CHARACTERIZATION OF THE MITOCHONDRIAL GENOME OF *ENDOCONIDIOPHORA RESINIFERA* EXPLAINING THE DYNAMICS OF MOBILE ELEMENTS

by

Abdullah Zubaer

A Thesis submitted to the Faculty of Graduate Studies of

The University of Manitoba

in partial fulfillment of the requirements of the degree of

MASTER OF SCIENCE

Department of Microbiology

University of Manitoba

Winnipeg, MB R3T 2N2

Abstract

Fungal mitochondrial genomes are evolutionarily important for its history of endosymbiosis and its mobile genetic elements. It is a good reservoir of group I and group II introns and intron-encoded proteins (IEPs). Those catalytic introns (ribozymes) and their IEPs have application in biotechnology and have contribution in the mitochondrial genome evolution. The mobility mechanism of the group I and group II introns and their IEPs is complicated as mitochondrial genome is a playground of evolutionary forces. Previous research proposed different models to explain evolutionary mechanism of mobile introns and IEPs, but not all cases were understood by those models. Recent rise of genome sequencing facilitated a lot of mitochondrial genome sequencing, but not enough for fungal species. We have sequenced mitochondrial genomes of four strains of *Endoconidiophora resinifera* collected from Europe and North America. Those mitochondrial genomes are among the largest ones of the fungal kingdom having the largest *cox1* gene reported so far. The study revealed that the gene or genome size expansion is mainly influenced by the number of intron. An updated model of the mobility mechanism of introns and IEPs is also proposed. An intron landscape is constructed by collecting and aligning *nad5* genes from Ascomycota and Basidiomycota to find out the intron homing sites and intron distribution among the fungal species. Overall, the study explored new genomes, analyzed intron distribution, developed a model of intron mobility and generated intron and IEP related information that would be useful in future application in biotechnology.

Acknowledgement

I am deeply indebted to my supervisor Dr. Georg Hausner for his supervision, research direction, support and care, especially in difficult times. I am thankful to my committee members Dr. Deborah Court and Dr. Olivier Tremblay-Savard for their thoughtful suggestions, critical comments and encouragement for this study.

Gratitude goes to Dr. Michele Piercey-Normore for her suggestions in my research proposal and to Dr. Teresa de Kievit for her consultation regarding my graduate study. I have received a tremendous help in learning laboratory techniques from my lab-mates – Alvan Wai and Talal G. Abboud. I thank to my former colleagues Dr. Tuhin K. Guha and Dr. Iman Bilto, also to the current lab members Aamer Kurdi and Nikki Patel.

I would like to thank Stephanie Carter and Jo Davies for their help and support on official issues regarding my MSc program.

Table of Content

List of Tables

List of Figures

List of Abbreviations

ATP	Adenosine triphosphate
BLAST	Basic Local Alignment Search Tool
CDS	Coding Sequence
CTAB	Cetyl-trimethyl-ammonium bromide
EDTA	Ethylene-diamine-tetra-acetic acid
ExPASy	Expert Protein Analysis System
GTP	Guanosine triphosphate
HEG	Homing Endonuclease Gene
HMM	Hidden Markov Model
IEP	Intron-encoded protein
ITS	Internal Transcribed Spacer
JGI	Joint Genome Institute
MEA	Malt Extract Agar
MSA	Multiple Sequence Alignment
mtDNA	Mitochondrial DNA
NADH	Nicotinamide adenine dinucleotide + Hydrogen
NCBI	National Center for Biotechnology Information
NGS	Next Generation Sequencing
ORF	Open Reading Frame
rRNA	Ribosomal RNA
SDS	Sodium dodecyl sulfate
SNP	Single Nucleotide Polymorphism
SSU	Small subunit rRNA
tRNA	Transfer RNA
TM	Trans-membrane
WIN(M)	Winnipeg(Manitoba); Notation for the fungal culture collection at the University of Manitoba

General Introduction

Upon the slide, with best of light
And lenses “5” or “7”,
I see a sight as wondrous as
The Milky Way in heaven.
…...
How wondrous is a mighty sun,
That lights a boundless chasm!
More wondrous still I deem a speck
Of living protoplasm.

-Arthur Henry Reginald Buller
(http://www.mhs.mb.ca/docs/mb_history/47/poetscientist.shtml)

Blue stain fungi are economically important because they can cause sap wood to be stained due to melanin pigments. This “blue-stain” is undesirable for the timber industries as it limits export opportunities. Many species of *Ceratocystis sensu lato* have the ability of causing stain on wood. In addition there are a number of species of *Ceratocystis* that are plant pathogens. With regards to genomic explorations this genus so far has not been investigated in great detail. *Endoconidiophora* (= *Ceratocystis*) *resinifera* is one of the blue stain fungi that has been described fairly recently and most of the work so far has focused on morphologically features and some limited exploration of its rDNA region and mtDNA *rnl* segment.

Mitochondrial genome in fungi are of great interest as in some cases mtDNA defects have to been linked to hypovirulence and fungal mtDNAs contain mobile elements such as group I and group II introns. These mobile introns frequently encode intron-encoded open reading frames (ORFs) such as homing endonucleases or in the case of group II introns, reverse transcriptases. These elements appear to be quite invasive and thus can be of evolutionary importance with regards to fungal mtDNA evolution. In addition these elements are self-splicing and thus represent ribozymes. Ribozymes and intron encoded proteins have shown to have application in biotechnology such as being used as "RNA scissors" (ribozymes) and genome editing tools [homing endonucleases and group II introns (their engineered targetron counterparts)].

Because there are only limited number of mtDNAs for members of *Ceratocystis sensu lato* available and the potential of novel mobile elements within these mtDNAs the goal of this thesis was to study the mtDNA of strains for *E. resinifera*. The entailed next generation sequencing of the mtDNA and the annotation of the mitogenomes. The hope was that this work would also contribute towards gaining a better understanding of the evolutionary dynamics of the mtDNA introns, such as their survival mechanisms.

A brief compilation of the literature on mobile elements found in fungal mitochondrial genomes and the techniques used in sequencing, assembling and annotating fungal mitogenomics is provided in the Chapter 1 of this thesis. Chapter 2 contains my work on four strains of *Endoconidiophora resinifera*. A complete annotation of the mitochondrial genome of this fungus is presented along with a comparative analysis with mitogenomes of other *Ceratocystis* species.

Based on this analysis a new model of intron dynamics is presented that could explain why introns persist or potentially can interact with each other to enhance their survival. Chapter 3 presents a more detailed analysis of the NADH dehydrogenase subunit 5 (*nad5*) gene; here 186 fungal species have been examined with regards to their *nad5* genes, the sequences were collected and an intron map was created for the *nad5* gene. Intron landscapes can be useful tools with regards to characterizing introns and their encoded ORFs providing a resource for others with regards to locating introns and mtDNA annotation.

Overall this thesis will give the audience a better understanding of fungal mitochondrial genomes and its mobile elements. This work is the first to explore the mtDNA for the genus *Endoconidiophora* and also the first attempt to generate an intron landscape for the *nad5* gene.

Chapter 1: Literature Review

1.1 *Ceratocystis resinifera* – Taxonomic history

The organism investigated in this study is named *Ceratocystis resinifera* (recently renamed as *Endoconidiophora resinifera*) and it belongs to the order Microascales placed within the Phylum Ascomycota.

In 1890, Halsted proposed a genus named *Ceratocystis* to accommodate the organisms that cause black rot disease in sweet potato (Halsted, 1890). Although it was initially considered as a distinct genus there has been a long debate of its relation with the morphologically similar genus *Ophiostoma*, a debate which remained until 1993 when Hausner presented the large subunit ribosomal gene (LSU) data in a phylogenetic tree that showed that *Ceratocystis* and *Ophiostoma* are distinct (Hausner et al., 1993). Wingfield and coworkers coined the polyphyletic term ophiostomatoid fungi, to accommodate both species of *Ophiostoma* and *Ceratocystis* under one "general" umbrella (Wingfield et al., 1993). But more evidence by Visser (Visser et al., 1995) using the internal transcribed spacer (ITS) and Jones and Blackwell (Jones and Blackwell, 1998) using the small subunit ribosomal gene (SSU) data support the work by Hausner et al. (1993). Later, Witthuhn et al. (1998) studied the species complex of *C. coerulescens* using ITS rDNA sequences and based on the data it was separated into *C. douglasii*, *C. rufipenni* and three different forms of *C. coerulescens* (that were also examined by (Harrington and Wingfield, 1998) and these were referred to as *C. coerulescens* sp. A, sp. B and sp. C. These forms of *C. coerulescens* (A, B, C) were designated as *C. pinicola*, *C. coerulescens* and *C. resinifera*,

respectively. Harrington and Wingfield's paper (1998) was the first to recognize *C. resinifera* as a distinct species and their study elaborated on its morphology. Another extensive taxonomic study by de Beer (de Beer et al., 2014) based on molecular data redefined the genus *Ceratocystis* into various new genera and proposed a homotypic (i.e. synonym) name for *Ceratocystis resinifera*, which is *Endoconidiophora resinifera*.

1.2 Origin of mitochondria and its genome

Mitochondria are double-membraned organelles present in all eukaryotic cells providing a platform for rapid ATP-synthesis with a set of membrane bound proteins that form an electron transport chain and generate an ion gradient across the membrane that is used to synthesize ATP. Because of the presence of mitochondria (and evolutionarily-related organelles – hydrogenosomes and mitosomes) in every eukaryotic organism, it is considered to be an essential organelle and assumed to have evolved before the diversification of the eukaryotes about 2.7 billion years ago (Brocks, 1999). It is well accepted that the mitochondrion has an endobacterial origin, where the pre-eukaryotic cells acquired bacterial cells through a process called endosymbiosis. It has also been noted that the incidence of the endosymbiosis event that gave rise to the proto-mitochondria probably only happened once. It is now accepted that mitochondria originated from eubacteria (Lang, 2004). Genomics, proteomics and phylogenetic approaches helped in understanding the metabolic system of the proto-mitochondria and its host dependency provided strong evidence that these organelles originated from alpha-proteobacteria (Gabaldon and Huynen, 2003; Esser 2004). The transformation of a living cell to an organelle is assumed to be a complex process where the host cell takes over proto-mitochondrial authority of

the metabolic system and control of most protein synthesis (Gabaldon, 2007). This process involved massive gene transfers from the endosymbiont to the host genome. Currently the most popular theory suggests that the evolution of the eukaryotes involved the merger of an Archaea bacterium with a eubacteria, where the latter eventually evolved into the mitochondrial organelle (Martin and Koonin, 2006).

1.3 Mitochondrial genomes – general features

The endosymbiont ancestor essentially lost its independence due to the transfer of genes from the mitochondria to the nucleus, resulting in a huge reduction of mtDNA content. Typically fungal mitochondrial genomes express 15 protein coding genes, 2 rRNA genes, and a few (20 – 30) tRNAs. In most organisms, the protein coding genes are ATP synthase subunit genes (*atp6*, *atp8*, *atp9*), NADH dehydrogenase subunit genes (*nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5* and *nad6*), cytochrome oxidase subunit genes (*cox1*, *cox2*, *cox3*, *cob*), and ribosomal protein coding gene (*rps3*). Gene content for fungal mtDNAs can vary with some fungal groups lacking genes encoding components of the NADH complex or ATP synthase (reviewed in Hausner, 2003). The small ribosomal subunit gene (*rns*) and large ribosomal subunit gene (*rnl*) are present along with the tRNA genes in a mitochondrial genome (Abboud et al., 2018). Among the genes the *atp9* gene is considered as optional as it is not consistent in its presence in some fungal mitogenomes. In a few cases the *atp8* gene was also noted to be absent. For example, *Shiraia bambusicola* and *Stemphylium lycopersici* do not have the *atp8* gene in their mtDNA (Franco et al., 2017). The ribosomal protein coding gene (*rps3*) in many filamentous ascomycetous fungi is encoded in a group I intron in the *rnl* gene (reviewed in Hausner, 2012), but there are a few exceptions

(Sethuraman et al., 2009). Fungal mitochondrial genomes have also been shown to accumulate segments of plasmids encoding RNA and/or DNA polymerases (reviewed in Hausner, 2003).

1.4 Diversity of mitochondrial genome architectures among the fungi

With a roughly a constant number of genes, the mitogenome in all fungal species could be expected to be the same. But in reality the mitogenome size is highly variable among the fungi. It was reported that the smallest fungal mitochondrial genome size is 11 kb in *Hanseniaspora uvarum* (Pramateftaki et al., 2006) and largest one is 235 kb in *Rhizoctonia solani* (Losada et al., 2014). This size variation is due to various factors such as:

- Mobile elements (plasmids, mobile introns and homing endonucleases)
- Intergenic spacers
- Gene duplication and repeats

The protein coding genes and the rRNA genes contain introns and the number of introns can be variable for a gene throughout a particular species. For example, the *cytb* gene (cytochrome b; also known as the *cob* gene) from 129 fungal species showed different intron arrangements. Some of the intron insertion sites in *cytb* are abundantly populated with introns, some others sites harbor very few introns (Guha et al., 2017). This would suggest that some intron insertion sites and their associated introns are conserved and others sites are under selection pressure (possible purifying selection) to avoid intron accumulation. Genome size is also influenced by the presence of intergenic spacers and uncharacterized open reading frames (Joardar et al., 2012). Another study showed the occurrence of repetitive DNA could also influence genome sizes (Li et al., 2015). Mitogenomes in fungi are diverse not only due to genome size variation but also due

to highly variable gene orders. Gene order variability is assumed to be a product of gene rearrangement through recombination events that might be promoted by repeat accumulation or due to the activity of mobile elements (Aguileta et al., 2014). Thus mitogenomic variation can be directly or indirectly connected with the mobile elements present within the mtDNAs. Mobile elements in fungal mtDNA typically are mobile introns and homing endonuclease genes (HEGs). Mobility of these introns is promoted by the proteins they encode, in addition these introns are autocatalytic at the RNA level that permits them to self-splice (i.e. ribozymes). However, it has been shown that for splicing under *in vivo* conditions these element will recruit both intron encoded and host factors to facilitate the formation of splicing competent RNA folds. These elements will be discussed separately.

1.5 Mitochondrial genome evolution

Mitochondrial genome evolution is in part "driven" by its origin as an endosymbiont and the transfer of genetic information to the nucleus. Within the fungi and other eukaryotes (but not most metazoans) organellar genomes tend to contain mobile elements such as group I and group II introns. These mobile introns can encode ORFs that express enzymes that probably help them to survive within the mitochondrial genome. The origin of these types of introns is probably linked to the eubacterial ancestor that gave rise to the mitochondria; however in some eukaryotic lineages for some unknown reasons these group I and group II introns were lost (see Hausner, 2012). There are other factors that drive the evolution of the mitochondrial genome, such as – drift (by accumulation of mutation), intra and inter molecular recombination (among mtDNAs of

a mitochondrion), transfer of genetic information between the organelles and the nucleus, and horizontal gene transfer (Korovesi et al., 2018).

Previous studies showed that mitochondrial genome variation is in part driven by the mobile introns and HEGs (encoding homing endonucleases) (Wu and Hao, 2014; Repar and Warnecke, 2017); there are some models with regards to the maintenance of these types of elements. The Goddard and Burt model usually referred to as the HEG lifecycle is one of the most insightful models in explaining the mobility of the mobile intron and intron-encoded ORFs. This model was based on the mobility of the omega-intron (group I intron located within the *rnl* gene) in various yeast species where they found three scenarios: (1) the presence of a complete intron with HEG, (2) intron with degenerated/eroded HEG (i.e. premature stop codons), and (3) the absence of the intron at the same spot (intron insertion site). Based on these observations they designed a model for the homing life cycle where they proposed that an intron encoding a functional HEG invades an "intron-homing-site" and populates that site in all available mtDNAs within the same individual, species and possible other species via horizontal transfer. However, as these elements are neutral they cannot be selected for and thus can accumulate mutations through the process of random drift. This inevitably results in the loss of the HEG. Once the HEG has been eroded it will limit the potential for the intron to be mobilized into new locations. The intron (without an ORF) can persist but any mutation that would make splicing less efficient would cause the intron to be rapidly lost as such mutations would be deleterious and negative selection would promote its rapid loss. However, the loss of the intron would re-establish a homing site that could be invaded again by a functional composite intron (intron plus HEG). The availability of sites for intron invasion is an important consideration for these elements to persist,

they essentially have to outpace the rate of neutral mutations (drift) in order to achieve a net gain within population. Introns and HEGs can escape this cycle by invading new (ectopic) sites, or by gaining new functions that benefit the host (Goddard and Burt, 1999; Guha et al., 2017). Data has shown that group I and their HEG partners can evolve independently from each other or they can become more integrated into each other by incorporating sequences (referred to as core creep) of the core intron (i.e. exonization) (Edgell et al., 2011; Guha et al., 2017).

1.6 Catalytic introns/ ribozymes/ mobile elements/ selfish elements

Discovery of ribozymes was a breakthrough by Sidney Altman and Thomas R. Cech that was awarded a Nobel Prize in 1989. The characterization of a group I intron from a protozaon yielded insights in the RNA-based splicing mechanisms of group I introns (Cech, 1990). Dujon in the late 1970s noted the omega intron of yeast and eventually determined that group I introns could be mobilized to cognate intron-less alleles by means of the homing endonuclease (Dujon, 1980). Mobile introns in fungi can be categorized into two groups: group I intron and group II intron (Michel et al., 1990). Some basic features of group I and II introns are presented in Table 1.1. Group II introns tend to encode reverse transcriptases and have a splicing mechanism that resembles that of nuclear spliceosomal introns (Zimmerly and Lambowitz, 2011). Many fungal mitochondrial genomes are rich in group I introns that appear to be self-splicing introns. In general these introns are considered to be selfish elements because of their propagation in the genome having no impact on the host genome's fitness (i.e. these are neutral elements). However, some evidence has been found that introns play important role in the biology of the organism such as modulating gene expression (Rudan et al., 2018). For example, a group II

intron in *Cryphonectria parasitica* (Chestnut blight fungus) contributes to the organism being hypovirulent (Baidyaroy et al., 2000; Baidyaroy et al., 2011), which allows the host tree (chestnut) to thrive, but it also ensures that the fungus does not over-exploit its host which would lead to the extinction of both the host and the "parasitic" fungus.

Table 1.1: A snapshot of the features of group I and group II introns.

Features	Group I intron	Group II intron
Distribution	In all three domains of life, if located within nuclear genomes restricted to rDNA	Not present in nuclear genomes and bacteriophages
Structure	10 “domains” (P1 to P10) in structure	6 domains (D1 to D6)
Splicing	Nucleophilic reaction with external guanosine (G)	Two step transesterification reaction for splicing with an internal adenosine (A) in DVI initiating splicing
Intron encoded protein (IEP)	Homing endonuclease (and or maturase)	Reverse transcriptase / in some instance homing endonuclease (Toor and Zimmerly, 2002)
Mobility	Intron homing by homologous recombination (mediated by double-strand break DNA repair)	Retro-homing into conserved target sites via an RNA intermediate followed by reverse transcription

1.6.1 Group I introns

Group I introns, although first discovered in a protozoan, are frequently encountered in fungal mitochondrial DNA. They are also found in plant and algal organellar genomes (mitochondria and plastid), bacteria, bacteriophages, but rarely in lower animals (sponges and sea corals) (Haugen et al., 2005, Lang et al., 2007). Group I introns can also be found in nuclear rRNA genes in some fungi and protozoans. It had been reported that group I introns are not found in the archaea (Belfort et al., 2002) but recent analysis suggests that group I introns are indeed present in all three domains of life (Nawrocki et al. 2018).

At the DNA level, intron sequences are not conserved among the fungal species. But the complex tertiary RNA structures (after self-splicing) are very similar for all the group I introns (Michel et al. 1990). The RNA secondary structure of group I introns can be described by its 10 characteristic helical (or paired regions) domains (referred to as P1 – P10; see Figure 1.1). In group I introns, an intron-encoded ORF usually is inserted within a loop region of the intron, but it can also be inserted into the intron's core sequence and be fused to the upstream exon (Sethuraman et al., 2008; Edgell et al. 2011). Although, superficially, the secondary structures of group I introns are very similar, they can be categorized into five different classes and further subdivided into subclasses (1, 2 etc.) on the basis of minor variations in intron core sequences and peripheral structure such as additional helical regions or absence of some helical components (Michel et al., 1990; Zhou et al, 2008). The group I classes are A to E, and subclasses are usually designated as 1 to 3, so that group I intron can be classified from group IA1 to group IE3 (Lang et al., 2007).

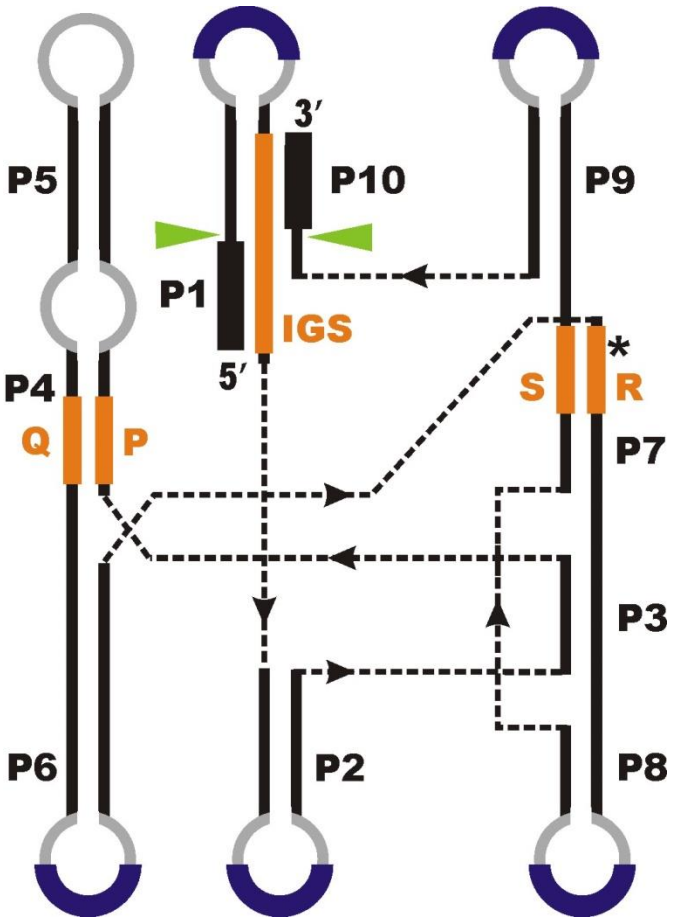

3′
P5
P10
P9
P1
IGS
5′
*
P4
S
R
Q
P
P7
P3
P6
P2
P8

Figure 1.1: Generalized RNA fold for a group I intron, all loop regions (in blue or grey) can harbour ORFs encoding homing endonucleases. (Adapted from Hausner et al., 2014; Mobile DNA 5:8 – open access). IGS = internal guide sequence, P, Q, R and S refers to conserved sequence elements found in some group I introns (see Michel et al. 1990). P1 to P10 refers to "pair regions" (or helices) that contribute toward to the three dimensional structure of group I introns.

The splicing of group I introns is autocatalytic – after the precursor RNA formation, it splices itself out mediated by the formation of RNA tertiary structure. The splicing requires an external GTP that can associate with the intron RNA and provide a nucleophile. The splicing mechanism can be modulated by intron-encoded homing endonucleases (that act as maturases) and nuclear encoded RNA chaperones (Halls et al., 2007). The overall splicing contains two transesterification reactions, the first of which is mediated by the external Guanosine (alpha-G) bound to the GTP-binding pocket in the P7 region; the alpha-G uses its 3'OH group to attack the upstream intron/exon junction. A second transesterification reaction happens by the free upstream exon's 3'OH group attacking the downstream intron/exon junction; thus freeing the intron and joining the exons together. These reactions are mediated by the RNA fold that promotes the formation of the internal RNA guide sequence (part of P1) and the formation of the P10 interaction (usually the downstream exon base pairing with an unpaired region within P1), which combined bring all the components that have to interact into close proximity with each other. The intron RNA undergoes a series of fragmentation reactions, which will remove the bound alpha-G (the original external Guanosine (G)) and therefore prevents the reverse splicing of the intron RNA into the matured RNA (Michel et al., 1989, 1990; Cech, 1990; Cech et al., 1994; Hausner et al., 2014). It is noteworthy that the primary sequence of a group I intron ends with a G, and usually the upstream exon ends with a Thymidine (T) (Michel et al., 1989).

1.6.2 Group II introns

Group II introns are less frequently noted compared to group I introns among fungal mtDNAs. They sometimes encode reverse transcriptase and are considered to be retro-elements.

They are also assumed to be the originator of spliceosomal introns and retrotransposons in the nuclear genomes of eukaryotes. Group II introns are found in bacteria and organellar genomes of protists, plants, fungi and less frequently in archaea and lower animals (Ferat and Michel, 1993). However, group II introns are never found in a nuclear genome and so far they have not been observed in bacteriophages (Lambowitz and Zimmerly, 2011).

Group II introns at the RNA level maintain a conserved RNA structure possessing six conserved domains designated as DI to DVI. The intron-encoded ORF (if present) typically resides in the DIV (Domain IV) (Michel et al., 1990; Toor et al., 2001), or sometimes in DII (Domain II) (Simon et al., 2008; Hafez and Hausner, 2011) or DIII (Toor and Zimmerly, 2002). Different domains of group II introns have different functions in folding – such as Domain I acts as the scaffold domain that initiates the RNA folding process that results in a "six fingered hand". Short intron sequences referred to as exon binding sequences (EBS) can interact (base pair) with short sequences in the flanking exons (intron binding sequences or IBS). These interactions stabilize the RNA fold and set up a splicing competent structure (i.e. the ribozyme).

Overall the splicing mechanism involves a two-step transesterification reaction. First, an internal adenosine's 3' OH group (A 3'OH) (= branch point) located within an unpaired region of the DVI domain reacts with the upstream intron/exon junction; second the liberated upstream exon's 3'OH attacks the downstream intron/exon junction. This liberates the intron RNA and ligate the exons together. The intron RNA is released in a lariat configuration as the internal branch point forms a 2'5' linkage with the first nucleotide of the group II intron (Michel and

Ferat, 1995; Daniels et al., 1996; Toor et al., 2008, 2010). Table 1.1 provides an overview with regards to features of group I and group II introns.

