splice site and the branch point, as opposed to being confined to the region between the branch point and the 3′ splice site, as is the case in higher eukaryotes (Kupfer *et al*., 2004; Figure 2).

2.2.3 The Spliceosome and splicing

All four of the sequence motifs or signatures that characterize Spliceosomal introns are involved in the splicing process. In addition, splicing of Spliceosomal introns from the pre-mRNA requires the action of the spliceosomal complex or spliceosome. The spliceosome consists of five small nuclear RNA (snRNA) molecules (U1, U2, U4, U5 and U6) and over 150 proteins (Jurica and Moore, 2003). One snRNA assembles with many proteins that forms small complexes, which are called small nuclear ribonucleoprotein complexes (snRNPs) (Ast, 2004). These snRNPs in turn attach to and excise the Spliceosomal introns, and also assist in joining together the flanking exons (Staley and Guthrie, 1998).

Each round of splicing requires the binding and release of the snRNPs. This happens in a sequential manner (Staley and Guthrie, 1998; Russell, 2006). The beginning of assembly is marked by the binding of U1 snRNP to the 5′ splice site of an intron *via* base-pairing. The U2 snRNP then binds to the branch site. U4 and U6 snRNP form a complex and then associate with the U5 snRNP, and together this complex binds to the U1 and U2 snRNPs already bound to the intron. This reaction causes the intron to form a loop which brings the two ends of the intron together. U4 snRNP then dissociates, and the remaining structure is an active spliceosome. The spliceosome then separates the first exon from the intron by cutting the intron at the 5′ spice site. The 5′ end of the intron is now free and it binds to the Adenosine nucleotide, called the branch point, in the branch site. The RNA lariat structure is stabilized by the formation of a phosphodiester bond between the 2′ OH of the branch point and the 5′ phosphate of the Guanosine nucleotide at the free 5′ end of the intron. The intron is then removed by cleavage at the 3′ splice site, and the second exon is ligated to the first.

2.2.4 Mechanisms that define the region to be spliced

Eukaryotes utilize different splicing mechanisms due to the specific features of their intron and/or exon (Kupfer *et al*., 2004; Deutsch and Long, 1999; Lim and Burge, 2001). This is particularly in terms of factors such as the amount of information at the intron 5′ and 3′ splice sites and at the exon regions flanking the intron, the branch point, and the positions of the polypyrimidine tracts (Kupfer *et al*., 2004). Dependent on these features, there are two prevalent splicing mechanisms for defining the region to be spliced. These are referred to as exon definition (ED) and intron definition (ID) (Berget, 1995).

In ED, splicing factors or splicing regulatory (serine-arginine rich; SR) proteins bind to degenerate sites on the exons and introns called splicing enhancer and silencers or to splice sites flanking an exon or an intron (McGuire *et al*., 2008). This can either promote or inhibit the binding of the spliceosome to the splice sites. ED is generally employed by higher eukaryotes, because their short exons require a highly specialized recognition mechanism for exon bridging (Berget, 1995; Romfo *et al*., 2000). ID is employed by lower eukaryotes, where the SR proteins bind to splice sites flanking an intron to facilitate splicing (McGuire *et al*., 2008). This allows the

short introns of lower eukaryotes that are flanked by long exons to be accurately recognized for excision (Talerico and Berget, 1994).

The regulation of splicing is directly linked to the mechanism of splicing through SR proteins. The manner in which SR proteins interact with the splicing machinery to promote their association with intronic splicing signatures is well studied (Kohtz *et al*., 1994; Zahler and Roth, 1995; Laviguer *et al*., 1993; Valcarcel *et al*., 1996; Wang et al., 1995; Zuo and Maniatis, 1996). By making use of electron microscopy it has also been demonstrated that SR protein mediated exon-exon interactions are highly specific and that SR proteins mediate the association between the 5′ and 3′ splice sites during splicing (Stark *et al.*, 1998). Furthermore, because the binding of the SR proteins facilitates either the promotion or inhibition of the binding of the spliceosome to respective intron or exon signatures, they are required for both alternative and constitutive splicing (see below).

2.2.5 Genomic density and spread of Spliceosomal introns

The density of introns varies widely in the genomes of Eukaryota. The human genome contains approximately 140 000 introns, the yeast *Saccharomyces cerevisiae* has about 253 introns, whereas the microsporidian fungus, *Encepalitozoon cuniculi*, has only 15 introns (Ast, 2004). This profound variation in intron density between eukaryotic genomes suggests that some introns were gained and some were lost during evolution (reviewed by Jeffares *et al*., 2006).

The increase of Spliceosomal intron density in a genome has been linked to their spread in genomes (Logsdon, 1998). There are two processes through which Spliceosomal introns spread in genomes. The spread can manifest either through duplication of a pre-existing intron or through transposable element/Group II intron insertion (Logsdon, 1998). In 1994 the first clear case of the spread of Spliceosomal introns *via* transposable element insertion was reported in a maize gene where an inserted transposon could be precisely spliced, thus reflecting the presence of splice signals on the transposon (Giroux *et al*., 1994).

The Ascomycetes lineage has somehow acquired many spliceosomal introns in the ribosomal RNA (rRNA) gene region relatively recently (Bhattacharya *et al*., 2000). These introns are thought to have been acquired through the insertion of "free" introns into rRNAs by the spliceosome (Bhattacharya *et al*., 2000), thus implying that the Ascomycetes spliceosomal machinery are interacting with the nucleolus (Bhattacharya *et al*., 2000). Furthermore, a sequence pattern (AG|G) similar to the proto site (MAG|R; Logsdon, 1998) has been found in Spliceosomal introns of the small- and large-subunit rDNA genes of a few Ascomycetes. This sequence has been postulated for intron insertion, supporting the introns-late theory (see below) (Bhattacharya *et al*., 2000). There are two possible explanations for the conservancy of the AG|G sequence pattern flanking the introns: (1) this motif is a favoured site for intron insertion or (2) the motif is a result of selection pressure after insertion so as to facilitate spliceosome-mediated intron excision (Bhattacharya *et al*., 2000).

Throughout evolution there have been gains and losses of introns. Intron gain and loss rates differ significantly among lineages (Jeffares *et al*. 2006) and can vary up to ten-fold between species (Jeffares *et al*. 2006). An analysis of orthologs from distantly related eukaryotes (the crown group and *Plasmodium*), conducted by Jeffares *et al*. (2006), showed intron gain rates which varied ten-fold. In contrast, rates of intron loss between the crown group and *Plasmodium* varied by eight-fold. The intron loss/gain ratio varied by more than 20-fold. This study suggests that the high intron density observed in some species is due to intron gain rather than an inherited state of intron rich ancestors (reviewed by Jeffares *et al*., 2006).

Eukaryotic intron evolution is a very dynamic process (Jeffares *et al*., 2006). Whether an intron will be lost or fixed following gain depends on the intron itself and the gene and organism in which the intron resides (Jeffares *et al*., 2006). The function of an intron also affects its evolutionary fate (Johnson, 2003). For example, introns found in genes involved in cell communication and enzyme regulation are more likely to be favoured by selection pressure so as to generate more protein variants through alternative splicing (Johnson, 2003). Different evolutionary rules apply to different introns (Jeffares *et al*., 2006).

2.2.6 Within-gene position and density of Spliceosomal introns

Depending on the location of an intron in the DNA sequence of a gene, an intron is considered to be in one of three phases: phase-0, phase-1 or phase-2. A phase-0 intron is located between two codons (Nguyen *et al*., 2006). Phase-1 and phase-2 introns are located within a single codon where a phase-1 intron is found between the first two and a phase-2 intron between the last two nucleotides of a codon (Nguyen *et al*., 2006). Intron phase appears to be conserved during evolution, because a change in intron phase requires simultaneous mutations that alter the intron's 5′ and 3′ ends in a complementary manner (Fedorov *et al*., 1992). Phase-0 introns occur most frequently while phase-1 introns are more frequent than phase-2 introns, making the distribution of intron phase non-uniform (Long *et al*., 1995; Long *et al*., 1998).

Within a gene, introns may display some positional bias. In *S. cerevisiae*, this was initially observed in the introns of protein coding genes that are predominantly located closer to the 5′ end (Fink, 1987). Later studies showed that this was not only true for intron-poor genomes (Sakuria *et al*., 2002; Mourier and Jeffares, 2003). During the course of evolution, introns closer to the 3′ end of some genes are lost (Roy and Gilbert, 2004), which was revealed by analyses of intron loss in 684 groups of orthologous genes from seven completely sequenced eukaryotic genomes (Roy and Gilbert, 2004). However, when Nielsen *et al*. (2004) conducted a study on four filamentous fungi (*N. crassa*, *Aspergillus nidulans*, *Magnaporthe grisea* and *Fusarium graminearum*), they found no positional bias of introns towards the 5′ end of genes. They observed intron loss in the middle of the genes in these fungi (Nielsen *et al*., 2004), reinforcing the idea of distinct patterns of intron loss among eukaryotes.

The introns of housekeeping genes have a 5′ positional bias (Lin and Zhang 2005), whether they occur in intron-rich or intron-poor genomes. These genes perform biological functions in the cell and their expression is generally constitutive. Because they are more highly expressed than any other type of genes in the genome, the 3′ intron loss bias of housekeeping genes is apparently more pronounced (Lin and Zhang 2005). In most Eukaryotes, the frequency of introns located at the 5′ end of the housekeeping genes has been found to be higher than that of introns located at the 3′ end (Lin and Zhang 2005).

Different mechanisms have been proposed for the observed positional bias of introns towards to the 5′ end of a gene. The most widely accepted is homologous recombination where recombination between a genomic copy and a spliced poly-adenylated mRNA leads to the loss of introns located closer to the 3′ end (Roy and Gilbert, 2004). Lin and Zhang (2005) postulated that introns could also be fixed at the 5′ end just after they have been gained. The mechanisms underlying intron gain and loss are, however, not clearly understood (Lin and Zhang, 2005).

3. Intron Origins

The origin of introns and the spliceosome, the spread of Spliceosomal introns, and the origin of alternative splicing have been popular topics of debate in evolutionary biology. These topics have been reviewed by several authors (Dibb and Newman, 1989; Giroux *et al*., 1994; Long *et al*., 1998; Logsdon, 1998; Bhattacharya *et al*., 2000; Roy *et al*., 2001; Jeffares *et al*., 2006; Nguyen *et al*., 2006; Roy and Gilbert). The literature points to at least two opposing models to explain intron evolution, *i.e.*, the introns-early theory and the introns-late theory. According to the introns-early theory introns already existed in the last common ancestor (LCA) of Eukaryota, Archaea and Bacteria where introns facilitated gene construction (Roy and Gilbert, 2005). Based on the introns-late theory, the genes of this LCA were intronless, and these elements appeared after the evolution of Eukaryotes (Nguyen *et al*., 2006). Currently neither of the theories is disregarded because experimental evidence for both has been presented (Jeffares *et al*., 2006; Nguyen *et al*., 2006).

3.1 The Introns-early theory

The introns-early theory explains the uneven distribution of introns by speculating that 35% of modern introns represent an inherited state (Roy, 2003; de Souza *et al*., 1998). Under this theory, these introns facilitated the assembly of the first gene(s) in the so-called “progenote” (Nguyen *et al*., 2006). This theory suggests that since exons are remnants of primitive mini genes, most of these ancient introns must lie between two codons, resulting in the current excess of phase-0 introns (Nguyen *et al*., 2006). However, why phase-1 introns are more frequent than phase-2 introns is not explained by this theory (Nguyen *et al*., 2006). Organisms with highly reduced genomes have lost the majority of their introns, probably due to selection pressure, suggesting that bacterial genomes might have originated from an ancestor that contained introns (Jeffares *et al*., 2006). If this is the case, it is expected that the spliceosome is an ancient part of the cell (Jeffares *et al*., 2006).

Should the introns-early theory be indeed correct, phase-0 introns should generally decrease over time as new introns are inserted into random positions. To explain the excess of phase-0 introns, Roy *et al*. (2001) performed a direct test by inferring the evolution of intron phase distributions from a dataset of 280 ancient genes. The authors divided the introns of these genes into two categories: lineage-specific introns and phylogenetically widely distributed introns to represent rough estimates of recently gained and ancestral introns, respectively (Roy *et al*., 2001). The authors found a stronger phase-0 bias ascribed to the ancestral introns. However, these findings were not supported by the results of another study that used a dataset of 79 apparently ancient ribosomal protein gene families from *Drosophila melanogaster*, *Anopheles gambiae*, *Caenorhabditis elegans*, *Homo sapiens*, *Schizosaccharomyces pombe*, *Arabidopsis thaliana*, and *Plasmodium falciparum* (Nguyen *et al*., 2006). The debate surrounding these findings and the introns-early theory remain vigorous (Roy and Gilbert, 2006; Martin and Koonin, 2006).

3.2 The Introns-late theory

According to the introns-late theory, intron insertions have given rise to Spliceosomal introns. This theory explains the non-uniformity of phase distribution by speculating that intron insertions occur at specific points in eukaryotic genes (Logsdon, 1998). A defined sequence pattern coined the "proto-splice site" has been the proposed site for intron insertions (Dibb and Newman, 1989; Long *et al*., 1998). A number of proto-splice sites have been proposed: MAG|R (Dibb and Newman, 1989); G|G, AG|G, AG|GT (Long *et al*., 1998); and MAG|GT (Dibb and Newman, 1989; Sverdlov *et al*., 2004) (where M = A or C, R = A or G, and the vertical line represents the intron-exon junction where an intron can be inserted).

A study conducted on representative eukaryotes to test for the origin of introns revealed an increasing proportion of phase-0 introns as a result of gained introns (Nguyen *et al*., 2006). Significant differences were found between phase distribution of gained and ancestral introns from the ancestor of the crown eukaryotes to *Arabidopsis thaliana* and also from the ancestor of the ecdysozoa to *C. elegans*, respectively (Figure 3). Differences between lost and ancestral introns were statistically insignificant. These results suggested that non-random intron insertions have resulted in the non-uniformity of phase distributions. However, it remains to be confirmed

that intron insertions have led to the observed distributions of intron phase (Long *et al*., 1998; Ruvinsky *et al*., 2005). As mentioned before, the AG|G sequence pattern similar to the proto site (MAG|R; Logsdon, 1998) found in Spliceosomal introns of the small- and large-subunit rRNA genes of Ascomycetes further supports the introns-late theory (Bhattacharya *et al*., 2000).

Both the introns-early and introns-late theories could be equally valid. The discrepancy of the introns-early theory lies in its failure to explain the reasons underlying phase-1 introns being more frequent than phase-2 introns. Should it be confirmed that the observed distributions of intron phase are due to intron insertions, the introns-late theory would be more believable.

4. Intron Functions

Group I, Group II and Spliceosomal introns are not merely interruptive sequences, they also have functions. As ribozymes, Group I and Group II introns can catalyse a variety of molecular reactions. These include endonucleolytic cleavage of RNA and DNA, RNA polymerization, nucleotide transfer, templated RNA ligation, and aminoacyl-ester cleavage (Cech, 1990; Piccirilli, 1992). Several Group I and Group II introns in yeast mtDNA also have open reading frames (ORF) (Saldanha *et al*., 1993), which encode maturases that function in splicing the intron encoding them (Burke, J.M. 1988; Lambowitz *et al*., 1990). Spliceosomal introns also render the eukaryotic cell services, *i.e.*, alternative splicing and intron mediated enhancement. These two functions of introns play a role in gene expression directly and indirectly, respectively, and will be further discussed below.

4.1 Alternative Splicing

Different combinations of exons can result in the production of multiple mRNA isoforms from one gene (Gilbert, 1978; Berget *et al*., 1977; Chow *et al*, 1977). This phenomenon is called alternative splicing, and was discovered during the 1970s. The mechanism of alternative splicing can have important consequences for gene function evolution, *i.e*., two mRNA isoforms can have different functions without any deleterious effect on the organism (Chow *et al.*, 1977). Thus, there would be no need for a second copy of the gene with a new function (Chow *et al*., 1977).

The frequency of alternative splicing in a given organism is determined by the intron density in the genes of that organism (Ast, 2004). The frequency of alternative splicing in higher eukaryotes is generally much higher than in lower eukaryotes. Although alternative splicing has been long thought to be absent in fungi (Barrass and Bleggs, 2003), EST data and microarray studies have provided a genome-wide picture of alternative splicing in a number of organisms, including fungi (Ebbole *et al*., 2004; Loftus *et al*., 2005). For example, 4.5 % of the genes in *Cryptococcus neoformans* undergo alternative splicing (Loftus *et al*., 2005), while a number of genes in the rice blast fungus, *M. grisea*, also undergo alternative splicing (Ebbole *et al.*, 2004).

4.1.1 The mechanisms of alternative splicing

Constitutive and alternative splicing uses the same machinery (Graveley, 2001). The production of different mRNA isoforms from a single pre-mRNA during alternative splicing is due to differences in regulation between the two types of splicing (Graveley, 2001).Similar to constitutive splicing, the information required for alternative splicing in higher eukaryotes is located on the exons. Lower eukaryotes, in contrast, have short introns and their alternative splicing information, similar to their constitutive splicing information, is located on their introns (Deutsch and Long, 1999; Lander, 2001).

The mechanisms involved in the generation of the mRNA isoforms vary among organisms (Figure 4). There are four known mechanisms responsible for the formation of mRNA isoforms (Smith and Valcarcel, 2000). The first is exon skipping, where some of the exons of a pre-mRNA contain weakened splicing signatures causing them not to be recognised by the spliceosome, which causes them to be skipped. In alternative 5′ or 3′ splicing, one exon contains

two 5′ splice sites or 3′ splice sites that can be used alternatively during splicing. In mutually exclusive exons, two different ESTs of the same pre-mRNA have one or more exons that overlap. Finally, in intron retention an intron contains weakened splicing signatures such that the intron does not get spliced out.

The mechanism that an organism uses for the identification and processing of its splice sites influences the profile of splice variants of that organism. Splice variants may be divided into four classes based on differences in intron processing: retained introns (RIs), cassette introns (CEs), and introns with competing 5′ splice sites, and competing 3′ splice sites (McGuire *et al*., 2008). McGuire *et al*. (2008) conducted a study to investigate splice variants in 42 eukaryotes (13 animals, 6 plants, 14 fungi, and 9 protists) with genome assemblies and a large number of publicly available ESTs (McGuire *et al*., 2008). Their study showed that most eukaryotes employ all types of alternative splicing as they produce all four classes of splice variants.

Despite the fact that all eukaryotes appear to utilize all types of alternative splicing, the respective mechanisms involved are not used to the same extent. For example, the 13 animals analysed by McGuire *et al.* (2008) had 1.3 times more cassette exons than retained introns. The fraction of cassette exons ranged from 28% for *Schistosoma mansoni* to 95% for *Branchiostoma floride*. Overall, fungi and protists were found to have 37 times more retained introns than cassette exons. The preference for intron retention in fungi is consistent with previous findings in *S. cerevisiae* and the intron-rich fungus *C. neoformans*, indicating that retained intron dominance in fungi is not coupled with intron density (McGuire *et al*., 2008). Plants appeared intermediate between animals and fungi in their amounts of retained introns and cassette exons (McGuire *et al*., 2008). The prevalence of cassette introns in higher eukaryotes is therefore consistent with their preferred use of ED for defining the region to be spliced. In contrast, retained introns are predominant in lower eukaryotes as they use ID for defining the region to be spliced (McGuire *et al*., 2008).

Variably spliced regions of the pre-mRNA exhibit size constraints (McGuire *et al*., 2008). In organisms such as fungi where cassette exons are rare, these exons are shorter than constitutive exons (McGuire *et al*., 2008). In organisms with rare retained introns, such as animals, those

introns are shorter than constitutive introns (McGuire *et al*., 2008). It has therefore been suggested that retained introns and cassette exons both tend to be shorter than their constitutively spliced counterparts, with the length difference most noticeable in organisms in which each splice variant is uncommon (McGuire *et al*., 2008).

Not all splice variants are functional. Some human cassette exons have been discovered to introduce stop codons and to alter the reading frame. In a study conducted by McGuire *et al.* (2008) most retained introns did not show a preference to preserve the reading frames as only 3% of the retained introns were evenly divisible by three. These results may have been due to artefacts associated with EST library construction. But, the universality of such artefacts across the many independent data sets and eukaryotes analysed by McGuire *et al.* (2008) could be suggesting that these events occur naturally and frequently.

4.1.2 The origin of alternative splicing

There are currently two models for explaining the evolution of alternative splicing. The first model speculates that mutations in the DNA sequences of exons and introns lead to changes in splicing patterns (Ast, 2004). The second model hypothesizes that the evolution of splicing regulatory factors (SR and heterogeneous nuclear RNP-pre-mRNA binding-proteins) induces selective pressure on constitutive exons to become alternative exons (Ast, 2004). In this model, the binding of SR proteins in proximity to a constitutively spliced exon weakens the selection on that exon (Ast, 2004). This releases the selective pressure from the splice sites, resulting in mutations that weaken them, leading to the exon being skipped in the pre-mRNA. According to Ast (2004), the mechanism remains speculative as this model has not received adequate experimental attention.

According to the first model for the origin of alternative splicing, mutated splice sites can lead to alternative splicing. The production of these weak splice sites through mutations would be adequate for the splicing machinery to skip an internal exon or to retain an internal intron during a series of splicing events (Ast, 2004). This gives the cell the potential to produce new transcripts with novel functions. It has been shown that alternative exons have weaker splice sites than constitutive exons (Stamm *et al*., 1994; Carmel *et al*., 2004), which allows for suboptimal

recognition of exons by the splicing machinery and leads to alternative splicing (Ast, 2004). Also, experiments on both yeast and *Drosophila* have shown that, when splice sites are presumably recognized by ID, mutating a single splice site diffuses splicing of the intron proximal to the mutation (McGuire *et al*., 2008). This leads to retention of the adjacent intron while splicing of the nearby intron remains unaffected (Romfo *et al*., 2000; Talerico and Berget, 1994). In contrast, when splice sites are presumably recognized by ED, mutating a splice site affects the splicing of both the intron proximal to and the intron on the other side of the mutated exonic splice site (McGuire *et al*., 2008). This results in the skipping of the exon with the mutated splice site (Talerico and Berget, 1990; Berget, 1995).

4.2 Intron-Mediated Enhancement

Introns have been reported to enhance gene expression in certain organisms. In 1990, Mascarenhas *et al*. reported that certain plant introns stimulated gene expression. This type of expression enhancement was termed intron-mediated enhancement (IME) and is different to that caused by transcriptional enhancer elements. For IME to occur, introns must be within the transcribed sequences and in the correct orientation. Another distinction between introns and transcriptional enhancer elements is that introns increase the accumulation of mRNA transcripts without affecting the rate of transcription initiation (Rose, 2002). Currently the mechanism of IME is unclear, but is suggested to operate at a co- or post-transcriptional level (Rose, 2002).

One of the best characterized examples of IME was found in *A. thaliana*, which involves the first intron of the tryptophan biosynthetic pathway gene *PAT1* (Rose and Last, 2000). This 110-nucleotide intron stimulated a five-fold increase in the steady-state level of mRNA expression without affecting the transcription of *PAT1* fused to a GUS (β-glucoronidase) reporter gene in transgenic *A. thaliana* (Rose and Beliakoff, 2000). When signatures of the intron were sequentially mutated, no enhancement in expression was observed (Rose and Beliakoff, 2000) proving that this intron was responsible for the enhancement of expression.

The position of an intron in a gene whose expression is enhanced is important. In addition to the first intron from *PAT1* (Rose and Last, 2000), studies conducted on yeast and mammals revealed

that introns more than 50-55 nucleotides upstream of a stop codon can cause that stop codon to be prematurely recognized, triggering non-sense mediated RNA decay – a rapid degradation of mRNAs containing premature translational stop codons (Conti and Izaurralde, 2005). This could explain an observed failure of introns located in the 3′ untranslated region (UTR) of any gene to boost expression (Conti and Izaurralde, 2005**)**. Rose (2006) showed that a similar system could be operating in plants. When the first intron of *PAT1* was inserted after the 25th or the 80th nucleotide downstream of the stop codon in the 3′ UTR of a *PAT1:GUS* fusion, the intron was spliced efficiently from the proximal location but with variable efficiency from the distal location (Rose, 2006). The ability of this intron to enhance expression was greatly reduced when it was moved to the 3′ UTR. These results show the importance of the position of introns involved in the enhancement of gene expression.

5. Future Prospects

As EST data for eukaryotes increases, the level of alternative splicing that occurs in these organisms appears to be much more than originally thought. This area of research is advancing for higher eukaryotes, due to more experiments being conducted on gene expression. Only few fungi are currently known to undergo alternative splicing. IME has been well investigated in plants, but not in fungi. Thus, there remains much room for research on alternative splicing and IME in fungi in order to gain a clearer picture of these intron functions. Furthermore, differences in both the structure and characteristics of fungal Spliceosomal introns necessitate the development of more defined fungal models. Since information relating to the architecture and distribution of Spliceosomal introns are used in *ab initio* genome annotation methods, investigating the intron structure in many more fungal species can aid in the improvement of gene prediction methods for these organisms.

6. References

1. Ast G: How did alternative splicing evolve? Nature Reviews Genetics 2004, 5(10):773-782.
2. Barrass JD, Beggs JD: Splicing goes global. Trends in Genetics 2003, 19(6):295-298.
3. Berget SM: Exon recognition in vertebrate splicing. Journal of Biological Chemistry 1995, 270(6):2411.
4. Berget SM, Moore C, Sharp PA: Spliced segments at the 5'terminus of adenovirus 2 late mRNA. Proceedings of the National Academy of Sciences 1977, 74(8):3171.
5. Berglund JA, Chua K, Abovich N, Reed R, Rosbash M: The splicing factor BBP interacts specifically with the pre-mRNA branch point sequence UACUAAC. Cell 1997, 89(5):781-787.
6. Bhattacharya D, Lutzoni F, Reeb V, Simon D, Nason J, Fernandez F: Widespread occurrence of spliceosomal introns in the rDNA genes of ascomycetes. Molecular Biology and Evolution 2000, 17(12):1971-1984.
7. Blake CCF: Do genes-in-pieces imply proteins-in-pieces? Nature 1978, 273:267.
8. Bon E, Casaregola S, Blandin G, Llorente B, Neuvéglise C, Munsterkotter M, Guldener U, Mewes HW, Van Helden J, Dujon B: Molecular evolution of eukaryotic genomes: hemiascomycetous yeast spliceosomal introns. Nucleic Acids Research 2003, 31(4):1121-1135.
9. Burke JM: Molecular genetics of group I introns: RNA structures and protein factors required for splicing--a review. Gene 1988, 73(2):273-294.
10. Carmel I, Tal S, Vig I, Ast G: Comparative analysis detects dependencies among the 5′ splice-site positions. RNA 2004, 10(5):828-840.
11. Cech TR: Self-splicing of group I introns. Annual Review of Biochemistry 1990, 59(1):543-568.
12. Chow LT, Gelinas RE, Broker TR, Roberts RJ: An amazing sequence arrangement at the 5'ends of adenovirus 2 messenger RNA. Cell 1977, 12(1):1-8.
13. Coolidge CJ, Seely RJ, Patton JG: Functional analysis of the polypyrimidine tract in pre-mRNA splicing. Nucleic Acids Research 1997, 25(4):888-896.

14. De Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W: Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. Proceedings of the National Academy of Sciences 1998, 95(9):5094.
15. Deutsch M, Long M: Intron-exon structures of eukaryotic model organisms. Nucleic acids research 1999, 27(15):3219.
16. Dibb N, Newman A: Evidence that introns arose at proto-splice sites. The EMBO journal 1989, 8(7):2015.
17. Doolittle WF: Genes in pieces: were they ever together? Nature 1978, 272(5654):581-582.
18. Fedorov A, Suboch G, Bujakov M, Fedorova L: Analysis of nonuniformity in intron phase distribution. Nucleic Acids Research 1992, 20(10):2553-2557.
19. Fink GR: Pseudogenes in yeast? Cell 1987, 49(1):5-6.
20. Galagan JE, Henn MR, Ma LJ, Cuomo CA, Birren B: Genomics of the fungal kingdom: insights into eukaryotic biology. Genome Research 2005, 15(12):1620-1631.
21. Gesteland RF, Cech T, Atkins JF: The RNA world: the nature of modern RNA suggests a prebiotic RNA world: Cold Spring Harbor Laboratory Pr; 2006.
22. Gilbert W: Why genes in pieces? Nature 1978, 271(5645):501.
23. Gilbert W: The exon theory of genes. In: 1987: Cold Spring Harbor Laboratory Press; 1987: 901-905.
24. Giroux MJ, Clancy M, Baier J, Ingham L, McCarty D, Hannah LC: De novo synthesis of an intron by the maize transposable element Dissociation. Proceedings of the National Academy of Sciences 1994, 91(25):12150.
25. Grate L, Ares M: Searching yeast intron data at Ares lab Web site. Methods in enzymology 2002, 350:380-392.
26. Graveley BR: Alternative splicing: increasing diversity in the proteomic world. Trends in Genetics 2001, 17(2):100-107.
27. Haugen P, Simon DM, Bhattacharya D: The natural history of group I introns. Trends in Genetics 2005, 21(2):111-119.
28. Hirschman JE, Balakrishnan R, Christie KR, Costanzo MC, Dwight SS, Engel SR, Fisk DG, Hong EL, Livstone MS, Nash R: Genome Snapshot: a new resource at the Saccharomyces Genome Database (SGD) presenting an overview of the *Saccharomyces cerevisiae* genome. Nucleic acids research 2006, 34(suppl 1):D442-D445.

29. Irimia M, Roy SW: Spliceosomal introns as tools for genomic and evolutionary analysis. Nucleic Acids Research 2008, 36(5):1703-1712.
30. Jeffares DC, Mourier T, Penny D: The biology of intron gain and loss. Trends in Genetics 2006, 22(1):16-22.
31. Jeon JS, Lee S, Jung KH, Jun SH, Kim C, An G: Tissue-preferential expression of a rice α-tubulin gene, OsTubA1, mediated by the first intron. Plant Physiology 2000, 123(3):1005-1014.
32. Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD: Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 2003, 302(5653):2141-2144.
33. Jurica MS, Moore MJ: Pre-mRNA Splicing:: Awash in a Sea of Proteins. Molecular Cell 2003, 12(1):5-14.
34. Kim E, Goren A, Ast G: Alternative splicing: current perspectives. Bioessays 2008, 30(1):38-47.
35. Kohtz JD, Jamison SF, Will CL, Zuo P, Lührmann R, Garcia-Blanco MA, Manley JL: Protein–protein interactions and 5'-splice-site recognition in mammalian mRNA precursors. 1994.
36. Kupfer DM, Drabenstot SD, Buchanan KL, Lai H, Zhu H, Dyer DW, Roe BA, Murphy JW: Introns and splicing elements of five diverse fungi. Eukaryotic Cell 2004, 3(5):1088-1100.
37. Lambowitz AM, Perlman PS: Involvement of aminoacyl-tRNA synthetases and other proteins in group I and group II intron splicing. Trends in biochemical sciences 1990, 15(11):440-444.
38. Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W: Initial sequencing and analysis of the human genome. Nature 2001, 409(6822):860-921.
39. Lim LP, Burge CB: A computational analysis of sequence features involved in recognition of short introns. Proceedings of the National Academy of Sciences 2001, 98(20):11193-11198.
40. Lin K, Zhang DY: The excess of 5′ introns in eukaryotic genomes. Nucleic Acids Research 2005, 33(20):6522-6527.

