10-12-2021 12:30 PM

Metschnikowia mitochondria

Dong Kyung Lee, *The University of Western Ontario*

Supervisor: Smith, David R., *The University of Western Ontario*
Co-Supervisor: Lachance, MA., *The University of Western Ontario*
A thesis submitted in partial fulfillment of the requirements for the Master of Science degree in Biology

Abstract

Mitochondrial genomes are known for their diverse characteristics and are an attractive model to study genome evolution. Draft nuclear genomes of 71 *Metschnikowia* yeast strains are publicly available but their mitochondrial genome assemblies are incomplete, thereby making genome studies difficult. To remediate this shortcoming, complete mitochondrial genomes of 71 *Metschnikowia* strains were assembled from the draft nuclear genomes. *Metschnikowia* mitochondrial genomes exhibit an unprecedented amount of diversity, particularly with respect to the frequency and distribution of introns, which is often reflected upon overall genome size variations. Additionally, loss of synteny between strains of the same species further strengthens the notion that mitochondrial genomes evolve differently from their host genomes. Diversities shown from multiple genome characteristics explored in this thesis therefore highlights importance of mitochondrial genomes for studying evolution and diversity of genomes that were often neglected.

Keywords

Genome diversity, *Metschnikowia*, Mitochondrial genome, mtDNA, yeast.

Summary for Lay Audience

Mitochondria are important organelles responsible for the production of energy in eukaryotic cells. Mitochondria differ from many other organelles by possessing their own genome. Early studies have focused on nuclear genomes, but there is now increasing interest in mitochondrial genomes, as the two are presumed to share similar, but somewhat independent evolutionary histories. *Metschnikowia* species are yeasts commonly found in the guts of beetles that inhabit morning glories and other flowers around the world. The nuclear genomes from various *Metschnikowia* are available, but their mitochondrial counterpart is known mostly from raw sequence reads that require careful assembly in order to extract useful information on their diversity. My thesis focused on constructing and analyzing mitochondrial genomes of these yeasts. I have found that the mitochondrial genomes of these species exhibit a remarkable amount of architectural diversity, such as how genes are oriented across mitochondrial genomes and how their overall distributions differ from other species, as well as how segments that are not converted into proteins (introns) are present within a gene. All these diversities contribute towards incredible overall size differences of mitochondrial genomes of *Metschnikowia* species. In many cases, even individuals belong to same species showed vast differences that were rare in nuclear genomes. In conclusion, my thesis shed a light on how amazing these non-nuclear genomes are and why they should be studied more in detail in hopes of further our understanding of how genomes evolve.

Co-Authorship Statement

An account of mitochondrial genome size morphology has been published (Lee et al. 2020). My co-authors Prof. D. R. Smith, Prof. M. A. Lachance, Prof. T. Hsiang and I contributed to the design of the project, the nuclear genome assemblies, and manuscript writing and editing. I have performed all mitochondrial genome analyses. Another manuscript with the same co-authors, covering the rest of the thesis, will be submitted for publication in due course. Under supervision of Prof. M. A. Lachance in University of Western Ontario, I have co-authored additional publications that were not directly related to the thesis topic (mitochondria) itself, but still involved nuclear genomes of the large-spored clade of *Metschnikowia* species. I have contributed *in-silico* analyses for descriptions of new species (Santos et al. 2020; Lee et al. 2020), a study of yeast pheromones (Lee et al. 2018), the development of a novel method for gene transformation (Gordon et al. 2019), and the application of whole genome divergence in the delineation of *Metschnikowia* species (Lachance et al. 2020).

Acknowledgments

I thank my supervisors, Prof. M. A. Lachance and Prof. D. R. Smith as well as my advisors Prof. A. Poon and Prof. V. Tai for their continuous guidance and support before and during my candidacy to the Master's degree. I also thank Prof. R. Gardiner and the members of the Biotron Experimental Climate Change Research Centre for helping with scanning electron microscopy and transmission electron microscopy.

Table of Contents

List of Abbreviations

ATP	Adenosine triphosphate
COB	Cytochrome b
COX	Cytochrome c oxidase
DNA	Deoxyribonucleic acid
EBD	Estación Biológica de Doñana
LSC	Large spored clade
mtDNA	Mitochondrial DNA
MUCL	Mycothèque de l'Université catholique de Louvain
NAD	Nitotinamide adenine dinucleotide
NRRL	Northern Regional Research Laboratory
ORF	Open reading frame
PCR	Polymerase chain reaction
ProCARs	Progressive contiguous ancestral regions
RNA	Ribonucleic acid
rnl	Ribosomal RNA large subunit
rns	Ribosomal RNA small subunit
rRNA	Ribosomal ribonucleic acid
SEM	Scanning electron microscopy
TEM	Transmission electron microscopy
tRNA	Transfer ribonucleic acid
UWOPS	University of Western Ontario Plant Science
UFMG	University of Federal de Minas Gerais

List of Tables

List of Figures

List of Appendices

Chapter 1

1 Introduction

1.1 Mitochondria

The mitochondrion is an important organelle for eukaryotes, participating in various essential metabolic processes, from generating energy to modulating intracellular signals (Chandel 2015). The first recorded observation of mitochondria dates back to the 19th century when internal cellular structures were being observed with microscopes (Ernster and Schatz 1981). The term mitochondrion, a combination of the words mito and chondros, meaning thread and granules in Greek, respectively, was created by Carl Benda who noticed chain-like structures (Ernster and Schatz 1981). Although Richard Altmann in the 19th century speculated correctly that mitochondria play a vital role in the cell, the primary role of mitochondria as ATP generators through oxidative phosphorylation was not discovered until the 20th century by Paul Boyer (Boyer 1965). Other roles of mitochondria were subsequently discovered. At present, mitochondria are known to be involved in energy production, apoptosis, calcium storage, heat production, regulation of cellular signals (Chandel 2015, Nunnary and Suomalainen 2012, Tzameli 2012), and other processes. Genetic defects in mitochondria are associated with numerous diseases in host organisms. So, it is no surprise that mitochondria are studied extensively worldwide, as shown from over 220 k literature hits in the PubMed Database with the term alone.

1.2 Eukaryogenesis

Eukaryogenesis refers to the combination of multiple evolutionary changes of cellular structures from those of prokaryote-like states into those of eukaryotic cells we know today (Roger et al. 2021). One key process that led to the rise of eukaryotic cells is the creation of the mitochondrion. The mitochondrion is hypothesized to be the result of an endosymbiosis between α-proteobacteria and a progenitor of eukaryotes three billion

years ago. The endosymbiotic proteobacteria are theorized to have evolved into the mitochondrion (Lopez-Garcia and Moreira 2015). Recent studies further suggest that the proto-mitochondrion was previously an organotrophic organism that had a symbiotic relationship with and later became endosymbiotic within an archaeal species (Spang et al. 2015, Imachi et al. 2020). The discovery of Lokiarchaea as a bridge that could shed more light on how prokaryotes evolved into eukaryotes by Spang et al., relied on concatenated sequences of metagenomic DNA data from the Atlantic Ocean floor. The author's initial discovery had cast doubts on the placement of Lokiarchaea due to conflicting phylogenetic tree topologies of some of individual sequences, but the recent isolation of a benthic species from Japanese deep sea floor sediments strongly supports the emergence of eukaryotes from an archaeal origin (Da Cunha et al. 2017, Imachi et al. 2020).

1.3 Mitochondrial Genomes

Endosymbiotic origin explains why mitochondria, in contrast to most organelles, contain their own, separate genome. Mitochondrial genomes, however, are only fragments of the related, free-living bacterial genomes, due to a major loss of gene content through three billion years of evolution. To put this in perspective, up to 8000 genes have been discovered in the genomes of modern α-proteobacteria whereas only a handful of loci can be found in mitochondria (Roger et al. 2017). The human mitochondrial genome, for instance, only encodes 13 protein genes (Ingman and Gyllensten 2001). A reduction in gene content is a common theme across most endosymbionts and mitochondria are no exception (Schneider and Ebert 2004). The underlying reasons for genome reduction are likely multi-factored. For example, a smaller genome size would provide replicative advantages of efficiency in both resources and time (Selosse et al. 2001). Gene loss is inevitable during genome reduction but the loss of genes, essential or not, may not be fatal if functional redundancy can be provided by the nuclear genome. Hence, selection can allow genome reduction to proceed without hindrance, thereby further accelerating the process. An alternative to gene loss is the transfer of genes from the mitochondrion to the nucleus, which has also been a continuous process (Brandvain and Wade 2009). Since gene transfer from the mitochondrion to the nucleus is far more likely than the reverse

due to differences in size, integrity, and genome complexity, the movement of genes is likely to be unidirectional (Selosse et al. 2001). As a result, mitochondrial genes are exposed to the possibility of being lost without significant consequences, due to functional redundancies that can be provided from newly transferred or duplicated genes in the nucleus. Furthermore, the higher evolutionary rates of mitochondrial genomes in some lineages can facilitate the above phenomena and push evolution further. For example, mitochondrial substitution rates in primate mitochondria are five to ten times those in the nucleus (Brown et al. 1982). Multiple factors accelerate mutation rates in mitochondria, including, but not limited to, the oxidative environment, drift, the high demands of adaptive evolution, and frequent genome recombination (James et al. 2016, Lynch 2010, Fritsch et al. 2014). As a result, mitochondrial genomes, at least in animals, have only retained genes encoding proteins that serve critical roles in mitochondrial function. Some may be critical enough that it was advantageous to maintain fidelity through close proximity to their destination, and some may be in a process of being transferred and/or lost from a mitochondrion in future. Examples include the different subunits of five transmembrane protein complexes in the mitochondrial membrane, Complex 1 / NADH dehydrogenase (*nad*), Complex III / cytochrome b (*cob*), Complex IV / cytochrome c oxidase (*cox*) and Complex V / ATP synthase (*atp*) (Hahn and Zuryn 2019). Mitochondrial genomes also contain tRNA and rRNA genes that participate in translation within mitochondria.

1.4 Yeast Mitochondrial Genome

Both the reduction in size and the high mutation rates make mitochondrial genomes ideal systems for investigating genome diversity and evolution. High evolutionary rates also allow the mitochondrial genomes of closely related species to vary in characteristics such as size, base composition, and topology, thereby further reinforcing the idea that mitochondrial genomes could provide critical insights for genome studies (Smith and Keeling 2015). One obvious approach to studying mitochondrial genomes is DNA sequencing. The emergence of next-generation sequencing has allowed researchers to sequence mitochondrial genomes in a time and cost-efficient manner. Over the past

decade, increasing numbers of mitochondrial genomes have been sequenced and made available in public databases (Smith and Keeling 2015). Unfortunately, the vast majority of the deposits are from metazoans and do not differ greatly in size or gene content, even between moderately related species. For example, near 3000 fish mitochondrial genome deposits can be found in the NCBI database, but the smallest genome is 15.6 kilobases (kb) and the largest is 24.9 kb. Single-celled eukaryotes such as yeasts, on the other hand, show a wider diversity in those characters, making them an attractive alternative to those of metazoans.

Compared to the mega-sized plant mitochondrial genomes, which can contain over ten million nucleotides, those of yeasts are easier to assemble and analyze in detail because many are around hundred thousand nucleotides. Also, yeast mitochondrial genomes contain more non-coding regions compared to those of vertebrates, which generally lack introns and possess smaller intergenic regions. In addition, even though the dominant pattern of inheritance in eukaryotic mitochondrial genomes is uniparental, some fungal mitochondrial genomes are inherited differently. For example, mitochondrial genomes of the model ascomycetous yeast *Saccharomyces cerevisiae* are inherited biparentally, where recombination of two parental mitochondrial genomes is possible (Wilson and Xu 2012). As a result, yeast mitochondrial genomes are more likely to show more diversity, which adds to the complexity of genome evolution studies in these organisms.

1.5 Yeasts of the genus *Metschnikowia*

Among ascomycetous yeasts, *Metschnikowia* species are characterized by their two needle-shaped ascospores, in sharp contrast to the spheroidal ascospores of most other Saccharomycetes (Lachance 2016). Within the genus *Metschnikowia*, many species that occur as haploid mating types have been collected extensively over the last 30 years and their draft genome DNA sequences are available on GenBank (Lachance et al. 2016). They offer the advantage in that their nuclear genome sequences are not complicated by heterozygosity. Compatible mating types can mate, diploidize, and undergo meiosis in a short time, which can be used as a powerful means of assigning individuals to species

that are defined on the basis of reproductive isolation. Although the haplontic *Metschnikowia* species have been found on all continents, individual species exhibit considerable endemism, making it possible to study the geographical effects of species on a global scale. Within the haplontic species, a large clade consists of species that form unusually large ascospores. These have been termed species of the large-spored clade (LSC). The LSC consists of subclades that correspond to the biogeographic regions where the species prevail (Lachance et al. 2016). The continental subclade is endemic to the American continent and can be further divided into northern and south-central species. Hawaiian endemics appear to comprise a more ancestral (paleo) and more recently derived (neo) species. Together, the continental and Hawaiian subclade constitute the large-spored species in the strict sense due to their early discovery (sensu stricto). The Arizonensis subclade coexists with the sensu stricto group and features species that form asci of intermediate size. Other subclades include species found in East Africa and Australia in one case, and South African species in the other. The evidence amassed so far indicates that the haplontic species form a monophyletic assemblage. The remaining haplontic species are termed non-LSC species. The genus also contains diplontic species that are associated with various habitats such as fruits, nectar, tree sap exudates, or marine invertebrates. These tend not to exhibit the same degree of endemism and because of their diplontic life cycle, cannot easily be assigned to species based on mating compatibility. Only haplontic species are considered in this study.

All of the publicly available nuclear genome assemblies of haplontic *Metschnikowia* species are of good quality and represent most, if not all the sequence content, as exemplified by the 90-93% complete BUSCO (Benchmarking Universal Single-Copy Orthologs, Waterhouse et al. 2017) orthologs found for *M. amazonensis, M. caudata* and *M. agaves* (Santos et al. 2020). None of these deposits, however, contain a complete mitochondrial genome in a single scaffold for reasons unclear. My preliminary studies conducted before undertaking this thesis showed that in the best cases, the mitochondrial genome was fragmented into fewer than 10 scaffolds, some of them over 20 kb long, but in the worst instances, parts of mitochondrial genomes were either incorrectly fused with nuclear contigs or fragmented into dozens of contigs that were less than 1 kb long. One common observation in the latter case was the presence of multiple small (<300 kb)

contigs that were identical in sequence except for one or a few nucleotides. These variations were commonly found in homorepeats or in the middle of those sequences. The former is speculated to be a technical problem with the sequencing itself upon encountering single base repeats and the latter is potentially single base polymorphisms. Another preliminary observation was that the connecting regions between some contigs was present in the read database (i.e., in the raw sequencing data) but not in the assembly, suggesting that read coverage may have not been high enough to assemble those regions into contigs. As a result, comparative mitochondrial genome studies of size, shape, or synteny could not be conducted from the assemblies. Furthermore, some yeast mitochondrial genes, such as *cox1*, are known to contain multiple introns that are larger than the coding regions, making it difficult to obtain complete coding sequences for phylogenetic or other studies, unless a complete mitochondrial genome is assembled.

1.6 Thesis

This thesis is an exploration of a rich, untapped source of shotgun genome sequence data aimed at reconstructing complete mitochondrial genomes. The genomes are reconstructed for 71 strains belonging to 38 *Metschnikowia* species, based on assemblies and raw reads generated in multiple studies (Lachance et al. 2016, Lee et al. 2018, Santos et al. 2020, Lee et al. 2020).

The variation encountered in genome size, synteny, and topology, as well as the distribution of introns is examined in a phylogenetic context. The work is primarily hypothesis-generating and as such constitutes pure discovery. However, among questions to be answered were (1) whether it is possible to reconstruct all mitochondrial genomes from whole genome shotgun sequence data intended primarily to assemble nuclear genomes, and (2) whether the phylogenies inferred from mitochondrial genes can reinforce those obtained from nuclear genes.

Chapter 2

2 Methods

2.1 Sequence Acquisition

Assembly and read files of 71 *Metschnikowia* strains used in the study were previously reported by Lachance et al. (2016), Lee et al. (2018), Lee et al. (2020), and Santos et al. (2020) and made available in GenBank. The strains and their GenBank accession numbers for assemblies, reads, and mitochondrial genomes are listed in **Table 1.**

Table 1: GenBank accession numbers for nuclear genome assemblies, sequence reads, and mitochondrial genome assemblies (mtDNA) for 71 *Metschnikowia* strains. Designated codes for figures are also listed.

Species	Strain Designation	Code	BioProject	Assembly	Read	mtDNA
M. aberdeeniae	UWOPS 07-202.1	abe+	PRJNA312754	NLAG01	SRX1647343	MT433106
M. aberdeeniae	SUB 05-213.18	abe-	PRJNA312754	NKZB01	SRX1647359	MT433105
M. agaves	UWOPS 92-207.1	aga+	PRJNA411829	VATI01	SRX11095251	MT442067
M. agaves	UWOPS 92-210.1	aga-	PRJNA411829	VATH01	SRX11095250	MT442068
M. amazonensis	UFMG-CM-Y6309	ama+	PRJNA411829	VATE01	SRX11095253	MT421952
M. amazonensis	UFMG-CM-Y6307	ama-	PRJNA411829	VATD01	SRX11095252	MT421953
M. arizonensis	UWOPS 99-103.3.1	ari+	PRJNA312754	NLAT01	SRX1647355	MT449701
M. arizonensis	UWOPS 99-103.4	ari-	PRJNA312754	NLAU01	SRX1647356	MT449702
M. borealis	SUB 99-207.1	bor+	PRJNA312754	NKZC01	SRX2982267	MT433104
M. borealis	UWOPS 96-101.1	bor-	PRJNA312754	NLAR01	SRX2982268	MT433103
M. bowlesiae	UWOPS 04-243x5	bow+	PRJNA312754	NLAE01	SRX2982266	MT444157
M. bowlesiae	UWOPS 12-611.1	bow-a	PRJNA312754	NLAJ01	SRX1647345	MT447077
M. bowlesiae	UWOPS 12-619.1	bow-b	PRJNA312754	NLAK01	SRX1647346	MT447074
M. caudata	EBDC CdV SA08.1	cau+	PRJNA411829	VATG01	SRX11095254	MT421957
M. caudata	EBDC CdV SA57.2	cau-	PRJNA411829	VATF01	SRX11095255	MT421958
M. cerradonensis	UFMG 03-T68.1	cer+	PRJNA312754	NKZE01	SRX1647361	MT442069
M. cerradonensis	UFMG 03-T67.1	cer-	PRJNA312754	NKZD01	SRX1647360	MT442070
Metschnikowia sp. M2Y3	EBD CdV M2Y3	cla+	PRJNA312754	NKYV01	SRX2982230	MT433117
M. colocasiae	UWOPS 03-134.2	col+	PRJNA312754	NKZQ01	SRX1647332	MT449698
M. colocasiae	UWOPS 03-202.1	col-	PRJNA312754	NKZV01	SRX2985829	MT447078
M. continentalis	UWOPS 96-173	con+	PRJNA312754	NLAS01	SRX1647354	MT442060
M. continentalis	UWOPS 95-402.1	con-	PRJNA312754	NLAO01	SRX1647349	MT442071
Metschnikowia sp. 03-147.1	UWOPS 03-147.1	cos-	PRJNA312754	NKZR01	SRX2982270	MT442053
M. cubensis	MUCL 45753	cub+	PRJNA312754	NKZA01	SRX1647358	MT449703
M. cubensis	MUCL 45751	cub-	PRJNA312754	NKYZ01	SRX1647353	MT449704
M. dekortorum	UWOPS 01-142b3	dek+	PRJNA312754	NKZM01	SRX1647329	MT442072
M. dekortorum	UWOPS 01-522a5	dek-	PRJNA312754	WFIU01	SRX6979404	MT442062
M. dekortorum	UFMG-CM-Y6306	dekY	PRJNA312754	WFIW01	SRX6979406	MT442061
M. drakensbergensis	EBD-CdVSA09-2	dra+	PRJNA312754	NKYW01	SRX1647324	MT442055
M. drakensbergensis	EBD-CdVSA10-2	dra-	PRJNA312754	NKYX01	SRX1647334	MT442054
M. drosophilae	UWOPS 83-1143.1	dro+	PRJNA411829	NXHC01	SRX3206839	MT421955
M. drosophilae	UWOPS 83-1135.3	dro-	PRJNA411829	NXHB01	SRX3206836	MT421956
Metschnikowia sp. 13-106.1	UWOPS 13-106.1	flo+	PRJNA312754	NLAL01	SRX2982269	MT433102
M. hamakuensis	UWOPS 04-207.1	ham+	PRJNA312754	NLAB01	SRX1647340	MT449696
M. hamakuensis	UWOPS 04-199.1	ham-	PRJNA312754	NKZZ01	SRX1647339	MT449697
M. hawaiiana	UWOPS 91-698.3	han+	PRJNA411829	NXGY01	SRX3206837	MT421961
M. hawaiiensis	UWOPS 91-745.2	haw+	PRJNA312754	NLAN01	SRX1647348	MT442059
M. hawaiiensis	UWOPS 87-2203.2	haw-	PRJNA312754	NLAM01	SRX1647347	MT442058

M. hibisci	UWOPS 95-797.2	hib+	PRJNA312754	NLAP01	SRX1647350	MT447079
M. hibisci	UWOPS 95-805.1	hib-	PRJNA312754	NLAQ01	SRX1647351	MT447080
M. ipomoeae	UWOPS 10-104.1	ipo+	PRJNA312754	NLAI01	SRX2982263	MT433113
M. ipomoeae	UWOPS 99-324.1	ipo-a	PRJNA312754	NLAV01	SRX2982264	MT433112
M. ipomoeae	UWOPS 01-141c3	ipov	PRJNA312754	NKZK01	SRX2982273	MT433111
M. kamakouana	UWOPS 04-112.5	kam+	PRJNA312754	NKZX01	SRX1647337	MT433119
M. kamakouana	UWOPS 04-206.1	kam-	PRJNA312754	NLAA01	SRX2982227	MT433120
M. kipukae	UWOPS 00-669.2	kip-	PRJNA312754	NKZJ01	SRX2982274	MT421951
M. lacustris	UWOPS 12-619.2	lac+	PRJNA312754	WFIT01	SRX6979403	MT442063
M. lacustris	UWOPS 03-172.2	lac-	PRJNA312754	NKZU01	SRX1647335	MT442065
M. lacustris	UWOPS 03-167b3	lacb	PRJNA312754	WFIV01	SRX6979405	MT442064
M. lochheadii	UWOPS 03-167a3	loc+	PRJNA312754	NKZT01	SRX1647333	MT433101
M. lochheadii	UWOPS 99-661.1	loc-	PRJNA312754	NLAW01	SRX1647357	MT433100
M. matae var. *maris*	UFMG-CM-Y397	mar-	PRJNA312754	NKZH01	SRX1647327	MT433114
M. matae var. *matae*	UFMG-CM-Y395	mat+	PRJNA312754	NKZG01	SRX1647326	MT433116
M. matae var. *matae*	UFMG-CM-Y391	mat-	PRJNA312754	NKZF01	SRX1647325	MT433115
M. mauinuiana	UWOPS 04-190.1	mau+	PRJNA312754	NKZY01	SRX1647338	MT449699
M. mauinuiana	UWOPS 04-110.4	mau-	PRJNA312754	NKZW01	SRX1647336	MT449700
Metschnikowia sp. 00-154.1	UWOPS 00-154.1	mer-	PRJNA312754	NKZI01	SRX2982229	MT442066
Metschnikowia orientalis	UWOPS 99-745.6	ori+	PRJNA411829	NXGZ01	SRX3206838	MT421959
Metschnikowia orientalis	UWOPS 05-269.1	ori-	PRJNA411829	NXHA01	SRX3206835	MT421960
Metschnikowia sp. 04-218.3	UWOPS 04-218.3	pal+	PRJNA312754	NLAC01	SRX2982272	MT449695
Metschnikowia sp. 04-226.1	UWOPS 04-226.1	pil-	PRJNA312754	NLAD01	SRX2982271	MT447081
M. proteae	EBD-T1Y1	pro+	PRJNA312754	NKYY01	SRX1647342	MT442057
M. proteae	EBD-A10Y1	pro-	PRJNA312754	NKYU01	SRX1647323	MT442056
*M. santacecilia*e	UWOPS 01-517a1	sce+	PRJNA312754	NKZN01	SRX1647330	MT433110
*M. santacecilia*e	UWOPS 01-142b1	sce-	PRJNA312754	NKZL01	SRX1647328	MT433109
M. shivogae	UWOPS 04-310.1	shi+	PRJNA312754	NLAF01	SRX1647341	MT433108
M. shivogae	UWOPS 07-203.2	shi-	PRJNA312754	NLAH01	SRX1647344	MT433107
M. similis	UWOPS 03-158.2	sim+	PRJNA312754	NKZS01	SRX1647523	MT447075
M. similis	UWOPS 03-133.4	sim-	PRJNA312754	NKZP01	SRX1647331	MT447076
Metschnikowia sp. 01-655c1	UWOPS 01-655c1	sma-	PRJNA312754	NKZO01	SRX2982265	MT433118
M. torresii	CBS 5152	tor-	PRJNA411829	NXHD01	SRX3206834	MT421954

2.2 Mitochondrial Genome Construction

A general flowchart of the mitochondrial genome construction pipeline from nuclear genome data is illustrated in **Figure 1**. Candidate mitochondrial reads and assemblies were identified and isolated by comparing read coverages through the Bowtie2 short read mapper v.7.2.1 plug-in of Geneious v.8.1.7. Read coverages in most species were later corroborated with the SPAdes v.3.12 assembler program. The isolated reads were reassembled through the Velvet v.1.2.10 plug-in of Geneious to create mitochondrial contigs. The resultant contigs were extended and bridged to other contigs whenever necessary, by iterative rounds of sequencing read mapping using the Geneious read mapper with a medium-low sensitivity setting and five iterations. When assembled mitochondrial genomes could not be further refined *in silico*, PCR amplification and Sanger sequencing were used to test the linearity of certain genomes.

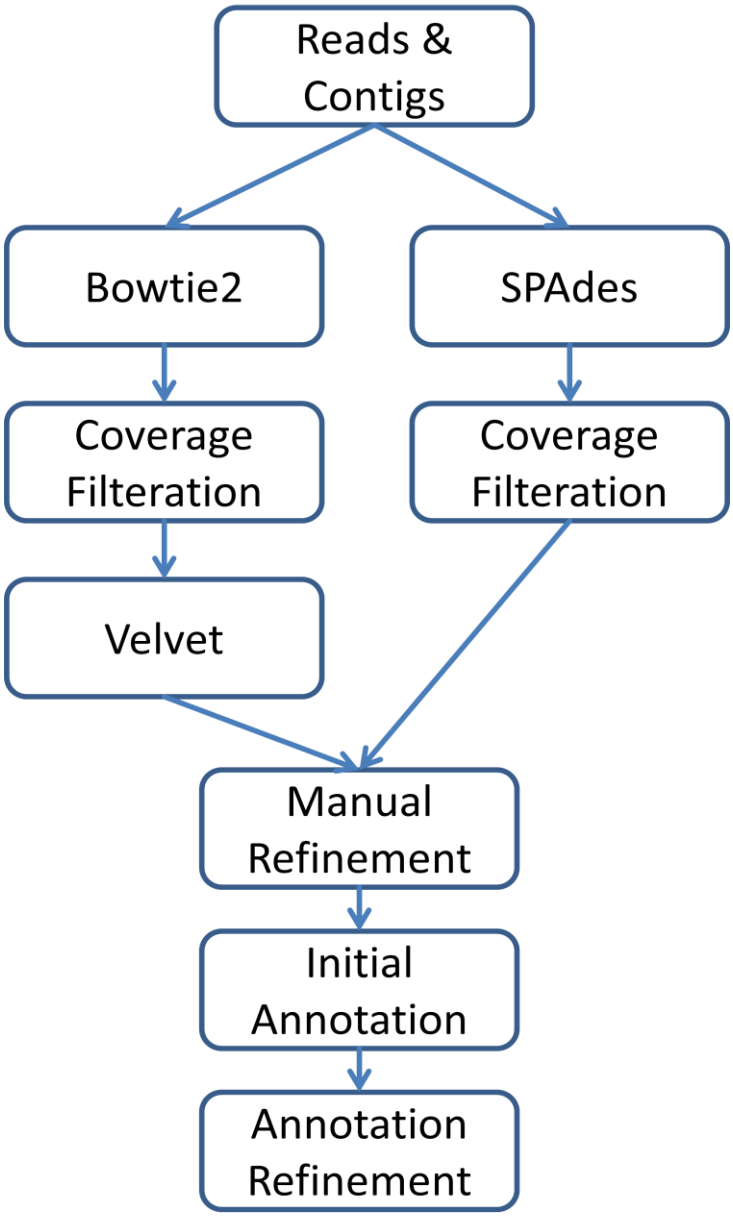

Figure 1: Example workflow pipeline of mitochondrial genome construction for *Metschnikowia* species. Manual refinement includes PCR amplifications and Sanger sequencing of amplicons.

2.3 Annotation

Mitochondrial genes were first annotated with MFannot (Lang et al. 2007) and then refined through alignments with orthologs of related species in MEGA5 (Tamura et al. 2011). Boundaries of each candidate gene suggested by MFannot were extended to a first Methionine codon following an upstream stop codon and to a stop codon downstream of the last suggested exon sequence. Candidate protein sequences using the yeast mitochondrial code (translation table 3) were subsequently prepared, aligned and compared with orthologous protein sequences of known ascomycetes obtained from public databases such as UniProt (www.uniprot.org), the *Saccharomyces* Genome Database (www.yeastgenome.org) and the *Candida* Genome Database (www.candidagenome.org). For further refinement, the genes of *Candida albicans*, *Saccharomyces cerevisiae*, *Candida lusitaniae*, and *Meyerozyma guilliermondii* were used. The orthologous genes were aligned with those of other *Metschnikowia* species used in the study with MAFFT v.7.017 via Geneious (scoring matrix = 200PAM, gap open penalty = 1.53, and offset value = 0.123) for the final refinements for genes, exons, and introns boundaries. As sequence alignments of rRNA genes with those of other non-*Metschnikowia* species were inconclusive in localizing the precise ends of the rRNA loci, the exact boundaries of rRNA coding regions were estimated by aligning rRNA loci (*rns* and *rnl* were separately aligned) determined by MFannot plus their surrounding regions (~100 bps from both ends) for all 71 haplontic *Metschnikowia* strains used in this study and trimming ends that are not conserved among *Metschnikowia* strains. Introns were individually queried with the GenBank web BLAST search engine (https://blast.ncbi.nlm.nih.gov/Blast.cgi) (Altschul et al. 1990) to test for the presence or absence of intron-encoded proteins such as endonucleases and maturases. Details of the parameters used for phylogenetic inference are given in the captions to the relevant figures.

2.4 PCR Analysis

InVitroGen Platinum II Hot-Start PCR Master Mix Taq (2X) was supplemented with primers (0.2μM), whole cells or purified genomic DNA, and water to the final reaction volume. Primers were designed using the Primer3 plug-in of Geneious and aimed to target either 1) confirmed regions of mtDNA as a control or 2) adjacent to putative telomere sequences to extend and sequence the region beyond telomere regions. Cycling conditions consisted of a 2 min initial denaturation step at 92ºC, followed by 35 cycles of denaturation for 15 sec, annealing at 55ºC for 15 sec, and a 15 sec/kb extension at 68ºC, adjusted as a function of the expected length of the sequence to be amplified. The presence of a PCR product was determined by agarose gel electrophoresis. Purified PCR products were sent to the London Regional Genomic Centre, Robarts Institute, for sequencing.

PCR analysis was utilized when 1) there were ambiguous nucleotides within mitochondrial assemblies that were unable to identify with reads and/or contig data alone or when 2) candidates of linear mitochondrial genomes were attempted to extend and sequence beyond ends of assemblies to test their linearities. Primer sequences are listed in **Table S1**.

2.5 Phylogenetic Analysis

Phylogenies of all sequenced *Metschnikowia* species as well as 71 haplontic *Metschnikowia* strains were constructed from aligned concatenation of genes using the RAxML plug-in of Geneious program to provide a reference for comparing diversities of genome characters.

2.6 Ancestral Gene Order Construction

Progressive Contiguous Ancestral Regions (ProCARs) program under default parameters on Ubuntu platform was used to calculate potential ancestral states of mitochondrial gene orders of unknown ancestors of haplontic *Metschnikowia* species (Perrin et al. 2015).

Chapter 3

3 Results and Discussion

3.1 Assembled Mitochondrial Genomes

The mitochondrial genomes of 71 strains belonging to 38 species were successfully assembled. The complete assemblies were deposited in GenBank (Lee et al. 2020). GenBank accession numbers for each mitochondrial genome can be found in **Table 1**. An initial report, which mainly focused on genome size diversity, has been published (Lee et al 2020). Overall genome diagrams for 71 strains are as visualized by Geneious and can be found in the appendix **(Figs. S1-S71)**. A general summary of the sizes, shapes, base compositions, spacer and intron abundances, and gene orders is given in **Figure 2.**

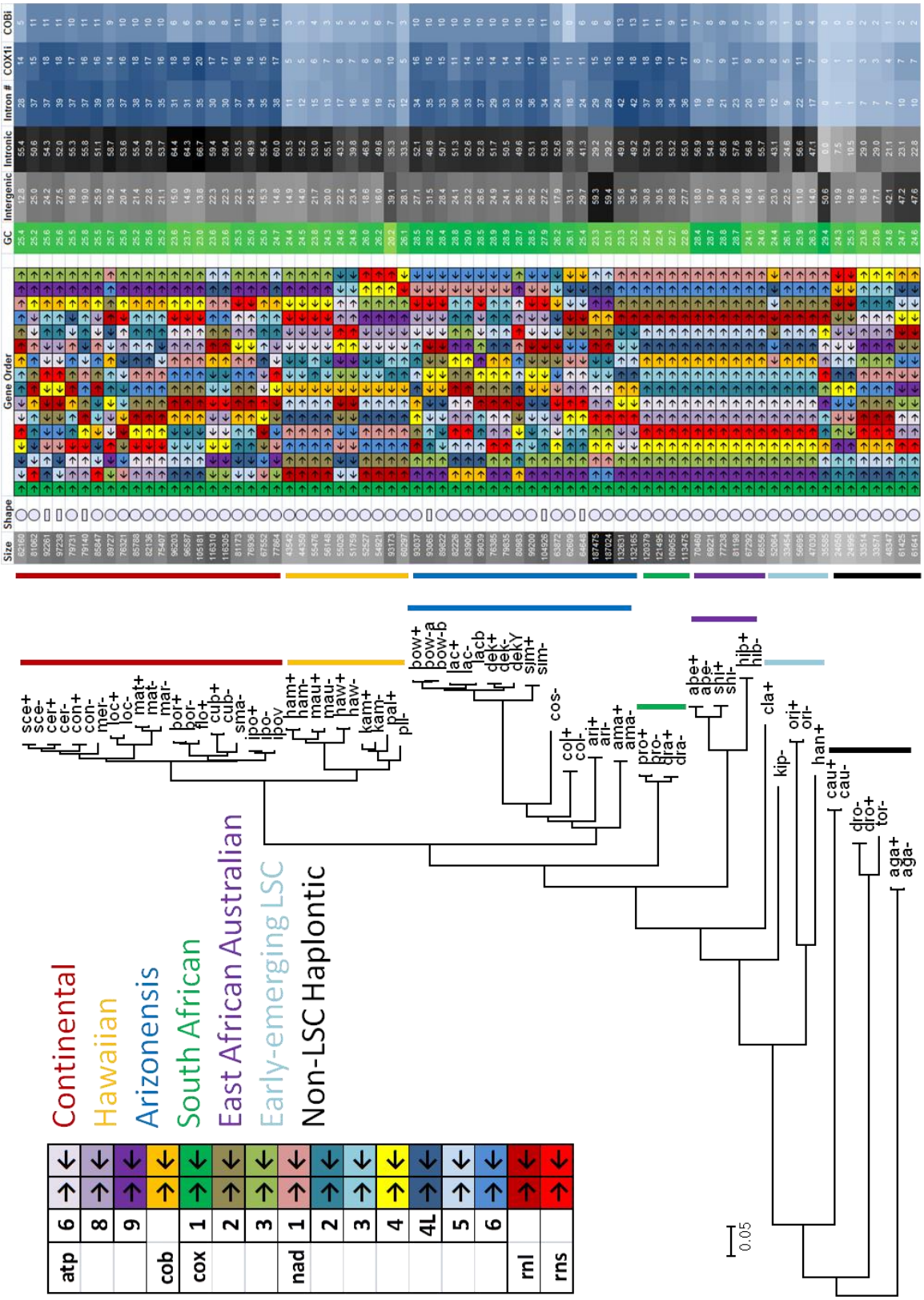
Size
Shape
Gene Order
GC
Intergenic
Intronic
Intron #
COX1ti
COBi
Continental
Hawaiian
Arizonensis
South African
East African Australian
Early-emerging LSC
Non-LSC Haplontic
atp 6
8
9
cob
cox 1
2
3
nad 1
2
3
4
4L
5
6
rnl
rns
sce+
sce-
cer+
cer-
con+
con-
mer-
loc+
loc-
mat+
mat-
mar-
bor+
bor-
flo+
cub+
cub-
sma-
ipo+
ipo-
ipov
ham+
ham-
mau+
mau-
haw+
haw-
kam+
kam-
pal+
pil-
bow+
bow-a
bow-b
lac+
lac-
lacb
dek+
dek-
dekY
sim+
sim-
cos-
col+
col-
ari+
ari-
ama+
ama-
pro+
pro-
dra+
dra-
abe+
abe-
shi+
shi-
hip+
hip-
cla+
kip-
ori+
ori-
han+
cau+
cau-
dro-
dro+
tor-
aga+
aga-
0.05

Figure 2: A general summary of the sizes (bp) , shapes, base compositions (%), spacer and intron abundances (%), gene orders, total number of introns, and number of introns in *cox1* and *cob* genes of mtDNA of 71 *Metschnikowia strains.*

3.2 Gene content

Three *cox* genes (1/2/3), three *atp* genes (6/8/9), seven *nad* genes (1/2/3/4/4L/5/6), and one *cob* gene, as well as two rRNA (*rnl/rns*) genes were found in all 71 mitochondrial genomes of the *Metschnikowia* species analyzed, thereby showing that the gene content is identical across all species **(Table** 2). Differences were found in the number of tRNA genes and in the copy number of certain genes. The presence of *nad* genes and the absence of a *var1* gene is characteristic of the CUG-ser1 clade within the Saccharomycetes and distinguishes its members from *Saccharomyces cerevisiae* (Freel et al. 2015). The CUG-Ser1 clade is one of two groups of yeasts with a mutated tRNA that causes the CUG codon to code for serine instead of leucine (Riley et al. 2016, Krassowski et al. 2018). The clade contains the families Metschnikowiaceae and Debaryomycetaceae, which includes the well-known pathogen *Candida albicans.* In contrast, *S. cerevisiae* belongs to the family Saccharomycetaceae, where CUG is translated to leucine. The translation of CUG to serine in haplontic *Metschnikowia* species was suspected as the mutation had been identified in the distant relative *Metschnikowia bicuspidata* (Riley et al. 2016). It should be noted that *Metschnikowia bicuspidata*, unlike all other members of the CUG-Ser1 group tested in Riley et al. (2016), lacked a ser identity element in the $tRNA_{CAG}$ sequence such that it was unclear whether all members of the genus *Metschnikowia* featured the mutated codon usage. Gordon et al. (2019) recently confirmed that *M. borealis* and several other *Metschnikowia* species use alternative codons in the nucleus.

Table 2: Copy number of mitochondrial genes found in 71 *Metschnikowia* mitochondrial genomes. See Table 1 for a key to the strain codes.

Code	ATP			COX			NAD							rRNA		COB
	6	8	9	1	2	3	1	2	3	4	4L	5	6	rnl	rns	
abe+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
abe-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
aga+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
aga-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ama+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ama-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ari+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ari-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
bor+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
bor-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
bow+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
bow-a	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
bow-b	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
cau+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
cau-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
cer+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
cer-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
cla+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
col+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
col-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
con+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
con-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
cos-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
cub+	1	1	1	1	1	1	1	2	2	1	1	1	1	1	1	1
cub-	1	1	1	1	1	1	1	2	2	1	1	1	1	1	1	1
dek+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
dek-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
dekY	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
dra+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
dra-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
dro+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
dro-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
flo+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ham+	2	2	2	1	1	1	1	2	2	2	1	1	1	1	2	1
ham-	2	2	2	1	1	1	1	2	2	2	1	1	1	1	2	1
han+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
haw+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
haw-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

hib+	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1
hib-	1	1	1	1	1	1	1	1	2	1	1	1	1	1	1	1
ipo+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ipo-a	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ipov	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
kam+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
kam-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
kip-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
lac+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
lac-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
lacb	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
loc+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
loc-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
mar-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
mat+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
mat-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
mau+	2	2	2	1	1	2	1	2	2	2	1	1	1	1	2	1
mau-	2	2	2	1	1	2	1	2	2	2	1	1	1	1	2	1
mer-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ori+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
ori-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
pal+	2	2	2	1	1	2	1	2	2	1	1	1	1	1	1	1
pil-	2	2	2	1	1	2	1	2	2	1	1	1	1	1	1	1
pro+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
pro-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
sce+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
sce-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
shi+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
shi-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
sim+	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
sim-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
sma-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
tor-	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1

3.3 Phylogeny

A phylogenetic analysis using a concatenation of the largest orthologous nuclear genes was constructed to provide a robust estimate of the phylogeny of the species being studied (**Figure 3** and **Figure 4**). The analysis showed that *M. agaves, M. caudata*, *M. drosophilae* and *M. torresii*, which had not yet been sequenced for the robust phylogeny of Lachance et al. (2016), do not cluster with species of the large-spored clade (LSC) and that the sequence divergence between the two groups is large (**Figure 3**). The four species instead form separate clades with other distantly related, diplontic *Metschnikowia* species such as *M. pulcherrima* and *M. fructicola.* This phylogeny identifies *M. orientalis* and *M. hawaiiana* as the outermost members of the so-called large-spored clade, as suggested by Lee et al. (2018). The two species will be treated as such for the rest of the thesis.

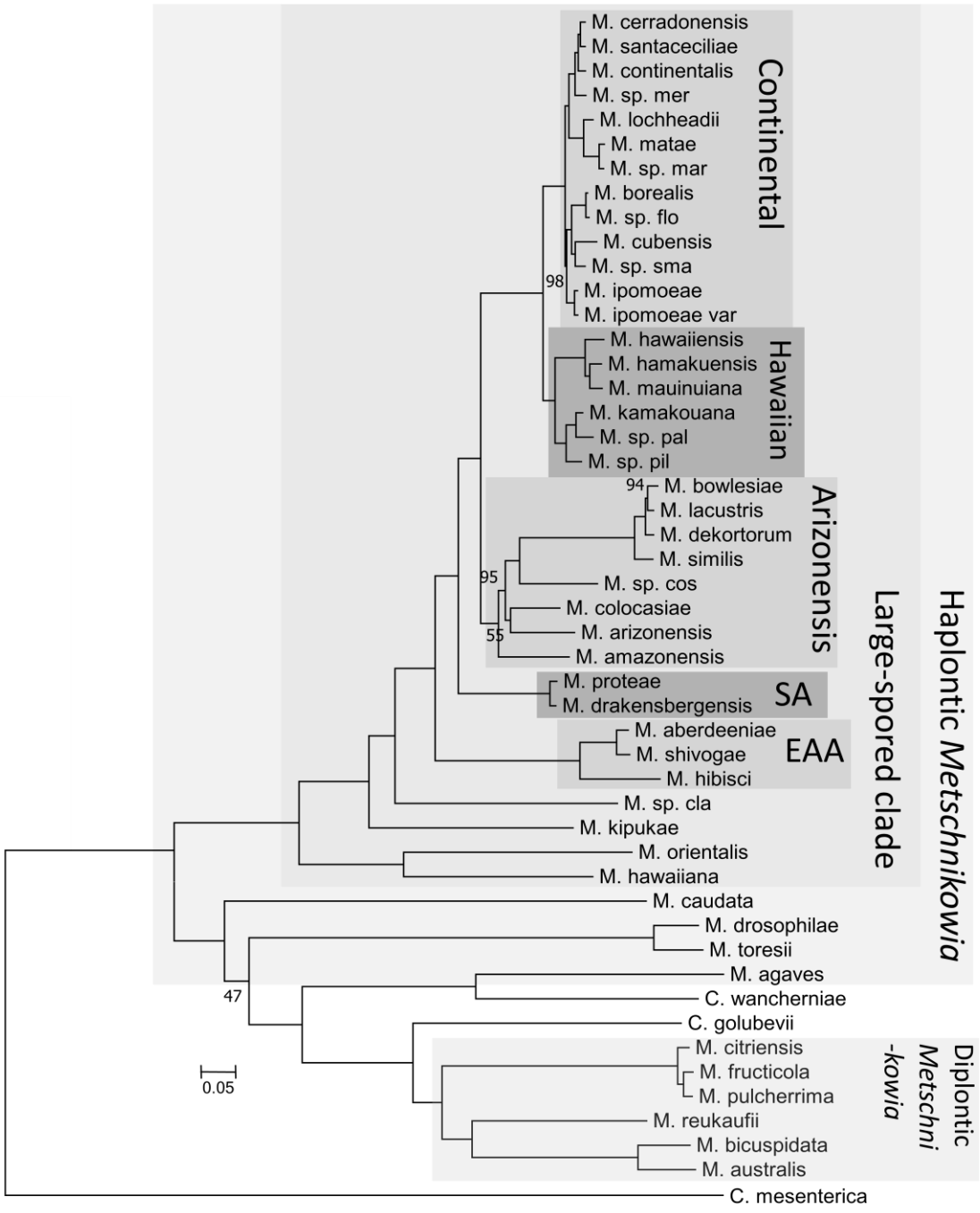

Figure 3: Phylogenetic tree of *Metschnikowia* species whose draft genomes are available on GenBank. Each species is represented by a single strain. The tree was generated using maximum likelihood with the General Time Reversible distance correction and gamma-distributed rates on a concatenation of nuclear genes that shared orthologs in all species included in the analysis. The chosen genes were based on Lee et al. (2020) and from that data set, a group of genes that could be found in

every *Metschnikowia* species was selected. The dataset spanned over 194,043 aligned positions. *Candida mesenterica* was chosen as an outgroup. Species from *M. agaves* and above are haplontic, and those from *M. citriensis* and below are diplontic. Species from *M. hawaiiana* and above are referred to as species of the large-spored clade (LSC). Subclades within the LSC are also labeled in accordance with Lachance et al. (2016). The names of the South African (SA) and East African and Australian (EAA) subclades are abbreviated. Bootstraps from 100 iterations are shown only for values less than 100%.

The near-ubiquity of mitochondria in eukaryotes brought considerable interest in using mitochondrial genes as phylogenetic markers for taxonomic purposes. For example, there are numerous reports of the use of cytochrome oxidase 1 gene sequences to build phylogenies of a wide variety of eukaryotes, as shown from the over 2.5 million NCBI deposits in 2017 as well as the near five thousand literature hits in the PubMed database when a combination of cytochrome oxidase 1 and taxonomy are entered (Hebert et al. 2003, Molitor et al. 2010, Porter and Hajibabei 2018). It was therefore of interest to determine the usefulness of mitochondrial genes in inferring a phylogeny of *Metschnikowia* species, and how such a phylogeny would compare with a robust phylogeny built on nuclear genes such as that published by Lachance et al. (2016) and the updated trees constructed in this thesis (**Figure 3 and Figure 4**). As mitochondrial and nuclear genomes are physically isolated, it was expected that the two have evolved under different constraints, for example different mutation rates, which might result in different branch lengths and saturation profiles, but there was no obvious reason why one would have expected different topologies.

A phylogenetic analysis of 14 concatenated mitochondrial coding sequences resulted in a tree (**Figure 4**) that differs in topology from trees inferred from nuclear genes (**Figure 5**). A direct comparison is given in **Figure 6**. Although in most cases the species were grouped in similar subclades, the internal structure of the subclades differed considerably. Noteworthy is the position of the Continental subclade closer to the Arizonensis subclade

in the mitochondrial phylogeny and closer to the Hawaiian subclade in the nuclear phylogeny (**Figure 6**). In addition, the phylogenetic tree based on mitochondrial genes often failed to place conspecific strains in the Continental and Arizonensis clades as sisters, unlike phylogenies inferred from nuclear genomes. The most surprising discrepancy was the placement of *M. arizonensis* outside of the large-spored species, evidence that their mitochondria have experienced a unique, puzzling evolutionary path (**Figure 6**). One possible mechanism would be horizontal transfer of mitochondria during hybridization, which has been reported in *Saccharomyces* yeast (Marinoni et al. 1999). It should be noted, however, that hybrids producing viable offspring are rare and in many cases only mitochondria from one parent are present in those hybrids and hybrid genomes are often unstable, making them unlikely to arise (Marinoni et al. 1999).

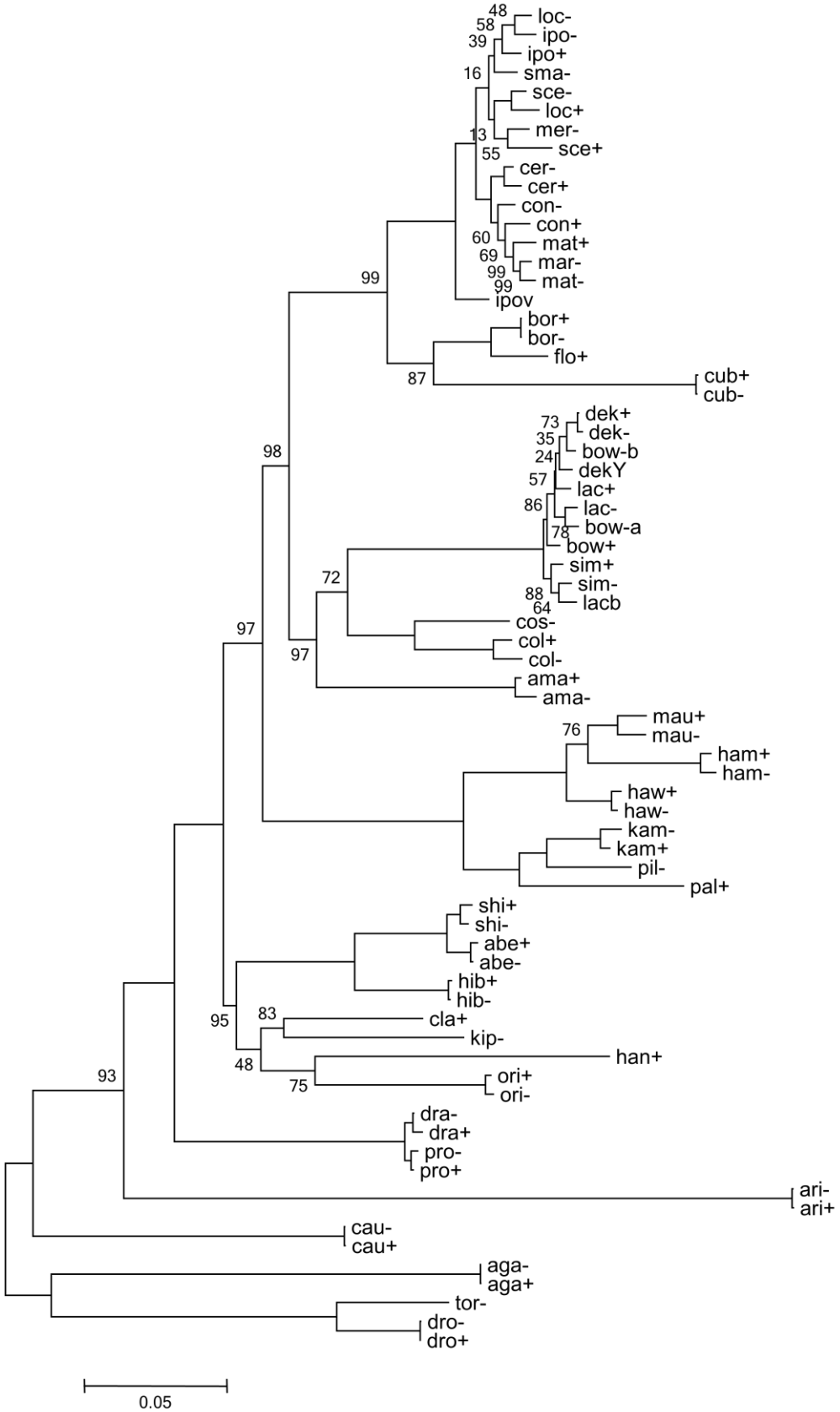
48
58
39
16
13
55
60
69
99
99
99
98
97
97
72
73
35
24
57
86
78
88
64
76
95
83
48
75
93
87
loc-
ipo-
ipo+
sma-
sce-
loc+
mer-
sce+
cer-
cer+
con-
con+
mat+
mar-
mat-
ipov
bor+
bor-
flo+
cub+
cub-
dek+
dek-
bow-b
dekY
lac+
lac-
bow-a
bow+
sim+
sim-
lacb
cos-
col+
col-
ama+
ama-
mau+
mau-
ham+
ham-
haw+
haw-
kam-
kam+
pil-
pal+
shi+
shi-
abe+
abe-
hib+
hib-
cla+
kip-
han+
ori+
ori-
dra-
dra+
pro-
pro+
ari-
ari+
cau-
cau+
aga-
aga+
tor-
dro-
dro+
0.05

Figure 4: Phylogenetic tree of 71 individual *Metschnikowia* strains inferred from a concatenation of coding sequences of 14 mitochondrial genes. Ribosomal RNA genes *rns* and *rnl* were not included. The tree was generated using maximum likelihood with the General Time Reversible distance correction and gamma-distributed rates on an alignment of 12650 positions. Bootstraps from 100 iterations are shown only for values less than 100%.