1.7 Intron-encoded ORFs

The mobile introns can have ORFs embedded in their sequences. Usually, group I introns harbour homing endonuclease genes (HEGs) and group II contain reverse transcriptase genes (RTs). However, group II introns also can also encode HEGs instead of RTs. Those intron-encoded genes are presumed to get translated after intron splicing and thus give rise to the intron-encoded proteins (IEPs). IEPs can be categorized according to their functions:

1. Homing endonuclease (some can also act as maturase that promote RNA folding)
2. Reverse transcriptase (multifunctional as they can nick DNA and promote RNA folding)

Like the mobile introns, some HEGs encoding IEPs are also mobile; that means HEGs can move independently of their host introns. Reverse transcriptases encoded by group II introns appear to have a closer association with their host introns and data suggests that RTs co-evolve with their introns (Toor et al., 2001).

Homing endonucleases are a type of endonuclease (DNase) enzyme that can cut DNA generating staggered ends. Homing endonucleases can be assigned into several different protein families (see Hafez and Hausner, 2012) but with regards to fungal mtDNAs they belong to two families defined by short amino acids motifs: LAGLIDADG and GIY YIG homing endonucleases. These enzymes are site specific and cut at or near their DNA binding sites, which

can be extensive ranging from 14 – 40 bp nucleotide sequence (depending on the particular enzyme) (Stoddard, 2005). This specificity ensures that these enzymes guide their host introns to insert into cognate alleles that lack introns. This specificity has also made these enzymes popular with regards to developing genome editing tools that require site specific cuts (Stoddard, 2005, 2014; Hafez and Hausner, 2012; Guha et al., 2017).

Group II encoded RTs are multi-domain/multi-functional (maturase, DNA binding, endonuclease and reverse transcriptase) enzymes that help their host introns to splice out from the transcript and promote intron mobility (retrohoming of the intron).

1.8 Mechanism of intron and HEGs mobility

The term "homing" is used in combination with intron-encoded endonucleases as these enzymes catalyze of the mobility of introns from an intron-plus allele into a cognate intron-minus allele, i.e. "intron" is going "home" (i.e. intron homing). A particular spot – sequence – or position in a gene where an intron is located is considered as the "home" of that intron. Studies of intron landscape in a gene have revealed intron locations are quite conserved among species (Ferandon et al., 2010; Hafez et al., 2013; Guha et al., 2017). Sometimes unrelated eukaryotes can share introns at some positions within mtDNA genes. This suggests that intron insertion sites are usually conserved segments within genes and they allow introns to move within species and horizontally to unrelated species. The mechanisms of horizontal gene transfers are still unknown (Hausner, 2012).

The homing endonuclease is the initiator of intron-homing or intron-mobility. It can recognize a very specific sequences near the homing-site and it can generate a double-stranded break in the DNA which eventually triggers the DNA repair system of the cell. First after the DNA has been cut by the HE, cellular exonucleases will remove additional nucleotides generating the template for DNA repair. The repair involves homologous recombination that requires a homologous template (intron-plus allele) to repair the "broken" intron-minus allele. So if there is a homologous template with a intron+ or HEG+ allele available then the intron (with HEG) will be copied to the new location by the repair mechanism (Belfort and Perlman, 1995; Chevalier and Stoddard, 2001); i.e. a non-reciprocal transfer of the intron sequence into the gap generated by the HE activity. The repair process usually leads to co-conversion of flanking markers (Belfort et al., 2002).

Retro-homing is mediated by the enzyme reverse transcriptase. This mechanism is also similar to the concept of homing, but the mechanism is different as it is mediated by the reverse transcriptase enzyme. Reverse transcriptase is a multifunctional enzyme having maturase (X), DNA binding (D), endonuclease (En) and reverse transcriptase (RT) domains in its structure. Retro-homing starts after the splicing of an intron with the aid of maturase activity of the reverse transcriptase enzyme. The intron lariat and the RT protein combine to form a ribonucleoprotein. Further the RT enzyme makes a nick in the sense strand of the retro-homing site in the DNA via the 3'OH end of the introns lariat. The RT uses its endonuclease domain to nick the antisense strand. Now the intron RNA can be reverse spliced into the sense strand of the DNA double-stranded gap. The RT can use the 3' end of the antisense strand as a primer to generate a cDNA using the intron RNA as a template. This generates an intermediate where a DNA-RNA hybrid is

now located at the homing site. That original intron RNA is replaced with DNA by the host cell's DNA repair system (Lambowitz and Zimmerly 2004; Edgell et al. 2011; Hafez and Hausner, 2015). So in both cases, homing and retro-homing, the host cell's DNA repair system is required. In both systems there are requirements for factors that promote RNA folding (maturases, splicing factors that have RNA chaperone functions etc.; divalent cations such as Mg^{+2}).

1.9 Importance and application of introns and HEGs

Mobile introns in mitochondrial DNA are sometimes referred to as selfish elements or molecular parasites that are not beneficial for the organism. But they are found to be influential on some occasions – such as in the case of hypovirulence in *Cryphonectria parasitica* where they can reduce the organism's virulence (Baidyaroy et al., 2000; Baidyaroy et al., 2011). Fungi can sometimes contact each other and form transient hyphal fusions permitting the exchange of cytoplasm along with organelles. This allows mitochondria with introns to enter individuals that lacked the intron. Mitochondria can fuse and thus introns can mobilize into empty sites within the recipient genome. This allows for the spread of neutral, beneficial and also deleterious introns.

In *Podospora anserina*, group II introns have been associated with fungal senescence (Begel et al., 1999). Recent studies have shown that mtDNA introns may not be as neutral as suggested in the literature. In yeast, mitochondrial introns are essential in fine tuning gene regulation; they appear to prevent the overexpression of certain genes in the mitochondria (Rudan et al., 2018).

The IEPs such as homing endonucleases are responsible for intron mobility and otherwise no biological significance has been noted for them. But they are very important for their applications in biotechnology. Homing endonucleases are great tools in genome editing and have been used in genome engineering (reviewed in Guha et al., 2017; Kleinstiver et al., 2012). Homing endonucleases can be engineered to recognize target sites and generate specific DNA cleavage that can subsequently be edited by the host repair mechanism by using homologous recombination in the presence of a "repair template". Homing endonucleases can be applied towards gene replacement strategies required for gene therapy or gene drive of certain alleles through populations (reviewed in Hafez and Hausner, 2012) but homing endonucleases can also be used as simple tools to generate site-specific DNA mutations. In the absence of DNA repair templates, double-stranded breaks in eukaryotes tend to be repaired by the non-homologous end joining repair systems which are error prone (Hafez and Hausner, 2012; Guha et al., 2017).

1.10 Fungal Mitochondrial Genomes: Recent advancements

It is now widely accepted that mitochondria originated from alpha-proteobacteria through endosymbiosis and an extensive reduction in genome size has taken place as the nuclear genome has taken control of proto-mitochondrial protein synthesis and metabolism. Studies showed the transfer of genes from mitochondria to the nucleus (Martin, 2003; Brandvain and Wade, 2009). This transfer of genes is part of the endosymbiont theory that posits that as the mitochondrial ancestor transitioned from being a facultative to an obligate endosymbiont and eventually reached organelle status the DNA content of the mitochondria "shrank" to a minimum set of core genes.

Mitochondrial sequences have been used for species identification and genealogy (Girish et al., 2004, 2005; Parson et al., 2000; Blouin, 2002; Ward et al., 2005). Initially, sequencing of mtDNAs was somewhat challenging requiring mtDNA isolation, cloning of mtDNA fragments and Sanger sequencing. This approach is effective for small metazoan type mtDNAs such as human mitochondrial DNA (reviewed in Smith, 2016). Recently, with Next Generation Sequencing (NGS) approaches, genomics became easier as whole DNA can be used to prepare sequencing libraries. So, a lot of mitochondrial genomes have been sequenced from different organisms. Most of the mitochondrial genomes presented in public databases are from Metazoans. These comprise 92% of all mitochondrial genomes sequenced until 2015. In comparison only few members of the Phylum Mycota have been sequenced so far (Smith, 2016). For fungi, mtDNAs for members of the Saccharomycetales are now quite common. Moreover, economically important members of the Ascomycota have been explored. There are also a few examples found that represent the Basidiomycota, Zygomycota and Chytridiomycota in NCBI genome database.

Fungal mtDNA sequences have confirmed earlier data based on restriction enzyme analysis or limited sequence analysis that fungal mtDNAs vary in size and gene order, and most of the studies found those variations are related to the mobile introns (group I and group II introns) and intron-encoded ORFs (Freel et al. 2015).

The advancement in fungal mitogenome research is mainly technological with regards to NGS strategies and new computational tools have been invented to assemble and annotate fungal mitochondrial genomes more easily and efficiently (see Table 1.2). Fungal mtDNAs are

challenging to annotate due the large numbers of introns and their AT-richness. Fungal mtDNAs are similar to repetitive DNA due to their AT-richness and are sometimes difficult to assemble as some programs by default tend to filter out repetitive DNA and short mtDNA reads get lost among scaffolds containing nuclear AT-rich repetitive DNAs. Table 1.2 provides an overview of various technologies (TECH) and bioinformatics tools that have been applied in sequencing fungal mtDNAs.

Table 1.2: Strategies (Tech) presented in recent mitogenome projects.

Organism	Genome size (kbp)	Sequencing Tech	Assembly Tech	Annotation Tech	Reference
Candida subhashii	29.795	1. Restriction enzyme (*Bam*HI and *Hin*dIII) and Cloning and primer-walking 2. Macrogen (http://www.macrogen.com/).	1. Vector NTI Advance v. 10.1.1 (Invitrogen) 2. Geneious v. 4.8.5 (Biomatters)	1. BLAST 2. TestCode 3. tRNAscan-SE	Fricova et al., 2010
***Jaminaea angkorensis* (Basidiomycetous yeast)**	29.999	Illumina HiSeq2000	Velvet	1. MFannot 2. Geneious v5.6.6	Hegedusova et al., 2013
Agaricus bisporus	135.005	JGI	JGI	1. BLAST 2. RNAweasel 3. tRNAscan-SE 4. ClustalW	Ferandon et al., 2013
Phlebia radiata	156.348	454 (GS FLX Titanium)	Newbler (Roche, 454 Life Sciences)	1. Artemis 2. tRNAscan-SE 3. RNAweasel	Salavirta et al., 2014
Madurella mycetomatis	45.590	454 (GS junior titanium)	GS de novo assembler of Roche	1. CLC sequence viewer version 6.5.1 2. tRNAscan-SE 3. ARAGORN 4. ARWEN 5. RNAweasel	van de Sande, 2012
***Aspergillus* spp.**	27.0-33.0	1. 454 GS FLX Titanium instrument (Roche) 2. Illumina Genome Analyzer II	Celera Assembler at JCVI	1. Artemis 2. BLAST 3. tRNAscan-	Joardar et al., 2012

				SE	
Aspergillus *terreus*	24.658	Sanger	Celera Assembler at JCVI	1. Artemis 2. BLAST 3. tRNAscan-SE	Joardar et al., 2012
Stemphylium lycopersici	75.911	Illumina Hiseq 2000 platform	Geneious 9.1.2 *de novo* assembler	1. MFannot and MSA 2. tRNAscan-SE 3. BLAST2GO 4. OGDRAW 5. Vmatch	Franco et al., 2017
Ophiocordyceps sinensis	157.510	PacBio RS II sequencing platform	1. BLASR 2. Celera Assembler 3. Quiver 4. SAMTools	1. BLAST 2. Clustal W 3. tRNAscan 4. ARAGORN	Li et al., 2015
Cordyceps militaris	33.277	Illumina	Newbler	(not specified)	Sung, 2015
Hypomyces aurantius	71.638	Illumina HiSeq 2500	Velvet 1.2.03	1. Mfannot v 1.33 2. tRNAscan-SE 1.31 3. BLASTX 4. CG view	Deng et al., 2016
***Rhynchosporium* spp.**	49.539 – 68.904	Roche/454 GS FLX	Newbler (gsAssembler – GUI)	1. BLAST 2. Geneious v5.5 3. RNAweasel 4. tRNAscan-SE v.1.21	Torriani et al., 2014
Engyodontium album	28.081	Illumina HiSeq2000	1. SPAdes v 3.6.1 2. Bandage 0.7.1	1. ORF-finder 2. tRNAscan-	Yuan et al., 2017

				SE v1.21 3. BLAST	
Phialocephala scopiformis	36.919	Illumina HiSeq	Geneious v8.0.5	1. MITOS web server 2. tRNAscan-SE v1.21 3. ARWEN v1.2 4. Clustal Omega 5. BLAST 6. MFannot	Robicheau et al., 2017
Colletotrichum lindemuthianum	37.446	Illumina HISeq2500 platform with paired-end reads	CLC Genomics Workbench 6.5.1	1. MITOS Web Server 2. MFannot 3. BLAST 4. ARWEN 5. GenomeVx	de Queiroz et al., 2018
Ophiostoma novo-ulmi	65.095	Illumina MiSeq	A5-miseq pipeline	1. MFannot 2. RNAweasel 3. BLAST 4. MAFFT 5. Artemis	Abboud et al., 2018

These days genome sequencing is based on next generation sequencing (NGS) technologies, provided by Illumina, Roche 454, and PacBio SMRT (Buermans and den Dunnen, 2014). With regards to sequence assemblies there are a lot of freeware available along with commercial programs. However, the problem is to find suitable programs for a particular genome. From the literature it appears that Velvet (Zerbino and Birney, 2008), Newbler (Margulies et al., 2005), SPAdes (Bankevich et al., 2012), a5 pipeline (Tritt et al., 2012) are the popular free programs that been used by many researchers in this discipline. In addition, there have been refinements to deal with small repetitive components within whole genomes, such as mtDNAs. For examples, MIRA (including MIRA-bait) (Chevreux et al., 1999) and MITObim (Hahn et al., 2013) programs are somewhat optimized for small genomes like mitochondrial genomes.

With regards to organellar DNAs such as fungal mtDNA, most of the thorough annotations have been with MFannot and RNAweasel (Lang et al., 2007) in addition to the BLAST suite of programs (Altschul et al., 1990). Mfannot is built specifically to find out the mitochondrial genes and introns, whereas BLAST is a general but a very powerful tool available to search for matching strings. The use of MFannot and RNAweasel helps to find out group I and group II introns and the intron-exon border, whereas other general gene finding programs (developed for nuclear or prokaryotic genomes) are not that specific for recognizing mobile introns so these fail to find the correct gene architectures. Although MFannot is good for mtDNA gene finding, it can sometimes fail in recognizing the correct intron-exon boundaries. This is because of the complicated structure of mitochondrial genes as they can have very small exons. For example in the placozoan *Trichoplax adhaerens*, there is an exon in the *cox1* gene that

consists of only one base pair, which is the smallest exon possible (Osigus et al., 2017). In addition, most programs are not efficient in detecting twintrons (Hafez and Hausner 2015). Twintrons are composed of multiple introns that can be either side-by- side or nested within a resident intron. Some of the studies used the MITOS program (Bernt et al., 2013) for mitogenome annotation, but this program is also not very efficient in defining intron-exon borders. In all cases manual inspection is still inevitable for mitochondrial genome annotation. Here some basic tools such as the multiple sequence alignment program MAFFT (Katoh and Standley, 2013) combined with blast can resolve issues. With regards to MAFFT, the E-INS-I algorithm allows for alignment of sequences with "large gaps" (i.e. introns); this allows for confirming intron/exon junctions.

Mitochondrial genomes that have been sequenced with paired-end sequencing techniques usually reveal the presence of repeat sequences in a genome. To identify these repeat regions certain bioinformatics tool can be applied such as Tandem Repeats Finder (TRF) (Benson, 1999), REPuter (Kurtz et al., 2001), MIcroSAtellite (MISA) identification tool (Beier et al., 2017) etc. Additional tools that allow for genome comparisons are programs such as OrthoMCL (Li et al., 2003) or Mauve program (Darling et al., 2010); these allow for determining changes in gene order and detecting genome rearrangements, deletions and insertions.

1.11 Challenges in studying mitochondrial genomes

The main challenge in fungal mitochondrial DNA analysis is annotation; difficulties arise due to the large number of introns, the variability of intron sizes and numbers. This variability in

genes has made fungal mitochondrial *cox1* gene incompatible for use as a DNA barcode; *cox1* has been applied as DNA barcode in other eukaryotes such as the metazoans (Rodrigues et al., 2017; Robba et al., 2006).

Mitochondrial genome annotation is currently done with the combination of the BLAST program and MFannot. In some cases, MITOS web server has also been used. MFannot and RNAweasel work differently than the other gene finder programs (such as Augustus; Stanke and Morgenstern, 2005). MFannot tries to define the introns compared to standard gene finder programs. However, improvements are needed that allow for the detection of small exons along with complex intron arrangements, such as nested introns or twintrons. In additions introns can harbour complex ORFs, fusion ORFs and ORFs that are fused to the upstream exons. These add complexity with regards to defining exons.

As costs for other methodologies make them more affordable, the incorporation of data from RNAseq (RNA sequencing) and proteomics experiments would allow for better predictions of exons and whether certain proteins are actually expressed. It should also be noted that in some cases fungal mtDNA genes are known to use non-standard start codons (i.e. not AUG): these are serviced by special tRNAs that can deliver Methionine (M) but recognize codons usually reserved for Valine (V) or Leucine (L).

The major objectives of this work are (1) to characterize the mitochondrial genomes for selected members of the genus *Ceratocystis*; and (2) determine the intron landscape for the mtDNA *nad5* gene in order to gain some insights into the dynamics of intron evolution. The

hypothesis of this work is that introns are the major source of polymorphisms observed among fungal genomes.

==========

[**Note:** Chapter 2 has been submitted to Scientific Reports as: Zubaer A., Wai A., and Hausner G. The mitogenome of *Ceratocystis resinifera*: A tale of many introns. Alvan Wai contributed towards the characterization of mitochondrial introns and Georg Hausner contributed towards the experimental design and editing of the manuscript. I performed all the experimental work including the mitogenome annotation and analysis and assembled the first draft of this manuscript.]

Chapter 2: The complete mitochondrial genome of *Endoconidiophora resinifera*: a tale of many introns

Abstract

Endoconidiophora resinifera (= *Ceratocystis resinifera*) is a blue-stain fungus that occurs on conifers. Four strains (WIN(M)79, WIN(M)1409A, WIN(M)1410B and WIN(M)1411) of this species were cultivated and total DNA was extracted and the samples were subjected to next generation sequencing utilizing the MiSeq (Illumina) platform. After scaffold were assembled large contigs could be recognized that yielded long continuous sequences with similarity to fungal mitochondrial genomes. The data showed that the *Endoconidiophora resinifera* mitogenome is one of the largest mitochondrial genomes (>220 kb) so far reported among members of the Ascomycota. An exceptionally large number of introns (81) were noted and differences among the four strains were restricted to minor variations in intron numbers and a few indels and single nucleotide polymorphisms. The major differences among the four strains examined are due to size polymorphisms generated by the absence or presence of mitochondrial introns. Also, these mitogenomes encode the largest cytochrome oxidase subunit 1 gene (47.5 kb) reported so far among the fungi. The large size for this gene again can be attributed to the large number of intron insertions. This study reports the first mitogenome for the genus *Endoconidiophora*, a species previously assigned to the genus *Ceratocystis*. The latter genus has recently undergone extensive taxonomic revisions and the mitogenome might provide loci that could be applied as molecular markers assisting in the identification of taxa within this group of

economically important fungi. The large mitogenome also may provide some insight on mechanisms that can lead to mitochondrial genome expansion.

2.1 Introduction

Endoconidiophora resinifera (= *Ceratocystis resinifera*) is a fungus that belongs to the Ceratocystidaceae (Sordariomycetes, Microascales). It is associated with blue stain on lumber and sapwood, which is an undesirable character for exporting timber related products. Some members of the genus *Ceratocystis sensu lato* (recently subdivided into several new genera including *Endoconidiophora*; de Beer et al., 2014) are known for causing infections such as black rot disease in sweet potato (*Ceratocystis fimbriata*; Halsted and Fairchild, 1891), oak wilt (*Ceratocystis fagacearum*; Juzwik et al., 2008), wilt in cacao plant (*Ceratocystis cacaofunesta*; Engelbrecht et al., 2007), canker stain of plane trees (*Ceratocystis platani*; Tsopelas et al., 2017), and sapstreak in maple tree (*Ceratocystis virescens*; Houston, 1993). However, *E. resinifera*, so far has not been associated with any pathogenicity, but this insect-vectored fungus can colonize bark beetle galleries and wounds in species of *Picea* (Harrington and Wingfield, 1998). Species of *Ceratocystis s.l.* have been studied with regards to their taxonomy, blue-staining ability, and pathology (Wingfield et al., 1993), but with regards to genetic or genomic investigations, no thorough study has been done so far except for *Ceratocystis cacaofunesta* (Ambrosio et al., 2013). Previous studies on this group of fungi with regards to mitochondrial DNA focused on the rRNA genes and these displayed a large variety of intron insertions among various *Ceratocystis s.l.* species (Hafez et al., 2013; Sethuraman et al., 2013). Additional mitochondrial genomes have recently been sequenced for members of *Ceratocystis*, but so far a detailed annotation is only

available for the mitogenome of *C. cacaofunesta* (Ambrosio et al., 2013; Wilken et al., 2013; van der Nest et al., 2014; Wingfield et al., 2016a, 2016b).

Fungal mitochondrial genomes usually encode genes involved in translation [small and large ribosomal subunit RNAs (*rns* and *rnl*) and tRNAs], proteins involved in the respiratory chain [subunits for Complex III and Complex IV (*cob*, *cox1*, *cox2*, and *cox3*)], subunits of NADH dehydrogenase (*nad1* to *nad6* and *nad4L*; except for members of the Taphrinomycota and some members of the Saccharomycetales), plus some of the components of the ATP synthase (*atp6*, *atp8*, and usually *atp9*), and in some instances the ribosomal protein RPS3 (Hausner, 2003; Freel et al., 2015). Mitochondrial genome sizes among the fungi are quite variable ranging from 18.9 kb (in *Schizosaccharomyces pombe*; Anziano et al., 1983) to 235 kb (in *Rhizoctonia solani*; Losada et al., 2014). Mitogenome size variation has also been reported among closely related species (Joardar et al., 2012). The size variations are mainly due to the number and sizes of intron insertions and size of intergenic spacers. Gene order, repeats, and in some instances other types of elements such as plasmid insertions are additional sources that generate variability among fungal mitogenomes (Aguileta et al., 2014).

Fungal mitochondrial introns, based on structure and splicing mechanisms, can be assigned to either group I or group II introns. These elements are potential ribozymes that can in part catalyze their own removal from transcripts; in addition these introns can encode open reading frames for so-called intron-encoded proteins (IEPs). Fungal mitochondrial group I introns tend to encode GIY-YIG or LAGLIDADG homing endonuclease genes (HEGs), and group II introns typically encode reverse transcriptase (RT) genes (Lang et al, 2007). These IEPs

tend to catalyze the mobility of their respective introns from an intron-plus to an intron-minus cognate allele. Some IEPs have been shown to assist in the splicing of their host intron by providing so called maturase activity; i.e. these IEPs promote the folding of the intron RNA into a splicing competent structure (Belfort et al., 2002; Belfort, 2003). Various host genome-encoded factors also have been co-opted to assist in the splicing of mitochondrial introns (Edgell et al., 2011).

Group I introns primarily mobilize via a DNA-based mechanism that involves its IEP [homing endonucleases (HEases)] generating a double-stranded cut at a cognate allele that is repaired by the double-strand break repair system. This involves homologous recombination using the intron-plus allele as the repair template and in a nonreciprocal manner the intron sequence and sometimes some flanking markers are transferred to repair the double-stranded break. Group II introns primarily act like retroelements where the mobility pathway utilizes an RNA intermediate and reverse transcriptase activity. In general, mobility of these group I and group II introns is referred to as homing or retrohoming, respectively, as they tend to invade cognate alleles that have not yet been invaded at a particular site. However, these elements can potentially insert into new locations (ectopic integration) and here the terms transposition or retro-transposition are applicable (Hausner, 2003). Mobile introns and homing endonuclease genes are sometimes referred to as selfish DNAs or selfish genes as they do not appear to benefit the host genome. It is generally assumed that these elements are neutral with regards to phenotype thus ensuring their survival (Edgell et al., 2011). However, being “neutral” (i.e. evolving by drift), can result in the rapid degeneration of these elements due to a lack of selection. In order for long term persistence, these elements have to invade intron-less alleles or

invade new sites (Goddard and Burt, 1999). There is also considerable evidence that these elements move horizontally across species barriers ensuring their long term persistence within populations or among fungal mitochondrial genomes (Wu and Hao, 2014).

The current study characterized the mitochondrial genomes of four strains of *E. resinifera*. The most noteworthy findings for these large (>214 kbp) genomes were the large numbers of introns and their IEPs. In addition, other components such as the protein-coding genes and other genetic components such as intergenic regions and remnants of inserted plasmids are also being described. These intron-rich genomes provide an opportunity to examine the mobility of group I introns and their HEGs. Currently, the actual mechanisms for intron acquisition and loss are still poorly understood. With regards to intron content one could speculate that ancestral mtDNAs were intron-rich and they are gradually being eroded and lost or alternatively, introns are continuously lost and reacquired by outcrossing or horizontal transfer (Goddard and Burt, 1999; Wu and Hao, 2014). The *E. resinifera* mitochondrial genomes were compared among the strains, and also with three related mitogenomes from *Ceratocystis platani, C. cacaofunesta and C. fimbriata* in order to gain a better understanding of the evolutionary mechanisms that could promote intron loss or gain and mitogenome rearrangements.