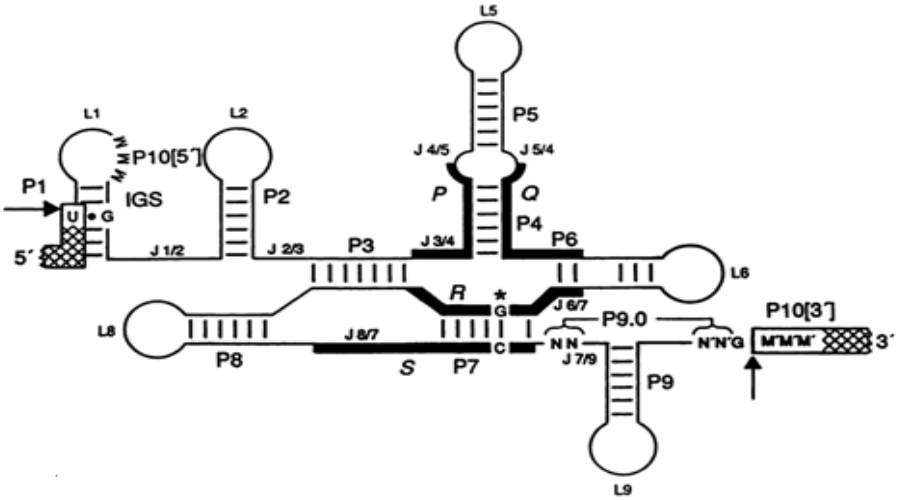
41. Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA: The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. Science 2005, 307(5713):1321.
42. Logsdon JM: The recent origins of spliceosomal introns revisited. Current Opinion in Genetics & Development 1998, 8(6):637-648.
43. Logsdon Jr JM, Stoltzfus A, Doolittle WF: Molecular evolution: Recent cases of spliceosomal intron gain? Current Biology 1998, 8(16):R560-R563.
44. Long M, De Souza SJ, Rosenberg C, Gilbert W: Relationship between "proto-splice sites" and intron phases: evidence from dicodon analysis. Proceedings of the National Academy of Sciences 1998, 95(1):219.
45. Long M, Rosenberg C, Gilbert W: Intron phase correlations and the evolution of the intron/exon structure of genes. Proceedings of the National Academy of Sciences 1995, 92(26):12495.
46. Lynch M: Intron evolution as a population-genetic process. Proceedings of the National Academy of Sciences 2002, 99(9):6118.
47. Martin W, Koonin EV: Introns and the origin of nucleus–cytosol compartmentalization. Nature 2006, 440(7080):41-45.
48. Martinez P, Martin W, Cerff R: Structure, evolution and anaerobic regulation of a nuclear gene encoding cytosolic glyceraldehyde-3-phosphate dehydrogenase from maize. Journal of Molecular Biology 1989, 208(4):551-565.
49. Mascarenhas D, Mettler IJ, Pierce DA, Lowe HW: Intron-mediated enhancement of heterologous gene expression in maize. Plant Molecular Biology 1990, 15(6):913-920.
50. McGuire AM, Pearson MD, Neafsey DE, Galagan JE: Cross-kingdom patterns of alternative splicing and splice recognition. Genome Biology 2008, 9(3):R50.
51. Michel F, Westhof E: Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. Journal of Molecular Biology 1990, 216(3):585-610.
52. Mourier T, Jeffares DC: Eukaryotic intron loss. Science 2003, 300(5624):1393.
53. Nguyen H, Yoshihama M, Kenmochi N: Phase distribution of spliceosomal introns: implications for intron origin. BMC Evolutionary Biology 2006, 6(1):69.

54. Nielsen CB, Friedman B, Birren B, Burge CB, Galagan JE: Patterns of intron gain and loss in fungi. PLoS Biology 2004, 2(12):e422.
55. Palmer JD, Logsdon JM: The recent origins of introns. Current Opinion in Genetics & Development 1991, 1(4):470-477.
56. Parra G, Bradnam K, Rose AB, Korf I: Comparative and functional analysis of intron-mediated enhancement signals reveals conserved features among plants. Nucleic Acids Research 2011, 39(13):5328-5337.
57. Piccirilli JA, McConnell TS, Zaug AJ, Noller HF, Cech TR: Aminoacyl esterase activity of the Tetrahymena ribozyme. Science 1992, 256(5062):1420.
58. Purugganan M, Wessler S: The splicing of transposable elements and its role in intron evolution. Genetica 1992, 86(1):295-303.
59. Romfo CM, Alvarez CJ, Van Heeckeren WJ, Webb CJ, Wise JA: Evidence for splice site pairing via intron definition in *Schizosaccharomyces pombe*. Molecular and cellular Biology 2000, 20(21):7955-7970.
60. Roscigno R, Weiner M, Garcia-Blanco M: A mutational analysis of the polypyrimidine tract of introns. Effects of sequence differences in pyrimidine tracts on splicing. Journal of Biological Chemistry 1993, 268(15):11222-11229.
61. Rose AB: Requirements for intron-mediated enhancement of gene expression in Arabidopsis. RNA 2002, 8(11):1444-1453.
62. Rose AB, Beliakoff JA: Intron-mediated enhancement of gene expression independent of unique intron sequences and splicing. Plant Physiology 2000, 122(2):535-542.
63. Roy SW: Recent evidence for the exon theory of genes. Genetica 2003, 118(2):251-266.
64. Roy SW, Gilbert W: The pattern of intron loss. Proceedings of the National Academy of Sciences 2005, 102(3):713.
65. Roy SW, Gilbert W: Rates of intron loss and gain: implications for early eukaryotic evolution. Proceedings of the National Academy of Sciences 2005, 102(16):5773.
66. Roy SW, Gilbert W: The evolution of spliceosomal introns: patterns, puzzles and progress. Nature Reviews Genetics 2006, 7(3):211-221.
67. Roy SW, Lewis BP, Fedorov A, Gilbert W: Footprints of primordial introns on the eukaryotic genome. Trends in Genetics: TIG 2001, 17(9):496.
68. Russell PJ: iGenetics: a Mendelian approach: Pearson/Benjamin Cummings; 2006.

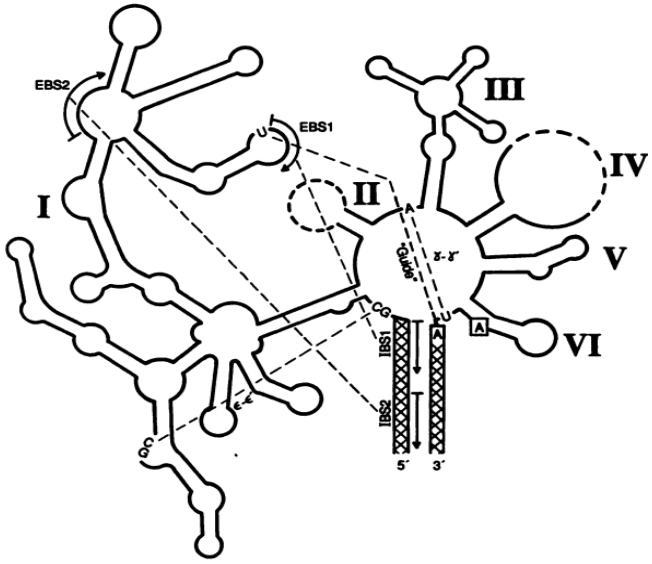
69. Ruvinsky A, Eskesen S, Eskesen F, Hurst L: Can codon usage bias explain intron phase distributions and exon symmetry? Journal of Molecular Evolution 2005, 60(1):99-104.
70. Sakurai A, Fujimori S, Kochiwa H, Kitamura-Abe S, Washio T, Saito R, Carninci P, Hayashizaki Y, Tomita M: On biased distribution of introns in various eukaryotes. Gene 2002, 300(1-2):89-95.
71. Saldanha R, Mohr G, Belfort M, Lambowitz AM: Group I and group II introns. The FASEB journal 1993, 7(1):15-24.
72. Sharp P, Roberts R: Adenovirus mazes at Cold Spring Harbor. Nature 1977, 268:101-104.
73. Sharp PA: On the origin of RNA splicing and introns. Cell 1985, 42(2):397-400.
74. Shelley CS, Baralle FE: Deletion analysis of a unique 3'splice site indicates that alternating guanine and thymine residues represent an efficient splicing signal. Nucleic Acids Research 1987, 15(9):3787-3799.
75. Singh R, Valcarcel J, Green MR: Distinct binding specificities and functions of higher eukaryotic polypyrimidine tract-binding proteins. Science 1995, 268(5214):1173-1176.
76. Smith CWJ, Valcárcel J: Alternative pre-mRNA splicing: the logic of combinatorial control. Trends in Biochemical Sciences 2000, 25(8):381-388.
77. Staley JP, Guthrie C: Mechanical Devices of the Spliceosome: Review Motors, Clocks, Springs, and Things. Cell 1998, 92:315-326.
78. Stamm S, Zhang MQ, Marr TG, Helfman DM: A sequence compilation and comparison of exons that are alternatively spliced in neurons. Nucleic Acids Research 1994, 22(9):1515-1526.
79. Stark JM, Bazett-Jones DP, Herfort M, Roth MB: SR proteins are sufficient for exon bridging across an intron. Proceedings of the National Academy of Sciences 1998, 95(5):2163-2168.
80. Talerico M, Berget SM: Effect of 5'splice site mutations on splicing of the preceding intron. Molecular and Cellular Biology 1990, 10(12):6299-6305.
81. Talerico M, Berget SM: Intron definition in splicing of small Drosophila introns. Molecular and Cellular Biology 1994, 14(5):3434-3445.
82. Valcárcel J, Gaur RK, Singh R, Green MR: Interaction of U2AF65 RS region with pre-mRNA of branch point and promotion base pairing with U2 snRNA. Science 1996, 273(5282):1706.

83. Vicens Q, Paukstelis PJ, Westhof E, Lambowitz AM, Cech TR: Toward predicting self-splicing and protein-facilitated splicing of group I introns. RNA 2008, 14(10):2013-2029.
84. Wang Z, Hoffmann H, Grabowski P: Intrinsic U2AF binding is modulated by exon enhancer signals in parallel with changes in splicing activity. RNA 1995, 1(1):21-35.
85. Will CL, Lührmann R: Protein functions in pre-mRNA splicing. Current Opinion in Cell Biology 1997, 9(3):320-328.
86. Woodley L, Valcárcel J: Regulation of alternative pre-mRNA splicing. Briefings in Functional Genomics & Proteomics 2002, 1(3):266-277.
87. Xing Y, Lee C: Alternative splicing and RNA selection pressure—evolutionary consequences for eukaryotic genomes. Nature Reviews Genetics 2006, 7(7):499-509.
88. Yarus M: How many catalytic RNAs? Ions and the Cheshire cat conjecture. The FASEB journal 1993, 7(1):31-39.
89. Zahler AM, Roth MB: Distinct functions of SR proteins in recruitment of U1 small nuclear ribonucleoprotein to alternative 5'splice sites. Proceedings of the National Academy of Sciences 1995, 92(7):2642-2646.
90. Zhou Z, Licklider LJ, Gygi SP, Reed R: Comprehensive proteomic analysis of the human spliceosome. Nature 2002, 419(6903):182-185.
91. Zuo P, Maniatis T: The splicing factor U2AF35 mediates critical protein-protein interactions in constitutive and enhancer-dependent splicing. Genes & Development 1996, 10(11):1356-1368.

7. Figures



A



B

Figure 1. The structure of Group I and II introns (Saldanha *et al*., 1993). **A**: The secondary and tertiary structure of a Group I intron. For splicing, the intron folds to bring together the 5' and 3' splice sites (indicated by arrows) and the external Guanosine (indicated with the asterisk). P1-10 represents the characteristic 10 paired segments of this intron type. **B**: For splicing of a Group II intron, its folding facilitates interaction (indicated by the dotted lines) between the 5' and 3' splice sites and the EBS (exon binding site) and IBS (intron binding site). I-VI represent the typical domains of this intron type.

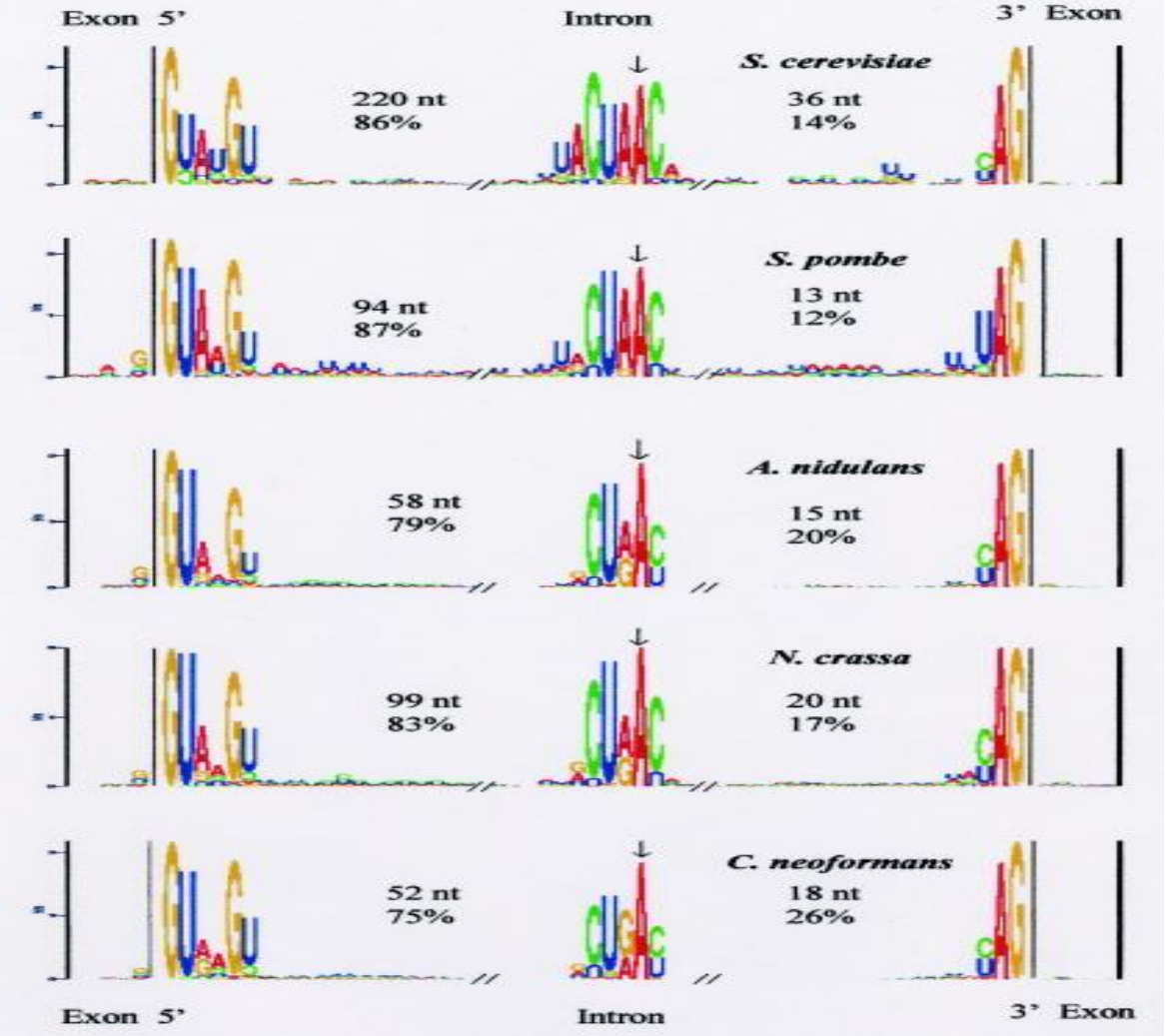


Figure 2. The consensus signature sequences for the 5′ splice site, the branch site and 3′ splice site proposed by Kupfer *et al*. (2004) for the Spliceosomal introns of *Cryptococcus neoformans*, *Neurospora crassa*, *Aspergillus nidulans*, *Saccharomyces cerevisiae* and *Schizosaccharomyces pombe*. Nucleotide positions are shown on the *x*-axis, while base frequencies are shown on the *y*-axis.

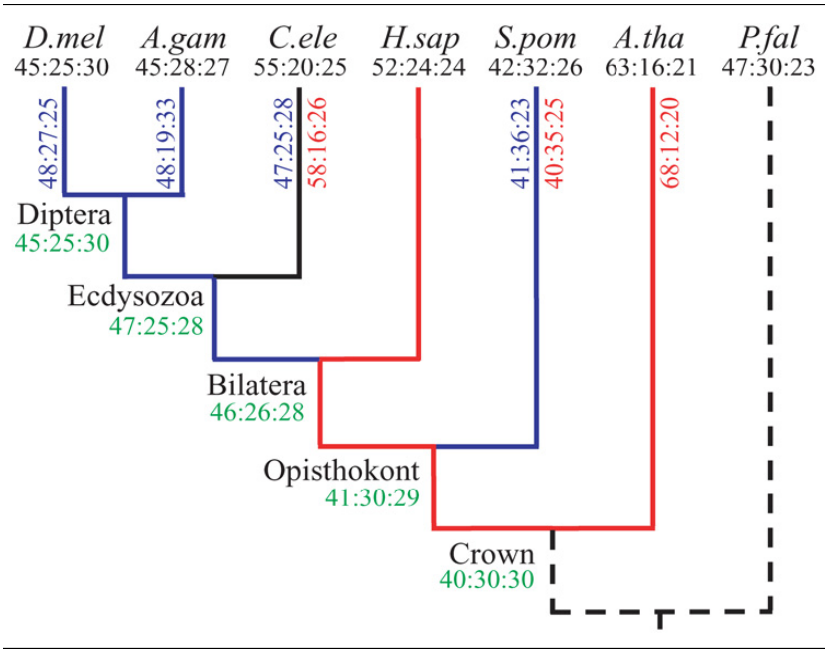


Figure 3. Schematic representation of the evolution of intron phase across the eukaryotic tree (Nguyen *et al*., 2006). The numbers in black show phase distributions (phase-0:phase-1:phase-2) of Spliceosomal introns in extant species. The inferred introns phase distributions in extinct ancestors are shown in green. Phase distributions of gained introns are indicated in red, while those of lost introns are shown in blue. Branches that experienced more gains than losses are shown in red, while those associated with more losses are shown in blue. Abbreviations are as follows: *D. mel=D. melanogaster*, *A. gam=A. gambiae*, *C. ele=C.elegans*, *H. sap=H. sapiens*, *S. pom=S. pombe*, *A. tha=A. thaliana*, *P. fal=P. falciparum*.

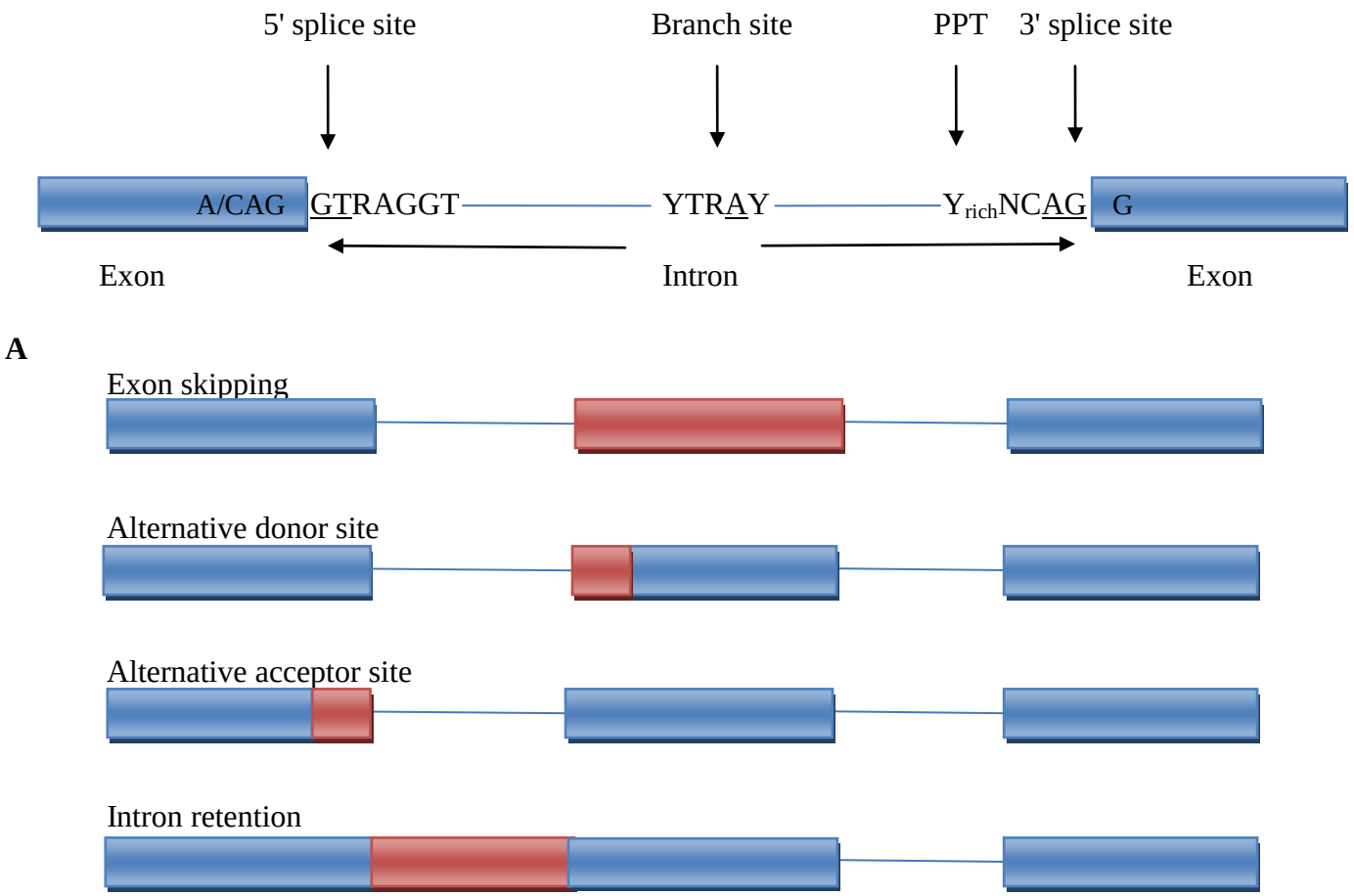


Figure 4. Alternative splicing of Spliceosomal introns (modified from Kim *et al*., 2007). **A**: The four intron signature sequences or *cis*-elements of Spliceosomal introns, where the nucleotides of the 5' and 3' splice sites and the branching point are underlined. The polypyrimidine tract is indicated by PPT. **B**: The four known forms of alternative splicing found in eukaryotes. The blue boxes indicate constitutive exons and solid lines indicate introns. Alternatively spliced transcripts are shown in red.

Chapter 2: The architecture and distribution of introns in housekeeping genes of four *Fusarium* species

Abstract

The removal of introns from transcribed RNA represents a crucial step during the production of mRNA in all eukaryotes. The availability of whole-genome sequences and expressed sequence tags (ESTs) has greatly increased our knowledge of this process and has revealed various commonalities among eukaryotes. However, these data have also shown that certain aspects of intron structure and diversity are taxon-specific, which greatly complicates the accuracy of *in silico* gene prediction methods. The aim of this study was therefore to use genomic and EST information to characterize the structure and distribution of spliceosomal introns in the agriculturally and medically important fungal genus *Fusarium*. For this purpose, we utilized the available sequence information for four species (*F. circinatum*, *F. verticillioides*, *F. oxysporum* and *F. graminearum*). The introns of Housekeeping genes were specifically examined to identify and compare the intron position, length, and presence of intron *cis*-elements among genes and species. Our results indicated that *Fusarium* species have the canonical 5' and 3' splice sites, but with subtle differences that are not shared with those of other fungal genera. The *Fusarium* polypyrimidine tract was also found to be highly divergent among the *Fusarium* species and the Housekeeping genes. Comparison of intron sequences also revealed that certain *Fusarium* introns potentially encode unique branch site motifs that contain the conserved Adenosine base required during the first step of splicing. The incorporation of these findings into current genome annotation software and pipelines will thus significantly improve the accuracy of gene prediction methods used for *Fusarium* species and other related fungi.

Introduction

Eukaryotic genes are often interrupted by introns, which are stretches of non-coding intervening DNA sequences (Bhattacharya *et al*., 2000). The three main types of intron that have been identified are Group I, Group II and Spliceosomal introns. Group I and Group II introns are present throughout the Tree of Life, where they disperse themselves within and between genomes, potentially contributing to the emergence of new genes (Saldanha, 1993). Both these types of introns also represent ribozymes, because they facilitate their own excision or splicing (Saldanha, 1993). In contrast, Spliceosomal introns, which are exclusively found in the nuclear protein-coding genes of Eukaryota, require the action of a large protein complex, the spliceosome, to facilitate their splicing (Jeffares *et al*., 2006).

The length of Spliceosomal introns vary depending on genome size and level of expression of the genes that harbour them. Larger genomes generally tend to harbour longer introns (Deutsch and Long, 1999), while highly expressed genes commonly harbour shorter introns (Castillo-Davis *et al*., 2002). The size of introns can also depend on their position within a gene. First introns, whether found in the 5′ untranslated region of a gene or in the open reading frame (ORF), are typically longer than those found downstream (Bradnam and Korf, 2008). However, those occurring in the 5′ untranslated gene region are usually longer than first introns in the coding region (Bradnam and Korf, 2008; Hawkin, 1988, Hong *et al*., 2006; Smith, 1988; Kriventseva and Gelfand, 1999). This is because introns in 5′ untranslated regions may contain functional elements such as expression enhancement motifs that accounts for their increased size (Gaffney and Keightley, 2006; Rose *et al*., 2002; Bradnam and Korf, 2008; Parra *et al*., 2011).

To maintain functionality, the minimum length of Spliceosomal introns is restricted. This is because Spliceosomal introns have to harbour a number of characteristic intron signature motifs or *cis*-elements to allow splicing (Deutsch and Long, 1999). These *cis*-elements or signatures include the 5' splice site, 3' splice site, the branch site containing the branching point and the polypyrimidine tract. Of these, the polypyrimidine tract has been reported to be optional (Kupfer *et al*., 2004). During splicing, the components of the spliceosome (that consists of small nuclear ribonucleoprotein complexes, snRNPs) bind sequentially to the 5' splice site, the branch site and

the 3′ splice site through RNA-DNA base pairing (Ast, 2004; Russell, 2006). The intron is then excised followed by the ligation of the two exons that flanked the intron (Russell, 2006).

Information regarding Spliceosomal intron architecture has been used to improve the accuracy of gene predictions (Bradnam and Korf, 2008). However, the annotations of many fungal genomes remain inadequate (Kupfer *et al*., 2004; Ter-Hovhannisyan *et al*., 2008). This is because intron position may be conserved among evolutionary distant lineages (Irimia and Roy, 2008), although the splicing signals between closely related genomes can be intrinsically different (Kupfer *et al*., 2004). For example, not all intron positions between closely related species are conserved and they may have dissimilar *cis*-elements. Therefore, most *ab initio* gene prediction methods (*i.e.*, methods that use characteristic gene sequences for genome annotations) generally lack high *cis*-element specificity, which limit their ability to deal with introns that are apparently less conserved (Ter-Hovhannisyan *et al*., 2008). This is especially true for fungal genomes as effective utilization of information relating to intron *cis*-elements, particularly the branch site, have been shown to improve the accuracy of splice site predictions by more than 90% (Ter-Hovhannisyan *et al*., 2008).

In this study, we considered the distribution and architecture of Spliceosomal introns in the fungal genus *Fusarium* (Phylum *Ascomycota*, Class *Sordariomycetes*, Order *Hypocreales*). The members of this taxon includes a wide range of medically, veterinary and agriculturally important fungi **(**Kvas *et al*., 2009**)**. Of these, the genomes of four plant pathogenic species have been sequenced (Wingfield *et al*., 2012; *Fusarium* Comparative Database, Broad Institute (http://www.broadinstitute.org/annotation/genome/fusarium_group). These are *F. circinatum* a pathogen of pine, *F. verticilioides* a pathogen of maize, *F. oxysporum* that is pathogenic to tomato and the wheat pathogen *F. graminearum* (Leslie and Summerell, 2006). To characterize the structure of Spliceosomal introns of these four *Fusarium* species, introns located in a standard set of Housekeeping (HK) genes were identified (Parra *et al*., 2007) and compared by making use of genome sequence and EST (Expressed Sequence Tag) data. Results from this study will broaden current knowledge regarding the organization of fungal Spliceosomal introns. This knowledge will also be used to improve gene prediction software for the accurate annotation of the genomes of *Fusarium* species, and possibly other fungal genera.

Materials and methods

Fusarium genomes

The genome sequences for *F. graminearum*, *F. verticilliodes* and *F. oxysporum* were accessed from the Broad Institute's *Fusarium* Comparative Database (http://www.broadinstitute.org/annotation/genome/fusarium_group), while the genome sequence for *F. circinatum* was available in-house (Forestry and Agricultural Biotechnology Institute [FABI], University of Pretoria; Wingfield *et al*., 2012). The *F. graminearum*, *F. verticillioides* and *F. oxysporum* genomes have respective sizes of 36 Mb, 42 Mb and 60 Mb (Ma *et al*., 2010), while that of *F. circinatum* is 44 Mb in size (Wingfield *et al*., 2012). Annotations of the *F. verticillioides*, *F. oxysporum* and *F. graminearum* genomes have been done using GENEid (Guigo *et al*., 1992) and FGENESH (Solovyev *et al*., 1994) gene prediction software. The *F. circinatum* genome was annotated using MAKER (Cantarel *et al*., 2008), which incorporates the programs GeneMark-ES (Ter-Hovhannisyan *et al*, 2008), Augustus (Stanke and Waack, 2003) and SNAP (Korf, 2004).

Identification of *Fusarium* HK Genes

The *Fusarium* HK genes employed in this study corresponded to a set of 458 protein-coding HK genes that are common to *Saccharomyces cerevisiae* and presumably all eukaryotes (Holt and Yandell, 2011). To compile the HK gene set for *Fusarium*, the *S. cerevisiae* HK genes were retrieved from the Σ-cegma (Core Eukaryotic Genes Mapping Approach; Parra *et al*., 2007) website (http://korflab.ucdavis.edu/Datasets/cegma/Appendix.html) and used in BLASTp searches against the *Fusarium* Comparative Database at the Broad Institute. FASTA files were subsequently compiled with the *F. verticillioides*, *F. oxysporum* and *F. graminearum* nucleotide sequences for the HK genes. The *F. verticillioides*, and in a few cases the *F. oxysporum* and *F. graminearum*, HK gene sequences were then used to identify and retrieve the corresponding sequences from the *F. circinatum* genome by making use of BLASTn, CLC Genomics Workbench version 3.7.1 (CLC bio A/S) and BioEdit version 7.0.9.0 (Hall, 1999).

Identification and annotation of introns

Alignments of the HK genes of the four *Fusarium* species were performed in BioEdit using the ClustalW Multiple Alignment tool (Thompson *et al*., 1994). Predicted intron positions in the HK genes of *F. verticillioides*, *F. oxysporum* and *F. graminearum* were assessed using the Broad Institute gene models. Intron positions in the *F. circinatum* HK genes were assessed by employing the MAKER gff annotations in the Apollo Genome Annotation Curation Tool (Lewis *et al*., 2002). The identified introns were then annotated in CLC Main Workbench 5.7 (CLC bio A/S).

ClustalW alignments of all HK genes with apparently incongruent (non-conserved) intron positions were analyzed in AUGUSTUS version 2.4 (Stanke and Waack, 2003; Stanke *et al*., 2008). This computer software predicts gene elements by using intrinsic *ab initio* algorithms and extrinsic experimental data such as EST sequences. The *F. graminearum* default transition matrix (*i.e.*, the matrix of transition probabilities for predicted gene models) was used because it is the only *Fusarium* species for which a matrix is available in AUGUSTUS. In addition to resolving incongruent intron positions, the EST data was also useful for identifying unusual *cis*-elements (see below).

Analysis of intron architecture and distribution

The three main *cis*-elements (5′ splice site, branch site with the branch point and 3′ splice site) of introns from all the genes in the dataset were examined. To achieve this, the annotated introns were extracted from each HK gene alignment using the "Extract Sequences" and then the "Extract Annotations" functions in CLC Genomics Workbench. In all cases, an additional five nucleotides in the upstream exon and five additional nucleotides in the downstream exon were extracted with the introns. The extracted introns were then aligned using the "Create Alignment" function after which they were manually annotated relative to the annotations (Figure 1) previously produced for fungal introns (Kupfer *et al*., 2004; Bhasi *et al*., 2007).

Using the introns from a preliminary set of seven arbitrarily chosen HK gene alignments, a *cis*-elements motif list was constructed from these alignments using the "Create Motif List" function. The motif list was then used to perform a motif search on the rest of the gene sequence

alignments, and sequence variants of these elements were continuously added when new motifs were encountered.