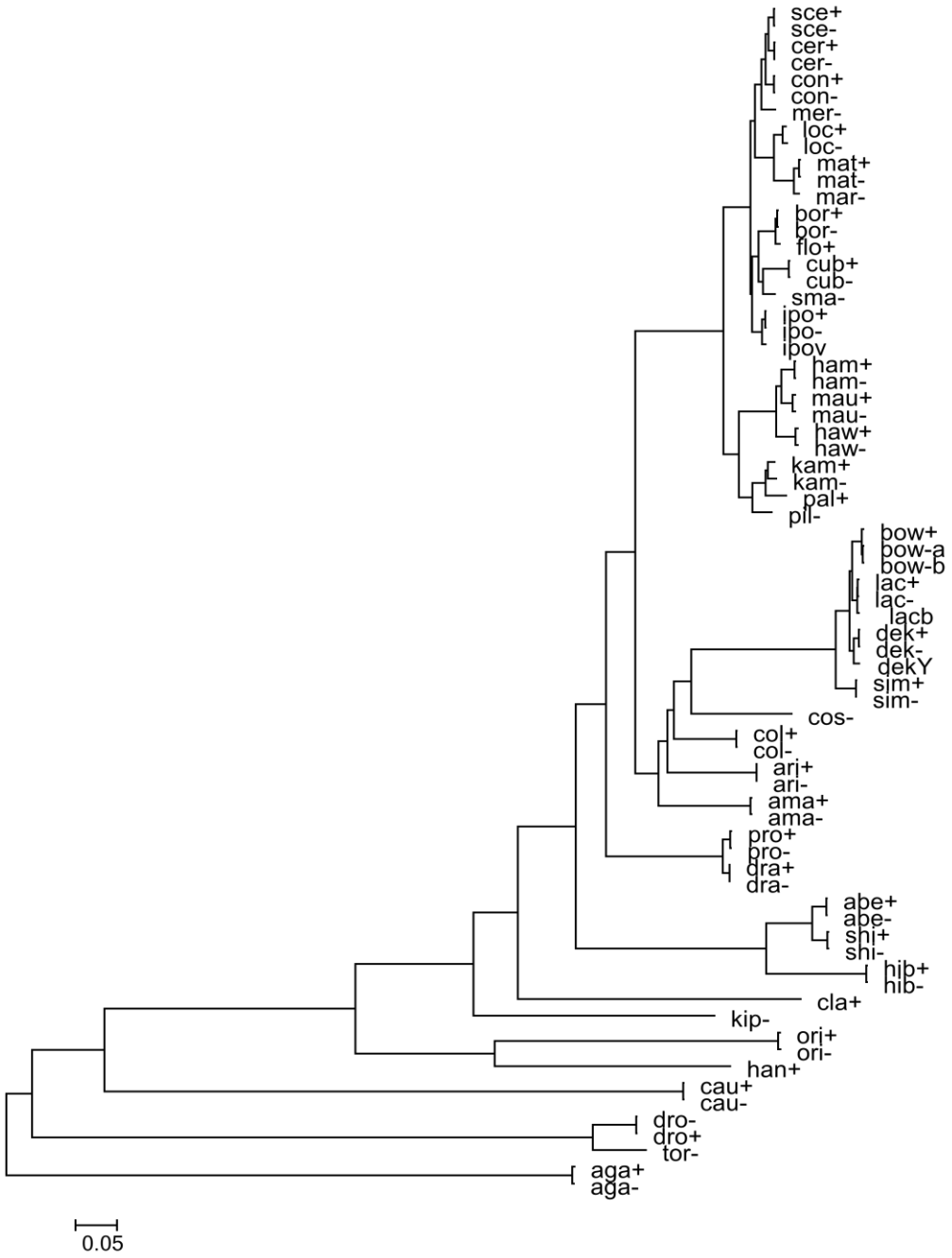

Figure 5: Phylogenetic tree showing all 71 *Metschnikowia* strains used in this study. The alignment was a concatenation of the 100 largest orthologous nuclear genes found in all strains. The tree was generated using maximum likelihood with the General Time Reversible distance correction and gamma-distributed rates on an alignment totaling 503,097 positions. Bootstraps from 100 iterations are shown only for values less than 100%.

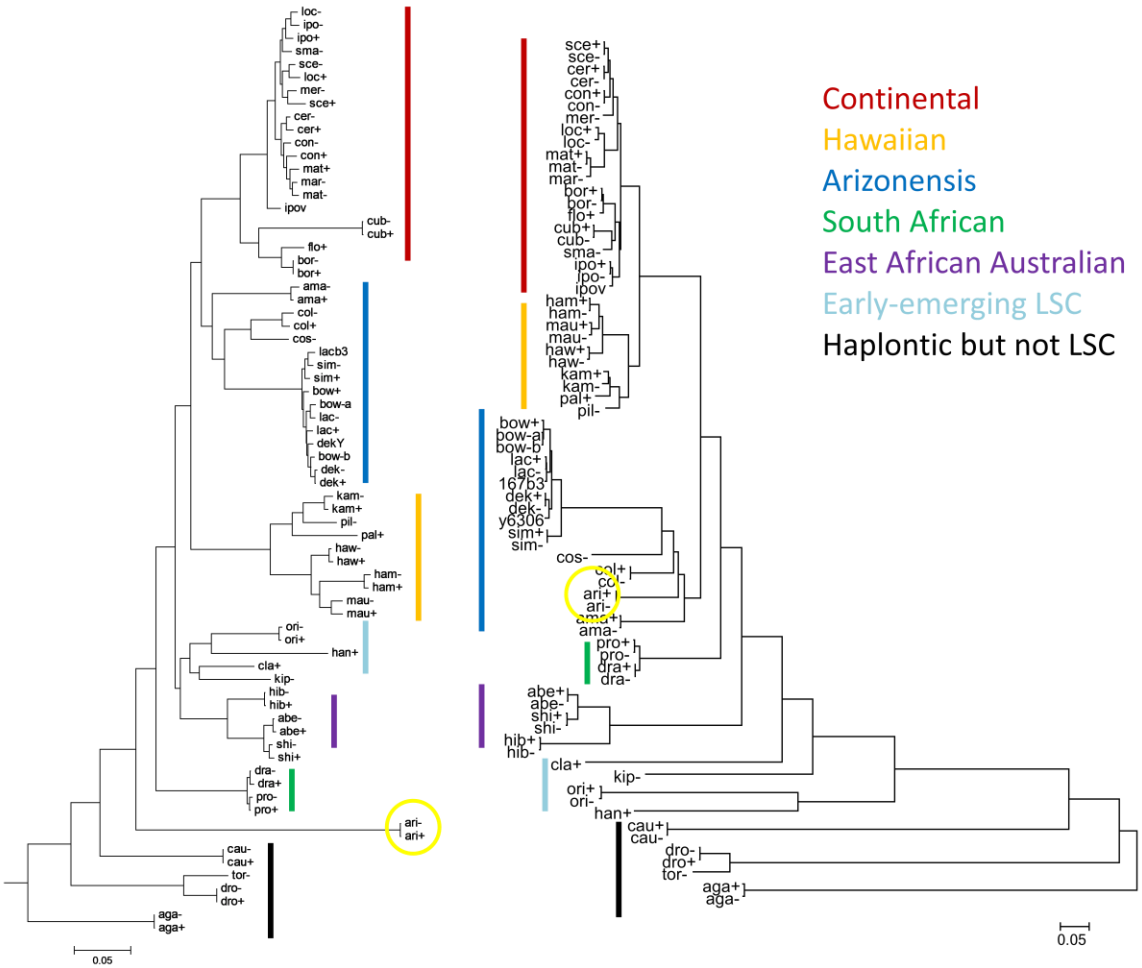

Figure 6: Comparison of phylogenetic trees built from mitochondrial (left) and nuclear (right) genes. Trees are modified from Figure 4 and 5 but retain their original topologies. Coloured bars identify subclades using the nomenclature introduced by Lachance et al. (2016). Discrepant placements of *M. arizonensis* are highlighted with yellow circles.

In individual phylogenetic trees using genes with larger (>1 kb) coding regions (*cox1*, *cob*, *nad4* and *nad5*), the topologies were broadly divergent (data not shown). It is therefore not entirely surprising that a phylogenetic tree inferred from a concatenation of all mitochondrial genes was not consistent with species relatedness established from nuclear gene phylogenies or even from mating success (Lachance et al. 2016, Lee et al. 2018). Incongruence between individual mitochondrial gene phylogenies shows that

mitochondrial genes would not be the first choice for reliably depicting relatedness among large-spored *Metschnikowia* species.

Discrepancies between the two phylogenies were not completely unexpected. Marinoni and Lachance (2004) compared phylogenies of a small group of *Metschnikowia* species of the Continental and Hawaiian subclades using three regions of the nuclear ribosomal DNA, a partial sequence of the mitochondrial small subunit rRNA, and restriction profiles of entire mitochondrial genomes. The five trees were different and those obtained from mitochondrial data bore the least resemblance with currently available multigene phylogenies (e.g., Lachance et al. 2020). Interestingly, the best phylogeny obtained by Marinoni and Lachance (2004) arose from an alignment of rDNA intergenic spacer sequences. It becomes clear that a larger amount of data does not necessarily guarantee better fine-grained phylogenies, although it has been argued that a phylogeny based on a sizeable proportion of the nuclear genome should be regarded as our best estimate of the true evolutionary history of the species (Lachance et al. 2016).

Phylogenetic disagreements between nuclear and organelle genes have been observed previously in other organisms such as plants (Leebens-Mack et al. 2019). Two possible explanations exist. First, mitochondrial genomes may have experienced different evolutionary histories from their nuclear counterparts. In such cases, the effects of different evolutionary paths should be reflected in other characteristics of the mitochondrial genomes as well. Second, mitochondrial genomes may have experienced saturation where rapid changes in sequences ended up masking repeated changes at the same sites and making it difficult for inference algorithms to identify the phylogenetic signal correctly, especially at the higher levels of the trees. Considering that *S. cerevisiae* contains multiple mitochondria per cell and multiple mtDNA copies per mitochondrion, *Metschnikowia* species may follow the same pattern, although this has yet to be confirmed (Solieri 2010). If that is so, mitochondrial genomes are expected to undergo frequent recombination between mtDNAs, which could have contributed to accelerating the rate of evolution.

The lack of congruence between phylogenies inferred from nuclear and mitochondrial genes observed in *Metschnikowia* species validates the choice, made by Kurtzman in 1989, to adopt ribosomal RNA genes as barcodes in yeast taxonomy, and not the *cox1* gene, which is more commonly used for other types of organisms, in particular insects (Hebert et al. 2003). The choice was guided in part by the relative ease of amplification and sequencing of the D1/D2 regions of the large subunit nuclear rRNA gene, which went on to serve as the standard yeast barcode (Kurtzman and Robnett 1998). Barcodes do not provide answers to all questions, however, and there is a need for supplementary approaches (Lachance et al. 2016, Lachance et al. 2020).

3.4 Mitochondrial tRNAs

The tRNA genes were in general located in the vicinity of either the *rnl* or the *rns* gene loci, which code for the large and small subunit ribosomal RNA genes, respectively. The clustering of tRNA and rRNA loci suggests that selection has favoured the simultaneous expression of both tRNA and rRNA genes through proximity, which is likely to prevent the disruption of protein translation. A disparity in tRNA distribution between the two rRNA loci was observed. Both the upstream and downstream regions of the *rns* locus are generally enriched in tRNA loci, without any protein genes in between. As for the *rnl* locus, on the other hand, nearby tRNA genes were separated by protein genes. For example, 15 of 25 tRNAs found in the mitochondrial genome of *M. borealis* were found adjacent to *rns* whereas only one tRNA coding for alanine was found next to *rnl* before protein genes (**Figs. S9 and S10**). Many non-alanine tRNA loci were located near *rnl* but some were intercepted by protein genes. Interestingly, the alanine-coding tRNA gene was found adjacent and in close proximity to *rnl* in all cases (<10 bases), suggesting that the two are co-transcribed and that their co-localization is ancestral. The reason for the disparity of tRNA proximity to the two rRNA loci is not clear. Some exceptions exist where tRNA genes are found far away from both the *rnl* and *rns* loci (e.g., *M. colocasiae*), which are mostly due to genome rearrangements.

Excluding cases of regional duplications, differences in the number of tRNA genes between species were minimal, with a range of 25 to 27 (**Table 3Error! Reference source not found.**). Early emerging species contained 25 tRNA genes whereas some late emerging species of both the Arizonensis and the Continental subclades contained extra copies. A single gene locus was present for the majority of mitochondrial tRNAs, but extra copies of tRNAs coding for arginine, leucine, serine and methionine were found in all *Metschnikowia* species studied. Methionine tRNAs, in particular, were most prevalent with three copies per genome. The presence of high copy numbers of methionine tRNAs is in accordance with how important the availability of methionine is to the proper translation of a gene due to it serving as the start codon. Moreover, methionine appears to be the second most prevalent amino acid present in *Metschnikowia* mitochondrial gene products, next only to leucine (data not shown). Likewise, two tRNA loci coding for leucine were found, probably in response to similar high demands. In contrast, the duplication of the tRNA for serine was unexpected; although the amino acid itself is present at a high frequency, the differences in frequency are minimal in reference to other amino acids like glycine and alanine with a single locus copy. Even more unexpected was the case for the tRNA for arginine, whose demand in protein products is low. The occasional duplication observed for tRNA gene loci were for asparagine or lysine. Duplication of the asparagine tRNA gene was found mostly in some members of the Continentalis subclade whereas the extra lysine gene was found in the *bow-dek-lac-sim* subclade of the Arizonensis subclade. The restricted distribution of the asparagine or lysine duplication within the two subclades suggests that it is a rare event that occurred independently in an ancestral species of each subclade (**Figure 7**).

Table 3: Number of tRNA loci found in each of 71 *Metschnikowia* mitochondrial genomes using tRNA-scanSE. Candidate tRNA loci that were detected by tRNA-scanSE but unable to determine their types were excluded.

Code	Ala	Arg	Asn	Asp	Cys	Gln	Glu	Gly	His	Ile	Leu	Lys	Met	Phe	Pro	Ser	Sup	Thr	Tyr	Val	Total
abe+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
abe-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
aga+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
aga-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
ama+	1	2	1	1	1	3	1	1	1	1	2	1	3	1	1	2	1	1	1	1	27
ama-	1	2	1	1	1	2	1	1	1	1	2	2	3	1	1	2	1	1	1	1	27
ari+	1	2	2	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	26
ari-	1	2	2	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	26
bor+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
bor-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
bow+	1	2	1	1	1	1	1	1	1	1	2	2	3	1	1	2	1	1	1	1	26
bow-a	1	2	1	1	1	1	1	1	1	1	2	2	3	1	1	2	1	1	1	1	26
bow-b	1	2	1	1	1	1	1	1	1	1	2	2	3	1	1	2	1	1	1	1	26
cau+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
cau-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
cer+	1	2	2	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	26
cer-	1	2	2	1	1	1	1	1	1	1	2	2	3	1	1	2	1	1	1	1	27
cla+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
col+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
col-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
con+	1	2	2	1	1	1	1	1	1	1	2	2	3	1	1	2	1	1	1	1	27
con-	1	2	2	1	1	1	2	1	2	1	3	1	4	1	1	2	1	2	2	1	32
cos-	1	2	2	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	26
cub+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
cub-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
dek+	1	2	1	1	1	1	1	1	1	1	2	2	3	1	1	2	1	1	1	1	26
dek-	1	2	1	1	1	1	1	1	1	1	2	2	3	1	1	2	1	1	1	1	26
dekY	1	2	1	1	1	1	1	1	1	1	2	2	3	1	1	2	1	1	1	1	26
dra+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
dra-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
dro+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
dro-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
flo+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
ham+	1	3	1	2	2	2	2	2	2	2	4	2	6	2	2	4	2	2	2	2	47
ham-	1	3	1	2	2	2	2	1	2	2	4	2	6	2	2	4	2	2	2	2	46
han+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
haw+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
haw-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
hib+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
hib-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
ipo+	1	2	2	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	26
ipo-a	1	2	2	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	26
ipov	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
kam+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
kam-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25

kip-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
lac+	1	2	1	1	1	1	1	1	1	1	2	2	3	1	1	2	1	1	1	1	26
lac-	1	2	1	1	1	1	1	1	1	1	2	2	3	1	1	2	1	1	1	1	26
lacb	1	2	1	1	1	1	1	1	1	1	2	2	3	1	1	2	1	1	1	1	26
loc+	1	2	1	1	1	1	1	1	1	1	2	2	3	1	1	2	1	1	1	1	26
loc-	1	2	2	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	26
mar-	1	2	2	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	26
mat+	1	2	3	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	27
mat-	1	2	2	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	26
mau+	1	3	1	2	2	2	2	2	2	2	4	2	6	2	2	4	2	2	2	2	47
mau-	1	3	1	2	2	2	2	2	2	2	4	2	6	2	2	4	2	2	2	2	47
mer-	1	2	2	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	26
ori+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
ori-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
pal+	1	2	1	1	2	1	1	2	1	1	3	2	3	2	2	3	2	1	1	2	34
pil-	1	2	1	1	2	1	1	2	1	1	2	2	3	2	2	2	2	1	1	2	32
pro+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
pro-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
sce+	1	2	2	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	26
sce-	1	2	2	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	26
shi+	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
shi-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25
sim+	1	2	1	1	1	1	1	1	1	1	2	2	3	1	1	2	1	1	1	1	26
sim-	1	2	1	1	1	1	1	1	1	1	2	2	3	1	1	2	1	1	1	1	26
sma-	1	2	2	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	26
tor-	1	2	1	1	1	1	1	1	1	1	2	1	3	1	1	2	1	1	1	1	25

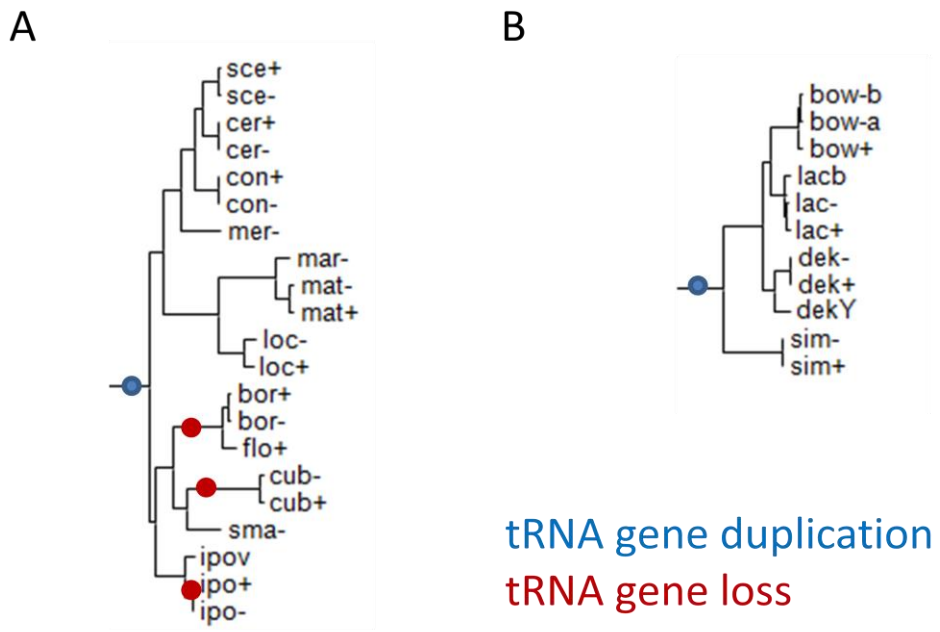

Figure 7: Hypothetical placement of tRNA gene duplications and losses in A) the asparagine-encoding tRNA in the Continental subclade and B) the lysine-encoding tRNA in the Arizonensis subclade. The subtrees are excerpted from Figure 3.

Regional duplications have led to an expansion in the number of tRNA genes in Hawaiian species, excluding *M. hawaiiensis* and *M. kamakouana*. The mitochondrial genomes of *M. hawaiiensis* contained duplicated regions that excluded the rRNA-tRNA cluster and those of *M. kamakouana* did not experience a regional duplication. As a result, tRNA numbers of the two species did not deviate from those of non-Hawaiian *Metschnikowia* species. *M. kamakouana* occupies an early-emerging position in the Hawaiian subclade, which suggests that the regional duplication event occurred soon after the emergence of the lineage that led to other Hawaiian endemics. This is in partial agreement with the proposition (Lachance et al. 2005) that speciation in this group took the form of successive peripatric events. Here, *M. kamakouana* is identified as the parent lineage, instead of *Metschnikowia* sp. 04-218.3 (‘*pal*’). The remaining Hawaiian species experienced expansion of the number of tRNA genes, up to 47 loci in the extreme cases

of *M. hamakuensis* and *M. mauinuiana*. Such a number vastly differs from the 25 to 27 tRNAs reported from other *Metschnikowia* species without regional duplications. Whether or not tRNA duplications have significant impacts on the fitness of the host organisms remains unclear. Strain UWOPS 95-402.1 of *M. continentalis* also contained an abnormal number of tRNA genes in its mitochondrial genome, suggesting that there could have been small regional duplications and/or recombinations, which could explain why the mitochondrial genome was linear, unlike the other sequenced strain of that species.

3.5 GC Content

Similar to those of parasitic organisms that require a host for survival, organelle genomes are typically high in adenine and thymine (Khachane et al. 2006). Enriched reactive oxygen species in mitochondria may also cause reductions in guanine-cytosine content due to the deanimation of 5-methylcytosine through oxidation, which generates thymine (Bjelland and Seeberg 2003). Mitochondrial genomes of *Metschnikowia* species were no exception as their GC content ranged from 20.2 to 29.4%, which is far lower than their nuclear counterpart, which mostly ranged from 40 to 45% (**Figure 2 and Table S2**). Although *M. caudata* had a much higher value (54%) for its nuclear genome, the GC content of its mitochondrial genome (~25%) was similar to those of other *Metschnikowia* species. Size or abundance of intergenic regions or intronic regions had no noticeable bearing on mitochondrial GC content (data not shown), thereby suggesting that neither is accountable for changes in GC content. The nuclear GC content of *M. caudata* is unusual not only for *Metschnikowia* species but also for Saccharomycetes in general. Of 332 ascomycetous yeast species whose genomes were examined by Shen et al. (2018), only two had GC contents exceeding 50%. Both were members of the CTG-Ser1 clade, namely *Candida wancherniae* and *Candida ascalaphidarum* with GC contents of 53%. The former is a sister species to *M. agaves*, a non-LSC haplontic species.

3.6 Mitochondrial Genome Size

The overall mitochondrial genome size for 71 *Metschnikowia* strains can be found in **Error! Reference source not found..** The mitochondrial genome size of 187 kb for *M. arizonensis* was the largest value found in the large-spored clade (and most likely the largest in all *Metschnikowia* species available on GenBank) and roughly sixfold higher than for *M. kipukae,* which had the smallest value of 33 kb in the large-spored clade (**Figure 2 and Table S2**). When other haplontic *Metschnikowia* species were considered, the difference increased to eightfold, as the size of the mitochondrial genome of *M. caudata* was 24 kb (**Figure 2 and Table S2**). As of June 2021, the *M. arizonensis* mitochondrial genome is the 14th largest in the Ascomycetes according to the GenBank database. The rank raises to 11th on a per species basis. Considering their identical gene contents and relatively close relatedness, such an enormous variation in size was unexpected and makes *Metschnikowia* species unique and interesting. By comparison, mitochondrial genome sizes among 33 species of *Kluyveromyces, Torulaspora, Lachancea* and other relatives of *S. cerevisiae* only varied up to fivefold, with extreme values of 20.1 kb for *Candida glabrata* and 107.1 kb for *Nakaseomyces bacilisporus* (Xiao et al. 2017). Duplicated genes found in some species could have explained genome expansion in the large-spored species but absence of gene duplications in very large (>120 kb) mitochondrial genomes suggested otherwise. Moreover, duplicated mitochondrial genes identified so far in a few members of Hawaiian subclade were small, such that their contribution to total genome size is not significant. In the Hawaiian clade, a regional duplication with multiple genes contributing to a significant portion of whole genome size was observed but these genomes were less than 100 kb in size. Therefore, the two elements most likely to affect diversity in mitochondrial genome size in *Metschnikowia* species are intergenic regions and introns.

It is currently unclear whether intergenic regions or introns contribute the most to genome size diversity in yeast mitochondria. Freel et al. (2015) compared 81 publicly available mitochondrial genomes across the Saccharomycetes and concluded that the intergenic regions were the major influencer of mitochondrial genome size variation. In contrast, Kanzi et al. (2016) analyzed mitochondrial genomes of *Chrysoporthe* species,

which belongs to a class of filamentous Ascomycota, the Sordariomycetes, and reported that introns were the key players. It should be noted that genome size distributions reported in the two studies were not similar. Of the 81 genomes used in Freel et al. (2015), 69 were less than 60 kb in length. On the other hand, the genomes in Kanzi et al. ranged in size from 89 kb to 190 kb. It is therefore possible that the observations of Freel et al. (2015) do not apply to the very large genomes studied by Kanzi et al. (2016) and vice versa. In comparison, *Metschnikowia* mitochondrial DNAs contain both small and very large genomes without a bias toward either extreme, thereby providing an appealing alternative to determine which of the two factors has the greater influence over mitochondrial genome sizes.

The largest mitochondrial genomes considered were those of *M. arizonensis*, and in that species the intergenic regions constituted more than half the mitochondrial genome (59.3 and 59.4%, **Figure 2 and Table S2**). In other *Metschnikowia* species, however, such large proportions of intergenic regions were more commonly found in those with the smaller genomes and not those with larger genomes. For example, the size of the *M. hawaiiana* mtDNA was 35.6 kb with 50.6% intergenic regions, whereas *M. amazonensis* mtDNAs were 132 kb long with 35.4 & 35.6% intergenic content. The possibility that the largest spacers contained unknown genes or mobile elements was eliminated when neither BLASTx nor BLASTn searches detected any reasonable hits within the intergenic regions of *M. arizonensis* (data not shown). In addition, potential ORFs predicted by Geneious did not match any orthologs in the GenBank database and most putative ORFs were too small (<400 bases) to form meaningful coding sequences that would contribute greatly to the overall genome size.

The conclusion reached by Freel et al. (2015) that the intergenic region better accounts for mitochondrial genome size diversity was drawn from determining correlations from r^2 values between overall genome size and intergenic or intronic regions. Applying same approach to *Metschnikowia* species, however, I found similar r^2 values for intergenic and intronic regions when compared to genome size, which suggests that neither factor is more influential than the other in *Metschnikowia* species (r^2 = 0.718 and 0.705, respectively, **Figure 8AB**). It should be noted, however, that *M. arizonensis* in both cases

was an outlier. When *M. arizonensis* was excluded from the analysis, different coefficients were obtained (r^2 = 0.654 and 0.889, respectively, **Figure 8CD**), suggesting that introns were the main determinant of genome size. Consistent with organelle phylogeny, genome size, and natural history, *M. arizonensis* stands out as the exception. It is currently unclear what makes *M. arizonensis* unique from the rest of the species. *M. arizonensis* has been reported to be unstable in terms of viability (Lachance and Bowles, 2002). It has been repeatedly isolated from only a single site in the Sonoran desert of Arizona in spite of extensive samplings of similar habitats in that region.

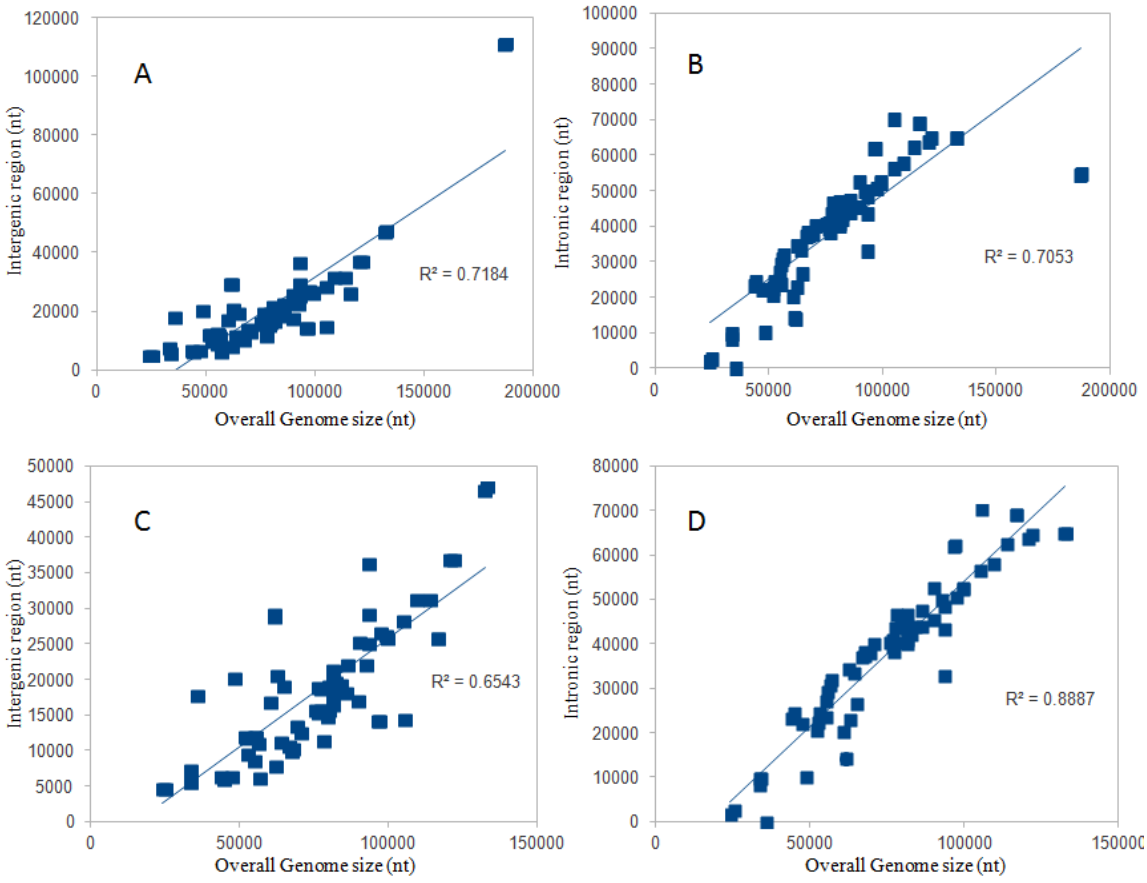

Figure 8: Joint distribution of intergenic (A, C) or intronic (B, D) mitochondrial contents and overall mitochondrial genome size, based on the approach of Freel et al. (2015), with (A,B) or without *M. arizonensis* (C, D).

Genome size was uncorrelated with species relatedness at the subclade level. Specifically, named subclades (**Figure 2, Figure 9 and Table S2**) cannot be identified with particular size patterns. Genome sizes were similar in some closely related species (e.g., *M. proteae* and *M. drakensbergensis* or the three Neo-Hawaiian species), but differences could also be substantial among members of the same subclade (e.g., *M. arizonensis* in the Arizonensis subclade and most of the Continental subclade). Furthermore, there was no visible pattern of changes in genome size between subclades as early emerging species of the South African subclade had larger genomes than most but not all late emerging species of the Continental subclade (**Figure 9**). Average genome size is therefore not an appropriate indicator to distinguish subclades.

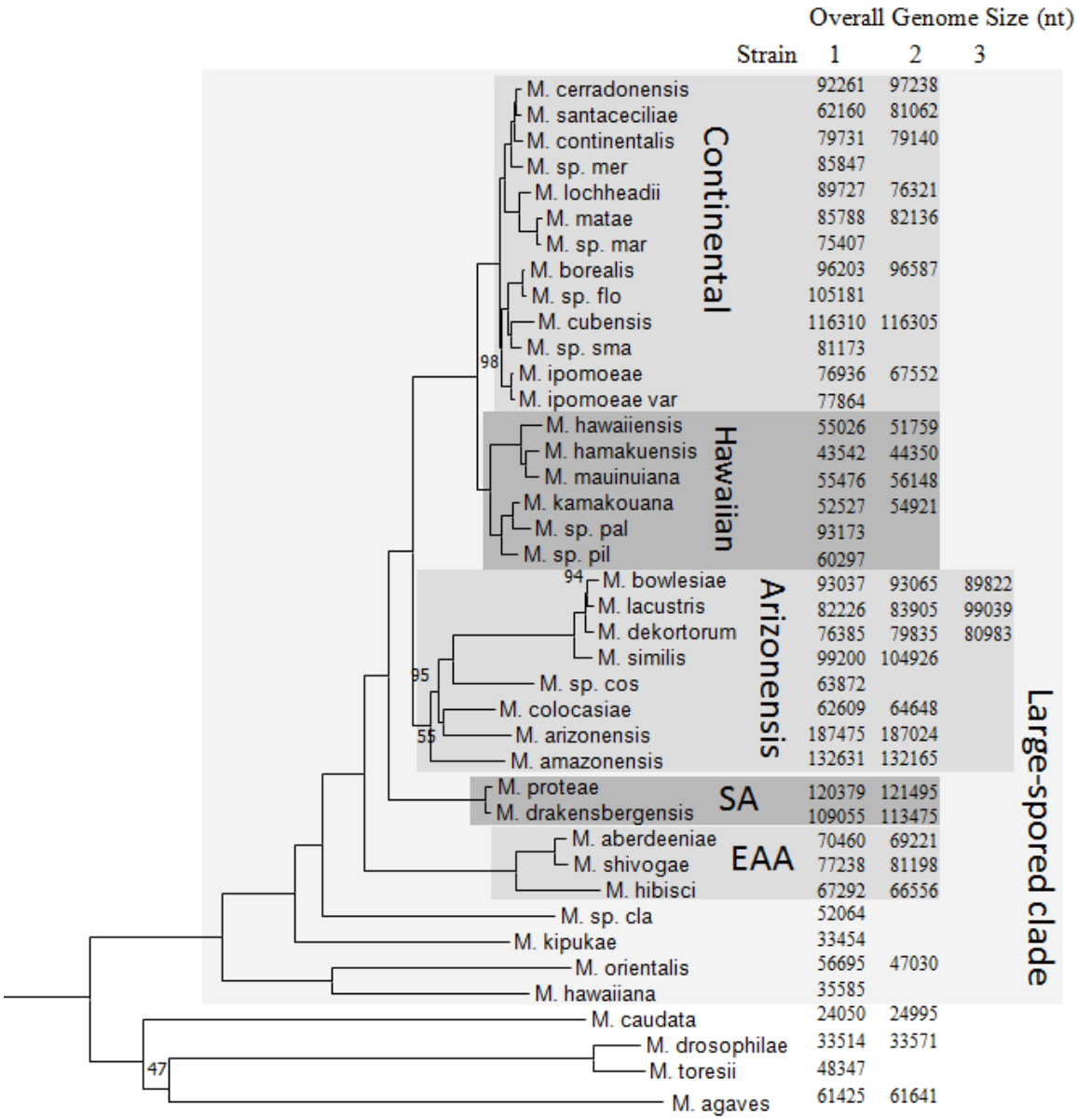

Figure 9: Overall mitochondrial genome size (nt) rearranged from Table S2 to match a robust phylogeny of haplontic *Metschnikowia* species (from Figure 3).

3.7 Mitochondrial Genome Morphology

Mitochondrial genomes harbor a variety of topologies. Circular, linear, multipartite, and concatenated circular genomes have been found across eukaryotes (Smith and Keeling 2015). Yeasts are no exception, although their variation in topology is confined mostly to either a circular or a linear configuration (Valach et al. 2011, Freel et al. 2015). Linear genomes are thought to arise from complications during genome replication (Nosek et al. 1998). The ends of linear genomes are usually enclosed by a telomere with or without additional structures such as hairpins or tandem repeats (Smith and Keeling 2015). The majority of *Metschnikowia* species sequenced have a circular genome (**Figure 2**) and given that all mitochondrial genomes of early emerging *Metschnikowia* species are circular, it is likely that the ancestral state for the large-spored clade was circular.

Six strains are suspected to have linear genomes. *M. similis* UWOPS 03-133.4, *M. continentalis* UWOPS 95-402.1, *M. colocasiae* UWOPS 03-202.1, *M. cerradonensis* UFMG 03-T68.1, UFMG 03-T67.1 and *M. bowlesiae* UWOPS 12-611.1 have telomeres with inverted repeats of varying lengths, ranging from 204 to 668 nucleotides, at the end of their assemblies (**Table 4**).

Table 4: Length of putative telomere sequences in *Metschnikowia* strains with linear mitochondrial genomes.

***Metschnikowia* species**	**Strain**	**Telomere length (nt)**
M. bowlesiae	UWOPS 12-611.1	232
M. cerradonensis	UFMG 03-T68.1	349
M. cerradonensis	UFMG 03-T67.1	204
M. colocasiae	UWOPS 03-202.1	351
M. continentalis	UWOPS 95-402.1	668
M. similis	UWOPS 03-133.4	251

PCR analysis and sequencing designed to extend non-telomeric regions failed to generate any products (**Figure 10A**). Sequencing read coverage analysis using the Bowtie2 plug-in of Geneious also showed that aligned reads abruptly disappeared at the ends of the assemblies instead of being continuous, as shown for circular genomes (**Figure 10B**). These observations do not irrefutably demonstrate the linearity of these genomes, as the distance between primers may be larger than anticipated due to some artefact of the

Illumina sequencing coverage. Barring such artefacts, it is reasonable to conclude that the six genomes are in fact linear. The appearance of linear genomes only in the Continental and Arizonensis subclades suggests that linearization events are recent and have occurred multiple times, independently, which is in agreement with previous findings that linearization occurs somewhat randomly (Nosek et al. 1998, Valach et al. 2011).

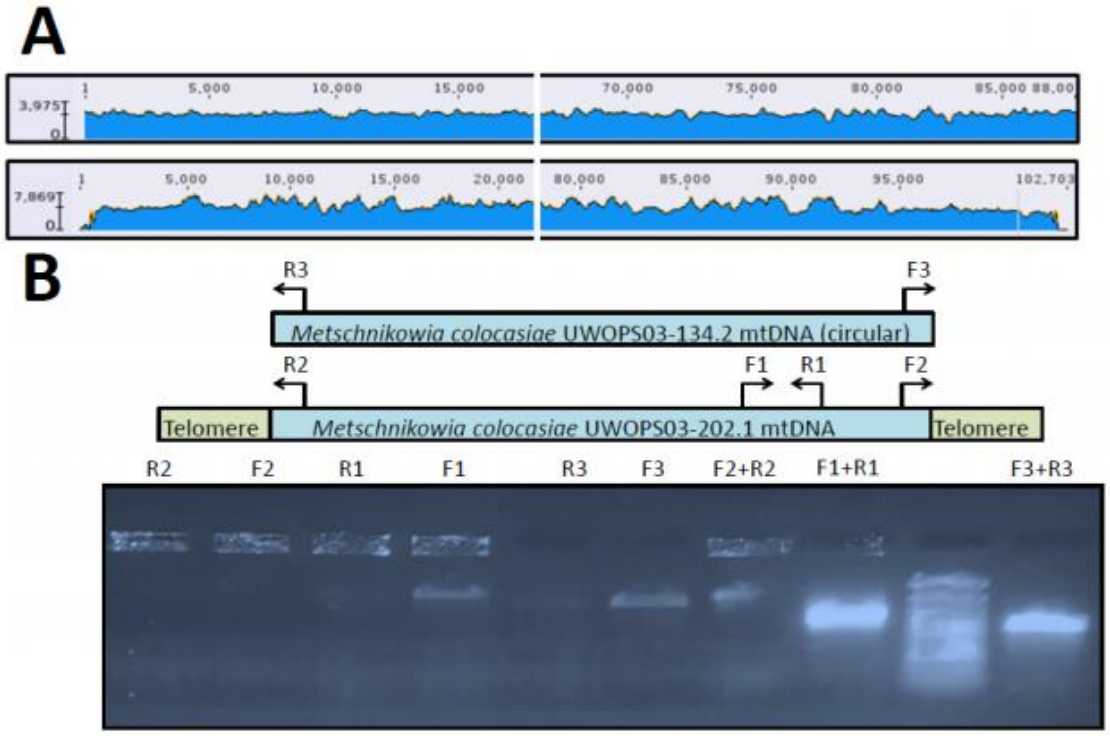

Figure 10: A) Sequence reads depth of coverage map of the circular mitochondrial genome of *Metschnikowia dekortorum* UFMG CM Y6306 (top) and the linear genome of *Metschnikowia cerradonensis* UFMG 03-T68.1 (bottom). B) Attempts to bridge the hypothetical connection of the two ends of the linear genome of two *Metschnikowia colocasiae* strains. In strain UWOPS03-134.2, a PCR reaction with primers F3 and R3 gave rise to a fragment with the expected length of 1.7k bp. In strain UWOPS03-202.1, primers F2 and R2 failed to yield a fragment of 3.2k bp, but the amplicon generated with the control primers F1 and R1 matched the predicted 1.5k bp result. The Thermo Gene Ruler 100bp plus (300-5000 bp, penultimate lane) was used as a reference.

3.8 Mitochondrial Gene Duplication

The majority of *Metschnikowia* species examined contained a single copy of each mitochondrial gene in their genomes. There were, however, ten genomes with duplicated genes and among those, some had a very large portion of their genome duplicated, resulting in the duplication of multiple genes (**Table 2**). For instance, both strains of *M. cubensis* contained an extra *nad3* gene. Similarly, strains of *M. hibisci* contained duplicated *nad2* and *nad3* genes. The extent of duplication was more pronounced in Hawaiian species. *Metschnikowia* sp. *pal* and *Metschnikowia* sp. *pil* had six genes (*atp6, atp8, atp9, cox3, nad2,* and *nad3*) duplicated whereas strains of *M. hamakuensis* and *M. mauinuiana* contained duplications in the same six genes as well as *nad4* and *rns,* for a total of eight duplicated genes. Errors in the assembly are possible, as the draft genomes of species in the Neo-Hawaiian subclades were determined earlier and are of lesser quality, making them more difficult to solve. That said, mitochondrial gene duplication is not uncommon and has been reported for other organisms such as birds, frogs, and lobsters (Kang et al. 2018, Kurabayashi and Sumida 2013, Gan et al. 2019). In Ascomycetes, duplication of a region containing six genes and five tRNAs has been found in *Candida sojae* (Valach et al. 2011), a member of the Debaryomycetaceae, another family in the CTG-Ser1 clade. Also, the detection of gene duplications in both strains of the same species, which were sequenced and assembled separately, increases one’s confidence in the results.

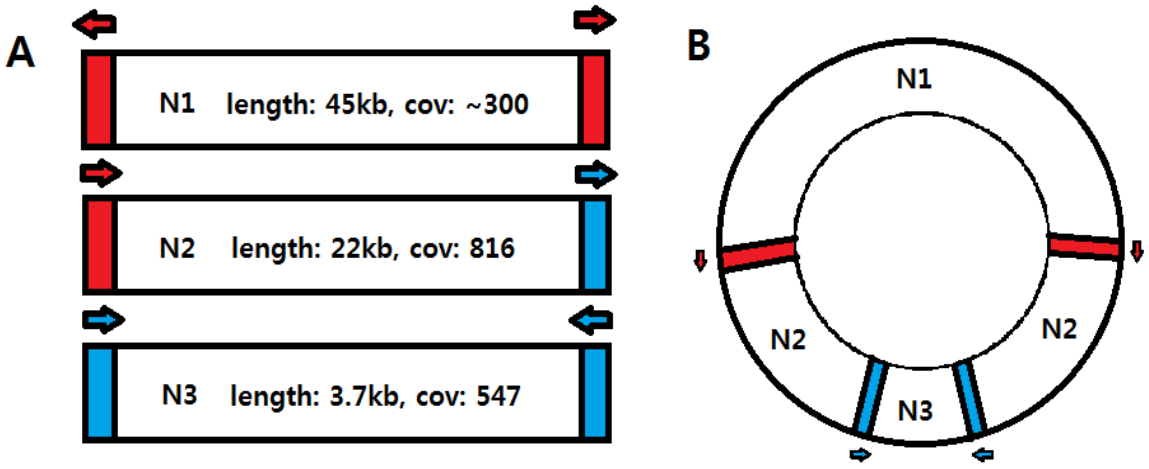

Figure 11: A) The final mitochondrial assembly for *Metschnikowia* sp. *pal* UWOPS04-218.3 yields three contigs that differ in mean depth of coverage. Areas with same color are identical and their directions are shown by arrows. B) A possible model of the whole genome connected by complementary sequences, assuming that the mitochondrial genome is not multipartite.

The process of detection of large duplications is illustrated for *Metschnikowia.* sp. *pal* UWOPS04-218.3 (**Figure 11**). In *Metschnikowia.* sp. *pal*, the mitochondrial assembly was narrowed down to three contigs. Each contig shared overlapping regions with the other two, suggesting that they were in fact connected. When the ends were connected based on complementarity of the overlapping regions, the presence of aligned overlaps confirmed the reconstruction (data not shown). The ends of the N1 contig (**Figure 11A**) were the reverse complements of each other. A similar pattern was observed in the N2 contig, although the overlapping sequences differed. The ends of contigs N1 and N3 were complementary to the ends of the N2 contig, favoring an N1-N2-N3 assembly with a linear structure as a preliminary result. The mean depth of coverage of each contig did not fit this interpretation perfectly. The coverage of contig N2 was more than double that of contig N1. Mitochondrial contigs often possess higher mean depths of coverage than their nuclear counterparts, most likely due to differences in the numbers of genomes between the nucleus (1 per cell) and the mitochondria (multiple genomes per

mitochondrion and multiple mitochondria per cell). The difference was pronounced in *Metschnikowia* species and was used extensively in this study to isolate contigs and reads associated with mitochondrial genomes. Hence, the most likely reason why the mean depth of coverage of contig N2 was approximately double that of contig N1, is that the N2 region is represented twice compared to the N1 region. Taking into account the presence of reverse-complementary regions at the two ends of contig N1, the most plausible configuration that takes into account double N2 regions would be rcN2-N1-N2-N3 (where rc indicates the reverse complement of the whole N2 contig; **Figure 11B**). Linking the two leftover ends of N2 with N3 will then close the circle and complete the circular genome structure. The relative coverage density might also suggest that the N3 region was also duplicated but the much higher coverage seen in N2 does not favor that interpretation. Alternatively, these genomes could be linear if the reverse complementary ends are interpreted as telomeres, as is typical in linear genomes (Smith and Keeling 2013). The ends of contigs N1 and N3 fit this characteristic, but those of N2 do not. Moreover, such an assumption would confer upon the mitochondrial genome of *Metschnikowia.* sp. *pal* a linear multipartite morphology. This is not unheard of in yeasts (Valach et al. 2011), but has not been found in *Metschnikowia* and related species. A multipartite structure would also be nearly impossible to demonstrate *in silico.* Another alternate morphology would be rcN3-rcN2-N1-N2-N3, which would satisfy the telomere requirement, but is not supported by the lower mean depth of coverage of contig N3 compared to contig N2. *Metschnikowia* sp. *pil* UWOPS04-226.1, a sister species of *Metschnikowia* sp. *pal*, exhibits a similar phenomenon, where the putative duplicated regions have higher mean depths of coverage compared to other mitochondrial contigs from the same strain (data not shown). Therefore, the ten genomes of the Hawaiian species are likely to be circular with duplicated regions and will be interpreted as such for the purpose of this thesis.

In the Hawaiian subclade, the duplicated regions of the six species are nearly syntenic. Accordingly, the duplication of six genes (*atp6, atp8*, *atp9, cox3, nad2*, and *nad3*) is likely to have occurred in the common ancestor of the Hawaiian subclade and the extra duplication involving *rns* and *nad4* must have occurred in the common ancestor of *M. hamakuensis* and *M. mauinuiana*, which is consistent with their inferred sisterhood.

Neither *M. kamakouana* nor *M. hawaiiensis* on the other hand, underwent the *rns/nad4* duplication or removed the duplicated regions post speciation. While *M. kamakouana* did not contain the *rns/nad4* duplicated regions, *M. hawaiiensis* did contain other duplicated regions (~4 kb each) that were devoid of genes. The most likely phylogenetic explanation would be that both *M. kamakouana* and *M. hawaiiensis* lost or fixed duplicated regions individually during their histories whereas other Hawaiian species retained their duplicated regions. This is far more probable than the hypothesis that Hawaiian species with duplicated regions acquired them by independent duplications for regions containing same the genes.

In other subclades, the duplication events in *M. cubensis* and *M. hibisci* must have occurred independently after their speciation, as evidenced by the lack of duplicated genes in their close relatives. It is interesting that in both species *nad3*, which is localized adjacent to *nad2* in almost all *Metschnikowia* strains, was included in the duplicated regions. One could postulate that the intergenic region near *nad3* could have been prone to recombination and/or rearrangements due to homologous regions or introns, but until more samples of similar duplications are obtained, the conclusion must be that they are coincidental.

3.9 Mitochondrial Gene Synteny

The concept of synteny began to be widely used in the literature as genome studies slowly gained popularity in the 1990s (Drillon and Fischer 2011). As whole genome sequencing gradually replaced traditional gene mapping techniques, the precise definition of synteny shifted from the simple physical linkage of markers on a same chromosome to include the conserved order of orthologs between species, also known as collinearity. Comparison of gene order between species provides valuable insights into genome evolution. For example, a loss of synteny between related species hints that genomes experienced one or more rearrangements after speciation. Also, the location and history of rearrangement events can be estimated from comparative studies. Fungal mitochondrial genomes have been reported previously to have a diverse mitochondrial

gene organization, suggesting the presence of frequent genome rearrangements (Aguileta et al. 2014). Evaluation of synteny between close relative has been conducted in multiple studies for *Saccharomyces* species but much less so in other yeasts, with the total number of species studied less than 15 (Ruan et al. 2017, Sulo et al. 2017).

The overall mtDNA gene orders of all 71 *Metschnikowia* strains were illustrated in **Figure 2**.