2.2 Results

2.2.1 Mitochondrial genome of *E. resinifera*

The mitochondrial genomes for four *E. resinifera* strains [WIN(M)79, WIN(M)1409A, WIN(M)1410B, WIN(M)1411] were sequenced, assembled and annotated. The mitogenome for

strain WIN(M)79 was annotated first and this genome was used as the reference mtDNA (Genbank accession number: MH551223) for this study; overall, the four *E. resinifera* mtDNAs only differed by the absence or presence of 4 introns and 2 single nucleotide polymorphisms and a few indels (see Section 2.2.9 and Table 2.4 for details). The genome of *E. resinifera* is composed of protein-coding genes such as (*atp6*, *atp8*, *cob*, *cox1-3*, *nad1-6* and *nad4L*) and rRNA (*rns* and *rnl*) and tRNA (27 tRNAs) genes. Most of the protein and rRNA -coding genes were noted to be populated with group I and group II introns, and most introns encode open reading frames (iORFs). The mitogenomes of the *E. resinifera* strains were annotated in Artemis and visualized [WIN(M)79] in Circos (Figure 2.1).


Endoconidiophora resinifera
WIN(M)79 mtDNA
215,106 bp
0
20
40
60
80
100
120
140
160
180
200
cob*
cox1
rnap
trnR(ucu)
nad1
nad4
atp8
atp6
atp6*
rns
cox3
nad6
rnl
rps3
trnL(uaa)
trnM(cau)
trnI(aau)
trnE(uuc)
trnT(ugu)
trnW(uca)
trnM(cau)
trnH(gug)
trnQ(uug)
trnL(uag)
trnF(gaa)
trnA(ugc)
nad3
nad2
atp9
cox2
nad4L
trnR(acg)
nad5
cob
trnC(gca)
trnP(ugg)
trnN(guu)
trnS(uga)
trnI(gau)
trnV(uac)
trnW(uca)
trnS(gcu)
trnD(guc)
trnK(uuu)
trnG(ucc)
trnM(cau)
trnY(gua)

Figure 2.1: The annotated mitochondrial genome of *E. resinifera* [strain WIN(M)79]. The total size of this circular genome is 215 kb (represented by the scale). The position of the tRNAs are shown on the outer track, with the positions connecting to the scale with red lines. The first inner circle represents the position, size and the names of the protein-coding and rRNA genes. The introns are shown in the second inner circle and are colour coded according to the intron types/subtypes: group II (yellow), group IA (purple), group IB and group I derived (very light red), group IC (orange), and group ID (dark red). The third inner circle is to visualize the presence of the LAGLIDADG (green) or GIY-YIG (black) homing endonuclease genes encoded by the introns. The innermost circle is the GC plot of this genome; calculating GC% of genome features.

2.2.2 Protein-coding, rRNA and tRNA genes

The mitogenome of *E. resinifera* contains 14 protein-coding genes and this includes NADH dehydrogenase subunits (*nad1*, *nad2*, *nad3*, *nad4*, *nad4L*, *nad5* and *nad6*) which contribute towards the electron transport complex I, cytochrome oxidase subunits (*cob*, *cox1*, *cox2* and *cox3*) that are part of complex III and complex IV, ATP synthase subunits (*atp6* and *atp8*) and the gene that encodes the 40S ribosomal protein S3 (*rps3*). The genome contains the small and large ribosomal RNA genes (*rns* and *rnl*). A total of 27 tRNA genes were identified of which most are located around the *rnl* gene; 10 tRNA genes are upstream and 12 are downstream of the *rnl* gene although the upstream tRNA cluster is interrupted by the *nad6* gene. The remaining 5 tRNA genes are dispersed along the genome. All of the protein-coding genes, rRNA and tRNA genes reside on one of the strands of the mtDNA. The gene sizes, positions and arrangements are depicted in Figure 2.1.

2.2.3 Introns and intron-encoded ORFs

Eighty-one introns were found within the mitogenome of *E. resinifera* WIN(M)79 and 72 of them contain open reading frames coding for homing endonucleases (HEs). Among the 81 introns, according to RNAweasel (Beck and Lang, 2009), 12 can be assigned to group-IA, 32 to group-IB, 17 to group-IC, 8 to group-ID, 9 to group-I(derived) introns (Table 2.1) and three introns have features diagnostic for group II introns (Michel and Westhof, 1990; Toor and Zimmerly, 2002; Hausner et al., 2014). Seventy-two group I introns in *E. resinifera* encode one or two intron-encoded open reading frames (iORFs). In total, 76 LAGLIDADG and 15 GIY-YIG type ORFs were identified within the *E. resinifera* group I introns. Group II introns can code for

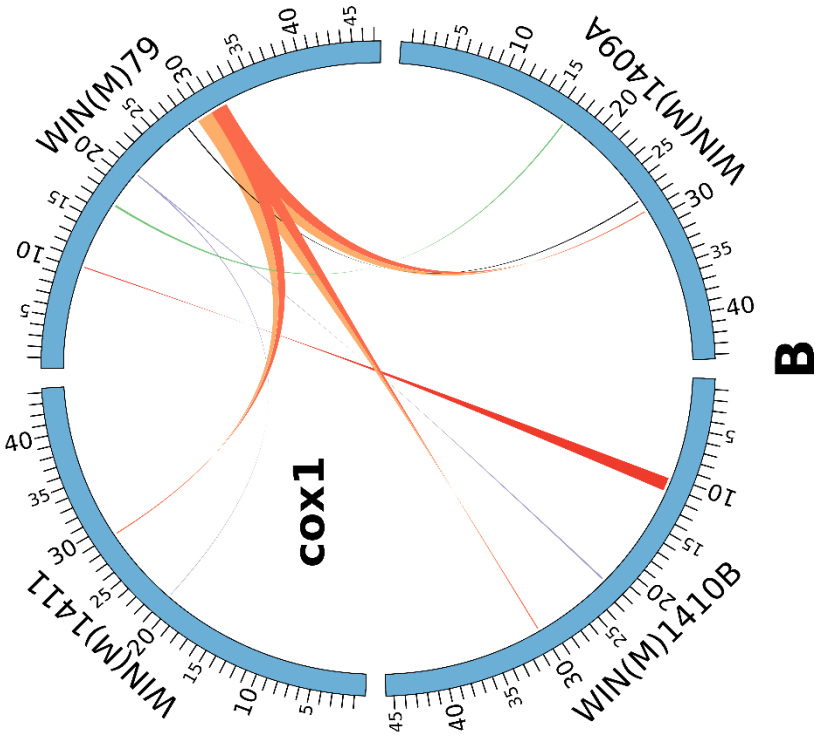
either homing endonuclease or reverse transcriptases (Toor and Zimmerly, 2002). In this study we observed three group II introns. Two of them (located in *rns* and *rnl* gene) encoded a LAGLIDADG type ORF. Another group II intron (located in *cox1* gene) appeared to have no ORF.

Table 2.1: List of introns (group I and group II) in each gene in *E. resinifera* strain WIN(M)79.

Name of the gene	Group IA introns	Group IB introns	Group IC introns	Group ID introns	Group I (derived) introns	Total Group I introns	Group II introns	Total number of Introns
cox1	1	17	1	1	2	22	1	**23**
nad1	2	3	-	-	-	5	-	**5**
nad4	-	1	1	-	-	2	-	**2**
atp8	-	-	-	-	-	-	-	**-**
atp6	-	-	-	-	-	-	-	**-**
rns	-	-	1	1	2	4	1	**5**
cox3	3	2	2	2	-	9	-	**9**
nad6	-	-	-	1	-	1	-	**1**
rnl	3	1	4	1	1	9	1	**11**
rps3	-	-	-	-	-	-	-	**-**
nad2	-	-	4	-	1	5	-	**5**
nad3	-	-	-	-	-	-	-	**-**
atp9	-	-	-	-	-	-	-	**-**
cox2	-	2	2	-	1	5	-	**5**
nad4L	-	-	1	-	-	1	-	**1**
nad5	-	3	-	2	2	7	-	**7**
cob	3	3	1	-	-	7	-	**7**
Total	**12**	**32**	**17**	**8**	**9**	**77**	**3**	**81**

2.2.4 A twintron (mS917a and b) in the *rns* gene

Among the four strains of *E. resinifera*, the WIN1410B strain showed a unique type of intron arrangement within the *rns* gene (Figure. 2.2A). All strains analyzed for this species have a group ID intron at position S917 (intron insertion sites designated according to Johansen and Haugen, 2001). The available *Ceratocystis* species also contain the mS917 intron. However, *E. resinifera* WIN(M)1410B has two group ID introns (instead of one) side by side (a and b) without any apparent exon sequence separating them. This special arrangement can be termed "tandem intron" or a "side by side" twintron. The mS917a and mS917b intron-encoded ORFs of the tandem intron are both related to the mS917 clade of LAGLIDADG ORFs previously characterized (Hafez et al., 2013; Bilto et al., 2017). Phylogenetic analysis showed that the components of the tandem intron mS917a (twintron front part) and mS917b (twintron back part) ORFs are paralogues and other members of this clade can be located in group ID introns located within the *rnl*, *nad5*, *nad6* and *cox3* genes. Moreover, the intron ORF from the mS917b component groups with the intron ORFs located in the *rnl* intron (mL2029) clade, whereas the mS917a ORF groups with orthologues located within the *rns* mS917 intron. Based on the phylogenetic distribution of members of the 917 family of HEs it would appear that the mS917b intron/ORF is derived from a version of the mL2029 intron that has inserted (ectopically) immediately after the mS917a intron (Figure 2.3).


WIN(M)1409A
WIN(M)1410B
WIN(M)1411
WIN(M)79
cox1
B

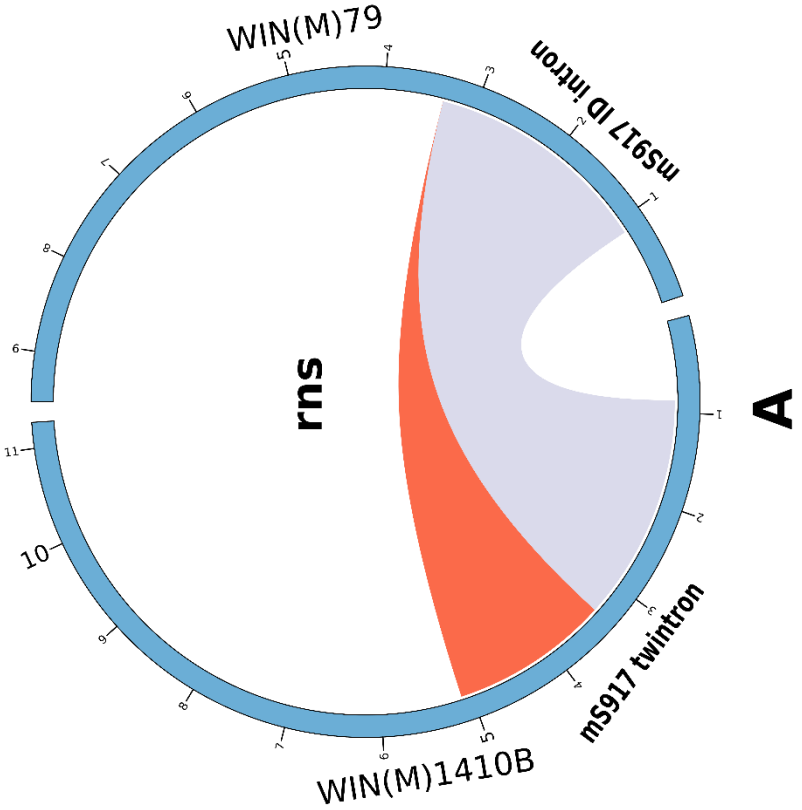

WIN(M)79
mS917 ID intron
mS917 twintron
WIN(M)1410B
rns
A

Figure 2.2: Circos generated diagrams showing the differences between the *rns* and *cox* genes among members of *E. resinifera*. The red lines show the presence of additions introns. The comparison of *rns* and *cox1* gene from four strains [WIN(M)79, WIN(M)1409A, WIN(M)1410B, WIN(M)1411] of *E. resinifera* considering the strain WIN(M)79 as a reference. (A) Comparing the *rns* gene showed that there is one novel group ID intron in WIN(M)1410B which is referred as a side by side twintron (mS917 twintron). (B) Comparison of *cox1* genes showing that there are two additional introns in WIN(M)79 and one additional intron in WIN(M)1410B, moreover there are small indels in different intronic regions.

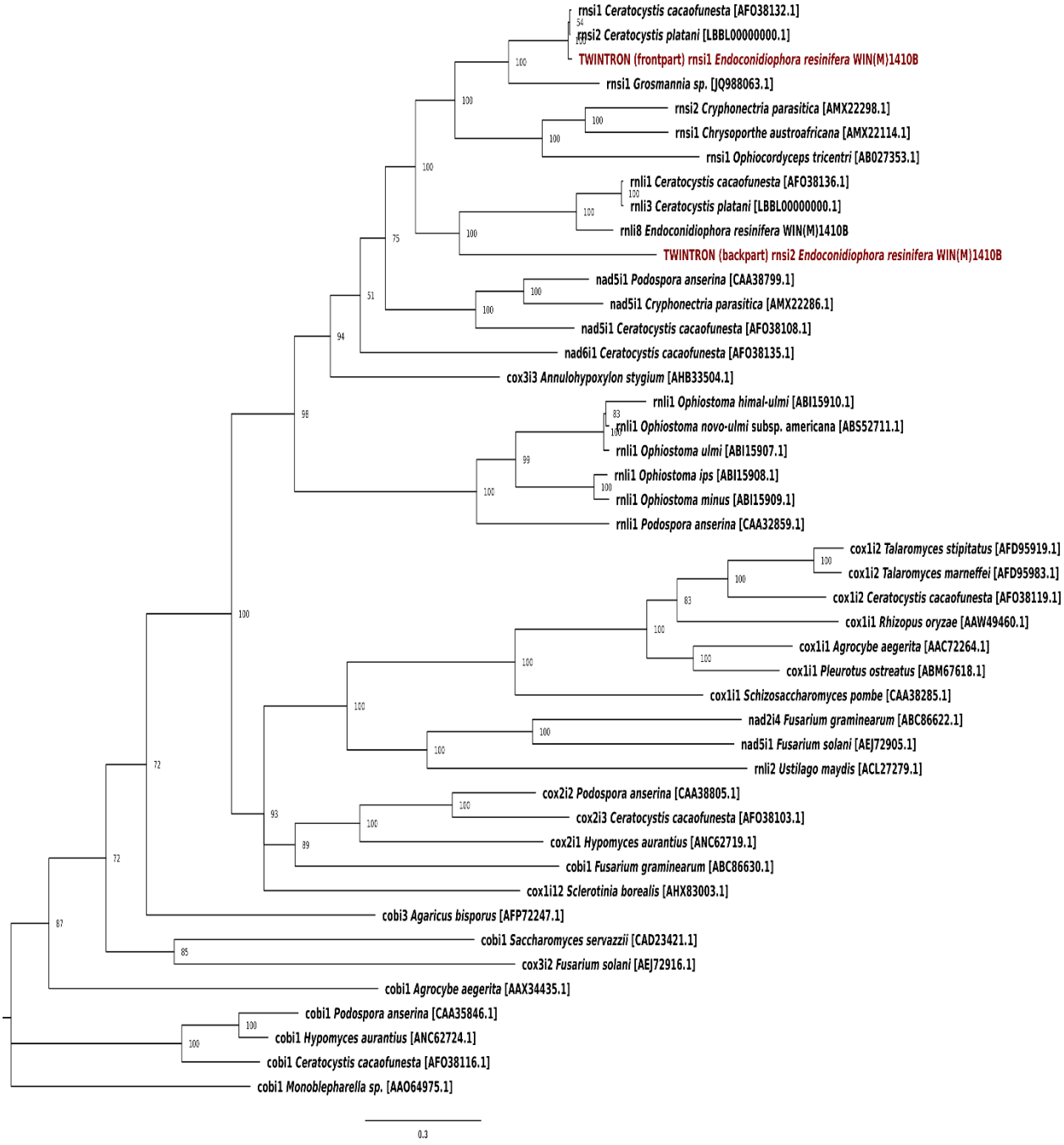

rnsi1 Ceratocystis cacaofunesta [AFO38132.1]
rnsi2 Ceratocystis platani [LBBL00000000.1]
TWINTRON (frontpart) rnsi1 Endoconidiophora resinifera WIN(M)1410B
rnsi1 Grosmannia sp. [JQ988063.1]
rnsi2 Cryphonectria parasitica [AMX22298.1]
rnsi1 Chrysoporthe austroafricana [AMX22114.1]
rnsi1 Ophiocordyceps tricentri [AB027353.1]
rnli1 Ceratocystis cacaofunesta [AFO38136.1]
rnli3 Ceratocystis platani [LBBL00000000.1]
rnli8 Endoconidiophora resinifera WIN(M)1410B
TWINTRON (backpart) rnsi2 Endoconidiophora resinifera WIN(M)1410B
nad5i1 Podospora anserina [CAA38799.1]
nad5i1 Cryphonectria parasitica [AMX22286.1]
nad5i1 Ceratocystis cacaofunesta [AFO38108.1]
nad6i1 Ceratocystis cacaofunesta [AFO38135.1]
cox3i3 Annulohypoxylon stygium [AHB33504.1]
rnli1 Ophiostoma himal-ulmi [ABI15910.1]
rnli1 Ophiostoma novo-ulmi subsp. americana [ABS52711.1]
rnli1 Ophiostoma ulmi [ABI15907.1]
rnli1 Ophiostoma ips [ABI15908.1]
rnli1 Ophiostoma minus [ABI15909.1]
rnli1 Podospora anserina [CAA32859.1]
cox1i2 Talaromyces stipitatus [AFD95919.1]
cox1i2 Talaromyces marneffei [AFD95983.1]
cox1i2 Ceratocystis cacaofunesta [AFO38119.1]
cox1i1 Rhizopus oryzae [AAW49460.1]
cox1i1 Agrocybe aegerita [AAC72264.1]
cox1i1 Pleurotus ostreatus [ABM67618.1]
cox1i1 Schizosaccharomyces pombe [CAA38285.1]
nad2i4 Fusarium graminearum [ABC86622.1]
nad5i1 Fusarium solani [AEJ72905.1]
rnli2 Ustilago maydis [ACL27279.1]
cox2i2 Podospora anserina [CAA38805.1]
cox2i3 Ceratocystis cacaofunesta [AFO38103.1]
cox2i1 Hypomyces aurantius [ANC62719.1]
cobi1 Fusarium graminearum [ABC86630.1]
cox1i12 Sclerotinia borealis [AHX83003.1]
cobi3 Agaricus bisporus [AFP72247.1]
cobi1 Saccharomyces servazzii [CAD23421.1]
cox3i2 Fusarium solani [AEJ72916.1]
cobi1 Agrocybe aegerita [AAX34435.1]
cobi1 Podospora anserina [CAA35846.1]
cobi1 Hypomyces aurantius [ANC62724.1]
cobi1 Ceratocystis cacaofunesta [AFO38116.1]
cobi1 Monoblepharella sp. [AAO64975.1]
0.3

Figure 2.3: Phylogenetic tree of the mS917 group ID introns encoded ORFs (from *rns* gene of *Endoconidiophora resinifera* WIN(M)1410B) and its homologues from different fungal species. The dataset included the amino acid sequences of 43 intron-encoded ORFs (collected from Bilto et al., 2017) and the twintron front and back parts. The aligned dataset was processed for 1 million generation in MrBayes with the mixed-model default setting and the consensus tree was generated. Gene names (with its intron number), species name and NCBI accession number are mentioned for each species. The front part (mS917) of the "side-by-side" twintron is found in the *rns* intronic ORF clade whereas the twintron back-part clustered with the *rnl* intronic ORFs. The numbers at the nodes refer to "posterior probability values"; these are analogous to "bootstrap support values" which provide a measure of the level of confidence that can be placed at nodes. The branch lengths are based on MrBayes analysis and are proportional to the mean number of substitutions per site (see scale bar).

2.2.5 GC percentage and composition of the genome

The mitogenome of *E. resinifera* is AT-rich (71%). The average GC content of the genome is 29%, which is maintained across the mtDNA. The tRNA genes, however, have a higher GC (up to 50%) content compared to the rest of the genome. The mt-genome is composed of genes (exons and introns), rRNAs (exons and introns), tRNAs, intron-encoded ORFs, and intergenic regions. However, the majority of the genome is comprised of introns and intron-encoded ORFs (68% of the genome). The introns embedded within the protein-coding genes make up 56% of the entire mitogenome. The second major component of the mitogenome is made up of the introns present within the rRNA genes (12%). The nucleotide sequences for protein-coding sequences (CDS) and rRNAs comprise 7% and 3% of the mitogenome, respectively. The tRNAs make up 1% of the genome, and 21% of the mitogenome is comprised of intergenic sequences. With regards to the 68% intron component, it can be arranged as follows for the intron subtypes: group IA, IB, IC, ID, I (derived) and group II as follows 14%, 42%, 22%, 9%, 10% and 3% respectively. The iORFs occupy half (50%) of the intron bases. The LAGLIDADG ORFs comprise 42% and the GIY-YIG ORFs make up 8% of all the bases that make up the introns (Figure 2.4).

7%
56%
21%
1%
3%
12%
CDS
Intron in genes
rRNA introns
rRNA exon
tRNA
Non-coding DNA
A
14%
42%
3%
10%
9%
22%
Group IA
Group IB
Group IC
Group ID
Group I (derived)
Group II
B
42%
8%
50%
LAGLIDADG
GIY-YIG
Non-coding bases
C

Figure 2.4: Composition of the mitochondrial genome of *Endoconidiophora resinifera*. (A) The proportion of the protein coding genes (CDS and introns), rRNA genes (exon and intron), tRNA genes and intergenic (non-coding) sequences. (B) Introns (in protein coding and rRNA genes) are further categorised into their subtypes (such as group IA, IB, IC, ID, derived I, and group II). (C) Composition of the introns in terms of their encoded ORFs (either LAGLIDADG or GIY-YIG homing endonuclease gene) and non-coding bases in introns.

2.2.6 Largest *cox1* gene encoded within the largest mtDNA recorded so far among the ascomycetes

The mitogenome of *E. resinifera* is one of the largest [215 kb for WIN(M)79 and 220 kb for WIN(M)1410B] mitogenomes so far reported for a member of the Ascomycota and it also contains the largest *cox1* gene (47.5 kb) recorded so far for any fungus. The size of this genome is for the most part due to the large number of introns (81 introns) and the *E. resinifera* mitogenome in comparison to the other *Ceratocystis* spp. appears to have higher numbers of introns (Table 2.2). The *cox1* gene also appears to have expanded in *E. resinifera* to 47.5 kb due to the large number of introns (23 introns). Among the 23 introns, 22 are group I introns and one group II intron was identified. The members of group IB are the most abundant (17 group IB introns) intron type in the *cox1* gene (Table 2.1). This gene is rich in introns compared to other genes in the mitogenome. The *cox1* gene is a conserved gene (at the CDS (Coding Sequence) level), but the intron numbers are variable among different strains examined for *E. resinifera* and also variable among species of *Ceratocystis* (Table 2.2).

Table 2.2: Comparison of the mitochondrial genome of the *E. resinifera*, *C. cacaofunesta* (JX185564.1), *C. platani* (LBBL00000000.1) and *C. fimbriata* (APWK03000239.1).

Organism	**mtDNA size (bps)**	**GC%**	**Number of introns (per gene)**													**Total introns**
			cox1	***nad1***	***nad4***	***atp6***	***rns***	***cox3***	***nad6***	***rnl***	***nad2***	***cox2***	***nad4L***	***nad5***	***cob***	
C. cacaofunesta	103,147	26	10	1	1	2	1	2	1	3	0	6	1	7	2	37
C. platani	116,162	27	9	2	1	2	2	4	0	5	1	5	1	6	3	41
C. fimbriata	141,204	27	12	3	1	2	1	4	0	4	2	6	1	10	5	51
E. resinifera WIN(M)79	215,106	29	23	5	2	0	5	9	1	11	5	5	1	7	7	81
WIN(M)1409A	215,081	29	21	5	2	0	5	9	1	11	5	5	1	7	7	79
WIN(M)1410B	220,224	29	22	5	2	0	6	9	1	11	5	5	1	7	7	81
WIN(M)1411	214,750	29	21	5	2	0	5	9	1	11	5	5	1	7	7	79

2.2.7 Open Reading Frames (ORFs) and gene fragments within the intergenic spacers: HEGs and a plasmid-derived RNA polymerase

Twenty ORFs were detected in the intergenic regions of the mitogenome of *E. resinifera*. A BLASTp search of those ORFs against the NCBI non-redundant database showed the presence of a partial DNA-dependent RNA polymerase (*rnap*) gene, which showed homology with a mitochondrial plasmid encoded *rnap* gene in *Neurospora intermedia* (Xu et al., 1999). Previously, a degenerated RNA polymerase gene was reported from *C. cacaofunesta*. Nine degenerated GIY-YIG and eight LAGLIDADG (degenerated) ORFs along with partial duplication of the *atp6* and *cob* genes were also recorded from the intergenic spacers (Table 2.3). The partial duplication (C-terminal segment) of the *cob* gene is located downstream of the *cob* gene (genomic position: 213807-214152). There are three degenerated LAGLIDADG ORFs situated in the intergenic space between *cob* and the partial C-terminal duplication of the *cob* gene. The *atp6* gene is followed by a C-terminal duplication of the *atp6* gene and this duplicated segment is located between the *atp6* and *rns* genes (genomic position: 81091-81417). The duplicated *atp6* segment is flanked by 5 GIY-YIG and 2 LAGLIDADG type ORFs. Partial C-terminal duplications of the *cob* and *atp6* genes were noted in all four strains of *E. resinifera*. In strain WIN(M)1411 the complete duplication of the *trnA* gene was recorded and the duplicated version also included some of the upstream bases associated with the original copy of *trnA* (genomic position 149128-149198 duplicated at 149958-150028).

Table 2.3: List of the genetic components found in intergenic regions. Most of them are degenerated HEGs, however gene duplications and a RNA polymerase gene were found.