To examine the structure of the polypyrimidine tract, the introns of 10 randomly selected genes were used. These genes were arbitrarily chosen and only ten were used due to the high variation observed in position, length and potential number per intron between polypyrimidine tracts of the four *Fusarium* species. A minimal definition of six consecutive nucleotides with at least three Thymidines (*i.e.*, Uridines in the transcribed sequence) and no Adenosines was used for this purpose (Kupfer *et al*., 2004). The polypyrimidine tracts were analyzed using the sequence, position and intron region occupied by the tract (*i.e.*, the polypyrimidine tract may occupy the region between the 5′ splice site and the branch site or the region between the branch site and the 3′ splice site) (Kupfer *et al*., 2004).

For the analysis of intron length, frequency and distribution, a subset of genes were used. This subset included 226 HK gene sequence alignments that were arbitrarily selected. Each alignment contained one gene from each of the four *Fusarium* species. Any gene that showed intron position incongruence among the four *Fusarium* species without EST data support was excluded from this dataset.

The dependence of intron length on its position in the coding sequences (CDSs) of a gene was tested using analysis of variance (ANOVA) (http://www.physics.csbsju.edu/stats/anova.html) and hypothesis testing was done using the *F* test where the null (H_o) hypothesis was that all the mean lengths of the introns in different positions are equal (Samuels and Witmer, 2003). For this purpose the *F* statistic was compared to an *F* distribution critical value at a 99.99% confidence level (*P* = probability = 0.001) with 6 (numerator) and 500 (denominator; the exact denominator values for *F. verticillioides*, *F. circinatum*, *F. oxysporum* and *F. graminearum* were 497, 498, 498 and 496, respectively) degrees of freedom (Samuels and Witmer, 2003). To test whether the observed differences between means were significant, Tukey's Honesty Significant Difference (HSD) test was used. For this purpose, Tukey's statistic value equation $q = \frac{M1 - M2}{\sqrt{MSw\left(\frac{1}{n}\right)}}$ was used (q is Tukey's statistic, M_1 and M_2 are the two means being tested, MS_w is the mean of squares

within the data, and *n* is the number of samples per treatment group). Critical values for Tukey's HSD (*q*) were determined using the degrees of freedom (∞; the exact values for *F. verticillioides*, *F. circinatum*, *F. oxysporum* and *F. graminearum* were 497, 498, 498 and 496, respectively) and the number of treatments (7), which is the number of intron positions compared.

For the analyses of relationships between gene length and intron length, and gene length and number of introns per gene, Scatter plots were generated (Microsoft Excel 2010). The significance of regression lines were tested using the Student's *t* test, where H_o and H_a were, respectively $\beta_1 = 0$ and $\beta_1 \neq 0$ (β_1 is the slope of the regression line) (Samuels and Witmer, 2003). When $\beta_1 = 0$, there is no correlation between gene length and intron length or gene length and number of introns per gene, and the null hypothesis will be accepted. The *t* statistic was obtained with the $t_s = b_1/SE_{b1}$ equation, where t_s is the *t* statistic, b_1 is the slope of the regression line and SE_{b1} is the standard error of the slope. For these calculations, the Excel Regression Tool found in Excel Analysis ToolPack and Excel Analysis ToolPack-VBA Add-ins were used. The *t* critical values were obtained from the Student's *t* distribution table at a probability value of $P = 0.10$ and 0.05 with $n - 2$ degrees of freedom (Samuels and Witmer, 2003).

For the analysis of intron phase, 50 genes were randomly selected from among the 226 HK dataset. Intron phases were examined manually in BioEdit, using the CDSs together with the protein and genomic sequences. The "Toggle Translation" function in BioEdit was used to determine whether introns were in phase 0 (between two codons), phase 1 (between the first two nucleotides of a codon) or phase 2 (between the last two nucleotides of a codon).

Results

Identification of Fusarium HK genes

The initial set of 458 *S. cerevisiae* core genes common to all eukaryotes (Parra *et al*., 2007) produced 458 significant BLASTn hits from the *Fusarium* Comparative Database at the Broad Institute. However, twelve of the 458 genes were absent in the *F. verticillioides* genome. To compensate for these missing genes, genes from *F. graminearum* (9) and *F. oxysporum* (3) were used for BLASTn searches against the *F. circinatum* genome instead. These searches allowed for the identification of the full set of 458 HK gene homologues in the *F. circinatum* genome, although 22 were separated over multiple contigs in the assembly. These gene homologues were excluded, thus resulting in a final dataset containing 436 HK gene sequences for the four *Fusarium* species.

Identification and annotation of introns

Of the 436 HK genes examined, 152 appeared be to non-conserved with respect to intron position. To ensure that this lack of conservation was not due to the different gene prediction methods used during annotation of the four genomes, all 152 genes were re-analyzed with AUGUSTUS using the available EST data. The analysis revealed that 54% (82) of the 152 genes were wrongly annotated. In part this is due to the fact that dissimilar gene prediction software were used for the four *Fusarium* genomes. Also, AUGUSTUS does not recognize GC-AG introns as was previously also showed for *Armillaria mellea* by Misiek and Hoffmeister (2008). The re-analysis with AUGUSTUS showed that these genes actually harboured positionally conserved introns that correlated with the positions in the genomes of at least one of the three other species. About 2% of the *F. ciricnatum* HK genes had very short introns of 5-23 nucleotides (nt) in the MAKER annotations, which were resolved using AUGUSTUS and the available EST data. After this process of resolving intron positions based on EST data, introns with non-conserved positions were present in 70 of the 436 HK gene alignments. Therefore the large majority (84%) of the HK genes examined harboured introns with conserved positions.

For the 70 genes with introns that were apparently not conserved with respect to position, EST data (from the four *Fusarium* species) were available for 24 genes only. Examination of the 46 gene alignments lacking EST support revealed five genes for which positional incongruences among introns were due to nucleotide substitutions. In one case, *F. graminearum* had an AT instead of the canonical 5′ splice site GT at the beginning of the intron, as was observed for the other three *Fusarium* species. In another case *F. graminearum* had a TT instead of a GT and in two other cases a GC instead of a GT at the 5′ splice site. *F. oxysporum* had one case of a GT to GA substitution. Seventeen gene alignments had non-conserved intron positions due to intron insertions/deletions (indels) (seven from *F. graminearum* and *F. oxysporum* each, and three from *F. verticillioides*). A number of gene alignments had shorter or longer predicted ORFs (four shorter ORFs in *F. graminearum*, two shorter ORFs in each of *F. verticillioides* and *F. oxysporum*, three shorter ORFs in *F. circinatum* and two longer ORFs in *F. circinatum*). Eleven gene alignments contained at least one truncated (*i.e.*, not fully sequenced) homologue of the gene in one of the species examined. All the above-mentioned gene alignments were excluded from the dataset.

The remaining 24 EST data-supported HK gene alignments with non-conserved intron positions were the result of intron indels, shorter ORFs, or the genes were highly divergent in terms of their nucleotide composition. Of these, 15 genes from *F. graminearum* had whole-intron indels, while one had a Thymidine to Guanosine (T>G) substitution in the second nucleotide of the 5′ splice site. Of the six intron position incongruences observed in *F. oxysporum*, three were the result of indels, one had a shorter ORF, while two were divergent in nucleotide sequence in comparison to the other *Fusarium* species. Of the three genes with non-conserved intron positions in *F. circinatum*, one gene had a whole-intron indel, two had shorter ORFs and one gene sequence was highly divergent compared to the other three species.

Analysis of intron architecture and distribution

Intron frequency per HK gene

The number of introns within each of the randomly selected 226 HK genes ranged from 0 to 15 introns per gene (Figure 2; Appendix), with an average density of 2.53 introns per gene. Seven of the examined HK genes harbored no introns in any of the four *Fusarium* species (Appendix) and included genes encoding pre-mRNA splicing factor clf, Aminomethyltransferase- mitochondrial precursor, Seryl tRNA synthetase, DNA pantothenate metabolism flavoprotein 2, Sol1 family protein, Uridylate kinase and DNA repair helicase RAD25. The gene encoding CTP synthase had the highest number of introns (15 in all four *Fusarium* species), followed by that encoding Glutamine dependent NAD+ synthetase with 14 introns in *F. verticillioides*, *F. circinatum* and *F. oxysporum*, and 12 in *F. graminearum* (Appendix).

No significant relationship was found between gene length and the number of introns per gene. We found *t* statistic values of 0.76, 0.72, 0.70 and 0.77 for *F. verticillioides*, *F. circinatum*, *F. oxysporum* and *F. graminiearum*, respectively. At probability values of $P = 0.10$ and 0.05 with $n - 2$ degrees of freedom, the respective *t* critical values were 1.65 and 1.96. The H_o (*i.e.*, $\beta_1 \neq 0$) thus could not be rejected as the number of introns per gene does not appear to be significantly associated with gene length (Figure 3).

Intron length

Intron length within the set of 226 HK genes of the four *Fusarium* species examined was on average 75.4 nt and ranged from 42 to 529 nt (Figure 4). The mean intron lengths for *F. verticillioides*, *F. circinatum*, *F. oxysporum* and *F. graminearum* were respectively 75.8 (44-529) nt, 76.3 (42-520) nt, 75.8 (42-529) nt and 75.7 (43-525) nt, where the numbers in parentheses denote intron size range. In almost all gene alignments the lengths of introns in the HK genes of *F. graminearum* were different from those of *F. verticillioides, F. circinatum and F. oxysporum* (See supplementary spreadsheet on disc provided).

Longer introns were mostly located at the 5' end of genes. This phenomenon was more pronounced when first introns were compared with the rest of the introns in the HK genes. The mean intron length of the first introns from all the four *Fusarium* species was approximately 93 nt (42-529). The rest of the introns at positions 2-7 had mean intron lengths between 54 and 72 nt (introns in positions 8-15 were excluded due to inadequate data to perform an ANOVA on them). A similar trend was observed for all four species when analyzed independently (Figure 5). Respective *F* statistic values of 6.277, 6.801, 6.374 and 6.027 were obtained for *F. verticillioides*, *F. circinatum*, *F. oxysporum* and *F. graminearum* from the ANOVA test. The *F* distribution critical value for all the species was 3.81, which showed that the means were significantly different. The null hypotheses, where the mean intron lengths are equal, were thus rejected for all species at $P = 0.001$. Tukey's HSD tests indicated that the mean sizes of introns in first positions were significantly different from those for introns in downstream positions. The Tukey's statistic values for *F. verticillioides*, *F. circinatum*, *F. oxysporum* and *F. graminearum* when M_1 (mean size of first introns) was compared to M_2 to M_7 (mean intron sizes at positions 2 through to 7, respectively) in a pairwise manner were found to be greater than 6.67, 6.73, 6.73 and 5.71, respectively. All other comparisons that excluded the mean sizes of the first introns produced values less than 1.94. The critical value at $P = 0.05$ was 4.17, indicating that the null hypotheses were only rejected when M_1 was compared with M_2 to M_7.

A statistically significant negative correlation between gene length and intron length was observed, where longer genes had shorter introns (Figure 6). We found *t* statistic values of 1.67, 1.98, 1.62 and 1.90 for *F. verticillioides*, *F. circinatum*, *F. oxysporum* and *F. graminiearum*, respectively. At 90 and 95% confidence levels (0.10 and 0.05 *P* values, respectively) with $n - 2$ degrees of freedom the *t* critical values for all four species were 1.65 and 1.96. H_o was rejected for *F. verticillioides* at 90% confidence levels, *F. circinatum* at both confidence levels and *F. graminiearum* at 90% confidence levels as their *t* statistic values were greater than the *t* critical values at the respective confidence levels. However, H_o was accepted for *F. oxysporum* because its *t* statistic value was less than both *t* critical values at the two levels of confidence.

Within-gene intron distribution

Analysis of the distribution of introns within the HK genes revealed a 5′ region positional bias (Figure 7). The largest proportion of the genes examined showed this intron positional bias where either all introns of a gene were located in the first third of a gene or more than 50% of its introns were located in this region. This bias was seen for genes harbouring one through to 15 introns. Equal numbers of introns in the first and last third of the gene (i.e. in the 5' and 3' region) were observed only in two and four-intron genes, while an approximately even distibution of introns accross the gene was observed for a small number of four-, five- and six-intron genes.

Intron phase

The analysis of intron phase allowed identification of potential sequencing errors involving mononucleotide tracts in five introns. In all but one of these instances, the sequence had one too many bases that, when deleted, generated an intron phase similar to those in the other species. The second intron of the *F. circinatum* gene encoding Heat Shock 70 kDA protein had an AAA motif and deletion of an Adenosine resulted in a phase 2 intron. The *F. circinatum* Superoxide Dismutase- mitochondrial precursor 2 gene had an extra Guanosine at position 47 in the gene and when deleted the first intron in this gene was phase 1. The *F. oxysporum* gene encoding Protein Phosphatase PP2A Regulatory Subunit A had a CCCC motif causing a stop codon in the middle of the gene and deletion of one Cytidine showed that the fourth intron of this gene was phase 0. The *F. graminearum* Glucose 6 Phosphate 1 Dehydrogenase gene had an extra Cytidine at position 336 and when deleted changed the phase of the second intron to 0. In the case of the *F. circinatum* gene encoding the alpha subunit of the AP 2 adaptor complex the possible correction of the Adenosine to a Guanosine at position 876 resulted in a phase 2 intron in this gene, similar to those of the other species.

Examination of intron phase in 50 HK genes revealed that phase 0, 1 and 2 introns were present in all four *Fusarium* species, albeit with different ratios (Supplementary material). Within our dataset, most introns were phase 0, which appeared to be true also for the three-, four-, five-seven-, eight-, nine- and twelve-intron genes (Figure 8). The respective percentage ratios of the three intron phases (0:1:2) in the four fungi were as follows: 39:34:27 for the examined 187 introns of *F. verticilliodes*, 39:35:26 for the 185 introns examined in *F. circinatum*, 39:36:25 for the 181 introns of *F. oxysporum* and 40:37:23 for the 177 introns in the *F. graminearum* 50 HK

gene dataset. The phases of all the introns examined were conserved between the four *Fusarium* species. The only two exceptions were from *F. graminearum*. For the gene encoding Glutamine Dependent NAD + Synthetase the eighth intron was phase 1, while the same intron in the other species was phase 0. Similarly, the fourteenth intron of this gene was phase 0 in *F. graminearum*, while it was phase 2 in the other three species. In addition, the analysis of the distribution of intron phase within the genes examined showed that phase 0 introns were mostly located closer to the 5' end of the genes than phase 1 and 2 introns (Figure 8A), and a non-uniform distribution of the intron phases was observed when the different intron positions were compared (Figure 8B).

For those introns (15) that were absent in one or two of the four *Fusarium* species, most appeared to be phase 2, three with phase 1 introns and two with phase 0 introns. Ten of these "missing" introns were ascribed to *F. graminearum* and five to *F. oxysporum*. Of the latter, two were shared with *F. graminearum* and two with *F. verticillioides*. None of the discrepancies were associated with *F. circinatum*. The genes included Proteasome Regulatory Particle subunit Rpn, ATP Synthase subunit beta, T Complex Protein 1 subunit zeta and Chromosome Segregation Protein sudA from *F. oxysporum*, and FK506 Binding Protein 1B, Glutamine dependent NAD + Synthetase, Adenosylhomocysteinase 2, Leukotriene A4 Hydrolase, Glucose 6 Phosphate 1 Dehydrogenase, Aspartyl tRNA Synthetase 1 and AP 1 Complex subunit mu from *F. graminearum*.

Intron cis-elements

Alignments of the 5′ splice site were performed with two sample sizes, first with 100 introns and second with 2022 introns. For the smaller dataset, analysis resulted in 5' splice site consensus with the sequence $A_{39}A_{38}G_{55}|G_{100}T_{100}A_{71}A_{40}G_{91}T_{57}$, where subscripts denote the percentage conservation of a base at a particular position in the 5′ splice site alignment and the vertical bar represents the exon-intron junction. Analysis of the larger dataset allowed refinement of this consensus by improving its resolution (Table 1). For example, the second nucleotide of the intron is in actual fact not always a Thymidine but sometimes a Cytidine.

Analysis of the 2022-introns dataset also allowed refinement of the 3′ splice and branch site motifs (Table 1). Alignments of these introns generated a well-resolved 3′ splice site motif consensus $Y_{93}A_{100}G_{100}|R_{59}$ (Y and R denote nucleotides with pyrimidine and purine bases) but revealed the presence of two variants (RAG|Y or RAG|R) in a small proportion of the introns (115, 5.69%) examined. With regards to the branch site motif, more than 90% of the sites had the consensus CTRAY, while three additional variants (TTRAY, CTHAH and ATCAT; H denotes a nucleotide with an adenosine, a Cytidine or a Thymidine base) were also detected in a small number of introns. The CTHAH branch site appears to be unique to *Fusarium*.

The polypyrimidine tract (a minimum of six consecutive nucleotides with at least three Thymidines and no Adenosines) was diverse among the four *Fusarium* species. In the dataset containing the 166 introns from 10 genes, the length of the predicted polypyrimidine tracts ranged from 6 to 25 nt. The number of predicted polypyrimidine tracts per intron also varied considerably. In some cases as many as nine were predicted, while in 24 instances no sequence resembling this element could be detected. More than half (53%) of the introns analyzed had either one or two polypyrimidine tracts. The predicted polypyrimidine tracts occupied two intron regions – the 5' region and the 3' region, where the branch site was the reference point (Kupfer *et al*., 2004). In all four *Fusarium* species, the majority of the predicted polypyrimidine tracts occupied the 5' region (*i.e*., 8 to 40 nt away from the 5' splice site). Only 14-20% of the predicted polypyrimidine tracts occupied the 3' region. Of the latter, *F. graminearum* had most predicted polypyrimidine tracts in the 3' region, while *F. oxysporum* had the fewest. In total, 61% of the examined introns had predicted polypyrimidine tracts exclussively in their 5' region, 35% exclussively in their 3' region and 5% in both the 5' and the 3' regions (Table 2).

For a subset of genes (28) with multiple introns, we also investigated whether specific forms of the four *cis*-elements found in one intron of a gene were similar to those of another intron in that gene. Overall we could not observe any such within-gene associations in terms of the specific form of the *cis*-elements (Table 1) and the size of the intron in the four *Fusarium* species. Therefore, the architecture of an intron in one position in a gene is not dependent on that of another intron in that gene.

Discussion

In this study, the basic structure and distribution of Spliceosomal introns were determined in four important *Fusarium* species. By making use of a set of HK genes we produced highly defined consensus sequences for the splice site junctions and internal *cis*-elements of Spliceosomal introns in these species. Apart from demonstrating a positional bias of introns in genes, we also showed that the length of introns is associated with both their positions in genes and the length of genes. In addition to improving our knowledge regarding the architecture of fungal Spliceosomal introns, incorporation of the findings presented here in *ab initio* gene prediction approaches will significantly improve the accuracy of genome annotation in *Fusarium* and related species.

No correlation was observed between genome size and intron length in the four *Fusarium* examined. Deutsch and Long (1999) suggested that an organism's genome size can influence the fixation of longer introns, e.g., the 3400 Mb human genome has a mean intron length of 3413.1 nt while the 13 Mb *Aspergillus* genome has a mean intron length of 72.2 nt. However, the mean intron length for all four of the *Fusarium* species were about 76 nt, despite the fact that their genome sizes differ up to 1.6-fold (Ter-Hovhannisyan *et al*, 2008; Ma *et al*. 2010; Wingfield *et al*., 2012). However, compared to higher eukaryotes, the *Fusarium* introns and those of other so-called lower eukaryotes are considerably shorter. For example, in fungi like *S. pombe* and *Aspergillus* (Lim and Burge, 2001) and protists (Russell *et al*., 1994), introns sizes are also relatively low ranging from 13 nt to 75 nt. The short introns in these lower eukaryotes are likely linked to their use of intron definition for locating intronic regions (Lim and Burge, 2001), which is dependent on optimally spaced splice junctions and internal *cis*-elements in the intron (Lim and Burge, 2001; Burge and Karlin, 1997; Salamov and Solovyev, 2000; Borodovsky and Mclninch, 1993). This is contrast to what happens in higher eukaryotes with longer introns, where splicing is facilitated *via* exon definition in which the spliceosome and other splicing regulatory proteins locate the exons before splicing occurs (Berget, 1995; Lim and Burge, 2001).

The mean lengths for the introns in *Fusarium* HK genes are shorter than those reported as a general average for fungal Spliceosomal introns. The latter has been suggested to be about 85 nt in length (Hawkin, 1988), which is comparable to the 83 nt mean intron length reported for *F.*

graminearum (Wong *et al*., 2011). The *Fusarium* HK genes, however, harbour introns with a mean length of 76 nt. This difference probably reflects the fact that constitutively expressed genes, such as the HK genes, have shorter introns to facilitate more efficient splicing (Comeron and Kreitman, 2000; Hurst *et al*., 1999). According to Castillo-Davis *et al*. (2002), natural selection could act to retain shorter introns or purge longer introns. In fact, with a eukaryotic transcription rate of 20 nt/sec (Ucker and Yamamoto, 1984; Izban and Luse, 1992) at an energy cost of two ATP/nt (Lehninger, 1985), shorter introns would allow for much shorter processing times and more efficient or rapid expression of the genes in which they occur. In our dataset, those 3% of genes lacking introns altogether would be the extreme case in point of expression streamlining.

Compared to other eukaryotes, the genes of *Fusarium* species and other fungi are characterized by relatively low Spliceosomal intron densities. Despite being within range of the 1.5 introns per gene that have been observed for other fungi (Jeffares *et al*., 2006), the average density of 2.53 introns per HK genes in *Fusarium* is much lower than those observed in higher eukaryotes. According to Jeffares *et al*. (2006), the relatively low intron densities in fungi correlate with their low complexity and short generation time. For example, *A. thaliana* and *Homo sapiens* with their considerably longer generation times, respectively, have 4.3 and 8.82 introns per gene (Jeffares *et al*., 2006). Also, the complexity of these higher eukaryotes has also been linked to the extra levels of gene expression regulation, where introns effect on transcription initiation, pre-mRNA polyadenylation, mRNA decay, mRNA transport and translation (Le Hir *et al*., 2003). However, not much is known regarding the contribution of intron-mediated expression regulation in fungi, and it remains to be determined whether the relatively low intron densities in these organisms are linked to their apparent independence of this form of expression regulation.

Consistent with what has been found in other eukaryotes (Bradnam and Korf, 2008), the first introns of the *Fusarium* HK genes were longer than those occurring in downstream positions. Because first introns are mostly in the 5' end of genes and could be “early” introns, which are thought to have been in existence in the progenote before the diversification of the three Domains of Life (Nguyen *et al*., 2006), they could have had adequate time to accumulate extra (“junk”) DNA (Bradnam and Korf, 2008; Wang and DePasse, 2012), which would have

increased their length. Furthermore, the first introns in the CDS and the 5′ untranslated region (which have not been analyzed in this study), have been hypothesized to help in gene expression (intron mediated enhancement) as they contain additional regulatory elements (Bradnam and Korf, 2008; Wang and DePasse, 2012). This form of expression enhancement has so far only been reported for plants (Parra *et al*., 2011) and further analysis of the sequences of the first introns in the HK of *Fusarium* species could therefore lead to interesting and insightful information in the future.

In this study we detected a 5′ positional bias for introns in the majority of *Fusarium* HK genes. Lin and Zhang (2005) attributed this to a preferential loss of introns in the 3′ region of genes, especially HK genes, during evolution. To explain how such losses may occur they proposed a mechanism based on homologous recombination between the genomic copies of the genes and their reverse transcribed spliced mRNAs (Lin and Zhang, 2005), although the molecular basis of this hypothesis remains to be determined. Within the *Fusarium* HK gene dataset, a number of genes (specifically those harbouring two and four introns) had a near-equal distribution of introns at both the 5′ and the 3′ regions. A similar trend was observed in certain genes of *N. crassa*, *M. grisea* and *F. graminearum*, where it was suggested that directed intron loss occurred in the middle of the genes (Nielsen *et al*., 2004).

Our comparison of intron positions in the four *Fusarium* species revealed that the positional conservation extended to the exact phase of the intron. Of 553 introns examined for phase, 551 were conserved between the four *Fusarium* species. Within the dataset of 50 genes, intron phase was also non-uniformly distributed, where phase 0 introns were more than phase 1 and 2 introns, and phase 1 introns were more than phase 2 introns. Both these findings are consistent with what has been observed before (Irimia and Roy, 2008; Fedorov *et al*., 1992; Qiu *et al*. 2004). In addition, most of the phase 0 introns examined for *Fusarium* appeared to be closer to the 5' end of the genes. The observation of an excess of phase 0 introns is in a genome has been used as support for the so-called “introns-late” theory which speculates that introns existed in the progenote before the diversification of the three Domains of Life (Nguyen *et al*., 2006; Roy, 2003). Also, 35% of extant introns are phase 0 and speculated to be an inherited state (Roy, 2003; De Souza *et al*., 1998) suggesting that phase 0 introns will be more prevalent than in phase

1 and 2. However "introns-late" theory does not explain why phase 1 introns are more prevalent than phase 2 introns. Therefore, proponents of the "introns-late" theory have proposed a site (the "proto- splice site") in which introns are inserted non-randomly to explain the high proportion of phase 0 introns and why phase 1 introns are more than phase 2 introns (Logsdon, 1998; Dibb and Newman, 1989; Long *et al*., 1998). In our data, we observed a higher rate of phase 2 intron loss in *F. graminearum* and *F.oxysporum*. Although there is no literature on phase 2 intron loss, it is possible that since these were introns located towards the 3' end of genes it can be explained by the homologous recombination between the genomic copies of the genes and their reverse transcribed spliced mRNAs mechanism (Qiu *et al*., 2004; Lin and Zhang, 2005).

As expected (Iwata and Gotoh, 2011), the 5′ splice site of the *Fusarium* introns was the most degenerate of the three main *cis*-elements required for Spliceosomal intron splicing. This is probably due to more nucleotides being part of the motif. However, when comparisons were made between the *Fusarium cis*-elements and those found in the genomes of other fungal genera, differences were found mainly in the 5′ splice site. *Aspergillus fumigatus, C. albicans, Cryptococcus neoformans, S. pombe* and *S. cerevisiae* have been reported to have the 5′ splice site sequence $A_{35}A_{39}G_{47}|G_{100}T_{99}R_{90}A_{56}G_{90}T_{72}$ (Bhasi *et al*., 2007). The major difference between the latter 5′ splice site consensus and that for the four *Fusarium* species is found at the third position of the intron where 90% of the time it is either a Guanosine or an Adenosine in the above *Aspergillus, Candida, Cryptococcus, Schizosaccharomyces* and *Saccharomyces* species but it is an Adenosine 74% of the time in *Fusarium* species.

Of the four motifs analyzed, the 3′ splice site was the least diverse. The *Fusarium* 3′ splice site HK gene consensus motif ($Y_{93}A_{100}G_{100}|R_{59}$) was highly similar to that found in *A. fumigatus, C. albicans, C. neoformans, S. pombe* and *S. cerevisiae* with their 3′ splice site consensus motif being $Y_{89}A_{100}G_{100}|R_{58}$ (Bhasi *et al*., 2007; Kupfer *et al*., 2004). The YAG motif has also been reported for introns in of Metazoa (Mount, 1982; Gates *et al*., 2011). In the current study, we also identify two additional 3' splice site motifs which are both supported by EST evidence. Both contain a RAG motif instead of the YAG motif, although a small proportion of those with the RAG motif has a Cytidine or a Thymidine base 3' to the spice site. Thus, the minimal requirement for a 3′ splice site for all fungal species analyzed thus far is four nucleotides

containing an Adenosine followed by a Guanosine at the second and third nucleotides, respectively.

The intron branch site motifs examined for the four *Fusarium* species was also relatively conserved. In addition to the most common CTRAY motif, our HK genes dataset also included introns with branch site motifs TTRAY (4.99 % of introns) and CTHAH (3.96 % of introns), as well as an ACCAT motif that occurred in 0.05% introns. Of these, the TTRAY motif has been reported for other fungi. Kupfer *et al*. (2004) indicated that this sequence present a secondary fungal branch site motif based on the genomes of *S. cerevisiae*, *S. pombe*, *A. nidulans*, *N. crassa*, and *Cryptococcus neoformans*. The CTHAH and ACCAT motifs appear to be new to Science. Whether these motifs are unique to *Fusarium* is not yet known.

The sequences of the 5' splice site and the branch site define the type of spliceosome needed for the splicing of the introns bearing these *cis*-elements (Wahl *et al*., 2009). During splicing the U1 and U2 spliceosomal components bind to the 5' splice site and the branch site in a sequential manner (Russel, 2006; Newman and Nagai 2010). In yeast, where the 5' splice site and the branch site are highly conserved, the sequence of these *cis*-elements is the only defining factor. However, in higher eukaryotes, where these sequences are more degenerate, additional factors such as splicing enhancers and silencers present on the pre-mRNA also influence the type of spliceosome needed (Wahl *et al*., 2009). Since the 5' splice site has been found to be degenerate and secondary branch site motifs have been found in the four *Fusarium* species examined, pre-mRNA splicing enhancers and silencers could also be involved in the definition of the type of spliceosome needed for the splicing in these fungi. Further research on pre-mRNA splicing enhancers and silencers in *Fusarium* species could shed light on this subject.

Consistent with what has been observed for other fungi (Kupfer *et al*., 2004), the polypyrimidine tract was the most diverse intron *cis*-element examined in the four *Fusarium* species. In the *Fusarium* HK genes dataset, this diversity was further emphasized by the multiplicity of potential polypyrimidine tracts predicted for single genes. For example, in one alignment of the HK genes of the four species, *F. verticillioides* had six different predicted polypyrimidine tracts, *F. circinatum* nine, *F. oxysporum* four and *F. graminearum* six. However, the predicted

polypyrimidine tract was predominantly found at the 5′ region of introns in *Fusarium*. This is in contrast to what has been found in metazoa where the polypyrimidine tract is found mainly in the 3′ region of introns (Banerjee *et al*., 2004). The predominance of the polypyrimidine tract at the 5′ region of introns has also been reported for, and in the introns of *S. cerevisiae, S. pombe, A. nidulans, N. crassa* and *C. neoformans* (Kupfer *et al*., 2004). Such diversity in the sequence, length and position of the polypyrimidine tract, suggests that the spliceosomal machinery for different organisms differ markedly, either in terms of the constituents of the spliceosome itself, or in terms of the specificity. For example, experimental work has shown that polypyrimidine tracts are not always essential for splicing, but when they are present protein U2AF^{65} of the spliceosome binds to it, subsequently allowing more efficient splicing (Banerjee *et al*., 2004). A more detailed analysis of the polypyrimidine tracts in the Spliceosomal itrons of *Fusarium* and other fungi would undoubtedly shed light not only on the splicing mechanisms in these organisms, but potentially allow identification of novel regulatory targets for gene expression.

When the overall intron structure of the HK genes of the four *Fusarium* species was compared, *F. graminearum* often had a different structure. It contributed to about 60% of the intron position incongruences between the four species and its intron lengths were mostly divergent from those of the other three species. Also, in instances where the other three species had an alternative branch site sequence, *F. gramineurum* had the canonical CTRAY motif, and *vice versa*. In contrast, the structure of introns in *F. circinatum*, *F. verticillioides* and *F. oxysporum* often resembled one another, with the structure and sequence of introns in *F. circinatum* being highly similar to those of *F. verticillioides*. Indeed, these similarities and differences reflect the known evolutionary relationships among these fungi (Ma *et al*., 2010; Kumar *et al*., 2010; Wingfield *et al*., 2012; http://www.broadinstitute.org/annotation/genome/fusarium_group).