Given the variable genome sizes observed among *Metschnikowia* species, their mitochondrial gene topology was expected to be complex. It therefore came as a surprise that synteny was preserved across most of the early-emerging species of the large-spored clade, consisting of species from Africa, Australia, Hawaii, and Malaysia (**Figure 12A**).

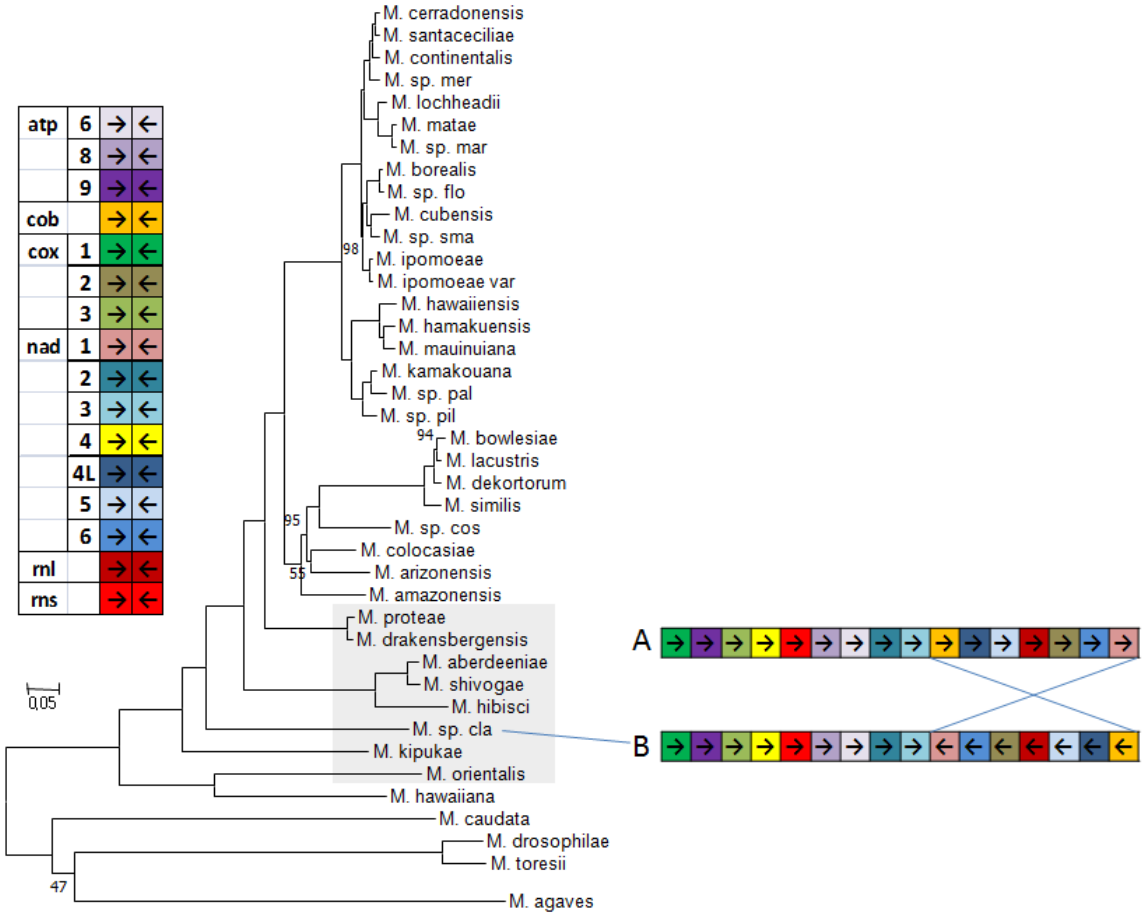

Figure 12: Mitochondrial gene synteny among early-emerging species of large-spored *Metschnikowia*. The tree was modified from Figure 3 so that only species studied in this thesis are shown. Coloured boxes refer to mitochondrial genes with arrows to indicate direction of their transcription. A list of genes and corresponding boxes can be found in the legend (left). The gene *cox1* was selected as a point of reference for comparative purposes. The species boxed in grey are syntenic (A) except for *Metschnikowia* sp. *cla* EBD CdV M2Y3 (B).

Considering the larger sizes of the mitochondrial genomes of *M. proteae* and *M. drakenbergensis* (over 100 kb), preservation of synteny among species from three different continents was unexpected. Furthermore, in the genomes of syntenic species, the genes were unidirectional. This suggests that the synteny observed in early-emerging

species is likely to be an ancestral condition in the large-spored *Metschnikowia* species. *Metschnikowia* sp. cla was an exception, where a single inversion was observed in the region between *cob* and *nad1* (**Figure 12B**).

Similarly, *M. hawaiiana*, a close relative of *M. orientalis*, was another exception. The mitochondrial genes of *M. hawaiiana* were unidirectional but the gene order was intermediate between that of the early-emerging large-spored species and that of the sister species *M. drosophilae* and *M. torresii* (**Figure 13**).

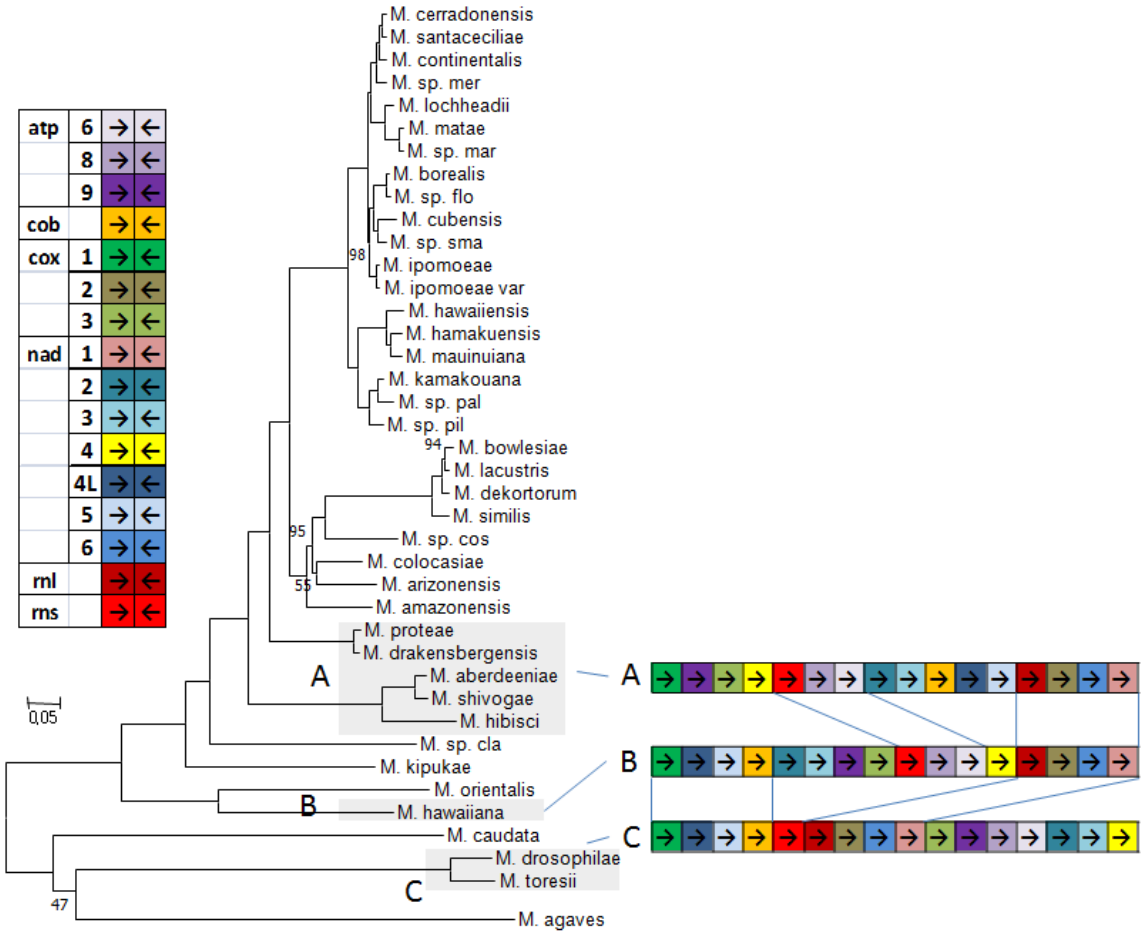

Figure 13: Differences in mitochondrial gene order in early-emerging large-spored species (A), *M. hawaiiana* (B), and *M. drosophilae* and *M. torresii* (C). The tree was modified from Figure 3 so only species studied in this thesis are shown. Coloured boxes refer to mitochondrial genes with arrows to indicate direction of their transcription. A list of genes and corresponding boxes can be found on the legend (left). The gene *cox1* was selected as a point of reference for comparative purposes. Species of interest are boxed in grey.

The *cox1-nad4L-nad5-cob* arrangement of *M. hawaiiana* was identical to that seen in *M. drosophilae and M. torresii* whereas the *rnl-cox2-nad6-nad1-cox1* concatenation characteristic of *M. hawaiiana* was identical to that of the early-emerging large-spored clade species. Two scenarios are hypothesized. One, *M. hawaiiana* is a remnant of an

intermediate state where mitochondrial gene order is on its way to becoming the ancestral state of the large-spored clade. Two, *M. hawaiiana,* after speciation, underwent rapid evolution to acquire the current gene order. Preservation of local synteny in two different lineages supports the former model, but the uniqueness of *M. hawaiiana* in having no introns, which suggests possibility of dramatic changes in their genomes in the past, supports the latter. Although the absence of introns in the *cob* gene is not unique among the species studied in this thesis, as both *M. caudata* and *M. colocasiae* UWOPS 03-134.2 also lack them, the complete absence of introns in any mitochondrial gene is unique to *M. hawaiiana.* It is currently not clear how *M. hawaiiana* would have remained the only haplontic *Metschnikowia* species studied with no introns in its mitochondrial genome, especially when other earlier-emerging species like *M. agaves* or *M. drosophilae* contain multiple introns. It remains likely that having no intron was the ancestral state of *Metschnikowia* species in general, which *M. hawaiiana* happened to retain while others did not. It is also possible that some unknown selection pressure favoured the elimination of introns in *M. hawaiiana* alone. An accurate picture of the state of synteny in *M. hawaiiana* will remain elusive until close relatives of *M. hawaiiana* are sequenced and analyzed. The recently described *Metschnikowia miensis* (Shibayama et al. 2019) would fit that bill.

Among non-LSC species, the mitochondrial genes of *M. agaves*, were not unidirectional. The region between *nad5* and *cox3* was reversed relative to the rest of the genome. *Candida wancherniae* NRRL Y-48709 is the closest relative of *M. agaves* (**Figure 14**). Its mitochondrial genome (WGS project PPOZ02) was therefore searched and found to be syntenic to that of *M. agaves* except that the region between *nad5* and *cox3* was reversed, thereby making the whole genome unidirectional (**Figure 14)**. An inversion that included the region compassing *nad5* and cox3 has clearly taken place after divergence of the two species. Further insight on rearrangements in the mitochondrial genome of *M. caudata* must await the characterization of genomes of related species such as *Candida hainanensis* and *Metschnikowia lopburiensis.*

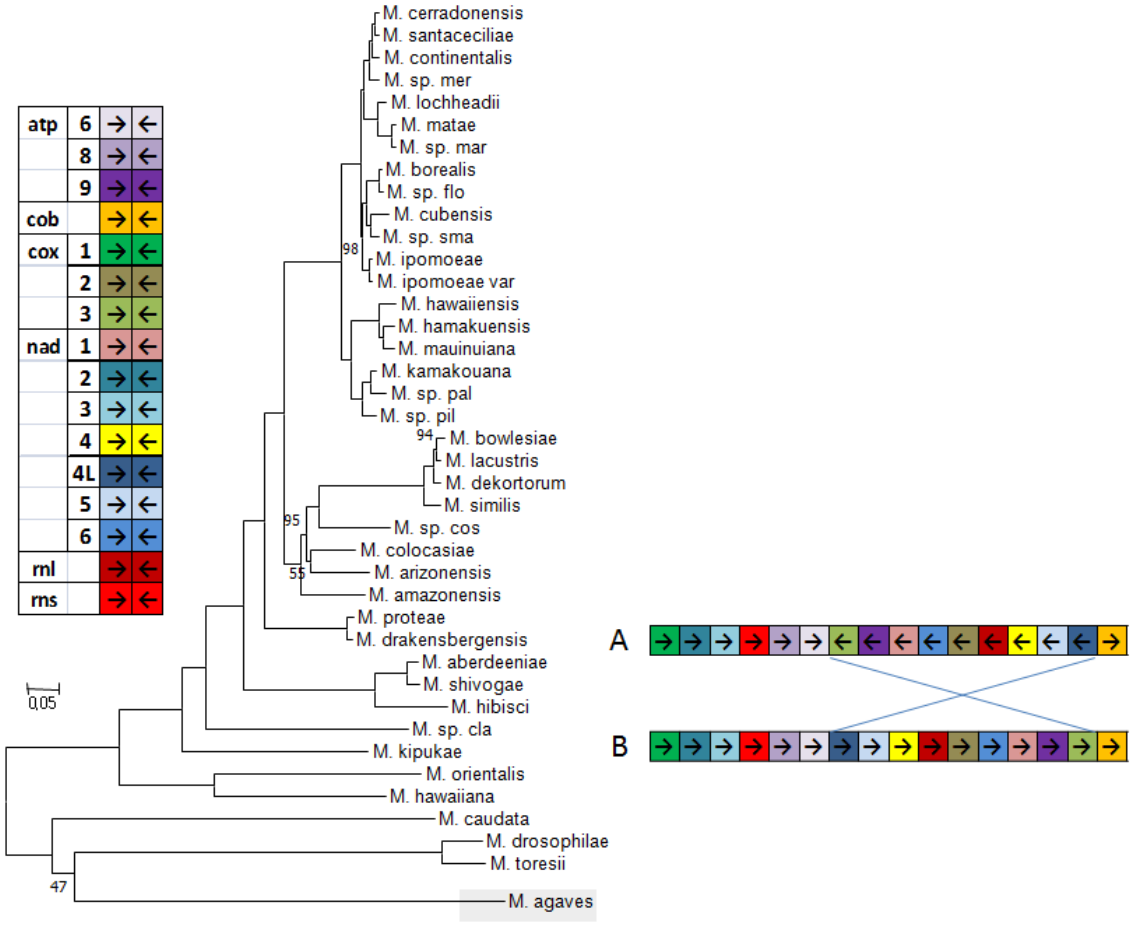

Figure 14: Differences in mitochondrial gene order in *M. agaves* (A) and *Candida wancherniae* NRRL Y-48709 (B, not included in the tree). The tree was modified from Figure 3 so only species studied in this thesis are shown. Coloured boxes refer to mitochondrial genes with arrows to indicate direction of their transcription. A list of genes and corresponding boxes can be found on the legend (left). The gene *cox1* was selected as a point of reference for comparative purposes.

In contrast to early emerging species, New World *Metschnikowia* species (sensu-stricto and Arizonensis subclades) have undergone such extensive rearrangements that the reconstruction of events is feasible only among species within each small subclade. Another noticeable change was that mitochondrial genes in the New World species are no longer unidirectional.

In the Hawaiian subclade, *M. hawaiiensis* features two inversions with respect to *M. mauinuiana* UWOPS 04-190.1, in the regions between *cob* and *cox1* (**Figure 15AB**).

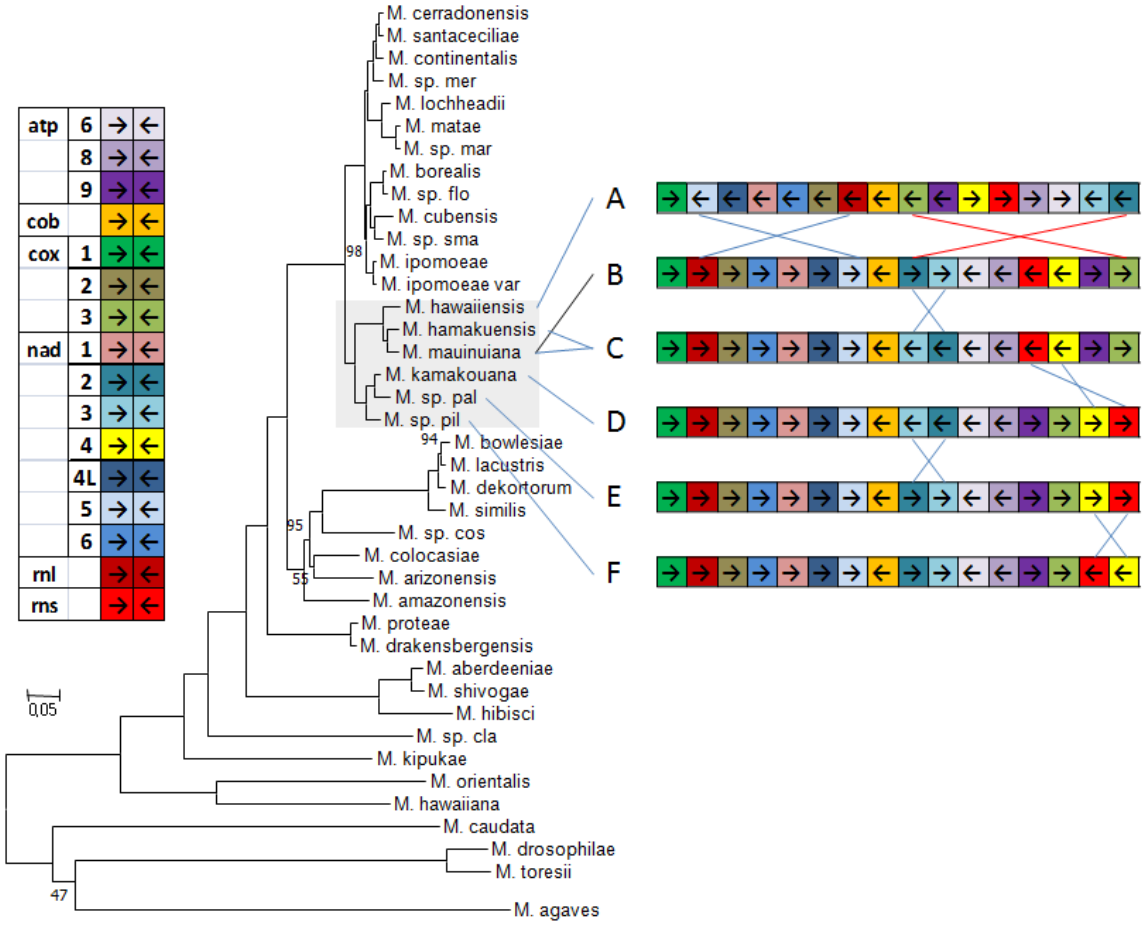

Figure 15: Mitochondrial gene order in the Hawaiian subclade. The tree was modified from Figure 3 so only species studied in this thesis are shown. Coloured boxes refer to mitochondrial genes with arrows to indicate direction of their transcription. A list of genes and corresponding boxes can be found on the legend (left). The gene *cox1* was selected as a point of reference for comparative purposes. The species of interest are boxed in grey. No two species of the six are fully syntenic. The gene order is shown for *M. hawaiiensis* (A), *M. mauinuiana* UWOPS 04-190.1 (B), *M. mauinuiana* UWOPS 04-110.4 and *M. hamakuensis* (C), *M. kamakouana* (D) *Metschnikowia* sp. *pal* UWOPS04-218.3 (E), and *Metschnikowia* sp. *pil* UWOPS04-226.1 (F).

Interestingly, the two strains of *M. mauinuiana* show a minor inversion between them in the *nad2/3* region. Otherwise, the mitochondrial genome of *M. mauinuiana* UWOPS04-110.4 is syntenic to that of *M. hamakuensis* (**Figure 15BC**). Furthermore, one strain of *M. mauinuiana* and both strains of *M. hamakuensis* differ from *M. kamakouana* in a single translocation and inversion of the *nad4/rns* region (**Figure 15CD**). Gene order variations in the “Paleo-Hawaiian” species (*M. kamakouana* and the undescribed species '*pal*' and '*pil*') were inversions of either the *nad2/3* or the *nad4/rns* regions (**Figure 15DEF**).

Inferring rearrangement events becomes comparatively more difficult in both the Continental and Arizonensis subclades due to their increasing diversity, both between species and among strains of the same species. Species delineation and phylogenetic placement in the Arizonensis subclade has been particularly difficult, leading to frequent readjustments from the description of the first species (Lachance & Bowles 2002) to the most recent (Lee et al. 2020). This is reflected in the widespread but fine-grained loss of synteny observed among the closely related species of the *bow-dek-lac-sim* group. The Hawaiian strain of *M. bowlesiae* had an inversion compared to the other two strains, which were isolated in Belize (**Figure 16AB**).

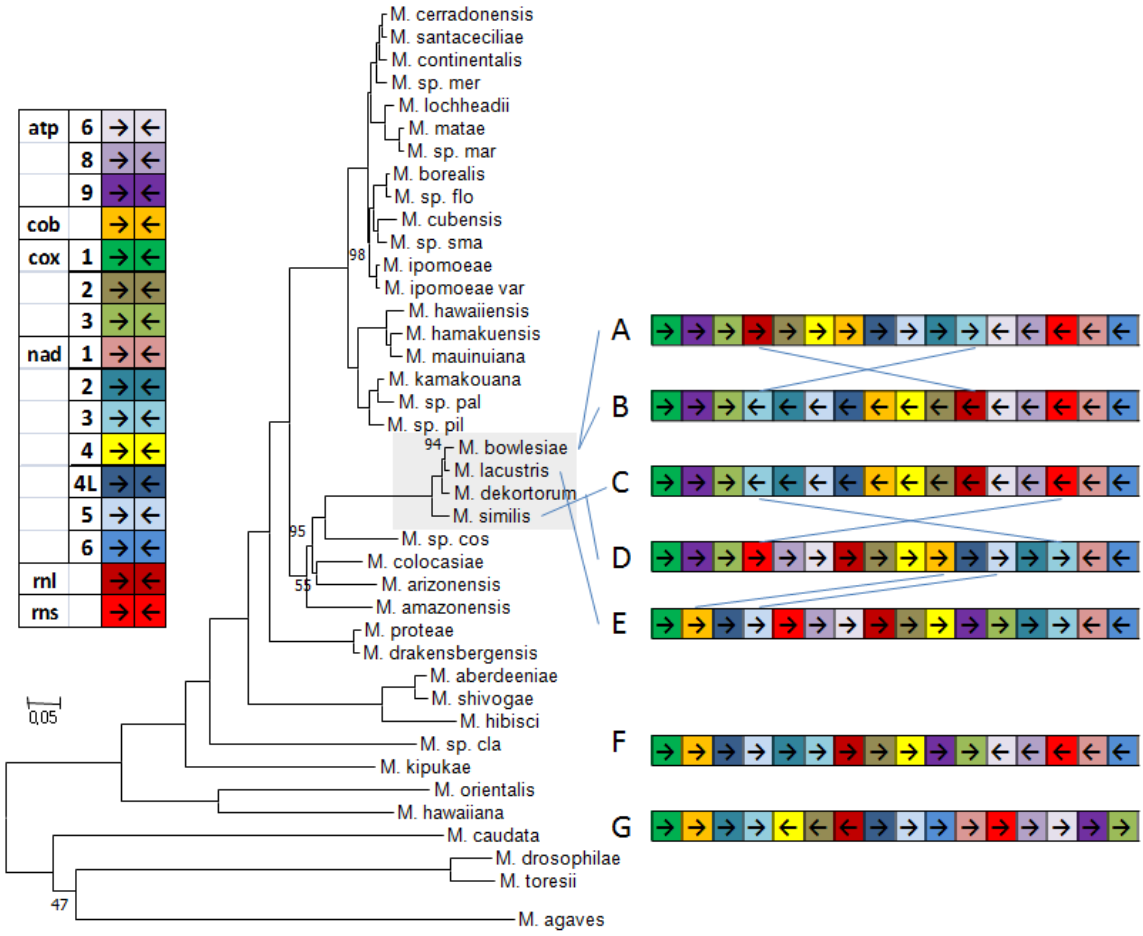

Figure 16: A comparison of mitochondrial gene order among species of the '*bow-dek-lac-sim* subclade. *M. bowlesiae* UWOPS 04-243x5 (Hawaii, A) two other *M. bowlesiae* strains (Belize, B), *M. similis* UWOPS 03-158.2 (C), two strains of *M. dekortorum* (Costa Rica, D), two strains of *M. lacustris* (E), *M. lacustris* strain UWOPS 03-167b3 (F), *M. dekortorum* UFMG-CM-Y6306 (Amazon basin). The tree was modified from Figure 3 so only species studied in this thesis are shown. Coloured boxes refer to mitochondrial genes with arrows to indicate direction of their transcription. A list of genes and corresponding boxes can be found on the legend (left). The gene *cox1* was selected as a point of reference for comparative purposes. Species of interest are boxed in grey.

The two Belizean strains were syntenic with *M. similis* whereas *M. dekortorum* had an inversion of a large region when compared to the previous four strains (**Figure 16BCD**). *M. lacustris* had a translocation of the *cob/nad4L/nad5* region when compared with *M. dekortorum* (**Figure 16DE**). Conversely, *M. lacustris* strain UWOPS03-167b3 (Arenal) and *M. dekortorum* strain UFMG-CM-Y6306 (Amazon basin) had unique gene orders, probably attributable to their distinct isolation localities, which may have contributed to making their initial species assignments problematic (Lee et al. 2020, **Figure 16FG**). There was no intraspecies variation in the rest of the species in the Arizonensis clade (*M. arizonensis, M. colocasiae, M. amazonensis* and the undescribed species "*cos*") but the gene order diverges extensively between species, so much as to preclude hypothetical reconstructions of the events (**Figure 17**). Tracing the history of loss of synteny in these unique species will become more feasible when more closely related species are discovered and sequenced in the future.

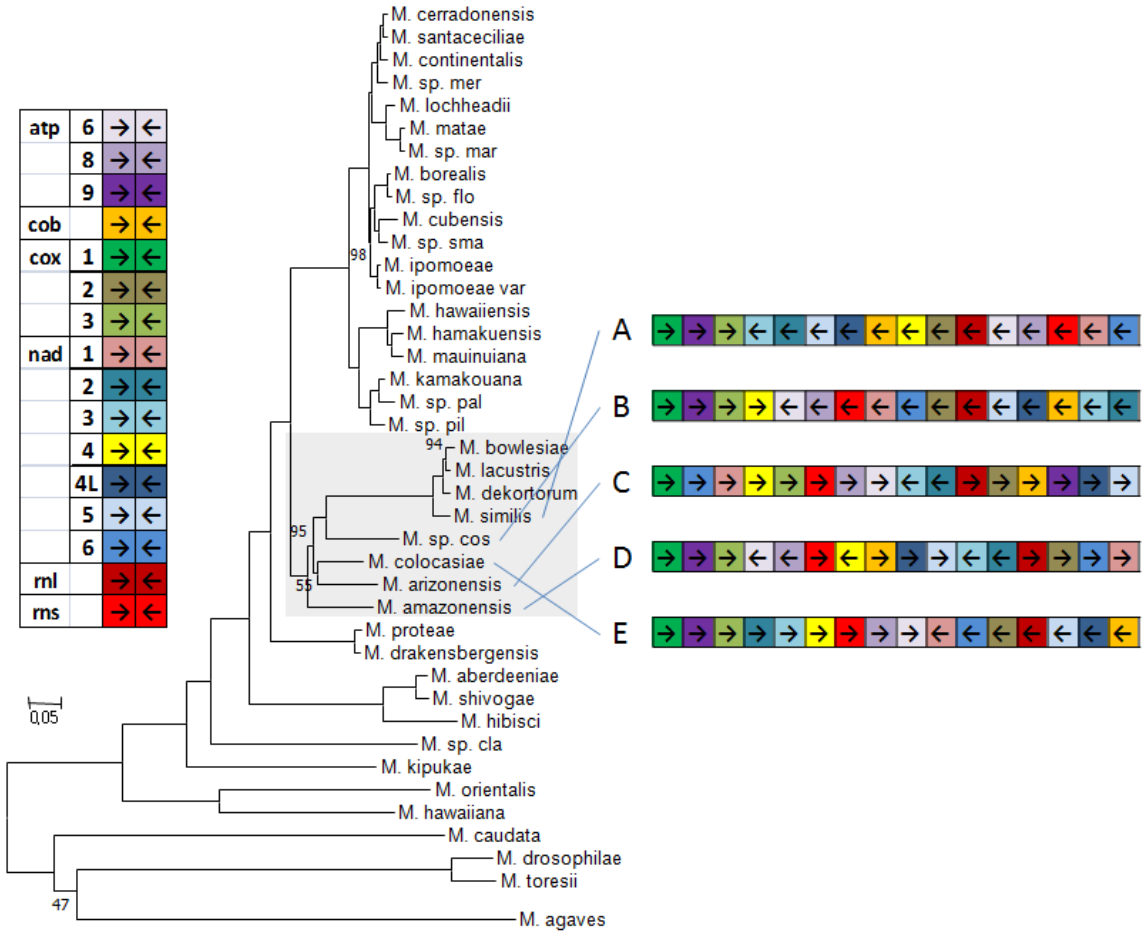

Figure 17: A comparison of mitochondrial gene order in *M. bowlesiae* UWOPS 12-611.1 from Belize and *M. similis* (A), *Metschnikowia* sp. *cos* UWOPS03-147.1 (B), *M. arizonensis* (C), *M. amazonensis* (D), and *M. colocasiae* (E). The tree was modified from Figure 3 so only species studied in this thesis are shown. Coloured boxes refer to mitochondrial genes with arrows to indicate direction of their loci. A list of genes and corresponding boxes can be found on the legend (left). The gene *cox1* was selected as a point of reference for comparative purposes. Species of interest are boxed in grey.

In the Continental subclade, synteny of *rnl-cox2-nad6-nad1* (and sometimes *cob*) was retained in all members but the rest of the genomes have undergone multiple rearrangements. Some species, for example *M. lochheadii* and *M. santaceciliae*, even

showed loss of synteny among conspecific strains, suggesting that the mitochondrial genomes are evolving faster than the nuclear genomes (**Figure 18)**.

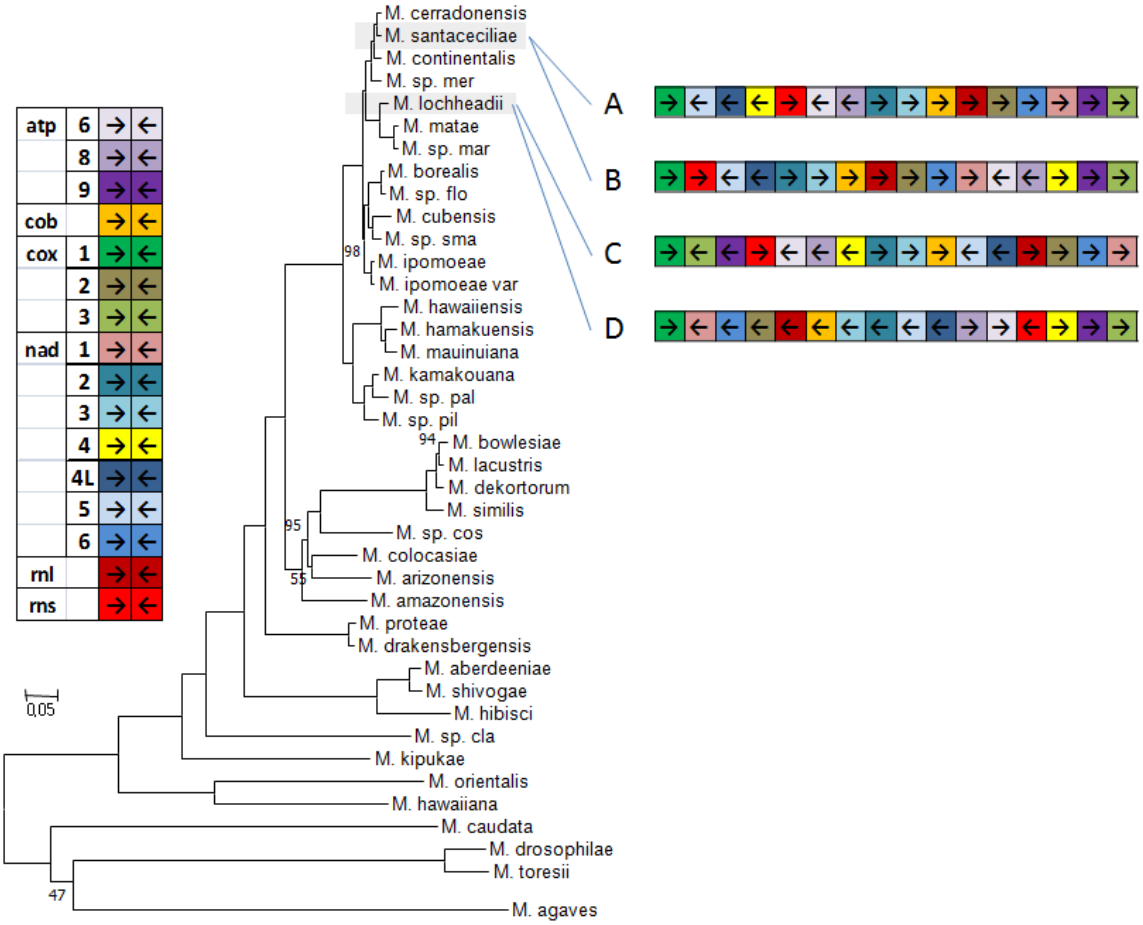

Figure 18: Loss of mitochondrial gene synteny in *M. santaceciliae* UWOPS 01-517a1 (A), *M. santaceciliae* UWOPS 01-142b1 (B), *M. lochheadii* UWOPS 03-167a3 (C), and *M. lochheadii* UWOPS 99-661.1 (D). The tree was modified from Figure 3 so only species studied in this thesis are shown. Coloured boxes refer to mitochondrial genes with arrows to indicate direction of their loci. A list of genes and corresponding boxes can be found on the legend (left). The gene *cox1* was selected as a point of reference for comparative purposes. Species of interest are boxed in grey.

Strain UWOPS03-167a3 was collected from an isolated locality and found to be sufficiently distant, genetically, from other strains of *M. lochheadii* (Lachance et al. 2008, 2020) to raise doubts on its conspecificity. Geographical isolation and reduced fertility of strain UWOPS03-167a3 when mated with other strains of *M. lochheadii* strains are consistent with the incipient speciation suggested by the loss of synteny in the mitochondrial genome. Similarly, all three strains of *M. ipomoeae* possessed different gene orders. The genome of one of the strains was syntenic with that of the undescribed species "*sma"* (**Figure 19**).

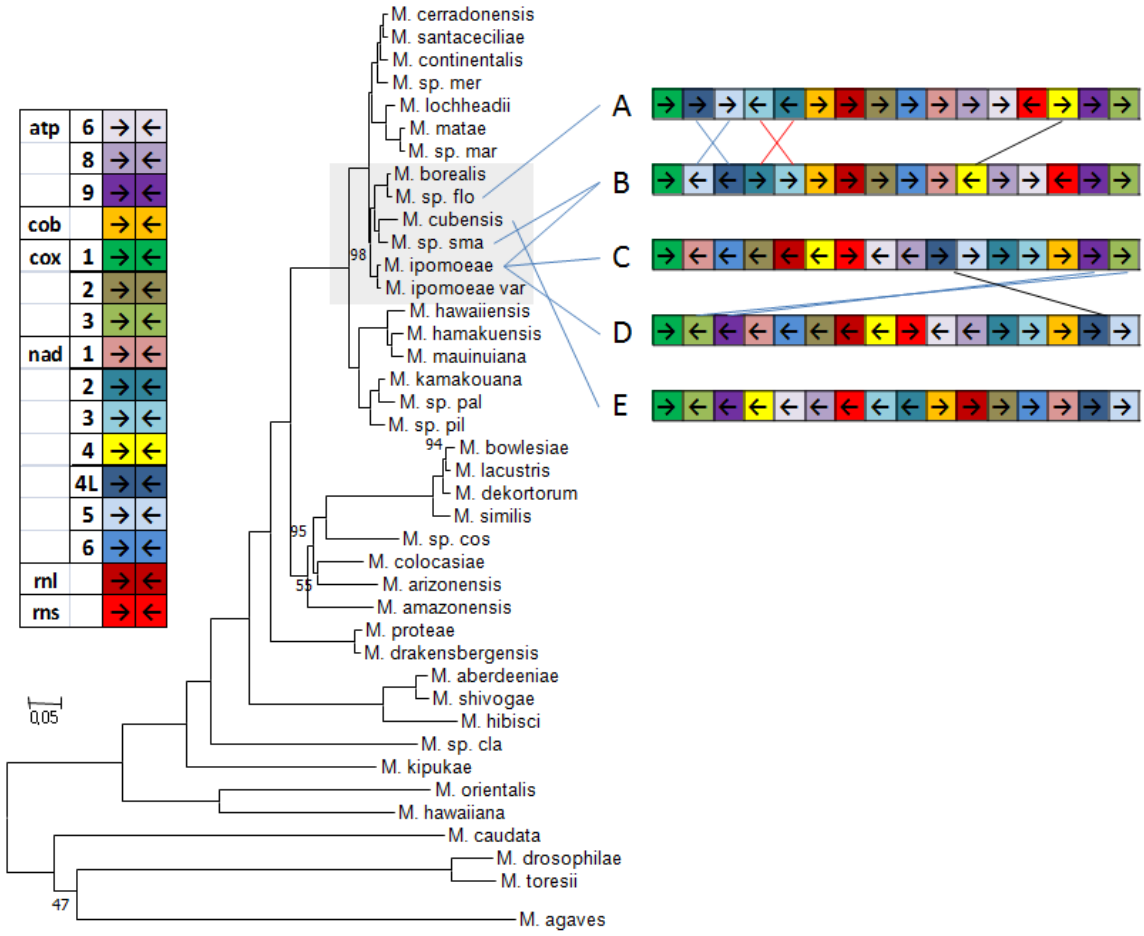

Figure 19: Conservation of mitochondrial gene synteny in *M. borealis* and *Metschnikowia* sp. *flo* UWOPS 13-106.1 (A), and *Metschnikowia* sp. *sma* UWOPS 01-665c1 and *M. ipomoeae* UWOPS 10-104.1 (B). Loss of synteny in *M. ipomoeae* UWOPS 99-324.1 (C), *M. ipomoeae* UWOPS 01-141c3 (D), and *M. cubensis* (E). The tree was modified from Figure 3 so only species studied in this thesis are shown. Coloured boxes refer to mitochondrial genes with arrows to indicate direction of their transcription. A list of genes and corresponding boxes can be found on the legend (left). The gene *cox1* was selected as a point of reference for comparative purposes. Species of interest are boxed in grey.

M. cubensis featured a unique gene order that differed from those of its closest relatives (**Figure 19**). Only after eliminating these unique strains, was it possible to reconstruct a hypothetical tracing of gene order change in species of the continental subclade that minimizes the number of inversions and translocations (**Figure 20**).

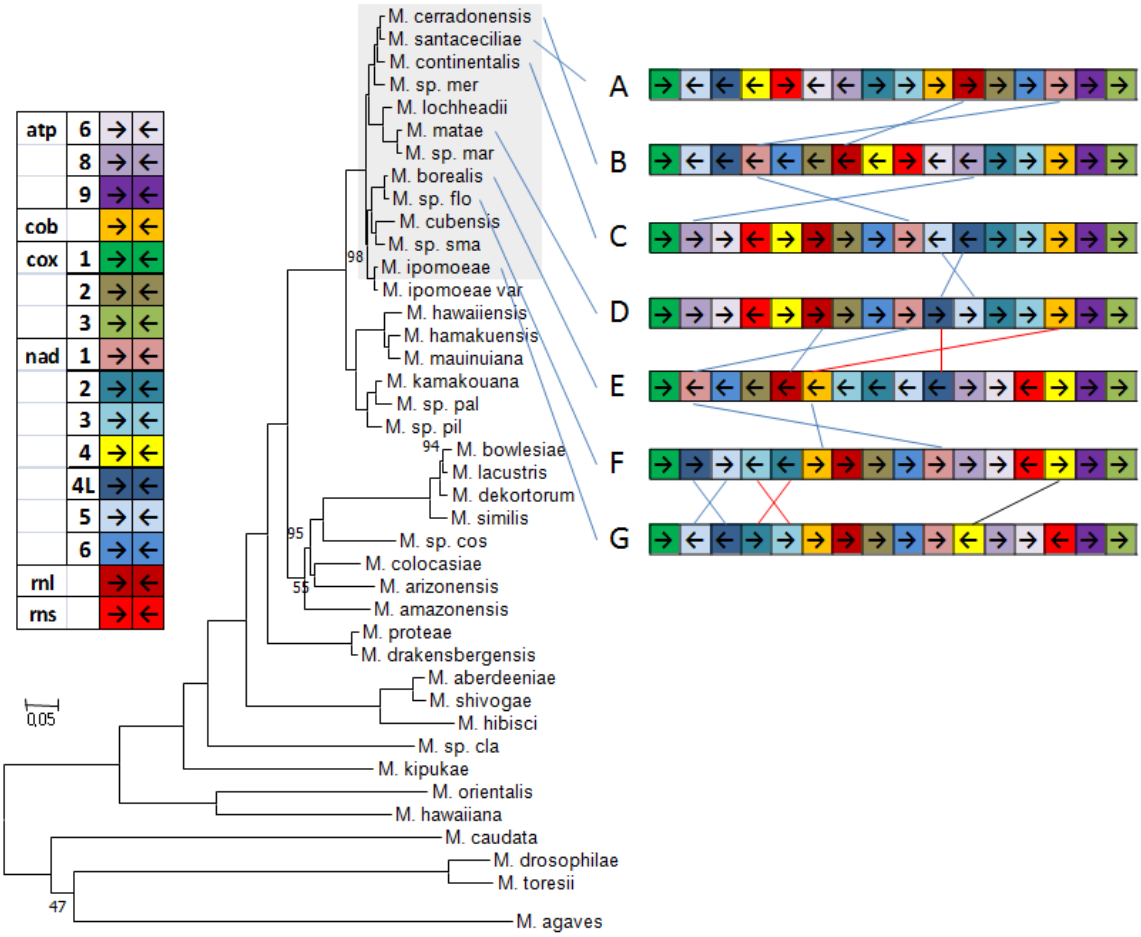

Figure 20: Hypothetical reconstruction of loss of synteny in mitochondrial genes synteny of *M. santaceciliae* UWOPS 01-517a1 (A), *M. cerradonensis* UFMG 03-T67.1 (B), *M. continentalis* UWOPS 95-402.1 (C) *M. matae* (D)*M. lochheadii* 99-661.1 (E), *M. borealis* (both strains) and *Metschnikowia* sp. *flo* UWOPS 13-106.1 (F), and *Metschnikowia* sp. *sma* UWOPS 01-665c1 and *M. ipomoeae* UWOPS 10-104.1 (G). The tree was modified from Figure 3 so only species studied in this thesis are shown. Coloured boxes refer to mitochondrial genes with arrows to indicate direction of their transcription. A list of

genes and corresponding boxes can be found on the legend (left). The gene *cox1* was selected as a point of reference for comparative purposes. Species of interest are boxed in grey.

It is not clear why the New World *Metschnikowia* species show such extensive variation in gene order compared to the early emerging species, even though the New World species are phylogenetically more closely interrelated. One possibility is that the New World species evolved to undergo more frequent genome recombination events, due either to having lost the ability to conserve synteny or having gained the ability to tolerate the loss of synteny and its fitness effects. Another possibility is that the New World environment favoured evolution of rapid replication and/or maintenance of high copy number of mitochondria, which could have enhanced the rate of evolution of mitochondria within New World species. Exchange of mitochondria by introgression among a large number of closely related species is also not to be discounted.

3.10 Ancestral Genome Reconstruction of mitochondria of haplontic *Metschnikowia* species

Reconstructing the ancestral state of a trait is an important aspect of understanding evolution, as it provides a glimpse into the evolutionary history of species. One trait of particular interest is the order of homologous genes. Currently, two types of algorithms are available. Some build a genome that minimizes the number of events required under a given evolutionary model and others construct a genome using conserved regions (Feng et al. 2017, Perrin et al. 2015). Higher resource demands and error rates of the event/distance-based algorithms have caused preferences to shift towards homology/adjacency-based algorithms in recent times (Feng et al. 2017).

In the preceding section, I have used intuitions and deductions to infer ancestral gene order changes. Here, candidate ancestral mitochondrial gene orders of common ancestors were explored using ProCARs (Progressive Contiguous Ancestral Regions), a program

that is based on homology/adjacency approaches (Perrin et al. 2015). Reconstructed ancestral mitochondrial gene orders using ProCARs are listed in **Figure 21**. An example of evolution of mitochondrial gene orders of Hawaiian subclade, with assistance of ProCARs, is illustrated in **Figure 22**.

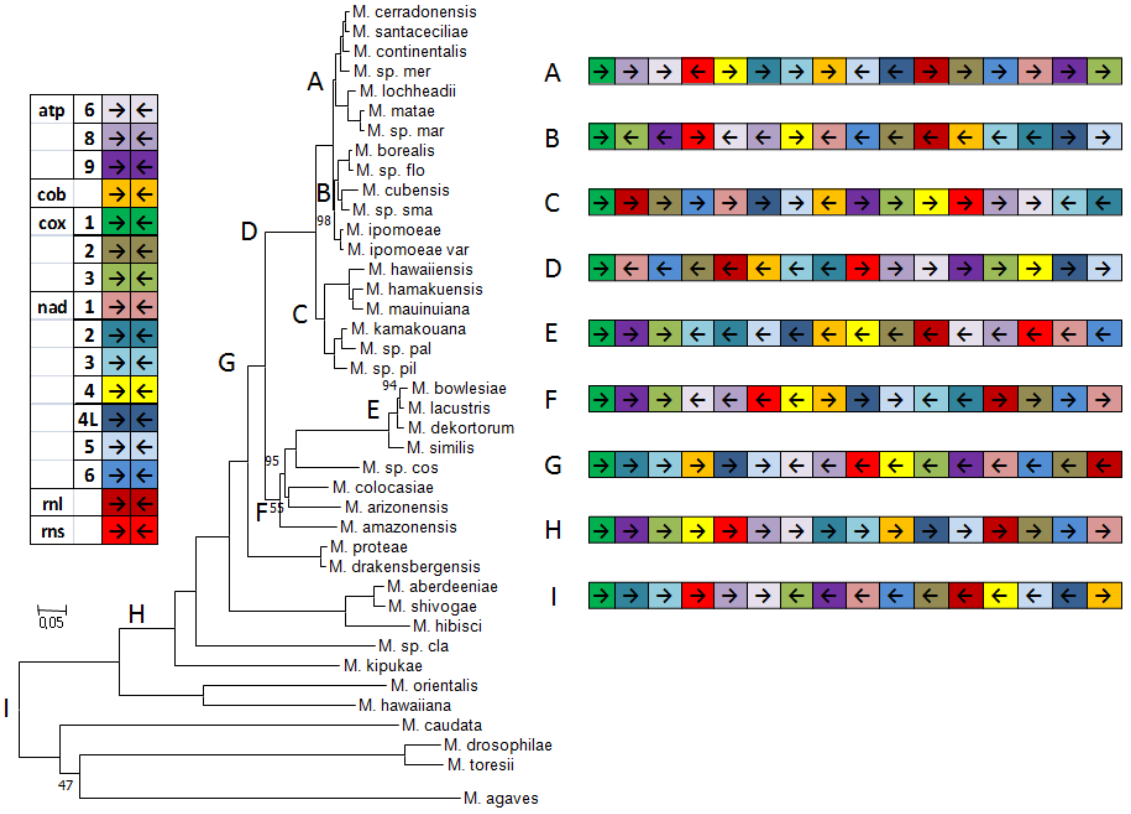

Figure 21: Ancestral mitochondrial gene order suggested by the program ProCARs (Perrin et al. 2015). From top to bottom, subclades containing *M. cerradonensis* to *M. matae* var. *maris* (A), *M. borealis* to *M. ipomoeae var* (B), the Hawaiian species (C), all *sensu stricto* strains (D), *M. bowlesiae* to *M. similis* (E), all Arizonensis species (F), both *sensu stricto* and Arizonensis species (G), all large-spored species excluding *M. hawaiiana* (H) and all haplontic *Metschnikowia* species studied (I) are shown. The tree was modified from Figure 3 so only species studied in this thesis are shown. Coloured boxes each refer to a mitochondrial gene, with arrows to indicate the direction of their transcription. A list of genes and the corresponding colours is

shown on the left. The gene *cox1* was selected as a point of reference for comparative purposes.

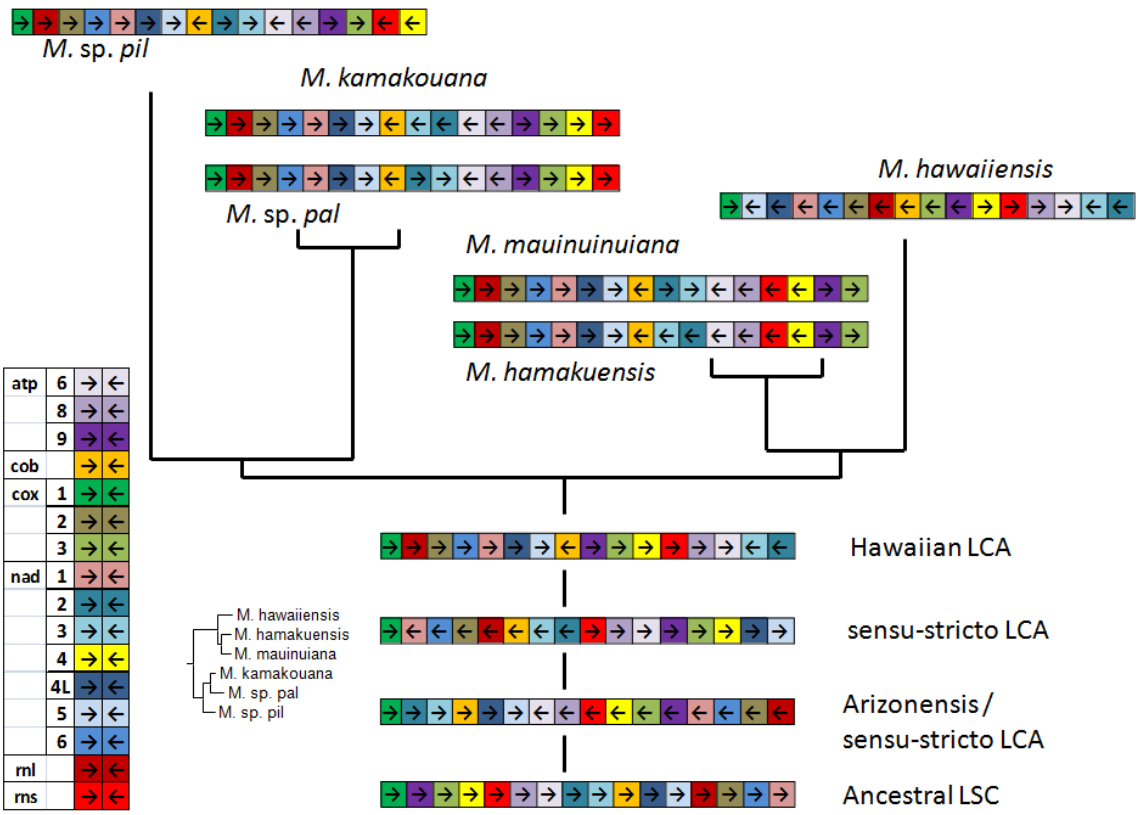

Figure 22: Evolutionary history of mitochondrial gene orders of Hawaiian subclade as an example of revised analysis using ProCARs. ProCARs provided gene orders of three unknown last common ancestors (LCA) between early-emerging LSC and Hawaiian species.