Gene name	Start (genomic position)	End (genomic position)
RNA Polymerase (*rnap*)	50255	50575
HEG (GIY-YIG)	75590	76156
HEG (GIY-YIG)	76179	76667
HEG (GIY-YIG)	76570	76890
HEG (GIY-YIG)	78102	78674
HEG (GIY-YIG)	78675	79124
HEG (LAGLIDADG)	79896	80213
HEG (LAGLIDADG)	80323	80727
ATP Synthase su.6 (*atp6*)	81091	81417
HEG (GIY-YIG)	92287	92598
HEG (LAGLIDADG)	111547	112203
HEG (LAGLIDADG)	112002	112985
HEG (GIY-YIG)	145826	146419
HEG (GIY-YIG)	146368	146736
HEG (GIY-YIG)	146670	147176
HEG (LAGLIDADG)	152425	153078
HEG (LAGLIDADG)	211869	212420
HEG (LAGLIDADG)	212342	212764
HEG (LAGLIDADG)	212868	213275
Cytochrome b (*cob*)	213807	214152

2.2.8 Degenerated *atp9*

The ATP synthase subunit 9 (*atp9*) gene sequence is present in the mitogenome of *E. resinifera* [genomic position: 163871-164096 in WIN(M)79] but it appears to have degenerated due to the presence of a premature stop codon. Blastx (BLAST – see original reference) analysis showed a strong match (70% identity) to the *atp9* gene of *C. cacaofunesta* (GenBank accession YP_007507043.1). The same phenomenon was noted for *C. platani* (GenBank accession LBBL00000000.1) where its *atp9* gene sequence showed near 100% identity with the *atp9* sequence of *C. cacaofunesta*, but the *C. platani atp9* sequence was also interrupted by a premature stop codon. It is noteworthy that the *atp9* gene is absent in the *C. fimbriata* mitogenome (GenBank accession APWK03000239.1).

2.2.9 Genome comparison

The genomes of *E. resinifera* strains are highly conserved with polymorphism mainly due to the presence or absence of introns along with some short indels (insertions) and nucleotide substitutions in the non-coding sequences. The differences among the strains in gene sequences (including intron and exon) are compiled in Table 2.4. Briefly, it was found that the CDS of gene sequences were highly conserved among the strains; only one silent mutation was noted in the *cox1* gene in the WIN(M)1410B strain (Figure 2.5). The *rnl* gene of WIN(M)1409A and WIN(M)1410B showed small insertions and the *trnA* gene of WIN(M)1411 showed a small insertion (see Table 2.4). But sources of mitogenomic variability among the strains are due to variations of the number of introns along with some small indels in the intronic sequences.

Additional introns were noted in the *rns*, *cox1* and *cox2* genes of strain WIN(M)1410B (Fig. 2.2A and B). Moreover, two additional introns were found in the *cox1* gene of strain WIN(M)79 (Figure 2.2B). We found no significant variations with regards to intergenic regions.

Table 2.4: Notable differences in genes (exons and introns) sequences from different *E. resinifera* strains.

Exon sequence in Gene (name)	Nucleotide position in the genome				
	WIN(M)79 (reference)	**WIN(M)1409A**	**WIN(M)1410B**	**WIN(M)1411**	**Remarks**
rnl	131931	128784-130899	132034-134149		Insertion
	132449	131416-135650			Insertion (intron exon overlap)
cox1	10217		10217		Substitution (G to T)
trnA	150313			149242-150088	Insertion
Intron sequence in Gene (name)	**Nucleotide position in the genome**				
	WIN(M)79 (reference)	**WIN(M)1409A**	**WIN(M)1410B**	**WIN(M)1411**	**Remarks**
cox1	10226		10226-11548		Novel intron
	17008	17008-17041			Insertion
	20796		22117-22146	20796-20825	Insertion
	27592	27653			Deletion
	28800-30348	28859	30149	28828	Novel intron in WIN(M)79

	30360-32013	28869	30159	28838	Novel intron in WIN(M)79
rns	85745		83902-85657		mS917 Twintron
rnl	131931			128745-130860	Insertion
	132443-132457		134660-135555		Insertion
	132597		135695-135737		Insertion
cox2	168549		171691-173721		Novel intron
cob	203413-203430		208583		Deletion
	204218-204220		209371		Deletion
nad1	62101			58934	Substitution (A to T)

WIN(M)79/1-215106
WIN(M)1409/1-215081
WIN(M)1410/1-220224
WIN(M)1411/1-214790
Consensus
10200
10209
10219
10224
TCGGCTTGTATTGAAGGGGGTGCTG

Figure 2.5: Multiple sequence alignment of the four strains of *Endoconidiophora resinifera*. The represented region is a part of a CDS in *cox1* gene where the WIN(M)1410B strain has a single nucleotide polymorphism (SNP), a transversion of G to T. That is the only SNP found in the CDS region among those strains.

For a more detailed comparison among *Ceratocystis sensu lato* species, we have collected the mitogenome data for *C. cacaofunesta, C. platani* and *C. fimbriata*. *Ceratocystis cacaofunesta* is fully annotated (GenBank accession: JX185564.1), and *C. platani* is available as one contig (but was not annotated) in GenBank (LBBL00000000.1) and in ENA database (GCA_000978885.1). The mitogenome for *C. fimbriata* is available in GenBank as a single contig (APWK03000239.1) but also not annotated. We have annotated *C. platani* and *C. fimbriata* for this study. The protein-coding regions were translated and extracted to compile a concatenated dataset that allowed these fungi to be compared with each other along with other members of the Ascomycota. The phylogenetic analysis showed that the *Ceratocystis s. l.* spp. are distinct from each other and they do comprise a separate clade (Microscales) in the phylogenetic tree based on concatenated mtDNA protein-coding sequences (Figure 2.6). All four members of *Ceratocystis s.l.* grouped into one clade with *E. resinifera* forming the basal member and sequences for *C. platani* and *C. fimbriata* grouping together. The phylogenetic tree overall showed strong node support values for the major nodes and the Microascales were positioned between the following Orders Hypocreales and Glomerellales (Figure 2.6).

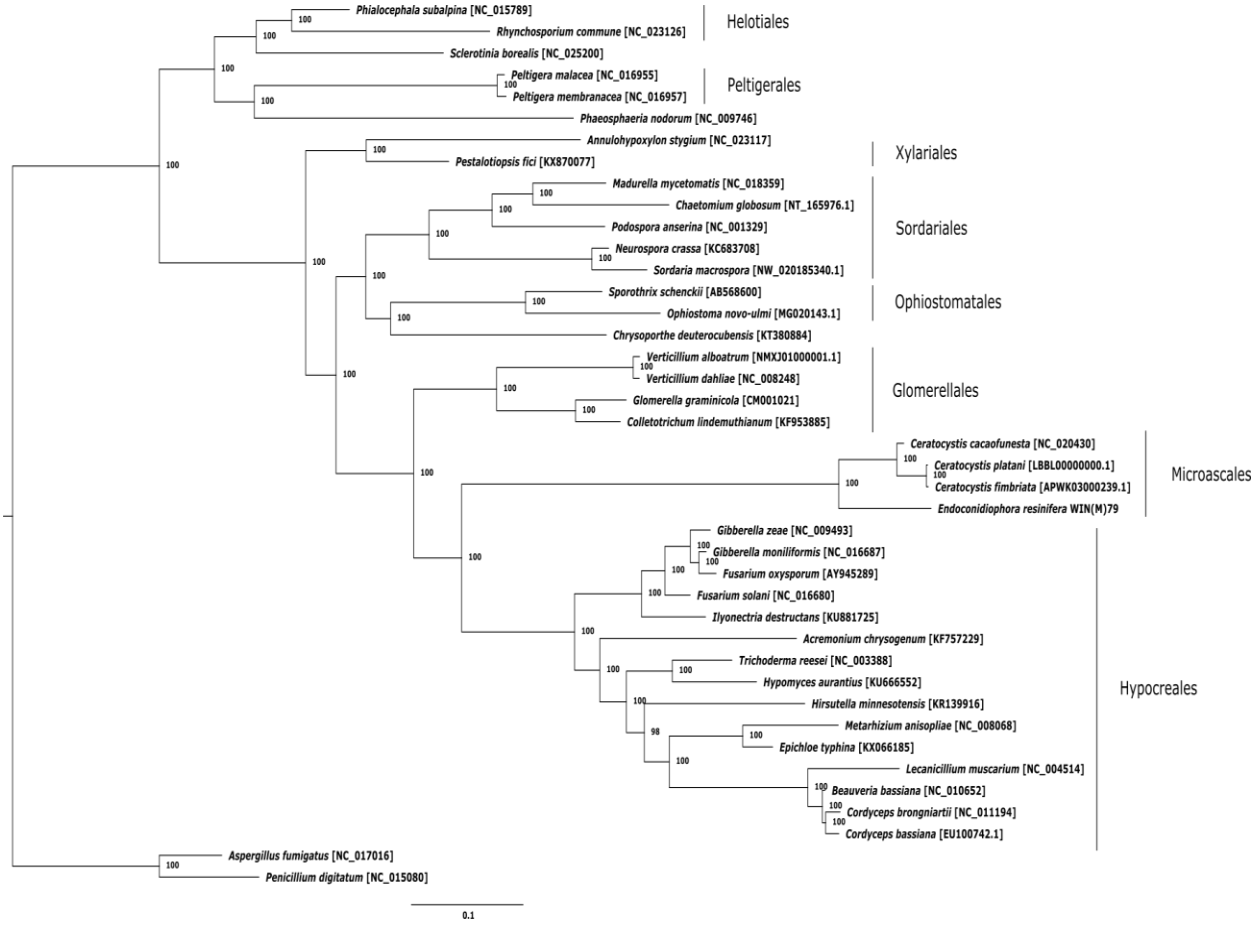

Phialocephala subalpina [NC_015789]
Rhynchosporium commune [NC_023126]
Sclerotinia borealis [NC_025200]
Helotiales
Peltigera malacea [NC_016955]
Peltigera membranacea [NC_016957]
Peltigerales
Phaeosphaeria nodorum [NC_009746]
Annulohypoxylon stygium [NC_023117]
Pestalotiopsis fici [KX870077]
Xylariales
Madurella mycetomatis [NC_018359]
Chaetomium globosum [NT_165976.1]
Podospora anserina [NC_001329]
Neurospora crassa [KC683708]
Sordaria macrospora [NW_020185340.1]
Sordariales
Sporothrix schenckii [AB568600]
Ophiostoma novo-ulmi [MG020143.1]
Ophiostomatales
Chrysoporthe deuterocubensis [KT380884]
Verticillium alboatrum [NMXJ01000001.1]
Verticillium dahliae [NC_008248]
Glomerella graminicola [CM001021]
Colletotrichum lindemuthianum [KF953885]
Glomerellales
Ceratocystis cacaofunesta [NC_020430]
Ceratocystis platani [LBBL00000000.1]
Ceratocystis fimbriata [APWK03000239.1]
Endoconidiophora resinifera WIN(M)79
Microascales
Gibberella zeae [NC_009493]
Gibberella moniliformis [NC_016687]
Fusarium oxysporum [AY945289]
Fusarium solani [NC_016680]
Ilyonectria destructans [KU881725]
Acremonium chrysogenum [KF757229]
Trichoderma reesei [NC_003388]
Hypomyces aurantius [KU666552]
Hirsutella minnesotensis [KR139916]
Metarhizium anisopliae [NC_008068]
Epichloe typhina [KX066185]
Lecanicillium muscarium [NC_004514]
Beauveria bassiana [NC_010652]
Cordyceps brongniartii [NC_011194]
Cordyceps bassiana [EU100742.1]
Hypocreales
Aspergillus fumigatus [NC_017016]
Penicillium digitatum [NC_015080]
0.1

Figure 2.6: Phylogenetic position of *Endoconidiophora resinifera* among the Ascomycota. The tree was inferred by Bayesian method in MrBayes using 1 million generation with gamma-distributed rates model and generated a consensus tree. The dataset was prepared by concatenating amino-acid sequences of the mitochondrial proteins extracted from the NCBI genome database. The fungal species are clustered according to their orders in the tree rooted with *Aspergillus* and *Penicillium*. The *Endoconidiophora resinifera* placed with its related taxa *Ceratocystis* species in a distinct clade of Microascales.

The variation in genome sizes, GC content and the presence of introns in every gene among the *Ceratocystis s.l.* species are listed in Table 2.2. The data showed that the genome size and intron number of *E. resinifera* is about double compared to other *Ceratocystis* species. A comparative alignment of all these mitogenomes was done with the Mauve program (Figure 2.7). It clearly showed the homologous blocks shared between these genomes and it also showed a linear relationship among the genes, which implies that the gene order or synteny of these genomes is conserved. Besides the gene synteny for protein and rRNA coding genes, the comparison of the tRNA genes also showed conservation of gene order. However, the number of tRNA genes is not the same among the examined species: 25 tRNAs in *C. fimbriata,* 26 in *C. platani*, 27 in *E. resinifera* and 30 in *C. cacaofunesta* (Table 2.5).

Table 2.5: The gene order of tRNA genes (trn*) from four different species. [The asterisk (*) represent an amino acid that is the product of the trn* gene]

Organism	**tRNA gene (trn*) order**																														
E. resinifera	R		M	Y	G	K	D	S	W	V		I	S	N	P	T	E	I	M	L	A	F	L		Q	H	M		W	R	C
C. cacaofunesta	R	I	M	Y	G	K	D	S	W	V	W	I	S	N	P	T	E	L	M	L	A	F	L	L	Q	H	M	M		R	C
C. platani	R		M	Y	G	K	D	S	W	V	W	I	S	N	P	T	E	L	M	L	A	F	L	L	Q	H	M				
C. fimbriata	R	I	M	Y	G	K	D	S		V	W			N	P	T	E	L	M	L	A	F	L	L		H	M	M			C

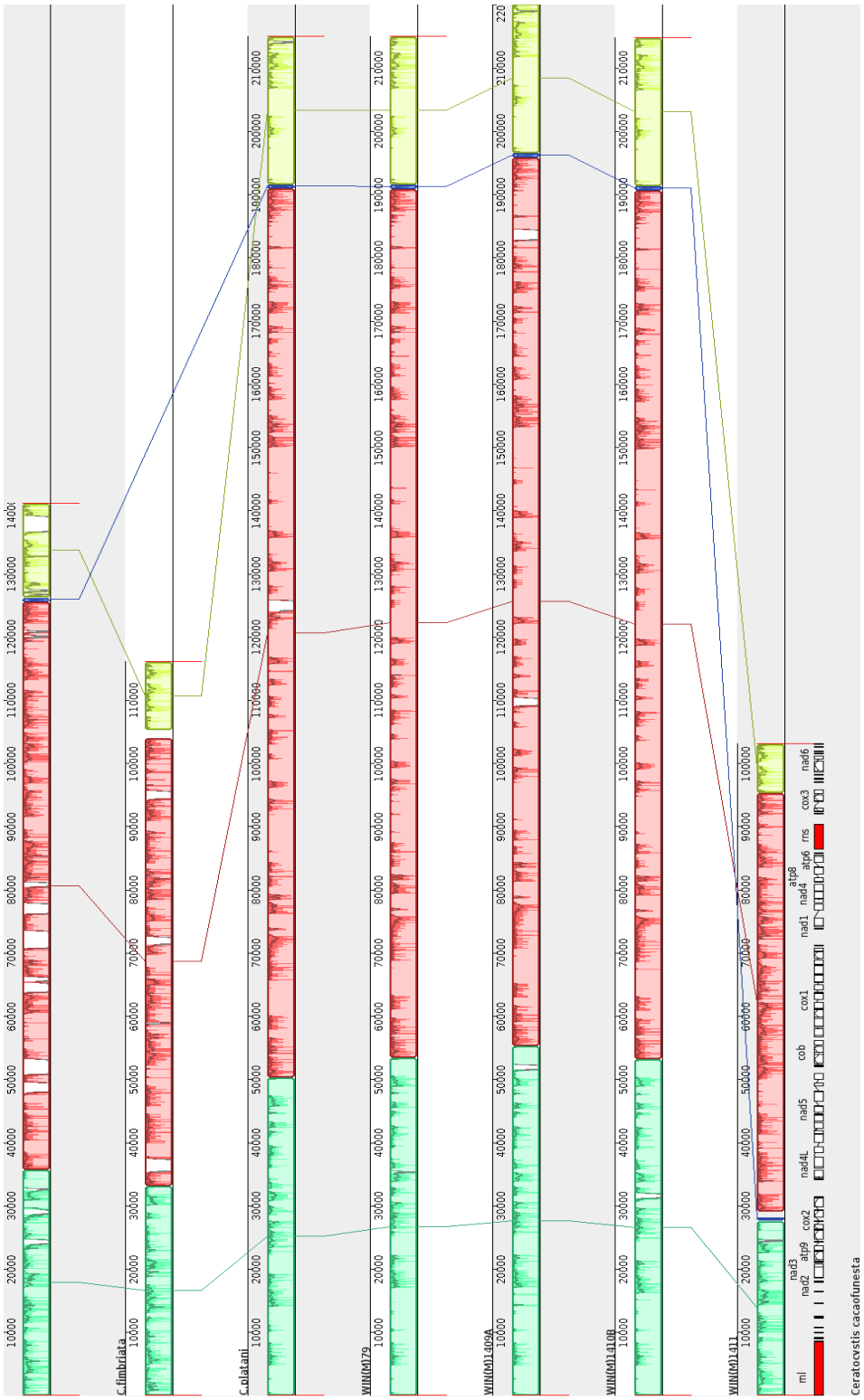

C.fimbriata
C.platani
WIN(M)79
WIN(M)1409A
WIN(M)1410B
WIN(M)1411
rnl
nad2
nad3
atp9
cox2
nad4L
nad5
cob
cox1
nad1
nad4
atp8
atp6
rns
cox3
nad6
Ceratocystis cacaofunesta

Figure 2.7: Genome-wide comparison for species of *Ceratocystis* using Mauve. The progressiveMauve alignment (in Mauve program) shows the homologous blocks (in different colors) shared among the mitogenomes and it also connected these blocks with lines, indicating corresponding position among the homologous blocks in order to visualize the gene arrangement. The GenBank annotation of *Ceratocystis cacaofunesta* has been incorporated at the bottom to show the gene order in different blocks.

2.3 Discussion

2.3.1 Mitochondrial genome architecture among members of *Ceratocystis sensu lato*

The phylogenetic tree generated for ascomycetes fungi, based on concatenated mitochondrial protein sequences, and generated a well-supported topology consistent with previously published reports (Schoch et al., 2009). *Ceratocystis* and allied taxa belong to the Microascales and are distinct from species that can be assigned to other orders such as the Hypocreales, Glomerellales, Xylariales, Sordariales, Ophiostomatales etc. The assembly of the mitogenomes for the tested strains of *E. resinifera* can be represented as circular molecules ranging in size from 214,750 to 220,224 nucleotides. It should be noted that fungal mtDNAs could also have linear topologies and have been proposed to occur as long concatemers, possibly products of a recombination-dependent rolling circle-type DNA replication mechanism (Bendich, 1993; Baidyaroy et al., 2012; Hausner et al., 2006).

The mitochondrial genomes of *E. resinifera* are the largest mitogenomes reported so far for members of the Ascomycota; yet with regards to the standard mtDNA core genes these genomes do not offer additional genes compared to other fungal mitogenomes (Salavirta, 2014; Kang et al., 2017; Mardanov et al., 2014). The *E. resinifera* mitogenome contains 15 protein coding genes, 2 rRNA genes and 27 tRNA genes similar to other fungal mitogenomes. Moreover, the *rps3* gene is embedded within an *rnl* group IA type intron (mL2449), which is a common feature in many filamentous ascomycetes fungi mitogenomes (Sethuraman et al., 2009). The *atp9* gene apparently is found to be present in some fungi and but not in others (van de Sande, 2012). With regards to the mitogenomes examined in this study the *atp9* gene is present in *C.*

cacaofunesta, but in *E. resinifera* and *C. platani* the *atp9* gene has accumulated mutations that generated a premature stop codon and in *C. fimbriata* the *atp9* gene is missing. This would suggest that the *atp9* gene is drifting in some species and a nuclear counterpart might be available that can compensate for the loss of the mitochondrial version of the *atp9* gene.

The progressiveMauve alignment of the mitogenomes for the several *Ceratocystis s.l.* species showed that gene synteny is conserved and variations in mtDNA and gene sizes are mostly due to the expanding numbers of introns. Variation among various strains of *E. resinifera* is restricted to one synonymous substitution in the *cox1* gene, a few SNPs within the intronic sequences or other non-CDS bases along with a few indels within the intergenic regions. Similar to what has been observed in other fungi such as *Chrysoporthe* species (Kanzi et al., 2016), *Aspergillus* and *Penicillium* species (Joardar et al., 2012), *Saccharomyces sensu stricto* species (Ruan et al., 2017), *Tolypocladium inflatum* (Zhang et al., 2017) we observed intron derived size polymorphism among strains of *E. resinifera*.

Other noteworthy features are the fusions of several gene pairs typically involving a one nucleotide overlap among the two reading frames. The overlap of *nad2* with *nad3* genes and *nad4L* with *nad5* by one nucleotide has been noted in other fungi (Aguileta et al., 2014).

2.3.2 Mobile elements and genome expansion (duplication and degeneration)

Recent papers have noted that fungal mitochondrial genomes are dynamic with regards to their structure and composition due to the presence of mobile elements (such as group I and group II introns) and duplication events (Losada et al., 2014; Mardanov et al., 2014; Jalalzadeh et al., 2015). This study found 81 introns in the mtDNAs of *E. resinifera* strains examined which is nearly double the number of introns compared to the other species of *Ceratocystis*. Most noteworthy is the *cox1* gene from *E. resinifera* strain WIN(M)79 that has 23 introns and this gene is 45.7 kb long. The size of the *cox1* gene alone exceeds the sizes of many complete fungal mitochondrial genomes (Deng et al., 2016). The *E. resinifera cox1* gene is the largest reported so far, previously the *cox1* gene from *Agaricus bisporus* at ~30 kb long with 19 introns was reported to be the largest *cox1* gene among the fungi (Ferandon et al., 2010). In *E. resinifera* the *cox1* gene is considerably longer and acquired 4 more introns combined with the intron encoded ORFs this expanded the size of the gene to 45.7 kb. The *cox1* gene has been utilized as a DNA barcoding marker in metazoans, but the presence of potentially large numbers of introns makes the *cox1* gene not very suitable for fungal DNA- based barcoding (Schoch et al., 2012). Mobile elements that require specific target sequences such as group I and group II introns favor genes that are under functional constraints and are highly conserved, making intra- and intergenomic mobility more feasible.

Examples of degenerated intron ORFs were noted and these are to be expected as according to Goddard and Burt (1999) introns and encoded ORFs such as homing endonucleases are not subject to natural selection; thus their sequences drift and can accumulate deleterious

mutations. Neutral evolution is thus a plausible model to explain the potential genome expansion noted among some members of *Ceratocystis s.l.* Although introns appear to be the major factor that contributes towards mtDNA size expansion in *E. resinifera*, insertion of plasmid components (such as the *rnap* gene), and gene duplication events (partial duplication of *atp6*, *cob* and HEGs and the expansion of intergenic spacers also contribute towards the size of the mitogenome.

Overall, the examined mitogenomes for *Ceratocystis* and *Endoconidiophora* species appear to evolve rapidly in gene structure (i.e. intron composition) but slowly in sequence and gene order. This has also been observed plant mitogenomes and some fungi (Kanzi et al., 2016; Palmer et al., 2000). Therefore, our findings show that mtDNA polymorphisms are mostly due to the presence and absence of introns.

2.3.3 "Side-by-side" twintron complex at mS917

Twintrons, or nested introns, have been described from various fungal mitogenomes with various combinations of group I or group II introns nested inside each other. These elements may require that during RNA processing the internal member has to splice first before the external member can be excised from the transcript (Hafez and Hausner, 2015). Deng et al. (2016) noted that in *Hypomyces aurantius* the *cox3* gene harbored a twintron (cox3-i2) that is a "side-by-side twintron" where two group IA introns are arranged in tandem. The *rns* gene in *E. resinifera* [WIN(M)1410B] contains a twintron where two group ID introns are placed next to each other at the S917 position of the *rns* gene. This position (S917) has previously been noted to be invaded

in some fungi by a group ID intron that expresses active HEases (Bilto et al., 2017), in addition it has been recorded that this location in *Cryphonectria parasitica* can be occupied by a nested intron arrangement (or twintron) where a group ID intron that encodes a double motif LAGLIDADG-type ORF is inserted into an ORF-less external group ID intron (Hafez et al., 2013; Monteiro-Vitorello et al., 2009). This arrangement differs from that observed in *E. resinifera* (strain WIN(M)1410B) where two LAGLIDADG ORF-encoding group ID introns are situated next to each other. Based on the phylogenetic relationships between the two members of this side-by-side twintron, it appears the 5′ member is the original resident of S917 site and the 3′ component is due to an ectopic integration event whereby a paralog of the mS917 HEG which was probably located in the *rnl* gene reinvaded the mS917 position. This intron arrangement warrants further characterization in future studies with regards to its splicing pattern and the target preferences for the encoded HEases.

2.3.4 Evolutionary dynamics of the introns and HEGs and the mitochondrial genome

Introns comprise 68% of the mitogenome in *E. resinifera*, and most of the introns contain ORFs encoding putative homing endonucleases. Those ORFs comprise 50% of the size of the introns. Group I introns can move to cognate alleles that lack introns or in some instances, ectopically integrate into new sites, as they encode homing endonucleases (Belfort, 2003). Intron-loss can be mediated when a reverse-transcribed mature transcript replaces the original intron-containing gene (Hausner, 2012). Deletion of introns can also be due to intra- or intergenomic recombination events (Hepburn et al., 2012). The evolutionary dynamics of introns

and homing endonucleases is quite complex, the gain and loss of introns and their encoded ORFs tends to be attributed to a HEG lifecycle (Goddard and Burt, 1999) that is based on neutral evolution. The model is based on the observation that among members of the Saccharomycetales the omega intron (*rnl* gene introns) appears to undergo a cycle of invasion and degeneration; as there is no selection, introns and encoded ORFs accumulate mutations that eventually lead to their erosion and loss. To persist these elements have to keep invading cognate intron-minus alleles, transpose into new sites or move horizontally into new genomes, or gain a function that benefits the host genome (Goddard and Burt, 1999; Wu and Hao, 2014).