In this study, we show that publicly available genome annotations for the four *Fusarium* species include numerous erroneously annotated introns. Although an annotated reference genome can aid in the annotation of the newly sequenced genome of one or more close relatives (Irimia and Roy, 2008), gene prediction software may still make mistakes. This is because the genomes of the relatives may not experience similar evolutionary rates and pressures, and also because existing gene prediction methods mostly have been developed from previously sequenced

genomes. Therefore, by incorporating the findings of this study into existing gene finding procedures, it should be possible to produce highly accurate annotations for *Fusarium* species. This is specifically true in terms of the branch site motif (*i.e*., CTRAY, TTRAY, CTHAH and ACCAT), the 3' splice site that should contain a minimum of four nucleotides with the canonical AG dinucleotide at the second and third positions, and the 5′ splice site that should contain a minimum nine nucleotides: three from the downstream exon and the intron dinucleotide which can either be a GT or a GC together with the four nucleotides following it. These data on the minimal requirements for splicing of the introns could also be used to restrict intron length in gene prediction programs (Lim and Burge, 2001; Burge and Karlin, 1997; Salamov and Solovyev, 2000; Borodovsky and Mclninch, 1993). According to Nguyen *et al*. (2006), the incorporation of intron-phase prediction algorithms in annotation software can also increase accuracy and even help correct sequencing errors that appear as indels or substitutions. Many research projects depend on publicly available genomes, and these improvements to fungal gene prediction methods will reduce the descrepencies resulting from unspecificity during fungal genome annotations thereby increasing their reliability.

References

1. Arthur J: QI Macros for All Versions of Excel 2000-2011. 2012:1-36.
2. Banerjee H, Rahn A, Gawade B, Guth S, Valcarcel J, Singh R: The conserved RNA recognition motif 3 of U2 snRNA auxiliary factor (U2AF65) is essential in vivo but dispensable for activity in vitro. RNA 2004, 10(2):240-253.
3. Berget SM: Exon recognition in vertebrate splicing. Journal of biological Chemistry 1995, 270(6):2411-2431.
4. Bhasi A, Pandey RV, Utharasamy SP, Senapathy P: EuSplice: a unified resource for the analysis of splice signals and alternative splicing in eukaryotic genes. Bioinformatics 2007, 23(14):1815-1823.
5. Bhattacharya D, Lutzoni F, Reeb V, Simon D, Nason J, Fernandez F: Widespread occurrence of spliceosomal introns in the rDNA genes of ascomycetes. Molecular Biology and Evolution 2000, 17(12):1971-1984.
6. Borodovsky M, McIninch J: GENEMARK: parallel gene recognition for both DNA strands. Computers & Chemistry 1993, 17(2):123-133.
7. Bradnam KR, Korf I: Longer first introns are a general property of eukaryotic gene structure. PLoS One 2008, 3(8):e3093.
8. Burge C, Karlin S: Prediction of complete gene structures in human genomic DNA1. Journal of Molecular Biology 1997, 268(1):78-94.
9. Cantarel BL, Korf I, Robb SMC, Parra G, Ross E, Moore B, Holt C, Alvarado AS, Yandell M: MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Research 2008, 18(1):188-196.
10. Castillo-Davis CI, Mekhedov SL, Hartl DL, Koonin EV, Kondrashov FA: Selection for short introns in highly expressed genes. Nature Genetics 2002, 31(4):415-418.
11. Comeron JM, Kreitman M: The correlation between intron length and recombination in Drosophila: dynamic equilibrium between mutational and selective forces. Genetics 2000, 156(3):1175-1190.
12. De Souza SJ, Long M, Klein RJ, Roy S, Lin S, Gilbert W: Toward a resolution of the introns early/late debate: only phase zero introns are correlated with the structure of ancient proteins. Proceedings of the National Academy of Sciences 1998, 95(9):5094-5099.

13. Deutsch M, Long M: Intron-exon structures of eukaryotic model organisms. Nucleic Acids Research 1999, 27(15):3219-3228.
14. Dibb N, Newman A: Evidence that introns arose at proto-splice sites. The EMBO Journal 1989, 8(7):2015-2021.
15. Fedorov A, Suboch G, Bujakov M, Fedorova L: Analysis of nonuniformity in intron phase distribution. Nucleic acids research 1992, 20(10):2553-2557.
16. Gaffney DJ, Keightley PD: Genomic selective constraints in murid noncoding DNA. PLoS Genetics 2006, 2(11):e204.
17. Gates DP, Coonrod LA, Berglund JA: Autoregulated Splicing of muscleblind-like 1 (MBNL1) Pre-mRNA. Journal of Biological Chemistry 2011, 286(39):34224-34233.
18. Hall TA: BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In: 1999; 1999: 95-98.
19. Hawkins JD: A survey on intron and exon lengths. Nucleic Acids Research 1988, 16(21):9893-9908.
20. Holt C, Yandell M: MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. BMC Bioinformatics 2011, 12(1):491-504.
21. Hong X, Scofield DG, Lynch M: Intron size, abundance, and distribution within untranslated regions of genes. Molecular Biology and Evolution 2006, 23(12):2392-2407.
22. Hurst LD, Brunton C, Smith N: Small introns tend to occur in GC-rich regions in some but not all vertebrates. Trends In Genetics 1999, 15(11):437-439.
23. Irimia M, Roy SW: Spliceosomal introns as tools for genomic and evolutionary analysis. Nucleic Acids Research 2008, 36(5):1703-1712.
24. Iwata H, Gotoh O: Comparative analysis of information contents relevant to recognition of introns in many species. BMC genomics 2011, 12(1):45-61.
25. Izban M, Luse D: Factor-stimulated RNA polymerase II transcribes at physiological elongation rates on naked DNA but very poorly on chromatin templates. Journal of Biological Chemistry 1992, 267(19):13647-13655.
26. Jeffares DC, Mourier T, Penny D: The biology of intron gain and loss. TRENDS in Genetics 2006, 22(1):16-22.
27. Korf I: Gene finding in novel genomes. Bmc Bioinformatics 2004, 5(1):59-67.

28. Kriventseva E, Gelfand M: Statistical analysis of the exon-intron structure of higher and lower eukaryote genes. Journal of Biomolecular Structure & Dynamics 1999, 17(2):281-288.
29. Kumar L, Breakspear A, Kistler C, Ma LJ, Xie X: Systematic discovery of regulatory motifs in *Fusarium graminearum* by comparing four *Fusarium* genomes. BMC genomics 2010, 11(1):208-220.
30. Kupfer DM, Drabenstot SD, Buchanan KL, Lai H, Zhu H, Dyer DW, Roe BA, Murphy JW: Introns and splicing elements of five diverse fungi. Eukaryotic Cell 2004, 3(5):1088-1100.
31. Kvas M, Marasas W, Wingfield BD, Wingfield MJ, Steenkamp ET: Diversity and evolution of Fusarium species in the Gibberella fujikuroi complex. Fungal Diversity 2009, 34:1-21.
32. Le Hir H, Nott A, Moore MJ: How introns influence and enhance eukaryotic gene expression. Trends in biochemical sciences 2003, 28(4):215-220.
33. Lehninger A: Principles of biochemistry.(1982). Translated under the title Osnovy biokhimii, Moscow: Mir 1985, 1:176-184.
34. Leslie JF, Summerell BA, Bullock S: The Fusarium laboratory manual, vol. 2: Wiley Online Library; 2006.
35. Lewis SE, Searle S, Harris N, Gibson M, Lyer V, Richter J, Wiel C, Bayraktaroglir L, Birney E, Crosby M: Apollo: a sequence annotation editor. Genome Biol 2002, 3(12):1-14.
36. Lim LP, Burge CB: A computational analysis of sequence features involved in recognition of short introns. Proceedings of the National Academy of Sciences of the United States of America 2001, 98(20):11193-11198.
37. Lin K, Zhang DY: The excess of 50 introns in eukaryotic genomes. Nucleic Acids Research 2005, 33:6522–6527.
38. Logsdon JM: The recent origins of spliceosomal introns revisited. Current opinion in genetics & development 1998, 8(6):637-648.
39. Long M, De Souza SJ, Rosenberg C, Gilbert W: Relationship between “proto-splice sites” and intron phases: evidence from dicodon analysis. Proceedings of the National Academy of Sciences 1998, 95(1):219-223.
40. Ma LJ, Van Der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, Di Pietro A, Dufresne M, Freitag M, Grabherr M, Henrissat B: Comparative genomics reveals mobile pathogenicity chromosomes in Fusarium. Nature 2010, 464(7287):367-373.

41. Misiek M, Hoffmeister D: Processing sites involved in intron splicing of Armillaria natural product genes. Mycological Research 2008, 112(2):216-224.
42. Mount SM: A catalogue of splice junction sequences. Nucleic Acids Research 1982, 10(2):459-472.
43. Newman AJ, Nagai K: Structural studies of the spliceosome: blind men and an elephant. Current opinion in structural biology 2010, 20(1):82-89.
44. Nguyen H, Yoshihama M, Kenmochi N: Phase distribution of spliceosomal introns: implications for intron origin. BMC Evolutionary Biology 2006, 6(1):69-77.
45. Nielsen CB, Friedman B, Birren B, Burge CB, Galagan JE: Patterns of intron gain and loss in fungi. PLoS Biology 2004, 2(12):e422.
46. Parra G, Bradnam K, Korf I: CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes. Bioinformatics 2007, 23(9):1061-1067.
47. Parra G, Bradnam K, Rose AB, Korf I: Comparative and functional analysis of intron-mediated enhancement signals reveals conserved features among plants. Nucleic Acids Research 2011, 39(13):5328-5337.
48. Qiu WG, Schisler N, Stoltzfus A: The evolutionary gain of spliceosomal introns: sequence and phase preferences. Molecular Biology and Evolution 2004, 21(7):1252-1263.
49. Rose AB: Requirements for intron-mediated enhancement of gene expression in *Arabidopsis*. RNA 2002, 8(11):1444-1453.
50. Roy SW: Recent evidence for the exon theory of genes. Genetica 2003, 118(2):251-266.
51. Russell CB, Fraga D, Hinrichsen RD: Extremely short 20–33 nucleotide introns are the standard length in Paramecium tetraurelia. Nucleic Acids Research 1994, 22(7):1221-1225.
52. Russell PJ: iGenetics: a Mendelian approach: Pearson/Benjamin Cummings; 2006.
53. Salamov AA, Solovyev VV: *Ab initio* gene finding in Drosophila genomic DNA. Genome Research 2000, 10(4):516-522.
54. Saldanha R, Mohr G, Belfort M, Lambowitz AM: Group I and group II introns. The FASEB journal 1993, 7(1):15-24.
55. Samuels ML: Statistics for life sciences, Third edn; 2003.
56. Smith M: Structure of vertebrate genes: a statistical analysis implicating selection. Journal of molecular evolution 1988, 27(1):45-55.

57. Solovyev VV, Salamov AA, Lawrence CB: The prediction of human exons by oligonucleotide composition and discriminant analysis of spliceable open reading frames. In: 1994; 1994: 354-362.
58. Stanke M, Diekhans M, Baertsch R, Haussler D: Using native and syntenically mapped cDNA alignments to improve de novo gene finding. Bioinformatics 2008, 24(5):637-644.
59. Stanke M, Waack S: Gene prediction with a hidden Markov model and a new intron submodel. Bioinformatics 2003, 19(suppl 2):ii215-ii225.
60. Ter-Hovhannisyan V, Lomsadze A, Chernoff YO, Borodovsky M: Gene prediction in novel fungal genomes using an *ab initio* algorithm with unsupervised training. Genome Research 2008, 18(12):1979-1990.
61. Thompson JD, Higgins DG, Gibson TJ: CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Research 1994, 22(22):4673-4680.
62. Ucker D, Yamamoto K: Early events in the stimulation of mammary tumor virus RNA synthesis by glucocorticoids. Novel assays of transcription rates. Journal of Biological Chemistry 1984, 259(12):7416-7420.
63. Wahl MC, Will CL, Lührmann R: The spliceosome: design principles of a dynamic RNP machine. Cell 2009, 136(4):701-718.
64. Wang B, Mason DePasse J, Watt WB: Evolutionary Genomics of Colias Phosphoglucose Isomerase (PGI) Introns. Journal of Molecular evolution 2012:1-16.
65. Wingfield BD, Steenkamp ET, Santana QC, Coetzee M, Bam S, Barnes I, Beukes CW, Chan WY, De Vos L, Fourie G, *et al.*: First fungal genome sequence from Africa: A preliminary analysis. South African Journal of Science 2012, 108(1/2):9 pages.
66. Wong P, Walter M, Lee W, Mannhaupt G, Münsterkötter M, Mewes HW, Adam G, Güldener U: FGDB: revisiting the genome annotation of the plant pathogen Fusarium graminearum. Nucleic Acids Research 2011, 39(suppl 1):D637.

Tables

Table 1. The length of introns and a summary of the motifs[a] examined in 2022 introns from 226 Housekeeping genes in four *Fusarium* species

Intron length	***5′ splice site motif***	***Polypyrimidine tract (PPT)***	***Branch site motif*** [b]	***3′ splice site motif*** [c]
42-529 nucleotides	$A_{38}A_{38}G_{53}\|G_{100}T_{99}A_{74}A_{42}G_{93}T_{66}$	83% located between 5' splice site and the branch site 17% located between the branch site and the 3' splice site	CTR<u>A</u>Y (91%) CTH<u>A</u>H (4.99%) TTR<u>A</u>Y (3.96%) ACC<u>A</u>T (0.05%)	$Y_{93}A_{100}G_{100}\|R_{59}$ YAG|R (94.31%) RAG|R (3.51%) RAG|Y (2.18%)

[a] Subscript digits following individual bases indicate the proportion (in percentage) of occurrence of the base in that position.
[b] The proportion of the introns in which a specific branch site motif was observed is indicated in parentheses. Alternative branch site sequences: CTHAH represents CTTAC, CTCAA, CTAAA and CTCAT; TTRAY represents TTAAC, TTAAT, TTGAC, and TTGAT*. All the predicted branch site motifs were supported by EST data, except for TTGAT. Within the sequences, R, H and Y represent standard IUPAC codes for degenerate nucleotides, where R represent a nucleotide with either Guanosine or Adesnosine bases, Y represents either Cytdine or Thymidine bases, and H is a nucleotide an Adesnosine, Cytidine, or Thymidine base.
[c] The proportion of the introns in which a specific 3’ splice site was observed is indicated in parentheses.

Table 2. The percentage of introns with polypyrimidine tracts (PPTs) in the 5' region only, in the 3' region only, and in both the 5' and 3' regions in the four species of *Fusarium*.

Species	% of introns with PPTs in 5' region	% of introns with PPTs in 3' region	% of introns with PPTs in both the 5' and 3' regions
F. verticillioides	60	5	35
F. circinatum	58	11	31
F. oxysporum	70	0	30
F.graminearum	55	0	45

The 5′ region is the region between the 5′ splice site and the branch site and the 3′ region is the region between the branch site and the 3′ splice site. PPTs = polypyrimidine tracts.

Figures

Figure 1. The three main *cis*-elements found in Eukaryotes. The 5' splice site, branch site and 3' splice site are shown from left to right. The main nucleotides of the intron splice sites are underlined. Blue boxes represent exons and the solid lines represent the flanked intron. R= purine and Y= pyrimidine. (Kupfer *et al*., 2004; Bhasi *et al*., 2007)

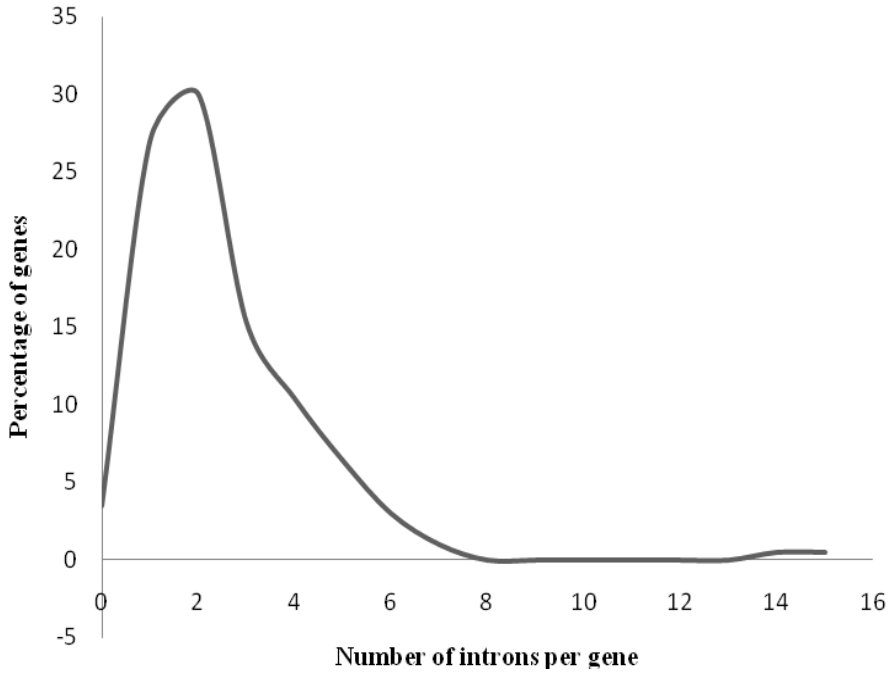


Figure 2. The frequency of introns in 226 HK genes of four *Fusarium* species.

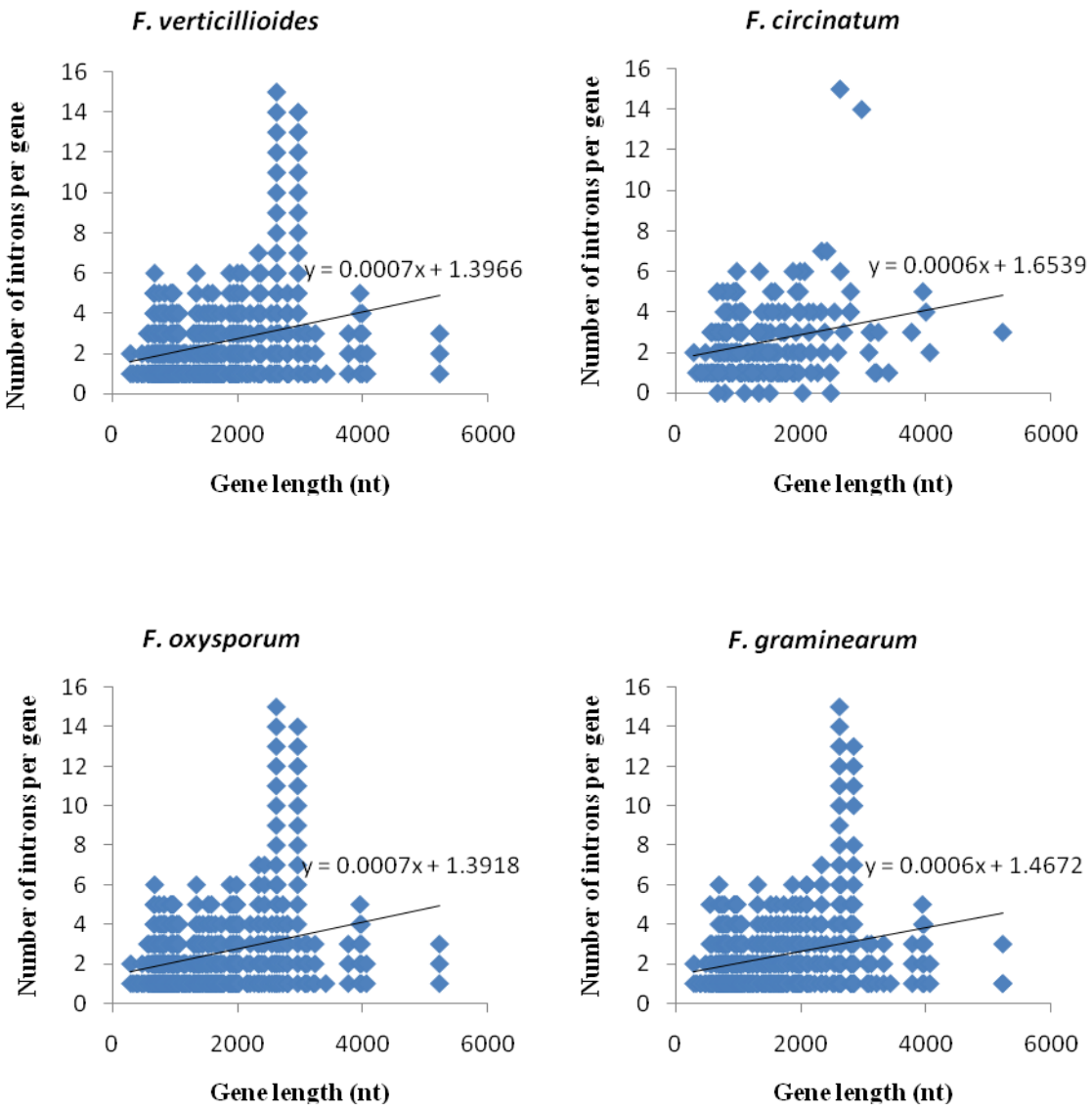


Figure 3. The relationship between gene length and number of introns per gene of *Fusarium* HK genes. The regression line and its equation are shown in the figures.

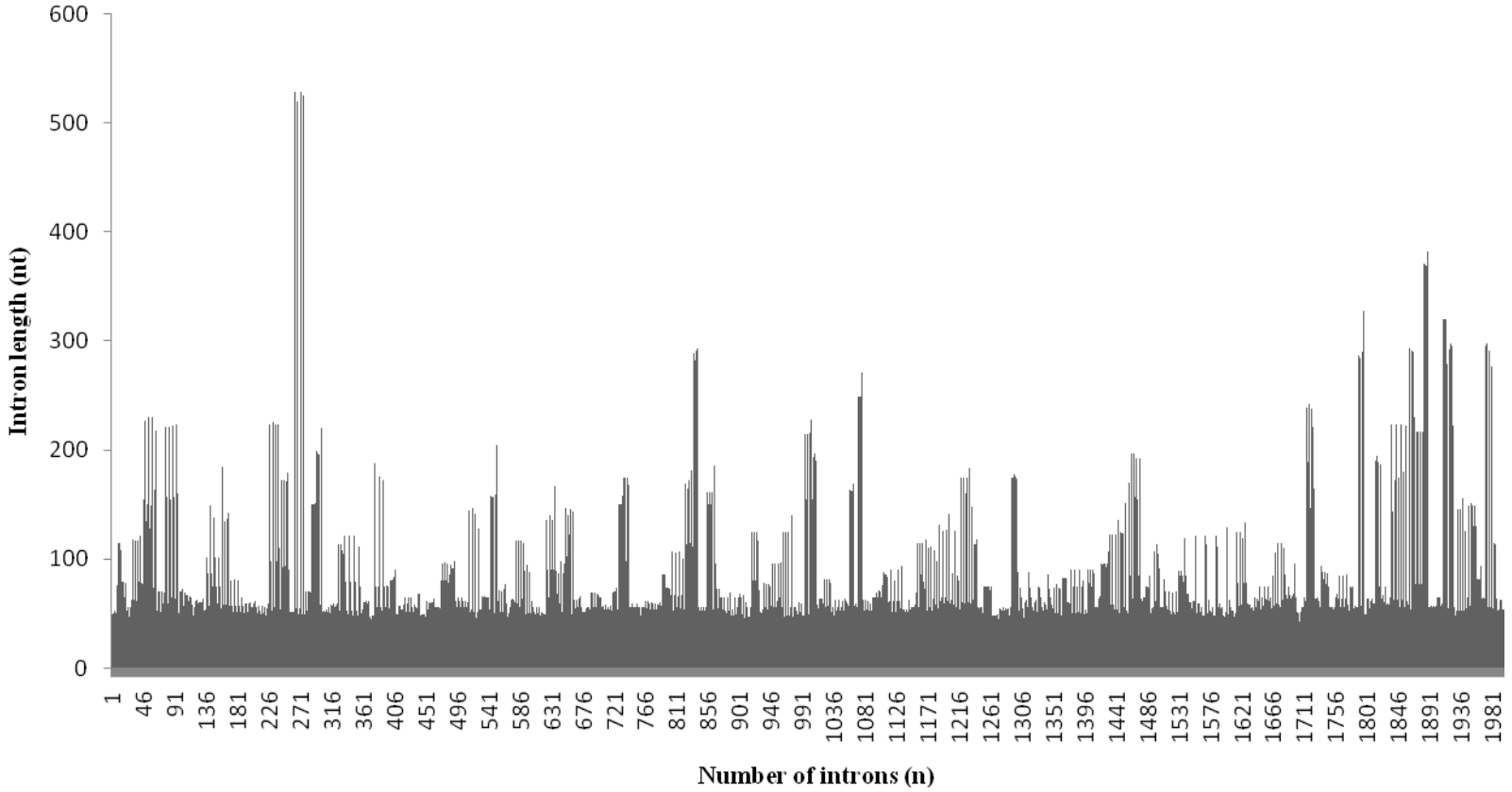


Figure 4. The lengths of all introns within the set of 226 HK genes of four species of *Fusarium*.

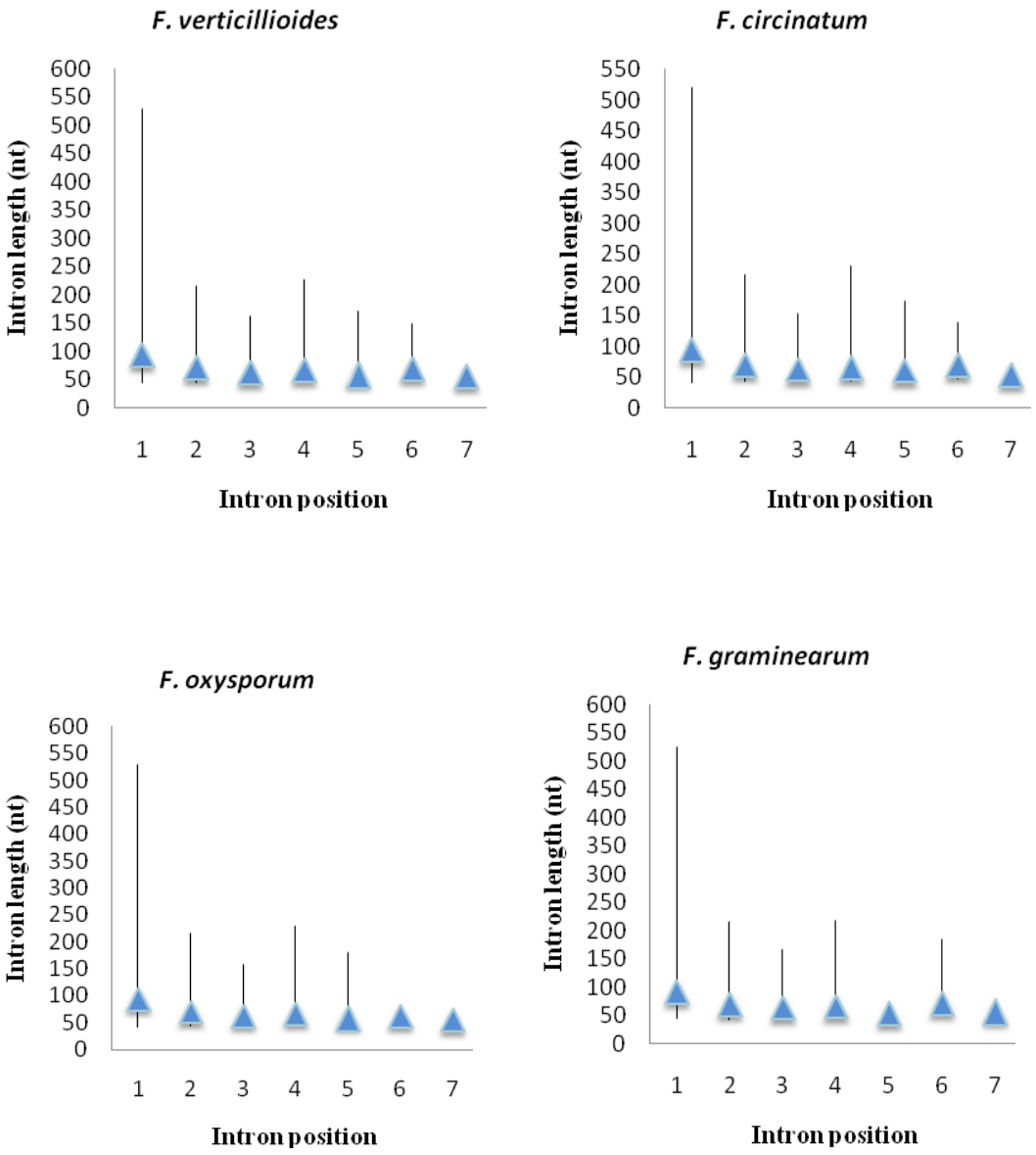


Figure 5. The relationship between intron position and intron length in the four *Fusarium* species. The charts were plotted with three data points (High, Low and Mean intron lengths) on the *y axes* for each intron position on the *x axes*. The vertical lines represent the High and Low intron lengths and the blue triangles represent the mean values. An Analysis of Variance (ANOVA) and Tukey's Honesty Significant Difference (HSD) tests showed that first-intron mean lengths were significantly different from intron 2 to intron 7 mean lengths ($P = 0.05$).

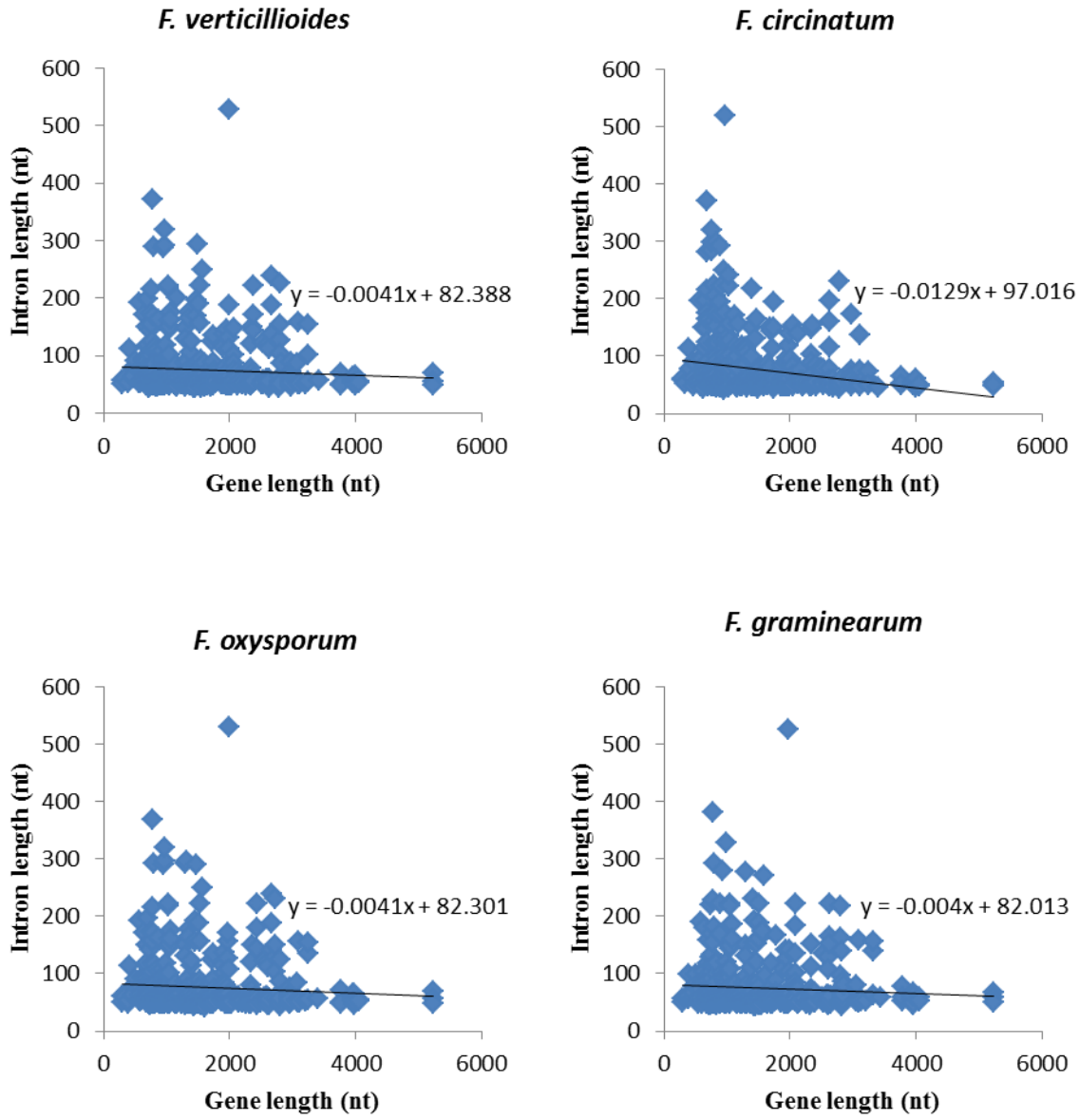


Figure 6. The relationship between gene and intron length within the HK genes of the four *Fusarium* species. The regression line and its equation are shown in the figures.