Given the conserved synteny among most of early-emerging species, it was not surprising that ProCARs also predicted an identical gene order for the reconstructed ancestral state of the large-spored clade up to *M. kipukae* (**Figure 21H**). The predicted ancestral gene order for the entire large-spored clade, including *M. orientalis* and *M. hawaiiana*, however, was inconclusive (data not shown). The predicted ancestral gene order for all strains analyzed as identical to that of *M. agaves* was unexpected; it may be due to that species being the most distant from all others (**Figure 21I**).

The reconstructed ancestral gene order suggests that an ancestor of all large-spored *Metschnikowia* species that shared synteny with other existing early-emerging species underwent at least three genome rearrangements including two inversions and one translocation (**Figure 21GH, Error! Reference source not found.**). ProCARs also suggested that the ancestors of the Arizonensis and the *sensu stricto* subclades experienced multiple rearrangements, at least four for *sensu stricto* and six for Arizonensis, with respect to the early-emerging species before they diversified further (**Error! Reference source not found.DFG**). The predicted ancestor of the *sensu stricto* subclade is thought to have experienced at least six rearrangements leading to the Hawaiian subclade(**Figure S74**).

It should be noted that ProCARs was inconclusive on an ancestral gene order for Continental subclade such that two nearby branching points within the subclade were used as a substitute (data not shown). Comparison of those substitute gene orders with that of the ancestor of the *sensu stricto* suggests five or more rearrangement events, which may or may not occurred before and after diversification of two major branches within the Continental subclade (**Figure S75**). A reconstructed ancestral gene order for the entire Continental subclade was inconclusive possibly due to the high diversity within the subclade.

Overall, ancestral gene order reconstruction by ProCARs provided an improved view of how the gene order of haplontic *Metschnikowia* species may have evolved. The analysis showed possible rearrangement events between subclades so that patterns and/or events were easier to grasp. It also showed that at least 4 or more rearrangement events have occurred between subclades thereby showing how rapidly mitochondrial genomes evolve. The analysis also showed limitations, as evidenced from two inconclusive predictions on the Continental clade and at the root of large-spored clade, most likely due to the high diversity present. Furthermore, there remains a chance that the proposed pathways may not reflect the true evolutionary history of gene orders, as evolution does not always proceed in an optimal manner without any detours or redundancies. As more related species get discovered and have their genomes sequenced, a clearer view of mitochondrial genome evolution may emerge.

3.11 Mitochondrial Introns

Self-splicing introns, which are commonly found in the genomes of bacteria and archaea, have been found in mitochondria of fungal species as well, and are thought to contribute to mitochondrial evolution (Repar and Warnecke 2017). Most introns found in fungal mitochondrial genes are either group I or group II introns, which differ by their splicing mechanisms (Lang et al. 2007). Some of these introns contain open reading frames that encode proteins such as homing endonucleases, reverse transcriptases, and maturases (Guha et al. 2018). There are six known families of endonucleases and most common ones are LAGLIDADG (Lys-Ala-Gly-Ile-Asp-Ala-Asp-Gly), HNH (His-Asn-His) and GIY (Gly-Ile-Tyr) families (Chevalier and Stoddard 2001).

During intron splicing, the homing endonuclease cleaves DNA sequences while the maturase facilitates the splicing process with its multiple domains (Zhao and Pyle 2017). The names maturase and reverse transcriptase will be used interchangeably in this thesis because most maturases contain a reverse transcriptase domain (Sultan et al. 2016). Assembled *Metschnikowia* mitochondrial genomes showed that introns and intron-encoded proteins are prevalent in *Metschnikowia* species as well. Intron insertions, as shown earlier, contribute greatly to size variation of genomes and genes. For example, the lengths of *cox1* protein sequences are more or less the same among *Metschnikowia* species, averaging around 1.6 kb. The length of the gene loci, however, varied up to 30-fold, as seen in *M. hawaiiana* and *M. cubensis* (**Table 5**).

Table 5: Total length of *cox1* and *cob* gene loci in 71 *Metschnikowia* strains, including both exons and introns.

***Metschnikowia* strains**	**Code**	***cox1*** (nt)	***cob*** (nt)
M. aberdeeniae	abe+	18467	17280
M. aberdeeniae	abe-	13914	19745
M. agaves	aga+	12347	4391
M. agaves	aga-	12233	4392
M. amazonensis	ama+	36051	22417
M. amazonensis	ama-	36041	22472
M. arizonensis	ari+	35735	14983
M. arizonensis	ari-	35511	14855
M. borealis	bor+	44379	12012
M. borealis	bor-	44365	12019
M. bowlesiae	bow+	25346	16658
M. bowlesiae	bow-a	22161	16177
M. bowlesiae	bow-b	22277	16672
M. caudata	cau+	3400	1152
M. caudata	cau-	4219	1152
M. cerradonensis	cer+	32554	14631
M. cerradonensis	cer-	32189	14631
Metschnikowia sp. M2Y3	cla+	16476	7244
M. colocasiae	col+	17186	1152
M. colocasiae	col-	15071	6856
M. continentalis	con+	28496	12002
M. continentalis	con-	25664	15063
Metschnikowia sp. 03-147.1	cos-	19196	10952
M. cubensis	cub+	49069	19130
M. cubensis	cub-	49088	19104
M. dekortorum	dek+	18820	16065
M. dekortorum	dek-	18875	16425
M. dekortorum	dekY	18420	16359
M. drakensbergensis	dra+	36010	18674
M. drakensbergensis	dra-	35993	23127
M. drosophilae	dro+	5891	4796
M. drosophilae	dro-	5887	4794
Metschnikowia sp. 13-106.1	flo+	46699	20081
M. hamakuensis	ham+	9206	3949
M. hamakuensis	ham-	9901	3938
M. hawaiiana	han+	1599	1155
M. hawaiiensis	haw+	16314	9391

M. hawaiiensis	haw-	13086	9438
M. hibisci	hib+	17436	16257
M. hibisci	hib-	17593	14596
M. ipomoeae	ipo+	24884	10089
M. ipomoeae	ipo-a	22003	12033
M. ipomoeae	ipov	29092	12857
M. kamakouana	kam+	15791	9915
M. kamakouana	kam-	15810	12378
M. kipukae	kip-	5778	2683
M. lacustris	lac+	20398	16726
M. lacustris	lac-	22867	16392
M. lacustris	lacb	25621	19739
M. lochheadii	loc+	32424	15149
M. lochheadii	loc-	24936	12482
M. matae var. *maris*	mar-	24878	12187
M. matae var. *matae*	mat+	30362	12085
M. matae var. *matae*	mat-	27132	12708
M. mauinuiana	mau+	9623	7761
M. mauinuiana	mau-	14270	5980
Metschnikowia sp. 00-154.1	mer-	26539	13980
Metschnikowia orientalis	ori+	17625	13578
Metschnikowia orientalis	ori-	11605	8648
Metschnikowia sp. 04-218.3	pal+	22003	11843
Metschnikowia sp. 04-226.1	pil-	13497	6428
M. proteae	pro+	37315	23095
M. proteae	pro-	38462	23054
*M. santacecilia*e	sce+	19290	10968
*M. santacecilia*e	sce-	23478	14265
M. shivogae	shi+	15327	21144
M. shivogae	shi-	23231	18905
M. similis	sim+	28905	16881
M. similis	sim-	30989	20266
Metschnikowia sp. 01-655c1	sma-	25986	14101
M. torresii	tor-	8347	2886

M. hawaiiana lacked introns in the *cox1* gene, such that the size was only 1.6 kb. Notably, *M. hawaiiana* was the only *Metschnikowia* species that completely lacked introns in its mitochondrial genome. In contrast, the *cox1* gene of *M. cubensis* is 49 kb long and is currently the second largest organelle gene ever sequenced in all eukaryotes (Lee et al. 2020). *Metschnikowia cubensis*, however, did not have the highest number of introns; it had larger introns. *Metschnikowia* sp. *flo* contained the highest number of *cox1* introns in the large-spored clade, with 20 introns. Another intron-rich mitochondrial gene, *cob*, exhibited less dramatic variation in both size and number of introns; length varied 22-fold (1.1 kb vs 23 kb) and the numbers ranged from 0 to 13. The *cob* genes of seven *Metschnikowia* species currently rank within the ten largest *cob* loci sequenced so far (Lee et al. 2020).

Introns were not unique to *cox1* and *cob* genes. *Metschnikowia* species also contained various numbers of introns in *rnl, rns, nad1, nad2,* and *nad5* genes, but in lower numbers. A maximum of one group I intron was found in *nad1* and *nad2* and up to three within *nad5* gene loci (**Table 6**).

Table 6: Number of group I, group II, and intron-encoded ORFs within the *nad1* and *nad5* loci of *71 Metschnikowia* strains.

Code	*nad1*			*nad5*		
	Group I	**Group II**	**ORF**	**Group I**	**Group II**	**ORF**
abe+	1	1	0	1	0	0
abe-	1	1	0	1	0	0
aga+	0	0	0	0	0	0
aga-	0	0	0	0	0	0
ama+	1	1	0	3	2	1
ama-	1	1	0	3	2	1
ari+	0	0	0	2	1	1
ari-	0	0	0	2	1	1
bor+	0	0	0	2	1	1
bor-	0	0	0	2	1	1
bow+	1	1	0	3	2	1
bow-a	1	1	0	2	2	0
bow-b	1	1	0	2	2	0
cau+	0	0	0	0	0	0
cau-	0	0	0	0	0	0
cer+	1	1	0	1	0	1
cer-	1	1	0	2	2	0
cla+	0	0	0	0	0	0
col+	0	0	0	2	2	0
col-	0	0	0	2	2	0
con+	1	1	0	2	2	0
con-	1	1	0	2	0	0
cos-	0	0	0	2	2	0
cub+	0	0	0	0	0	0
cub-	0	0	0	0	0	0
dek+	1	1	0	2	2	0
dek-	1	1	0	2	2	0
dekY	1	1	0	2	2	0
dra+	1	1	0	1	1	0
dra-	1	1	0	1	1	0
dro+	0	0	0	1	1	0
dro-	0	0	0	1	1	0
flo+	0	0	0	2	1	1
ham+	0	0	0	0	0	0
ham-	0	0	0	0	0	0
han+	0	0	0	0	0	0

haw+	1	1	0	0	0	0
haw-	1	1	0	0	0	0
hib+	1	1	0	1	1	0
hib-	1	1	0	1	1	0
ipo+	1	1	0	2	2	0
ipo-a	1	1	0	2	2	0
ipov	1	1	0	2	2	0
kam+	1	1	0	0	0	0
kam-	1	1	0	1	0	0
kip-	1	1	0	0	0	0
lac+	1	1	0	2	2	0
lac-	1	1	0	2	2	0
lacb	1	1	0	2	2	0
loc+	1	1	0	2	2	0
loc-	1	1	0	2	2	0
mar-	1	1	0	2	2	0
mat+	1	1	0	2	2	0
mat-	1	1	0	2	2	0
mau+	1	1	0	0	0	0
mau-	1	1	0	0	0	0
mer-	1	1	0	2	2	0
ori+	0	0	0	1	1	0
ori-	0	0	0	1	1	0
pal+	1	1	0	0	0	0
pil-	1	1	0	0	0	0
pro+	1	1	0	1	1	0
pro-	1	1	0	1	1	0
sce+	1	1	0	2	2	0
sce-	1	1	0	2	2	0
shi+	1	1	0	1	0	0
shi-	1	1	0	1	0	0
sim+	1	1	0	2	2	0
sim-	1	1	0	1	1	0
sma-	1	1	0	2	2	0
tor-	0	0	0	1	1	0

Group II introns had been identified in the small ribosomal subunit gene (*rns*) by Marinoni and Lachance (2004) in the five species of the continental subclade known at the time. Mitochondrial genome assembly confirmed these findings and a single *rns* intron was found in 36 out of 71 *Metschnikowia* strains (19 out of 39 species, including ones yet to be described). The distribution of *rns* introns is given in **Table 7**. It should be noted that despite their variable lengths, *rns* introns shared the same insertion site and were all group II introns. Considering how similar *rns* introns are in their overlapping regions when aligned and the presence of an intron in the early-emerging species *M. orientalis*, the most parsimonious explanation is that a single insertion event occurred when the large-spored clade first emerged. In the remaining 35 strains (23 out of 39 species), inserted introns were probably lost over time. The species *M. bowlesiae*, *M. orientalis*, and *M. santaceciliae* varied in the presence or absence of *rns* introns between strains of same species. The South African subclade species (*M. proteae* and *M. drakenbergensis*) and *M. colocasiae* contained intron-inside-intron motifs. The presence of an intron-encoded ORF for reverse transcriptase in *M. colocasiae* and the low sequence similarity with the intron-inside-intron motif of the South African species suggests that the *rns* loci of the South African species and *M. colocasiae* most likely arose as independent insertion events. The intron distribution within the large-ribosomal subunit gene (*rnl*), on the other hand, was more complex. Up to five introns, including both group I and group II, were detected (**Table 7**). Thirteen out of 71 strains also contained intron-encoded proteins that appear to have been acquired independently, except for an endonuclease (HNH peptide motif) found in the interrelated species *M. borealis*, *M. cubensis*, and the undescribed species "*flo*". The HNH endonuclease insertion event is likely to have occurred in the common ancestor of those three species.

Table 7: Number of group I, group II, and intron-encoded ORFs within the *rns* and *rnl* loci of *71 Metschnikowia* strains

Code	*rns*			*rnl*		
	Group I	**Group II**	**ORF**	**Group I**	**Group II**	**ORF**
abe+	0	0	0	1	0	0
abe-	0	0	0	1	0	0
aga+	0	0	0	1	0	0
aga-	0	0	0	1	0	0
ama+	0	1	0	3	2	0
ama-	0	1	0	3	2	0
ari+	0	1	0	2	1	1
ari-	0	1	0	2	1	1
bor+	0	0	0	1	2	2
bor-	0	0	0	1	2	2
bow+	0	0	0	2	1	0
bow-a	0	1	0	2	2	0
bow-b	0	1	0	1	1	1
cau+	0	0	0	0	0	0
cau-	0	0	0	0	0	0
cer+	0	1	0	3	2	0
cer-	0	1	0	3	2	0
cla+	0	0	0	0	0	0
col+	0	1	1	2	0	0
col-	0	1	1	2	0	0
con+	0	1	0	3	2	0
con-	0	1	0	3	2	0
cos-	0	0	0	3	0	0
cub+	0	0	0	2	1	2
cub-	0	0	0	2	1	2
dek+	0	1	0	0	0	0
dek-	0	1	0	2	2	0
dekY	0	1	0	2	1	0
dra+	0	1	0	3	1	0
dra-	0	1	0	3	1	0
dro+	0	0	0	1	0	0
dro-	0	0	0	1	0	0
flo+	0	0	0	1	1	2
ham+	0	0	0	2	1	0
ham-	0	0	0	2	1	0
han+	0	0	0	0	0	0

haw+	0	0	0	2	1	0
haw-	0	0	0	2	1	0
hib+	0	0	0	1	0	0
hib-	0	0	0	1	0	0
ipo+	0	1	0	3	2	0
ipo-a	0	1	0	3	2	0
ipov	0	1	0	3	2	0
kam+	0	0	0	2	0	0
kam-	0	0	0	2	1	0
kip-	0	0	0	2	0	1
lac+	0	1	0	2	2	0
lac-	0	1	0	2	2	0
lacb	0	1	0	2	2	0
loc+	0	1	0	3	2	0
loc-	0	1	0	3	2	0
mar-	0	1	0	2	1	0
mat+	0	1	0	3	2	0
mat-	0	1	0	3	2	0
mau+	0	0	0	2	0	2
mau-	0	0	0	1	1	0
mer-	0	1	0	3	2	0
ori+	0	0	0	3	0	1
ori-	0	1	0	3	0	1
pal+	0	0	0	2	1	0
pil-	0	0	0	1	1	1
pro+	0	1	0	3	1	0
pro-	0	1	0	3	1	0
sce+	0	0	0	3	2	0
sce-	0	1	0	3	2	0
shi+	0	0	0	1	0	0
shi-	0	0	0	1	0	0
sim+	0	1	0	2	2	0
sim-	0	1	0	1	2	0
sma-	0	1	0	3	2	0
tor-	0	0	0	1	0	0

3.12 Mitochondrial Intron Insertion Sites

Among all *Metschnikowia* species studied, there were 25 and 13 possible intron insertion sites in the *cox1* and *cob* genes, respectively. Many strains within the same species shared similar intron insertion patterns in those genes but there were also many cases where strains of the same species differed. It should be noted that the intron sequences within the same insertion sites were not necessarily identical between strains such that the presence and absence of intron-encoded proteins also varied between strains, thereby adding one more layer of complexity in mitochondrial diversity. In early-emerging LSC and non-LSC *Metschnkowia* species, not only was the number of *cox1* and *cob* introns within the single digits, but those introns also mostly contained homing endonucleases (**Figure 23**).

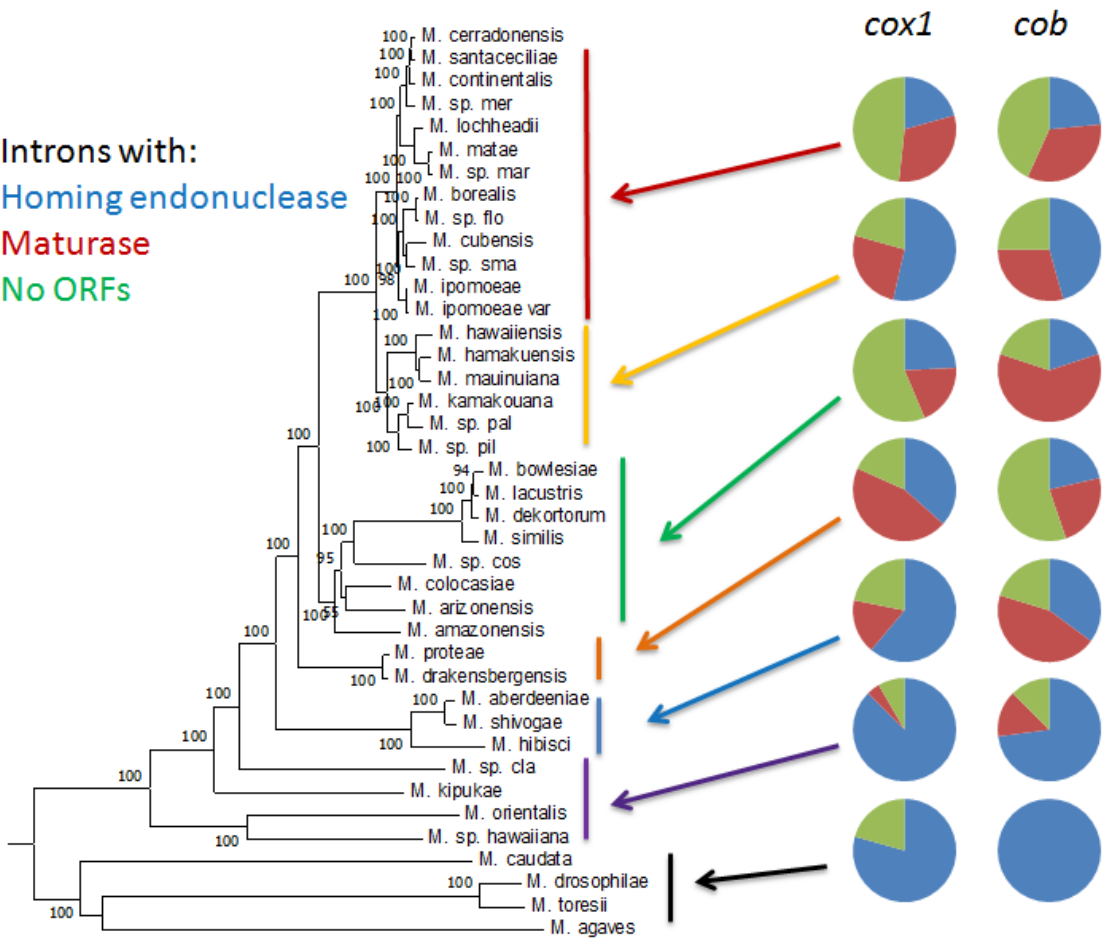

Figure 23: Proportion of protein-encoding (homing endonuclease, maturase/reverse transcriptase) introns or absence of an ORF in the mitochondrial genomes of 71 *Metschnikowia* strains. Each pie chart represents the average proportion of introns in each subclade.

The higher frequency of maturases and introns without ORFs in later-emerging species suggests that the mitochondrial genomes of these strains were invaded by group I introns and subsequently by group II introns or introns without ORFs. Small introns of less than 50 bases in length with no identifiable types might be remnants of previously inserted group I or II introns that experienced prolonged mutations and eventual deletions.

Intron insertion sites and the presence and absence of intron-encoded ORFs are shown in **Table 8**, **Figure 24** and **Figure 25**. Most of the endonucleases detected were of the LAGLIDADG, but there were a few cases of HNH and GIY endonuclease sequences in

cox1 introns. LAGLIDADG endonucleases inserted into the #16 and #22 potential *cox1* intron sites were nearly always present, suggesting an ancient insertion event. The #4 *cox1* intron site is a unique case, where two different introns respectively encoding a LAGLIDADG endonuclease and a maturase were found, suggesting that an intron has been inserted within an intron. It is currently not clear whether the presence of two genes is the result of two insertion events or the single insertion of a locus containing both maturase and endonuclease-encoding sequences. In contrast, *cob* introns contained more GIY endonucleases than their *cox1* counterparts due to the conserved insertion of GIY loci at the #6 *cob* intron site. With a few exceptions, most endonucleases were found within group I introns whereas maturases were found within group II introns. Intron site #6 of *cox1* is a special case where both maturases and endonucleases were found, sometimes simultaneously, which is unique in the mitochondrial genomes assembled in this thesis. Intron position #7 of *cob* was the only other case where either a maturase or an endonuclease was found within introns but not simultaneously.

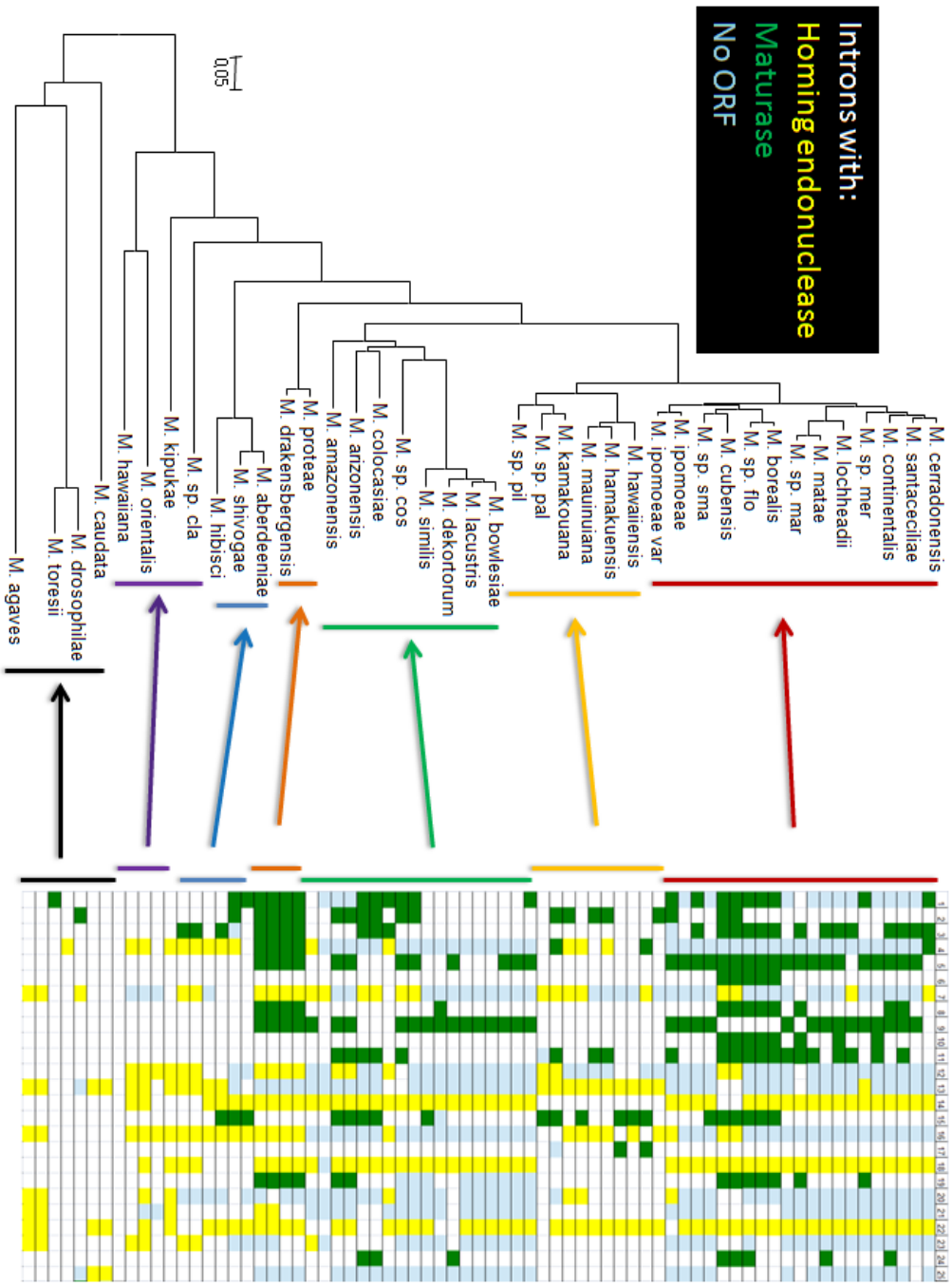

Figure 24: Intron insertion pattern map for the *cox1* gene of 71 mitochondrial genomes of *Metschnikowia* strains. Each box represents absence of introns (white), introns with a homing endonuclease (yellow), introns with a maturase/reverse transcriptase (green), and introns without any detectable ORFs (light blue) in one of 25 possible intron insertion sites. Coloured lines and arrows indicate patterns corresponding to matching subclades. The tree was modified from Figure 3

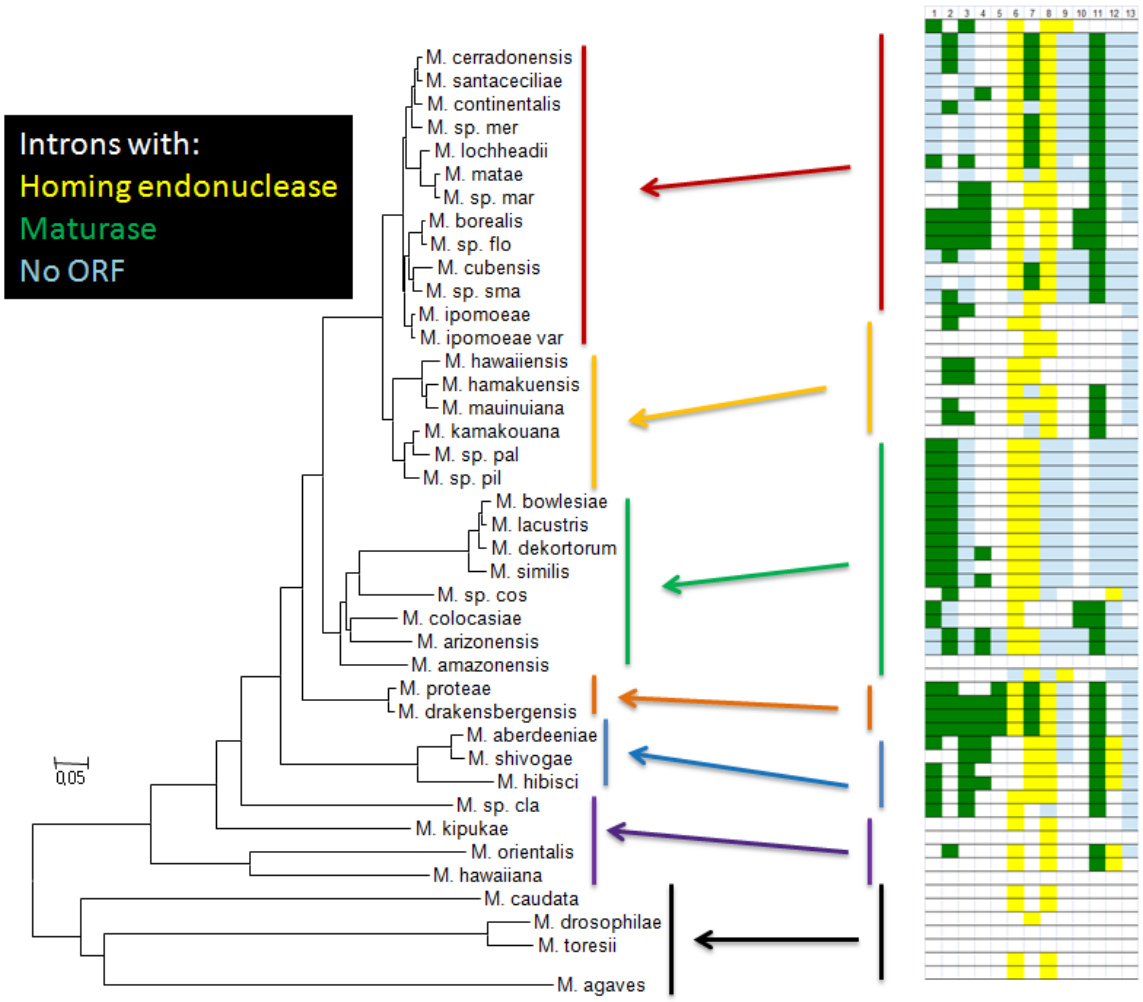

Figure 25: Intron insertion pattern map for the *cob* gene of 71 mitochondrial genomes of *Metschnikowia* strains. Each box represents absence of introns (white), introns with a homing endonuclease (yellow), introns with a maturase/reverse transcriptase (green), and introns without any detectable ORFs (light blue) in one of 13 possible intron insertion sites. Coloured lines and arrows indicate patterns corresponding to matching subclades. The tree was modified from Figure 3.

Table 8: Number of group I, group II, and intron-encoded ORFs within the *cox1* and *cob* loci of 71 *Metschnikowia* strains

Code	*cox1*			*cob*		
	Group I	Group II	ORF	Group I	Group II	ORF
abe+	4	3	7	4	3	6
abe-	5	1	5	3	4	7
aga+	7	0	7	2	0	2
aga-	7	0	7	2	0	2
ama+	8	9	11	5	7	5
ama-	8	9	11	5	7	5
ari+	6	8	11	2	4	5
ari-	6	8	11	2	4	5
bor+	7	11	15	3	3	5
bor-	7	11	15	3	3	5
bow+	10	6	6	5	4	4
bow-a	10	5	5	5	4	4
bow-b	10	5	5	5	4	4
cau+	1	0	1	0	0	0
cau-	0	1	1	0	0	0
cer+	8	9	10	6	5	5
cer-	9	10	10	6	5	5
cla+	8	1	8	2	0	2
col+	7	3	5	0	0	0
col-	7	2	6	5	0	2
con+	9	7	9	6	4	4
con-	10	6	10	6	5	5
cos-	8	3	9	4	1	4
cub+	4	13	18	2	6	8
cub-	4	13	18	2	6	8
dek+	10	4	4	5	4	4
dek-	10	4	4	5	4	4
dekY	10	4	4	5	4	4
dra+	8	9	14	4	4	7
dra-	8	9	14	4	6	9
dro+	3	0	3	2	0	2
dro-	3	0	3	2	0	2
flo+	8	13	15	3	6	8
ham+	4	0	4	3	0	2
ham-	3	1	4	3	0	2
han+	0	0	0	0	0	0

haw+	6	2	7	2	2	4
haw-	6	1	6	2	2	4
hib+	6	1	8	4	3	6
hib-	6	1	8	3	3	5
ipo+	10	6	9	5	2	4
ipo-a	10	5	7	6	4	4
ipov	9	7	12	6	5	4
kam+	5	1	7	3	1	4
kam-	5	2	7	4	2	5
kip-	5	0	2	1	0	1
lac+	7	4	5	5	4	4
lac-	10	2	5	5	4	4
lacb	10	5	6	5	5	5
loc+	6	8	11	6	3	6
loc-	10	6	8	6	4	4
mar-	10	6	8	6	4	4
mat+	9	8	10	6	4	4
mat-	10	7	9	6	4	4
mau+	5	1	4	2	2	3
mau-	4	3	6	1	1	3
mer-	10	6	9	6	5	5
ori+	10	1	9	3	2	5
ori-	7	0	6	2	1	4
pal+	4	4	8	4	3	6
pil-	4	0	4	2	1	3
pro+	9	9	15	4	6	9
pro-	10	9	15	4	6	9
sce+	8	4	7	2	2	5
sce-	9	6	8	6	5	5
shi+	5	2	5	5	4	7
shi-	6	5	8	5	3	6
sim+	10	7	8	5	4	4
sim-	8	6	9	5	5	5
sma-	10	6	8	6	5	5
tor-	3	1	2	1	0	1

In *cob,* intron insertions with a GIY-YIG endonuclease at position #6 and a LAGLIDADG endonuclease at position #8 appear to be the most ancestral insertions in haplontic *Metschnikowia* species, as these insertions are found in the large-spored species as well as in the outgroup species *M. agaves* and *M. drosophilae.* The remaining intron insertions were specific to the large-spored species, suggesting that something triggered the proliferation of intron insertions in *cob.* In contrast, the seven introns found in *cox1* of *M. agaves* showed that proliferation of intron insertions in *cox1* was not exclusive to the large-spored species. Interestingly, all seven introns found in *M. agaves* shared the same location as those of most other *Metschnikowia* species, suggesting that these insertions preceded the emergence of the large-spored clade. The disappearance of endonuclease loci from those sites in some large-spored species, with the exception of position #22 of *cox1* gene, suggests that late-emerging species are in the process of losing these endonuclease loci.

It is currently unclear why intron insertions are so prevalent in *Metschnikowia* species, leading to an extreme variation in gene sizes. Rudan et al. (2018) reported that intron splicing directly affects the expression levels of both *cox1* and *cob,* suggesting that the number of intron insertions bears on the fitness of the species. It was once postulated that if the number of introns is inversely proportional to the expression level of host genes, then species in a competitive environment would be selected for having fewer introns in these genes to allow mitochondria to produce more energy for better growth. In contrast, species in a harsh environment would be selected to have more introns in mitochondrial genes to slow down mitochondrial gene expression and the resulting metabolism to survive without wasting resources. As *Metschnikowia arizonensis* resides within a harsh environment (a hot desert in Arizona), it would be reasonable to expect that species to be exposed to a selection towards possessing a high number of introns. However, a comparison of the total number of introns, including ones with or without ORFs, between strains showed that the two strains of *M. arizonensis* did not differ dramatically from other late-emerging species. Their 15 introns were fewer than were encountered in some other species, suggesting that the harsh environment hypothesis should be set aside, at least for now (**Figure 2 and Table S3**).

Although less enticing, the opposite view that intron diversity in mitochondrial genomes is purely the result of introns proliferating and has little if any association with fitness, should be considered. For this to be true, intron splicing should be fast and efficient, and the effect of introns on the expression of mature *cob* and *cox1* mRNA should either be minimal or compensated by other mechanisms. More research on the impact of introns on mitochondrial genes is needed to provide a more definite answer to this question.

Chapter 4

4 Conclusion

This thesis constructed mitochondrial genomes of haplontic *Metschnikowia* species and analyzed their various characteristics *in silico*. The first question of whether constructing a complete mitochondrial genome *in silico* from whole genome data without additional sequencing of isolated mitochondria was answered as possible in this thesis, with completion of 71 mitochondrial genomes. Some aspects of the assemblies, such as duplicated regions and linear genomes, are potentially debatable, but in such cases additional sequencing from isolated mitochondria may not eliminate doubts, such that the genomes constructed in this thesis are the best approximation available at the moment. Additional support from PCR amplifications strengthened the idea of the presence of a few linear genomes. The second question of whether nuclear and mitochondrial genomes would share similar patterns of relatedness was answered negatively, suggesting that mitochondrial genomes experience different evolutionary histories from their nuclear counterparts. Other genome characteristics such as gene order or introns further showed that evolution of mitochondrial genomes deviated from those of nuclear genomes. Additional analyses showed that mitochondrial genomes are highly diverse in size, morphology, gene order, tRNAs, and introns, even though the gene contents were identical. Some characteristics were so diverse that the species boundaries and relatedness determined from nuclear genomes or mating studies did not completely explain the phenomena present. My thesis revealed considerable mitochondrial diversity at the strain level, thereby suggesting that mitochondrial genomes evolve much faster than their nuclear counterparts. My thesis therefore reinforced the notion that mitochondrial genomes are important in studying genome evolution and diversity. It also showed the need to go beyond *in silico* analyses to unravel the possible biological mechanism responsible for the observed patterns. More traditional experiments like transformation studies would be helpful in exploring the driving forces behind rapid genome evolution shown in *Metschnikowia*. Nevertheless, my thesis has shown that mitochondrial genomes of *Metschnikowia* yeasts are interesting and I hope that they are studied further in the future.

Bibliography

Aguileta, G. Vienne, DM. Ross, ON. Hood, ME. Giraud, T. Petit, E and Gabaldon, T. 2014. High variability of mitochondrial gene order among fungi. Genome Biol Evol 6(2): 451-465.

Altschul, SF. Gish, W. Miller, W. Myers, EW and Lipman, DJ. 1990. Basic local alignment search tool. J Mol Biol 215(3): 403-410.

Bjelland, S and Seeberg, E. 2003. Mutagenicity, toxicity and repair of DNA base damage induced by oxidation. Mutat Res 531: 37-80.

Boyer, PD. 1965. In oxidases and related redox systems. John Wiley & Sons Inc., New York, 994-1008.

Brandvain, Y and Wade, MJ. 2009. The functional transfer of genes from the mitochondria to the nucleus: the effects of selection, mutation, population size and rate of self-fertilization. Genetics. 182(4): 1129-1139.

Brown, WM. Prager, EM. Wang, A and Wilson, AC. 1982. Mitochondrial DNA sequences of primates: tempo and mode of evolution. J Mol Evol 18:225-239.

Chandel, NS. 2015. Evolution of mitochondria as signaling organelles. Cell Metab 22(2): 204-206.

Chevalier, BS and Stoddard, BL. 2001. Homing endonucleases: structural and functional insight into the catalysts of intron/intein mobility. Nucleic Acids Res 29(18): 3757-3774.

Da Cunha, V. Gaia, M. Nasir, A and Forterre, P. Asgard archaea do not close the debate about the universal tree of life topology. PLoS Genetics 14(3): e1007215.

Drillon, G and Fischer, G. 2010. comparative study on synteny between yeasts and vertebrates. Comptes Rendus Biologies 334: 629-638.

Ernster, L and Schatz, G. 1981. Mitochondria: A historical review. The J Cell Bio 91: 227-255.

Feng, B. Zhou, L and Tang, J. 2017. Ancestral genome reconstruction on whole genome level. Current Genomics 18:306-315.

Freel, KC. Friedrich, A and Schacherer, J. 2015. Mitochondrial genome evolution in yeasts: an all-encompassing view. FEMS Yeast Research 15: fov023.

Fritsch, ES. Chabbert, CD. Klaus, B and Steinmetz, LM. 2014. A genome-wide map of mitochondrial DNA recombination in yeast. Genetics 198(2): 755-771.

Gan, HM. Grandjean, F. Jenkins, TL and Austin, CM. 2019. Absence of evidence is not evidence of absence: Nanopore sequencing and complete assembly of the European lobster (*Homarus gammarus*) mitogenome uncovers the missing nad2 and a new major gene cluster duplication. BMC Genomics 20: 335.

Gordon, Z et al. 2019. Development of a transformation method for *Metschnikowia borealis* and other CUG-Serine yeasts. Genes 10:78.

Guha, TK. Wai, A. Mullineux, ST and Hausner, G. 2018. The intron landscape of the mtDNA cytb gene among the Ascomycota: introns and intron-encoded open reading frames. Mitochondrial DNA Part A. 29(7): 1015-1024.

Hahn, A and Zuryn, S. 2019. Mitochondrial genome (mtDNA) mutations that generate reactive oxygen species. Antioxidants 8(9): 392.

Hebert, PDN. Cywinska, A. Ball, SL and deWaard, JR. 2003. Biological identifications through DNA barcodes. Proc Biol Sci 270(1512): 313-21

Ingman, M and Gyllensten, U. 2001. Analysis of the complete human mtDNA genome: methodology and inferences for human evolution. Journal of Heredity 92(6): 454-461.

James, JE. Piganeau, G and Eyre-Walker, A. 2016. The rate of adaptive evolution in animal mitochondria. Mol Ecol 25(1): 67-78.

Kang, H. Li, B. Ma, X and Xu, Y. 2018. Evolutionary progression of mitochondrial gene rearrangements and phylogenetic relationships in Strigidae (Strigiformes). Gene 674: 8-14.

Kanzi, AM. Wingfield, BD. Steenkamp, ET. Naidoo, S and van der Merwe, NA. 2016. Intron derived size polymorphism in the mitochondrial genomes of closely related *Chrysoporthe* species. PloS One 11(6): e0156104.

Khachane, AN. Timmis, KN and Santos, VAPM. 2006. Dynamics of reductive genome evolution in mitochondria and obligate intracellular microbes. Mol Biol Evol 24(2): 449-56.

Krassowski, T et al. 2018. Evolutionary instability of CUG-Leu in the genetic code of budding yeasts. Nature communications 9: 1887.

Kurabayashi, A and Sumida, M. 2013. Afrobactrachian mitochondrial genomes: genome reorganization, gene rearrangement mechanisms, and evolutionary trends of duplicated and rearranged genes. BMC genomics 14: 633.

Kurtzman, CP and Robnett, CJ. 1998. Identification and phylogeny of ascomycetous yeasts from analysis of nuclear large subunit (26S) ribosomal DNA partial sequences. Antonie van Leeuwenhoek 73: 331-371.

Kurtzman, CP. 1989. Estimation of phylogenetic distances among ascomycetous yeasts from partial sequencing of ribosomal RNA. Yeast 5: 351-354

Lachance, MA and Bowles, JM. 2002. *Metschnikowia arizonensis* and *Metschnikowia dekortorum*, two new large-spored yeast species associated with floricolous beetles. FEMS Yeast Research 2: 81-86.

Lachance, MA. Ewing, CP. Bowles, JM and Starmer, WT. 2005. *Metschnikowia hamakuensis* sp. nov., *Metschnikowia kamakouana* sp. nov. and *Metschnikowia mauinuiana* sp. nov., three endemic yeasts from Hawaiian nitiduid beetles. Int J Syst Evol Microbiol 55: 1369-1377.

Lachance, MA. Lawrie, D. Dobson, J and Piggott, J. 2008. Biogeography and population structure of the Neotropical endemic yeast species *Metschnikowia lochheadii*. Antonie van Leeuwenhoek 94:403-414.

Lachance, MA. 2016. *Metschnikowia*: half tetrads, a regicide and the fountain of youth. Yeast. 33(11): 563-574.

Lachance, MA. Hurtado, E and Hsiang, T. 2016. A stable phylogeny of the large-spored *Metschnikowia* clade. Yeast 33: 261-275.

Lachance, MA, Lee, DK, and Hsiang, T. 2020. Delineating yeast species with genome average nucleotide identity: a calibration of ANI with haplontic, heterothallic *Metschnikowia* species. Antonie van Leeuwenhoek 113: 2097-2106.

Lang, BF. Laforest, MJ and Burger, G. 2007. Mitochondrial introns: a critical view. Trends Genet 23(3): 119-125.

Lee, DK. Hsiang, T and Lachance, MA. 2018. *Metschnikowia* mating genomics. Antonie van Leeuwenhoek 111: 1935-1953.

Lee, DK. Santos, ARO. Hsiang, T. Rosa, CA and Lachance, MA. 2020. Catching speciation in the act-act 2: *Metschnikowia lacustris* sp. Nov., a sister species to *Metschnikowia dekortorum*. Antonie van Leeuwenhoek. 113(6): 753-762.

Lee, DK. Hsiang, T. Lachance, MA and Smith, DR. 2020. The strange mitochondrial genomes of *Metschnikowia* yeasts. Curr Biol 30(14): 800-801.

Leebens-Mack, JH et al. 2019. One thousand plant transcriptomes and the phylogenomics of green plants. Nature 574: 679-685.

Lopez-Garcia, P and Moreira, D. 2015. Open questions on the origin of eukaryotes. Trends Ecol Evol 30(11): 697-708.

Lynch, M. 2010. Evolution of the mutation rate. Trends Genet 26(8): 345-352.

Marinoni, G and Lachance, MA. 2004. Speciation in the large-spored *Metschnikowia* clade and establishment of a new species, *Metschnikowia borealis* comb. nov. FEMS Yeast Research 4: 587-596.

Marinoni, G. Manuel, M. Petersen, RF. Hvidtfeidt, J. Sulo, P and Piskur, J. 1999. Horizontal transfer of genetic material among *Saccharomyces* Yeasts. J Bacteriology 181: 20

Molitor, C. Inthavong, B. Sage, L. Geremia, RA and Mouhamdou B. 2010. Potentiality of the cox1 gene in the taxonomic resolution of soil fungi. FEMS Microbiol Lett 302(1): 76-84.

Nosek, J. Tomaska, L. Fukuhara, H. Suyama, Y and Kovac, L. 1998. Linear mitochondrial genomes: 30 years down the line. Trends Genet 14(5): 184-188.

Nunnary, J and Soumalainen, A. 2012. Mitochondria: in sickness and in health. Cell 148(6): 1145-1159.

Perrin, A. Varre, JS. Blanquart, S and Ouangraoua, A. 2015. ProCARs: Progressive Reconstruction of Ancestral Gene Orders. BMC Genomics 16:S6.

Porter, TM and Hajibabaei, M. 2018. Over 2.5 million COI sequences in GenBank and growing. PLoS ONE 13(9): e0200177.

Repar, J and Warnecke, T. 2017. Mobile introns shape the genetic diversity of their host genes. Genetics 205: 1641-1648.

Riley, R et al. 2016. Comparative genomics of biotechnologically important yeasts. PNAS 113:35.

Roger, AJ. Munoz-Gomez, SA and Kamikawa, R. 2017. The origin and diversification of mitochondria. Curr Biol 27(21): 1177-1192.

Roger, AJ. Susko, E and Leger, MM. 2021. Evolution: Reconstructing the timeline of eukaryogenesis. Curr Biol 31: 186-214.

Rudan, M et al. 2018. Normal mitochondrial function in *Saccharomyces cerevisiae* has become dependent on inefficient splicing. Elife 7:e35330.

Ruan, J. Cheng, J. Zhang, T and Jiang, H. 2017. Mitochondrial genome evolution in the *Saccharomyces sensu stricto* complex. PloS ONE 12(8): e0183035.

Santos, ARO. Lee, DK. Ferreira, AG. Carmo, MC. Rondelli, VM. Barros, KO. Hsiang, T. Rosa, CA and Lachance, MA. 2020. The yeast community of *Conotelus* sp. (*Coleoptera*: Nitidulidae) in Brazilian passionfruit flowers (*Passiflora edulis*) and description of *Metschnikowia amazonensis* sp. nov., a large-spored clade yeast. Yeast. 37: 253-260.

Schneider, A and Ebert, D. 2004. Covaration of mitochondrial genome size with gene lengths: evidence for gene length reduction during mitochondrial evolution. J Mol Evol 59:90-96.

Selosse, MA. Albert, B and Godelle, B. 2001. Reducing the genome size of organelles favours gene transfer to the nucleus. Trends Ecol Evol 16(3):135-141.

Shen, X et al. 2018. Tempo and mode of genome evolution in the budding yeast subphylum. Cell 175: 1533-1545.

Shibayama, K. Otoguro M. Nakashima, C and Yanagida, F. *Metschnikowia miensis* f.a. sp. nov., isolated from flowers in Mie prefecture, Japan. Antonie van Leeuwenhoek 113: 321-329.

Smith, DR and Keeling, PJ. 2015. Mitochondrial and plastid genome architecture: reoccurring themes, but significant differences at the extremes. Proc Natl Acad Sci U S A 112(33): 10177-84.

Solieri, L. 2010. Mitochondrial inheritance in budding yeasts: towards an integrated understanding. Trends Microbiol 18(11): 521-530.

Spang, A et al. 2015. Complex archaea that bridge the gap between prokaryotes and eukaryotes. Nature 521(7551): 173-179.

Sulo, P. Szaboova, D. Bielik, P. Polakova, S. Soltys, K. Jatzova, K and Szemes, T. 2017. The evolutionary history of *Saccharomyces* species inferred from completed mitochondrial genomes and revision in the yeast mitochondrial genetic code. DNA Research 24(6): 571-583.

Sultan, LD et al. 2016. The reverse transcriptase/RNA maturase protein MatR is required for the splicing of various group II introns in Brassicaceae mitochondria. The Plant cell 28: 2805-2829.

Tamura, K. Peterson, D. Peterson, N. Stecher, G. Nei, M and Kumar, S. 2011. MEGA5: Molecular evolutionary genetic analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. Mol Biol Evol 28: 2731-2739.

Tzameli, I. 2012. The evolving role of mitochondria in metabolism. Trends Endocrinol Metab 23(9): 417-419.

Valach, M et al. 2011. Evolution of linear chromosomes and multipartite genomes in yeast mitochondria. Nucleic Acids Res 39(10): 4202-4219.

Waterhouse, RM et al. 2017. BUSCO applications from quality assessments to gene prediction and phylogenomics. Molecular Biology and Evolution 35: 543-548.Xiao, S. Nguyen, DT. Wu, B and Hao, W. 2017. Genetic drift and indel mutation in the evolution of yeast mitochondrial genome size. Genome Biol Evol 9(11): 3088-3099.

Wilson, AJ and Xu, J. 2012. Mitochondrial inheritance: diverse patterns and mechanisms with an emphasis on fungi. Mycology 3(2): 158-166.

Zhao, C and Pyle, AM. 2017. The group II intron maturase: a reverse transcriptase and splicing factor go hand in hand. Curr Opin Struct Biol 47: 30-39.

Appendices

Appendix A: Graphical representation of all mitochondrial genomes of haplontic *Metschnikowia* strains studied.

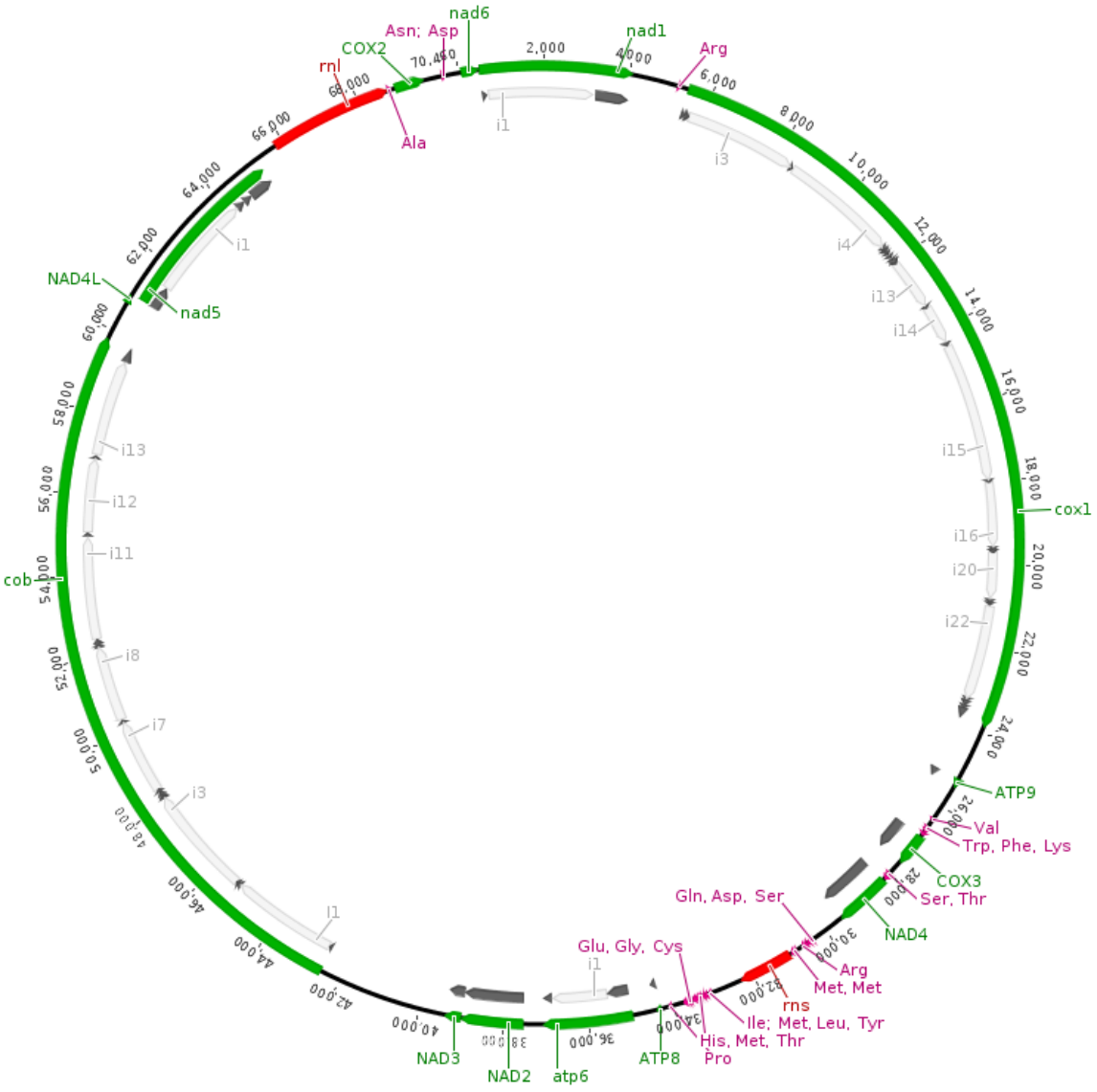

Figure S 1: Mitochondrial genome of *Metschnikowia aberdeeniae* UWOPS07-202.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA, respectively.