The large number of introns noted in some fungal mitochondrial genomes such as *Agaricus bisporus*, *Rhizoctonia solani*, *Cryphonectria parasitica*, *Sclerotinia borealis*, *E. resinifera* etc. are in contrast to small fungal mtDNAs encoding only one intron such as *Sporothrix schenckii* and *S. globosa*; this raises the question if drift is indeed the only possibly explanation for the distribution of introns within fungal mitogenomes? Also, some introns appear to be quite conserved such as the *rnl* intron that in many ascomycetes fungi encodes *rps3*. It is assumed that encoding a potentially essential host gene ensures these introns are not subject to neutral evolution. Other introns for less obvious reason also appear to be observed at relative higher frequencies compared to other introns, such as cob393 and cob490 and possibly these introns may provide some means for gene regulation and therefore selection may favour their maintenance (Guha et al., 2017). Rudan *et al.* (2018) recently presented data from *S. cerevisiae* that suggests the mtDNA introns are important in fine-tuning gene expression and they facilitate the generation of appropriate amounts of transcripts. Belfort (2017) has suggested that some self-

splicing elements could be bio-sensors that can modulate the expression of their host gene. Conversely, many introns (intron insertion sites) have a rather disjointed distribution among the fungi encoding ORFs at various states of degeneration; these introns may be excellent examples of neutral evolution as proposed by Goddard and Burt (1999).

Another category of introns is represented by those that appear to have invaded new sites within the same genome (Haugen and Bhattacharya, 2004), a temporary means of escaping the Goddard and Burt (1999) HEG lifecycle of invasion, decay and eventual loss. Related introns present within the same genome may still interact in *trans* in some collaborative fashion rendering them less prone to extinction (Dabbagh et al., 2017; Martínez-Rodríguez et al., 2014). Some intron encoded proteins (IEPs) can also act as maturases that facilitate the intron RNA to fold into a splicing-competent configuration (Lang et al., 2007). In *S. cerevisiae*, two homologous IEPs have been characterized, and cytb bI4 is required for splicing of both the cytb bI4 intron and the cox1 aI4 intron but the cox1 aI4 IEP can generate double-stranded cuts within *cox1* sequences (Delahodde et al., 1989). Trans-acting interactions between introns and free standing HEGs have been noted among phages; collaborative homing refers to scenarios where a HE can catalyze the mobility of an intron as both share the same insertion/target site (Bonocora and Shub, 2009; Zeng et al., 2009; Edgell, 2009). In scenarios such as the mS917 clade of introns where orthologous elements have spread into the *rns*, *rnl*, *nad5*, *nad6* and *cox3* genes (Bilto et al., 2017) one can propose that there might be some interactions among members of this clade. *E. resinifera* has mS917 members located within the *rns*, *rnl*, *nad5* and *nad6* genes in addition one strain (WIN(M)1410B) contains a side-by-side twintron located at mS917

composed of two members (*rns* and *rnl* version, Figure 2.2A) of this family of introns. When members of a HEG family are present within the same genome one can envision a “hypercycle”-like analogy (Eigen, 1971; Szostak et al., 2016); however, here dependencies are more lose as individual members can drift, become selfish and some members can be short circuited (Wolters et al., 2015). Members would not have to interact in a directional manner instead interactions such as trans-acting maturase activities or possible trans-acting homing activity most likely would be linked to those that recently diverged from each other. This arrangement would provide some degree of stability for the persistence of members of a HEG family allowing members to maintain their numbers or even spread to new locations “outpacing” drift as predicted by the Goddard and Burt model (1999). These types of interactions (Figure 2.8) may in part explain why some mitochondrial genomes have expanded by gaining or maintaining large numbers of introns.

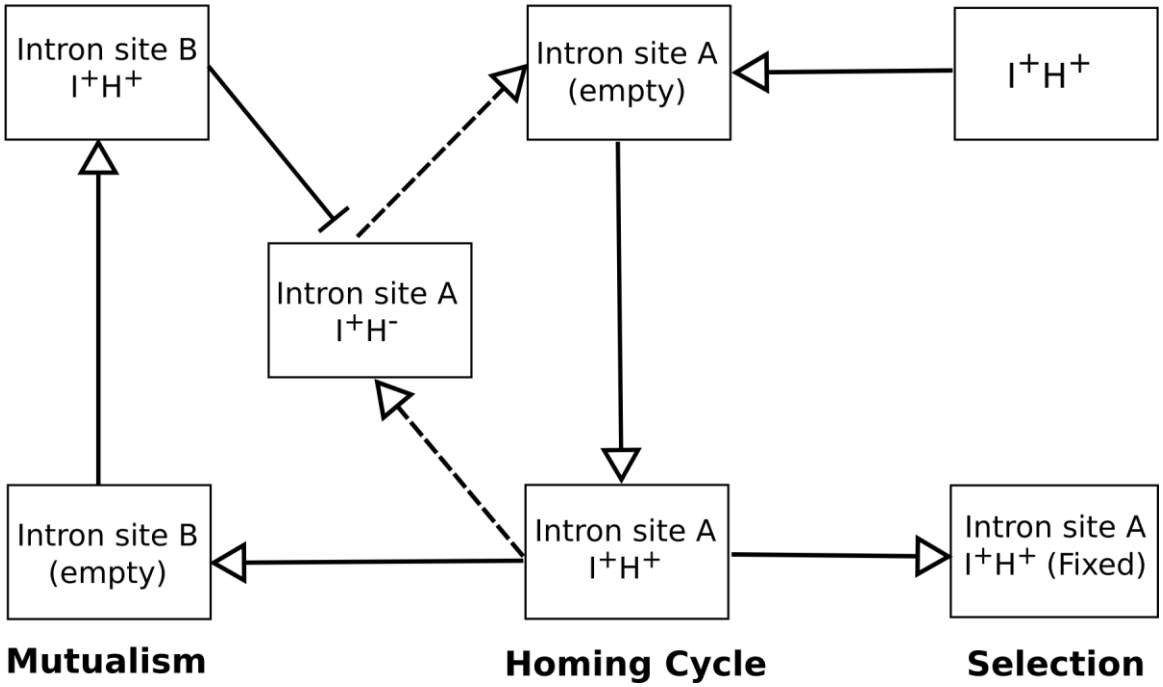

Intron site B
I+H+
Intron site A
(empty)
I+H+
Intron site A
I+H-
Intron site B
(empty)
Intron site A
I+H+
Intron site A
I+H+ (Fixed)
Mutualism
Homing Cycle
Selection

Figure 2.8: The fate of composite elements such as introns plus IEPs (I+H+). The composite element invades an empty site and from here it could possibly spread into other sites (only site B shown for simplicity) and related IEPs could still interact with their ancestral intron version possibly facilitating splicing or mobility thus enhancing the chances of the ancestral intron to persist. This form of mutualism could even complement situations where the ancestral intron ORF has started to accumulate deleterious mutations (H-). Other composite elements may be strictly subject to drift and first the ORF is degenerating and eventually the intron is lost from the genome and possibly from the population. There might be situations where the composite elements have been co-opted as maturases or regulatory elements or as platforms for expressing host genes (*rps3*) and these introns would be subject to adaptive selection and thus could become fixed in the genome and the population (Figure adapted from Gogarten and Hilario, 2006).

Horizontal gene transfer, inter and intra-genome intron mobility, plus gene conversion promoted by IEPs and host genome repair systems combined with drift are the mechanisms that appear to promote intron diversity and potentially intron expansion in fungi (Losada et al., 2014; Haugen and Bhattacharya, 2004; Wolters et al., 2015; Xiao et al., 2017). Why are some fungal genomes almost completely devoid of introns? This could be again due to drift and the biased transmission of mtDNAs that are reduced in size or replicative advantage of smaller mtDNAs or loss of mtDNA introns could be the result of nuclear genome rearrangements that result in the loss of factors that can promote intron splicing, intron RNA stability, intron homing or mtDNA DNA repair (recombination).

2.4 Conclusion

So far, a few mitogenomes are currently available for members of the Microascales. This study examines the mitogenome of *E. resinifera* a species that used to be assigned to the genus *Ceratocystis*. The latter is a genus that has recently undergone extensive taxonomic revisions (de Beer et al., 2017) and the mitogenome might offer mtDNA loci that could be developed into molecular markers assisting in the identification of taxa within this group of economically important fungi. Large mitochondrial genomes offer some insight on mechanisms that might cause these genomes to expand. With regards to *E. resinifera*, introns appear to be a major contributor towards genome expansion. Introns and their encoded homing endonucleases all assemble and initiate further invasions by drift (Guha et al., 2017) but once they have inserted into a host gene several mechanisms may determine their fate. Many probably do fit the model

proposed by Goddard and Burt (1999) whereby these elements drift and thus face eventual elimination due to the accumulation of mutations and persistence within a population requires re-invasion of new loci or loci that lost introns. However, there appears to be evidence that some introns may actually be beneficial either encoding proteins that are useful to the host genome (maturase activity, *rps3*) or introns that can act as gene regulators and thus these introns are maintained within a population. Finally, some introns appear to persist as they are co-operating with other introns promoting a system of mutualism that renders them less prone to extinction.

2.5 Materials and Methods

2.5.1 Culturing fungi – The fungi were grown at 25°C for 8-10 days on malt extract agar (MEA – 3% malt extract, 2% agar and 0.1% yeast extract) plates. Mycelium was scraped from these plates and transferred to 1 L flasks containing 250 mL yeast extract, peptone, dextrose broth (YPD – 0.1% yeast extract, 0.1% peptone, 0.3% dextrose). The YPD broth cultures were maintained as still cultures for 8-10 days at 25°C. The fungal strains of *E. resinifera* utilized in the study are listed in Table 1.

2.5.2 Isolation of Mitochondria – Fungal mycelia was collected by vacuum filtration using a Büchner funnel and Whatman® qualitative filter paper. The mycelium was disrupted by grinding with mortar and pestle with the addition of 2 mL of isolation buffer [10 mM Tris-Cl (pH 8.0), 440 mM sucrose, 5 mM ethylene-diamine-tetra-acetic acid (EDTA)] and 1.5 g of acid-washed sand for each 1g of mycelia. The fungal material was ground for about 5-10 minutes until the

mycelia/sand/buffer mixtures formed a slurry. This slurry was transferred to a 25 mL Corex® centrifuge tube (ThermoFisher) and centrifuged for 15 min at 3000 g using a Sorvall® SS-34 fixed angle rotor in a Sorvall® RC-5B Plus centrifuge at 4°C the pellet nuclei, cell debris and sand. The supernatant was transferred to a 25 mL Corex® centrifuge tube (ThermoFisher) and centrifuged at 20 000 g using a Sorvall® SS-34 fixed angle rotor in a Sorvall® RC-5B Plus centrifuge at 4°C for 30 min to pellet the mitochondria.

2.5.3 Mitochondrial DNA extraction – The mitochondrial-enriched pellet was suspended in 3 mL of extraction buffer [100 mM Tris-Cl (pH 8.0), 2% cetyl-trimethylammonium bromide (CTAB), 20 mM EDTA, 1.4 M NaCl] plus 330 µL of 20% sodium dodecyl sulfate (SDS) and nucleic acids were extracted based on Hausner et al. (1992). Briefly the mixture was incubated for 2 hours (or overnight) at 55-65°C and proteins and lipids were removed by adding an equal volume (~3 mL) of chloroform, after mixing the contents of the tube it was centrifuged at 3000 g for 5 min. All centrifugation steps were carried out with a IEC Centra CL2 centrifuge unless stated otherwise. The top aqueous layer was transferred to a new 15 mL centrifuge tube and mixed with 4 µL of RNase A (100 mg/ml; QIAGEN) and incubated in a 55-65°C water bath for 1 h to remove RNA. RNase was removed by addition of Chloroform in a 1:1 ratio in the tube and centrifuged at 3000 g for 20 min. The aqueous layer was transferred to a new tube and mixed with 2.5 volumes of 95% ethanol and placed in the freezer at -20°C for 1 hour. The mixture was then centrifuged at 3000 g for 15 min to pellet the DNA. Supernatant was removed and the DNA pellet was washed with 1 mL of 70% ethanol, the tube was centrifuged again at 3000 g for 5 min

and the ethanol was removed. The DNA pellet was air dried and suspended in 200 µL DNase/RNase-free water and placed in -20°C for storage.

2.5.4 Quantifying DNA – The extracted DNA was quantified with a NanoDrop 2000c UV-Vis Spectrophotometer and the quality was determined on the basis of the 260/280 and 260/230 OD ratio. Quantification was confirmed by gel electrophoresis of 10 µL of the extracted DNA sample on a 1% agarose gel.

2.5.5 Genome sequencing and assembly – The mitochondrial genomic DNA was sequenced and assembled by Génome Québec (Innovation Centre, McGill University). For each sample 75 µL of DNA (~ 1 µg) supplied within an Eppendorf 96-well twin.tec® PCR plate (Cat. No. 951020401) sealed with VWR® aluminum foil (Cat. No. 60941-074), was sent to Génome Québec for Illumina sequencing using the MiSeq platform. The DNAs from different fungal samples were barcoded and combined into a single MiSeq run. The quality of the sequence reads were verified by FastQC (Andrews, 2010). The sequenced reads generated from NGS sequencing were assembled by the a5-miseq-pipeline (Coil et al., 2015) – a MiSeq optimization of the original a5 pipeline (Tritt et al., 2012).

2.5.6 Genome Annotation – The assemblage of the genome by a5-miseq pipeline yielded a set of scaffolds. The scaffolds were sorted out on the basis of the scaffold-size and presence of mitochondrial genes and those scaffolds were used to join and construct an entire uninterrupted mitochondrial genome sequence by using custom python script (available upon request) and

NCBI-blast+ program as well as by manual inspection. The position of the protein-coding genes, rRNAs and tRNAs were identified by MFannot (Beck and Lang, 2010). tRNA genes were identified with tRNAscan-SE (Lowe and Eddy, 1997). Intron-exon junctions within protein-coding and rRNA genes were initially obtained in MFannot and verified by multiple sequence alignments (MSA) of a gene and aligning it to the CDS of the same gene from related species. Sequence alignments were performed with MAFFT (Katoh and Standley, 2013). Sequences were also analyzed with the RNAweasel program (Beck and Lang, 2009) to determine intron types and subtypes. Intron sequences were also examined with the ORF-Finder program (NCBI) to identify possible open reading frames (ORFs). Further, the Smart-BLAST program was used to determine the type of the intron-encoded ORFs. The coordinates of all genes, rRNA, tRNA, introns and intron-encoded ORFs and any other features were annotated using Artemis (Rutherford et al., 2000) and visualized in Circos (Krzywinski et al., 2009).

2.5.7 Genome comparison – The annotated *E. resinifera* mitogenomes were compared for their features by generating multiple sequence alignments (MSA). The MSA allowed for noting SNPs, indels and polymorphisms that relate to the presence and absence of introns. For comparative purposes the mitogenomes of *C. cacaofunesta*, *C. platani* and *C. fimbriata* were also included in the MSA.

A phylogenetic tree, based on a concatenated data set of 13 protein sequences (in alphabetical order: *atp6, atp8, cob, cox1-3, nad1-6*) was constructed in MrBayes (Ronquist et al., 2012) based on an alignment of 41 fungal species generated with MAFFT. Also, a comparative

map was generated in progressiveMauve (Darling et al., 2010) to visualize variations in the genomic architecture.

==========

Chapter 3: Mobile introns in *nad5* genes across Ascomycota and Basidiomycota: an intron landscape

Abstract

Although fungal mitochondrial genomes share a core set of genes they are quite variable in size (11 to >200 kb) in part due to intergenic regions and the presence of introns. Fungal mitochondrial introns are self-splicing introns commonly referred to as group I and group II introns. These introns are different from the nuclear spliceosomal introns as group I and II introns are ribozymes and they can encode proteins that assist in the mobility of group I and II introns. Group I intron typically encode homing endonuclease and group II intron encode reverse transcriptase. These introns are referred to as mobile introns as they can move from intron containing alleles to cognate alleles that lack introns. In this study the NADH dehydrogenase subunit 5 gene (*nad5*) has been examined for members of the Ascomycota and Basidiomycota with regards to the presence of introns. *In silico* analysis was performed on these sequences with regards to identifying introns, intron/exon junctions, intron open reading frames and the data have been summarized in order to generate an "intron landscape" for the *nad5* gene. Intron landscapes are a resource to those involved in annotating mitochondrial genomes and in bioprospecting for endonucleases and reverse transcriptases which have applications in biotechnology. However, intron landscapes also offer an opportunity to examine if there is a biased pattern to the insertion of mobile introns. Mobile introns are probably "parasitic DNAs" that have to balance their own survival against minimizing their impact on the host genome. This

sets up a rather complex scenario and the long term objective of this work is to gain a better understanding of the “evolutionary dynamics” of mobile introns and their impact on fungal mtDNA.

3.1 Introduction

Mitochondria are organelles present in all known eukaryotes that evolved through an endo-symbiosis process during the origin of eukaryotic cell. Mitochondrial genomes are thought to be remnants from the eubacterial endosymbiont that gave rise to this organelle and it went through a genome reduction process in part to function efficiently within a host eukaryotic cell. Fungal mitogenomes typically encode 15 protein coding genes, rRNA genes, and full set of tRNA genes. However, assuming group I and group II introns are of bacterial origins, mtDNAs retained their introns that are mobile elements and can survive as selfish elements in the mitogenome. Mobile introns in fungi are categorized into two types: group I introns and group II introns. Those introns are ribozymes (self-catalytic) that can splice out from the genes transcript, but usually these introns get assistance from intron-encoded proteins to make the process faster (Lang et al., 2007; Hausner, 2012). The self-catalysis of mobile introns is mediated by its specific tertiary structure (Lambowitz and Belfort 2015; Zimmerly and Semper 2015). Starting from a linear transcript group I and group II introns can form complex RNA tertiary structures due to extensive base pair complementary regions and the assistance of maturase activity.

Mitochondrial introns can contain open reading frames (ORFs), which can express endonuclease type enzymes. Intron-encoded proteins (IEPs) can either be homing endonucleases or reverse transcriptases. Group I introns typically contain ORFs for homing endonucleases, which promote intron mobility in the mitogenome; however these IEP can help in intron splicing with their maturase activity (Belfort, 2003). Group II introns usually encode reverse transcriptases but they can also encode for homing endonucleases (Toor and Zimmerly 2002).

Homing endonucleases initiate intron mobility by creating a double-stranded break or in some instances a single-stranded nick at the target site of the double-stranded DNA, which initiate the hosts double strand DNA repair that will repair the gap by a homologous recombination process (Belfort et al., 2002). In fungal mitochondria homing endonucleases can be of two types based on the following amino acids motifs: LAGLIDADG and GIY-YIG (Stoddard, 2011). On the other hand, group II intron encoded reverse transcriptases tend to be multifunctional enzymes that have maturase, DNA binding, endonuclease and reverse transcriptase activity. The process of retro-homing that involves an RNA intermediate is mediated by this enzyme.

The current study focuses on the NADH dehydrogenase subunit 5 (*nad5*) gene and its products. Previous studies investigated the *cob* and *rns* gene (Guha et al., 2017; Hafez et al., 2013), so a *nad* gene is a good candidate for the intron study. The intron landscape provides some insight into the structural variability of the N-terminus of the *nad5* gene, which is intron rich compared to the C-terminal region. The mapping of introns and the characterization of introns and their ORFs provides a useful resource to those interested in mitogenome annotation or ribozymes and other genome editing tools.

3.2 Results and Discussion

3.2.1 The *nad5* Intron Landscape

Of the 186 different *nad5* genes identified from members of the Ascomycota and Basidiomycota, 89 fungi were noted to have introns in their *nad5* gene. The multiple sequence alignment (MSA) of these *nad5* genes when compared against a reference sequence allowed for the identification of exons. The reference sequence for this study is the intron-minus version of the *nad5* gene of *Neurospora crassa* (a model organism for the filamentous fungi); the latter is a well-known model system The MSA revealed that there are 19 intron insertion sites in the *nad5* gene. Different species have different gene architectures with regards to the arrangements of the introns – i.e. Introns in different sites. But some intron sites appear to be more populated with introns within the pool of *nad5* genes examined in this study. For example, 51 species have an intron in the intron-site 717 (Reference nucleotide position 717). The complete list of intron arrangements of 89 fungal species is tabulated in Table 3.1. The recognized intron sites and the number of introns found in those sites are shown in Figure 3.1. The *N. crassa* reference nucleotide sequence size is 2148. The first intron site is at position 248 and the last one is at 1878. Introns are distributed in 19 different intron sites in the *nad5* gene, those are at the following positions: 248, 260, 324, 417, 426, 522, 570, 671, 710, 717, 747, 758, 924, 1000, 1032, 1065, 1152, 1479, and 1878. There was no intron found in the first 247 and the last 270 nucleotides of *nad5* genes. The 3' terminal part of the gene is also found to be variable in the multiple sequence alignment.

A group ID twintron was noted in the *nad5* gene of *Annulohypoxylon stygium* at position 710; the frequency of group II introns was low and only two group II introns were noticed in the *nad5* intron landscape. One group II intron was found in *Podospora anserina* at position 671 and in *Cryphonectria parasitica* at position 324. The rest of the introns reported in the 89 fungal species are group I introns and the total number of group I introns noted was 241 (excluding the twintron).

Table 3.1: Intron landscape of the *nad5* gene. The species having introns in their *nad5* gene are listed here and introns are reported according to their corresponding position of the reference sequence (For the *N. crassa nad5* sequence the intron sequences have been removed).

		248 – 249	260 – 261	324 – 325	417 – 418	426 – 427	522 – 523	570 – 571	671 – 672	710 – 711	717 – 718	747 – 748	758 – 759	924 – 925	1000 – 1001	1032 – 1033	1065 – 1066	1152 – 1153	1479 – 1480	1878 – 1879
	Phase	**2**	**2**	**0**	0	**0**	0	**0**	**2**	**2**	**0**	**0**	**2**	**0**	**1**	**0**	**0**	0	0	0
Species	**Introns**																			
Reference (gi|452883339:27595-27918, 29328-29720, 30858-32288)	0																			
***Neurospora crassa* (KC683708.1|:27595-32288)**	2			IC2							I (derived)									
***Annulohypoxylon stygium* (KF545917.1|:51039-63085)**	5			IC2		ID		IB		ID Twintron				IB						
***Dekkera bruxellensis* (GQ354526.1|:42829-45994)**	1										IB									
***Candida viswanathii* (EF536359.1|:29459-31487)**	1										IB									
***Candida chauliodes* (KF017574.1|:59678-61708)**	1										IB									
***Ceratocystis cacaofunesta* (JX185564.1|:36003-50959)**	8	ID	ID	IC2								IC2	IB	IB	IB		IA			
***Glomerella graminicola* (CM001021.1|:607-3589)**	1										IB									

		248 – 249	260 – 261	324 – 325	417 – 418	426 – 427	522 – 523	570 – 571	671 – 672	710 – 711	717 – 718	747 – 748	758 – 759	924 – 925	1000 – 1001	1032 – 1033	1065 – 1066	1152 – 1153	1479 – 1480	1878 – 1879
	Phase	**2**	**2**	**0**	0	**0**	0	**0**	**2**	**2**	**0**	**0**	**2**	**0**	**1**	**0**	**0**	0	0	0
Species	**Introns**																			
Reference (gi|452883339:27595-27918, 29328-29720, 30858-32288)	0																			
***Epidermophyton floccosum* (AY916130.1|:14469-18691)**	2										IB			IB						
***Fusarium culmorum* (KP827647.1|:45361-48361)**	1										IB									
***Fusarium qerlachii* (KM486533.1|:38230-41230)**	1										IB									
***Fusarium oxysporum* (AY945289.1|:11947-14945)**	1										IB									
***Hirsutella minnesotensis* (KR139916.1|:17525-20923)**	1									ID										
***Madurella mycetomatis* (JQ015302.1|:20521-24612)**	1													IB						
***Meyerozyma guilliermondii* (KC993176.1|:28808-30803)**	1										IB									

		248 – 249	260 – 261	324 – 325	417 – 418	426 – 427	522 – 523	570 – 571	671 – 672	710 – 711	717 – 718	747 – 748	758 – 759	924 – 925	1000 – 1001	1032 – 1033	1065 – 1066	1152 – 1153	1479 – 1480	1878 – 1879
	Phase	**2**	**2**	**0**	0	**0**	0	**0**	**2**	**2**	**0**	**0**	**2**	**0**	**1**	**0**	**0**	0	0	0
Species	**Introns**																			
Reference (gi|452883339:27595-27918, 29328-29720, 30858-32288)	0																			
***Pichia farinosa* (gi|257143744:11489-15502)**	2										IB			IB						
***Pichia kluyveri* (KC993182.1|:14668-18007)**	1													IB						
Peltigera malacea	1													IB						
***Peltigera membranacea* (JN088165.1|:21848-27103)**	3					I (derived)				ID	ID									
***Phaeosphaeria nodorum* (EU053989.1|:32519-35847)**	1			IC2																
Rhynchosporium secalis (KF650575.1|:54249-60424)	2												IB	IB						
Rhynchosporium agropyri	2												I (derived)	I (derived)						

		248 – 249	260 – 261	324 – 325	417 – 418	426 – 427	522 – 523	570 – 571	671 – 672	710 – 711	717 – 718	747 – 748	758 – 759	924 – 925	1000 – 1001	1032 – 1033	1065 – 1066	1152 – 1153	1479 – 1480	1878 – 1879
	Phase	**2**	**2**	**0**	0	**0**	0	**0**	**2**	**2**	**0**	**0**	**2**	**0**	**1**	**0**	**0**	0	0	0
Species	**Introns**																			
Reference (gi|452883339:27595-27918, 29328-29720, 30858-32288)	0																			
***Sclerotinia borealis* (KJ434027.1|:160677-171367)**	4	I (derived)		I (derived)						IB	IB									
***Yarrowia lipolytica* (AJ307410|:33838-38325)**	2										IB			IB						
***Podospora anserina* (gi|294489412:69221-79154)**	4	IA		IC2				IB	GROUP II											
***Ajellomyces dermatitidis* (gi|261187229:580-5054)**	2			IC2										IB						
***Cryphonectria parasitica* (AF456838.1|:553-7016)**	2	IC2		GROUP II																
***Flammulina velutipes* (JF799107.1|:52452-60143)**	2			IC2							IB									
***Ophiostoma novo-ulmi* (CM001753.1)**	1			IC2																