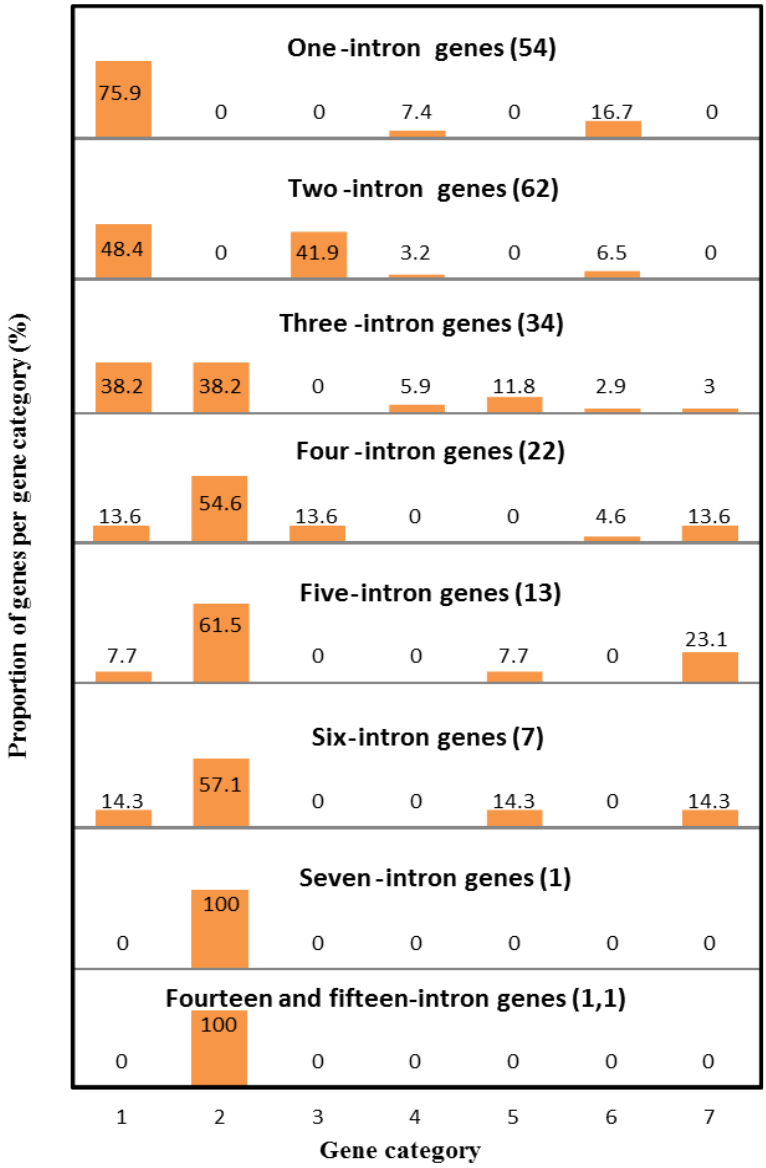
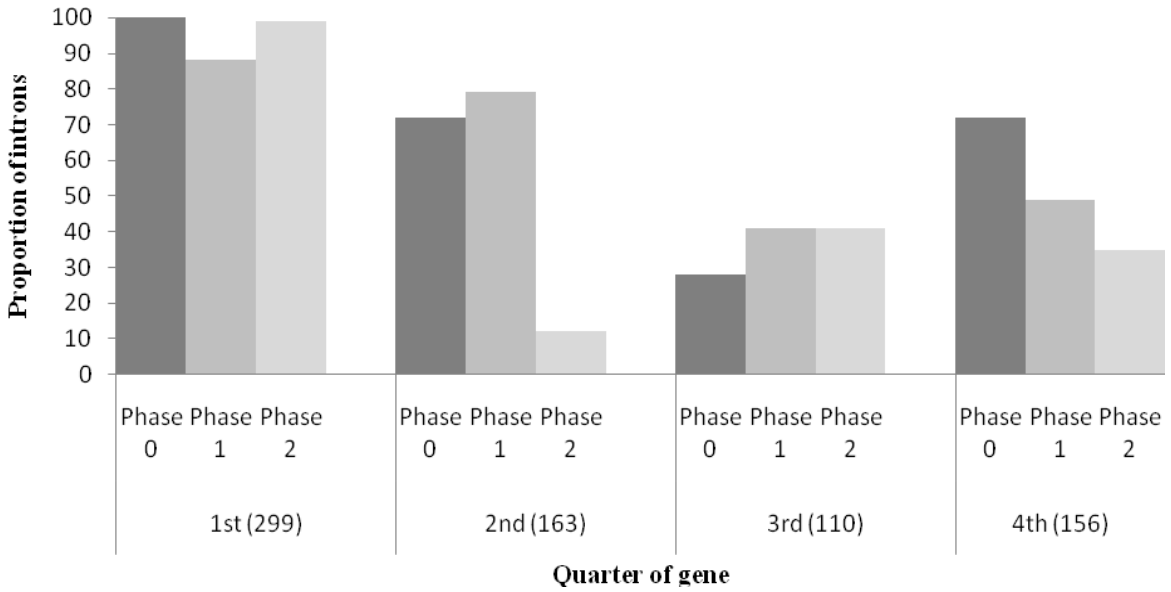
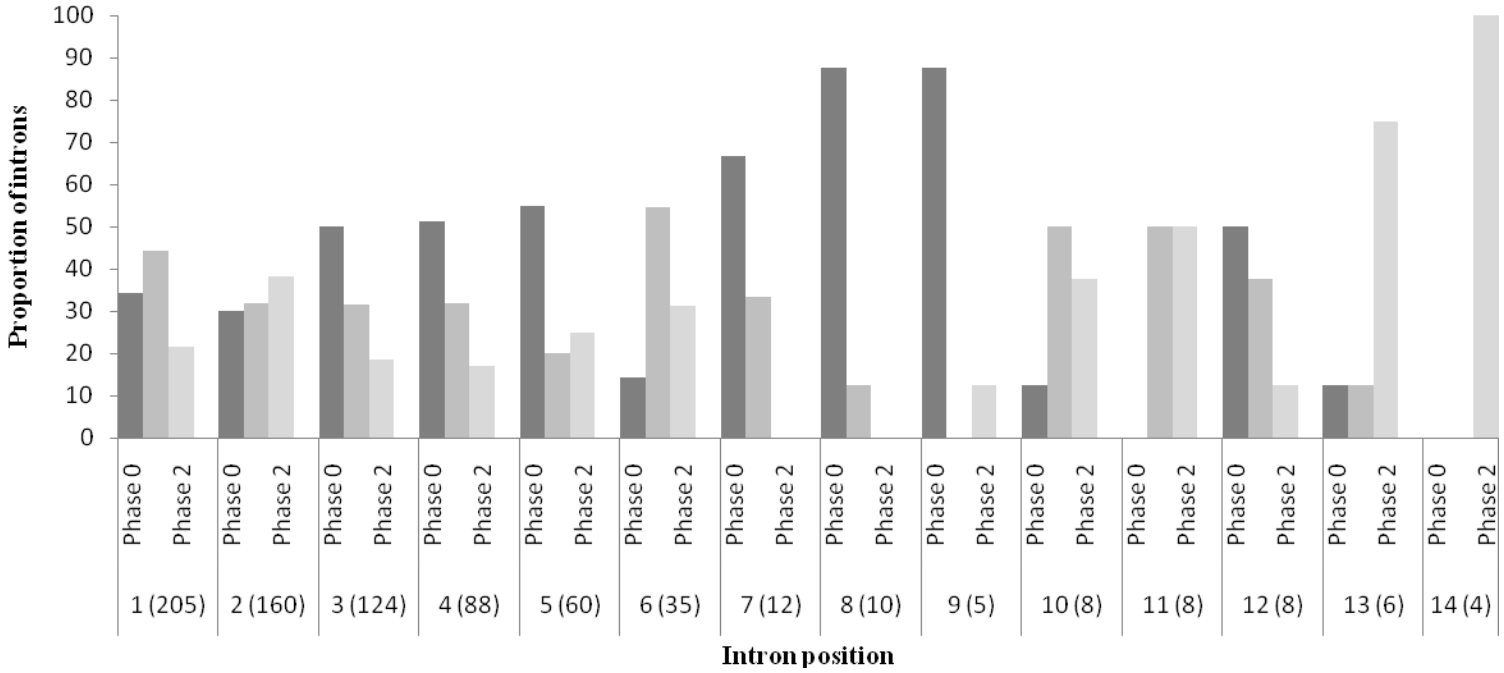


Figure 7. The distribution of introns within the set of 226 HK genes of the four *Fusarium* species. The genes were divided into three regions: the 5′ region (the first third of the ORF), the middle region (the second third of the ORF) and the 3′ region (the last third of the ORF). 1 = All introns are at 5′ region; 2 = >50% of the introns are in the 5′ region; 3 = 50% of the introns in the 5′ region and 50% in the 3′ region; 4 = All introns are in the middle region; 5 = >50% of the introns are in the 3′ region; 6 = All introns are in the 3′ region; 7 = Introns evenly distributed across gene (no concentration of introns at a particular gene region). The numbers in parentheses are the number of genes per gene category.



A



B

Figure 8. The distribution of intron phases within the set of 50 HK genes of the four *Fusarium* species. **A**) Distribution within the gene. **B**) Distribution in terms of intron position. The numbers in parentheses on the *x-axis* show the number of introns per gene location.

Chapter 3: *In silico* identification and characterization of alternative splicing in the *Fusarium circinatum* genome

Abstract

Alternative splicing is a phenomenon which provides eukaryotic cells with a diversity of proteins without the need for extra gene copies. Higher eukaryotes have more than 90% of their genes undergoing alternative splicing, and only up to 4.2% of alternatively spliced genes have been discovered for fungi. Alternative transcripts can sometimes harbour premature stop codons which may lead to the production of truncated proteins deleterious to the eukaryotic cell. These transcripts are degraded by a quality control system referred to as non-sense mediated mRNA decay. In this study, alternative splicing was sought in the genome of *Fusarium circinatum*. We used EST data generated from a strain of this organism grown under carbon and nitrogen starvation, as well as three additional strains grown in complete medium together with the genome sequence information for this fungus to detect alternative transcripts. We found alternative transcripts for genes involved in diverse metabolic processes, although we only detected transcripts with retained introns in *F. circinatum*. Transcripts bearing premature stop codons and genes involved in the non-sense mediated mRNA decay mechanism were also found. These results indicate that alternative splicing does occur in the *F. circinatum* genome on 0.3% of the genes and that potentially deleterious transcripts are under surveillance.

Introduction

Eukaryotic genes are made up of coding and non-coding sequences. The coding sequences or exons form part of the mature mRNAs that are translated into proteins. The non-coding sequences or introns are spliced from pre-mature mRNA before translation (Bhattacharya *et al*., 2000). Despite the fact that they are apparently non-coding, introns have a variety of roles in Eukaryotes. In yeast mitochondrial DNA, for example, several Group I and Group II introns have open reading frames that encode maturases, which function in the splicing of introns encoding them (Burke, J.M. 1988; Lambowitz *et al*., 1990; Saldanha *et al*., 1993). Group I and Group II introns have also been implicated in endonucleolytic cleavage of RNA and DNA, RNA polymerization, nucleotide transfer, templated RNA ligation, and aminoacyl-ester cleavage (Cech, 1990; Piccirilli, 1992). Spliceosomal introns have also been found to effect eukaryotic gene expression in a number of ways. These include transcription initiation, pre-mRNA polyadenylation, mRNA decay, mRNA transport and translation, intron mediated enhancement and alternative splicing (Mascarenhas *et al*., 1990; Le Hir *et al*., 2003; Conti and Izaurralde, 2005; Parra *et al*., 2011).

Alternative splicing was discovered in 1978 by Walter Gilbert, shortly after the discovery of introns. He saw that different combinations of the exons and introns of the Adenovirus *hexon* gene could lead to the production of mosaic mRNAs that could be translated into different proteins. This phenomenon in which changes in the processing of introns can lead to the production of multiple different mRNAs from a single gene was termed alternative splicing. Alternative intron splicing occurs in different frequencies in higher and lower eukaryotes. In humans, for example, as many as 90% of the genes undergo alternative intron splicing (Keren and Ast, 2010), while only 4.2% of fungal genes undergo alternative splicing (Loftus *et al*., 2005). These differences are probably due to differences in the number of introns and exons per gene and the average length of introns and exons (Modrek and Lee, 2003). During constitutive splicing, higher eukaryotes use exon definition (ED) to locate their short exons for inclusion in the mature mRNA, while lower eukaryotes use intron definition (ID) to locate their short introns for exclusion from the mature mRNA (Berget, 1995). The mechanism of constitutive splicing thus affects the mechanism of alternative splicing (Ast, 2004; Graveley, 2001).

Four mechanisms of alternative splicing are known: exon skipping, intron retention, alternative 5' and 3' splice site usage, and mutually exclusive exons (Smith and Valcarcel, 2000). Although higher and lower eukaryotes use these mechanisms in different ratios, alternative splicing is commonly achieved through the use of alternative 5' and 3' splice site usage and mutually exclusive exons at lower ratios than through exon skipping and intron retention where latter mechanisms are dominant (McGuire *et al*., 2008). In higher eukaryotes exon skipping occurs more frequently as these organisms employ ED during constitutive splicing (McGuire *et al*., 2008). This is because their splicing machinery identifies exons and if the exons have mutations near exon-intron boundaries they might be skipped. This is generally different from the alternatively spliced introns of lower eukaryotes where the ID-dependent splicing mechanism causes mRNAs to mostly retain introns (McGuire *et al*., 2008).

The processing of mRNA happens at different rates for different genes (Ho *et al*., 2007). This complicates the detection of alternative splicing. For instance, at the time of RNA extraction during expression experiments some transcripts may be unprocessed (*i.e*., containing all exons and introns) some partially processed and some fully processed (*i.e*., bearing the 5' cap and the poly-A tail) (Ho *et al*., 2007). To circumvent this problem in the generation of EST libraries, techniques have been developed to target processed mRNA using primers targeting the mRNA poly-A tail. In addition, some transcripts may have been degraded by the Non-sense mediated mRNA (NMD) system, which is a quality control system that degrades alternative transcripts bearing premature stop codons that could lead to the production of truncated proteins that may be deleterious to a cell or an organism (Mekouar *et al*., 2010). Therefore, unless experiments are explicitly designed for studying alternative splicing, results may be significantly influenced by the processing rates of mRNA molecules.

In the recent years developments in sequencing techniques have increased our knowledge regarding the frequency and mechanisms of alternative splicing in eukaryotes (Mardis, 2008; Metzker, 2009). Next generation sequencing technologies have led to an increased detection of different mRNA species in the human transcriptome (Pan *et al*., 2008; Wang *et al*., 2008). As more experiments are conducted under different cell growth conditions, the picture of alternative splicing of less complex organisms such as fungi is also increasingly becoming clearer (Mekouar

et al., 2010). However, to handle the large volumes of data produced by next generation sequencing technologies, data are often partitioned in *in silico* experiments, which might limit the detection of the true frequency of alternative splicing. Thus, a lot of research is needed in order to improve our ability to detect and quantify the contribution of alternative intron splicing to gene expression in eukaryotes, particularly lower eukaryotes.

Alternative splicing has long been thought not to occur in fungi as these organisms are less complex than higher eukaryotes (Barrass and Bleggs, 2003, Ho *et al*., 2007). So far this phenomenon has only been extensively analyzed in *Magnaporthe grisea* (Phylum *Ascomycota*, Class *Sordariomycetes*), *Cryptococccus neoformans* (Phylum *Basidiomycota*, Class *Tremellomycetes*), *Ustilago maydis* (Phylum *Basidiomycota*, Class *Ustilagomycetes*), *Aspergilus flavus* (Phylum *Ascomycota*, Class *Eurotiomycetes*) and *Yarrowia lipolytica* (Phylum *Ascomycota*, Class *Saccharomycetes*) (Ebbole *et al*., 2004; Loftus *et al*., 2005; Ho *et al*., 2007; Chang *et al*., 2010; Mekouar *et al*., 2010). The conclusion from these studies was that alternative splicing does occur in fungi, but at very low frequency ranging from about 1.6% in *M. grisea* to 4.2% in *C. neoformans* (Ebbole *et al*., 2004; Loftus *et al*., 2005). Whether these data could be extrapolated to other fungi remains unclear.

There has not been a thorough study of alternative splicing in *Fusarium* species (Phylum *Ascomycota*, Class *Sordariomycetes*). This is despite the fact that genomic DNA and EST data are available for four plant pathogenic *Fusarium* species with varying levels of host adaptation and specificity (Ma *et al*., 2010; Wingfield *et al*., 2012). The main aim of this study was therefore to determine whether or not alternative splicing does occur in *Fusarium* species by making use of *F. circiantum* as a reference. If this form of intron splicing indeed occurs in these species, our secondary aims were to determine the actual mechanism(s) employed and to evaluate the extent to which alternative transcripts might be subject to NMD by identifying premature stop codons in alternatively spliced transcripts and the genes responsible for NMD.

Materials and methods

Fusarium circinatum cDNA library construction and sequencing

The *F. circinatum* EST data used in this study were obtained from two independent studies. In the first study, a pathogenic strain of *F. circinatum* (FSP34) was used to generate two cDNA libraries from carbon and nitrogen starved cultures. The libraries were constructed by growing cultures first on complete medium and then transferring them onto minimal medium lacking a carbon and onto minimal medium lacking a nitrogen source. Total RNA was extracted using TRI Reagent (Sigma) and Pure Zol RNA Isolation Reagent (Bio-Rad). The total RNA was purified using RNeasy Mini Kit (Qiagen). mRNA was isolated from the total RNA using Oligotex mRNA Mini Kit (Qiagen). Genomic DNA contamination was eliminated using On-Column DNase Digestion (Qiagen) in combination with DNase I recombinant (Roche). cDNA synthesis was performed using cDNA Synthesis System (Roche). The cDNA clones were purified using MinElute PCR Purification (Qiagen). Subtraction Suppression Hybridization (SSH) was used to compare the cDNA libraries generated from mRNA populations obtained from the carbon and the nitrogen starved cultures. For this purpose, PCR-Select cDNA Subtraction Kit (Clontech) was used. DNA sequences of the cDNA clones from un-subtracted (prior to SSH) and subtracted carbon and nitrogen cultures were obtained through 454 pyrosequencing using the Roche 454 GS-FLX Titanium at Inqaba Biotec in South Africa.

In the second study, three *F. circinatum* strains GL 57, GL 100 and GL 101 were used for the construction of two cDNA libraries on complete medium. One library included cDNAs from strains GL 57 and GL 100 and the other library cDNAs from strain GL 101. The libraries were generated by growing the *F. circinatum* isolates in potato dextrose broth and extracting total RNA from the cultures using TRIzol (Invitrogen). Isolation and purification of mRNA from total RNA was carried out by adding total RNA to Dynabeads (Invitrogen) and Binding Buffer suspension, allowing mRNA to anneal to oligo $(dT)_{25}$ molecules on the beads, thereby creating an mRNA-bead complex. This was followed by mRNA elution (protocol modified by Marta Matvientko; University of California, Davis, Plant Science department). Fractionation of mRNA was performed using Ambion fractionation reagents. cDNA synthesis was done using Universal

RiboClone® cDNA Synthesis System (Promega C4360). The cDNA clones were purified using the MinElute PCR Purification Kit (Qiagen). The purified cDNA templates were enriched by PCR with two primers that anneal to the ends of cDNA adapters. The enriched cDNA clones were purified using the MinElute PCR Purification Kit (Qiagen). Library validation was done at the DNA Technology Facility at the University of California, Davis using the Agilent Bioanalyzer. Sequencing of cDNA clones was done using Illumina Sequencing platform GA II 26 Cycles.

Identification of genes showing alternative splicing

Two EST datasets were used in parallel to perform searches for alternatively spliced genes in the genome of *F. circinatum*. The first dataset included FASTA files with DNA sequence information for FSP34, while the second dataset included FASTA files with EST information for FSP34 in addition to those for GL 57, GL 100 and GL 101. The latter dataset was obtained from a CLC Genomics Workbench *de novo* assembly of all the sequence information generated for the various cDNA libraries constructed in this study. These datasets were used together with the 15 151 genes predicted in the *F. circinatum* genome (Wingfield *et al.*, 2012) to perform an RNA sequence Analysis which is a tool in CLC Genomics Workbench 4.8 (CLC bio A/S). This analysis produced an RNA sequence file containing gene-EST alignments, as well as expression values, exon lengths, and unique gene, exon, exon-exon, and intron-exon reads. Alignments that had at least two mRNA reads with intron-exon boundaries that were in conflict with those of the reference genes were used for further analyses. These alignments with conflicts were manually analyzed for possible alternative splicing signals by searching for places in the alignments where at least two ESTs gave conflicting alignment patterns to introns of the reference gene model.

To measure the chances of the identified ESTs being real alternative transcripts and not partially processed or sequenced transcripts, ESTs containing retained introns were categorized into two classes (Ho *et al*., 2007). The first class included ambiguous ESTs, which refers to partially sequenced ESTs, starting or ending with truncated introns. The second class was unambiguous ESTs, which contain full length introns.

Architecture of introns in genes showing alternative splicing

Two separate FASTA files of genes bearing putative alternative splicing signals were generated for each of the two EST datasets and imported into CLC Genomics Workbench 4.0.3 for the annotation of coding sequences (CDSs) and introns in the genes. For this purpose, the gene models from the annotated genome of *F. circinatum* were used. These annotations were confirmed by aligning the gene sequences with the CDSs of *F. verticillioides* (Broad Institute; http://www.broadinstitute.org/) using BioEdit version 7.0.9.0 (Hall, 1999).

All of the identified alternatively spliced introns were then analysed in terms of their phases and *cis*-elements by making use of procedures described before (Chapter 2 of this dissertation). For the analysis of *cis*-elements, alternatively spliced and constitutive introns for each gene were also compared. The CDSs with exons and alternatively spliced introns were also analysed for the presence of premature termination codons (TAA, TAG and TGA). This was done by counting multiples of three nucleotides from the start codon until an in-frame termination codon was encountered.

For the analysis of the relationship between CDS length and alternative intron length, a Scatter plots was generated (Microsoft Excel 2010). The significance of the regression line was tested using the Student's *t* test, where H_o and H_a were, respectively $\beta_1 = 0$ and $\beta_1 \neq 0$ (β_1 is the slope of the regression line) (Samuels and Witmer, 2003). The $t_s = b_1/SE_{b1}$ equation was used to obtain the *t* statistic (t_s is the *t* statistic, b_1 is the slope of the regression line and SE_{b1} is the standard error of the slope, which was calculated using the Excel Regression Tool found in Excel Analysis ToolPack and Excel Analysis ToolPack-VBA Add-ins) and *t* critical value was obtained from the Student's *t* distribution table at 90 and 95% confidence levels (0.10 and 0.05 *p* values, respectively) with $n - 2$ degrees of freedom (Samuels and Witmer, 2003).

Identification of homologues of genes with alternative splicing signals in the other Fusarium species

Homologues of the *F. circinatum* genes that showed alternative splicing were identified in *F. verticillioides*, *F. oxysporum* and *F. graminearum* using the Broad Institute's *Fusarium* Comparative database (http://www.broadinstitute.org/). Genome and EST data for *F. verticillioides*, *F. oxysporum* and *F. graminearum* were obtained from the *Fusarium* Comparative database at the Broad Institute (http://www.broadinstitute.org/). The cDNA libraries for *F. verticillioides*, *F. oxysporum* and *F. graminearum* were generated from carbon and nitrogen starved mycelia, mycelia grown on simple and complex medium and from cultures of maturing perithecia (Trail *et al*., 2003; http://www.broadinstitute.org/). For the identification of the homologues, nucleotide BLAST (BLASTn) searches were performed against those in the database. By making use of the corresponding EST data available in the database for these fungi the type of alternative splicing events and the number of ESTs showing alternative splicing for each gene were analyzed.

Identification of nonsense-mediated mRNA decay (NMD) genes

The genomes of *F. verticillioides*, *F. circinatum*, *F. oxysporum* and *F. graminearum* were searched for the presence of known NMD genes. For this purpose, the *Saccharomyces cerevisiae* NMD genes were first retrieved from the database of the Broad Institute and included *UPF1*, *UPF2*/*NMD2*, *UPF3*, *SUI1*, *PRT1* and *HRP1*/*NAB4*. These genes were subsequently used to perform BLASTn and BLASTx searches against the *Fusarium* Comparative database in order to obtain *Fusarium* homologues. Genes retrieved from the BLAST searches were used to perform a local BLASTn search against the *F. circinatum* genome (Wingfield *et al.*, 2012) using the CLC Genomics Workbench version 4.8 software.

Results

Identification of genes showing alternative splicing

By making use of the *F. circinatum* genome and the EST data generated for isolate FSP34 a total of 283 (out of 15 151) genes had at least one read showing one or more unique intron-exon junction when the ESTs were aligned to the reference gene models. Of these, 192 had at least two reads that showed unique intron-exon junctions relative to their reference genes. Among all these genes showing putative alternative intron splicing signals, 27 genes were represented by at least two ESTs with conflicting intron-exon junctions (Table 1).

When the dataset containing the assembled ESTs generated from all of the *F. circinatum* cDNA libraries included in this study was used, 200 genes were identified with at least one read having at least one unique intron-exon junction when compared to the gene model. Of the 200 genes, 133 had at least two reads that showed unique intron-exon junctions to their reference genes. Of the 133, 20 genes were found to have at least two ESTs with conflicting intron-exon junctions (Table 1).

Only three genes showing alternative splicing signals were common to the two datasets. The three genes included FCIRC_02105 (similar to FVEG_06773, which has an uncharacterized protein family UPF0057 domain), FCIRC_03285 (similar to FVEG_05139 40S Ribosomal Protein S6-B) and FCIRC_12064 (similar to gi|46107863|ref|XM_380991.1| *Gibberella zeae* PH-1 hypothetical protein partial mRNA Peptide Methionine Sulfoxide Reductase msrA 1). Therefore, analysis of the two EST datasets allowed identification of 44 genes (about 0.3% of all the genes in the genome) showing alternative splicing.

Within the set of 44 genes, 51 introns were alternative as they were retained in the ESTs. Among these alternative introns, 39 were supported by at least one EST and 12 by at least two ESTs (Table 1). In terms of the possibility that these data might be influenced by partial processing of mRNAs prior to preparation of the respective cDNA libraries or by partial sequencing, a large

proportion of the ESTs supporting the 51 alternative introns were classed as unambiguous, because they did not start or end with truncated introns (Table 1).

Analysis of genes showing alternative splicing

The 44 genes potentially harbouring alternatively spliced introns had CDS lengths ranging from 213 to 2540 nucleotides (nt), and had one to 16 introns per gene. The number of alternatively spliced introns per gene ranged from one to five, with the longest CDS (2540 nt; FCIRC_06581) harbouring five alternative introns. The longest alternative intron was 669 nt long and was found in the 5' untranslated region (UTR) of gene FCIRC_03285. This gene is similar to FVEG_05139, which encodes a 14-3-3 protein domain (Table 1). The shortest and longest alternative introns found within the CDSs were 45 and 191 nt long (Table 2) from genes FCIRC_04337 and FCIRC_05496, respectively. These are hypothetical proteins with uncharacterized protein domains. The shortest and longest constitutively spliced introns within the 44 genes were 44 and 274 nt long from FCIRC_14955, a hypothetical protein with uncharacterized protein domains, and FCIRC_00214 that is similar to FVEG_05230 encoding a Ribosomal L37ae protein family domain, respectively (data not shown).

There was no significant correlation between CDS length and alternative intron length (Figure 1). The *t* statistic was found to be 0.901, and *t* critical values of 1.68 and 2.01 were obtained at 90 and 95% confidence levels (0.10 and 0.05 *p* values, respectively) with *n* - 2 (50-2) degrees of freedom. The relationship between alternative intron length and position was also analyzed (Figure 2). Twelve out of 51 alternative introns were longer than 100 nt (109 to 669 nt). These were found in positions 1 and 2 of the 5' UTRs, and position 1, 2 and 3 of the CDSs. The remaining (39) introns shorter than 100 nt (45 to 87 nt) were found in position 1 of the 5' UTRs and positions 1 to 10 of the CDSs with no particular pattern.

Analysis of the intron phases showed that 45 of the alternative introns that were examined occur in different phases and that intron position is not correlated with intron phase. Six introns were excluded from the total of 51 alternative introns as they were in the 5' UTR. About 33% percent of the introns were in phase 0, 40% (18 out of 45) in phase 1 and 27% (12 out of 45) in phase 2. When the relationship between intron phase and intron position was analyzed, all three intron

phases were found in positions 1, 2, and 3 of the CDSs, which was not the case for those occurring in positions 4-10 (Figure 3 A). However, analysis of alternative intron phase distribution within genes revealed that phase 0 and 1 introns were more dominant than phase 2 introns in the 1st quarter of the genes and phase 2 introns were more dominant in the last quarter of the genes examined (Figure 3 B).

Investigation of the 45 alternative introns within CDSs showed the presence of premature termination codons in different positions of 33 introns. Ten alternative introns carried an in-frame premature TGA termination codon (Table 2 and Figure 4). Eight of these codons were located between the 5' splice site GT and the branch site where they were situated between 7 and 72 nt away from the 5' splice site (Figure 4). One of the termination codons formed part of the 5' splice site, one and formed part of the branch site. Nine alternative introns carried an in-frame premature TAG termination codon, of which five were located 13-162 nt away from the 5' splice site dinucleotide where they were situated between this site and the branch site, while two formed part of the 3' splice site and two formed part of the 5' splice site. Fourteen alternative introns carried in-frame premature TAA codons. Seven of these formed part of the 5' splice site, six formed part of the branch site, and one was located between the 5' splice site and the branch site.

The alternative introns in all three phases (phase 0, 1 and 2) also had all three premature termination codons (TGA, TAG and TAA) (Figure 4). Five of the 15 phase 0 introns had a TGA between the 5' splice site and the branch site. Five introns had a TAA, of which two were located at the 5' splice site, two at the branch site, and one was between the 5' splice site and the branch site. Only one intron had a premature termination codon which coincided with the 3' splice site. Seven of the 18 phase 1introns had premature termination codons, six between the 5' splice site and the branch site and one which coincided with the 3' splice site. Three introns had a TGA premature termination codon, two between the 5' splice site and the branch site and one which coincided with the 5' splice site. Two introns had a TAA premature termination codon, both coincided with the branch site. Seven of the 12 phase 2 introns had a TAA premature termination codon, five in the branch site and two between the 5' splice site and the branch site. Two phase 2 introns had a TGA premature termination codon, one coincided with the branch site and one was

between the 5' splice site and the branch site. Only one phase 2 intron generated a TAG premature termination codon between the 5' splice site and the branch site.

Twelve genes did not carry an in-frame premature termination codon and their CDSs utilized the same termination codons as their constitutive counterparts. These were Triosephosphate isomerase, Bacteriorhodopsin, a hypothetical protein with a Ctr Copper Transporter family domain, LICD Protein family, Amino Acid Permease, a hypothetical protein with Major Facilitator Superfamily and Multidrug Resistance protein domains, Aminotransferase class I and II, 60S ribosomal protein L10a, Tubulin alpha chain. Two of the 12 genes had no characterized domains and one did not match any of the genes in the *Fusarium* Broad Institute database.