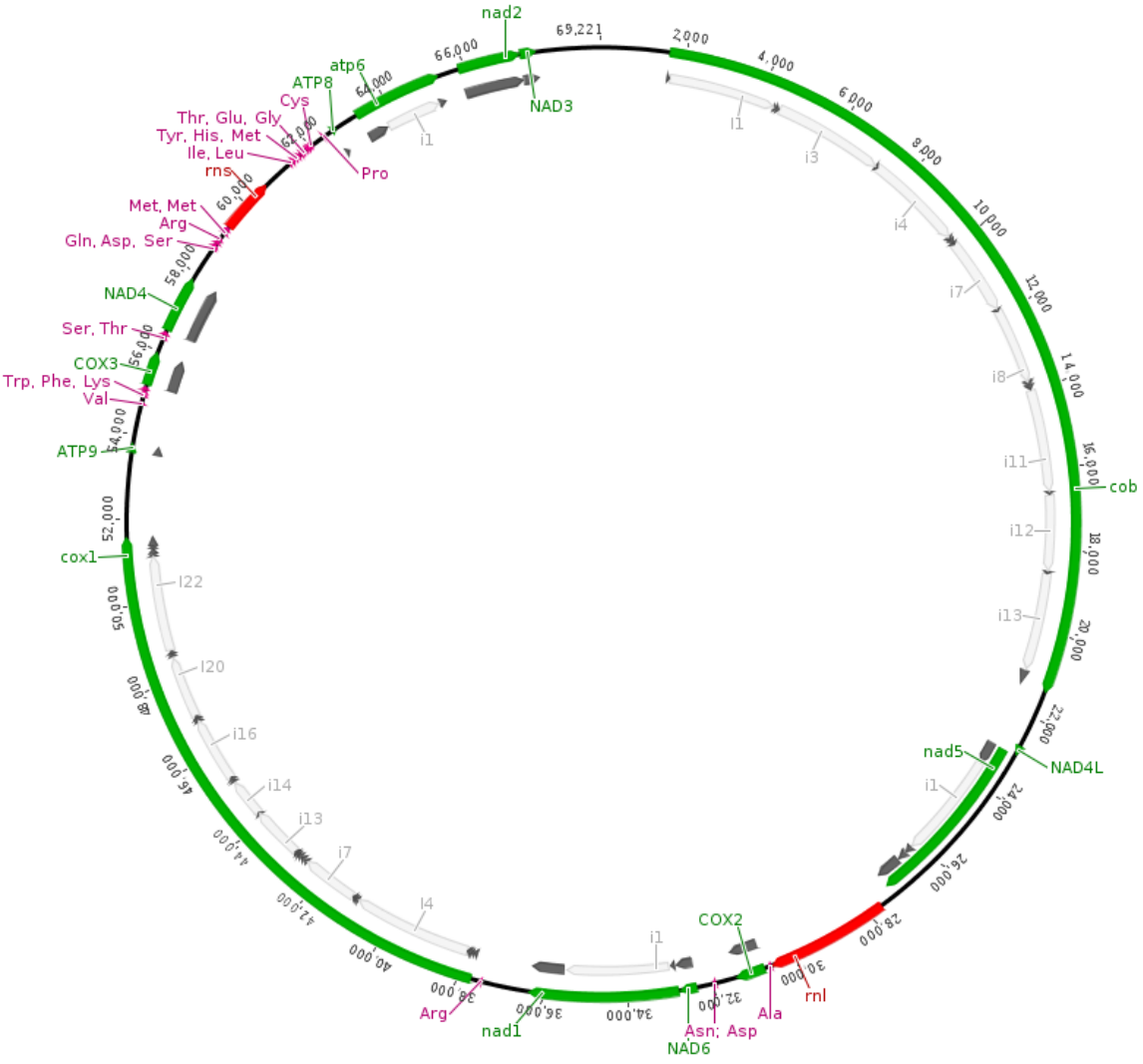

Figure S 2: Mitochondrial genome of *Metschnikowia aberdeeniae* SUB 05-213.18. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

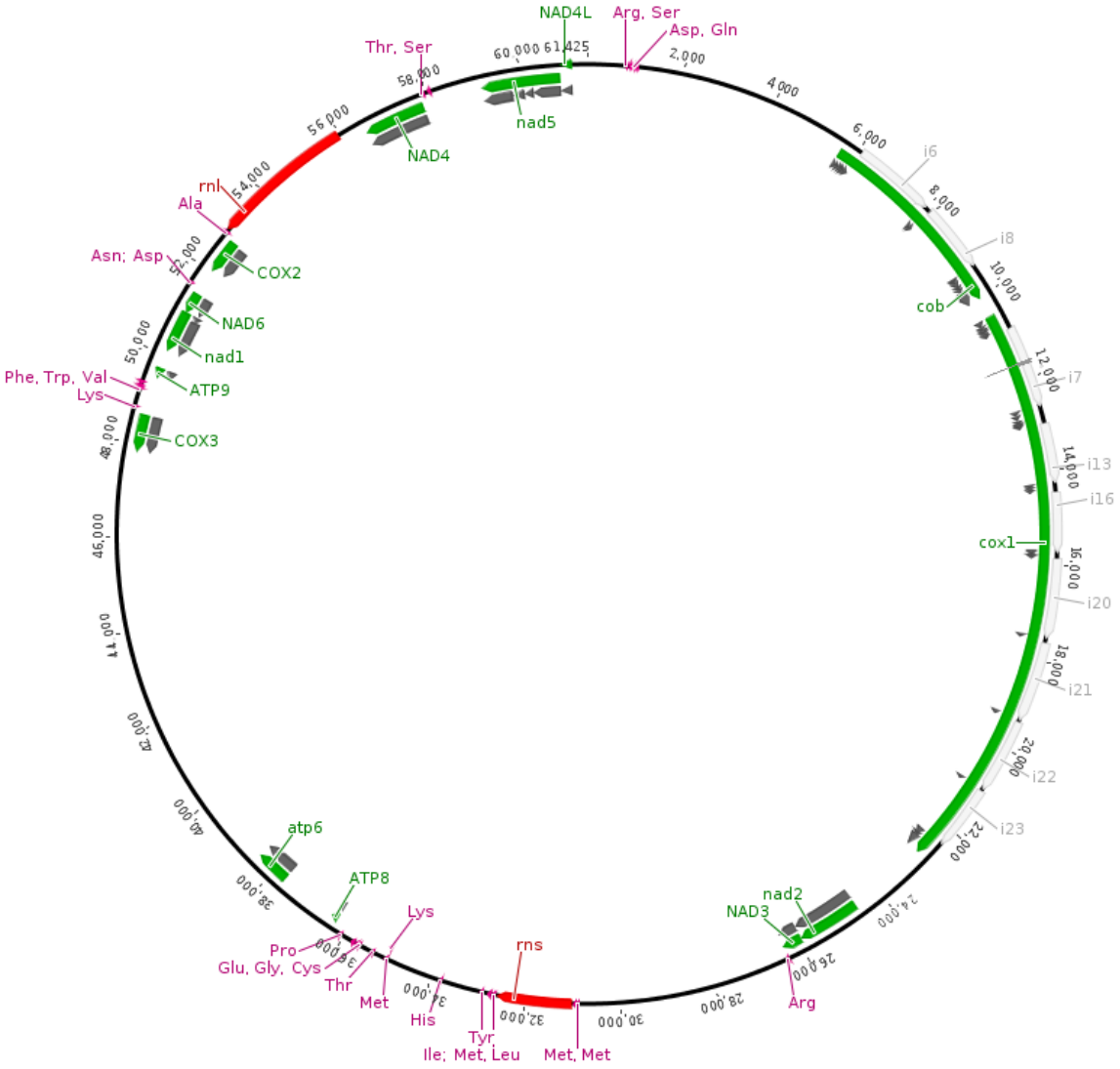

Figure S 3: Mitochondrial genome of *Metschnikowia agaves* UWOPS92-207.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

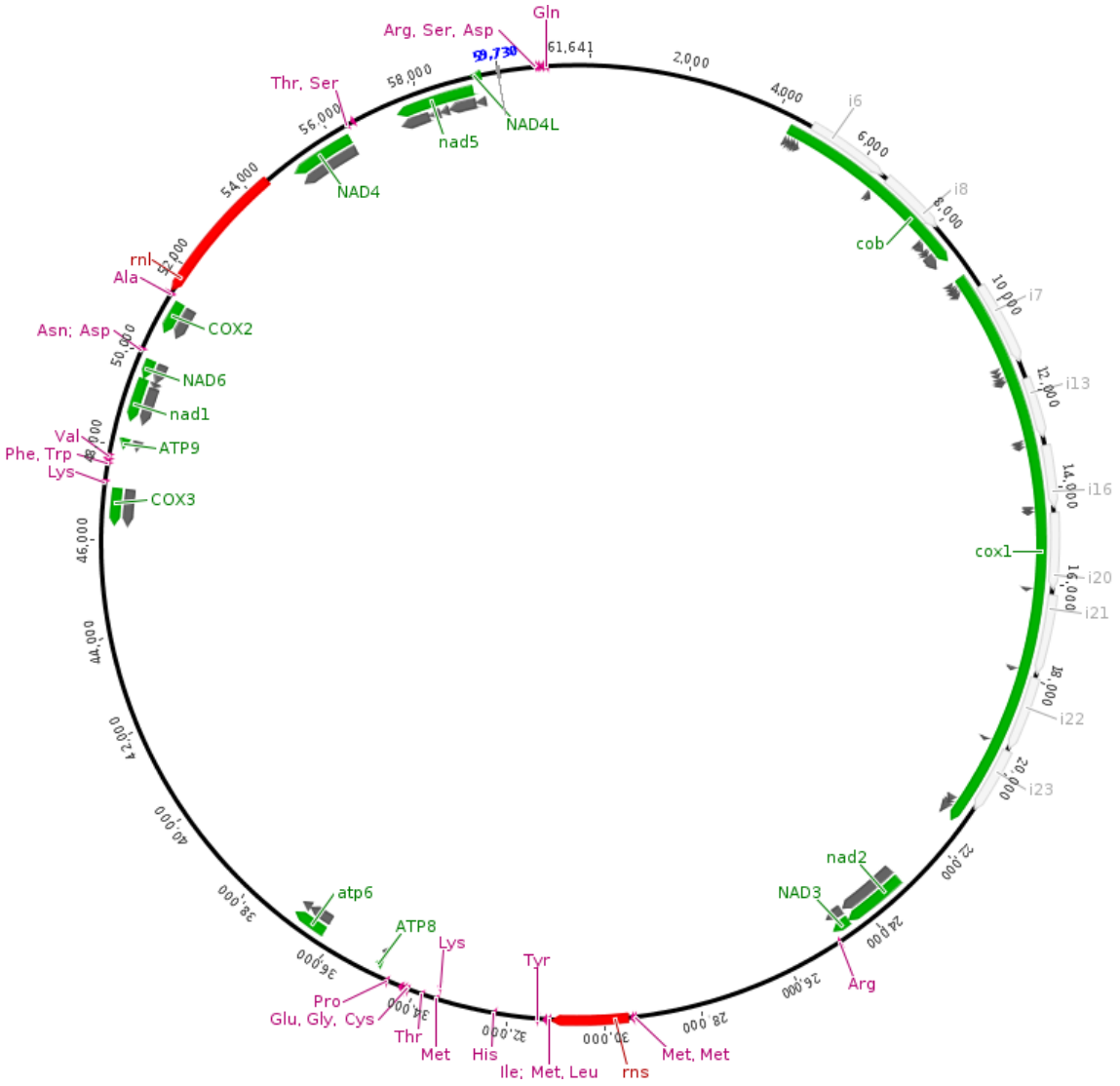

Figure S 4: Mitochondrial genome of *Metschnikowia agaves* UWOPS92-210.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

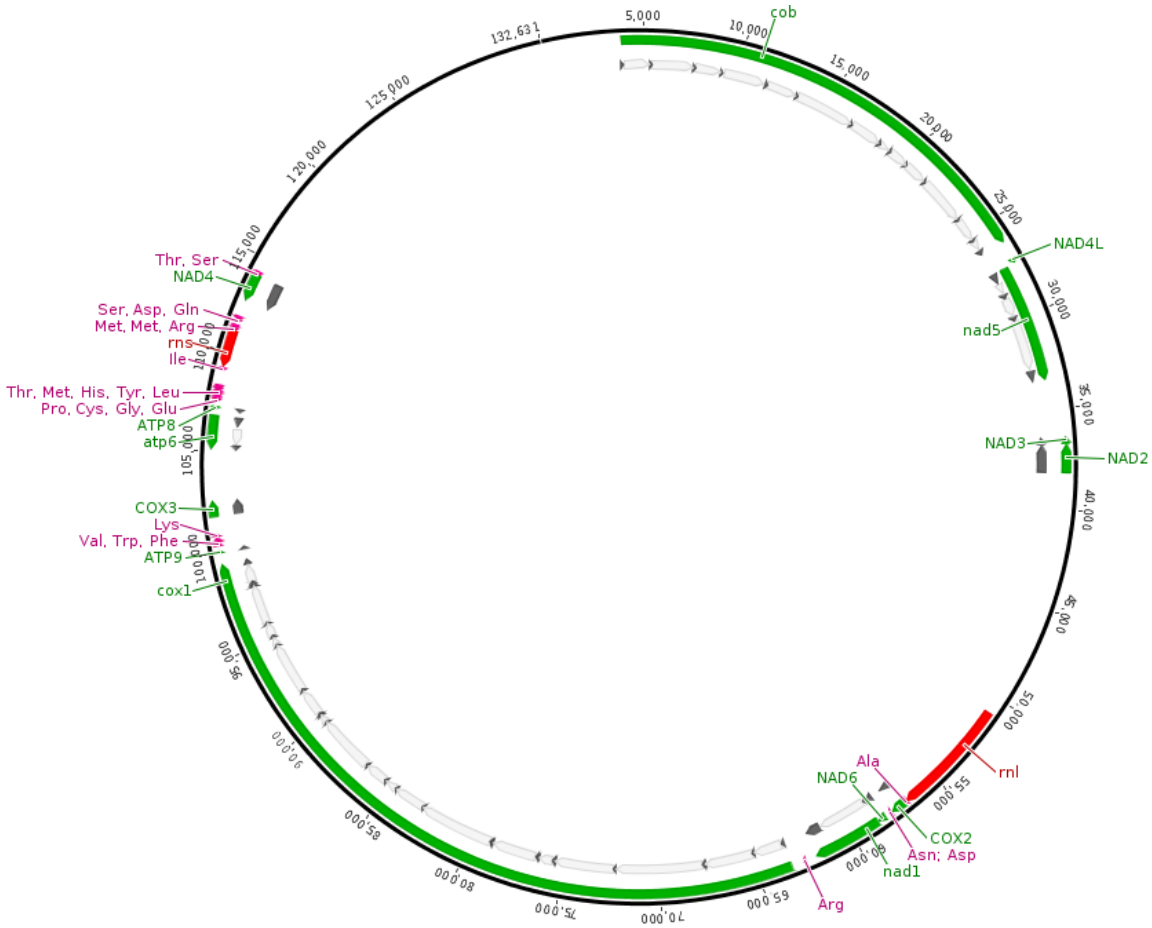

Figure S 5: Mitochondrial genome of *Metschnikowia amazonensis* UFMG-CM-Y6309. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

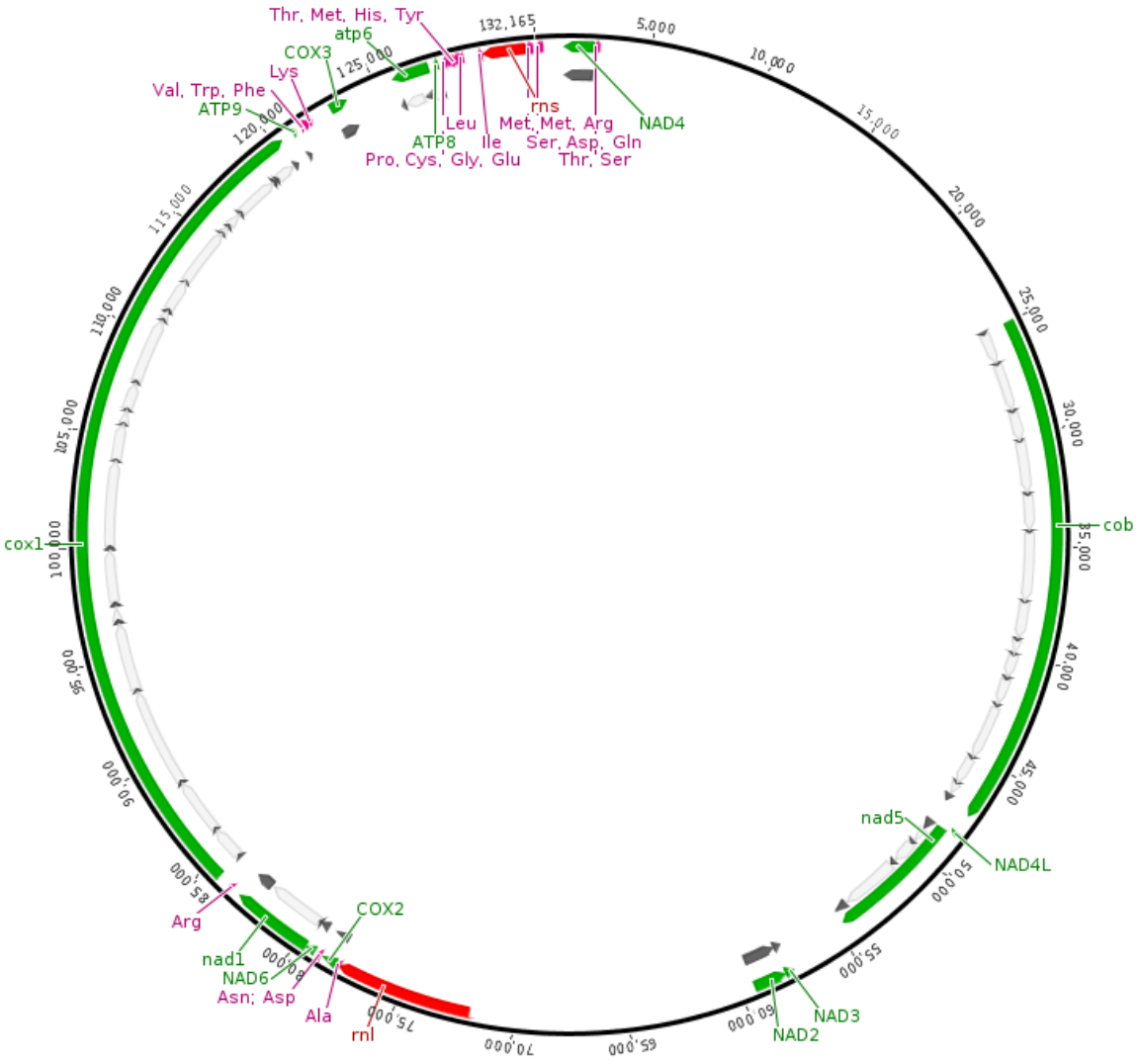

Figure S 6: Mitochondrial genome of *Metschnikowia amazonensis* UFMG-CM-Y6307. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

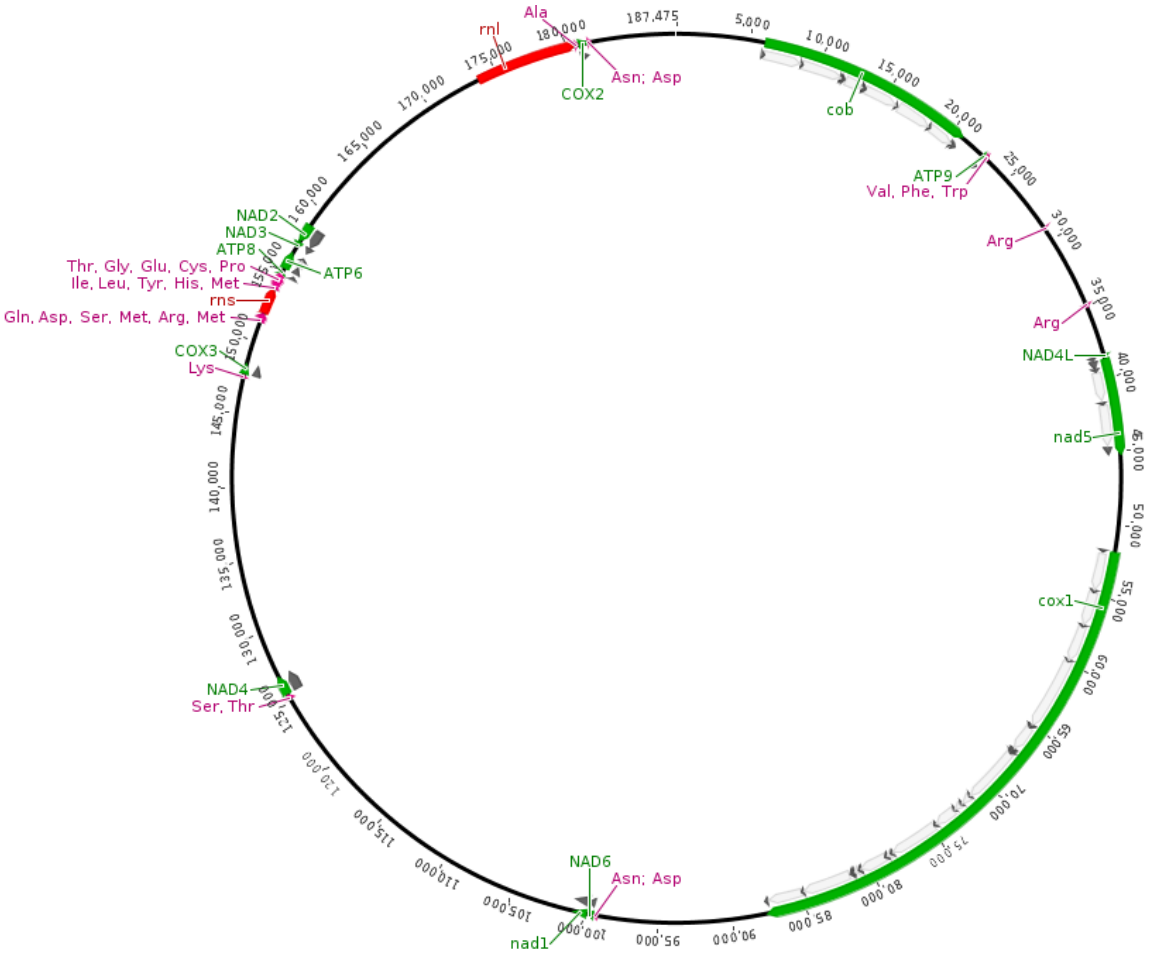

Figure S 7: Mitochondrial genome of *Metschnikowia arizonensis* UWOPS99-103.3.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

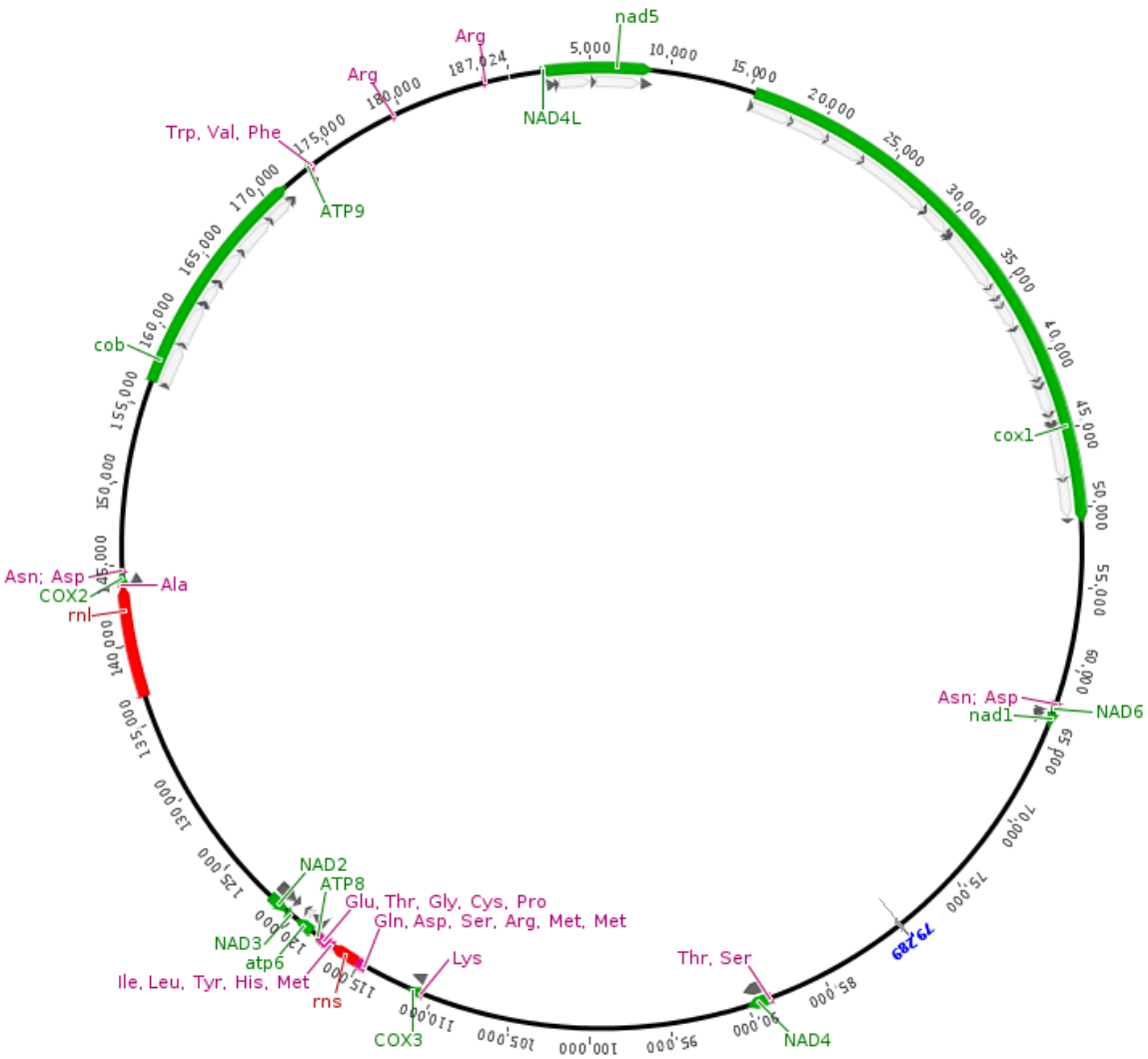

Figure S 8: Mitochondrial genome of *Metschnikowia arizonensis* UWOPS99-103.4. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

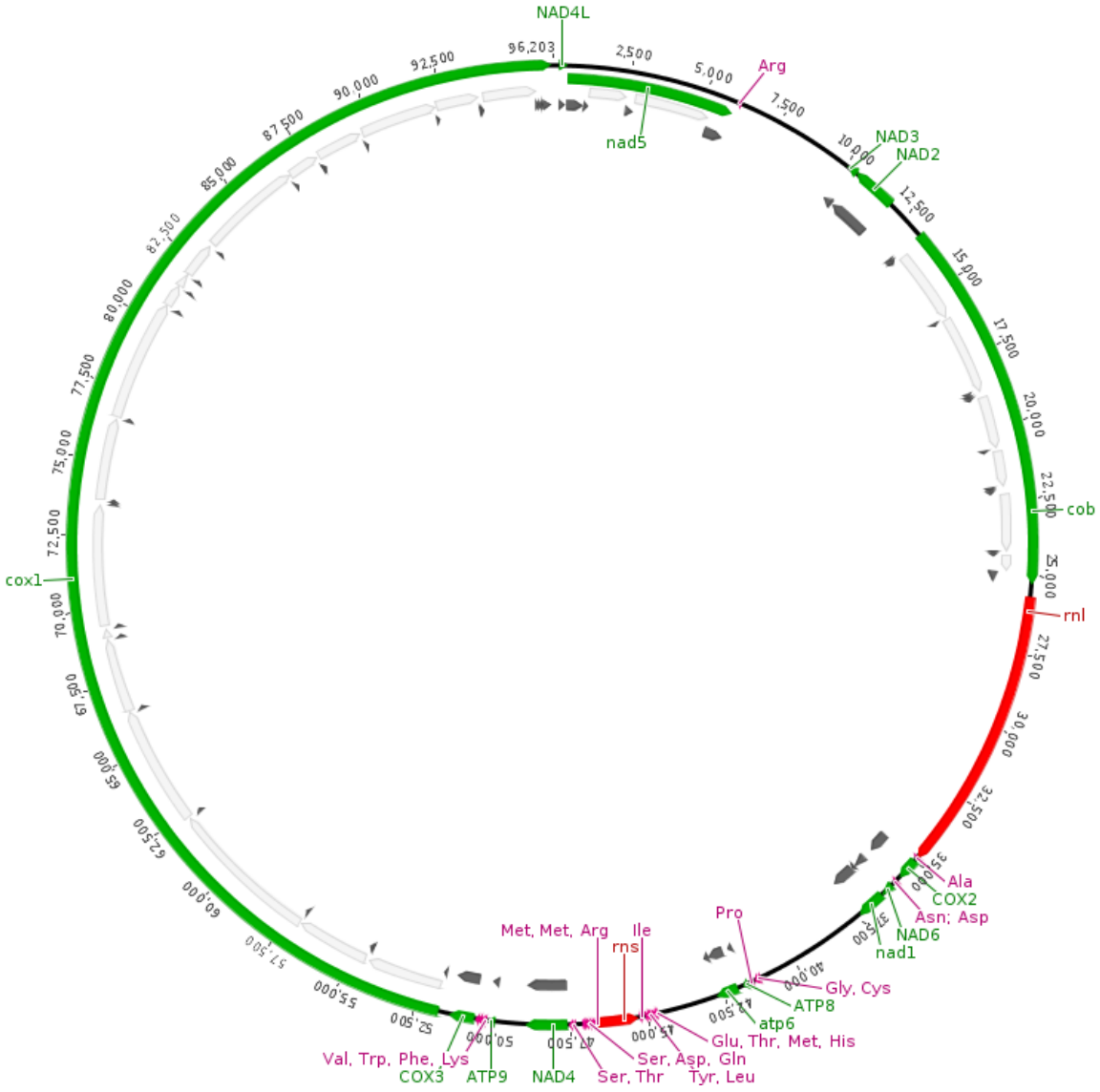

Figure S 9: Mitochondrial genome of *Metschnikowia borealis* SUB99-207.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

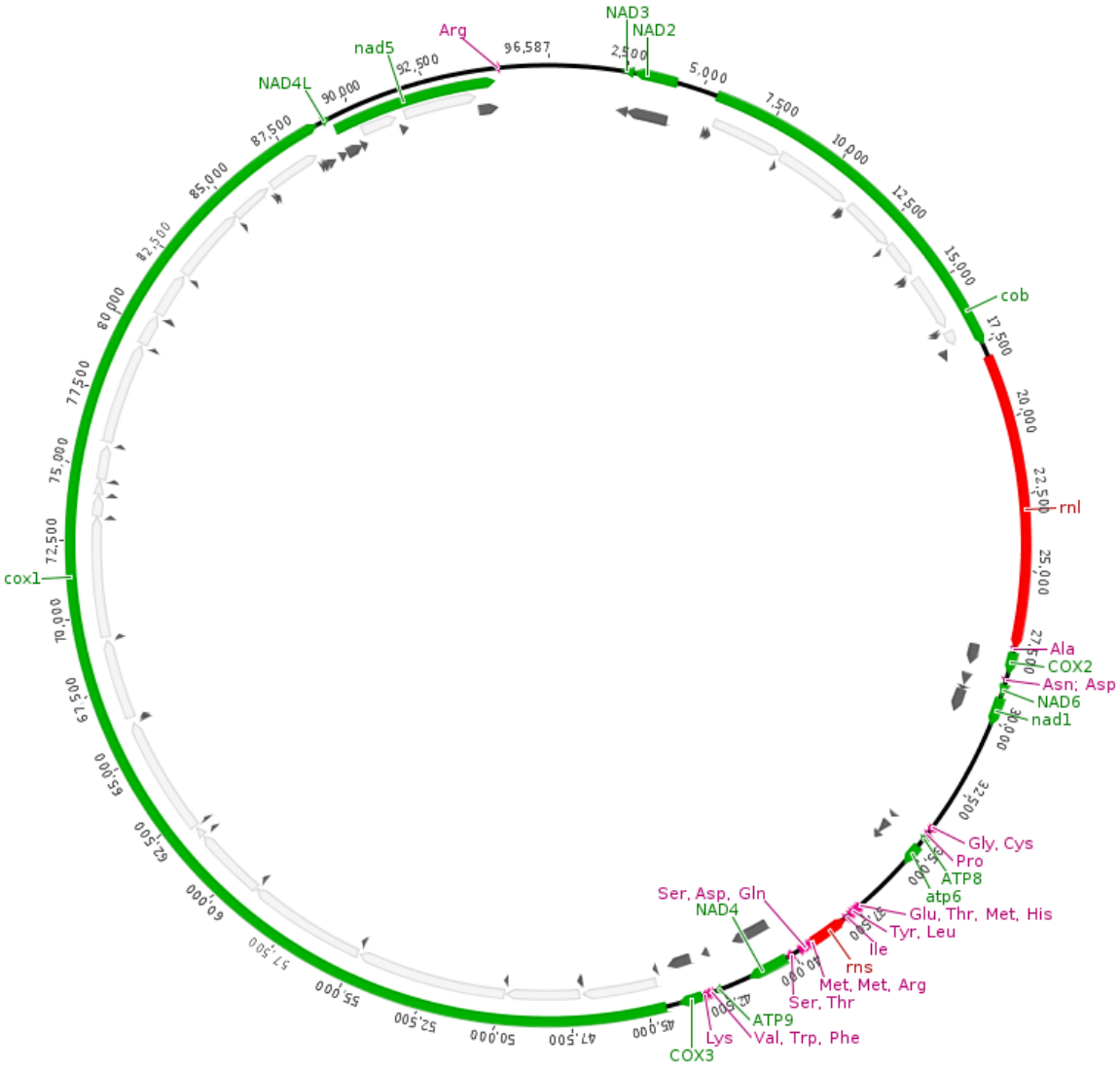

Figure S 10: Mitochondrial genome of *Metschnikowia borealis* UWOPS 96-101.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

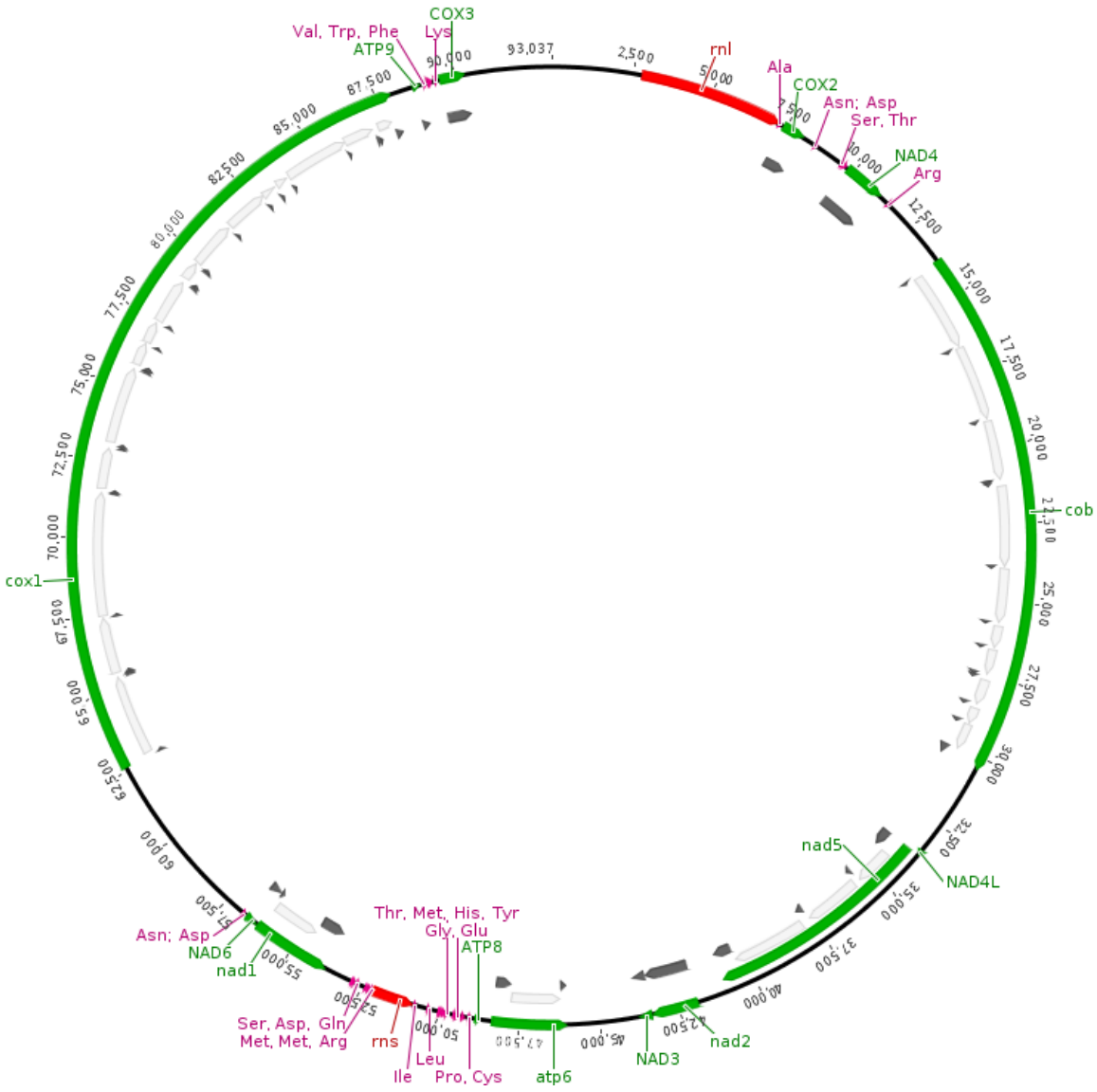

Figure S 11: Mitochondrial genome of *Metschnikowia bowlesiae* UWOPS 04-243x5. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

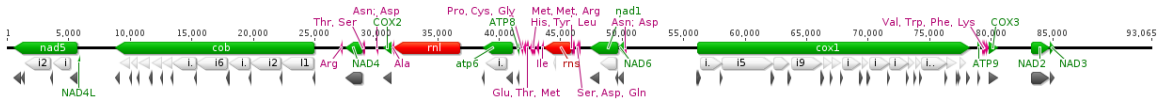

Figure S 12: Mitochondrial genome of *Metschnikowia bowlesiae* UWOPS 12-611.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

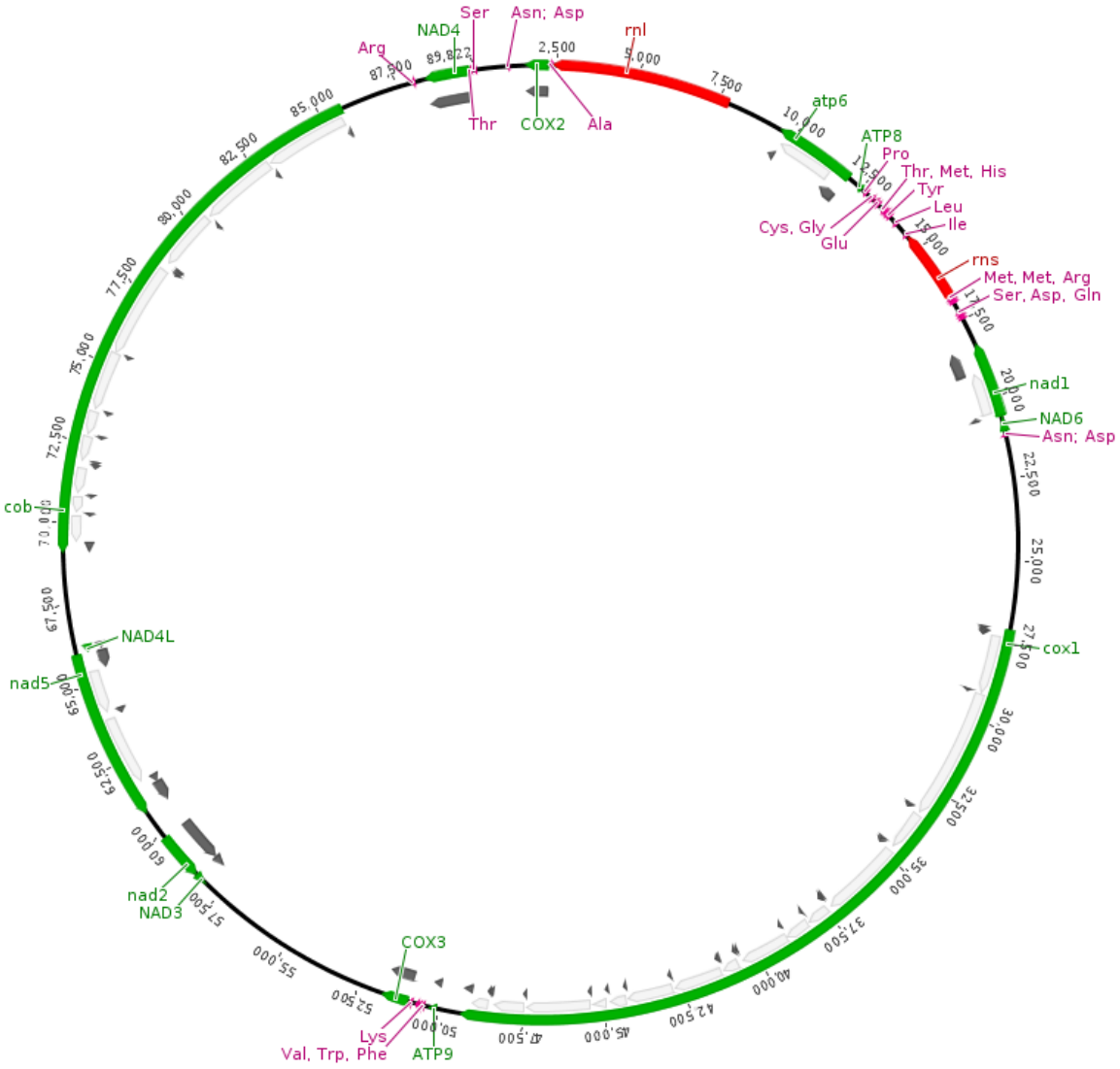

Figure S 13: Mitochondrial genome of *Metschnikowia bowlesiae* UWOPS 12-619.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

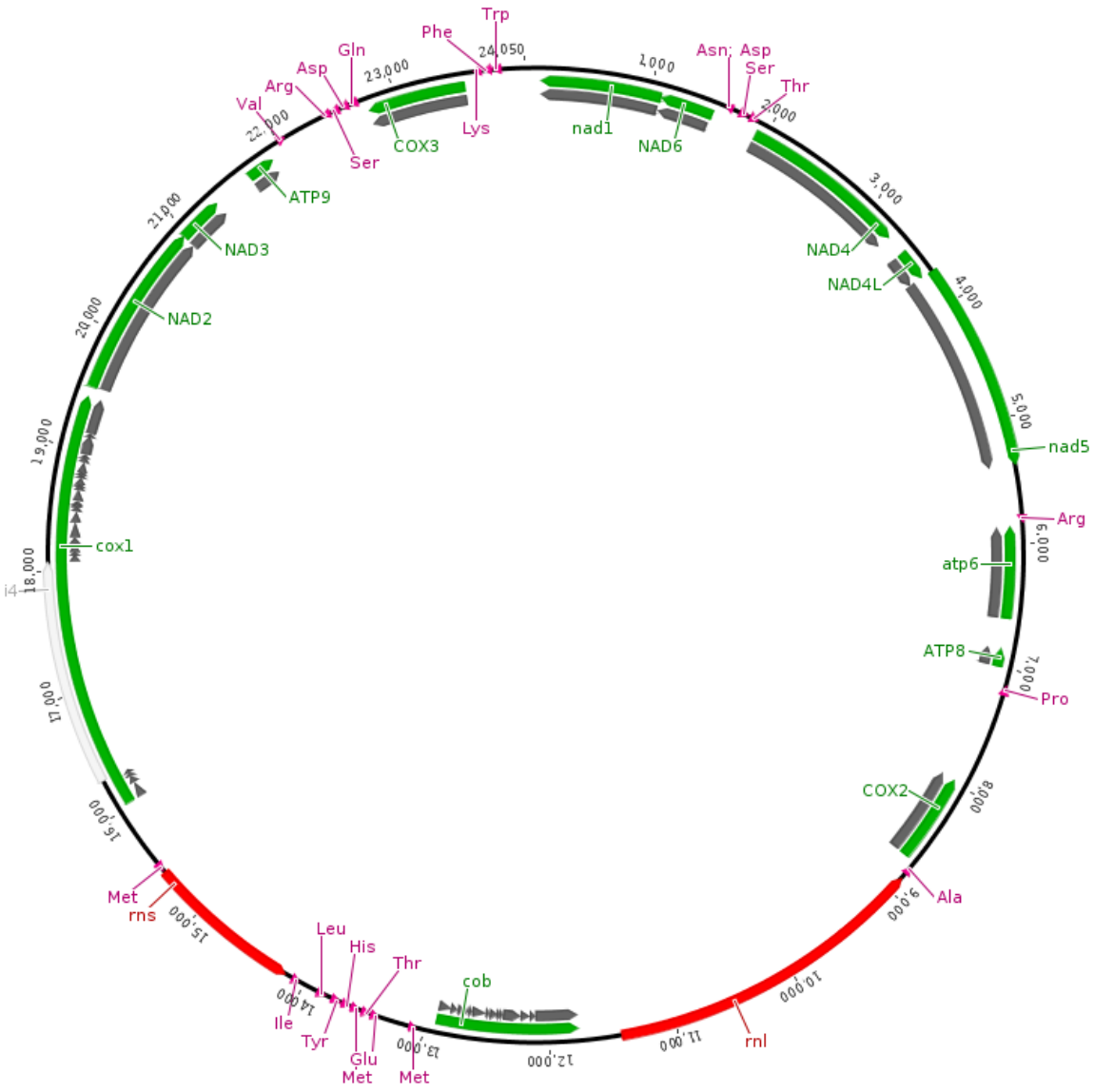

Figure S 14: Mitochondrial genome of *Metschnikowia caudata* EBDC CdV SA08.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

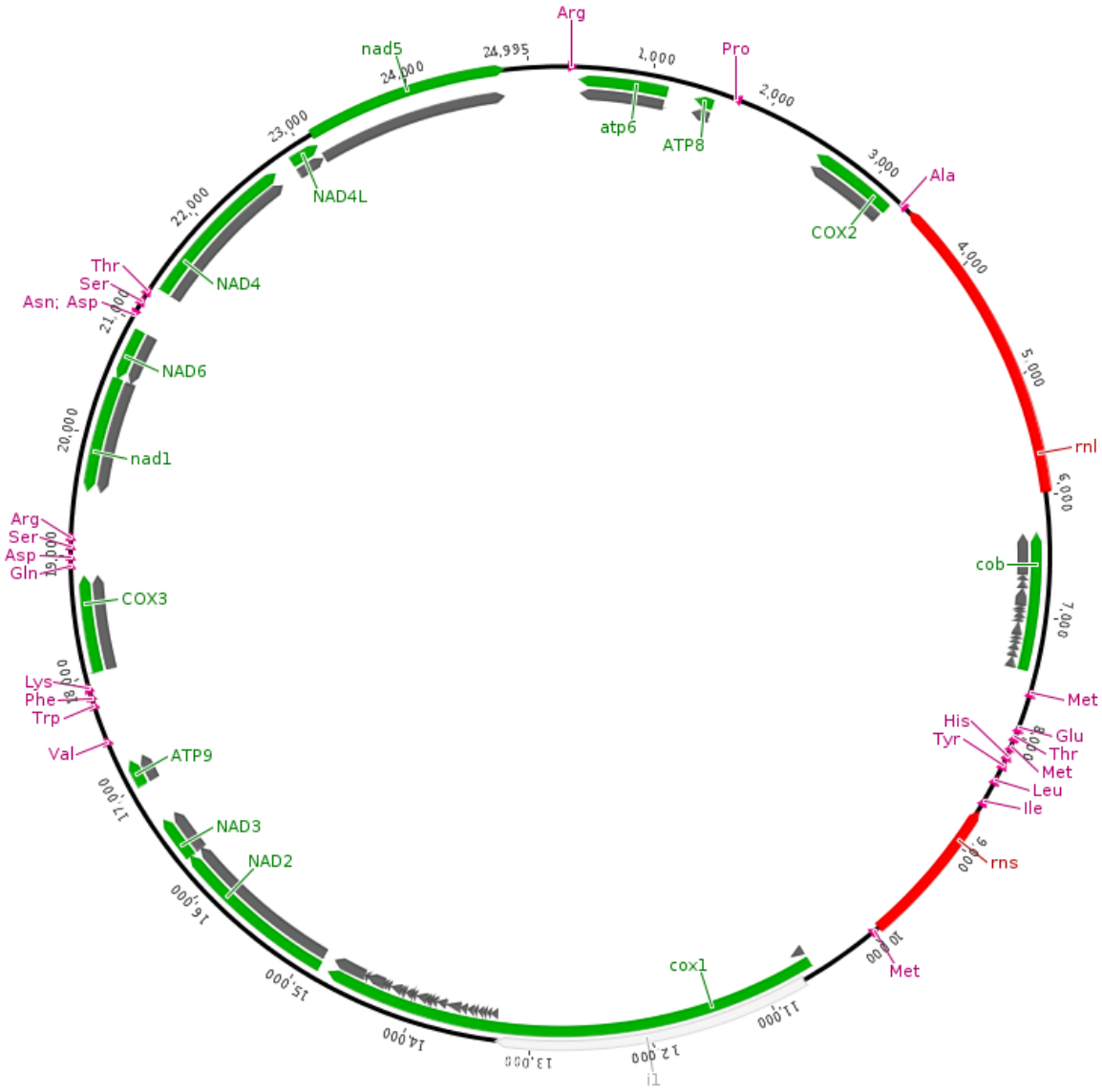

Figure S 15: Mitochondrial genome of *Metschnikowia caudata* EBDC CdV SA57.2. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