		248 – 249	**260 – 261**	**324 – 325**	**417 – 418**	**426 – 427**	**522 – 523**	**570 – 571**	**671 – 672**	**710 – 711**	**717 – 718**	**747 – 748**	**758 – 759**	**924 – 925**	**1000 – 1001**	**1032 – 1033**	**1065 – 1066**	**1152 – 1153**	**1479 – 1480**	**1878 – 1879**
	Phase	**2**	**2**	**0**	0	**0**	0	**0**	**2**	**2**	**0**	**0**	**2**	**0**	**1**	**0**	**0**	0	0	0
Species	**Introns**																			
Reference (gi|452883339:27595-27918, 29328-29720, 30858-32288)	0																			
Moniliophthora	2			I (derived)												I (derived)				
Ganoderma sinense (KF673550.1|:14880-19263)	2			I (derived)	I (derived)															
Ganoderma lucidum (HF570115.2|:34349-38732)	2			IC2	ID															
Ganoderma meredithae (KP410262.1|:14716-19031)	1				I (derived)															
Phlebia radiata (HE613568.1|:66153-76178)	4			I (derived)	I (derived)		I (derived)				I (derived)									
Agaricus bisporus (JX271275.1|:1-7859)	3				IB						ID				IB					
Microbotryum lychnidis (KC285586.1|:41281-46087)	3				IB						IB				IB					

		248 – 249	**260 – 261**	**324 – 325**	**417 – 418**	**426 – 427**	**522 – 523**	**570 – 571**	**671 – 672**	**710 – 711**	**717 – 718**	**747 – 748**	**758 – 759**	**924 – 925**	**1000 – 1001**	**1032 – 1033**	**1065 – 1066**	**1152 – 1153**	**1479 – 1480**	**1878 – 1879**
	Phase	**2**	**2**	**0**	0	**0**	0	**0**	**2**	**2**	**0**	**0**	**2**	**0**	**1**	**0**	**0**	0	0	0
Species	**Introns**																			
Reference (gi|452883339:27595-27918, 29328-29720, 30858-32288)	0																			
Sporisorium reilianum (FQ311469.1|:7888-13104)	2				IB						IB									
Tilletia indica (DQ993184.1|:25527-30010)	1				IB						IB									
Tilletia walkeri (EF536375.1|:24284-28767)	2				IB						IB									
Claviceps purpurea (FO082257.1|:14418-16632)	2				IB															
Pyronema omphalodes (KU707476.1|:138647-149672)	4				IB							I (derived)	IB		IB					
Parasitella parasitica (KM382275.1|:59785-63811)	1										IB									
Phycomyces blakesleeanus (KR809878.1|:48111-51046)	1										IB									

		248 – 249	**260 – 261**	**324 – 325**	**417 – 418**	**426 – 427**	**522 – 523**	**570 – 571**	**671 – 672**	**710 – 711**	**717 – 718**	**747 – 748**	**758 – 759**	**924 – 925**	**1000 – 1001**	**1032 – 1033**	**1065 – 1066**	**1152 – 1153**	**1479 – 1480**	**1878 – 1879**
	Phase	**2**	**2**	**0**	0	**0**	0	**0**	**2**	**2**	**0**	**0**	**2**	**0**	**1**	**0**	**0**	0	0	0
Species	**Introns**																			
Reference (gi|452883339:27595-27918, 29328-29720, 30858-32288)	0																			
Heterobasidion irregulare (KF957635.1|:39025-42400)	1										IB									
Ustilago maydis (DQ157700.1|:46103-49221)	1										IB									
Jaminaea angkoriensis (KC628747.1|:23908-26967)	1										IB									
Setosphaeria_turcica-NY001	12	ID		IC2	IB	ID		IC2		IB	ID	IB	IB		IB		I (derived)		IC2	
Pleomassaria_siparia	7	I (derived)		IC2	IB					ID	IB			IB	IB					
Massarina_eburnea-CBS_473.64	12	ID		IC2	IB			I (derived)		ID	IB	IC2	I (derived)	IB	IB			IC2	IC2	
Phyllosticta_citriasiana	3	ID		IC2	IB															

		248 – 249	**260 – 261**	**324 – 325**	**417 – 418**	**426 – 427**	**522 – 523**	**570 – 571**	**671 – 672**	**710 – 711**	**717 – 718**	**747 – 748**	**758 – 759**	**924 – 925**	**1000 – 1001**	**1032 – 1033**	**1065 – 1066**	**1152 – 1153**	**1479 – 1480**	**1878 – 1879**
	Phase	**2**	**2**	**0**	0	**0**	0	**0**	**2**	**2**	**0**	**0**	**2**	**0**	**1**	**0**	**0**	0	0	0
Species	**Introns**																			
Reference (gi|452883339:27595-27918, 29328-29720, 30858-32288)	0																			
Pseudovirgaria_hyperparasitica-CBS_121739	6			I (derived)	IB					ID	IB			IB			IA			
Septoria_musiva-SO2202	2			IC2									IB							
Teratosphaeria_nubilosa-CBS_116005	6			I (derived)				I (derived)	I (derived)	I (derived)	I (derived)		I (derived)	I (derived)						
Bimuria_novae-zelandiae-CBS_107.79	4			I (derived)	IB			IB			IB									
Mytilinidion_resinicola-CBS_304.34	1			IC2																
Lepidopterella_palustris	2			IC2				IB												
Clathrospora_elynae-CBS_161.51	3			IC2											IB				IC2	

		248 – 249	**260 – 261**	**324 – 325**	**417 – 418**	**426 – 427**	**522 – 523**	**570 – 571**	**671 – 672**	**710 – 711**	**717 – 718**	**747 – 748**	**758 – 759**	**924 – 925**	**1000 – 1001**	**1032 – 1033**	**1065 – 1066**	**1152 – 1153**	**1479 – 1480**	**1878 – 1879**
	Phase	**2**	**2**	**0**	0	**0**	0	**0**	**2**	**2**	**0**	**0**	**2**	**0**	**1**	**0**	**0**	0	0	0
Species	**Introns**																			
Reference (gi|452883339:27595-27918, 29328-29720, 30858-32288)	0																			
Melanomma_pulvis-pyrius	1			IC2																
Lindgomyces_ingoldianus-ATCC_200398	6			IC2	IB					ID	IB			IB	IB					
Lizonia_empirigonia-CBS_542.76	4			IC2				IB			IB			IB						
Ophiobolus_disseminans-CBS_113818	2			IC2											IB					
Ampelomyces_quisqualis-HMLAC05119	3			IC2										IB					IC2	
Decorospora_gaudefroyi	5							IB			ID			IB	IB				IC2	
Lophiostoma_macrostomum	2										IB				IB					

		248 – 249	260 – 261	324 – 325	417 – 418	426 – 427	522 – 523	570 – 571	671 – 672	710 – 711	717 – 718	747 – 748	758 – 759	924 – 925	1000 – 1001	1032 – 1033	1065 – 1066	1152 – 1153	1479 – 1480	1878 – 1879
	Phase	**2**	**2**	**0**	0	**0**	0	**0**	**2**	**2**	**0**	**0**	**2**	**0**	**1**	**0**	**0**	0	0	0
Species	**Introns**																			
Reference (gi|452883339:27595-27918, 29328-29720, 30858-32288)	0																			
Paraconiothyrium_sporulosum-AP3s5-JAC2a	2							IB			IB									
Elsinoe_ampelina-CECT_20119	1										IB									
Tothia_fuscella-CBS_130266	1										IB									
Cercospora_zeae-maydis	2										IB			IB						
Corynespora_cassiicola-CCP	1										IB									
Macroventuria_anomochaeta-CBS_525.71	3			IC2							IB			IB						
Pyrenochaeta_sp.-DS3sAY3a	1													IB						

		248 – 249	**260 – 261**	**324 – 325**	**417 – 418**	**426 – 427**	**522 – 523**	**570 – 571**	**671 – 672**	**710 – 711**	**717 – 718**	**747 – 748**	**758 – 759**	**924 – 925**	**1000 – 1001**	**1032 – 1033**	**1065 – 1066**	**1152 – 1153**	**1479 – 1480**	**1878 – 1879**
	Phase	**2**	**2**	**0**	0	**0**	0	**0**	**2**	**2**	**0**	**0**	**2**	**0**	**1**	**0**	**0**	0	0	0
Species	**Introns**																			
Reference (gi|452883339:27595-27918, 29328-29720, 30858-32288)	0																			
Setomelanomma_holmii-CBS_110217	1													IB						
Westerdykella_ornata-CBS_379.55	1													IB						
Lophium_mytilinum-CBS_269.34	1													IB						
Mycosphaerella_fijiensis	1													IB						
Clohesyomyces_aquaticus	1																		IC2	
Coccomyces_strobi-CBS_202.91	8	ID				ID				ID	IB			IB		IB		IC2		
Lindra_thalassiae-JK4322	7	I (derived)		IC2				IB			IB		IB	IB						I (derived)

		248 – 249	260 – 261	324 – 325	417 – 418	426 – 427	522 – 523	570 – 571	671 – 672	710 – 711	717 – 718	747 – 748	758 – 759	924 – 925	1000 – 1001	1032 – 1033	1065 – 1066	1152 – 1153	1479 – 1480	1878 – 1879
	Phase	**2**	**2**	**0**	0	**0**	0	**0**	**2**	**2**	**0**	**0**	**2**	**0**	**1**	**0**	**0**	0	0	0
Species	**Introns**																			
Reference (gi|452883339:27595-27918, 29328-29720, 30858-32288)	0																			
Magnaporthiopsis_poae-ATCC_64411	4	ID		IC2		ID		IB												
Caloscypha_fulgens-ATCC_42695	10			I (derived)		ID		IB		I (derived)	ID	IC2	I (derived)	I (derived)	IB				I (derived)	
Lollipopaia_minuta-P26-CBS_116597	4			IC2							IB			IB	IB					
Bulgaria_inquinans-CBS118.31	8			I (derived)		ID		IB		ID		IC2		IB	IB			IC2		
Loramyces_macrosporus-CBS235.53	5			IC2		ID					IB		IB	IB						
Morchella_importuna-SCYDJ1-A1	6				IB	IB		IB			IB		IB		IB					
Aspergillus_carbonarius-ITEM_5010	2							IB					IB							

		248 – 249	**260 – 261**	**324 – 325**	**417 – 418**	**426 – 427**	**522 – 523**	**570 – 571**	**671 – 672**	**710 – 711**	**717 – 718**	**747 – 748**	**758 – 759**	**924 – 925**	**1000 – 1001**	**1032 – 1033**	**1065 – 1066**	**1152 – 1153**	**1479 – 1480**	**1878 – 1879**
	Phase	**2**	**2**	**0**	0	**0**	0	**0**	**2**	**2**	**0**	**0**	**2**	**0**	**1**	**0**	**0**	0	0	0
Species	**Introns**																			
Reference (gi|452883339:27595-27918, 29328-29720, 30858-32288)	0																			
Colletotrichum_caudatum-CBS131602	1										IB									
Thelebolus_microsporus-ATCC_90970	1										IB									
Choiromyces_venosus-120613-1	3							IB			IB		IB						IC2	
Microdochium_bolleyi-J235TASD1	1													IB						
Gyromitra_esculenta-CBS101906	1																		IC2	
Total (89)	244	11	1	37	19	9	1	17	1	13	51	6	14	30	16	2	3	3	9	1

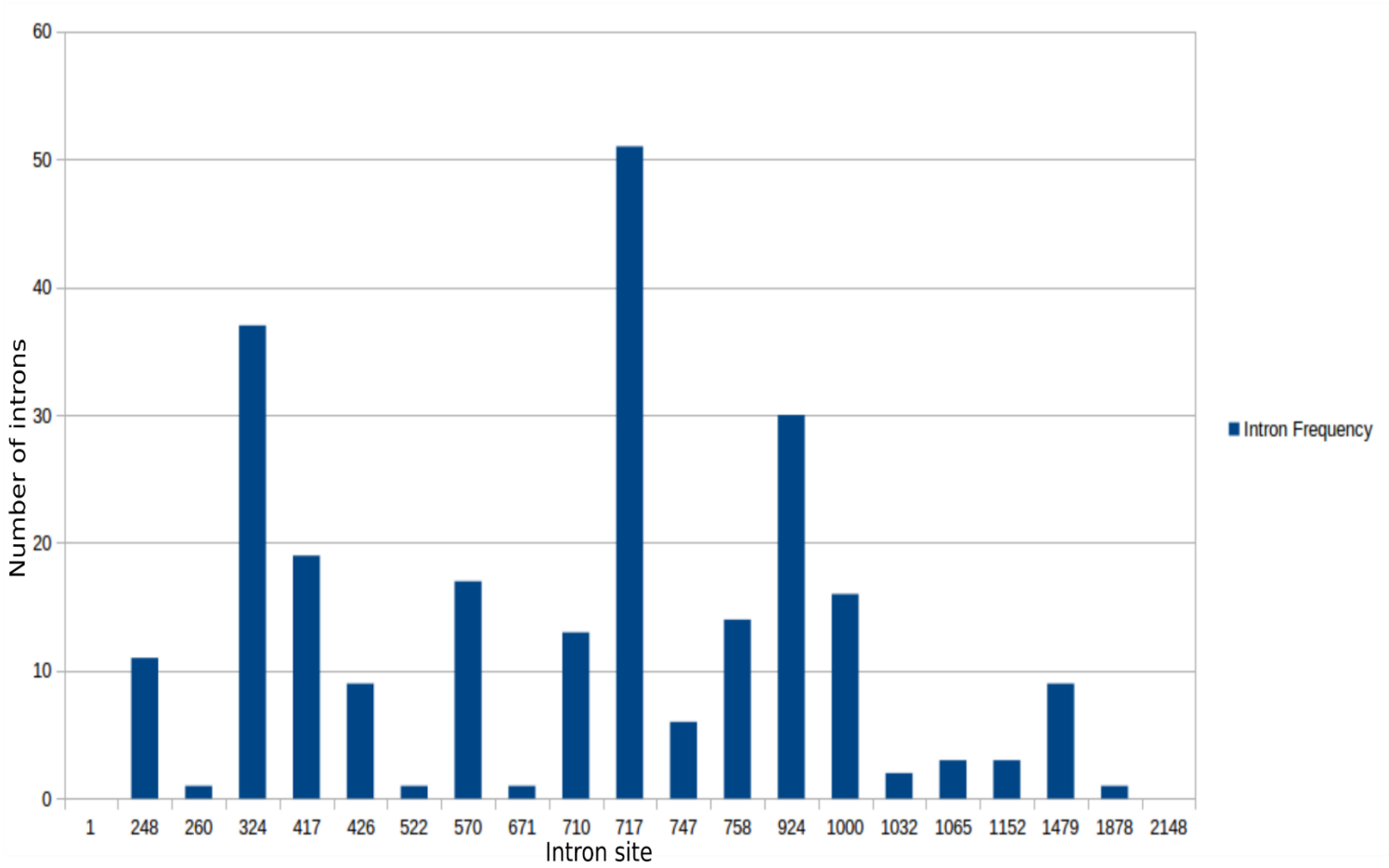

60
50
40
30
20
10
0
Number of introns
1
248
260
324
417
426
522
570
671
710
717
747
758
924
1000
1032
1065
1152
1479
1878
2148
Intron site
Intron Frequency

Figure 3.1: Number of introns found at the intron sites in different species within the *nad5* gene. The intron sites (in reference to *N. crassa*) are given in the x-axis, and the bars are showing the number of introns found.

In fungal mtDNAs gene as well as genome sizes, are variable because of the number of introns in the mitochondrial genes (Guha et al., 2017; Abboud et al. 2018; Hafez et al., 2013). From the intron landscape it was noted that there are certain positions within the gene that are more favourable for the presence of introns and some intron sites are more populated than others. The presence or absence of a specific intron can be explained by Goddard and Burt's (1999) model known as homing endonuclease life-cycle, a model that argues that introns and their HEGs are neutral elements that rapidly accumulate mutations due to the lack of selection. Therefore, these elements can be quickly lost and in order to persist they have to continuously move into empty homing sites or gain a function that is beneficial to the host. Although one would expect that most intron insertion sites should harbour introns at similar frequencies, instead one tends to observe biases. This might hint at the possibilities that some intron/HEG combinations are more efficient in mobility, outpacing "drift" that is eroding the intron ORF; or some introns may have some benefit and thus are maintained due to adaptive selection.

A recent study based on compiling introns that insert within the *cytb* gene provided an elaboration of the Goddard and Burt model introducing the concept of founder intron that initially inserts into a new site; however as the intron ORF becomes optimized for improved expression via "core creep" (see Edgell et al. 2011) the intron starts to be more invasive populating intronless cognate alleles and potentially moving into ectopic sites (Guha et al. 2017). However, this model also suggests that introns and HEGs are subject to drift and erode due to the accumulation of mutations due to lack of selection. Horizontal gene transfer (HGT) is also a factor influencing intron arrangement in mitochondrial genome (Wu & Hao, 2014), as one

observes frequently that unrelated fungi share introns within the same insertion sites. HGT is a key component in the long term persistence of mobile introns as they have to continuously invade new sites in order to outpace drift.

3.2.2 Structure analysis of the *nad5* encoded protein

It was noticed in the MSA that the tail part (3' end) of the *nad5* gene is quite variable among the different *nad5* genes. The alignment was tested for its sequence variability or disorder in GeneSilico MetaDisorder Server and it was found that the last 500 bps have increased sequence disorder and the last 100 bps show very high variability (Figure 3.2). Further the amino acid sequence of the reference *nad5* gene-product (ND5) was analyzed for membrane affinity in TMHMM program (Krogh et al., 2001). The result showed there are 16 transmembrane domains in the ND5 protein (Figure 3.3). The 3D structure of the NADH dehydrogenase subunit 5 protein (*nad5* gene product) was generated in the Phyre2 program (Figure 3.4) which used the crystal structure of the NADH oxidoreductase I from *E. coli* (PDB ID: 3RKO) as a template against the query (ND5 of *N. crassa*). The query coverage was 85% and the confidence score in Phyre2 was 100%. The structure analysis also showed the transmembrane helices and a long arm of alpha-helix (supernumerary unit) that interact with the ND4 and ND2 protein involved in the electron transport chain (Vinothkumar et al., 2014). The long arm of ND5 is the tail part (C-terminal) of the protein and that could be the cause of its tail part variability. Our results are in accordance to the previous structural study of the mammalian respiratory complex I (Vinothkumar et al., 2014). Data showed that the N-terminus coding sequences harboured more

intron insertion sites compared to the C-terminal coding part of the gene. Although, there were intron insertion sites in the C-terminal coding sequences but they appear to be less population with introns. HEGs due to their requirement for long target sites (14 to 40 nt) require conserved sequences in order to proliferate, therefore segments of genes under functional constrains (or sequence conservation) would be targeted more frequently by mobile introns.

3.3 Concluding comments

The intron landscape provides a resource for other workers with regards to mitogenome annotations and some insight with regards to biases that may exist in reference to intron insertion sites and gene architecture. The study also uncovered on novel "side by side" twintron that should be further examined in future studies as it may provide some insights on how these complex elements are formed and how they might be spliced in order to avoid interfering with the function of the host gene.

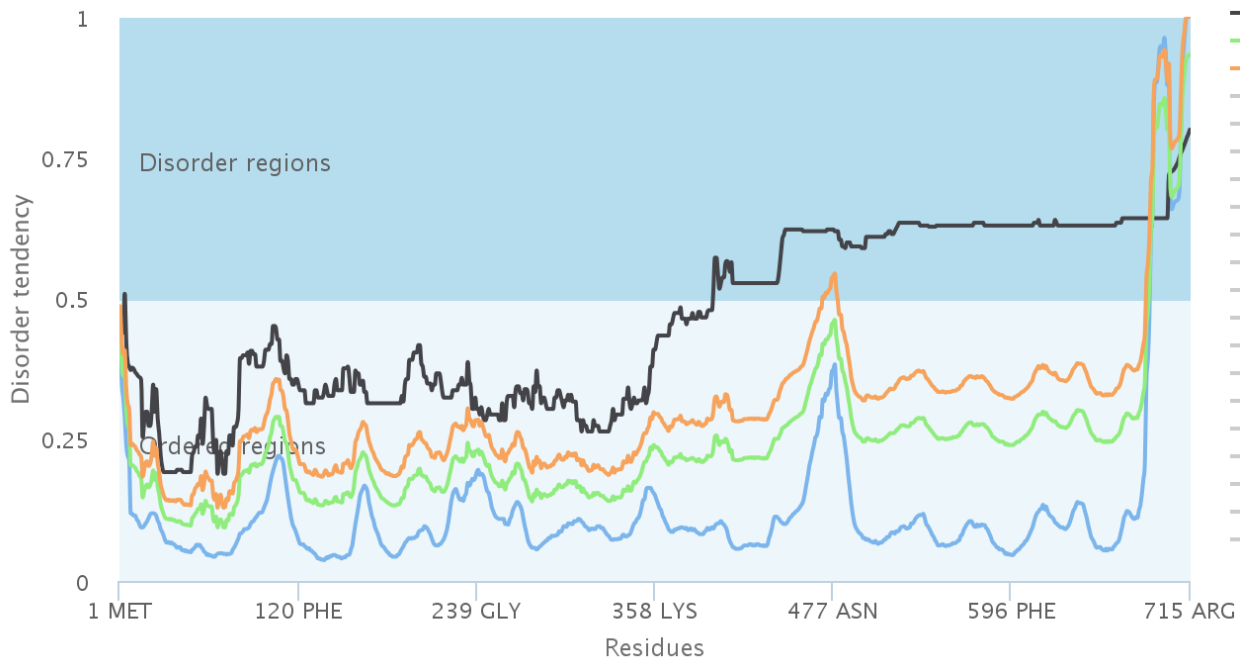

1
0.75
0.5
0.25
0
Disorder tendency
Disorder regions
Ordered regions
1 MET
120 PHE
239 GLY
358 LYS
477 ASN
596 PHE
715 ARG
Residues

Figure 3.2: On the basis of the sequence alignment of *nad5* exon sequences, sequence orderedness was measured at the amino acid level. A peptide that may not have a well-defined or firm 3D structure (due to interaction with other molecules) can be determined by meta-method including utilizing evolutionary information, energy functions, and various statistical and machine learning methods at GeneSilico MetaDisorder Server. The graph showed that the C-terminal of ND5 is in a disordered region. Different colored lines are representing different methods of calculating disorder (blue – MetaDisorder, black – MetaDisorder3D, green – MetaDisorderMD, red – MetaDisorderMD2).

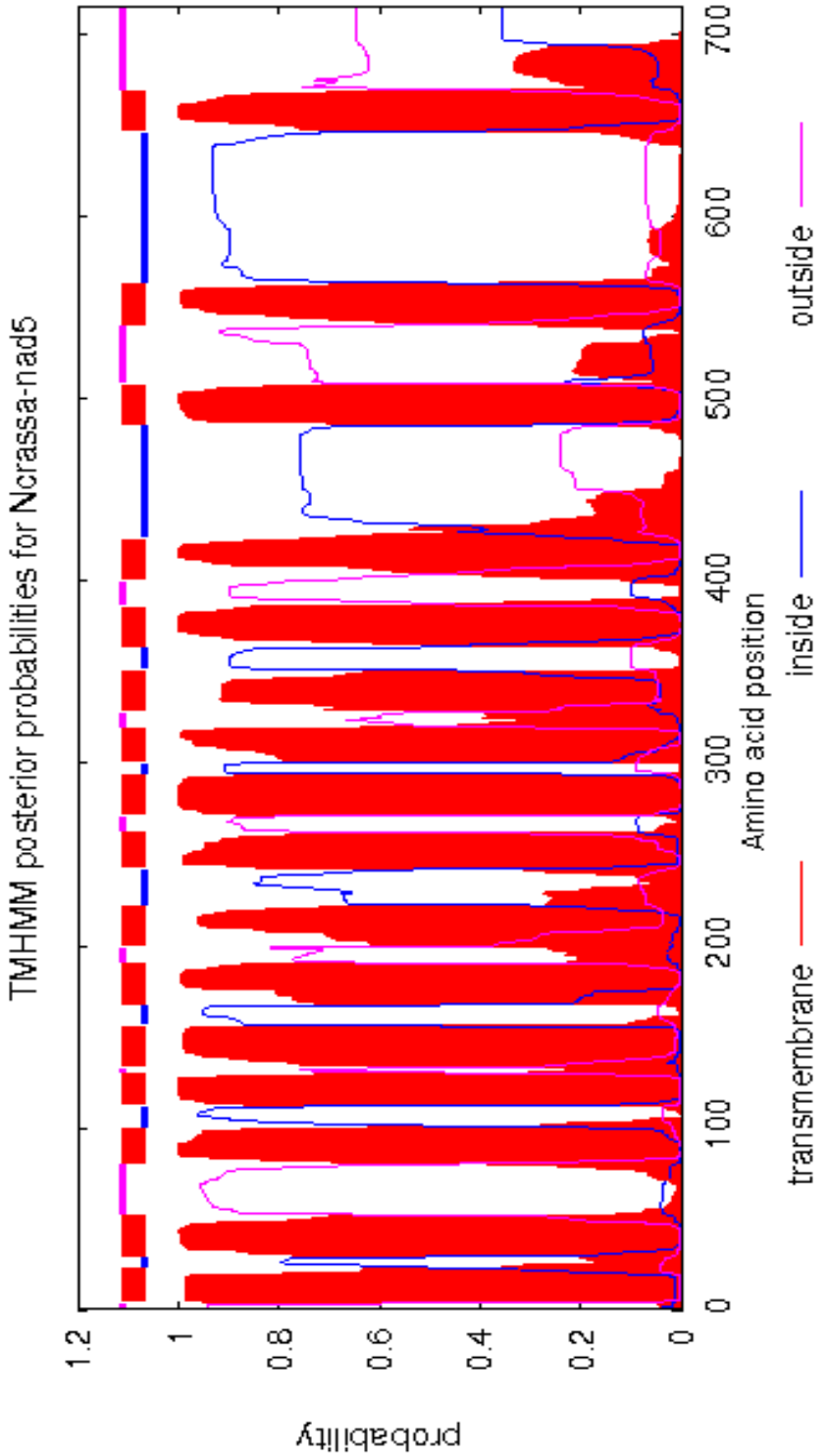

TMHMM posterior probabilities for Ncrassa-nad5
probability
1.2
1
0.8
0.6
0.4
0.2
0
0
100
200
300
400
500
600
700
Amino acid position
transmembrane
inside
outside

Figure 3.3: The transmembrane domains of *N. crassa* ND5 protein. The plot shows the posterior probabilities of inside/outside/trans membrane helix which is obtained by calculating the total probability that a residue sits in a helix, summed over all possible paths through the model, and generated at by TMHMM at ExPASy. We can observe 16 potential transmembrane domain in the ND5 protein where the first 12 are closely situated to each other and the last 3 are separated from each other by about 100 amino acids.