Consensus sequences of the 5' and 3' splice site obtained from the multiple sequence alignments of constitutive and alternative introns showed that alternative introns have on average more degenerate splice sites. The following 5' splice sites were detected in the FSP34 EST dataset: $C_{34}T_{35}G_{59}|G_{100}T_{97}A_{61}A_{69}G_{91}T_{73}$ and $A_{29}N_{35}G_{44}|G_{100}T_{100}A_{74}A_{49}G_{85}T_{74}$ for respectively constitutive and alternative introns. The 5' splice sites that were detected from all of the *F. circinatum* EST data (*i.e.*, generated for isolates FSP34, GL 57, GL 100 and GL 101) were $A_{31}A_{45}G_{62}|G_{100}T_{100}A_{55}A_{55}G_{86}T_{72}$ for constitutive introns and $A_{65}A_{45}G_{60}|G_{100}T_{100}A_{75}A_{45}G_{80}T_{60}$ for alternative introns. The 3' splice sites detected in the FSP34 EST data had the consensus sequences $T_{51}A_{100}G_{100}|T_{28}$ for constitutive introns and $C_{50}A_{100}G_{100}|G_{41}$ for alternative introns, while all the *F. circinatum* ESTs together allowed identification of 3' splice sites with consensus sequences $C_{52}A_{100}G_{100}|G_{34}$ for constitutive introns and $C_{50}A_{100}G_{100}|T_{40}$ for alternative introns.

The putative branch site sequences of the alternative introns were also more degenerate than those of constitutive introns (Table 3). Four branch site sequences (CTAAC, CTGAC, CTAAT and CTGAT) of the constitutive introns conformed to the canonical CURAY motif of the fungal branch site, while seven branch site sequences (TTAAC, TTGAC, TTGAT, CTAAA, CTTAC, CTCAC and CCAAC) were secondary (Chapter 2 of this dissertation). Three branch site sequences (CTAAC, CTGAC and CTAAT) of the alternative introns conformed to the canonical

CURAY motif, while seven sequences (CCAAT, TGAAT, TTAAC, TTGAC, TTGAT, CTAAG and CTCAC) were secondary.

Identification of homologues of genes with alternative splicing signals in the other Fusarium species

A total of 22 of the 44 genes that showed alternative splicing in *F. circinatum* had homologues in other *Fusarium* species. These genes also showed alternative splicing when their EST data was analyzed. In *F. verticillioides* 10 genes showing alternative splicing were identified, while 12 and 11 respectively were identified in *F. oxysporum* and *F. graminearum* (Table 1). Three alternative splicing mechanisms (intron retention, exon skipping, and 5' and 3' alternative splice site usage) were observed in these species, whereas only intron retention was observed for *F. circinatum*. The 60S Ribosomal Protein L10a (FCIRC_11342) had an alternative 5' splice site in its intron in both *F. oxysporum* and *F. graminearum*. The 1st intron of the gene encoding Alcohol dehydrogenase 1 (FCIRC_13422) and also the 1st intron of the gene encoding hypothetical protein FCIRC_04337 had alternative 3' splices site in, respectively, *F. oxysporum* and *F. graminearum*. The 1st intron of the gene encoding hypothetical protein FCIRC_08575 had both an alternative 5' splice site and an alternative 3' splice site in *F. graminearum*. The remaining 12 homologues had the same alternative introns as the *F. circinatum* genes, one of which was in the 5' UTR.

Identification of NMD genes

Five of the six known NMD genes (*UPF1*, *UPF2*/*NMD2*, *UPF3*, *SUI1*, *PRT1* and *HRP1*/*NAB4*) had homologues in the *Fusarium* genomes. These included *UPF1* (ATP-dependent Helicase NAM7), *NMD2* (nonsense-mediated mRNA decay protein 2), *SUI1* protein translation factor, *PRT1* (translation initiation factor eIF3 subunit) and HRP1 (nuclear polyadenylated RNA-binding protein 4) which were found in all three *Fusarium* species (*F. graminearum*, *F. oxysporum* and *F. graminearum*) in the Broad Institute database. All five genes were also found in the *F. circinatum* genome: FCIRG_03838 similar to *UPF1*, FCIRG_05623 similar to *NMD2*, FCIRG_10988 similar to *SUI1*, FCIRG_13253 similar to *PRT1* and FCIRG_09778 similar to *HRP1*.

Discussion

By making use of EST data for *F. circinatum* we were able to detect 44 genes (0.3% of all the genes encoded by this species) within the genome of this fungus that most likely experience alternative intron splicing. This number is low relative to the 277 genes (4.2%) found in *C. neoformans*, 224 genes (3.6%) in *U. maydis* and 134 genes (1.6%) in *M. grisea* (Loftus *et al*., 2005; Ho *et al*., 2007; Ebbole *et al*., 2004). One possible explanation for this is that the EST data available for this study was generated from mRNA harvested from mycelia grown under three conditions (*i.e*., carbon stress, nitrogen stress and complete medium only). Also, in some gene-EST sequence alignments, our EST data did not cover the entire length of the genes, which undoubtedly affected our ability to detect alternative intron splicing signals. Therefore, obtaining more EST data from spores and from other growth conditions, such as *in planta*, would help increase the probability of detecting more alternative splice variants (Graveley, 2001) in *F. circinatum*. Nevertheless, our data suggests that *Fusarium* species, and *F. circinatum* in particular are no different from other fungi in that alternative intron splicing is also included in its suite of mechanisms for regulating gene expression.

The only type of alternative splicing detected in *F. circinatum* was intron retention. However in both *F. verticilioides* and *F. graminearum* intron retention and exon skipping were observed, and alternative 5' and 3' splice site usage were detected in *F. oxysporum* and *F. graminearum*. Our data further showed that alternative splicing occurs at the same introns for some genes in the *Fusarium* species examined, although the mechanisms involved in the splicing at the same intron positions were not always similar. The fact that the large majority of the alternatively spliced genes examined in the four *Fusarium* species apparently utilize intron retention as a splicing mechanism was not surprising. This is because intron definition is the preferred mechanism of constitutive splicing in fungi and other lower eukaryotes, thereby usually leading to the dominance of intron retention in these organisms (McGuire *et al*., 2008; Ho *et al*., 2007). However, an unequivocal conclusion that intron retention is the only alternative splicing mechanism occurring in *F. circinatum* cannot be reached, as the available EST data were limited.

The genes that showed alternative splicing signals are involved in diverse cellular processes (Table 1). Proton transport, copper transport, amino acid metabolism, glycolysis, cell structure maintenance, ribosome biogenesis, rRNA processing, oxidative phosphorylation, and pathogenicity were among the processes that were identified. Twenty-seven genes showed alternative splicing when EST data from the carbon and nitrogen starvation study were used and 20 when the assembled reads from the carbon and nitrogen starvation study as well as from the complete medium study were used. Genes found to undergo alternative splicing in *C. neoformans* and *A. flavus* are also involved in diverse cellular processes similar those that have been mentioned above (Loftus *et al*., 2005; Chang *et al*., 2010).

Some of the genes that showed alternative splicing in *F. circinatum* and other *Fusarium* species have also been reported to undergo alternative splicing in other eukaryotes (Szabo *et al*., 1994; Chen and Thelen, 2010; Yamada *et al*., 2003; Iida *et al*., 2009; Kim and Gladyshev, 2006). Among those genes were Glutamate decarboxylase, Triose Phosphate isomerase, Amino Acid permease and Peptide Methionine Sulfoxide reductase. The different types of alternative splicing observed thus far for each of these genes are discussed below.

The gene encoding Glutamate decarboxylase (GAD), which catalyzes the decarboxylation of glutamate to GABA (gamma-aminobutyric acid) and CO during the Citric Acid cycle, has an alternative transcript with intron 2 retained in *F. circinatum*. The variant alternative transcripts of GAD have intron 3 retained in *F. verticillioides* and *F. oxysporum*. The mammalian Glutamate decarboxylase exists in two isoforms encoded by two different genes (*gad1* and *gad2*) (Szabo *et al*., 1994). GAD1 and GAD2 are both expressed in the brain and GAD2 in the pancreas (Szabo *et al*., 1994). However, in an embryonic brain two more isoforms are expressed (GAD25 and GAD44), which are encoded by alternative transcripts of GAD1 (I-80 and I-86) (Szabo *et al*., 1994). In *N. crassa*, GAD has been found in mature conidia with declining levels during germination of the conidia (Kumar *et al*., 1997). There have been indications that the mRNA synthesis of this gene is differentially regulated and this has been linked to the polypeptide being produced in mycelia but packed in conidia (Kumar *et al*., 1997). In-depth analysis of the alternative transcripts in *Fusarium* could potentially shed light on other functions of this gene during growth.

The second intron of the *F. circinatum* alternative transcript, which encodes Triose Phosphate isomerase (TPI) which catalyzes the conversion of acetone phosphate to glyceraldehyde-3-phosphate during glycolysis, is retained and the transcript is expressed under nitrogen stress conditions. *Fusarium verticillioides* has two additional alternative transcripts: one transcript in which intron 1 is retained and one in which exon 3 is skipped. *Arabidopsis thaliana* TPI has a plastid alternative transcript required for the postgerminative switch from heterotrophic to autotrophic growth (Chen and Thelen, 2010). *S. cerevisiae* and *Schizosaccharomyces pombe* have different transcription initiation sites for this gene (Russell, 1985). TPI is one of the proteins found to be expressed in *F. graminearum* during infection of barley with low levels of nitrogen (Yang *et al*., 2009). Again, further analysis of the alternative transcripts in *Fusarium* could shed light on other functions of this gene during growth and infection.

The *F. circinatum* alternative transcript encoding Amino Acid permease (AAP) has intron 4 retained and is expressed under nitrogen stress. In *F. verticillioides*, intron 7 of this gene is retained, while *F. oxysporum* and *F. graminearum* had no alternative transcripts. *A. thaliana* also has two isoforms of the protein due to alternative splicing (Yamada *et al*., 2003; Iida *et al*., 2009). Isoform 1 belongs to the Amino Acid/Polyamine transporter 2 family and the Amino Acid/Auxin permease (AAAP) subfamily (Yamada *et al*., 2003). Isoform 2 is a probable Amino Acid permease 7 (AAP7) and a stereospecific transporter with a broad specificity for neutral amino acids (Iida *et al*., 2009). AAP, as a nitrogen responsive gene, is expressed in *F. oxysporum* infection on tomato and it allows the uptake rare amino acids such as the plant defense-related Orninthine (Divon *et al*., 2005). In fungi in general, this protein also takes up plant defense-related metabolites such as GABA (Divon and Fluhr, 2006). Perhaps the alternative transcripts of this gene encode permeases that are specific to the different plant defense-related metabolites and could be further investigated in *Fusarium*.

The alternative transcript of Peptide Methionine Sulfoxide reductase *msrA* 1 in *F. circinatum* has intron 1 retained. The protein encoded by this gene catalyzes the reduction of oxidized Methionines in a Thiol dependent manner (Le *et al*., 2005). No evidence of alternative splicing was observed in the other *Fusarium* species, this transcript was supported by EST data from the carbon starvation and the combined libraries of *F. circinatum* used in this study. Alternative splicing of *msrA* and *msrB* genes has been reported in insects such as *Drosophila melanogaster, D. yakuba* and *D. pseudoobscura* (Kim and Gladyshev, 2006). It has also been suggested that alternative splicing regulates subcellular distribution of Methionine Sulfoxide reductases in mammals and other animals (Kim and Gladyshev, 2006). Further analysis of the different transcripts in *Fusarium* could possibly inform us on the different roles and specificities of this gene in fungal metabolism.

Some of the *F. circinatum* genes harbouring possibly alternative splicing signals had homologues in other organisms encoding protein isoforms produced by different genes or by post-translational modifications, and not through alternative splicing. Intron 2 of the alternative transcript of *F. circinatum* that encodes Cytochrome c Oxidase subunit V, which is involved in generating a proton gradient within the inner mitochondrial membrane for ATP synthesis and H_2O production (Burke and Poyton, 1998), is retained and *F. graminearum* has an additional transcript in which exon 3 is skipped. In yeast and mammals this protein has two differentially regulated isoforms that are encoded by two genes (Burke and Poyton, 1998; Waterland *et al*., 1990; Fukuda *et al*., 2007). Intron 3 of the alternative transcript encoding Neutral trehalase is retained in *F. circinatum* and *F. verticillioides*. This protein has a probable role in carbohydrate metabolism and its isoforms are encoded by two genes in *F. oxysporum* (Wolska-Mitaszko *et al*., 2007).

During this study the relationships between CDS length and alternative intron length and intron length and intron position were assessed. Linear regression analysis showed that there was no significant correlation between CDS length and alternative intron length. However, after analyzing intron length in relation to intron position, we found that introns longer than 100 nt were mainly found in the 5' UTRs and in position 1 and 2 of the CDSs. This finding is consistent with studies that were done on all introns of genes (or genomes) showing that introns in the 5'

UTR and in position 1 of eukaryotic CDSs were longer than those in the rest of the gene or CDS (Bradnam and Korf, 2008; Hawkin, 1988, Hong *et al*., 2006; Smith, 1988; Kriventseva and Gelfand, 1999). These longer introns have been suggested to play a role in enhancing gene expression as motifs have been found on them in some eukaryotes (Gaffney and Keightley, 2006; Parra *et al*., 2011).

The relationship between intron phase and intron position were determined for the alternative introns. All three intron phases (phase 0, 1 and 2) were found in positions 1, 2 and 3. The other intron positions (3-10) were represented by low intron numbers and therefore did not have all intron phases represented. It seems from these results that the phase of an alternative intron is not dependent on intron position. However, more phase 0 and phase 1 introns were closer to the 5' end of genes than phase 2 introns and more phase 2 introns were closer to the 3' end than phase 1 and 2 introns. This distribution of intron phases is similar to what is usually observed for introns in general (Ruvinsky and Ward, 2006; Chapter 2 of this dissertation). Perhaps the low number of phase 2 introns is due to the fact that they are mostly located at the 3' end of genes, making them more prone to loss *via* homologous recombination of a spliced mRNA and a genomic copy of that gene (Qiu *et al*., 2004; Lin and Zhang, 2005).

Irrespective of whether the alternative introns were between two codons (phase 0) or disrupting a codon (phase 1 and 2), they all generated all three types of in-frame premature termination codons (TGA, TAG and TAA). It is known that such termination codons sometimes lead to the production of antagonistic, non-functional or functional proteins when translated (Graveley, 2001; Chang *et al*., 2010). The transcripts encoding these proteins are translated as normal transcripts until the premature stop codon is encountered after which translation stops (Graveley, 2001). An untranslated alternative transcript bearing a premature termination codon can also play an indirect role in post-transcriptional gene regulation (Smith and Valcarcel, 2000).

Transcripts with premature stop codons that are due to alternative splicing are targets of NMD (Wang and Brendel, 2006). These potentially deleterious alternative transcripts are in most cases degraded before translation in order to prevent cell toxicity (Mekouar *et al*., 2010; Graveley, 2001). The high number of transcripts with premature stop codons observed in this study

indicated that an NMD system is required in *F. circinatum*. When NMD pathway genes were searched in the genomes of *F.circinatum* and its three relatives (*F. verticillioides*, *F. oxysporum* and *F. graminearum*), five (*UPF1*, *NMD2*, *SUI1*, *PRT1* and *HRP1*) of the six *S. cerevisiae* genes were found. *UPF3* was not found in any of the *Fusarium* species which could either be an indication of the in-completeness of these genomes as *UPF1*, *UPF2* and *UPF3* have been reported to be essential for NMD in organisms ranging from yeast, human to plants (Aronoff *et al*., 2001) or that NMD mechanisms in *Fusarium* species can function optimally without *UPF3*. Nonetheless, functional studies are required in order to determine whether these genes are functional as NMD genes in *Fusarium* or merely ancestral copies. In the nematode *Caenorhabditis elegans* a protein (*smg-1*) that degrades double-stranded RNA in a post-transcriptional gene regulation process (*i.e.* RNA interference) has also been implicated in the degradation of alternative transcripts bearing premature termination codons through NMD (Domeier *et al*., 2000). Whether this is the case in fungi still needs investigation. In addition, the high number of potential transcripts with premature stop codons observed in the currently study also warrants research of the efficacy of this quality control system in *Fusarium*.

The nature of the three main intron *cis*-elements (the 5' and 3' splice sites, and the branch site) of *F. circinatum* causes the elements to bear termination codons following alternative splicing, which lead to truncation of proteins during translation when in frame. The 5' splice site of alternatively spliced introns harboured all three classes of premature termination codons (for example NNN*GTAA*NN, NNN*GTAG*NN, and NNNGTNNN*TGA*; the underlined nucleotides represent the *cis*-element, N represents any base, and the nucleotides in italics are the premature termination codons). The branch sites of alternatively spliced introns had TAA and TGA premature termination codons (for example Y*TAA*B and Y*TGA*B; Y represents a pyrimidine and B represents a pyrimidine and a Guanosine), while their 3' splice sites contained TAG premature termination codon (*TAG*N). Fourteen of the identified alternative introns had in-frame premature termination codons between the 5' splice site and the branch and none bore in-frame-premature termination codons between the branch site and the 3' splice site. In yeast genomes, the sequence of the 5' splice site motif (GTATGT) does not allow for premature termination codon generation following alternative splicing (Mekouar *et al*., 2010). Currently, only *Y. lipolytica* is known to

have a 5' splice site (GTGAGT) that allows for premature termination codon generation following alternative splicing (Mekouar *et al*., 2010).

Not all of the genes in *F. circinatum* that potentially undergo alternative splicing harbour premature termination codons. In about 27% in these genes no in-frame termination codons were detected. Therefore, upon translation, the transcripts of these genes can either yield functional or non-functional proteins that may or may not be important in the regulation of cellular processes (Smith and Valcarcel, 2000). The significance of many alternative splices has not yet been discovered (Graveley, 2001). In order to know the significance of these transcripts, functional (expression and knockout) studies should be performed in future.

The question remains: What makes an intron prone to alternative splicing? Our comparison of the 5' and 3' splice sites of constitutive and alternative introns revealed that the consensus sequences motif of the alternative introns was more degenerate than that of constitutive introns. This was even more pronounced for the 3' splice site for which we detected a range of different signatures. Our results further indicate that alternative introns have a higher percentage of alternative branch site sequences (Table 3). Although the consensus sequences reported here could have been influenced by the difference in sample sizes, alternative introns definitely appear to tolerate more diversity within their three main *cis*-elements. This supports the idea that the branch site, together with the 5' and 3' splice site, influences alternative splicing (Ast, 2004).

References

1. Aronoff R, Baran R, Hodgkin J: Molecular identification of smg-4, required for mRNA surveillance in *C. elegans*. Gene 2001, 268(1-2):153-164.
2. Ast G: How did alternative splicing evolve? Nature Reviews Genetics 2004, 5(10):773-782.
3. Barrass JD, Beggs JD: Splicing goes global. Trends in Genetics 2003, 19(6):295-298.
4. Berget SM: Exon recognition in vertebrate splicing. Journal of Biological Chemistry 1995, 270(6):2411.
5. Bhattacharya D, Lutzoni F, Reeb V, Simon D, Nason J, Fernandez F: Widespread occurrence of spliceosomal introns in the rDNA genes of ascomycetes. Molecular Biology and Evolution 2000, 17(12):1971-1984.
6. Bradnam KR, Korf I: Longer first introns are a general property of eukaryotic gene structure. PLoS One 2008, 3(8):e3093.
7. Burke JM: Molecular genetics of group I introns: RNA structures and protein factors required for splicing--a review. Gene 1988, 73(2):273-294.
8. Burke PV, Poyton RO: Structure/function of oxygen-regulated isoforms in cytochrome c oxidase. Journal of Experimental Biology 1998, 201(8):1163.
9. Cech TR: Self-splicing of group I introns. Annual Review of Biochemistry 1990, 59(1):543-568.
10. Chang KY, Georgianna DR, Heber S, Payne GA, Muddiman DC: Detection of alternative splice variants at the proteome level in *Aspergillus flavus*. Journal of Proteome Research 2010, 9(3):1209-1217.
11. Chen M, Thelen JJ: The plastid isoform of triose phosphate isomerase is required for the postgerminative transition from heterotrophic to autotrophic growth in Arabidopsis. The Plant Cell Online 2010, 22(1):77.
12. Conti E, Izaurralde E: Nonsense-mediated mRNA decay: molecular insights and mechanistic variations across species. Current Opinion in Cell Biology 2005, 17(3):316-325.
13. Divon HH, Fluhr R: Nutrition acquisition strategies during fungal infection of plants. FEMS Microbiology Letters 2007, 266(1):65-74.

14. Dinon HH, Rothan-Denoyes B, Davydov O, Di Pietro A, Fluhr R: Nitrogen-responsive genes are differentially regulated in planta during *Fusarium oxyspsorum* f. sp. *lycopersici* infection. Molecular Plant Pathology 2005, 6(4):459-470.
15. Domeier ME, Morse DP, Knight SW, Portereiko M, Bass BL, Mango SE: A link between RNA interference and nonsense-mediated decay in *Caenorhabditis elegans*. Science 2000, 289(5486):1928-1930.
16. Ebbole DJ, Jin Y, Thon M, Pan H, Bhattarai E, Thomas T, Dean R: Gene discovery and gene expression in the rice blast fungus, *Magnaporthe grisea*: analysis of expressed sequence tags. Molecular Plant-microbe Interactions 2004, 17(12):1337-1347.
17. Fukuda R, Zhang H, Kim J, Shimoda L, Dang CV, Semenza GL: HIF-1 regulates cytochrome oxidase subunits to optimize efficiency of respiration in hypoxic cells. Cell 2007, 129(1):111-122.
18. Gaffney DJ, Keightley PD: Genomic selective constraints in murid noncoding DNA. PLoS Genetics 2006, 2(11):e204.
19. Graveley BR: Alternative splicing: increasing diversity in the proteomic world. Trends in Genetics 2001, 17(2):100-107.
20. Hall TA: BioEdit: a user-friendly biological sequence alignment editor and analysis program for Windows 95/98/NT. In: 1999; 1999: 95-98.
21. Hawkin JD: A survey on intron and exon lengths. Nucleic Acids Research 1988, 16(21):9893-9908.
22. Ho E, Cahill M, Saville B: Gene discovery and transcript analyses in the corn smut pathogen *Ustilago maydis*: expressed sequence tag and genome sequence comparison. BMC Genomics 2007, 8(1):334-355.
23. Hong X, Scofield DG, Lynch M: Intron size, abundance, and distribution within untranslated regions of genes. Molecular Biology and Evolution 2006, 23(12):2392.
24. Iida K, Fukami-Kobayashi K, Toyoda A, Sakaki Y, Kobayashi M, Seki M, Shinozaki K: Analysis of multiple occurrences of alternative splicing events in *Arabidopsis thaliana* using novel sequenced full-length cDNAs. DNA Research 2009, 16(3):155-164.
25. Keren H, Lev-Maor G, Ast G: Alternative splicing and evolution: diversification, exon definition and function. Nat Rev Genet 2010, 11(5):345-355.

26. Kim HY, Gladyshev V: Alternative first exon splicing regulates subcellular distribution of methionine sulfoxide reductases. BMC Molecular Biology 2006, 7(1):11.
27. Kriventseva E, Gelfand M: Statistical analysis of the exon-intron structure of higher and lower eukaryote genes. Journal of Biomolecular Structure & Dynamics 1999, 17(2):281.
28. Kumar S, Punekar NS: The metabolism of 4-aminobutyrate (GABA) in fungi. Mycological Research 1997, 101(4):403-409.
29. Lambowitz AM, Perlman PS: Involvement of aminoacyl-tRNA synthetases and other proteins in group I and group II intron splicing. Trends in Biochemical Sciences 1990, 15(11):440-444.
30. Le DT, Lee BC, Marino SM, Zhang Y, Fomenko DE, Kaya A, Hacioglu E, Kwak GH, Koc A, Kim HY: Functional analysis of free methionine-R-sulfoxide reductase from *Saccharomyces cerevisiae*. Journal of Biological Chemistry 2009, 284(7):4354-4364.
31. Le Hir H, Nott A, Moore MJ: How introns influence and enhance eukaryotic gene expression. Trends in Biochemical Sciences 2003, 28(4):215-220.
32. Lin K, Zhang DY: The excess of 50 introns in eukaryotic genomes. Nucleic Acids Res 2005, 33:6522–6527.
33. Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, Vamathevan J, Miranda M, Anderson IJ, Fraser JA: The genome of the basidiomycetous yeast and human pathogen *Cryptococcus neoformans*. Science 2005, 307(5713):1321.
34. Long M, De Souza SJ, Rosenberg C, Gilbert W: Relationship between "proto-splice sites" and intron phases: evidence from dicodon analysis. Proceedings of the National Academy of Sciences 1998, 95(1):219-223.
35. Ma LJ, Van Der Does HC, Borkovich KA, Coleman JJ, Daboussi MJ, Di Pietro A, Dufresne M, Freitag M, Grabherr M, Henrissat B: Comparative genomics reveals mobile pathogenicity chromosomes in *Fusarium*. Nature 2010, 464(7287):367-373.
36. Mardis ER: Next-generation DNA sequencing methods. Annu Rev Genomics Hum Genet 2008, 9:387-402.
37. Mascarenhas D, Mettler IJ, Pierce DA, Lowe HW: Intron-mediated enhancement of heterologous gene expression in maize. Plant Molecular Biology 1990, 15(6):913-920.
38. McGuire AM, Pearson MD, Neafsey DE, Galagan JE: Cross-kingdom patterns of alternative splicing and splice recognition. Genome Biology 2008, 9(3):R50.

39. Mekouar M, Blanc-Lenfle I, Ozanne C, Da Silva C, Cruaud C, Wincker P, Gaillardin C, Neuvéglise C: Detection and analysis of alternative splicing in *Yarrowia lipolytica* reveal structural constraints facilitating nonsense-mediated decay of intron-retaining transcripts. Genome Biology 2010, 11(6):R65.
40. Metzker ML: Sequencing technologies-the next generation. Nature Reviews Genetics 2009, 11(1):31-46.
41. Modrek B, Lee CJ: Alternative splicing in the human, mouse and rat genomes is associated with an increased frequency of exon creation and/or loss. Nature Genetics 2003, 34(2):177-180.
42. Pan Q, Shai O, Lee LJ, Frey BJ, Blencowe BJ: Deep surveying of alternative splicing complexity in the human transcriptome by high-throughput sequencing. Nature Genetics 2008, 40(12):1413-1415.
43. Parra G, Bradnam K, Rose AB, Korf I: Comparative and functional analysis of intron-mediated enhancement signals reveals conserved features among plants. Nucleic Acids Research 2011, 39(13):5328-5337.
44. Piccirilli JA, McConnell TS, Zaug AJ, Noller HF, Cech TR: Aminoacyl esterase activity of the Tetrahymena ribozyme. Science 1992, 256(5062):1420.
45. Qiu WG, Schisler N, Stoltzfus A: The evolutionary gain of spliceosomal introns: sequence and phase preferences. Molecular Biology and Evolution 2004, 21(7):1252.
46. Russell PR: Transcription of the triose-phosphate-isomerase gene of *Schizosaccharomyces pombe* initiates from a start point different from that in *Saccharomyces cerevisiae*. Gene 1985, 40(1):125-130.
47. Ruvinsky A, Ward W: A gradient in the distribution of introns in eukaryotic genes. Journal of Molecular Evolution 2006, 63(1):136-141.
48. Saldanha R, Mohr G, Belfort M, Lambowitz AM: Group I and group II introns. The FASEB journal 1993, 7(1):15-24.
49. Samuels ML: Statistics for life sciences, Third edn; 2003.
50. Smith CWJ, Valcárcel J: Alternative pre-mRNA splicing: the logic of combinatorial control. Trends in Biochemical Sciences 2000, 25(8):381-388.
51. Smith M: Structure of vertebrate genes: a statistical analysis implicating selection. Journal of Molecular Evolution 1988, 27(1):45-55.

52. Szabo G, Katarova Z, Greenspan R: Distinct protein forms are produced from alternatively spliced bicistronic glutamic acid decarboxylase mRNAs during development. Molecular and Cellular Biology 1994, 14(11):7535-7545.
53. Trail F, Xu JR, Miguel PS, Halgren RG, Corby Kistler H: Analysis of expressed sequence tags from *Gibberella zeae* (anamorph *Fusarium graminearum*). Fungal Genetics and Biology 2003, 38(2):187-197.
54. Wang BB, Brendel V: Genomewide comparative analysis of alternative splicing in plants. Proceedings of the National Academy of Sciences 2006, 103(18):7175-7180.
55. Wang ET, Sandberg R, Luo S, Khrebtukova I, Zhang L, Mayr C, Kingsmore SF, Schroth GP, Burge CB: Alternative isoform regulation in human tissue transcriptomes. Nature 2008, 456(7221):470-476.
56. Waterland R, Basu A, Chance B, Poyton R: The isoforms of yeast cytochrome c oxidase subunit V alter the in vivo kinetic properties of the holoenzyme. Journal of Biological Chemistry 1991, 266(7):4180.
57. Wingfield BD, Steenkamp ET, Santana QC, Coetzee M, Bam S, Barnes I, Beukes CW, Chan WY, de Vos L, Fourie G: First fungal genome sequence from Africa: A preliminary analysis. South African Journal of Science 2012, 108(1/2):9 pages.
58. Wolska-Mitaszko B, Jaroszuk-Scisel J, Pszeniczna K: Isoforms of trehalase and invertase of *Fusarium oxysporum*. Mycological Research 2007, 111(4):456-465.
59. Yamada K, Lim J, Dale JM, Chen H, Shinn P, Palm CJ, Southwick AM, Wu HC, Kim C, Nguyen M: Empirical analysis of transcriptional activity in the *Arabidopsis* genome. Science 2003, 302(5646):842.
60. Yang F, Jensen JD, Spliid NH, Svensson B, Jacobsen S, Jørgensen LN, Jørgensen HJL, Collinge DB, Finnie C: Investigation of the effect of nitrogen on severity of Fusarium Head Blight in barley. Journal of Proteomics 2010, 73(4):743-752.

Tables

Table 1. List of genes identified as having putative alternative splicing signals.