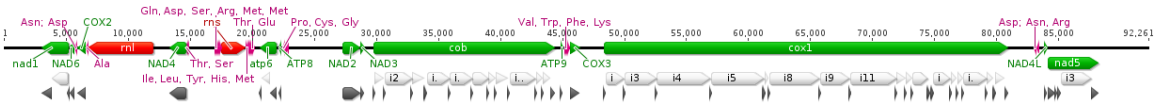

Figure S 16: Mitochondrial genome of *Metschnikowia cerradonensis* UFMG 03-T68.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

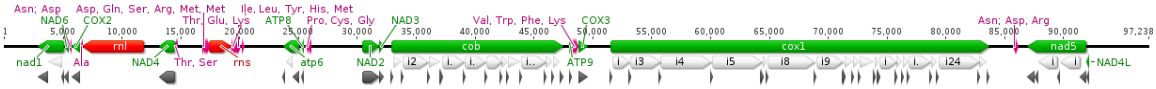

Figure S 17: Mitochondrial genome of *Metschnikowia cerradonensis* UFMG 03-T67.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

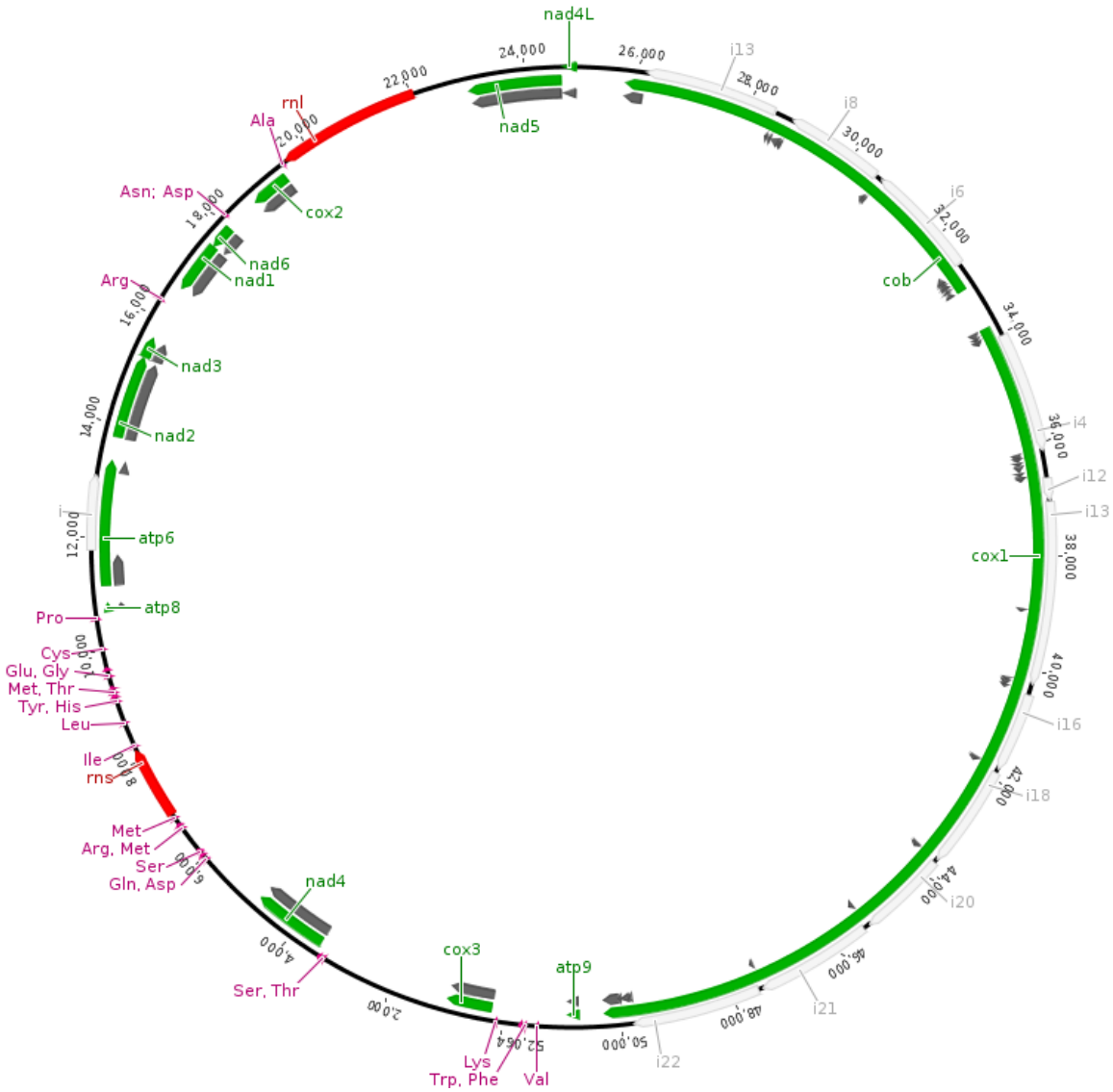

Figure S 18: Mitochondrial genome of *Metschnikowia* sp. M2Y3. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

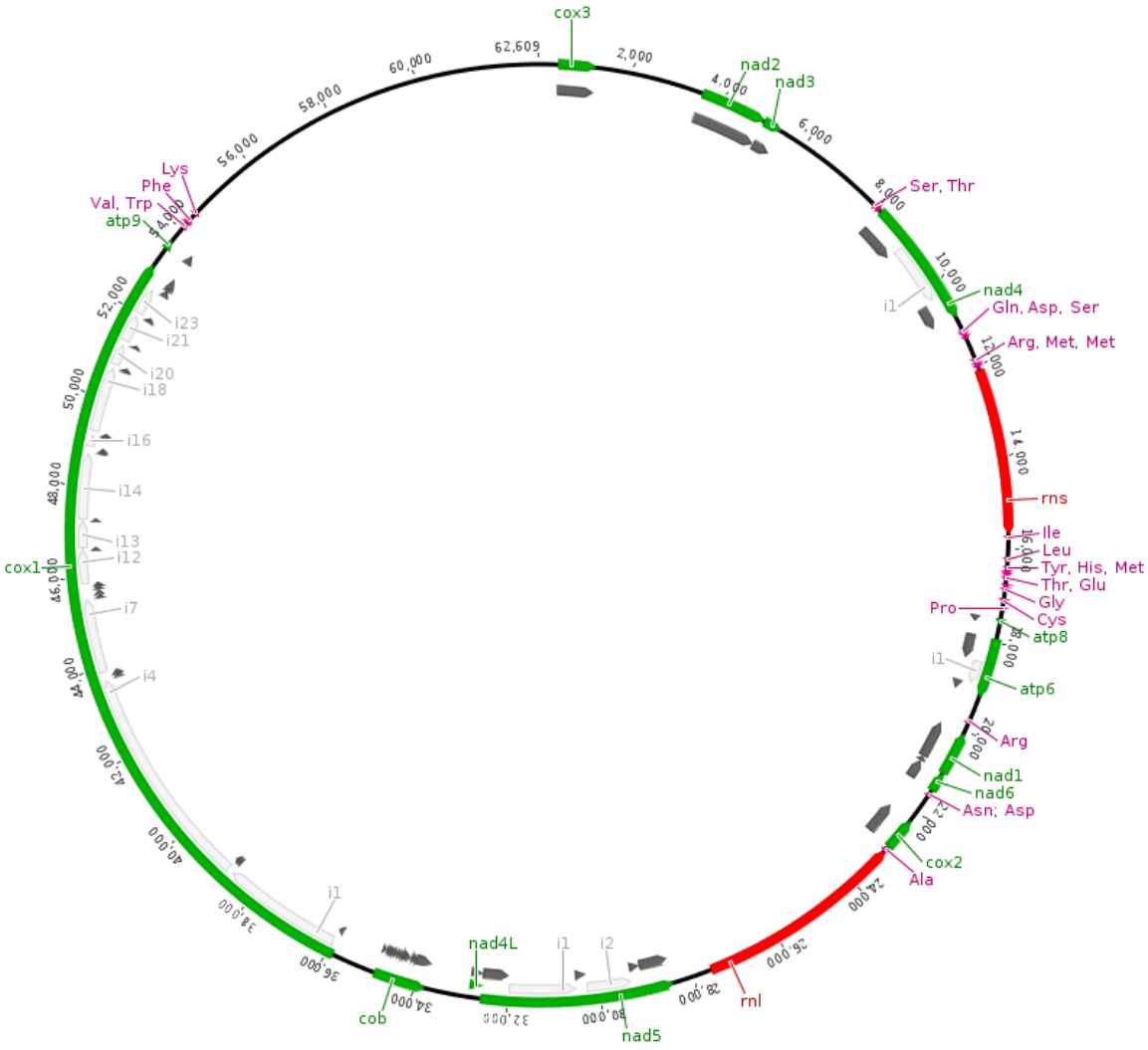

Figure S 19: Mitochondrial genome of *Metschnikowia colocasiae* UWOPS03-134.2. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

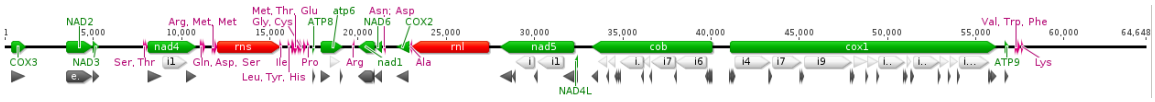

Figure S 20: Mitochondrial genome of *Metschnikowia colocasiae* UWOPS03-202.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

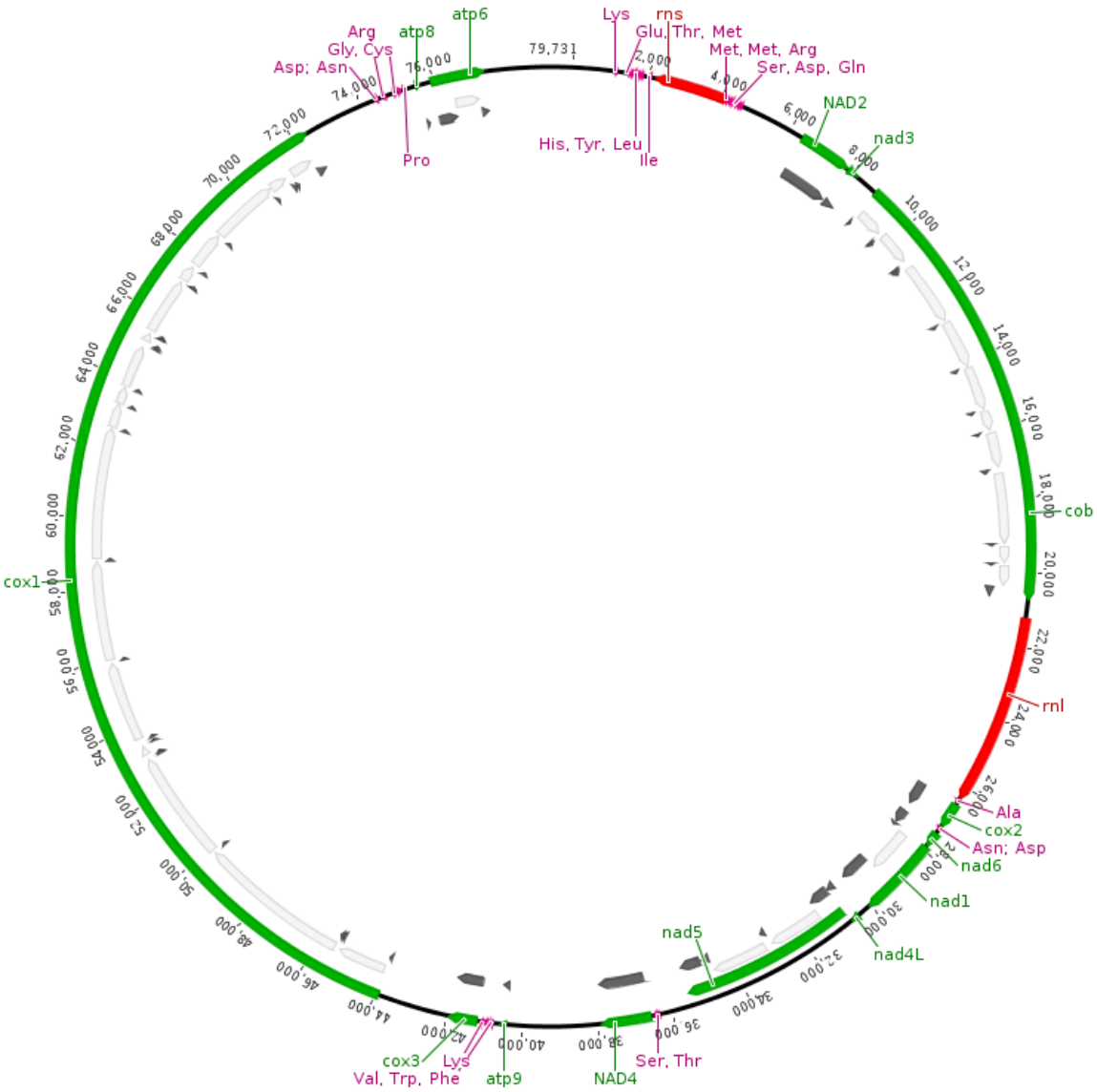

Figure S 21: Mitochondrial genome of *Metschnikowia continentalis* UWOPS96-173. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

Figure S 22: Mitochondrial genome of *Metschnikowia continentalis* UWOPS95-402.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

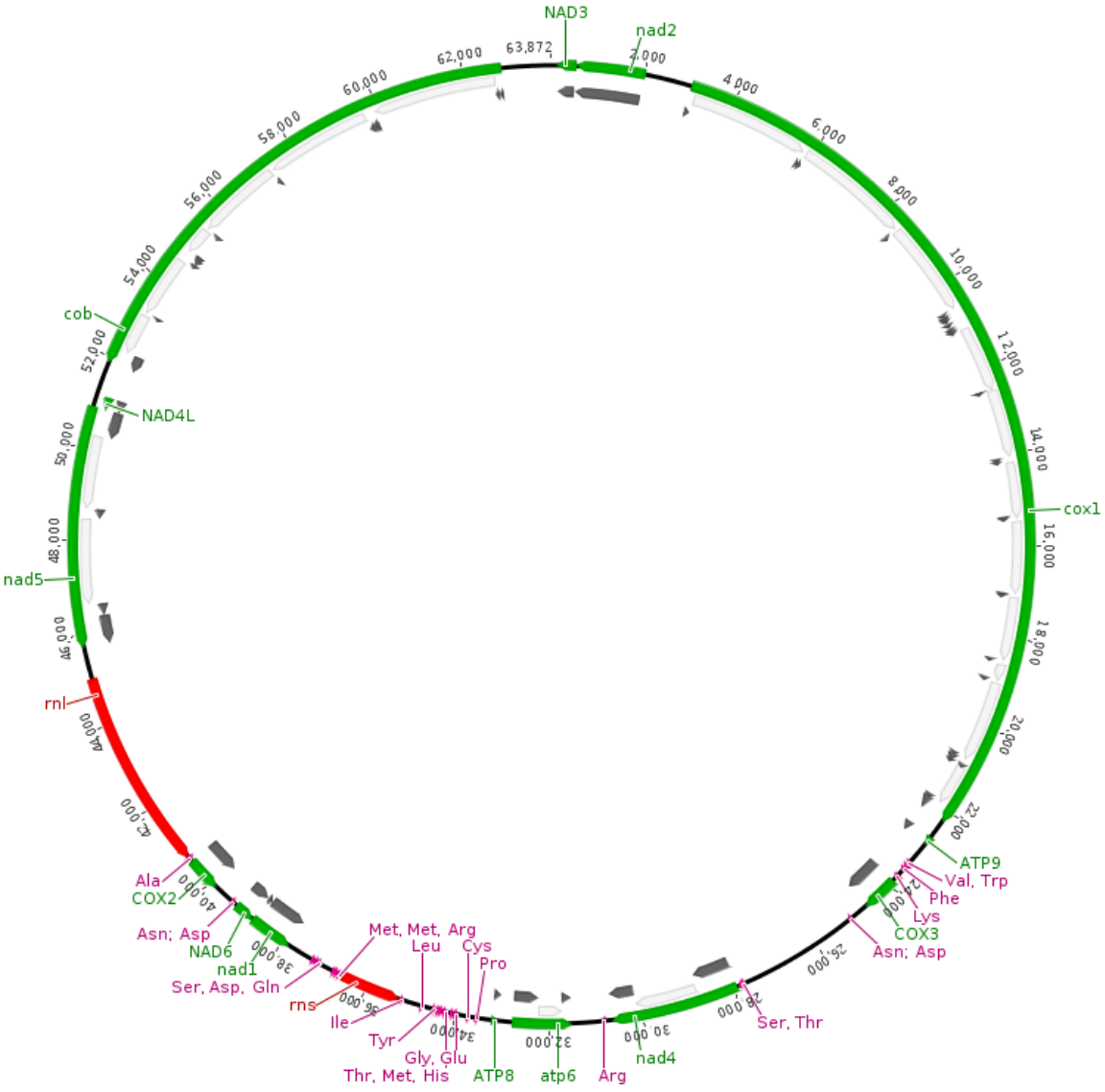

Figure S 23: Mitochondrial genome of *Metschnikowia* sp. UWOPS03-147.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

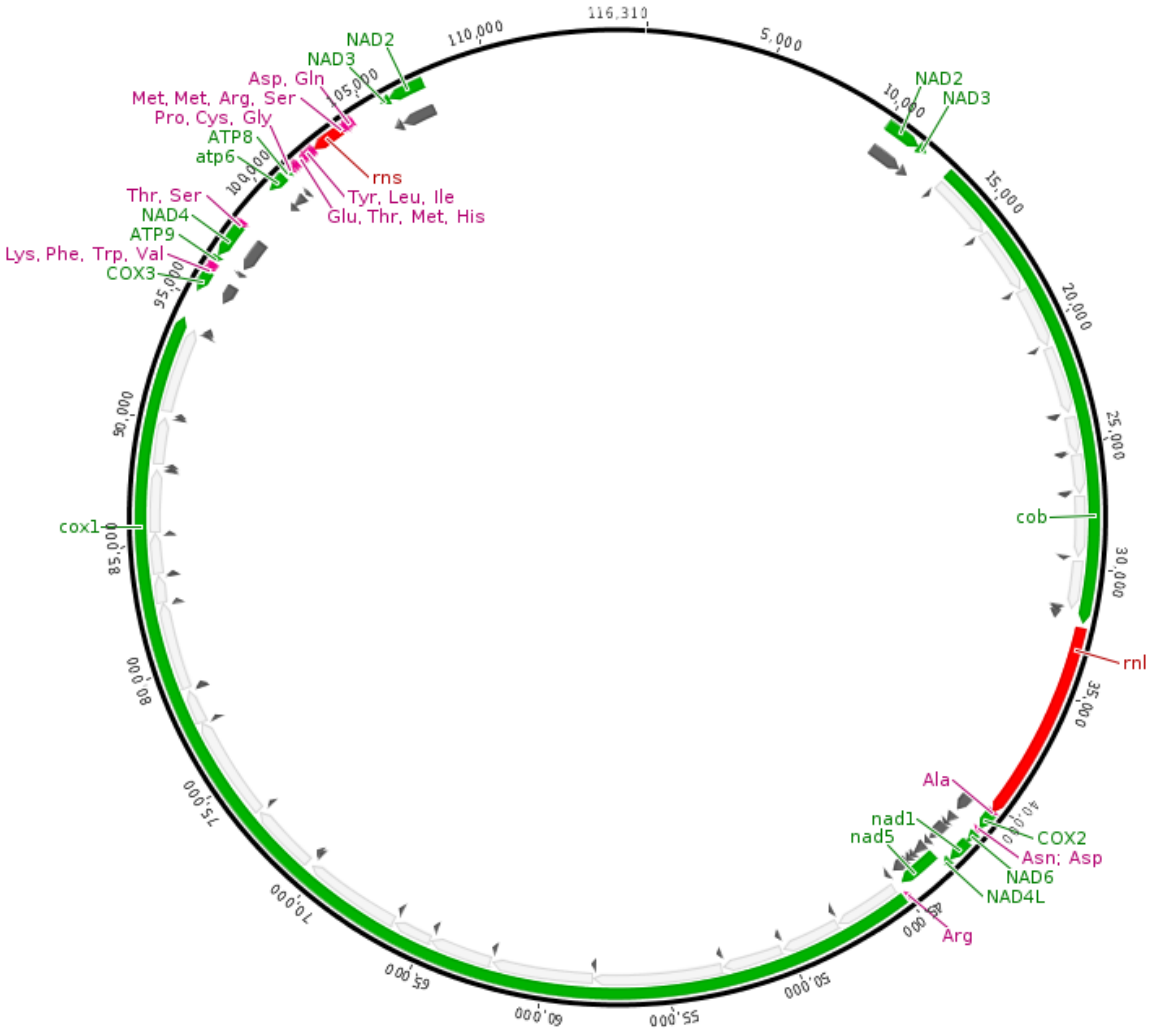

Figure S 24: Mitochondrial genome of *Metschnikowia cubensis* MUCL45753. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

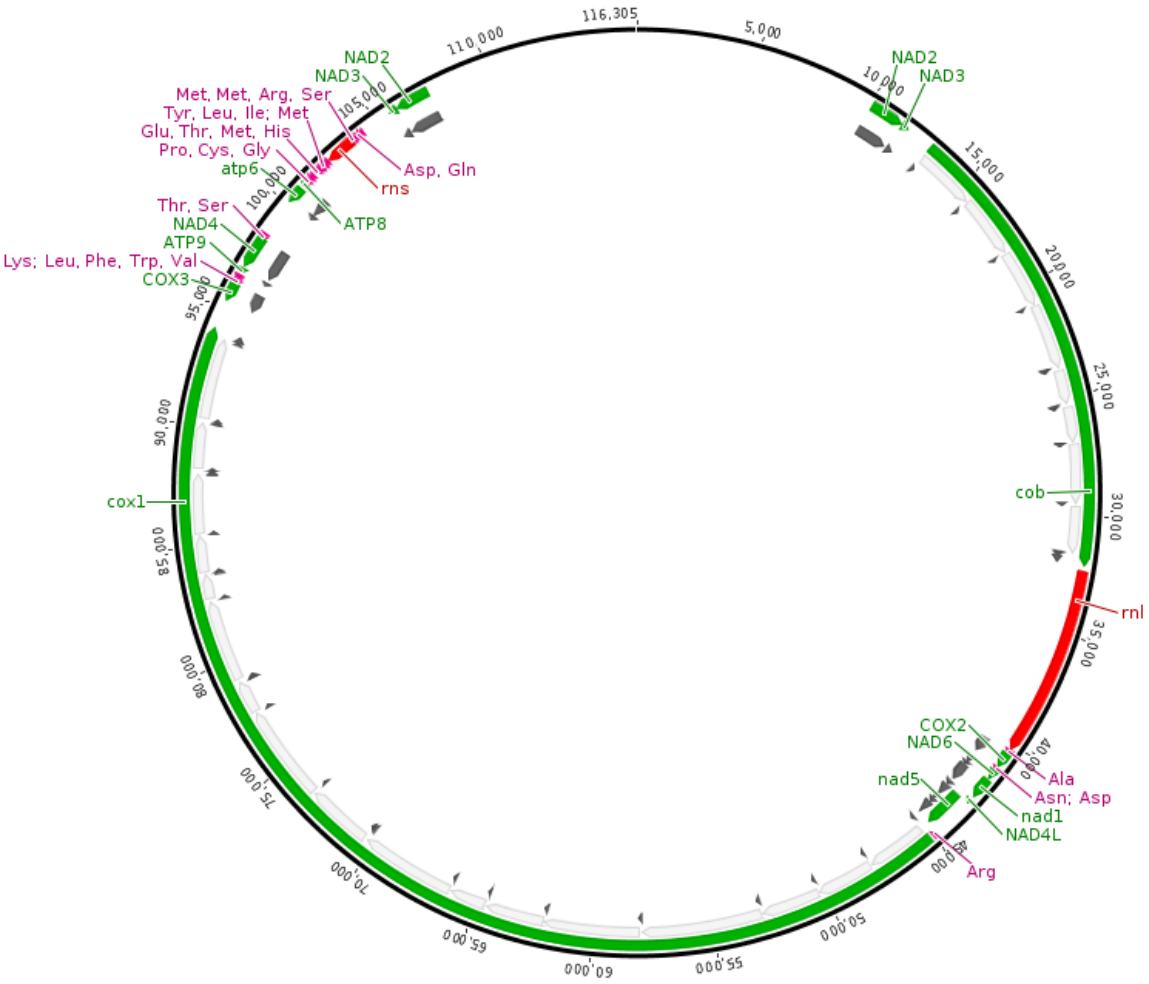

Figure S 25: Mitochondrial genome of *Metschnikowia cubensis* MUCL45751. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

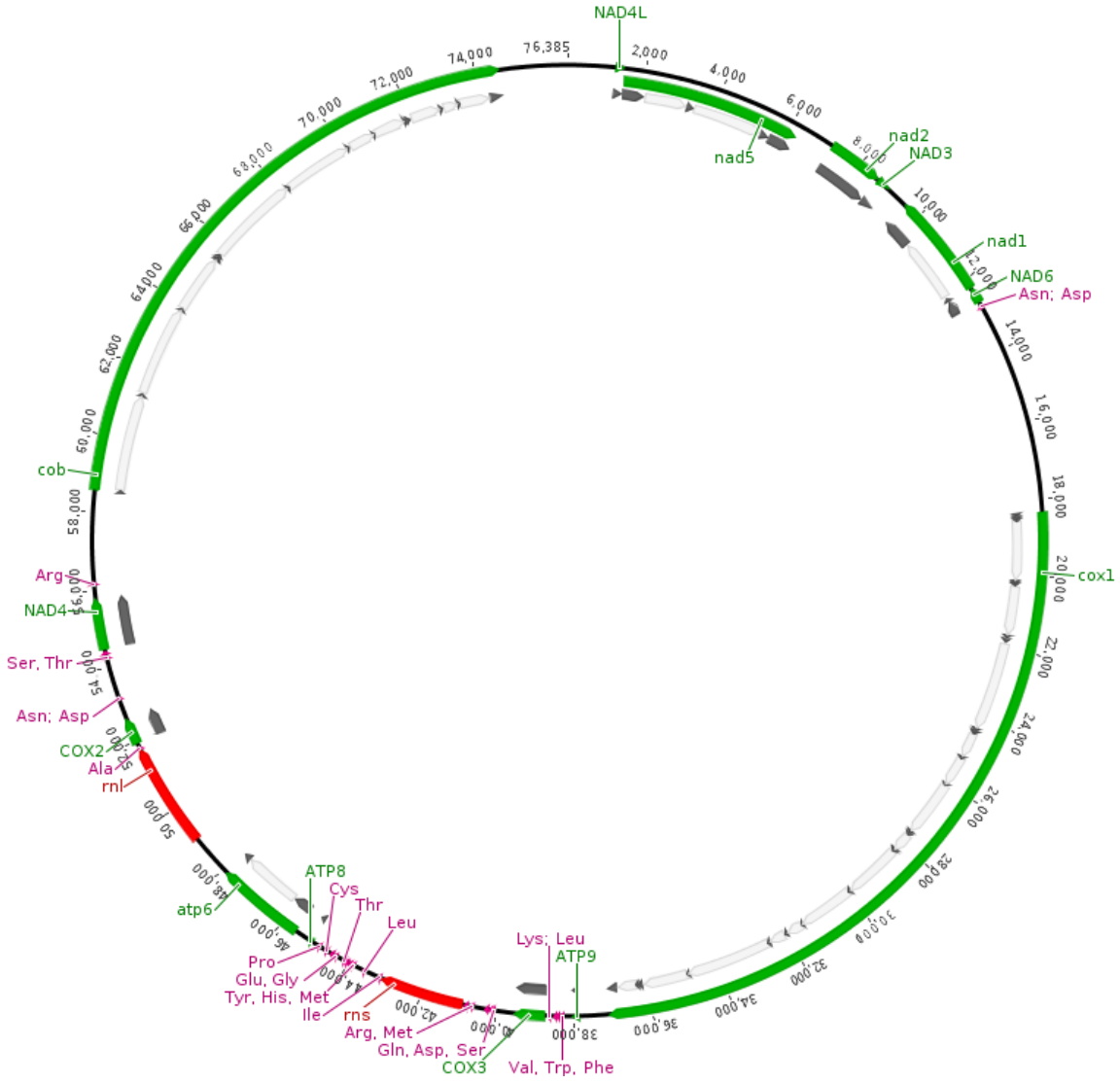

Figure S 26: Mitochondrial genome of *Metschnikowia dekortorum* UWOPS01-142b3. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

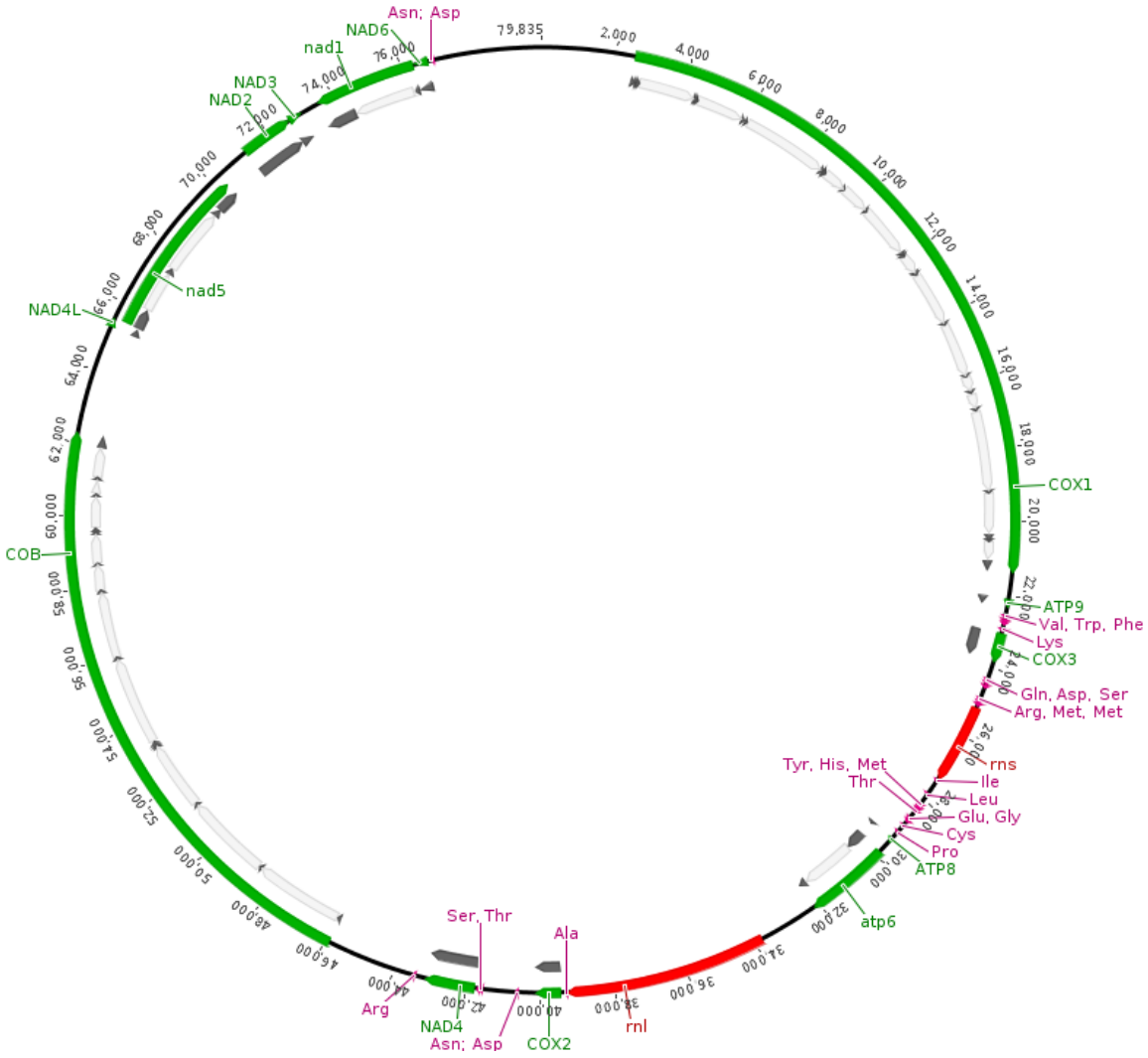

Figure S 27: Mitochondrial genome of *Metschnikowia dekortorum* UWOPS01-522a5. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

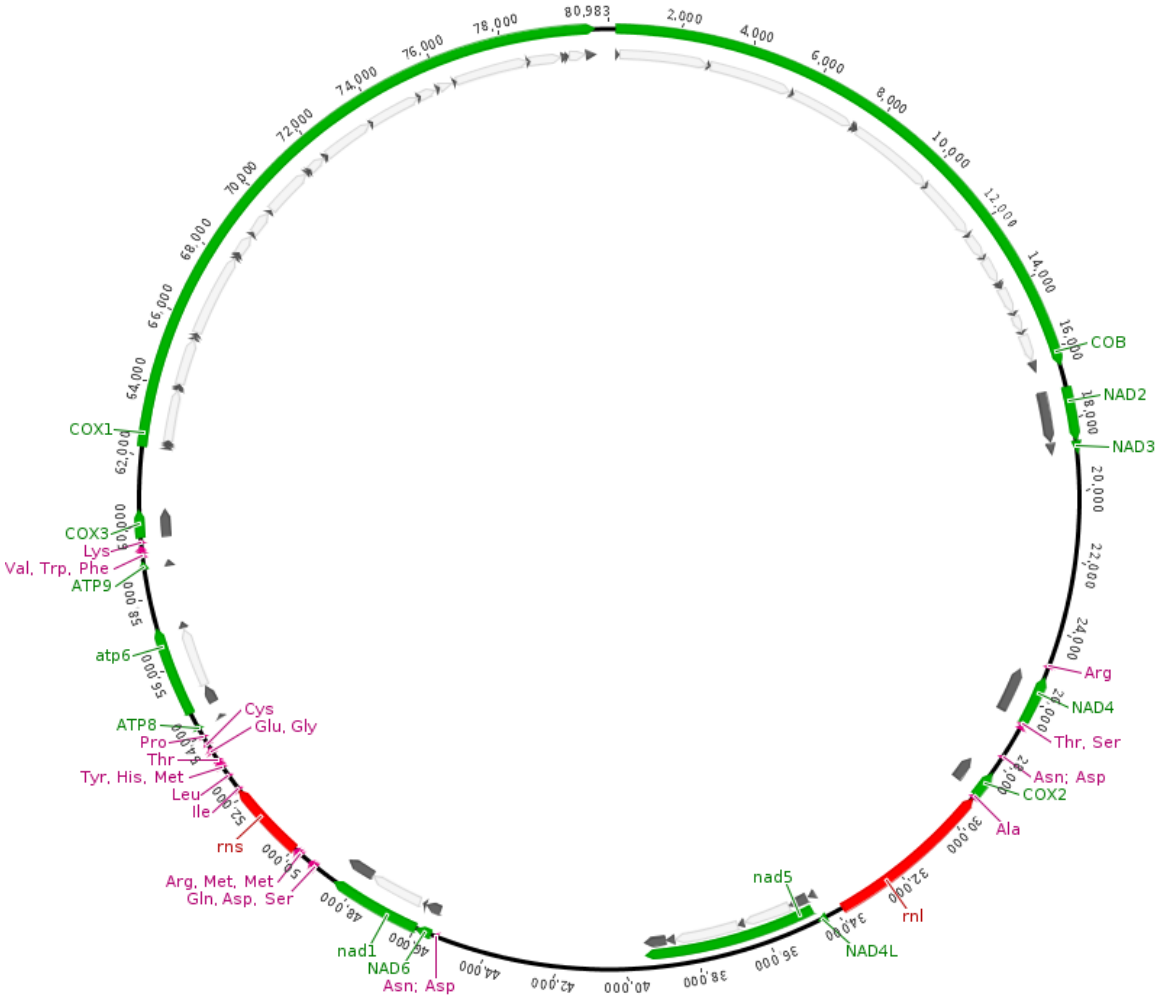

Figure S 28: Mitochondrial genome of *Metschnikowia dekortorum* UFMG-CM-Y6306. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

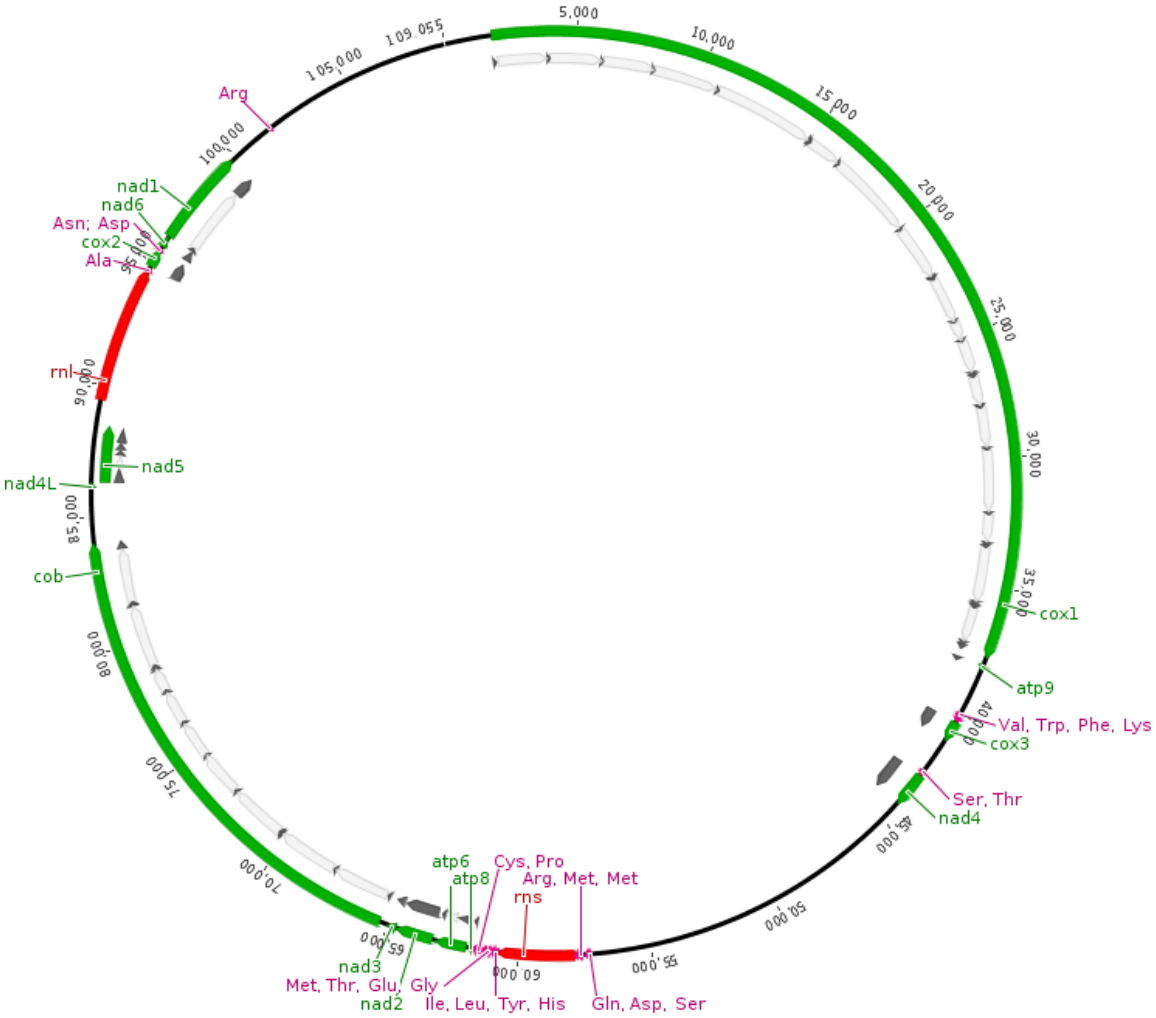

Figure S 29: Mitochondrial genome of *Metschnikowia drakensbergensis* EBD-CdVSA09-2. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

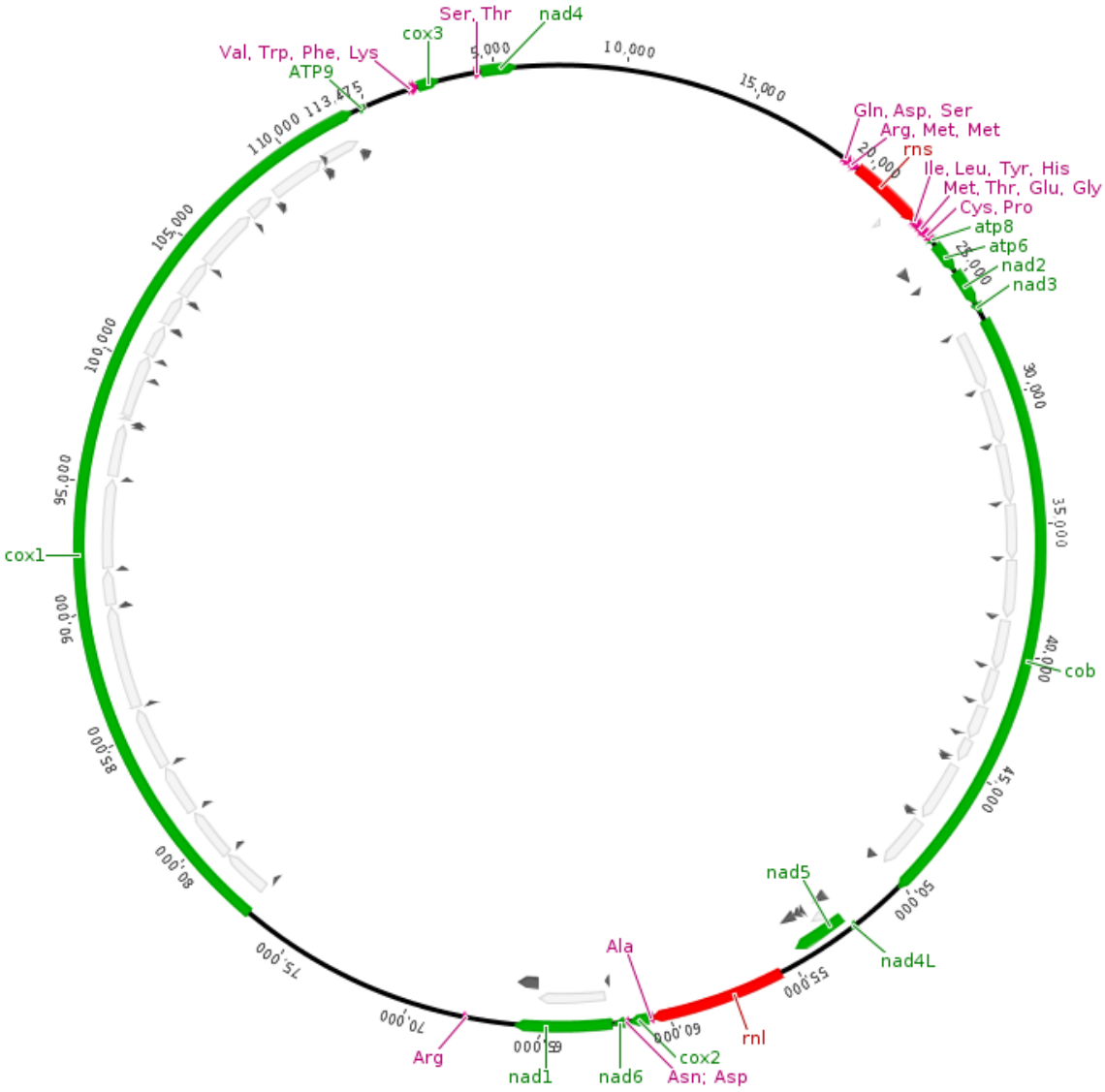

Figure S 30: Mitochondrial genome of *Metschnikowia drakensbergensis* EBD-CdVSA10-2. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

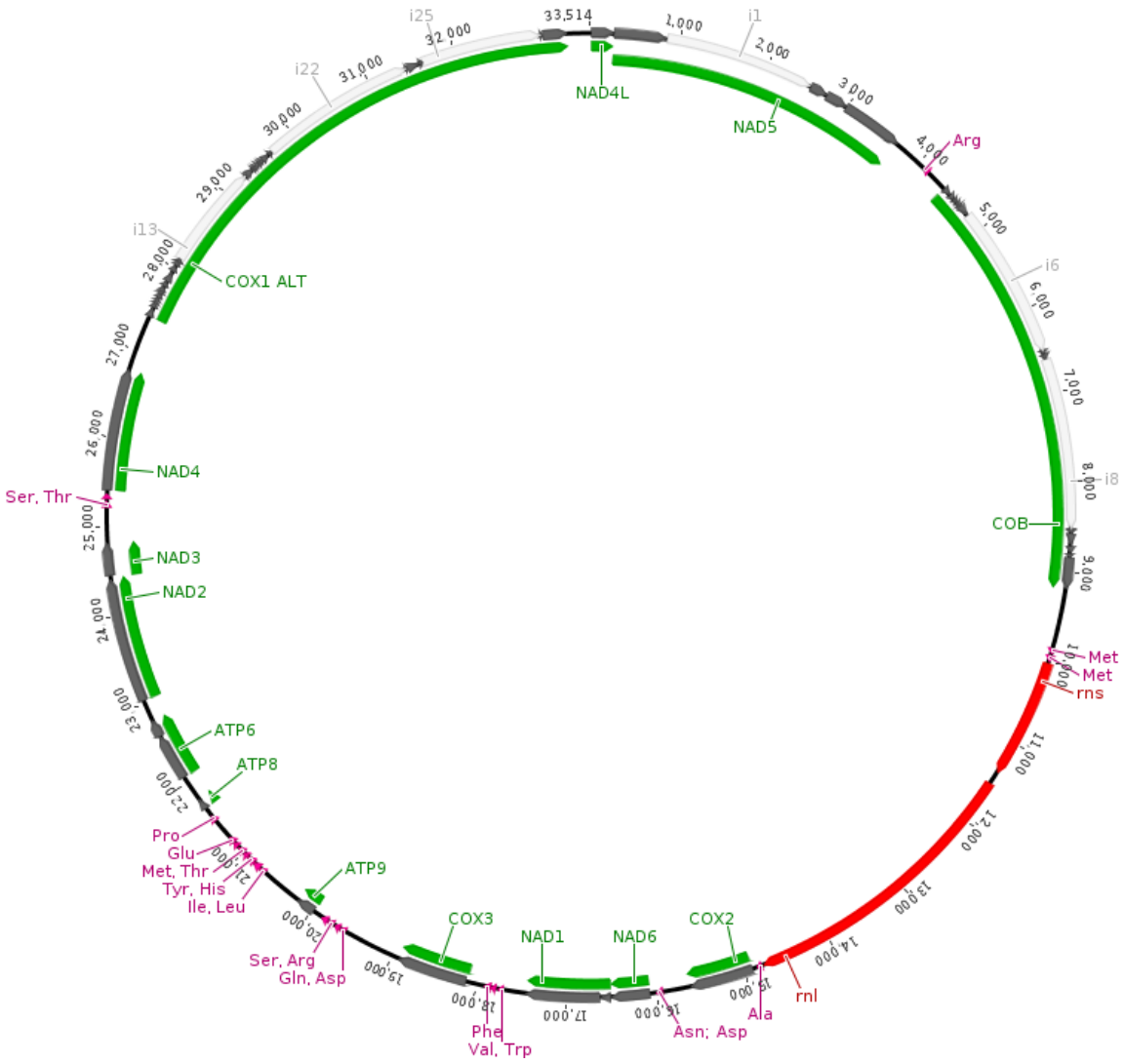

Figure S 31: Mitochondrial genome of *Metschnikowia drosophilae* UWOPS83-1143.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

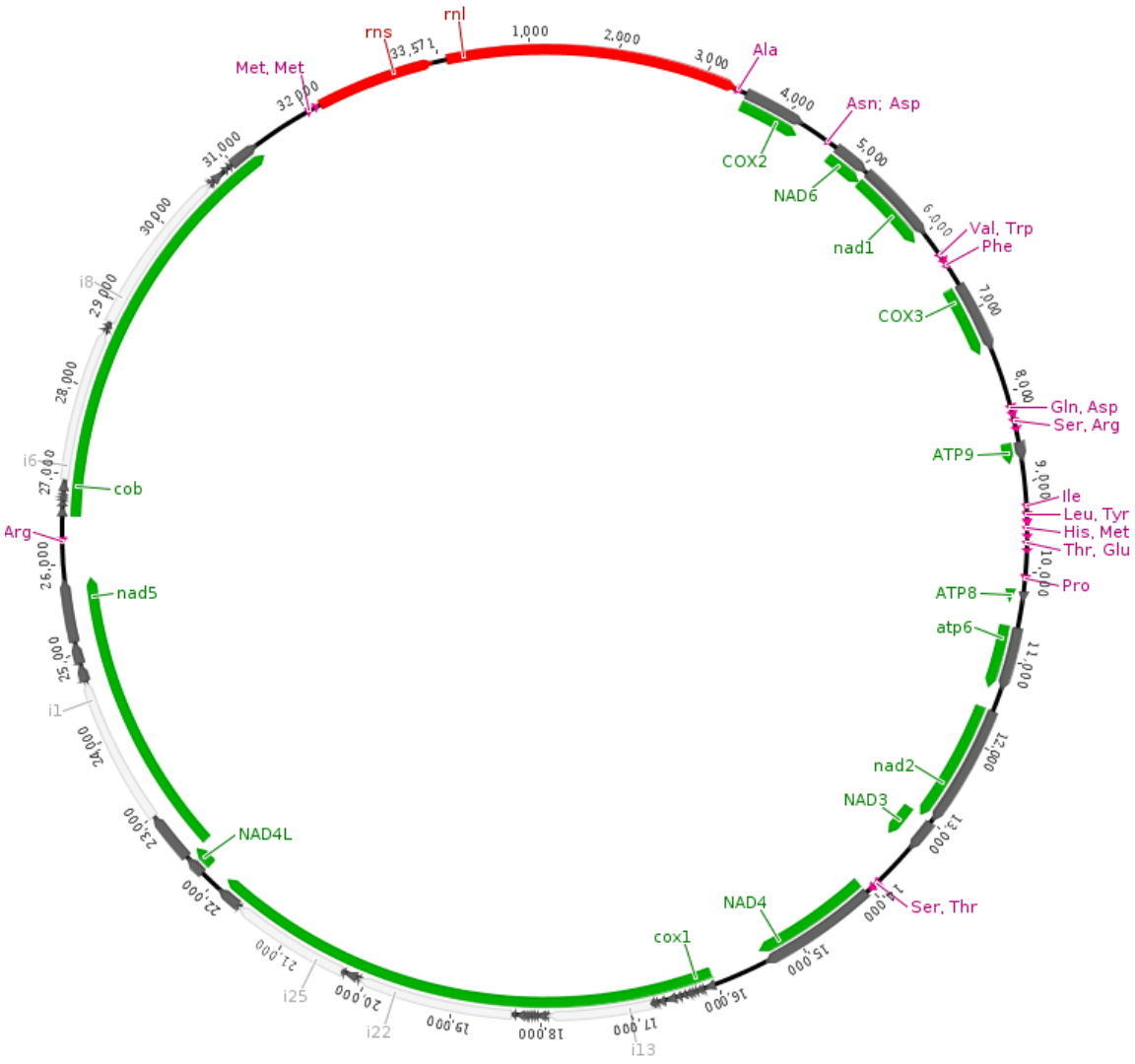

Figure S 32: Mitochondrial genome of *Metschnikowia drosophilae* UWOPS83-1135.3. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