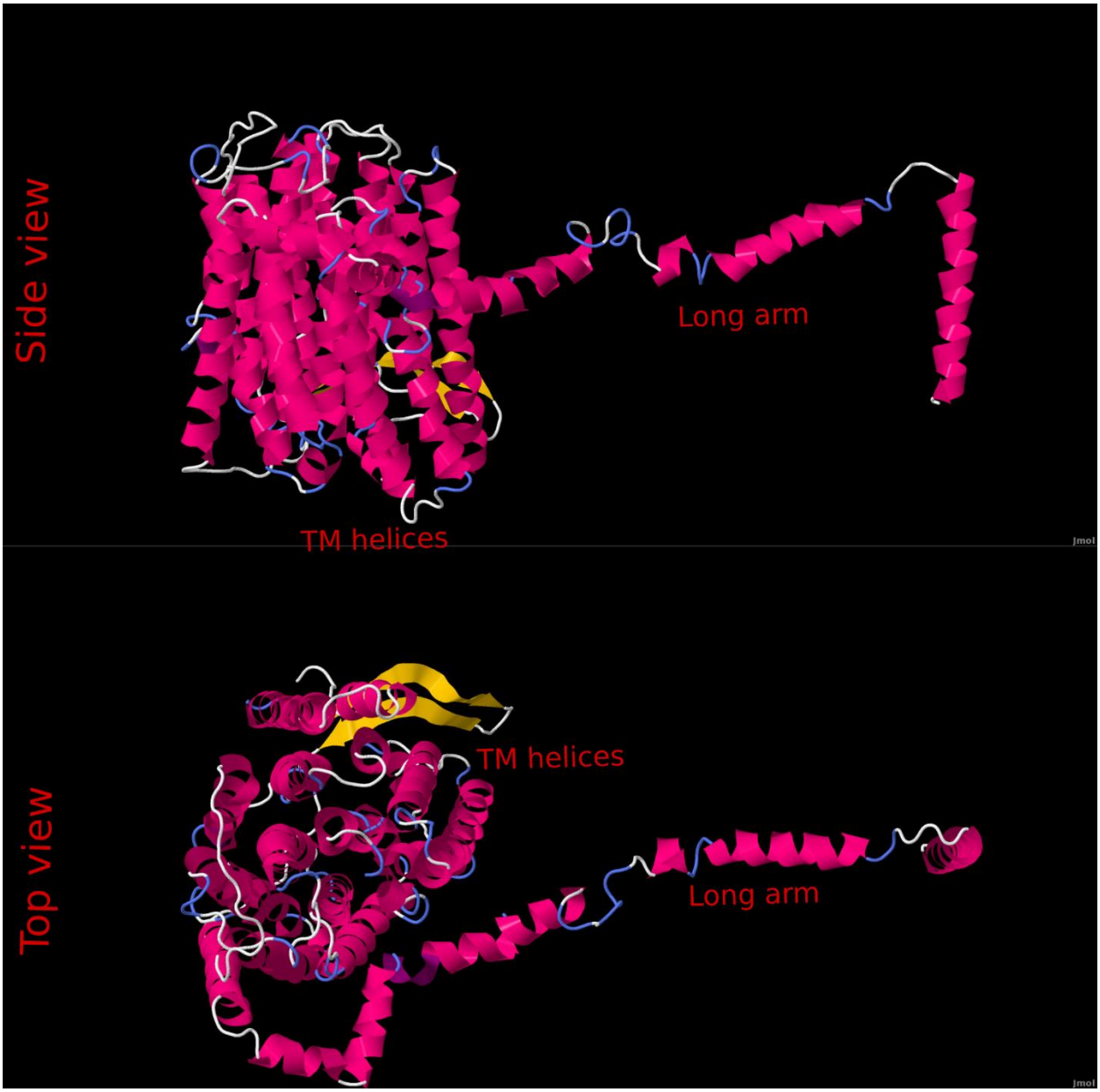

Side view
Long arm
TM helices
Top view
TM helices
Long arm

Figure 3.4: Predicted structure of ND5 protein of *N. crassa* showing transmembrane domains and the long arm with 3 alpha helices. The TM helices are found to be highly conserved among the ascomycetes and basidiomycetes fungi, but the long arm (supernumerary domain) was noted to be variable and disordered probably due to the arm's interacting with other proteins in the neighborhood. These interactions might be electrostatic (or indirect) in nature and thus not under evolutionary constraint. The 3D structure of ND5 was generated by Phyre2 which uses a profile-profile alignment using HMM algorithm for generating a 3D structure. The structure is visualized in Jmol (http://jmol.org/).

3.4 Materials and Methods

3.4.1 Collection of data

The NCBI genome database provides an information retrieval system for collecting genomes by browsing using the names of organisms. Selecting “organelle” and specifying “fungi” and “mitochondria” type provide a list of fungal mitochondrial genomes. The *nad5* gene sequences available in NCBI genome database were used to BLAST against the NCBI non-redundant database to extract other *nad5* genes available in that database. We have also collected some unannotated fungal genomes from the JGI database and annotated them with MFannot and collected the *nad5* gene from these newly annotated mtDNAs (from “*Setosphaeria turcica* NY001” to the end of the **Table 3.1**, 44 species listed that were annotated).

3.4.2 Multiple sequence alignment

A total of 186 nucleotide sequences representing the *nad5* gene from different fungal species were aligned with the MAFFT multiple sequence alignment program (Katoh and Standley, 2013). The E-INS-I algorithm was selected within MAFFT as it allows for short conserved sequences separated by long gaps. This particular feature facilitates the alignment of genes-without-introns with genes-with-introns; thus allowing for the identification of the CDS counterparts among the various genes being examined. This alignment strategy defines the position of introns within the context of the alignment. Introns were removed from the *nad5* gene of *Neurospora crassa* and this sequences was used it in the alignment as a reference sequence to

identify intron position in other sequences. Therefore, introns could be named according to intron insertions sites utilizing *N. crassa* nad5 CDS as the reference.

3.4.3 Intron Landscaping

The aligned sequences were visualized and edited (where necessary) in Jalview (Waterhouse et al., 2009). The position of the introns in every sequence were determined from the MSA and tabulated according to the nucleotide position of the *nad5* intronless sequence from *Neurospora crassa*. A histogram was generated to show the frequency of occurrence of a particular intron at a position within the gene among the sampled species. Further the intron sequences were submitted to the RNAweasel program online with default settings (http://megasun.bch.umontreal.ca/cgi-bin/RNAweasel/RNAweaselInterface.pl) to determine intron type and subtype. The ORFfinder program (https://www.ncbi.nlm.nih.gov/orffinder/) was used to identify the presence of any intron-encoded ORF and the putative nature of the ORF.

3.4.4 Structure analysis

The three dimensional structure of the *nad5* gene-product, the NADH dehydrogenase subunit 5 protein, was determined by the Phyre 2.0 program (Kelley et al., 2015). The transmembrane domain and disorder region of the protein was predicted by using the TMHMM (Krogh et al., 2001) and GeneSilico servers (Kurowski and Bujnicki, 2003).

General Conclusion

The current study examined some of the recent concepts of mobile introns, intron encoded ORFs and fungal mitochondrial genomic variations. It also briefly reviewed recent advancement in fungal mitochondrial genome sequencing, assembly and annotation technologies. Sequencing of the mitochondrial genome of the *Endoconidiophora resinifera* (=*Ceratocystis resinifera*), along with its assembly and annotation of the genes and mobile introns and the intron-encoded ORFs revealed very large mitogenomes. Comparative analysis of the mitochondrial genomes of four different strains of *E. resinifera* and other *Ceratocystis* spp. revealed that introns can make a significant contribution towards genome size and gene architecture. This study also raises questions with regards to the causes that may result in some mitochondrial genomes expanding due to gaining introns or why some mitogenomes are streamlined due to minimal numbers of introns and small intergenic regions. Clearly the latter could be the subject for future studies. In addition, an intron landscape was generated for the *nad5* gene from different species from the Ascomycota and Basidiomycota. The major findings of the current study are:

- Introns are found in the same site in different species. Intron hotspots are usually situated in the first half of a gene. This may in part relate to sequences that are under functional constraint thus more conserved. It may also be indicative of a potential avenue of intron-loss. The incorporation of cDNAs generated by reverse transcriptase (RT) activity. As RTs are known to possess low processivity and thus terminate early; 3' terminal regions are more prone to intron loss due to recombination events that replace them with partial cDNAs.

- Mitochondrial genome of *E. resinifera* is the largest in Ascomycota, also it has the largest *cox1* gene reported so far.

- Introns are the main factor in genome size expansion and genome size variation in fungal mitochondria.
- An updated model for the HEG life cycle including a component that allows for introns to co-operate with each other, thus enhancing their chance for persistence and thus the avoidance of extinction.

This is the first study which examines the mitochondrial genome for members of the genus *Endoconidiophora*. Based on comparisons with related *Ceratocystis* species it is clear that the mitochondrial genome size and the size of the *cox1* gene for *E. resinifera* are setting new milestones for future fungal mitochondrial genome studies. The annotation of the miogenome of *E. resinifera* will be a great resource for further research on fungal mitochondrial genomics. Future work should examine the transcriptome and proteome for the mitochondria of this fungus in order to examine the level of expression of introns and their encoded proteins.

A compilation of the introns in *nad5* genes from Ascomycota and Basidiomycota, plus the introns and twintron from *E. resinifera* provide a framework to modulate the current state of the HEG/intron mobility model (as initially defined by Goddard and Burt, 1999). In essence there appears to be some circumstantial evidence that some introns could co-operate with one

another and potentially in a hypercycle-like mutualistic relationship providing some insight why some families of HEGs are more successful compared to others.

References

Abboud TG, Zubaer A, Wai A, Hausner G. The complete mitochondrial genome of the Dutch elm disease fungus *Ophiostoma novo-ulmi* subsp *novo-ulmi*. Can J Microbiol. 2018 May;64(5):339-348. PubMed PMID: 29401406.

Aguileta G, de Vienne DM, Ross ON, Hood ME, Giraud T, et al. High variability of mitochondrial gene order among fungi. Genome Biol Evol. 2014 Feb;6(2):451-65. PubMed PMID: 24504088; PubMed Central PMCID: PMC3942027.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990 Oct 5;215(3):403-10. PubMed PMID: 2231712.

Ambrosio AB, do Nascimento LC, Oliveira BV, Teixeira PJ, Tiburcio RA, et al. Global analyses of *Ceratocystis cacaofunesta* mitochondria: from genome to proteome. BMC Genomics. 2013 Feb 11;14:91. PubMed PMID: 23394930; PubMed Central PMCID: PMC3605234.

Andrews S. FastQC: a quality control tool for high throughput sequence data. Babraham Bioinformatics. 2010; http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Anziano PQ, Perlman PS, Lang BF, Wolf K. The mitochondrial genome of the fission yeast *Schizosaccharomyces pombe*: I isolation and physical mapping of mitochondrial DNA. Curr Genet. 1983 Jul;7(4):273-84. PubMed PMID: 24173337.

Baidyaroy D, Hausner G, Bertrand H. *In vivo* conformation and replication intermediates of circular mitochondrial plasmids in *Neurospora* and *Cryphonectria* parasitica. Fungal Biol. 2012 Aug;116(8):919-31. PubMed PMID: 22862920.

Baidyaroy D, Huber DH, Fulbright DW, Bertrand H. Transmissible mitochondrial hypovirulence in a natural population of *Cryphonectria parasitica*. Mol Plant Microbe Interact. 2000 Jan;13(1):88-95. PubMed PMID: 10656589.

Baidyaroy D, Hausner G, Hafez M, Michel F, Fulbright DW, et al. A 971-bp insertion in the rns gene is associated with mitochondrial hypovirulence in a strain of *Cryphonectria parasitica* isolated from nature. Fungal Genet Biol. 2011 Aug;48(8):775-83. PubMed PMID: 21601643.

Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, et al. SPAdes: a new genome assembly algorithm and its applications to single-cell sequencing. J Comput Biol. 2012 May;19(5):455-77. PubMed PMID: 22506599; PubMed Central PMCID: PMC3342519.

Beck N, Lang BF. RNAweasel, a webserver for identification of mitochondrial, structured RNAs. 2009; http://megasun.bch.umontreal.ca/cgi-bin/RNAweasel/RNAweaselInterface.pl.

Beck N, Lang BF. MFannot, organelle genome annotation webserver. 2010; http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl.

Begel O, Boulay J, Albert B, Dufour E, Sainsard-Chanet A. Mitochondrial group II introns, cytochrome c oxidase, and senescence in *Podospora anserina*. Mol Cell Biol. 1999 Jun;19(6):4093-100. PubMed PMID: 10330149; PubMed Central PMCID: PMC104368.

Beier S, Thiel T, Münch T, Scholz U, Mascher M. MISA-web: a web server for microsatellite prediction. Bioinformatics. 2017 Aug 15;33(16):2583-2585. PubMed PMID: 28398459; PubMed Central PMCID: PMC5870701.

Belfort M. Two for the price of one: a bifunctional intron-encoded DNA endonuclease-RNA maturase. Genes Dev. 2003 Dec 1;17(23):2860-3. PubMed PMID: 14665667.

Belfort M. Mobile self-splicing introns and inteins as environmental sensors. Curr Opin Microbiol. 2017 Aug;38:51-58. PubMed PMID: 28482231; NIHMSID: NIHMS874341; PubMed Central PMCID: PMC5671916.

Belfort M, Derbyshire V, Parker MM, Cousineau B, Lambowitz AM. Mobile introns: pathways and proteins. In: Craig NL, Craigie R, Gellert M, Lambowitz AM, editors. Mobile DNA II. Washington DC: American Society for Microbiology Press; 2002 pp. 761-783.

Belfort M, Perlman PS. Mechanisms of intron mobility. J Biol Chem. 1995 Dec 22;270(51):30237-40. PubMed PMID: 8530436.

Bendich AJ. Reaching for the ring: the study of mitochondrial genome structure. Curr Genet. 1993 Oct;24(4):279-90. PubMed PMID: 8252636.

Benson G. Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res. 1999 Jan 15;27(2):573-80. PubMed PMID: 9862982; PubMed Central PMCID: PMC148217.

Bernt M, Donath A, Jühling F, Externbrink F, Florentz C, et al. MITOS: improved de novo metazoan mitochondrial genome annotation. Mol Phylogenet Evol. 2013 Nov;69(2):313-9. PubMed PMID: 22982435.

Bilto IM, Guha TK, Wai A, Hausner G. Three new active members of the I-OnuI family of homing endonucleases. Can J Microbiol. 2017 Aug;63(8):671-681. PubMed PMID: 28414922.

Blouin MS. Molecular prospecting for cryptic species of nematodes: mitochondrial DNA versus internal transcribed spacer. Int J Parasitol. 2002 May;32(5):527-31. PubMed PMID: 11943225.

Bonocora RP, Shub DA. A likely pathway for formation of mobile group I introns. Curr Biol. 2009 Feb 10;19(3):223-8. PubMed PMID: 19200727; NIHMSID: NIHMS95610; PubMed Central PMCID: PMC2856452.

Brandvain Y, Wade MJ. The functional transfer of genes from the mitochondria to the nucleus: the effects of selection, mutation, population size and rate of self-fertilization. Genetics. 2009 Aug;182(4):1129-39. PubMed PMID: 19448273; PubMed Central PMCID: PMC2728854.

Brocks JJ, Logan GA, Buick R, Summons RE. Archean molecular fossils and the early rise of eukaryotes. Science. 1999;285:1033-1036.

Buermans HP, den Dunnen JT. Next generation sequencing technology: Advances and applications. Biochim Biophys Acta. 2014 Oct;1842(10):1932-1941. PubMed PMID: 24995601.

Cech TR. Self-splicing of group I introns. Annu Rev Biochem. 1990;59:543-68. PubMed PMID: 2197983.

Cech TR, Damberger SH, Gutell RR. Representation of the secondary and tertiary structure of group I introns. Nat Struct Biol. 1994 May;1(5):273-80. PubMed PMID: 7545072.

Chevalier BS, Stoddard BL. Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. Nucleic Acids Res. 2001 Sep 15;29(18):3757-74. PubMed PMID: 11557808; PubMed Central PMCID: PMC55915.

Chevreux B, Wetter T, Suhai S. Genome Sequence Assembly Using Trace Signals and Additional Sequence Information. Proceedings of the German Conference on Bioinformatics (GCB);1999 p. 45-56.

Coil D, Jospin G, Darling AE. A5-miseq: an updated pipeline to assemble microbial genomes from Illumina MiSeq data. Bioinformatics. 2015 Feb 15;31(4):587-9. PubMed PMID: 25338718.

Dabbagh N, Bennett MS, Triemer RE, Preisfeld A. Chloroplast genome expansion by intron multiplication in the basal psychrophilic euglenoid *Eutreptiella pomquetensis*. PeerJ. 2017;5:e3725. PubMed PMID: 28852596; PubMed Central PMCID: PMC5572947.

Daniels DL, Michels WJ Jr, Pyle AM. Two competing pathways for self-splicing by group II introns: a quantitative analysis of in vitro reaction rates and products. J Mol Biol. 1996 Feb 16;256(1):31-49. PubMed PMID: 8609612.

Darling AE, Mau B, Perna NT. progressiveMauve: multiple genome alignment with gene gain, loss and rearrangement. PLoS One. 2010 Jun 25;5(6):e11147. PubMed PMID: 20593022; PubMed Central PMCID: PMC2892488.

de Beer ZW, Duong TA, Barnes I, Wingfield BD, Wingfield MJ. Redefining *Ceratocystis* and allied genera. Stud Mycol. 2014 Sep;79:187-219. PubMed PMID: 25492989; PubMed Central PMCID: PMC4255530.

de Beer ZW, Marincowitz S, Duong TA, Wingfield MJ. Bretziella, a new genus to accommodate the oak wilt fungus, *Ceratocystis fagacearum* (Microascales, Ascomycota). MycoKeys. 2017;27:1-19. doi: 10.3897/mycokeys.27.20657.

de Queiroz CB, Santana MF, Pereira Vidigal PM, de Queiroz MV. Comparative analysis of the mitochondrial genome of the fungus *Colletotrichum lindemuthianum*, the causal agent of anthracnose in common beans. Appl Microbiol Biotechnol. 2018 Mar;102(6):2763-2778. PubMed PMID: 29453633.

Delahodde A, Goguel V, Becam AM, Creusot F, Perea J, et al. Site-specific DNA endonuclease and RNA maturase activities of two homologous intron-encoded proteins from yeast mitochondria. Cell. 1989 Feb 10;56(3):431-41. PubMed PMID: 2536593.

Deng Y, Zhang Q, Ming R, Lin L, Lin X, et al. Analysis of the Mitochondrial Genome in *Hypomyces aurantius* Reveals a Novel Twintron Complex in Fungi. Int J Mol Sci. 2016 Jun 30;17(7)PubMed PMID: 27376282; PubMed Central PMCID: PMC4964425.

Dujon B. Sequence of the intron and flanking exons of the mitochondrial 21S rRNA gene of yeast strains having different alleles at the omega and rib-1 loci. Cell. 1980 May;20(1):185-97. PubMed PMID: 6156002.

Edgell DR, Chalamcharla VR, Belfort M. Learning to live together: mutualism between self-splicing introns and their hosts. BMC Biol. 2011 Apr 11;9:22. PubMed PMID: 21481283; PubMed Central PMCID: PMC3073962.

Edgell DR. Selfish DNA: homing endonucleases find a home. Curr Biol. 2009 Feb 10;19(3):R115-7. PubMed PMID: 19211047.

Eigen M. Selforganization of matter and the evolution of biological macromolecules. Naturwissenschaften. 1971 Oct;58(10):465-523. PubMed PMID: 4942363.

Engelbrecht CJ, Harrington TC, Alfenas A. *Ceratocystis* wilt of cacao-a disease of increasing importance. Phytopathology. 2007 Dec;97(12):1648-9. PubMed PMID: 18943727.

Esser C, Ahmadinejad N, Wiegand C, Rotte C, Sebastiani F, Gelius-Dietrich G et al. A genome phylogeny for mitochondria among alpha-proteobacteria and predominantly eubacterial ancestry of yeast nuclear genes. Mol Biol Evol. 2004;21:1643-1660.

Férandon C, Moukha S, Callac P, Benedetto JP, Castroviejo M, et al. The Agaricus bisporus cox1 gene: the longest mitochondrial gene and the largest reservoir of mitochondrial group i introns. PLoS One. 2010 Nov 18;5(11):e14048. PubMed PMID: 21124976; PubMed Central PMCID: PMC2987802.

Férandon C, Xu J, Barroso G. The 135 kbp mitochondrial genome of Agaricus bisporus is the largest known eukaryotic reservoir of group I introns and plasmid-related sequences. Fungal Genet Biol. 2013 Jun;55:85-91. PubMed PMID: 23428625.

Ferat JL, Michel F. Group II self-splicing introns in bacteria. Nature. 1993 Jul 22;364(6435):358-61. PubMed PMID: 7687328.

Franco MEE, López SMY, Medina R, Lucentini CG, Troncozo MI, et al. The mitochondrial genome of the plant-pathogenic fungus *Stemphylium lycopersici* uncovers a dynamic structure due to repetitive and mobile elements. PLoS One. 2017;12(10):e0185545. PubMed PMID: 28972995; PubMed Central PMCID: PMC5626475.

Freel KC, Friedrich A, Schacherer J. Mitochondrial genome evolution in yeasts: an all-encompassing view. FEMS Yeast Res. 2015 Jun;15(4):fov023. PubMed PMID: 25969454.

Fricova D, Valach M, Farkas Z, Pfeiffer I, Kucsera J, et al. The mitochondrial genome of the pathogenic yeast *Candida subhashii*: GC-rich linear DNA with a protein covalently attached to the 5' termini. Microbiology. 2010 Jul;156(Pt 7):2153-63. PubMed PMID: 20395267; PubMed Central PMCID: PMC3068681.

Gabaldon T, Huynen M. Reconstruction of the proto-mitochondrial metabolism. Science. 2003;301:609.

Gabaldon T, Huynen M. From endosymbiont to host-controlled organelle: the hijacking of the mitochondrial protein synthesis and metabolism. PLOS Comput Biol. 2007;3:e219.

Girish PS, Anjaneyulu AS, Viswas KN, Shivakumar BM, Anand M, et al. Meat species identification by polymerase chain reaction-restriction fragment length polymorphism (PCR-RFLP) of mitochondrial 12S rRNA gene. Meat Sci. 2005 May;70(1):107-12. PubMed PMID: 22063286.

Girish PS, Anjaneyulu AS, Viswas KN, Anand M, Rajkumar N, et al. Sequence analysis of mitochondrial 12S rRNA gene can identify meat species. Meat Sci. 2004 Mar;66(3):551-6. PubMed PMID: 22060864.

Goddard MR, Burt A. Recurrent invasion and extinction of a selfish gene. Proc Natl Acad Sci U S A. 1999 Nov 23;96(24):13880-5. PubMed PMID: 10570167; PubMed Central PMCID: PMC24159.

Gogarten JP, Hilario E. Inteins, introns, and homing endonucleases: recent revelations about the life cycle of parasitic genetic elements. BMC Evol Biol. 2006 Nov 13;6:94. PubMed PMID: 17101053; PubMed Central PMCID: PMC1654191.

Guha TK, Wai A, Mullineux ST, Hausner G. The intron landscape of the mtDNA *cytb* gene among the Ascomycota: introns and intron-encoded open reading frames. Mitochondrial DNA A DNA Mapp Seq Anal. 2017 Nov 20;:1-10. PubMed PMID: 29157056.

Hafez M, Majer A, Sethuraman J, Rudski SM, Michel F, et al. The mtDNA rns gene landscape in the Ophiostomatales and other fungal taxa: twintrons, introns, and intron-encoded proteins. Fungal Genet Biol. 2013 Apr;53:71-83. PubMed PMID: 23403360.

Hafez M, Hausner G. Convergent evolution of twintron-like configurations: One is never enough. RNA Biol. 2015;12(12):1275-88. PubMed PMID: 26513606; PubMed Central PMCID: PMC4829276.

Hafez M, Hausner G. The highly variable mitochondrial small-subunit ribosomal RNA gene of *Ophiostoma minus*. Fungal Biol. 2011 Nov;115(11):1122-37. PubMed PMID: 22036291.

Hafez M, Hausner G. Homing endonucleases: DNA scissors on a mission. Genome. 2012 Aug;55(8):553-69. PubMed PMID: 22891613.

Hahn C, Bachmann L, Chevreux B. Reconstructing mitochondrial genomes directly from genomic next-generation sequencing reads--a baiting and iterative mapping approach. Nucleic Acids Res. 2013 Jul;41(13):e129. PubMed PMID: 23661685; PubMed Central PMCID: PMC3711436.

Halls C, Mohr S, Del Campo M, Yang Q, Jankowsky E, et al. Involvement of DEAD-box proteins in group I and group II intron splicing Biochemical characterization of Mss116p, ATP hydrolysis-dependent and -independent mechanisms, and general RNA chaperone activity. J Mol Biol. 2007 Jan 19;365(3):835-55. PubMed PMID: 17081564; NIHMSID: NIHMS16168; PubMed Central PMCID: PMC1832103.