Gene Identification number and name [a]	Supporting ESTs [b]	Processing completeness [c]	Alternative intron position(s) [d]	Homologues in other *Fusarium* species [e]		
				F. verticillioides	*F. oxysporum*	*F. graminearum*
FSP34 ESTS						
FCIRC_00214_Ribosomal L37ae protein family Domain	1	Ambiguous	1			1 (ES)
FCIRC_01601	1	Unambiguous	1			4 (IR)
FCIRC_01845_glutamate decarboxylase 1	1	Unambiguous	2	2 (IR)	2 (IR)	
FCIRC_02843	1	Unambiguous	2	4 (IR)		
	1	Unambiguous	3			
FCIRC_02980_triosephosphate isomerase	1	Ambiguous	2	1; 1 (IR; ES)		
FCIRC_03286_Mitochondrial ATPase inhibitor, IATP Domain	2	Ambiguous	2			
FCIRC_04115_Amino acid permease	1	Ambiguous	4	5 (IR)		
	1	Ambiguous	5			
FCIRC_05491_tubulin alpha chain	1	Unambiguous	6	5 (IR)	1 (IR)	
FCIRC_05496_40S ribosomal protein S6-B	3	Ambiguous	3	3 (IR)	3 (IR)	
FCIRC_06155_Major Facilitator Superfamily and Multidrug resistance protein Domains	1	Ambiguous	2			
FCIRC_06257_Cytochrome c oxidase subunit Via	2	Ambiguous	2			2; 1 (IR; ES)
FCIRC_06323_neutral trehalase	1	Unambiguous	3	2 (IR)		
FCIRC_06526_FHA domain, Protein kinase domain, Protein tyrosine kinase	1	Ambiguous	4			
FCIRC_06581	1	Ambiguous	5			
	1	Unambiguous	6			
	1	Unambiguous	7			
	2	Unambiguous	9			

	2	Unambiguous	10			
FCIRC_06732_ATP synthase alpha/beta family, nucleotide-binding domain	1	Unambiguous	4	4 (IR)	13 (ES)	
FCIRC_07301_probable methyltransferase domain, EasF family	1	Ambiguous	2			
FCIRC_09592_Aminotransferase class I and II	2	Unambiguous	1			
FCIRC_09687_Mitochondrial carrier protein and EF hand Domains	1	Unambiguous	1			
FCIRC_10565_Prenylcysteine lyase	2	Unambiguous	1		2 (5A)	1 (5A)
FCIRC_10968	2	Ambiguous	1			
FCIRC_11342_60S ribosomal protein L10a	2	Unambiguous	2			
FCIRC_12064_peptide methionine sulfoxide reductase msrA 1	1	Unambiguous	1			
FCIRC_12163_Bacteriorhodopsin	1	Ambiguous	1			
FCIRC_12655_40S ribosomal protein S5-B	1	Unambiguous	2		2 (3A)	2 (3A)
FCIRC_13218 (No hit)	1	Ambiguous	1			
FCIRC_13422_alcohol dehydrogenase 1	2	Ambiguous	1		1 (3A)	
	1	Ambiguous	3			
FCIRC_13979	2	Ambiguous	3	2 (IR)	3 (IR)	

ASSEMBLED ESTs

FCIRC_01390 Ctr copper transporter family Domain	1	Ambiguous	2	4 (IR)	2 (IR)	
FCIRC_02105_Uncharacterized protein family UPF0057 Domain	1	Ambiguous	1			
FCIRC_02340_vacuolar ATP synthase subunit D	1	Ambiguous	1 (at 5' UTR)			1 (ES)
FCIRC_03285_14-3-3 protein Domain	2	Ambiguous	1 (at 5' UTR)		6 (IR)	
FCIRC_04337	1	Unambiguous	1			1; 2 (IR; 3A)
FCIRC_04669_Basic region leucine zipper and bZIP transcription factor Domains	1	Ambiguous	1			
FCIRC_04971_Profilin	1	Ambiguous	1			4 (ES)

FCIRC_05768_ARP2/3 complex 20 kDa subunit Domain	1	Ambiguous	1			
FCIRC_06028_LICD Protein Family	1	Ambiguous	3			
FCIRC_07299	1	Unambiguous	1 (at 5' UTR)			
FCIRC_08575	1	Ambiguous	1		1; 1 (3A; 5A)	
FCIRC_09654	1	Ambiguous	1			
FCIRC_10813_Thi4 family and thiazole biosynthesis enzyme Domains	1	Ambiguous	2 (at 5' UTR)	1 (IR)	1 (IR)	1 (IR)
FCIRC_11232	1	Ambiguous	1 (at 5' UTR)			
FCIRC_11597_Ubiquitin family Domain	1	Ambiguous	1 (at 5' UTR)			
FCIRC_14835_mitochondrial cytochrome c oxidase assembly factor	1	Ambiguous	1			
FCIRC_14955	1	Unambiguous	1			
FCIRC_15141 (No hit)	1	Unambiguous	1			

[a] Gene names and numbers refer to those in the Broad Institute and *F. circinatum* genome, respectively.
[b] The number of *F. circinatum* ESTs supporting the alternatively spliced introns in the various genes.
[c] As measure of the completeness of mRNA processing, ESTs are classed as either ambiguous (partially sequenced ESTs, starting or ending with truncated introns) or unambiguous ESTs (containing full-length introns) (Ho *et al*., 2007).
[d] Alternative intron positions (intron located at the 5' UTR are indicated in parentheses next to the intron position number).
[e] The number of ESTs that support a specific alternatively spliced intron is followed in parentheses by the type of alternatively splicing that occurs (intron retention = IR, exon skipping = ES, 5' and 3' alternative splice site usage = 5A and 3A).

Table 2. Architecture of the introns in the genes potentially showing alternative splicing.

Gene Identification number and name [a]	Alternative intron position(s) [d]	Intron length [e]	Intron phase	Termination codon [f]	Branch site [g]
FSP34 ESTS					
FCIRC_00214_Ribosomal L37ae protein family Domain	1	60	0	TGA nt 67-69 between 5' ss and BS	CTAAC
FCIRC_01601	1	48	1	TAG nt 15-17 between 5' ss and BS	CTAAA
FCIRC_01845_Glutamate decarboxylase 1	2	50	2	TAA nt 2-4 of 5' ss	CTAAC
FCIRC_02843	2	51	2	TAA nt 34-36 (nt 1-3 of BS)	CTAAC
	3	55	0	TGA nt 28-30 between 5' ss and BS	CTGAG*
FCIRC_02980_Triosephosphate isomerase	2	121	0	None	CTAAC
FCIRC_03286_Mitochondrial ATPase inhibitor, IATP Domain	2	63	0	TAG nt 60-63 (nt 1-3 of 3' ss)	CTGAC
FCIRC_04115_Amino acid permease	4	57	1	TAA nt 2-4 of 5' ss	CTAAC
	5	62	1	None	CTAAC
FCIRC_05491_Tubulin alpha chain	6	58	2	None	CTAAT
FCIRC_05496_40S ribosomal protein S6-B	3	191	1	TAG nt 129-131 between 5' ss and BS	CTAAT
FCIRC_06155_Major Facilitator Superfamily and Multidrug resistance protein Domains	2	47	1	None	CTAAC
FCIRC_06257_Cytochrome c oxidase subunit Via	2	123	1	TGA nt 72-74 between 5' ss and BS	CTAAC
FCIRC_06323_Neutral trehalase	3	55	0	TGA nt 7-9 between 5' ss and BS	CTGAC
FCIRC_06526_FHA domain, Protein kinase domain, Protein tyrosine kinase	4	50	1	TGA nt 6-8 (part of 5'ss)	CTAAT
FCIRC_06581	5	52	1	TGA nt 21-23 between 5' ss and BS	CTGAC
	6	55	1	None	CTAAC
	7	50	1	None	CTGAC
	9	49	0	TAA nt 2-4 of 5' ss	CTAAC
	10	46	0	TAA nt 2-4 of 5' ss	CTAAC
FCIRC_06732_ATP synthase alpha/beta family, nucleotide-binding domain	4	48	2	TAA nt 2-4 of 5' ss	CTAAC
FCIRC_07301_probable Methyltransferase domain, EasF family	2	117	2	TAA nt 2-4 of 5' ss	CTGAC
FCIRC_09592_Aminotransferase class I and II	1	70	1	None	CTAAC
FCIRC_09687_Mitochondrial carrier protein	1	78	0	TAA nt 64-66 (nt 1-3 of BS)	CTAAC

and EF hand Domains					
FCIRC_10565_Prenylcysteine lyase	1	130	1	TAG nt 24-26 between 5' ss and BS	CTAAC
FCIRC_10968	1	51	2	TAA nt 2-4 of 5' ss	TTGAT*
FCIRC_11342_60S Ribosomal protein L10a	2	54	1	None	CTAAC
FCIRC_12064_Peptide methionine sulfoxide reductase msrA 1	1	50	1	TAA nt 42-44 (nt 1-3 of BS)	CTAAC
FCIRC_12163_Bacteriorhodopsin	1	54	0	None	CTAAC
FCIRC_12655_40S Ribosomal protein S5-B	2	80	2	TAG nt 13-15 between 5' ss and BS	CTAAC
FCIRC_13218 (No hit)	1	55	2	None	CTAAC
FCIRC_13422_Alcohol dehydrogenase 1	1	84	1	TAA nt 24-26 (nt 1-3 of BS)	CTGAT
	3	51	2	TAA nt 37-39 (nt 1-3 of BS)	CTAAC
FCIRC_13979	3	58	1	TAG nt 30-32 between 5' ss and BS	CTGAT

ASSEMBLED ESTs

FCIRC_01390 Ctr copper transpoter family Domain	2	146	0	None	CTAAG*
FCIRC_02105_Uncharacterized protein family UPF0057 Domain	1	50	0	TAG nt 49-50 (nt 1-3 of 3' ss)	CTAAC
FCIRC_02340_Vacuolar ATP synthase subunit D	1 (at 5' UTR)	61		N/A	CTGAC
FCIRC_03285_14-3-3 Protein domain	1 (at 5' UTR)	669		N/A	CTAAC
FCIRC_04337	1	45	0	TAA nt 12-14 between 5'ss and BS	CTAAT
FCIRC_04669_Basic region leucine zipper and bZIP transcription factor Domains	1	52	1	TAA nt 39-41 (nt 1-3 of BS)	CTAAC
FCIRC_04971_Profilin	1	164	1	TAG nt 162-164 (nt 1-3 of 3' ss)	CTAAC or CTGAC
FCIRC_05496_40S Ribosomal protein S6-B	3	191	1	TAG nt 73-75 between 5'ss and BS	CTAAT
FCIRC_05768_FVEG_02524 - ARP2/3 complex 20 kDa subunit domain	1	60	0	TGA nt 13-15 between 5'ss and BS	CTGAC
FCIRC_06028_LICD Protein Family	3	56	0	None	CTGAC
FCIRC_07299	1 (at 5' UTR)	51		N/A	CTCAC*
FCIRC_08575	1	137	1	TAG nt 6-8 between 5'ss and BS	CTTAC*
FCIRC_09654	1	173	0	TAA at nt 154-156 (nt 2-4 of BS)	TTAAC*
FCIRC_10813_Thi4 family and thiazole biosynthesis enzyme Domains	2 (at 5' UTR)	179		N/A	CTAAC
FCIRC_11232	1 (at 5' UTR)	56		N/A	TTAAC*
FCIRC_11597_Ubiquitin family domain	1 (at 5' UTR)	87		N/A	CTGAC
FCIRC_12064_Peptide methionine sulfoxide reductase msrA 1	1	50	1	TAA nt 39-41 (nt 1-3 of BS)	CTAAT

FCIRC_14835_Mitochondrial cytochrome c oxidase assembly factor	1	109	0	TGA nt 64-66 between 5'ss and BS	CTGAC
FCIRC_14955	1	48	1	TAG nt 3-5 betweet 5'ss and BS	TTGAC*
FCIRC_15141 (No hit)	1	62	2	TGA nt 26-28 between 5'ss and BS	CTGAC

[a] Gene names and numbers refer to those in the Broad Institute and *F. circinatum* genome, respectively.
[b] The number of ESTs supporting the alternatively spliced introns in the various genes.
[c] As measure of the completeness of mRNA processing, ESTs are classed as either ambiguous (partially sequenced ESTs, starting or ending with truncated introns) or unambiguous ESTs (containing full-length introns) (Ho *et al*., 2007).
[d] Alternative intron positions (intron located at the 5' UTR are indicated in parentheses next to the intron position number).
[e] Intron length is indicated in nucleotides.
f Termination codons within genes with alternatively spliced introns are shown together with their positions relative to the 5' and 3' splice sites (ss) and the Branch Site (BS).
[g] Secondary branch sites are indicated with asterisks.

Table 3. The branch sites of alternative and constitutive introns identified using two *F. circinatum* EST datasets.

Putative branch site	Introns identified using the carbon and nitrogen starvation EST dataset		Introns identified using the assembled EST dataset	
	Constitutive	Alternative	Constitutive	Alternative
CTAAC	31/57; 54%	20/34; 59%	10/29; 34.5%	4/20; 20%
CTGAC	11/57; 19%	5/34; 15%	10/29; 34.5%	8/20; 40%
CTAAT	4/57; 7%	4/34; 12%	2/29; 7.0%	3/20; 15%
CTGAT	4/57; 7%		3/29; 10.3%	
CCAAT*		1/34; 3%		
TGAAT*		1/34; 3%		
TTAAC*	2/57; 4%	1/34; 3%		2/20; 10%
TTGAC*	1/57; 2%	1/34; 3%	1/29; 3.4%	1/20; 5%
TTGAT*		1/34; 3%		1/20; 5%
CTAAG*				1/20; 5%
CTCAC*	1/57;2%			1/20; 5%
CTAAA*	1/57; 2%		1/29; 3.4%	
CTTAC*	2/57; 4%		1/29; 3.4%	
CCAAC*			1/29; 3.4%	

* Secondary branch sites.

Figures

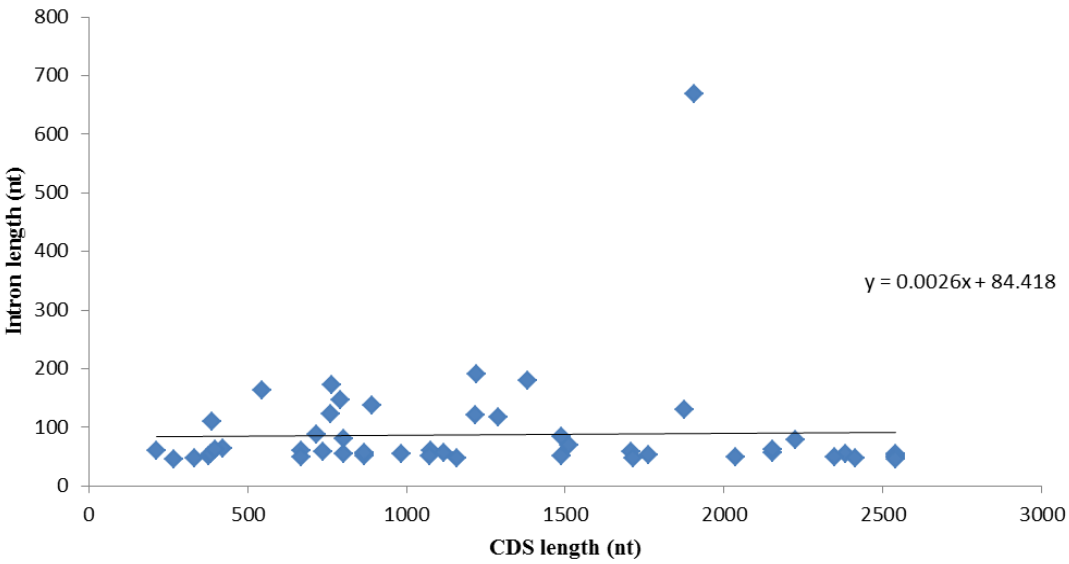


Figure 1. The relationship between CDS length and alternative intron length in *F. circinatum*.

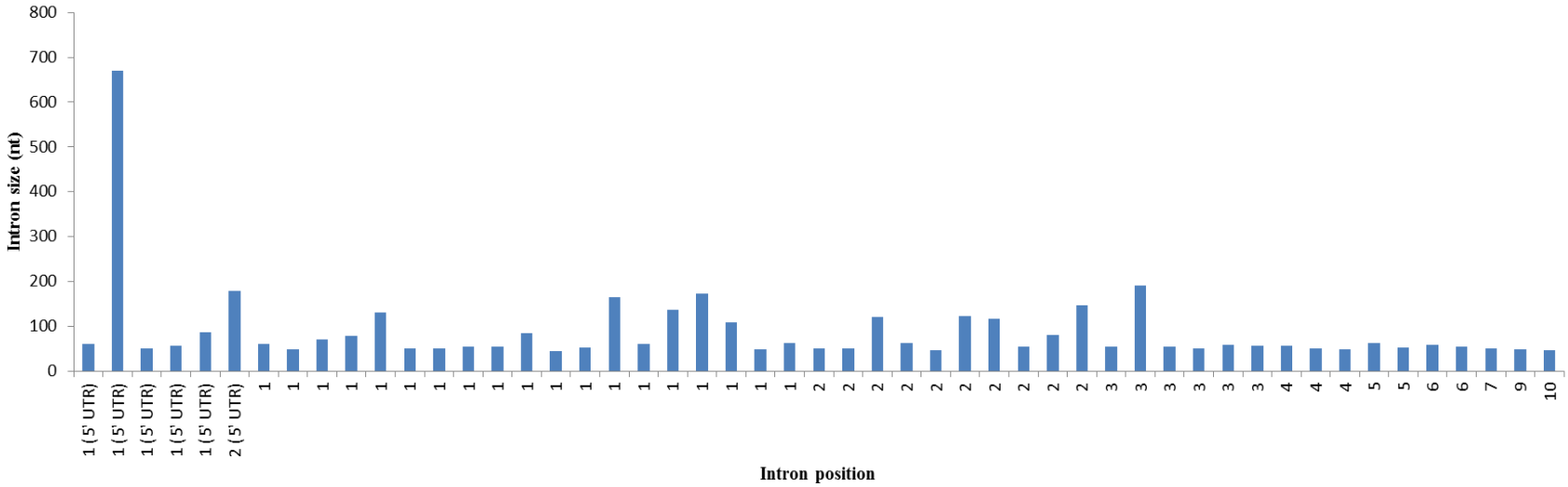
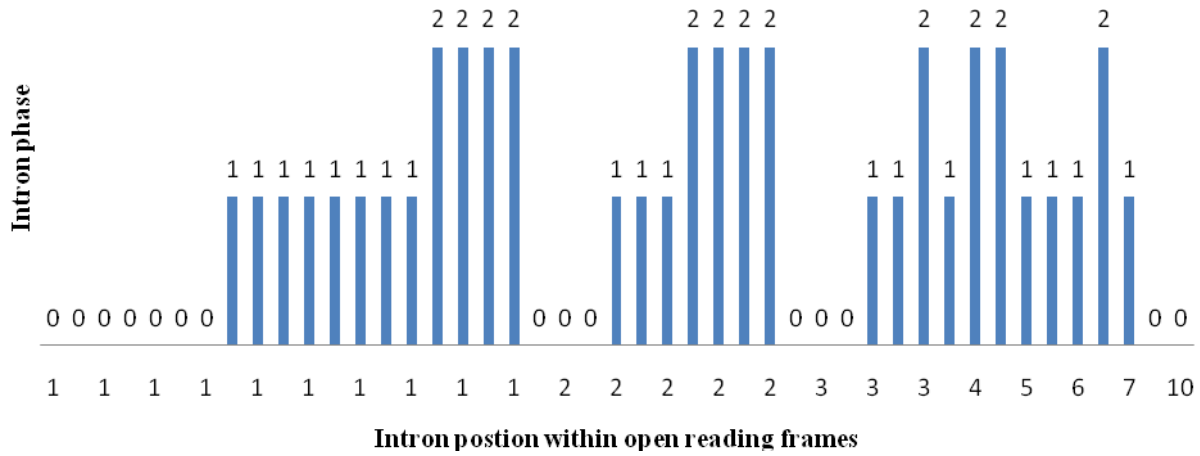
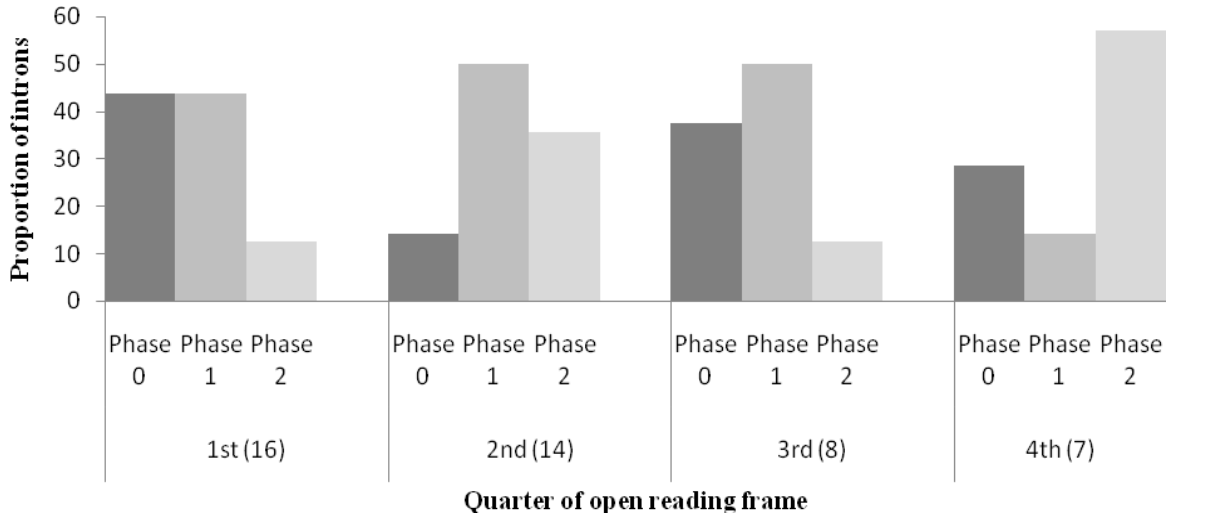


Figure 2. The relationship between alternative intron position and length in *F. circinatum*.



A



B

Figure 3. The distribution of alternative intron phases in *F. circinatum*. A) Distribution in terms of intron position. B) Distribution within the open reading frames.

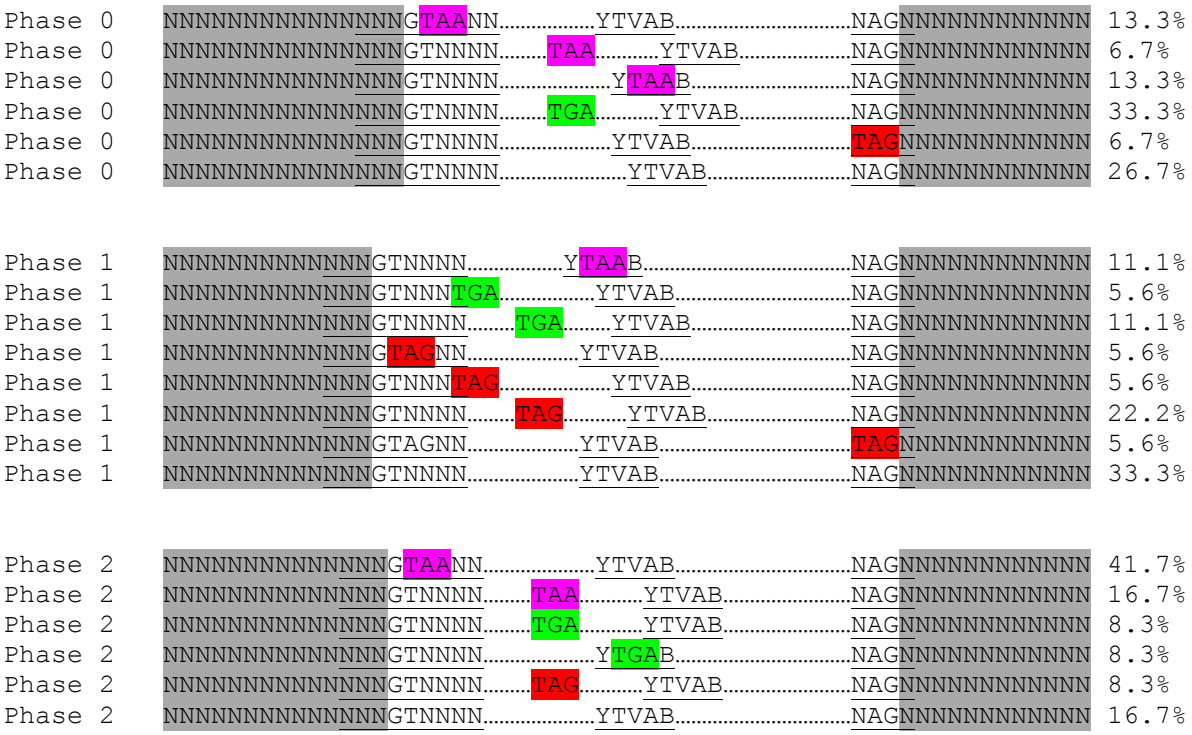


Figure 4. The distribution of premature termination codons in phase 0, 1 and 2 alternative introns of *F. circinatum*. Underlined nucleotides represent *cis*-elements (5' splice site, branch site and 3' splice site), colored nucleotides are premature termination codons, N represents any base, and flanking each intron are exons colored in grey. The last sequence in each phase represents introns that had no premature termination codons. R= AG, Y=CT, V=ACG, B=CGT. Phase 0 introns were 15 in total, phase 1 introns 18, and phase 2 introns 12. Percentages were derived from the latter total numbers.

Summary

The genus *Fusarium* constitutes fungi with diverse biological behaviours. This study focused on four plant pathogenic species. These were *F. verticillioides* which infects maize, *F. oxysporum* which infects tomato, *F. graminearum*, a pathogen of wheat and *F. circinatum*, which is pathogenic to pine. The genomes of *F. verticillioides*, *F. oxysporum*, *F. graminearum* and *F. circinatum* have been sequenced. These genomes were annotated using different gene prediction software. To study the architecture and distribution of Spliceosomal introns in these a set of Housekeeping (HK) genes common to all eukaryotes were used. These analyses revealed discrepancies in the annotations of these genomes, which most commonly included intron position incongruences, misidentified introns and sequencing errors.

Spliceosomal introns have four *cis*-elements which include the 5' and 3' splice sites, the branch site and the polypyrimidine tract. Analysis of the first three elements of Spliceosomal introns in the four *Fusarium* species and comparisons to those in other fungi showed significant differences in the consensus sequences of these elements. Two additional branch site motifs were also found for *Fusarium*, while the polypyrimidine tract of these species was found to be very diverse. The results also indicated that the first introns of the HK genes of the *Fusarium* species significantly longer, which is consistent with what have been found for genes of other eukaryotic. Also consistent with what is known for other eukaryotes, the analysed *Fusarium* genes had much lower intron densities than those observed in higher eukaryotes. An average of 2.53 introns per gene was observed in *Fusarium* and most of these introns were located closer to the 5' end of the HK genes. This average is low compared *A. thaliana* and *Homo sapiens* which have averages of 4.3 and 8.82 introns per gene, respectively.

With the aid of EST and genome data, *F. circinatum* was shown to harbour putative alternative introns at a frequency of approximately 0.3%. Homologues of a number of these genes from *F. verticillioides*, *F. oxysporum* and *F. graminearum* also harboured alternative splicing signals. Certain alternatively spliced transcripts harbored premature stop codons. These transcripts are targeted by non-sense mediated mRNA decay (NMD) system. The high rate at which these transcripts included premature stop codons suggested that such a quality control system is indeed

needed for these fungi. Overall, however, it remains to be investigated whether these alternative transcripts are functional as is the case with some of them in humans, plants and insects.

As more fungal genomes are being sequenced the need for accurate gene prediction methods is soaring. The incorporation of the findings of the architecture and distribution of Spliceosomal introns in *Fusarium* into gene prediction methods will thus increase the accuracy of such methods for *Fusarium* species, especially those related to *F. circinatum*. The identification of genes that are potentially regulated through alternative intron splicing also provide valuable targets for future studies on important biological processes such as pathogenicity and virulence.

Note: The disc provided contains a spreadsheet with all the data included in this dissertation.

Appendix

Table. The number of Spliceosomal introns per Housekeeping (HK) gene of *F. verticillioides*, *F. circinatum*, *F. oxysporum* and *F. graminearum*.

Gene name[a]	*S. cerevisiae* Id no.[b]	Biological process[c]	Number of introns[d]
Pre-mRNA-splicing factor clf1	YLR117c_KOG1915	Member of the NineTeen Complex (NTC) that contains Prp19p and stabilizes U6 snRNA in catalytic forms of the spliceosome containing U2, U5, and U6 snRNAs; interacts with U1 snRNP proteins	0
Aminomethyltransferase mitochondrial precursor	YDR019c_KOG2770	T subunit of the mitochondrial glycine decarboxylase complex, required for the catabolism of glycine to 5,10-methylene-THF; expression is regulated by levels of levels of 5,10-methylene-THF in the cytoplasm	0
Seryl-tRNA synthetase	YHR011w_KOG2509	Probable mitochondrial seryl-tRNA synthetase, mutant displays increased invasive and pseudohyphal growth	0
Uridylate kinase	YKL024c_KOG3079	Uridylate kinase, catalyzes the seventh enzymatic step in the de novo biosynthesis of pyrimidines, converting uridine monophosphate (UMP) into uridine-5'-diphosphate (UDP)	0
DNA repair helicase RAD25	YIL143c_KOG1123	Required, with Rad3p, for unwinding promoter DNA; involved in DNA repair	0
pre mRNA splicing factor Prp31	YGR091w_KOG2574	Splicing factor, component of the U4/U6-U5 snRNP complex	1
Multifunctional methyltransferase subunit TRM112	YNR046w_KOG1088	In combination with Trm9p and Trm11p; subunit of complex with Mtq2p that methylates Sup45p (eRF1) in the ternary complex eRF1-eRF3-GTP; deletion confers resistance to zymocin	1
2 oxoglutarate dehydrogenase E1 component mitochondrial precursor	YIL125w_KOG0450	Component of the mitochondrial alpha-ketoglutarate complex, which catalyzes a key step in the tricarboxylic acid (TCA) cycle, the oxidative decarboxylation of alpha-ketoglutarate to form succinyl-CoA	1
Geranylgeranyl transferase type-2 subunit beta	YPR176c_KOG0366	Required for vesicular transport between the endoplasmic reticulum and the Golgi; provides a	1

		membrane attachment moiety to Rab-like proteins Ypt1p and Sec4p	
Exonuclease	YKL113c_KOG2519	5' to 3' exonuclease, 5' flap endonuclease, required for Okazaki fragment processing and maturation as well as for long-patch base-excision repair; member of the S. pombe RAD2/FEN1 family	1
Methyltransferase	YCR047c_KOG1541	Methyltransferase, methylates residue G1575 of 18S rRNA; required for rRNA processing and nuclear export of 40S ribosomal subunits independently of methylation activity; diploid mutant displays random budding pattern	1
Glycerol 3 phosphate dehydrogenase	YOL059w_KOG2711	NAD-dependent glycerol 3-phosphate dehydrogenase, homolog of Gpd1p, expression is controlled by an oxygen-independent signaling pathway required to regulate metabolism under anoxic conditions; located in cytosol and mitochondria	1
Superoxide dismutase mitochondrial precursor	YHR008c_KOG0876	Mitochondrial manganese superoxide dismutase, protects cells against oxygen toxicity	1
Mitochondrial import inner membrane translocase subunit TIM16	YJL104w_KOG3442	Constituent of the import motor (PAM complex)component of the Translocase of the Inner Mitochondrial membrane (TIM23 complex); forms a 1:1 subcomplex with Pam18p and inhibits its cochaperone activity; contains a J-like domain	1
Mitochondrial inner membrane i-AAA protease supercomplex subunit YME1	YPR024w_KOG0734	Essential for degradation of unassembled subunit 2 of cytochrome c oxidase in yeast mitochondria	1
proteasome component C5	YBL041w_KOG0179	Required for the receptor-mediated retrieval of luminal ER proteins from the secretory pathway	1
60S acidic ribosomal protein P2-beta	YDR382w_KOG3449	Ribosomal protein P2 beta, a component of the ribosomal stalk, which is involved in the interaction between translational elongation factors and the ribosome; regulates the accumulation of P1 (Rpp1Ap and Rpp1Bp) in the cytoplasm	1
Adenosylhomocysteinase	YER043c_KOG1370	S-adenosyl-L-homocysteine hydrolase, catabolizes S-adenosyl-L-homocysteine which is formed after donation of the activated methyl group of S-adenosyl-L-methionine (AdoMet) to an acceptor	1****
Dolichyl pyrophosphate Man9GlcNAc2 alpha-	YOR002w_KOG2575	Involved in transfer of oligosaccharides from dolichyl	1

1,3-glucosyltransferase		pyrophosphate to asparagine residues of proteins during N-linked protein glycosylation	
Elongation factor Tu domain 2	YOR187w_KOG0460	Comprises both GTPase and guanine nucleotide exchange factor activities, while these activities are found in separate proteins in S. pombe and humans	1
pyridoxal phosphate binding			1
Ubiquinone biosynthesis monooxgenase Coq6	YGR255c_KOG3855	Putative flavin-dependent monooxygenase; involved in ubiquinone (Coenzyme Q) biosynthesis; localizes to the matrix face of the mitochondrial inner membrane in a large complex with other ubiquinone biosynthetic enzymes	1
Serine/threonine-protein kinase KIN28	YDL108w_KOG0659	Subunit of the transcription factor TFIIH; involved in transcription initiation at RNA polymerase II promoters	1
Mannose-6-phosphate isomerase	YER003c_KOG2757	Catalyzes the interconversion of fructose-6-P and mannose-6-P; required for early steps in protein mannosylation	1
Mitochondrial-processing peptidase alpha subunit	YHR024c_KOG2067	Essential processing enzyme that cleaves the N-terminal targeting sequences from mitochondrially imported proteins	1
U3 small nucleolar RNA-associated protein 7	YER082c_KOG1272	Nucleolar protein, component of the small subunit (SSU) processome containing the U3 snoRNA that is involved in processing of pre-18S rRNA	1
N-acetylglucosamine phosphate mutase	YEL058w_KOG2537	Essential N-acetylglucosamine-phosphate mutase; converts GlcNAc-6-P to GlcNAc-1-P, which is a precursor for the biosynthesis of chitin and for the formation of N-glycosylated mannoproteins and glycosylphosphatidylinositol anchors	1
Mitochondrial ornithine transporter 1	YOR130c_KOG0758	Ornithine transporter of the mitochondrial inner membrane, exports ornithine from mitochondria as part of arginine biosynthesis	1
Galactokinase	YBR020w_KOG0631	A divergent promoter region in the yeast gal gene cluster is responsible for the LeLoir pathway enzymes necessary for the utilization of galactose	1
ATP-dependent molecular chaperone HSC82/ Heat shock protein Hsp90 constitutive isoform	YMR186w_KOG0019	Required in higher concentrations for growth of cells at higher temperatures	1