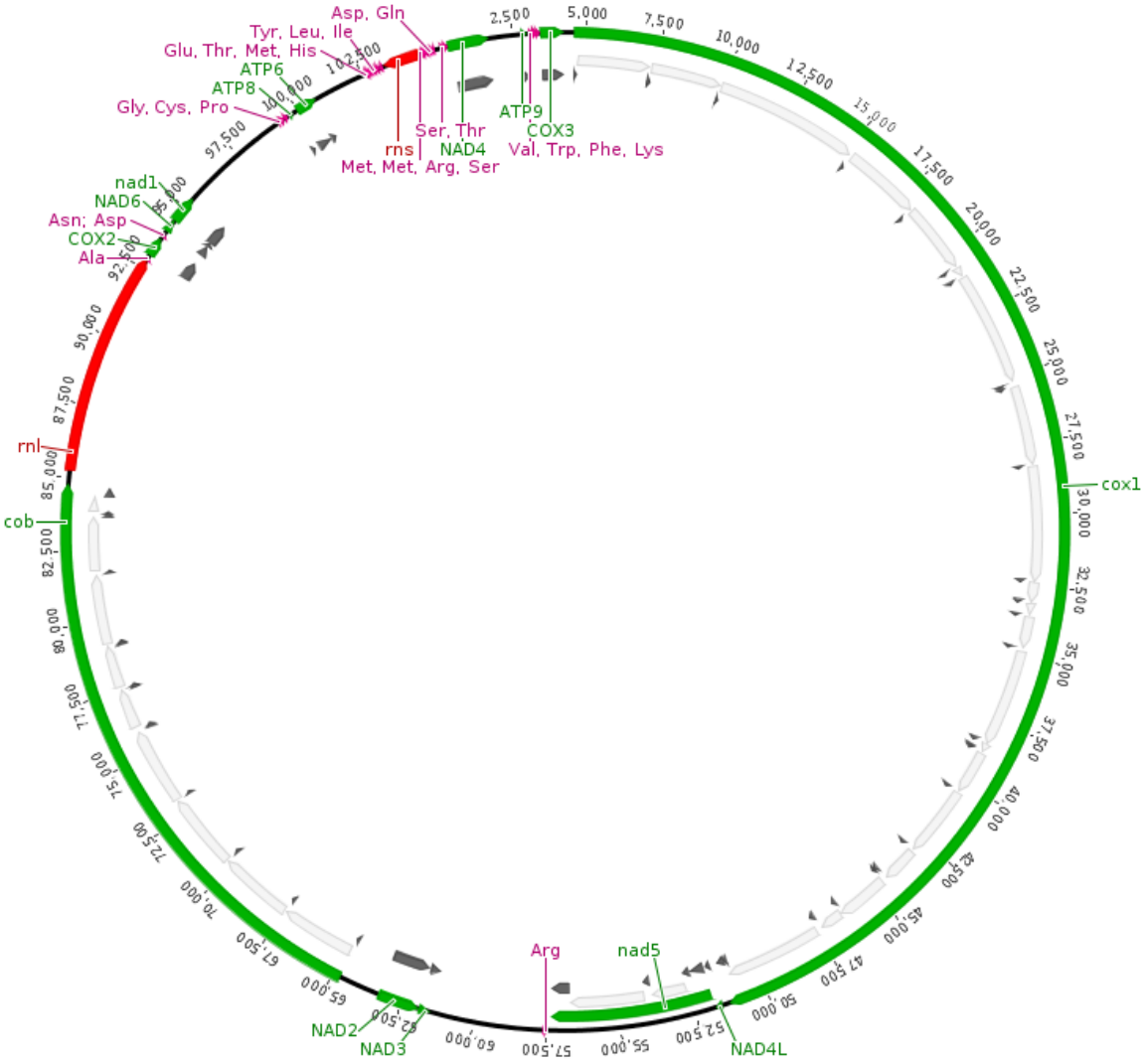

Figure S 33: Mitochondrial genome of *Metschnikowia* sp. UWOPS13-106.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

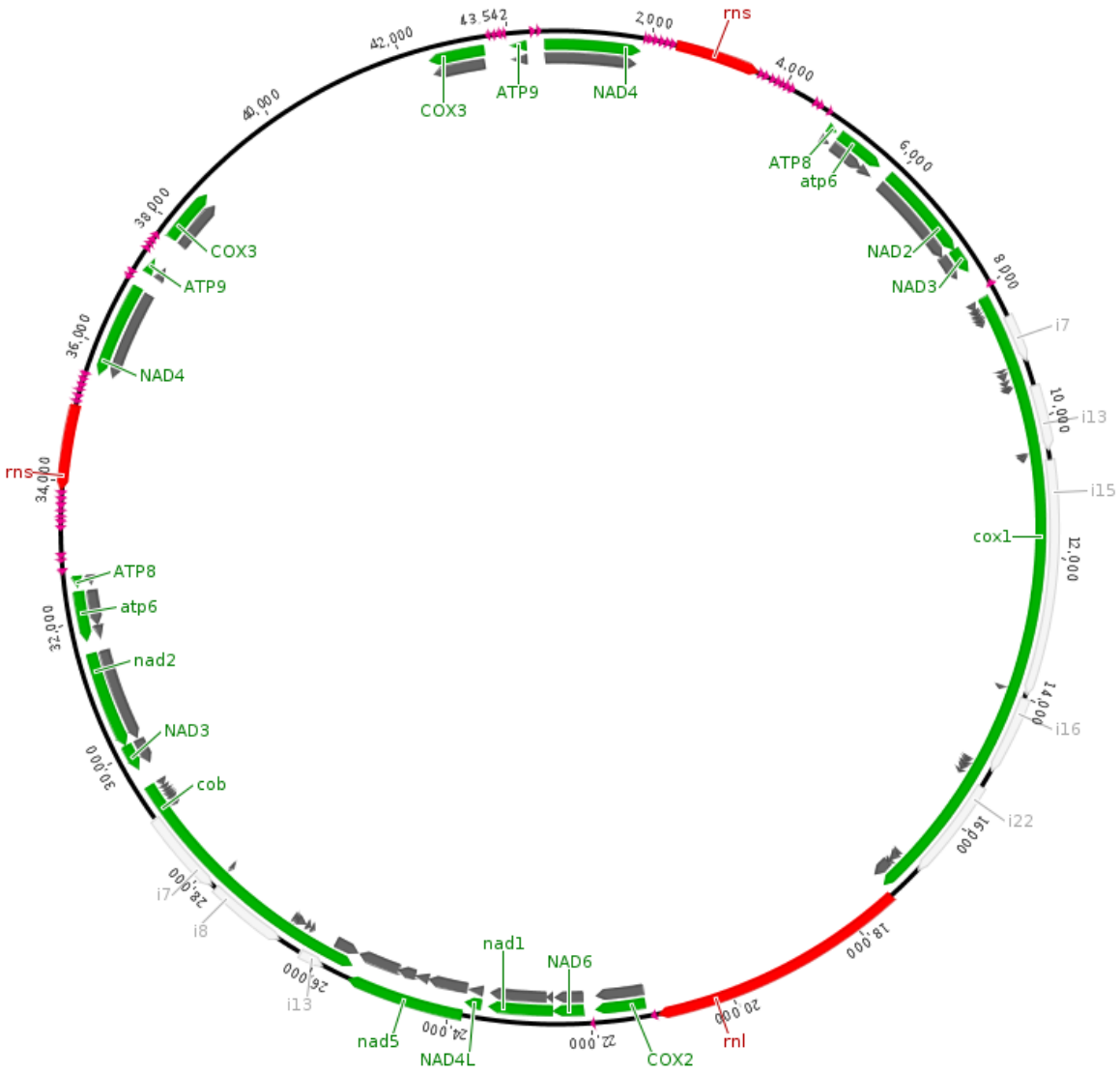

Figure S 34: Mitochondrial genome of *Metschnikowia hamakuensis* UWOPS04-207.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

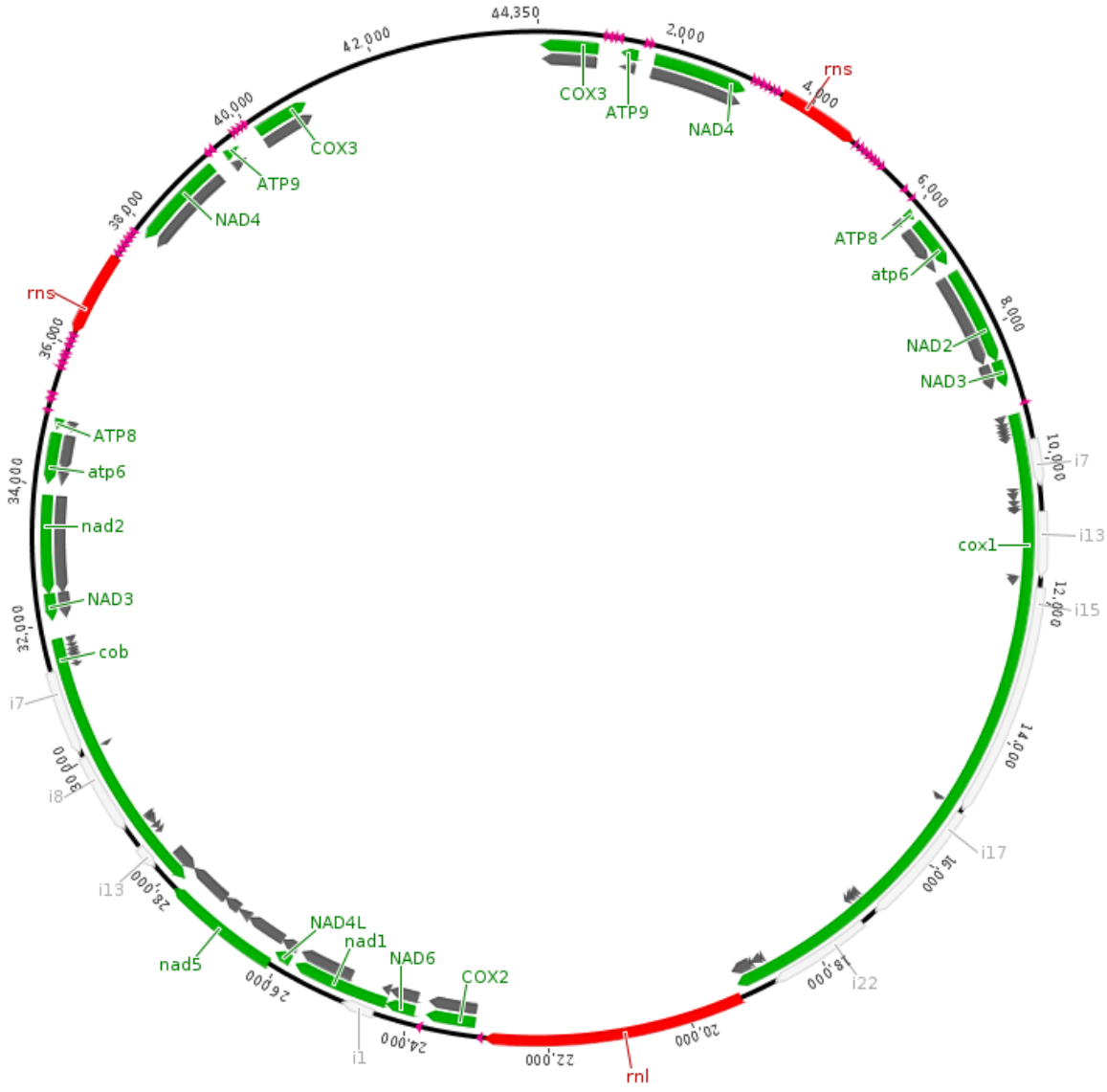

Figure S 35: Mitochondrial genome of *Metschnikowia hamakuensis* UWOPS04-199.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

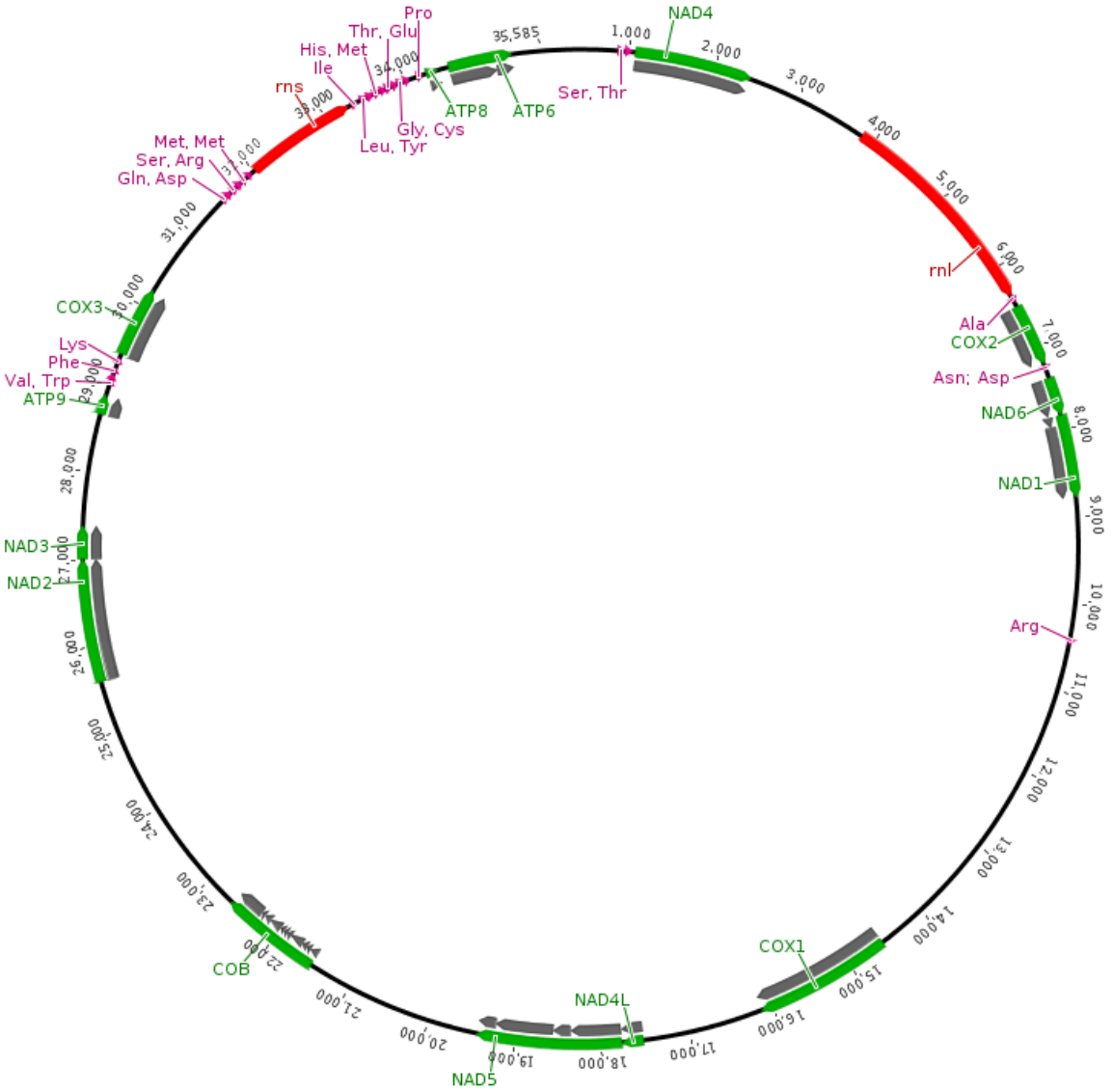

Figure S 36: Mitochondrial genome of *Metschnikowia hawaiiana* UWOPS91-698.3. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

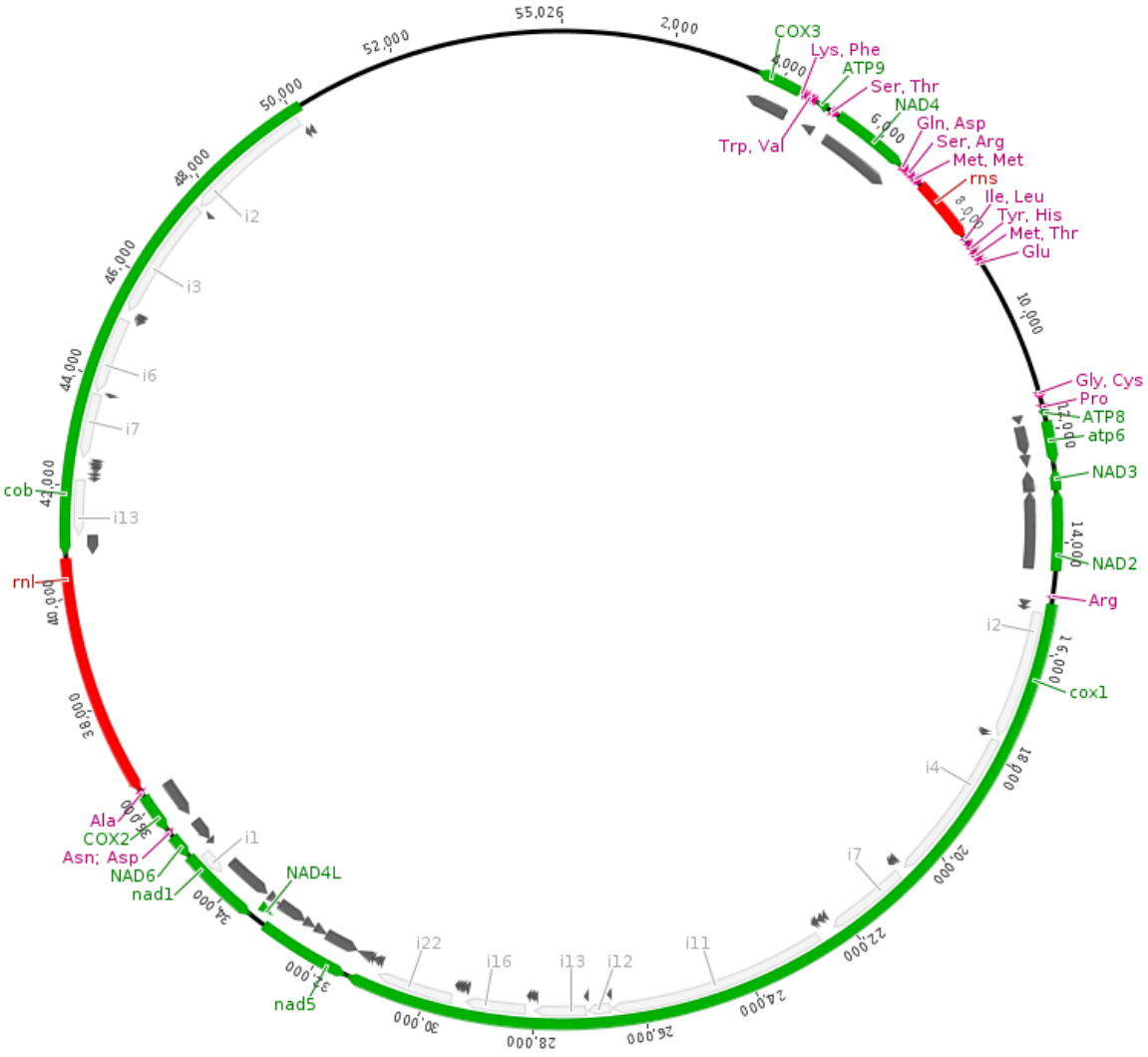

Figure S 37: Mitochondrial genome of *Metschnikowia hawaiiensis* UWOPS91-745.2. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

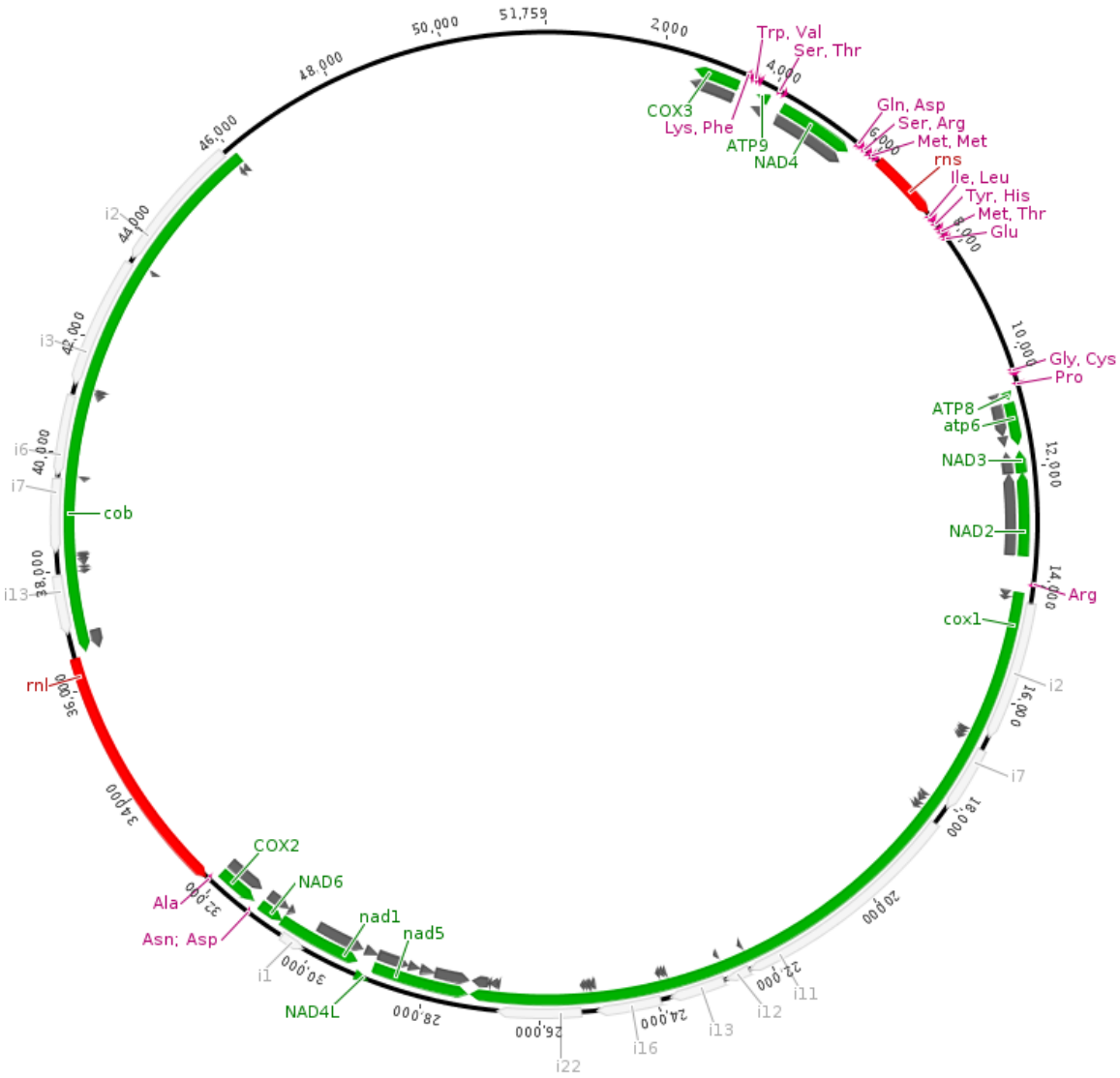

Figure S 38: Mitochondrial genome of *Metschnikowia hawaiiensis* UWOPS87-2203.2. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

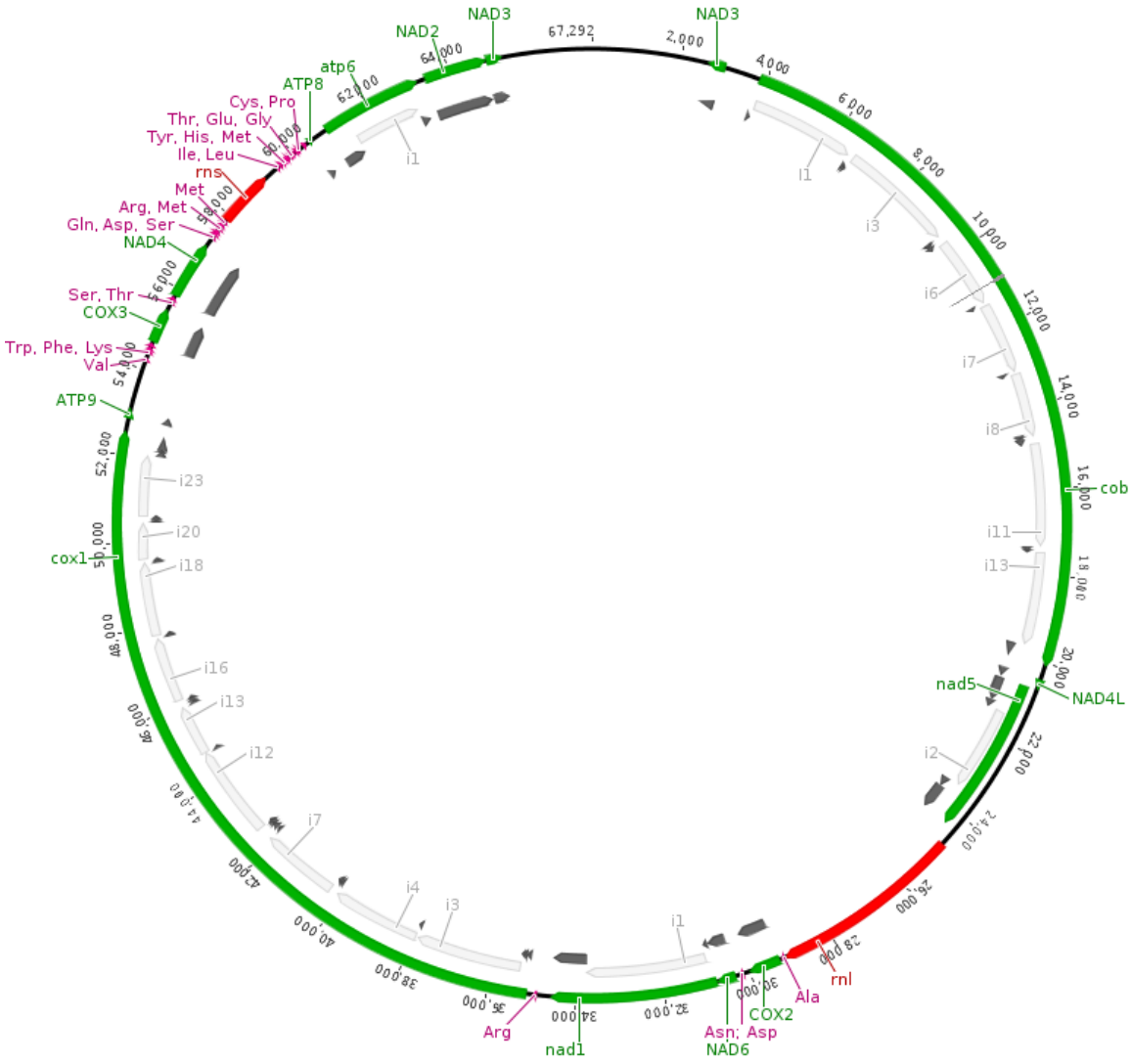

Figure S 39: Mitochondrial genome of *Metschnikowia hibisci* UWOPS95-797.2. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

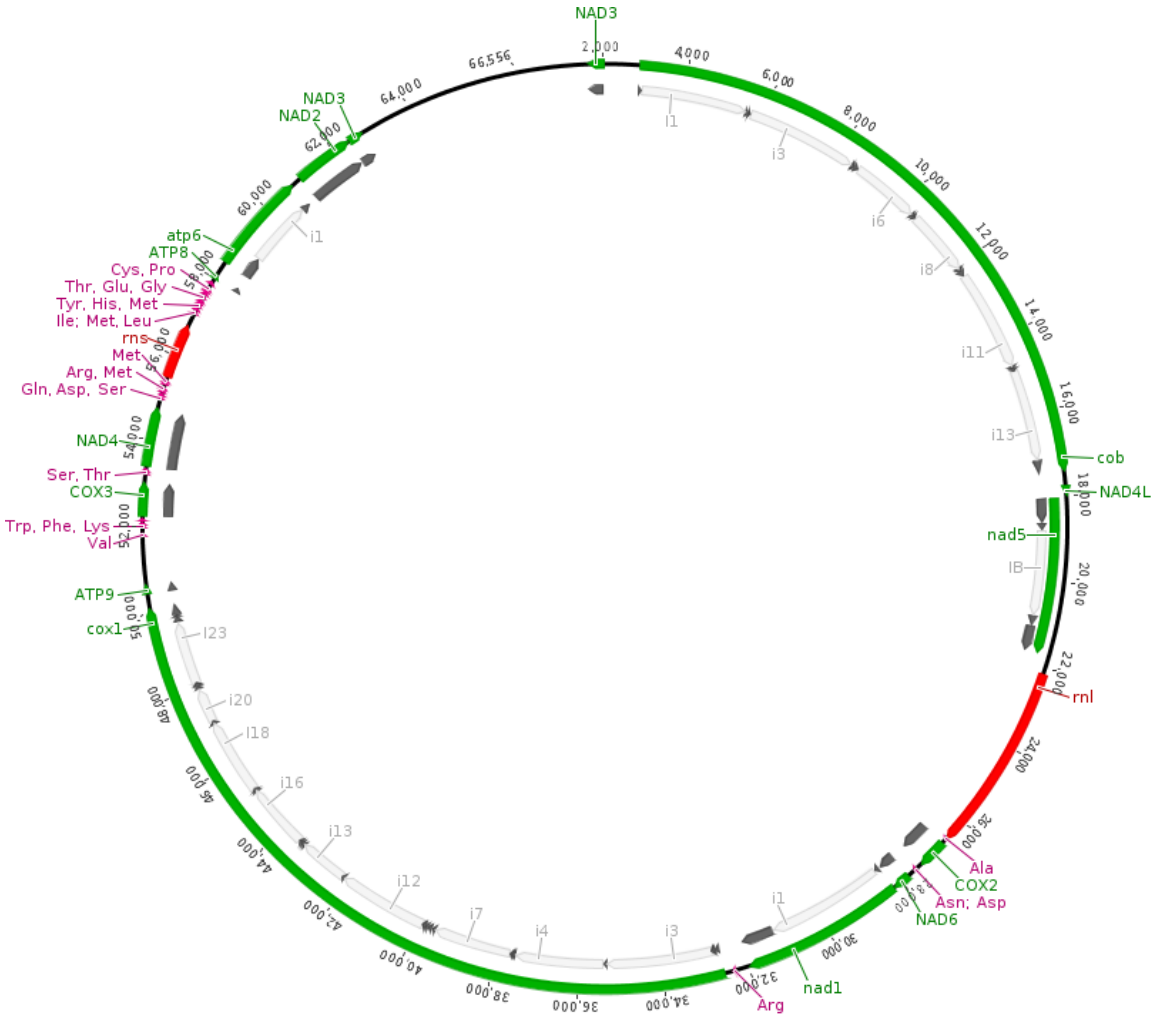

Figure S 40: Mitochondrial genome of *Metschnikowia hibisci* UWOPS95-805.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

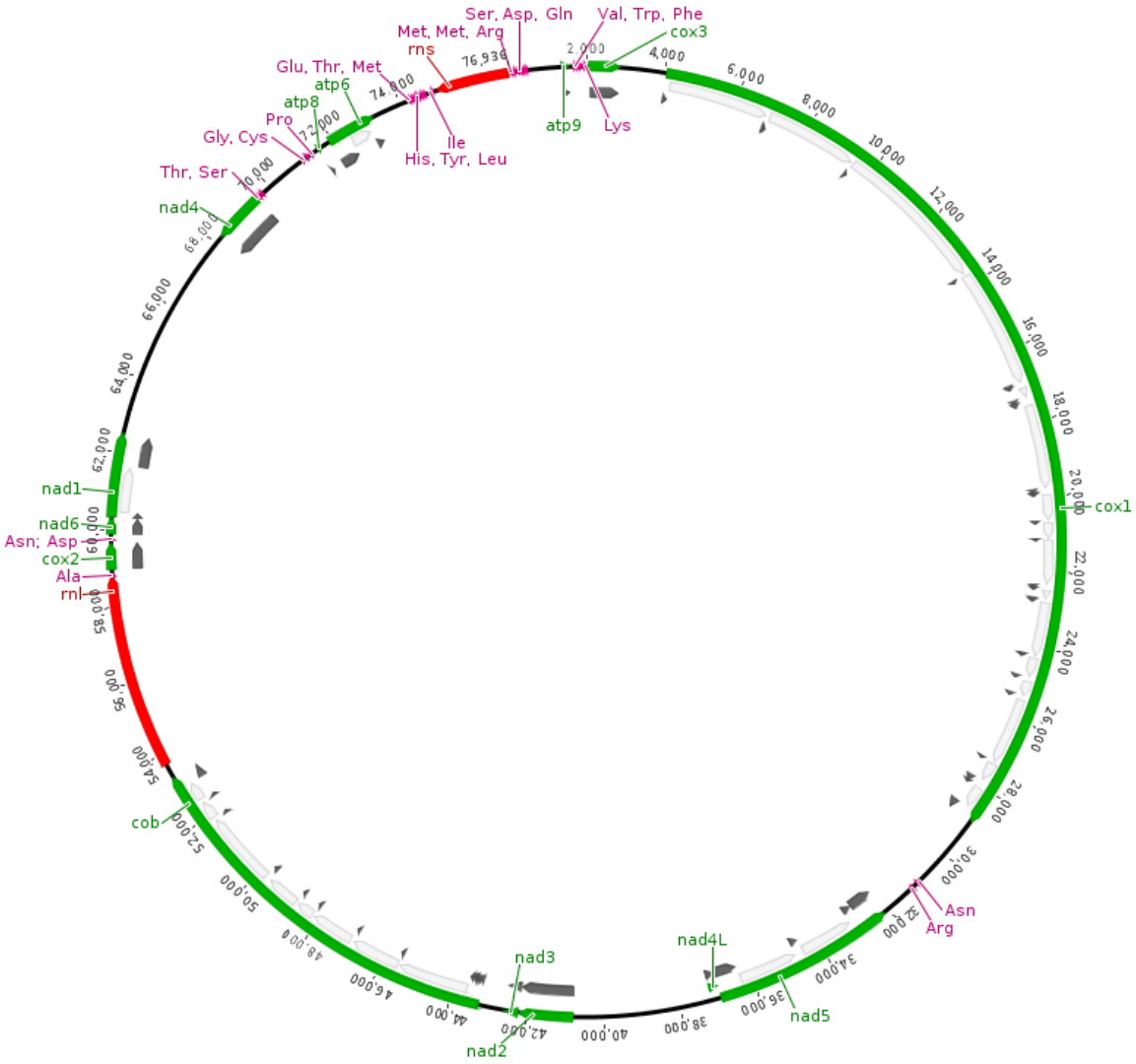

Figure S 41: Mitochondrial genome of *Metschnikowia ipomoeae* UWOPS10-104.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

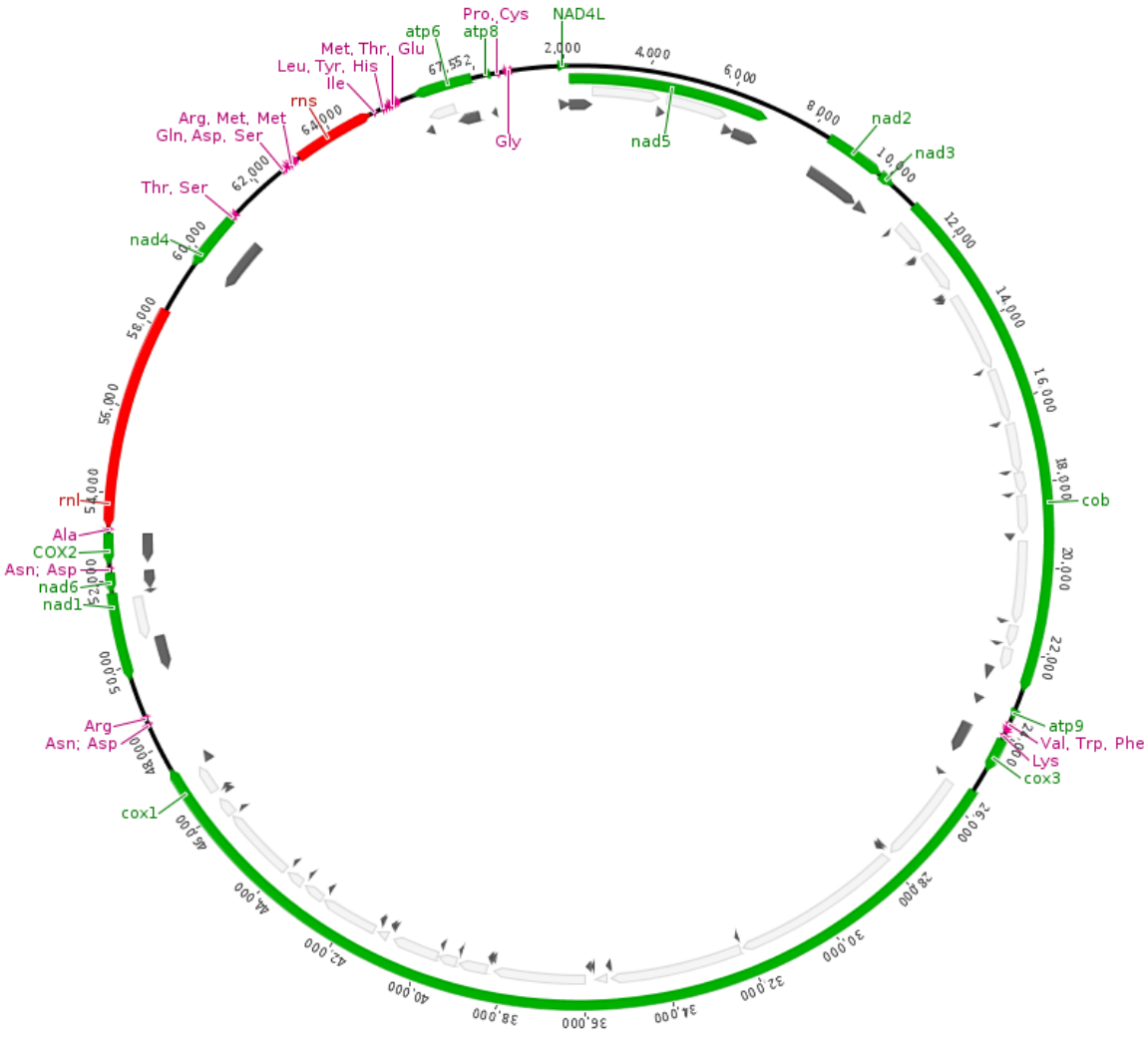

Figure S 42: Mitochondrial genome of *Metschnikowia ipomoeae* UWOPS99-324.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

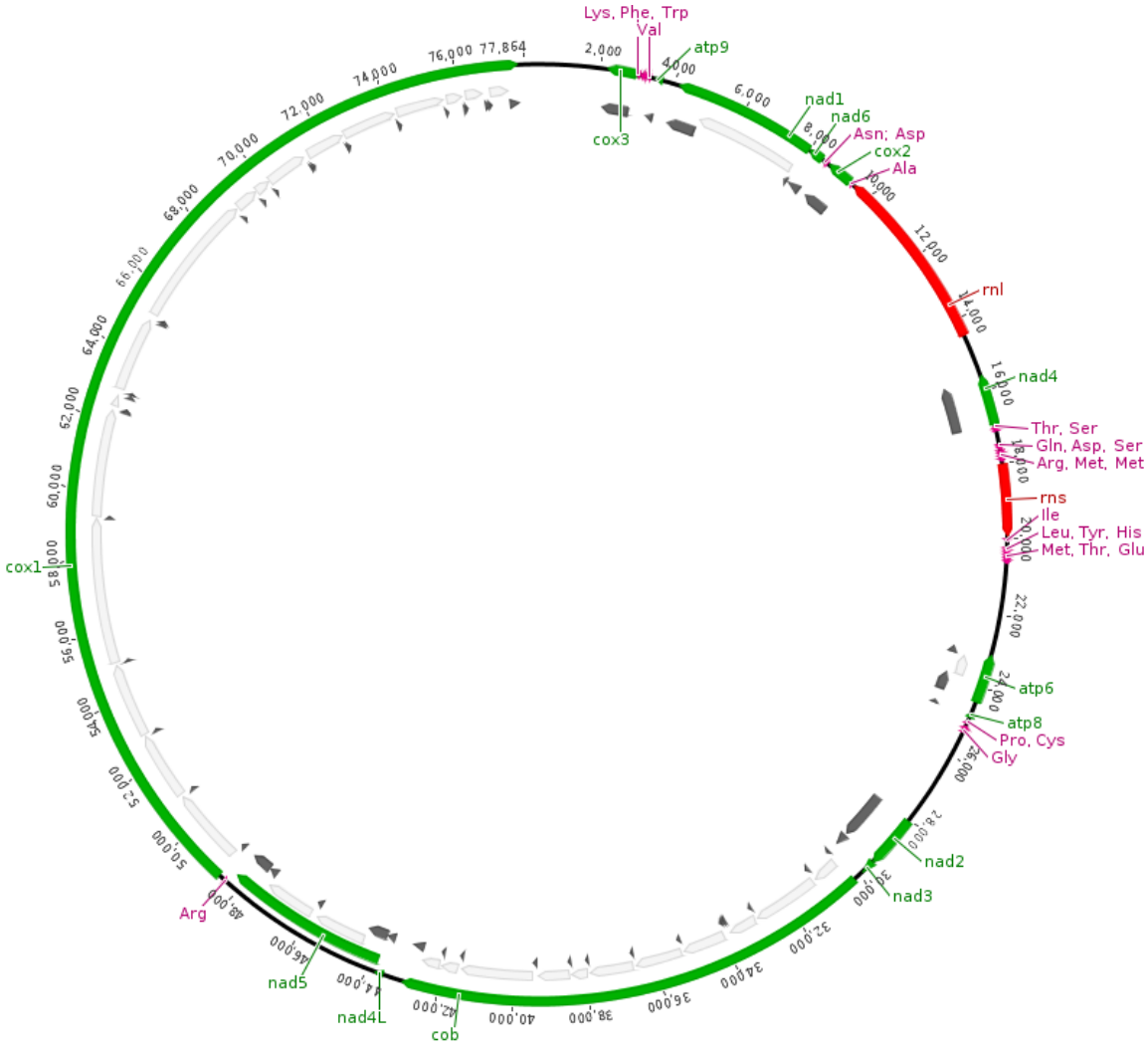

Figure S 43: Mitochondrial genome of *Metschnikowia ipomoeae* UWOPS01-141c3. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

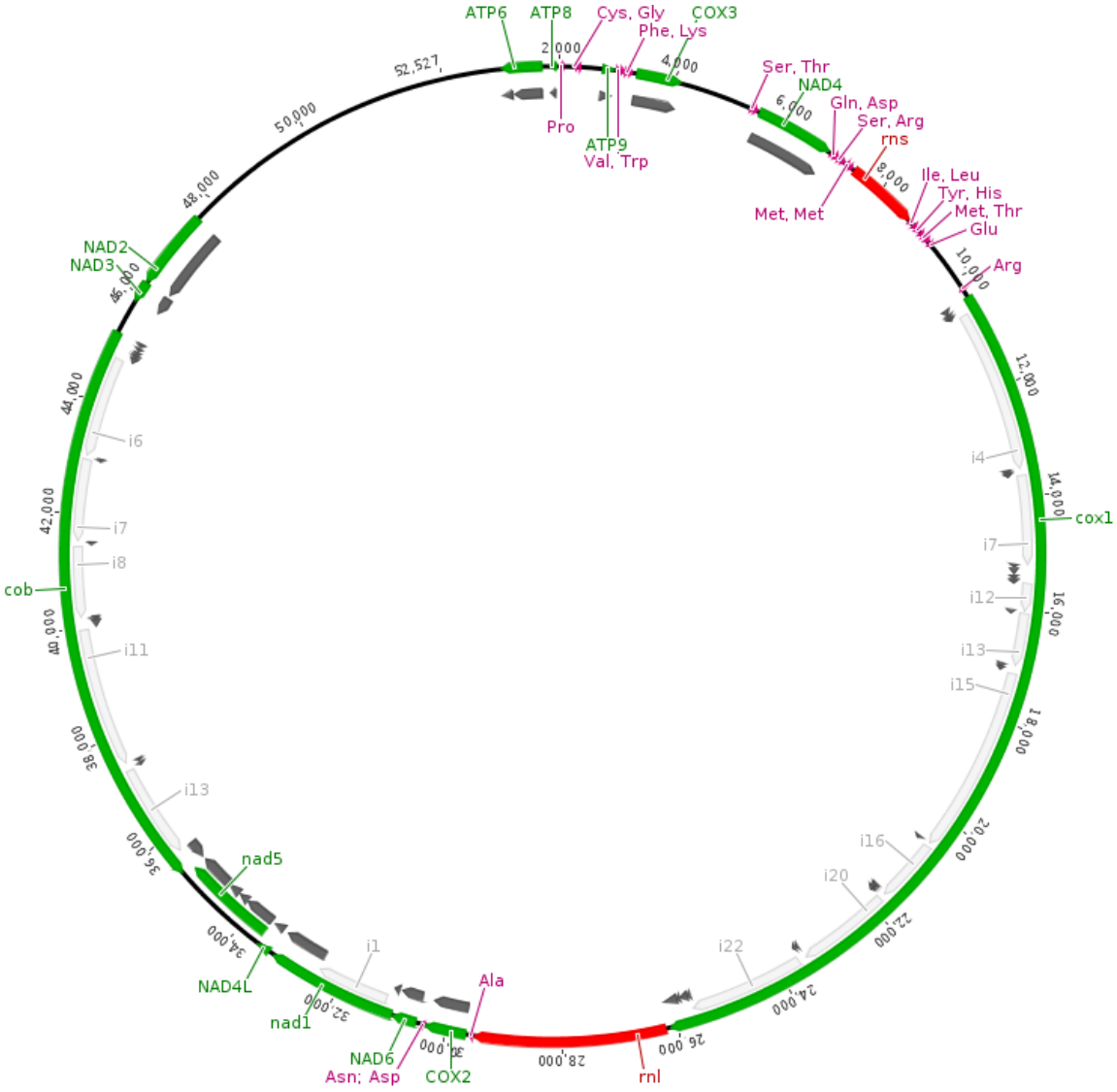

Figure S 44: Mitochondrial genome of *Metschnikowia kamakouana* UWOPS04-112.5. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

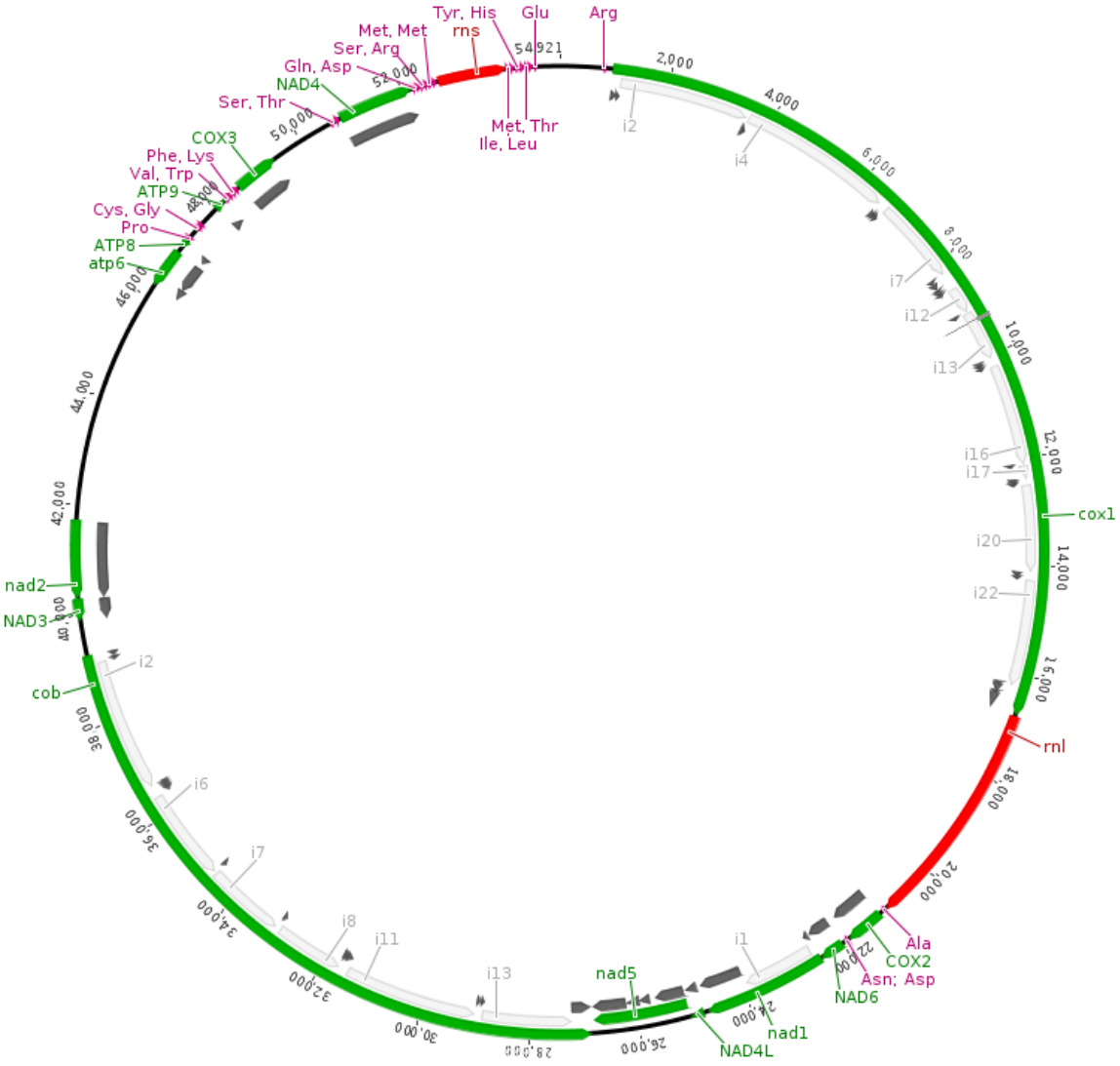

Figure S 45: Mitochondrial genome of *Metschnikowia kamakouana* UWOPS04-206.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