Halsted BD. Some fungous disease of sweet potato. New Jersey Agricultural College Experimental Station Bulletin. 1890;76:1-32.

Halsted BD, Fairchild DG. Sweet-potato black rot. J Mycol. 1891;7:1–11.

Harrington TC, Wingfield MJ. The *Ceratocystis* species on conifers. Can J Bot. 1998;76:1446-1447.

Haugen P, Bhattacharya D. The spread of LAGLIDADG homing endonuclease genes in rDNA. Nucleic Acids Res. 2004;32(6):2049-57. PubMed PMID: 15069127; PubMed Central PMCID: PMC390371.

Haugen P, Simon DM, Bhattacharya D. The natural history of group I introns. Trends Genet. 2005 Feb;21(2):111-9. PubMed PMID: 15661357.

Hausner G. Introns, mobile elements and plasmids. In: Bullerwell CE, editor. Organelle Genetics: Evolution of Organelle Genomes and Gene Expression. Springer Verlag; 2012 p. 329-358.

Hausner G. Fungal Mitochondrial Genomes, Introns and Plasmids. In: Arora DK, Khachatourians GG, editors. Applied Mycology and Biotechnology Volume III: Fungal Genomics. New York: Elsevier Science; 2003. p. 101-131.

Hausner G, Hafez M, Edgell DR. Bacterial group I introns: mobile RNA catalysts. Mob DNA. 2014 Mar 10;5(1):8. PubMed PMID: 24612670; PubMed Central PMCID: PMC3984707.

Hausner G, Nummy KA, Bertrand H. Asexual transmission, non-suppressiveness and meiotic extinction of small plasmid-like derivatives of the mitochondrial DNA in *Neurospora crassa*. Fungal Genet Biol. 2006 Feb;43(2):90-101. PubMed PMID: 16386438.

Hausner G, Reid J, Klassen GR. On the subdivision of *Ceratocystis* s.l., based on partial ribosomal DNA sequences. Can J Bot. 1993;71:52-63.

Hausner G, Reid J, Klassen GR. Do galeate-ascospore members of the Cephaloascaceae, Endomycetaceae and Ophiostomataceae share a common phylogeny? Mycologia. 1992;84:870–881.

Hegedusova E, Brejova B, Tomaska L, Sipiczki M, Nosek J. Mitochondrial genome of the basidiomycetous yeast *Jaminaea angkorensis*. Curr Genet. 2014 Feb;60(1):49-59. PubMed PMID: 24071901.

Hepburn NJ, Schmidt DW, Mower JP. Loss of two introns from the *Magnolia tripetala* mitochondrial *cox2* gene implicates horizontal gene transfer and gene conversion as a novel mechanism of intron loss. Mol Biol Evol. 2012 Oct;29(10):3111-20. PubMed PMID: 22593225.

Houston DR. Recognizing and managing sapstreak disease of sugar maple. Res Pap. 1993;NE-675:1 – 11.

Jalalzadeh B, Saré IC, Férandon C, Callac P, Farsi M, et al. The intraspecific variability of mitochondrial genes of *Agaricus bisporus* reveals an extensive group I intron mobility combined with low nucleotide substitution rates. Curr Genet. 2015 Feb;61(1):87-102. PubMed PMID: 25159526.

Joardar V, Abrams NF, Hostetler J, Paukstelis PJ, Pakala S, et al. Sequencing of mitochondrial genomes of nine *Aspergillus* and *Penicillium* species identifies mobile introns and accessory genes as main sources of genome size variability. BMC Genomics. 2012 Dec 12;13:698. PubMed PMID: 23234273; PubMed Central PMCID: PMC3562157.

Johansen S, Haugen P. A new nomenclature of group I introns in ribosomal DNA. RNA. 2001 Jul;7(7):935-6. PubMed PMID: 11453066; PubMed Central PMCID: PMC1370146.

Jones KG, Blackwell M. Phylogenetic analysis of the ambrosial species in the genus *Raffaelea* based on 18S rDNA sequences. Mycol Res. 1998;102:661-665.

Juzwik J, Harrington TC, MacDonald WL, Appel DN. The origin of *Ceratocystis fagacearum*, the oak wilt fungus. Annu Rev Phytopathol. 2008;46:13-26. PubMed PMID: 18680421.

Kang X, Hu L, Shen P, Li R, Liu D. SMRT Sequencing Revealed Mitogenome Characteristics and Mitogenome-Wide DNA Modification Pattern in *Ophiocordyceps sinensis*. Front Microbiol. 2017;8:1422. PubMed PMID: 28798740; PubMed Central PMCID: PMC5529405.

Kanzi AM, Wingfield BD, Steenkamp ET, Naidoo S, van der Merwe NA. Intron Derived Size Polymorphism in the Mitochondrial Genomes of Closely Related *Chrysoporthe* Species. PLoS One. 2016;11(6):e0156104. PubMed PMID: 27272523; PubMed Central PMCID: PMC4894602.

Katoh K, Standley DM. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol. 2013 Apr;30(4):772-80. PubMed PMID: 23329690; PubMed Central PMCID: PMC3603318.

Kelley LA, Mezulis S, Yates CM, Wass MN, Sternberg MJ. The Phyre2 web portal for protein modeling, prediction and analysis. Nat Protoc. 2015 Jun;10(6):845-58. PubMed PMID: 25950237; NIHMSID: EMS71337; PubMed Central PMCID: PMC5298202.

Kleinstiver BP, Wolfs JM, Kolaczyk T, Roberts AK, Hu SX, et al. Monomeric site-specific nucleases for genome editing. Proc Natl Acad Sci U S A. 2012 May 22;109(21):8061-6. PubMed PMID: 22566637; PubMed Central PMCID: PMC3361397.

Korovesi AG, Ntertilis M, Kouvelis VN. Mt-rps3 is an ancient gene which provides insight into the evolution of fungal mitochondrial genomes. Mol Phylogenet Evol. 2018 Oct;127:74-86. PubMed PMID: 29763662.

Krogh A, Larsson B, von Heijne G, Sonnhammer EL. Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol. 2001 Jan 19;305(3):567-80. PubMed PMID: 11152613.

Krzywinski M, Schein J, Birol I, Connors J, Gascoyne R, et al. Circos: an information aesthetic for comparative genomics. Genome Res. 2009 Sep;19(9):1639-45. PubMed PMID: 19541911; PubMed Central PMCID: PMC2752132.

Kurowski MA, Bujnicki JM. GeneSilico protein structure prediction meta-server. Nucleic Acids Res. 2003 Jul 1;31(13):3305-7. PubMed PMID: 12824313; PubMed Central PMCID: PMC168964.

Kurtz S, Choudhuri JV, Ohlebusch E, Schleiermacher C, Stoye J, et al. REPuter: the manifold applications of repeat analysis on a genomic scale. Nucleic Acids Res. 2001 Nov 15;29(22):4633-42. PubMed PMID: 11713313; PubMed Central PMCID: PMC92531.

Lambowitz AM, Zimmerly S. Group II introns: mobile ribozymes that invade DNA. Cold Spring Harb Perspect Biol. 2011 Aug 1;3(8):a003616. PubMed PMID: 20463000; PubMed Central PMCID: PMC3140690.

Lambowitz AM, Zimmerly S. Mobile group II introns. Annu Rev Genet. 2004;38:1-35. PubMed PMID: 15568970.

Lambowitz AM, Belfort M. Mobile Bacterial Group II Introns at the Crux of Eukaryotic Evolution. Microbiol Spectr. 2015 Feb;3(1):MDNA3-0050-2014. PubMed PMID: 26104554.

Lang BF, Laforest MJ, Burger G. Mitochondrial introns: a critical view. Trends Genet. 2007 Mar;23(3):119-25. PubMed PMID: 17280737.

Li L, Stoeckert CJ Jr, Roos DS. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 2003 Sep;13(9):2178-89. PubMed PMID: 12952885; PubMed Central PMCID: PMC403725.

Li Y, Hu XD, Yang RH, Hsiang T, Wang K, et al. Complete mitochondrial genome of the medicinal fungus *Ophiocordyceps sinensis*. Sci Rep. 2015 Sep 15;5:13892. PubMed PMID: 26370521; PubMed Central PMCID: PMC4570212.

Ling F, Shibata T. Recombination-dependent mtDNA partitioning: in vivo role of Mhr1p to promote pairing of homologous DNA. EMBO J. 2002 Sep 2;21(17):4730-40. PubMed PMID: 12198175; PubMed Central PMCID: PMC126199.

Losada L, Pakala SB, Fedorova ND, Joardar V, Shabalina SA, et al. Mobile elements and mitochondrial genome expansion in the soil fungus and potato pathogen *Rhizoctonia solani* AG-3. FEMS Microbiol Lett. 2014 Mar;352(2):165-73. PubMed PMID: 24461055.

Lowe TM, Eddy SR. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. 1997 Mar 1;25(5):955-64. PubMed PMID: 9023104; PubMed Central PMCID: PMC146525.

Maleszka R, Skelly PJ, Clark-Walker GD. Rolling circle replication of DNA in yeast mitochondria. EMBO J. 1991 Dec;10(12):3923-9. PubMed PMID: 1935911; PubMed Central PMCID: PMC453131.

Margulies M, Egholm M, Altman WE, Attiya S, Bader JS, et al. Genome sequencing in microfabricated high-density picolitre reactors. Nature. 2005 Sep 15;437(7057):376-80. PubMed PMID: 16056220; NIHMSID: NIHMS5166; PubMed Central PMCID: PMC1464427.

Martin W, Koonin EV. Introns and the origin of nucleus-cytosol compartmentalization. Nature. 2006 Mar 2;440(7080):41-5. PubMed PMID: 16511485.

Martin W. Gene transfer from organelles to the nucleus: frequent and in big chunks. Proc Natl Acad Sci U S A. 2003 Jul 22;100(15):8612-4. PubMed PMID: 12861078; PubMed Central PMCID: PMC166356.

Martínez-Rodríguez L, García-Rodríguez FM, Molina-Sánchez MD, Toro N, Martínez-Abarca F. Insights into the strategies used by related group II introns to adapt successfully for the colonisation of a bacterial genome. RNA Biol. 2014;11(8):1061-71. PubMed PMID: 25482895; PubMed Central PMCID: PMC4615759.

Michel F, Westhof E. Modelling of the three-dimensional architecture of group I catalytic introns based on comparative sequence analysis. J Mol Biol. 1990 Dec 5;216(3):585-610. PubMed PMID: 2258934.

Michel F, Umesono K, Ozeki H. Comparative and functional anatomy of group II catalytic introns--a review. Gene. 1989 Oct 15;82(1):5-30. PubMed PMID: 2684776.

Michel F, Ferat JL. Structure and activities of group II introns. Annu Rev Biochem. 1995;64:435-61. PubMed PMID: 7574489.

Monteiro-Vitorello CB, Hausner G, Searles DB, Gibb EA, Fulbright DW, et al. The *Cryphonectria parasitica* mitochondrial rns gene: plasmid-like elements, introns and homing endonucleases. Fungal Genet Biol. 2009 Nov;46(11):837-48. PubMed PMID: 19607930.

Nawrocki EP, Jones TA, Eddy SR. Group I introns are widespread in archaea. Nucleic Acids Res. 2018 May 18;PubMed PMID: 29788499.

Osigus HJ, Eitel M, Schierwater B. Deep RNA sequencing reveals the smallest known mitochondrial micro exon in animals: The placozoan cox1 single base pair exon. PLoS One. 2017;12(5):e0177959. PubMed PMID: 28542197; PubMed Central PMCID: PMC5436844.

Palmer JD, Adams KL, Cho Y, Parkinson CL, Qiu YL, et al. Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. Proc Natl Acad Sci U S A. 2000 Jun 20;97(13):6960-6. PubMed PMID: 10860957; PubMed Central PMCID: PMC34370.

Parson W, Pegoraro K, Niederstätter H, Föger M, Steinlechner M. Species identification by means of the cytochrome b gene. Int J Legal Med. 2000;114(1-2):23-8. PubMed PMID: 11197623.

Pramateftaki PV, Kouvelis VN, Lanaridis P, Typas MA. The mitochondrial genome of the wine yeast *Hanseniaspora uvarum*: a unique genome organization among yeast/fungal counterparts. FEMS Yeast Res. 2006 Jan;6(1):77-90. PubMed PMID: 16423073.

Repar J, Warnecke T. Mobile Introns Shape the Genetic Diversity of Their Host Genes. Genetics. 2017 Apr;205(4):1641-1648. PubMed PMID: 28193728; PubMed Central PMCID: PMC5378118.

Robba L, Russell SJ, Barker GL, Brodie J. Assessing the use of the mitochondrial *cox1* marker for use in DNA barcoding of red algae (Rhodophyta). Am J Bot. 2006 Aug;93(8):1101-8. PubMed PMID: 21642175.

Robicheau BM, Young AP, LaButti K, Grigoriev IV, Walker AK. The complete mitochondrial genome of the conifer needle endophyte, *Phialocephala scopiformis* DAOMC 229536 confirms evolutionary division within the fungal *Phialocephala fortinii* sl-Acephala appalanata species complex. Fungal Biol. 2017 Mar;121(3):212-221. PubMed PMID: 28215349.

Rodrigues MS, Morelli KA, Jansen AM. Cytochrome c oxidase subunit 1 gene as a DNA barcode for discriminating *Trypanosoma cruzi* DTUs and closely related species. Parasit

Vectors. 2017 Oct 16;10(1):488. PubMed PMID: 29037251; PubMed Central PMCID: PMC5644147.

Ronquist F, Teslenko M, van der Mark P, Ayres DL, Darling A, et al. MrBayes 32: efficient Bayesian phylogenetic inference and model choice across a large model space. Syst Biol. 2012 May;61(3):539-42. PubMed PMID: 22357727; PubMed Central PMCID: PMC3329765.

Ruan J, Cheng J, Zhang T, Jiang H. Mitochondrial genome evolution in the *Saccharomyces* sensu stricto complex. PLoS One. 2017;12(8):e0183035. PubMed PMID: 28813471; PubMed Central PMCID: PMC5558958.

Rudan M, Bou Dib P, Musa M, Kanunnikau M, Sobočanec S, et al. Normal mitochondrial function in *Saccharomyces cerevisiae* has become dependent on inefficient splicing. Elife. 2018 Mar 23;7PubMed PMID: 29570052; PubMed Central PMCID: PMC5898908.

Rutherford K, Parkhill J, Crook J, Horsnell T, Rice P, et al. Artemis: sequence visualization and annotation. Bioinformatics. 2000 Oct;16(10):944-5. PubMed PMID: 11120685.

Salavirta H, Oksanen I, Kuuskeri J, Mäkelä M, Laine P, et al. Mitochondrial genome of *Phlebia radiata* is the second largest (156 kbp) among fungi and features signs of genome flexibility and recent recombination events. PLoS One. 2014;9(5):e97141. PubMed PMID: 24824642; PubMed Central PMCID: PMC4019555.

Schoch CL, Sung GH, López-Giráldez F, Townsend JP, Miadlikowska J, et al. The Ascomycota tree of life: a phylum-wide phylogeny clarifies the origin and evolution of fundamental reproductive and ecological traits. Syst Biol. 2009 Apr;58(2):224-39. PubMed PMID: 20525580.

Schoch CL, Seifert KA, Huhndorf S, Robert V, Spouge JL, et al. Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. Proc Natl Acad Sci U S A. 2012 Apr 17;109(16):6241-6. PubMed PMID: 22454494; PubMed Central PMCID: PMC3341068.

Sethuraman J, Rudski SM, Wosnitza K, Hafez M, Guppy B, et al. Evolutionary dynamics of introns and their open reading frames in the U7 region of the mitochondrial rnl gene in species of *Ceratocystis*. Fungal Biol. 2013 Nov-Dec;117(11-12):791-806. PubMed PMID: 24295918.

Sethuraman J, Majer A, Iranpour M, Hausner G. Molecular evolution of the mtDNA encoded rps3 gene among filamentous ascomycetes fungi with an emphasis on the Ophiostomatoid fungi. J Mol Evol. 2009 Oct;69(4):372-85. PubMed PMID: 19826748.

Sethuraman J, Okoli CV, Majer A, Corkery TL, Hausner G. The sporadic occurrence of a group I intron-like element in the mtDNA rnl gene of *Ophiostoma novo-ulmi* subsp americana. Mycol Res. 2008 May;112(Pt 5):564-82. PubMed PMID: 18406119.

Simon DM, Clarke NA, McNeil BA, Johnson I, Pantuso D, et al. Group II introns in eubacteria and archaea: ORF-less introns and new varieties. RNA. 2008 Sep;14(9):1704-13. PubMed PMID: 18676618; PubMed Central PMCID: PMC2525955.

Smith DR. The past, present and future of mitochondrial genomics: have we sequenced enough mtDNAs?. Brief Funct Genomics. 2016 Jan;15(1):47-54. PubMed PMID: 26117139; PubMed Central PMCID: PMC4812591.

Stanke M, Morgenstern B. AUGUSTUS: a web server for gene prediction in eukaryotes that allows user-defined constraints. Nucleic Acids Res. 2005 Jul 1;33(Web Server issue):W465-7. PubMed PMID: 15980513; PubMed Central PMCID: PMC1160219.

Stoddard BL. Homing endonuclease structure and function. Q Rev Biophys. 2005 Feb;38(1):49-95. PubMed PMID: 16336743.

Stoddard BL. Homing endonucleases from mobile group I introns: discovery to genome engineering. Mob DNA. 2014 Mar 3;5(1):7. PubMed PMID: 24589358; PubMed Central PMCID: PMC3943268.

Stoddard BL. Homing endonucleases: from microbial genetic invaders to reagents for targeted DNA modification. Structure. 2011 Jan 12;19(1):7-15. PubMed PMID: 21220111; NIHMSID: NIHMS261821; PubMed Central PMCID: PMC3038549.

Sung GH. Complete mitochondrial DNA genome of the medicinal mushroom *Cordyceps militaris* (Ascomycota, Cordycipitaceae). Mitochondrial DNA. 2015;26(5):789-90. PubMed PMID: 24320617.

Szostak N, Wasik S, Blazewicz J. Hypercycle. PLoS Comput Biol. 2016 Apr;12(4):e1004853. PubMed PMID: 27054759; PubMed Central PMCID: PMC4824418.

Toor N, Zimmerly S. Identification of a family of group II introns encoding LAGLIDADG ORFs typical of group I introns. RNA. 2002 Nov;8(11):1373-7. PubMed PMID: 12458791; PubMed Central PMCID: PMC1370344.

Toor N, Hausner G, Zimmerly S. Coevolution of group II intron RNA structures with their intron-encoded reverse transcriptases. RNA. 2001 Aug;7(8):1142-52. PubMed PMID: 11497432; PubMed Central PMCID: PMC1370161.

Toor N, Keating KS, Fedorova O, Rajashankar K, Wang J, et al. Tertiary architecture of the *Oceanobacillus iheyensis* group II intron. RNA. 2010 Jan;16(1):57-69. PubMed PMID: 19952115; PubMed Central PMCID: PMC2802037.

Toor N, Keating KS, Taylor SD, Pyle AM. Crystal structure of a self-spliced group II intron. Science. 2008 Apr 4;320(5872):77-82. PubMed PMID: 18388288; NIHMSID: NIHMS680281; PubMed Central PMCID: PMC4406475.

Torriani SF, Penselin D, Knogge W, Felder M, Taudien S, et al. Comparative analysis of mitochondrial genomes from closely related *Rhynchosporium* species reveals extensive intron invasion. Fungal Genet Biol. 2014 Jan;62:34-42. PubMed PMID: 24240058.

Tritt A, Eisen JA, Facciotti MT, Darling AE. An integrated pipeline for *de novo* assembly of microbial genomes. PLoS One. 2012;7(9):e42304. PubMed PMID: 23028432; PubMed Central PMCID: PMC3441570.

Tsopelas P, Santini A, Wingfield MJ, de Beer ZW. Canker stain: A lethal disease destroying iconic plane trees. Plant Disease. 2017;101(5):645-658. doi: 10.1094/PDIS-09-16-1235-FE.

Tyack PL, Calambokidis J, Friedlaender A, Goldbogen J, Southall B. Formal Comment on Schorr GS, Falcone EA, Moretti DJ, Andrews RD (2014) First Long-Term Behavioral Records from Cuvier's Beaked Whales (Ziphius cavirostris) Reveal Record-Breaking Dives PLoS ONE 9(3): e92633 doi:101371/journalpone0092633. PLoS One. 2015;10(12):e0142287. PubMed PMID: 26678487; PubMed Central PMCID: PMC4683059.

van de Sande WW. Phylogenetic analysis of the complete mitochondrial genome of Madurella mycetomatis confirms its taxonomic position within the order Sordariales. PLoS One. 2012;7(6):e38654. PubMed PMID: 22701687; PubMed Central PMCID: PMC3368884.

van der Nest MA, Bihon W, De Vos L, Naidoo K, Roodt D, et al. IMA Genome-F 2: *Ceratocystis manginecans, Ceratocystis moniliformis, Diplodia sapinea*: Draft genome sequences of *Diplodia sapinea, Ceratocystis manginecans,* and *Ceratocystis moniliformis*. IMA Fungus. 2014 Jun;5(1):135-40. PubMed PMID: 25083413; PubMed Central PMCID: PMC4107891.

Vinothkumar KR, Zhu J, Hirst J. Architecture of mammalian respiratory complex I. Nature. 2014 Nov 6;515(7525):80-84. PubMed PMID: 25209663; NIHMSID: EMS59675; PubMed Central PMCID: PMC4224586.

Visser C, Wingfield MJ, Wingfield BD, Yamaoka Y. *Ophiostoma polonicum* is a species of *Ceratocystis* sensu stricto. Systematic and Applied Microbiol. 1995;18:403-409.

Ward RD, Zemlak TS, Innes BH, Last PR, Hebert PD. DNA barcoding Australia's fish species. Philos Trans R Soc Lond B Biol Sci. 2005 Oct 29;360(1462):1847-57. PubMed PMID: 16214743; PubMed Central PMCID: PMC1609232.

Waterhouse AM, Procter JB, Martin DM, Clamp M, Barton GJ. Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics. 2009 May 1;25(9):1189-91. PubMed PMID: 19151095; PubMed Central PMCID: PMC2672624.

Wilken PM, Steenkamp ET, Wingfield MJ, de Beer ZW, Wingfield BD. IMA Genome-F 1: *Ceratocystis fimbriata*: Draft nuclear genome sequence for the plant pathogen, *Ceratocystis fimbriata*. IMA Fungus. 2013 Dec;4(2):357-8. PubMed PMID: 24563841; PubMed Central PMCID: PMC3905947.

Wingfield BD, Duong TA, Hammerbacher A, van der Nest MA, Wilson A, et al. IMA Genome-F 7: Draft genome sequences for *Ceratocystis fagacearum*, *C harringtonii*, *Grosmannia penicillata*, and *Huntiella bhutanensis*. IMA Fungus. 2016 Dec;7(2):317-323. PubMed PMID: 27990338; PubMed Central PMCID: PMC5159602.

Wingfield MJ, Seifert KA, Webber J, editors. *Ceratocystis* and *Ophiostoma*: taxonomy, ecology and pathogenicity. American Phytopathological Society Press; 1993.

Witthuhn RC, Wingfield BD, Wingfield MJ. Monophyly of the conifer species in the *Ceratocystis coerulescens* complex based on DNA sequence data. Mycologia. 1998;90:96-101.

Wolters JF, Chiu K, Fiumera HL. Population structure of mitochondrial genomes in *Saccharomyces cerevisiae*. BMC Genomics. 2015 Jun 11;16:451. PubMed PMID: 26062918; PubMed Central PMCID: PMC4464245.

Wu B, Hao W. Horizontal transfer and gene conversion as an important driving force in shaping the landscape of mitochondrial introns. G3 (Bethesda). 2014 Apr 16;4(4):605-12. PubMed PMID: 24515269; PubMed Central PMCID: PMC4059233.

Xiao S, Nguyen DT, Wu B, Hao W. Genetic Drift and Indel Mutation in the Evolution of Yeast Mitochondrial Genome Size. Genome Biol Evol. 2017 Nov 1;9(11):3088-3099. PubMed PMID: 29126284; PubMed Central PMCID: PMC5714193.

Xu Y, Yang S, Turitsa I, Griffiths A. Divergence of a linear and a circular plasmid in disjunct natural isolates of the fungus *Neurospora*. Plasmid. 1999 Sep;42(2):115-25. PubMed PMID: 10489328.

Yuan XL, Mao XX, Liu XM, Cheng S, Zhang P, et al. The complete mitochondrial genome of *Engyodontium album* and comparative analyses with Ascomycota mitogenomes. Genet Mol Biol. 2017 Oct-Dec;40(4):844-854. PubMed PMID: 29064513; PubMed Central PMCID: PMC5738615.

Zeng Q, Bonocora RP, Shub DA. A free-standing homing endonuclease targets an intron insertion site in the psbA gene of cyanophages. Curr Biol. 2009 Feb 10;19(3):218-22. PubMed PMID: 19200728.

Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs. Genome Res. 2008 May;18(5):821-9. PubMed PMID: 18349386; PubMed Central PMCID: PMC2336801.

Zhang YJ, Yang XQ, Zhang S, Humber RA, Xu J. Genomic analyses reveal low mitochondrial and high nuclear diversity in the cyclosporin-producing fungus *Tolypocladium inflatum*. Appl Microbiol Biotechnol. 2017 Dec;101(23-24):8517-8531. PubMed PMID: 29034434.

Zhou Y, Lu C, Wu QJ, Wang Y, Sun ZT, et al. GISSD: Group I Intron Sequence and Structure Database. Nucleic Acids Res. 2008 Jan;36(Database issue):D31-7. PubMed PMID: 17942415; PubMed Central PMCID: PMC2238919.

Zimmerly S, Semper C. Evolution of group II introns. Mob DNA. 2015;6:7. PubMed PMID: 25960782; PubMed Central PMCID: PMC4424553.