Splicing factor 3a subunit 3/ Pre-mRNA-splicing factor PRP9	YDL030w_KOG2636	Subunit of the SF3a splicing factor complex,required for spliceosome assembly; acts after the formation of the U1 snRNP-pre-mRNA complex	1
U3 small nucleolar ribonucleoprotein protein IMP3	YHR148w_KOG4655	Component of the SSU processome, which is required for pre-18S rRNA processing, essential protein that interacts with Mpp10p and mediates interactions of Imp4p and Mpp10p with U3 snoRNA	1
Proteasome component PUP3	YER094c_KOG0180	Nucleotide sequence and transcriptional regulation of the yeast recombinational repair gene RAD51	1
U3 small nucleolar RNA-associated protein 11	YKL099c_KOG3237	Subunit of U3-containing Small Subunit (SSU) processome complex involved in production of 18S rRNA and assembly of small ribosomal subunit	1
26S protease regulatory subunit 8 homolog	YGL048c_KOG0728	Gene encodes a protein homologous to the human Tat-binding protein TBP-1	1
Adenine phosphoribosyltransferase	YML022w_KOG1712	Adenine phosphoribosyltransferase, catalyzes the formation of AMP from adenine and 5-phospho-ribosylpyrophosphate; involved in the salvage pathway of purine nucleotide biosynthesis	1
Signal recognition particle subunit SRP54 / Signal recognition particle 54 kDa protein homolog	YPR088c_KOG0780	Essential for growth	1
Ribosome assembly protein RRB1	YMR131c_KOG0302	Essential nuclear protein involved in early steps of ribosome biogenesis; physically interacts with the ribosomal protein Rpl3p	1
Mitochondrial pyruvate dehydrogenase kinase	YIL042c_KOG0787	Involved in negative regulation of pyruvate dehydrogenase complex activity by phosphorylating the ser-133 residue of the Pda1p subunit; acts in concert with kinase Pkp2p and phosphatases Ptc5p and Ptc6p	1
Proteosome component PRE2 precursor	YPR103w_KOG0175	Beta 5 subunit of the 20S proteasome, responsible for the chymotryptic activity of the proteasome	1
Pyruvate dehydrogenase E1 component subunit beta	YBR221c_KOG0524	E1 beta subunit of the pyruvate dehydrogenase (PDH) complex, which is an evolutionarily-conserved multi-protein complex found in mitochondria	1
ATP dependent protease La/ Lon protease homolog, mitochondrial	YBL022c_KOG2004	ATP-dependent Lon protease, involved in degradation of misfolded proteins in mitochondria; required for biogenesis and maintenance of mitochondria	1

GPI anchor transamide precursor	YDR331w_KOG1349	ER membrane glycoprotein subunit of the glycosyl-phosphatidylinositol transamidase complex that adds glycosylphosphatidylinositol (GPI) anchors to newly synthesized proteins	1
Ubiquitin-like protein SMT3	YDR510w_KOG1769	Ubiquitin-like protein of the SUMO family, conjugated to lysine residues of target proteins; regulates chromatid cohesion, chromosome segregation, APC-mediated proteolysis, DNA replication and septin ring dynamics; phosphorylated at Ser2	1
40S ribosomal protein S9	YBR189w_KOG3301	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps9Ap and has similarity to E. coli S4 and rat S9 ribosomal proteins	2
Proteasome regulatory particle subunit Rpn /26S proteasome regulatory subunit RPN7	YPR108w_KOG0687	Essential, non-ATPase regulatory subunit of the 26S proteasome, similar to another S. cerevisiae regulatory subunit, Rpn5p	2
Histidyl-tRNA synthetase, mitochondrial precursor	YPR033c_KOG1936	Cytoplasmic and mitochondrial histidine tRNA synthetase; encoded by a single nuclear gene that specifies two messages; efficient mitochondrial localization requires both a presequence and an amino-terminal sequence	2
Fructose-1, 6-bisphosphatase	YLR377c_KOG1458	Hey regulatory enzyme in the gluconeogenesis pathway, required for glucose metabolism; undergoes either proteasome-mediated or autophagy-mediated degradation depending on growth conditions; interacts with Vid30p	2
Phosphomannomutase	YFL045c_KOG3189	Involved in synthesis of GDP-mannose and dolichol-phosphate-mannose; required for folding and glycosylation of secretory proteins in the ER lumen	2
AP-1 complex subunit mu-1-I	YPL259c_KOG0937	Mu1-like medium subunit of the clathrin-associated protein complex (AP-1); binds clathrin; involved in clathrin-dependent Golgi protein sorting	2***
Ubiquitin fusion degradation protein 1	YGR048w_KOG1816	Protein that interacts with Cdc48p and Npl4p, involved in recognition of polyubiquitinated proteins and their presentation to the 26S proteasome for degradation; involved in transporting proteins from the ER to the cytosol	2
Aspartyl tRNA synthetase	YLL018c_KOG0556	Primarily cytoplasmic; homodimeric enzyme that catalyzes the specific aspartylation of tRNA(Asp); class	2****

		II aminoacyl tRNA synthetase; binding to its own mRNA may confer autoregulation	
DNA mismatch repair protein Mlh1	YMR167w_KOG1979	Protein required for mismatch repair in mitosis and meiosis as well as crossing over during meiosis; forms a complex with Pms1p and Msh2p-Msh3p during mismatch repair	2
60S ribosomal protein L28/ L27a	YGL103w_KOG1742	Has similarity to E. coli L15 and rat L27a ribosomal proteins; may have peptidyl transferase activity; can mutate to cycloheximide resistance	2
Vacuolar protein sorting-associated protein 28/ VPS28	YPL065w_KOG3284	Component of the ESCRT-I complex (Stp22p, Srn2p, Vps28p, and Mvb12p), which is involved in ubiquitin-dependent sorting of proteins into the endosome; conserved C-terminal domain interacts with ESCRT-III subunit Vps20p	2
Vacuolar protein sorting-associated protein 21	YOR089c_KOG0092	Rab family GTPase required for endocytic transport and for sorting of vacuolar hydrolases; localized in endocytic intermediates; detected in mitochondria; geranylgeranylation required for membrane association	2
Alanyl-tRNA synthetase; Mitochondrial	YOR335c_KOG0188	Required for protein synthesis; point mutation (cdc64-1 allele) causes cell cycle arrest at G1	2
Replication factor C subunit 5/ RFC5	YBR087w_KOG2035	Subunit of heteropentameric Replication factor C (RF-C), which is a DNA binding protein and ATPase that acts as a clamp loader of the proliferating cell nuclear antigen (PCNA) processivity factor for DNA polymerases delta and epsilon	2
Succinate dehydrogenase [ubiquinone] iron-sulfur subunit, mitochondrial	YLL041c_KOG3049	Iron-sulfur protein subunit of succinate dehydrogenase (Sdh1p, Sdh2p, Sdh3p, Sdh4p), which couples the oxidation of succinate to the transfer of electrons to ubiquinone as part of the TCA cycle and the mitochondrial respiratory chain	2
40S ribosomal protein S18-B	YML026c_KOG3311	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps18Ap and has similarity to E. coli S13 and rat S18 ribosomal proteins	2
60S ribosomal protein L24-A	YGL031c_KOG1722	Ribosomal protein L30 of the large (60S) ribosomal subunit, nearly identical to Rpl24Bp and has similarity to rat L24 ribosomal protein; not essential for translation but may be required	2

Protein transport protein SEC61	YLR378c_KOG1373	Essential subunit of Sec61 complex (Sec61p, Sbh1p, and Sss1p); forms a channel for SRP-dependent protein import and retrograde transport of misfolded proteins out of the ER; with Sec63 complex allows SRP-independent protein import into ER	2
Protein sco1	YBR037c_KOG2792	Copper-binding protein of the mitochondrial inner membrane, required for cytochrome c oxidase activity and respiration; may function to deliver copper to cytochrome c oxidase; has similarity to thioredoxins	2
40S ribosomal protein S19-A	YOL121c_KOG3411	Protein component of the small (40S) ribosomal subunit, required for assembly and maturation of pre-40 S particles; mutations in human RPS19 are associated with Diamond Blackfan anemia; nearly identical to Rps19Bp	2
Replication factor C subunit 4	YOL094c_KOG0991	Subunit of heteropentameric Replication factor C (RF-C), which is a DNA binding protein and ATPase that acts as a clamp loader of the proliferating cell nuclear antigen (PCNA) processivity factor for DNA polymerases delta and epsilon	2
Transcription initiation factor IIa small chain	YKL058w_KOG3463	TFIIA small subunit; involved in transcriptional activation, acts as antirepressor or as coactivator	2
60S ribosomal protein L10	YLR075w_KOG0857	Protein component of the large (60S) ribosomal subunit, responsible for joining the 40S and 60S subunits; regulates translation initiation; has similarity to rat L10 ribosomal protein and to members of the QM gene family	2
Phosphoglycerate kinase	YCR012w_KOG1367	Catalyzes transfer of high-energy phosphoryl groups from the acyl phosphate of 1,3-bisphosphoglycerate to ADP to produce ATP; key enzyme in glycolysis and gluconeogenesis	2
26S proteasome non ATPase regulatory subunit 11	YFR004W	Metalloprotease subunit of the 19S regulatory particle of the 26S proteasome lid; couples the deubiquitination and degradation of proteasome substrates; involved, independent of catalytic activity, in fission of mitochondria and peroxisomes	2
F actin capping protein subunit beta	YIL034c_KOG3174	Beta subunit of the capping protein (CP) heterodimer (Cap1p and Cap2p) which binds to the barbed ends of actin filaments preventing further polymerization;	2

		localized predominantly to cortical actin patches	
Mitochondrial phosphate carrier protein 2	YER053c_KOG0767	Imports inorganic phosphate into mitochondria; functionally redundant with Mir1p but less abundant than Mir1p under normal conditions; expression is induced at high temperature	2
Probable cation-transporting ATPase 1	YEL031w_KOG0209	Ion transporter of the ER membrane involved in ER function and Ca2+ homeostasis; required for regulating Hmg2p degradation; confers sensitivity to a killer toxin (SMKT) produced by *Pichia farinosa* KK1	2
small nuclear ribonucleoprotein LSM1 Sm-like protein LSm1?	YJL124c_KOG1782	Lsm (Like Sm) protein; forms heteroheptameric complex (with Lsm2p, Lsm3p, Lsm4p, Lsm5p, Lsm6p, and Lsm7p) involved in degradation of cytoplasmic mRNAs	2
ATP-dependent RNA helicase FAL1	YDR021w_KOG0328	Nucleolar protein required for maturation of 18S rRNA, member of the eIF4A subfamily of DEAD-box ATP-dependent RNA helicases	2
Malate dehydrogenase, mitochondrial	YKL085w_KOG1494	Catalyzes interconversion of malate and oxaloacetate; involved in the tricarboxylic acid (TCA) cycle; phosphorylated	2
Ribosomal protein S28e/ 40S ribosomal protein S28-A	YOR167c_KOG3502	Protein component of the small (40S) ribosomal subunit; nearly identical to Rps28Bp	2
S-adenosylmethionine synthetase	YLR180w_KOG1506	S-adenosylmethionine synthetase, catalyzes transfer of the adenosyl group of ATP to the sulfur atom of methionine; one of two differentially regulated isozymes (Sam1p and Sam2p)	2
Eukaryotic translation initiation factor 2 subunit gamma	YER025w_KOG0466	Involved in the identification of the start codon; binds GTP when forming the ternary complex with GTP and tRNAi-Met	2
Eukaryotic translation initiation factor 3 subunit G	YDR429c_KOG0122	eIF3g subunit of the core complex of translation initiation factor 3 (eIF3), which is essential for translation; stimulates resumption of ribosomal scanning during translation reinitiation	2
Peroxiredoxin TSA1	YML028w_KOG0852	Thioredoxin peroxidase, acts as both a ribosome-associated and free cytoplasmic antioxidant; self-associates to form a high-molecular weight chaperone complex under oxidative stress; deletion results in mutator phenotype	2

Protein farnesyltransferase/geranylgeranyltransferase type-1 subunit alpha	YKL019w_KOG0530	Catalyzes prenylation of proteins containing a CAAX consensus motif; essential protein required for membrane localization of Ras proteins and a-factor	2
Protein disulfide-isomerase	YCL043c_KOG0190	Multifunctional protein resident in the endoplasmic reticulum lumen, essential for the formation of disulfide bonds in secretory and cell-surface proteins, unscrambles non-native disulfide bonds	2
Proteasome component Pre4	YFR050c_KOG0185	Beta 7 subunit of the 20S proteasome	2
60S acidic ribosomal protein P0	YLR340w_KOG0815	Conserved ribosomal protein P0 of the ribosomal stalk, which is involved in interaction between translational elongation factors and the ribosome	2
40S ribosomal protein S7	YOR096w_KOG3320	Protein component of the small (40S) ribosomal subunit, nearly identical to Rps7Bp; interacts with Kti11p; deletion causes hypersensitivity to zymocin	2
rRNA processing protein Rrp20/ Pre-rRNA-processing protein PNO1	YOR145c_KOG3273	Essential nucleolar protein required for pre-18S rRNA processing, interacts with Dim1p, an 18S rRNA dimethyltransferase, and also with Nob1p, which is involved in proteasome biogenesis; contains a KH domain	2
Isocitrate dehydrogenase NADP dependent 1/ Isocitrate dehydrogenase [NADP] cytoplasmic	YLR174w_KOG1526	Cytosolic NADP-specific isocitrate dehydrogenase, catalyzes oxidation of isocitrate to alpha-ketoglutarate; levels are elevated during growth on non-fermentable carbon sources and reduced during growth on glucose	2
Heat shock protein SSC1, mitochondrial	YJR045c_KOG0102	Hsp70 family ATPase, constituent of the import motor component of the Translocase of the Inner Mitochondrial membrane (TIM23 complex); involved in protein translocation and folding; subunit of SceI endonuclease	2****
60S ribosomal protein L11-B	YGR085c_KOG0397	Nearly identical to Rpl11Ap; involved in ribosomal assembly; depletion causes degradation of proteins and RNA of the 60S subunit	2
Isocitrate dehydrogenase [NAD] subunit 1, mitochondrial	YNL037c_KOG0784	Catalyzes the oxidation of isocitrate to alpha-ketoglutarate in the TCA cycle	2
60S ribosomal protein L15-B	YMR121c_KOG1678	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl15Ap	2

Proteasome component c1	YOR362c_KOG0184	Alpha 7 subunit of the 20S proteasome	2
Peptidyl-prolyl cis-trans isomerase B (cyclophilin)	YHR057c_KOG0880	Catalyzes the cis-trans isomerization of peptide bonds N-terminal to proline residues; has a potential role in the secretory pathway	3
Methionine aminopeptidase 1	YLR244c_KOG2738	Catalyzes the cotranslational removal of N-terminal methionine from nascent polypeptides; function is partially redundant with that of Map2p	3
T complex protein 1 subunit eta	YJL111w_KOG0361	Subunit of the cytosolic chaperonin Cct ring complex, related to Tcp1p, required for the assembly of actin and tubulins in vivo; mutant has increased aneuploidy tolerance	3
60S ribosomal protein L37-A	YLR185w_KOG3475	Has similarity to Rpl37Bp and to rat L37 ribosomal protein	3
Serine palmitoyltransferase 1	YMR296c_KOG1358	Component of serine palmitoyltransferase, responsible along with Lcb2p for the first committed step in sphingolipid synthesis, which is the condensation of serine with palmitoyl-CoA to form 3-ketosphinganine	3
Structural maintenance of chromosomes protein 3	YJL074c_KOG0964	Subunit of the multiprotein cohesin complex required for sister chromatid cohesion in mitotic cells; also required, with Rec8p, for cohesion and recombination during meiosis; phylogenetically conserved SMC chromosomal ATPase family member	3
40S ribosomal protein S13	YDR064w_KOG0400	Protein component of the small (40S) ribosomal subunit; has similarity to E. coli S15 and rat S13 ribosomal proteins	3****(4)
Methionine aminopeptidase 2	YBL091c_KOG2775	Protein involved in regulation of phospholipid metabolism; homolog of Scs2p	3
Mannose-1-phosphate guanyltransferase (GDP-mannose pyrophosphorylase)	YDL055c_KOG1322	Synthesizes GDP-mannose from GTP and mannose-1-phosphate in cell wall biosynthesis; required for normal cell wall structure	3
Triosephosphate isomerase	YDR050c_KOG1643	Abundant glycolytic enzyme; mRNA half-life is regulated by iron availability; transcription is controlled by activators Reb1p, Gcr1p, and Rap1p through binding sites in the 5' non-coding region	3
Trafficking protein particle complex subunit	YKR068c_KOG3330	Hydrophilic protein that acts in conjunction with	3

BET3		SNARE proteins in targeting and fusion of ER to Golgi transport vesicles; component of the TRAPP (transport protein particle) complex	
Clathrin heavy chain	YGL206c_KOG0985	Subunit of the major coat protein involved in intracellular protein transport and endocytosis; two heavy chains form the clathrin triskelion structural component; the light chain (CLC1) is thought to regulate function	3
26S protease regulatory subunit 6B homolog	YDR394w_KOG0727	One of six ATPases of the 19S regulatory particle of the 26S proteasome involved in the degradation of ubiquitinated substrates; substrate of N-acetyltransferase B	3
40S ribosomal protein S15	YOL040c_KOG0898	Has similarity to E. coli S19 and rat S15 ribosomal proteins	3
AP-2 complex subunit alpha	YBL037w_KOG1077	Large subunit of the clathrin associated protein complex (AP-2); involved in vesicle mediated transport	3
ATP synthase subunit 5; mitochondrial	YDR298c_KOG1662	Subunit 5 of the stator stalk of mitochondrial F1F0 ATP synthase, which is an evolutionarily conserved enzyme complex required for ATP synthesis; phosphorylated	3
Centromere/microtubule-binding protein CBF5	YLR175w_KOG2529	Pseudouridine synthase catalytic subunit of box H/ACA small nucleolar ribonucleoprotein particles (snoRNPs), acts on both large and small rRNAs and on snRNA U2	3
Protein translation factor SUI1 / Eukaryotic translation initiation factor eIF-1	YNL244c_KOG1770	Translation initiation factor eIF1; component of a complex involved in recognition of the initiator codon; modulates translation accuracy at the initiation phase	3
V-type proton ATPase subunit c''	YHR026w_KOG0233	Functions in acidification of the vacuole; one of three proteolipid subunits of the V0 domain	3
ER lumen protein retaining receptor	YBL040c_KOG3106	HDEL receptor, an integral membrane protein that binds to the HDEL motif in proteins destined for retention in the endoplasmic reticulum; has a role in maintenance of normal levels of ER-resident proteins	3
Serine/threonine protein phosphatase PP1-1	YDL047w_KOG0373	Type 2A-related serine-threonine phosphatase that functions in the G1/S transition of the mitotic cycle; cytoplasmic and nuclear protein that modulates functions mediated by Pkc1p including cell wall and actin cytoskeleton organization	3
Coatomer subunit beta	YDR238c_KOG1058	Essential beta-coat protein of the COPI coatomer, involved in ER-to-Golgi protein trafficking and	3

		maintenance of normal ER morphology; shares 43% sequence identity with mammalian beta-coat protein (beta-COP)	
Small nuclear ribonucleoprotein Sm D2	YLR275w_KOG3459	Core Sm protein Sm D2; part of heteroheptameric complex (with Smb1p, Smd1p, Smd3p, Sme1p, Smx3p, and Smx2p) that is part of the spliceosomal U1, U2, U4, and U5 snRNPs	3
Histone deacetylase RPD3	YNL330c_KOG1342	Regulates transcription, silencing, and other processes by influencing chromatin remodeling; forms at least two different complexes which have distinct functions and members	3***
40S ribosomal protein S14-A	YCR031c_KOG0407	Ribosomal protein 59 of the small subunit, required for ribosome assembly and 20S pre-rRNA processing; mutations confer cryptopleurine resistance; nearly identical to Rps14Bp and similar to *E. coli* S11 and rat S14 ribosomal protein	3
Glucosamine--fructose-6-phosphate aminotransferase [isomerizing]	YKL104c_KOG1268	Catalyzes the formation of glucosamine-6-P and glutamate from fructose-6-P and glutamine in the first step of chitin biosynthesis	3
neutral trehalase	YDR001c_KOG0602	Degrades trehalose; required for thermotolerance and may mediate resistance to other cellular stresses; may be phosphorylated by Cdc28p	3
Elongation factor 1-beta	YAL003w_KOG1668	Translation elongation factor 1 beta; stimulates nucleotide exchange to regenerate EF-1 alpha-GTP for the next elongation cycle; part of the EF-1 complex, which facilitates binding of aminoacyl-tRNA to the ribosomal A site	3
DNA polymerase delta subunit	YJR006w_KOG2732	DNA polymerase III (delta) subunit, essential for cell viability; involved in DNA replication and DNA repair	3
Cytochrome c1 mitochondrial precursor/ Cytochrome c1, heme protein, mitochondrial	YOR065w_KOG3052	Component of the mitochondrial respiratory chain; expression is regulated by the heme-activated, glucose-repressed Hap2p/3p/4p/5p CCAAT-binding complex	3
ATP synthase gamma chain mitochondrial precursor/ ATP synthase subunit gamma, mitochondrial	YBR039w_KOG1531	Gamma subunit of the F1 sector of mitochondrial F1F0 ATP synthase, which is a large, evolutionarily conserved enzyme complex required for ATP synthesis	3
pfkB family carbohydrate kinase			3

Protein SYL1	YDR189w_KOG1301	Hydrophilic protein involved in vesicle trafficking between the ER and Golgi; SM (Sec1/Munc-18) family protein that binds the tSNARE Sed5p and stimulates its assembly into a trans-SNARE membrane-protein complex	4
60S ribosomal protein L16-A	YIL133c_KOG3204	N-terminally acetylated protein component of the large (60S) ribosomal subunit, binds to 5.8 S rRNA; transcriptionally regulated by Rap1p	4
ATP dependent RNA helicase SUB2	YDL084w_KOG0329	Component of the TREX complex required for nuclear mRNA export; member of the DEAD-box RNA helicase superfamily and is involved in early and late steps of spliceosome assembly	4
Heat shock 70 kDA protein/ Heat shock protein SSC1, mitochondrial/ mtHSP70	YJR045c_KOG0102	Hsp70 family ATPase, constituent of the import motor component of the Translocase of the Inner Mitochondrial membrane (TIM23 complex); involved in protein translocation and folding; subunit of SceI endonuclease	4
60S ribosomal protein L13-B	YMR142c_KOG3295	Protein component of the large (60S) ribosomal subunit, nearly identical to Rpl13Ap; not essential for viability; has similarity to rat L13 ribosomal protein	4
Leukotriene A 4 hydrolase	YNL045w_KOG1047	Leucyl aminopeptidase yscIV (leukotriene A4 hydrolase) with epoxide hydrolase activity, metalloenzyme containing one zinc atom; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm and nucleus	4****
CaaX farnesyltransferase beta subunit	YDL090c_KOG0365	Prenylates the a-factor mating pheromone and Ras proteins; required for the membrane localization of Ras proteins and a-factor	4
RNA polymerase III subunit Rpc25	YKL144c_KOG3297	RNA polymerase III subunit C25, required for transcription initiation; forms a heterodimer with Rpc17p; paralog of Rpb7p	4
Cohesin complex subunit Psm1/ Structural maintenance of chromosomes protein 1	YFL008w_KOG0018	Subunit of the multiprotein cohesin complex, essential protein involved in chromosome segregation and in double-strand DNA break repair; SMC chromosomal ATPase family member, binds DNA with a preference for DNA with secondary structure	4
Proteasome subunit alpha type 6/ Proteasome component C7-alpha	YGL011c_KOG0182	Alpha 1 subunit of the 20S proteasome involved in the degradation of ubiquitinated substrates; 20S	4

		proteasome is the core complex of the 26S proteasome; essential for growth; detected in the mitochondria	
Coatomer subunit beta	YDR238c_KOG1058	Essential beta-coat protein of the COPI coatomer, involved in ER-to-Golgi protein trafficking and maintenance of normal ER morphology	4
ATP synthase delta chain mitochondrial precursor/ ATP synthase subunit 5, mitochondrial	YDR298c_KOG1662	Subunit 5 of the stator stalk of mitochondrial F1F0 ATP synthase, which is an evolutionarily conserved enzyme complex required for ATP synthesis; phosphorylated	4
GTP-binding protein YPT1	YFL038c_KOG0084	Rab family GTPase, involved in the ER-to-Golgi step of the secretory pathway; complex formation with the Rab escort protein Mrs6p is required for prenylation of Ypt1p by protein geranylgeranyltransferase type II (Bet2p-Bet4p)	4
Cell cycle control protein Cwf8 (Pre-mRNA-splicing factor 19???)	YLL036c_KOG0289	Splicing factor associated with the spliceosome; contains a U-box, a motif found in a class of ubiquitin ligases	4
Oligosaccharyl transferase stt3 subunit	YGL022w_KOG2292	Catalyzes asparagine-linked glycosylation of newly synthesized proteins; forms a subcomplex with Ost3p and Ost4p and is directly involved in catalysis	4
Guanine nucleotide-binding protein subunit beta-like protein	YMR116c_KOG0279	G-protein beta subunit and guanine nucleotide dissociation inhibitor for Gpa2p; ortholog of RACK1 that inhibits translation; core component of the small (40S) ribosomal subunit; represses Gcn4p in the absence of amino acid starvation	4
Translationally controlled tumor protein homolog	YKL056c_KOG1727	Protein that associates with ribosomes; green fluorescent protein (GFP)-fusion protein localizes to the cytoplasm and relocates to the mitochondrial outer surface upon oxidative stress	4
AP-1 complex subunit sigma-1	YLR170c_KOG0934	Small subunit of the clathrin-associated adaptor complex AP-1, which is involved in protein sorting at the trans-Golgi network	4
Isocitrate dehydrogenase [NAD] subunit 1, mitochondrial	YNL037c_KOG0784	Subunit of mitochondrial NAD(+)-dependent isocitrate dehydrogenase, which catalyzes the oxidation of isocitrate to alpha-ketoglutarate in the TCA cycle	4
Phosphoglucomutase-2	YMR105c_KOG0625	Catalyzes the conversion from glucose-1-phosphate to glucose-6-phosphate, which is a key step in hexose metabolism; functions as the acceptor for a Glc-	4

		phosphotransferase	
Heat shock protein homolog SSE1	YPL106c_KOG0103	ATPase that is a component of the heat shock protein Hsp90 chaperone complex; binds unfolded proteins; member of the heat shock protein 70 (HSP70) family; localized to the cytoplasm	5
Transaldolase	YLR354c_KOG2772	Transaldolase, enzyme in the non-oxidative pentose phosphate pathway; converts sedoheptulose 7-phosphate and glyceraldehyde 3-phosphate to erythrose 4-phosphate and fructose 6-phosphate	5
T complex protein 1 subunit zeta	YDR188w_KOG0359	Subunit of the cytosolic chaperonin Cct ring complex, related to Tcp1p, essential protein that is required for the assembly of actin and tubulins in vivo; contains an ATP-binding motif	5
Proteosome component PRE3	YJL001w_KOG0174	Beta 1 subunit of the 20S proteasome, responsible for cleavage after acidic residues in peptides	5
Coatomer subunit alpha	YDL145c_KOG0292	Alpha subunit of COPI vesicle coatomer complex, which surrounds transport vesicles in the early secretory pathway	5
Hit family protein 1	YDL125c_KOG3275	Adenosine 5'-monophosphoramidase; interacts physically and genetically with Kin28p, a CDK and TFIIK subunit, and genetically with CAK1; member of the histidine triad (HIT) superfamily of nucleotide-binding proteins and similar to Hint	5
pac10 protein/ Prefoldin subunit 3	YGR078c_KOG3313	Part of the heteromeric co-chaperone GimC/prefoldin complex, which promotes efficient protein folding	5
Mitotic spindle checkpoint component MAD2	YJL030w_KOG3285	Delays the onset of anaphase in cells with defects in mitotic spindle assembly; forms a complex with Mad1p	5
GTP-binding protein YPT6	YLR262c_KOG0094	Rab family GTPase, Ras-like GTP binding protein involved in the secretory pathway, required for fusion of endosome-derived vesicles with the late Golgi, maturation of the vacuolar carboxypeptidase Y	5
T-complex protein 1 subunit gamma	YJL014w_KOG0364	Subunit of the cytosolic chaperonin Cct ring complex, related to Tcp1p, required for the assembly of actin and tubulins in vivo	5
Inosine triphosphate pyrophosphatase	YJR069c_KOG3222	Conserved protein with deoxyribonucleoside	5****

		triphosphate pyrophosphohydrolase activity, mediates exclusion of noncanonical purines from deoxy-ribonucleoside triphosphate pools; mutant is sensitive to the base analog 6-N-hydroxylaminopurine	
FK506-binding protein 1B	YNL135c_KOG0544	Peptidyl-prolyl cis-trans isomerase (PPIase), binds to the drugs FK506 and rapamycin; also binds to the nonhistone chromatin binding protein Hmo1p and may regulate its assembly or function	5****
ATP synthase subunit beta	YJR121w_KOG1350	Beta subunit of the F1 sector of mitochondrial F1F0 ATP synthase, which is a large, evolutionarily conserved enzyme complex required for ATP synthesis; phosphorylated	6***
Glucose-6-phosphate 1-dehydrogenase (G6PD)	YNL241c_KOG0563	Catalyzes the first step of the pentose phosphate pathway; involved in adapting to oxidatve stress	6****(4)
Protein transport protein SEC23	YPR181c_KOG1986	GTPase-activating protein, stimulates the GTPase activity of Sar1p; component of the Sec23p-Sec24p heterodimer of the COPII vesicle coat, involved in ER to Golgi transport	6
Ribose-phosphate pyrophosphokinase 3	YHL011c_KOG1448	5-phospho-ribosyl-1(alpha)-pyrophosphate synthetase, is required for nucleotide, histidine, and tryptophan biosynthesis; one of five related enzymes, which are active as heteromultimeric complexes	6
ATP synthase subunit alpha, mitochondrial	YBL099w_KOG1353	Alpha subunit of the F1 sector of mitochondrial F1F0 ATP synthase, which is a large, evolutionarily conserved enzyme complex required for ATP synthesis; phosphorylated	6**(7),***(7)
V-type proton ATPase subunit B	YBR127c_KOG1351	Subunit B of the eight-subunit V1 peripheral membrane domain of the vacuolar H+-ATPase (V-ATPase), an electrogenic proton pump found throughout the endomembrane system; contains nucleotide binding sites; also detected in the cytoplasm	6
Protein phosphatase PP2A regulatory subunit A	YAL016w_KOG0211	Regulatory subunit A of the heterotrimeric protein phosphatase 2A (PP2A), which also contains regulatory subunit Cdc55p and either catalytic subunit Pph21p or Pph22p; required for cell morphogenesis and transcription by RNA polymerase III	7

Glutamine-dependent NAD (+) synthetase	YHR074w_KOG2303	Essential for the formation of NAD(+) from nicotinic acid adenine dinucleotide	14****(12)
CTP synthase 2	YJR103w_KOG2387	Minor CTP synthase isozyme (see also URA7), catalyzes the ATP-dependent transfer of the amide nitrogen from glutamine to UTP, forming CTP, the final step in de novo biosynthesis of pyrimidines; involved in phospholipid biosynthesis	15

[a]Gene names and [c]Biological process are described as in the Gene Ontology Databse AmiGo (http://amigo.geneontology.org/cgi-bin/amigo/gp-details.cgi?gp=SGD:S000006380&session_id=8797amigo1327045088).
[b]*S. cerevisiae* Id no. indicates the *S. cerevisiae* identity numbers as found in CEGEMA.
[d]Asterisks show genes in which there was a difference in intron number within the four *Fusarium* species where one to four asterisks respectively indicates that *F. verticillioides*, *F. circinatum*, *F. oxysporum* or *F. graminearum* was the divergent species. The numbers in parentheses indicates the variant intron number in for the respective genes and species.