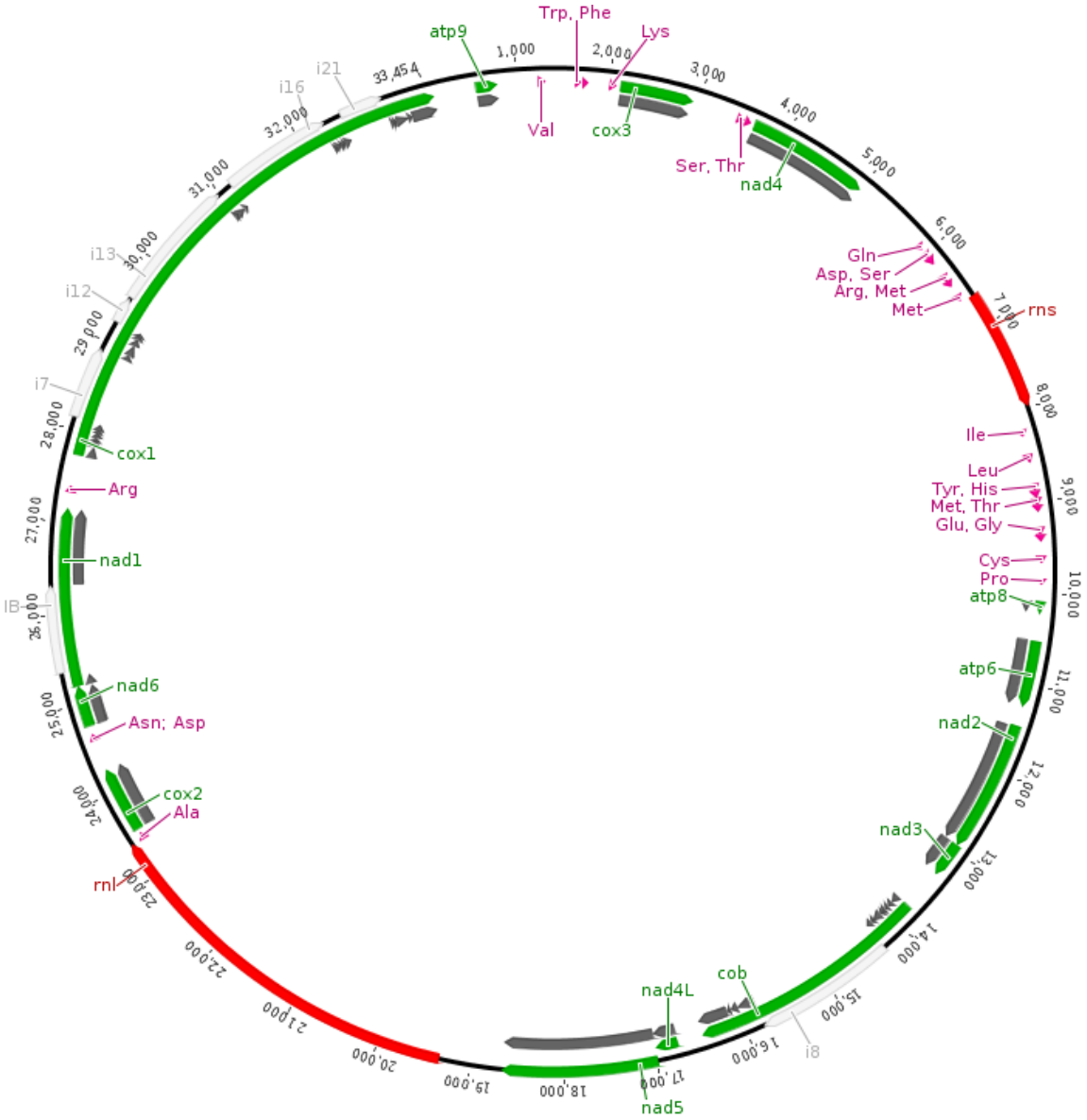

Figure S 46: Mitochondrial genome of *Metschnikowia kipukae* UWOPS00-669.2. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

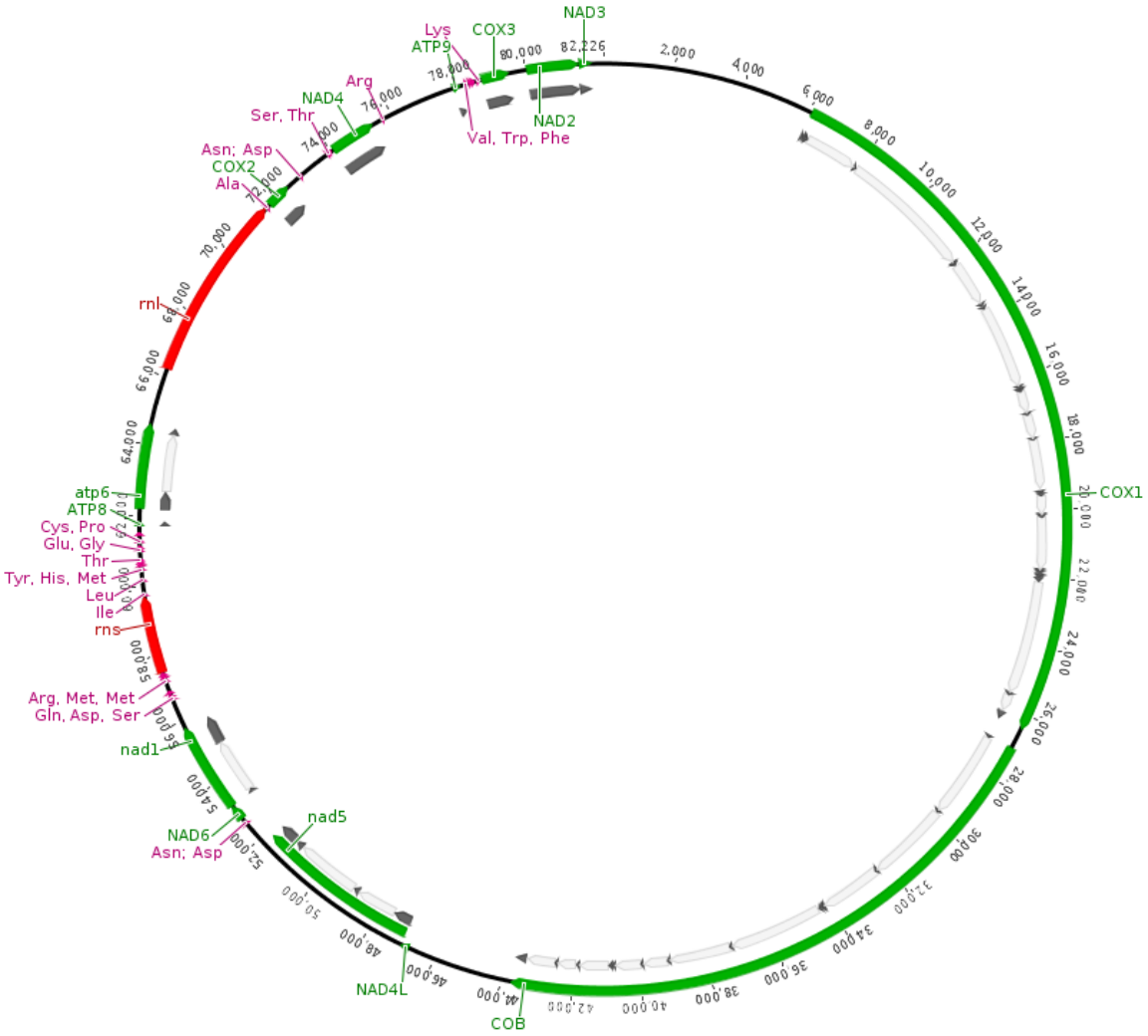

Figure S 47: Mitochondrial genome of *Metschnikowia lacustris* UWOPS12-619.2. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

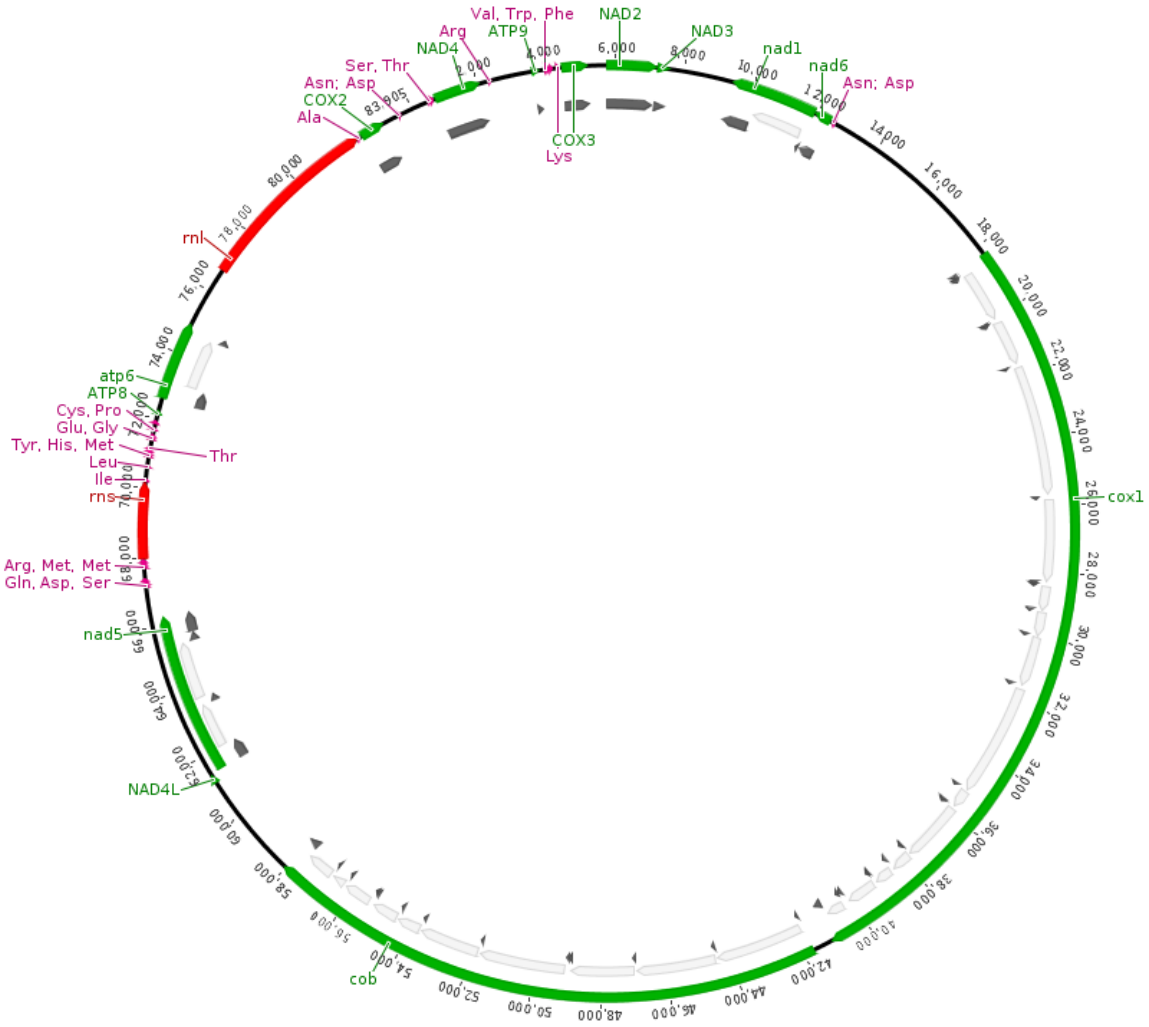

Figure S 48: Mitochondrial genome of *Metschnikowia lacustris* UWOPS03-172.2. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

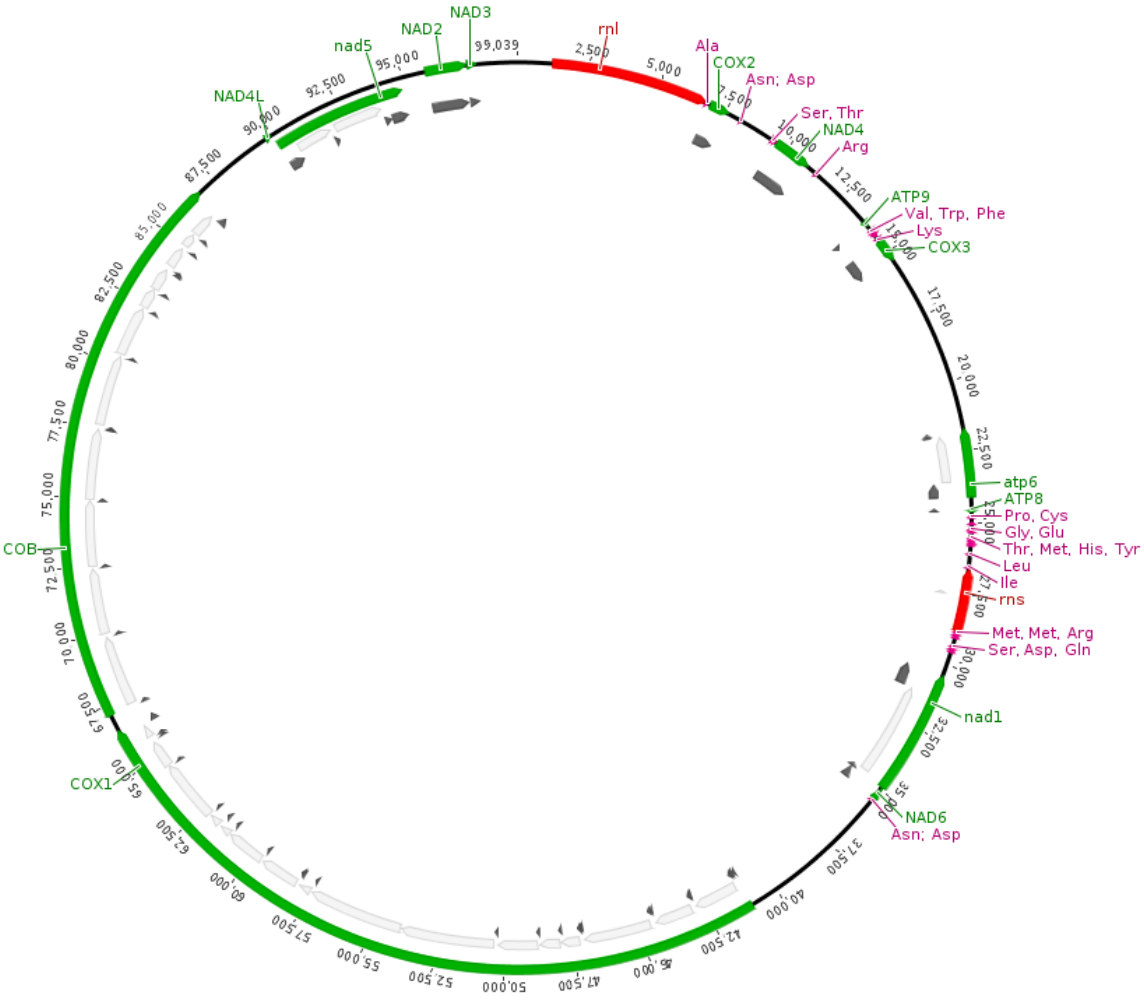

Figure S 49: Mitochondrial genome of *Metschnikowia lacustris* UWOPS03-167b3. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

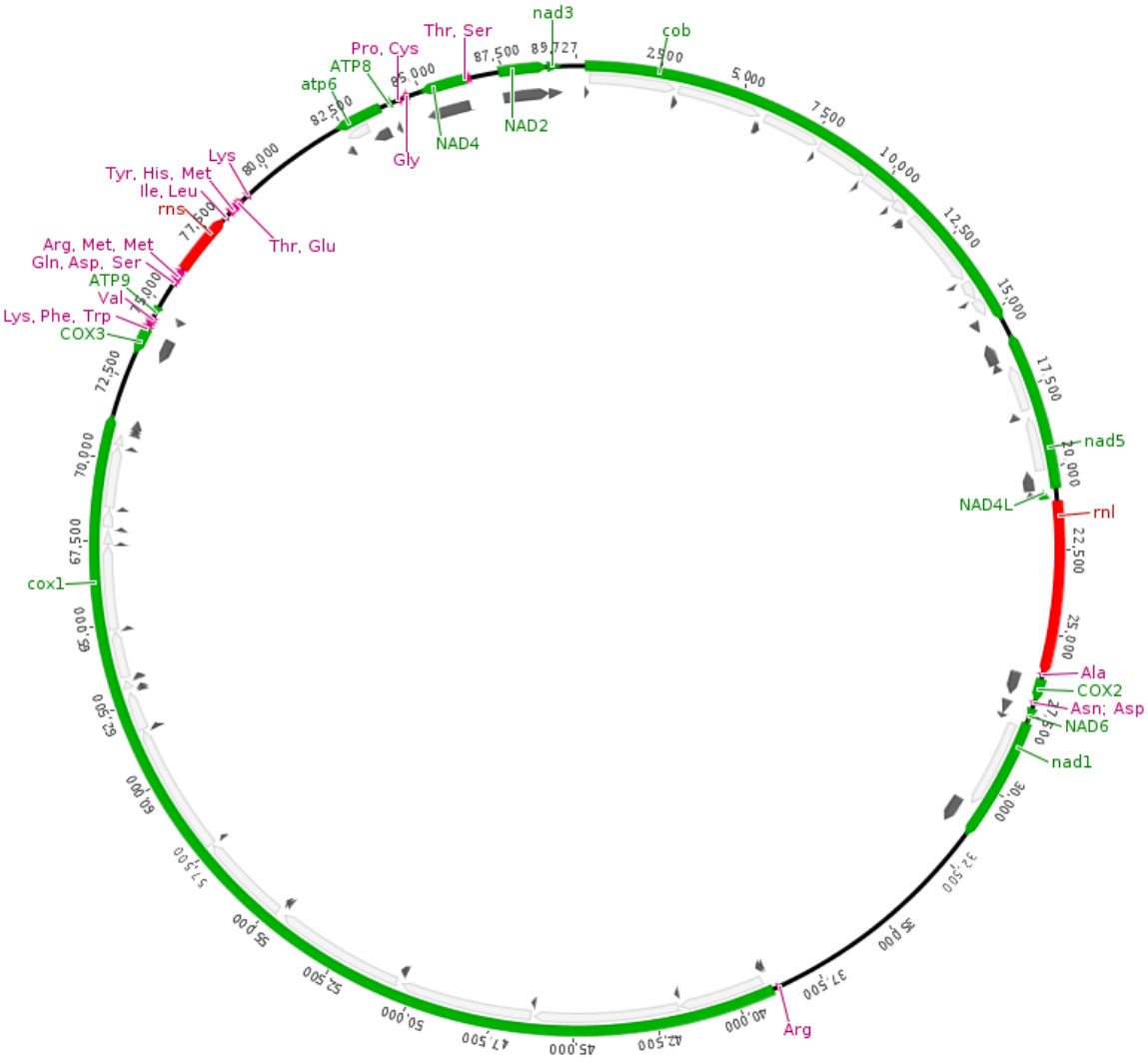

Figure S 50: Mitochondrial genome of *Metschnikowia lochheadii* UWOPS03-167a3. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

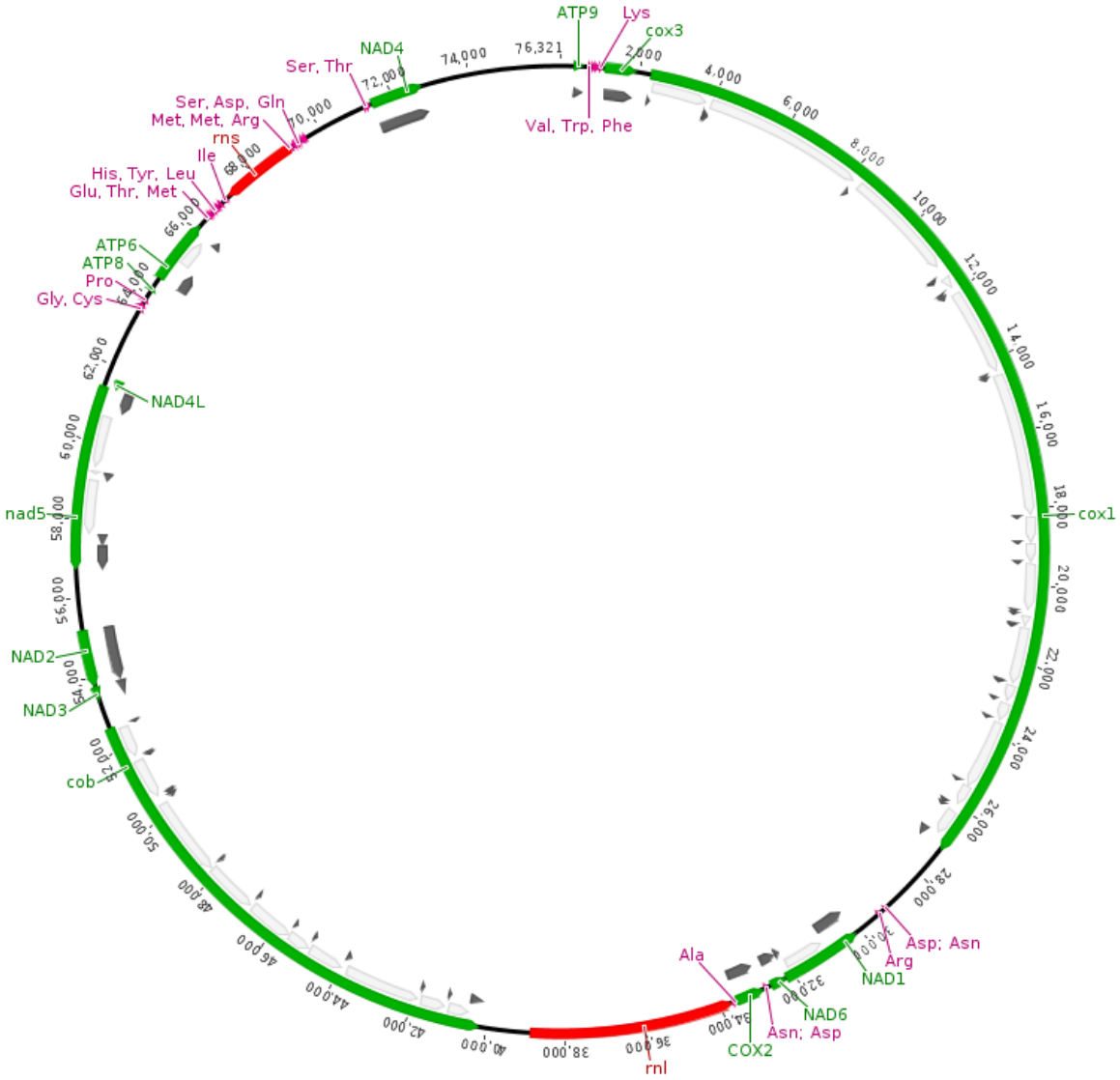

Figure S 51: Mitochondrial genome of *Metschnikowia lochheadii* UWOPS99-661.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

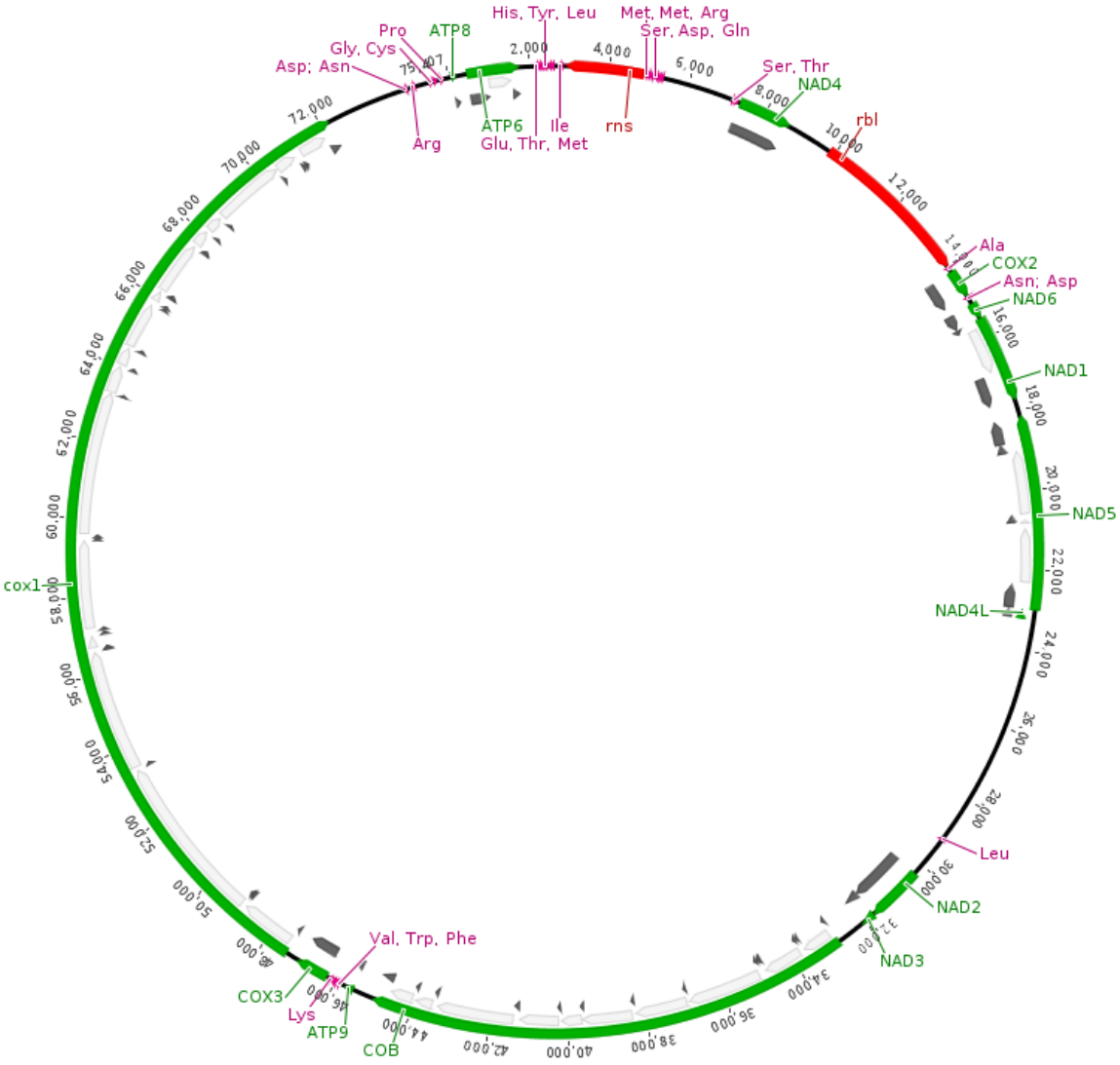

Figure S 52: Mitochondrial genome of *Metschnikowia matae* var *maris* UFMG-CM-Y397. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

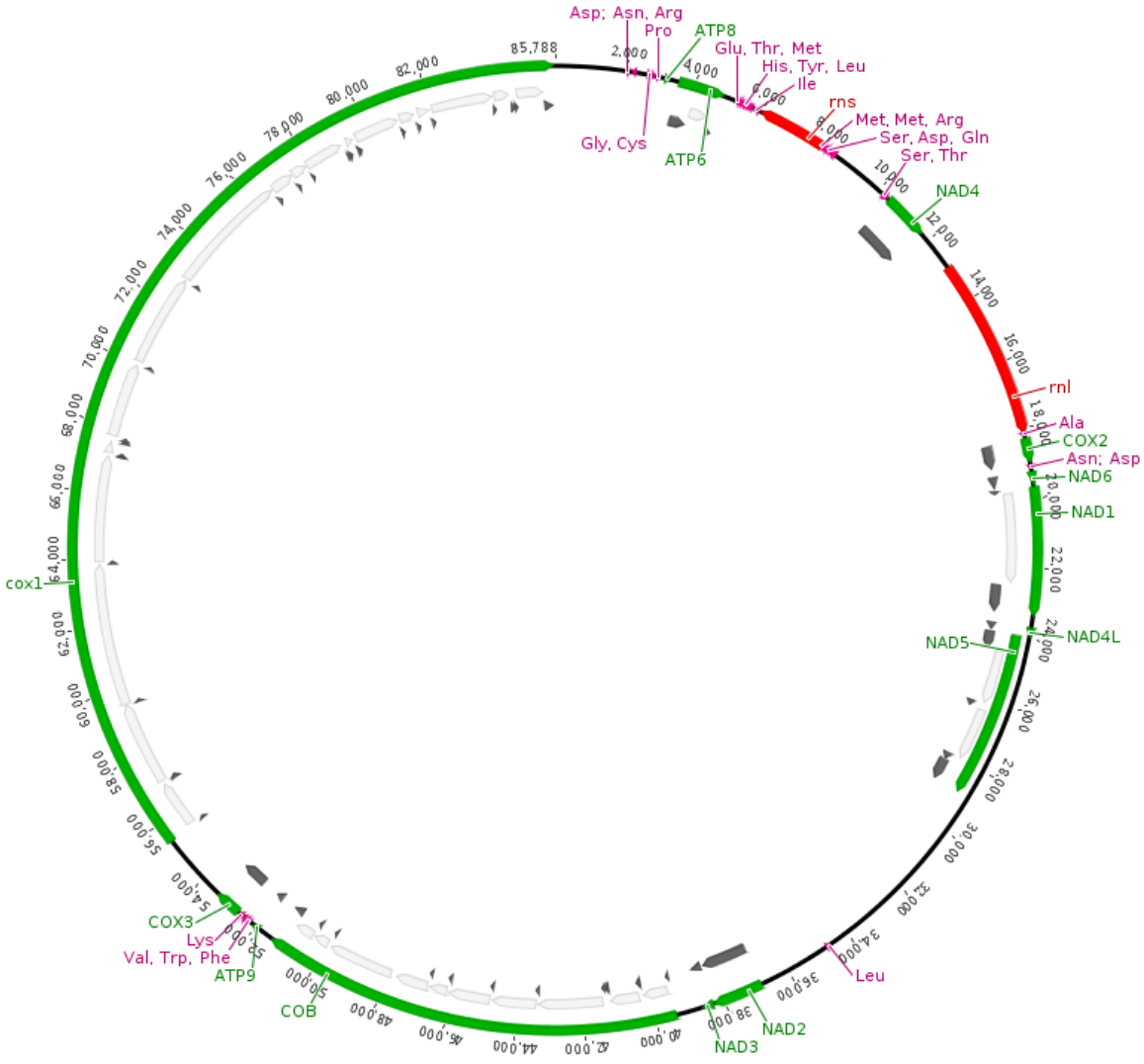

Figure S 53: Mitochondrial genome of *Metschnikowia matae* var *matae* UFMG-CM-Y395. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

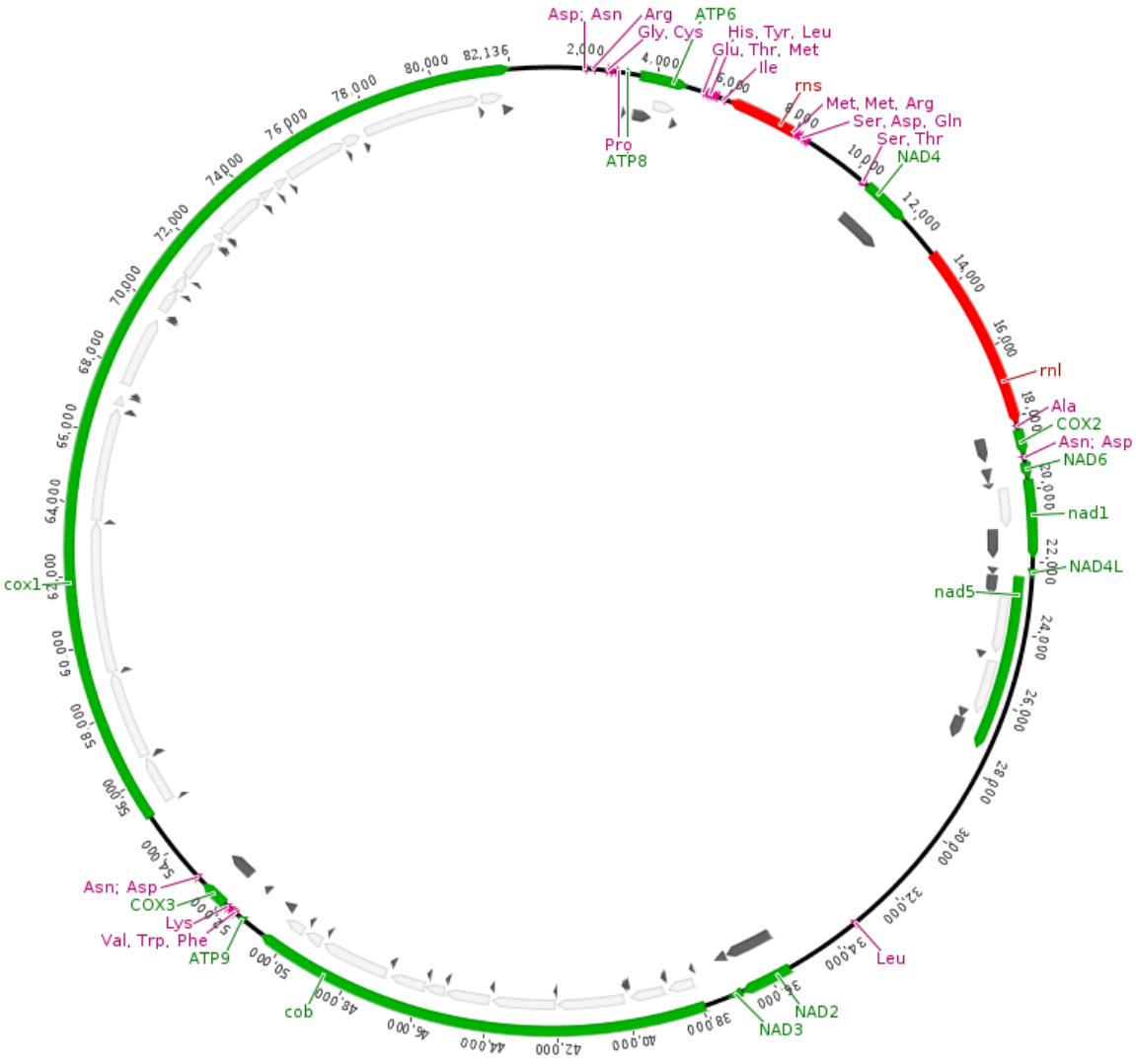

Figure S 54: Mitochondrial genome of *Metschnikowia matae* var *matae* UFMG-CM-Y391. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

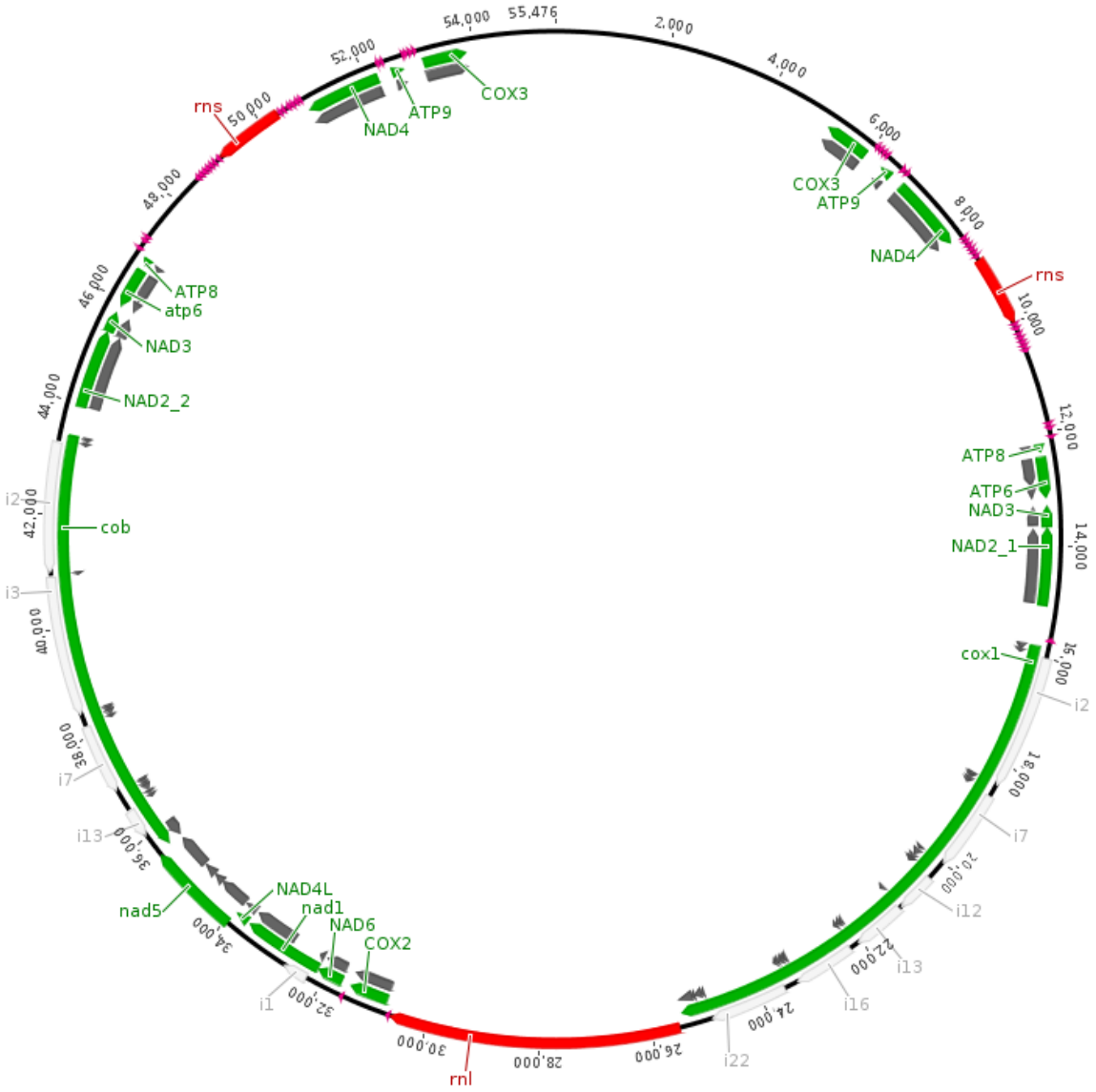

Figure S 55: Mitochondrial genome of *Metschnikowia mauinuiana* UWOPS04-190.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

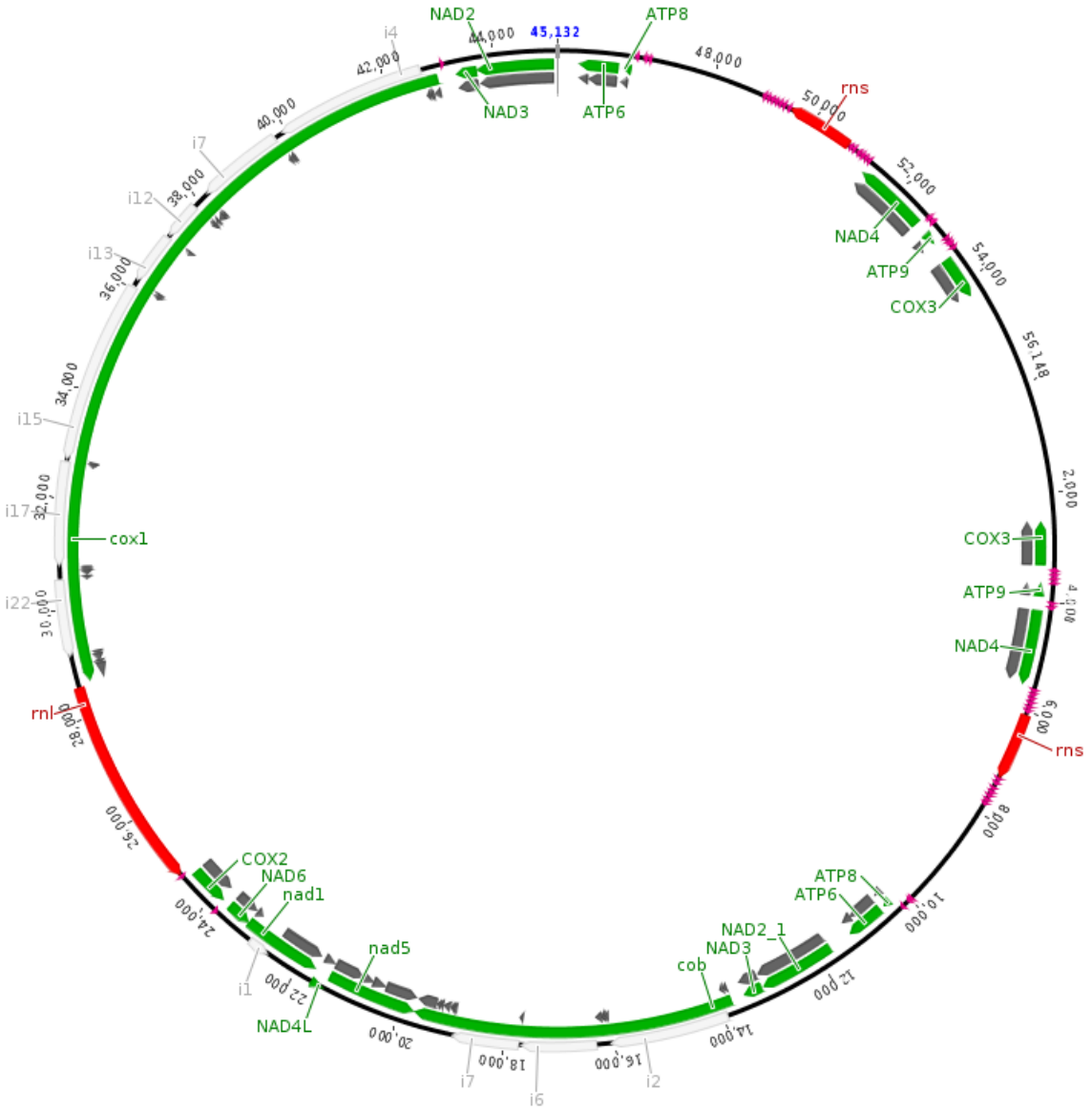

Figure S 56: Mitochondrial genome of *Metschnikowia mauinuiana* UWOPS04-110.4. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

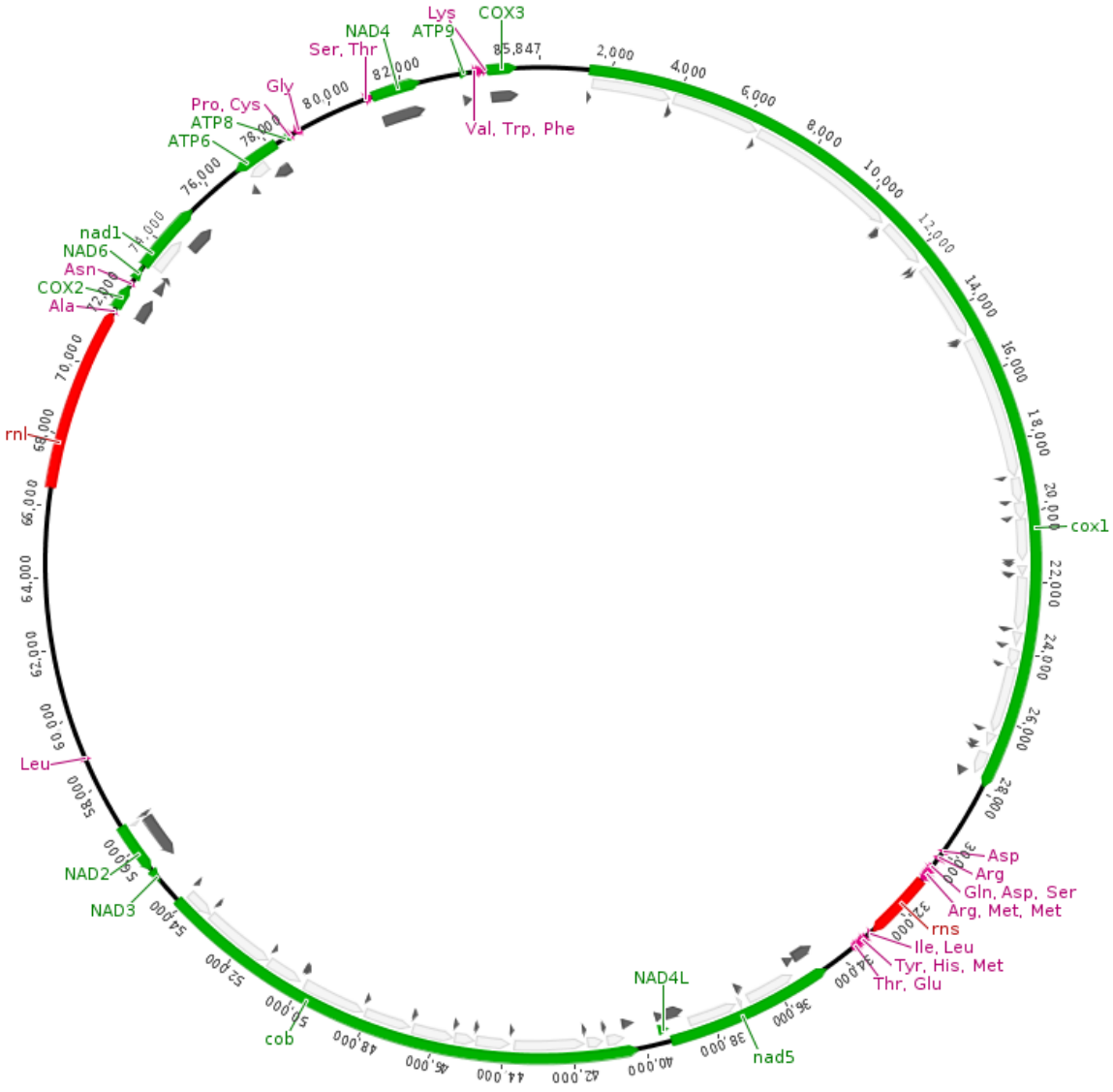

Figure S 57: Mitochondrial genome of *Metschnikowia* sp. UWOPS00-154.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

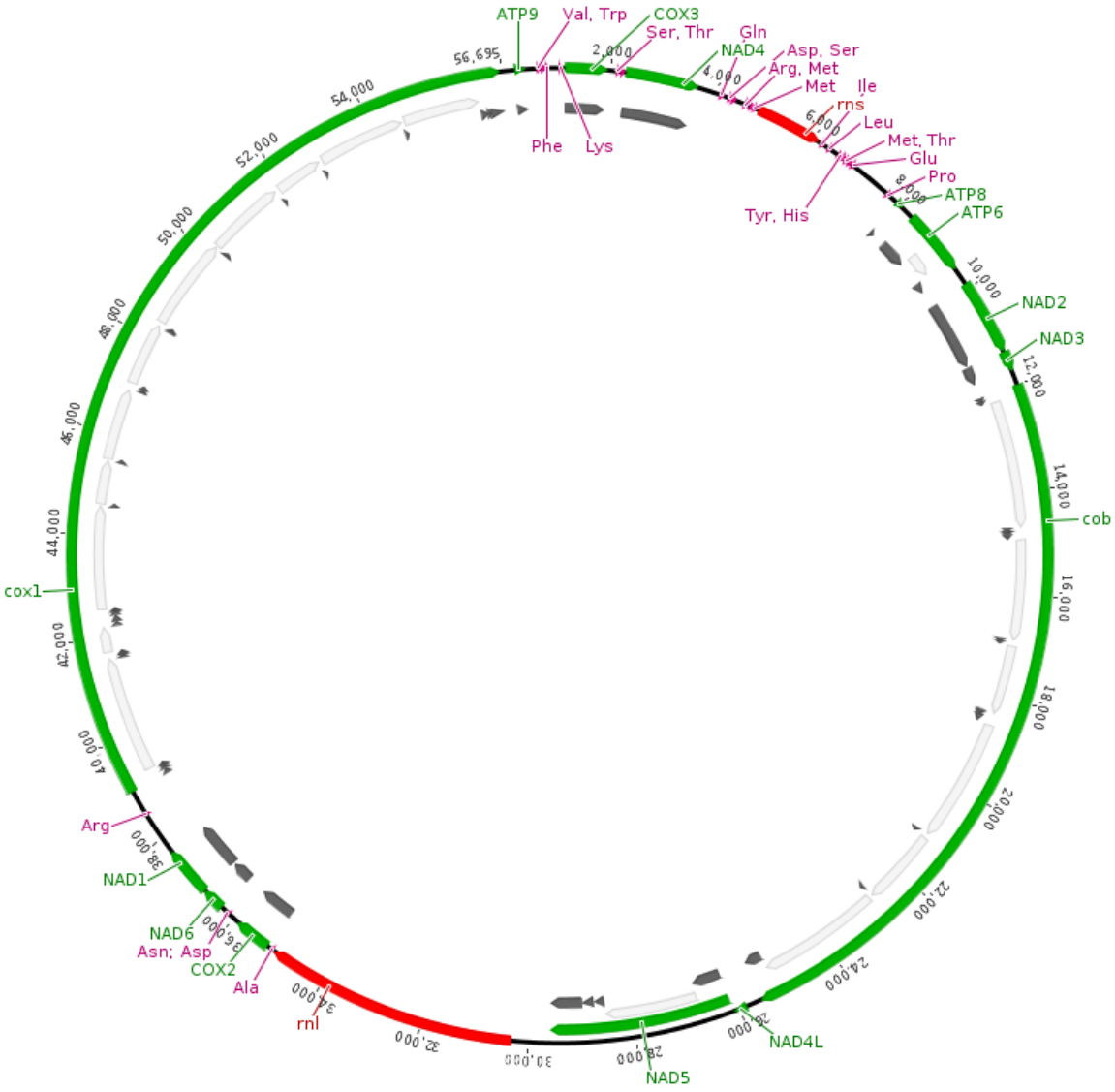

Figure S 58: Mitochondrial genome of *Metschnikowia orientalis* UWOPS99-745.6. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

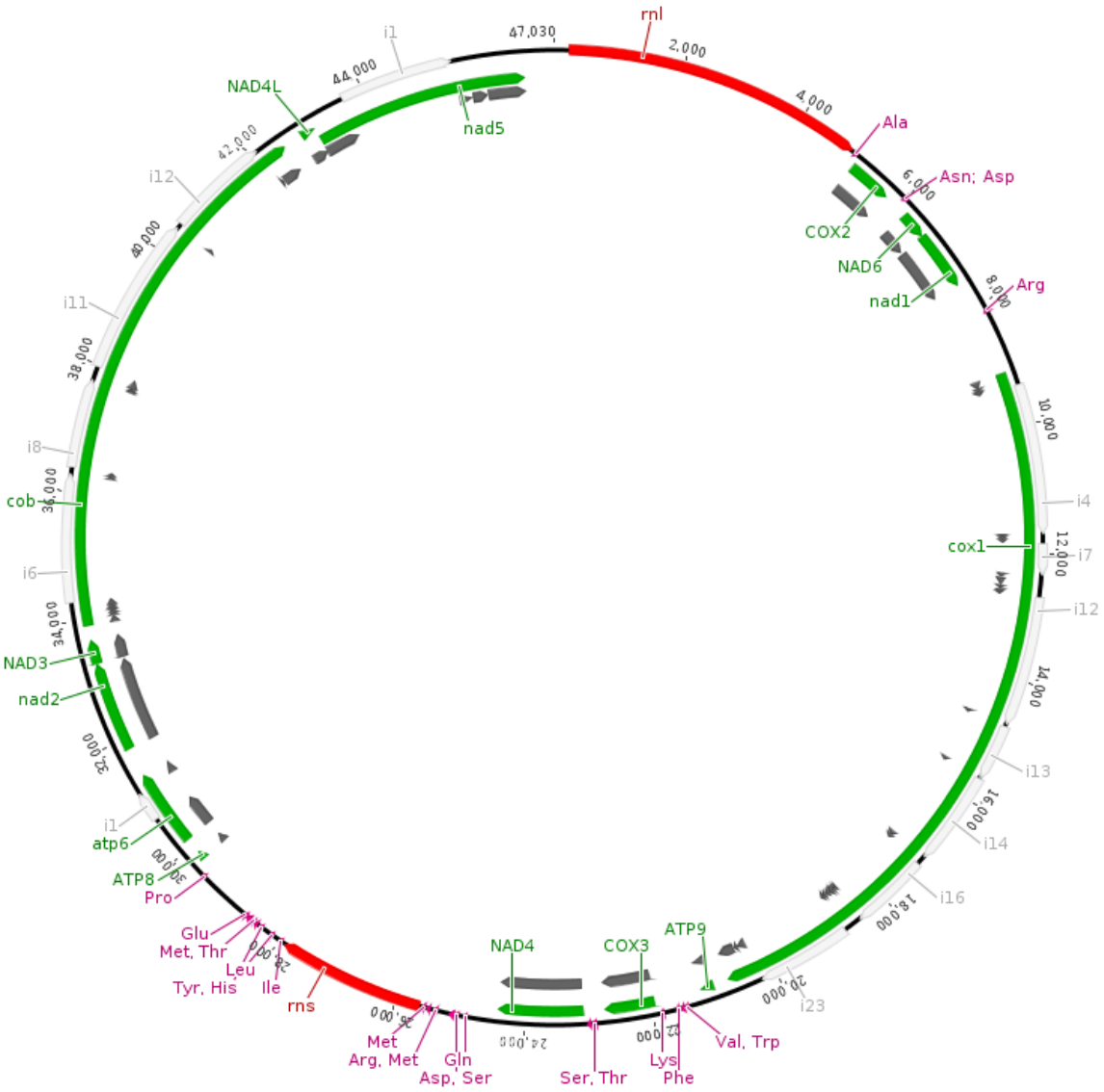

Figure S 59: Mitochondrial genome of *Metschnikowia orientalis* UWOPS05-269.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

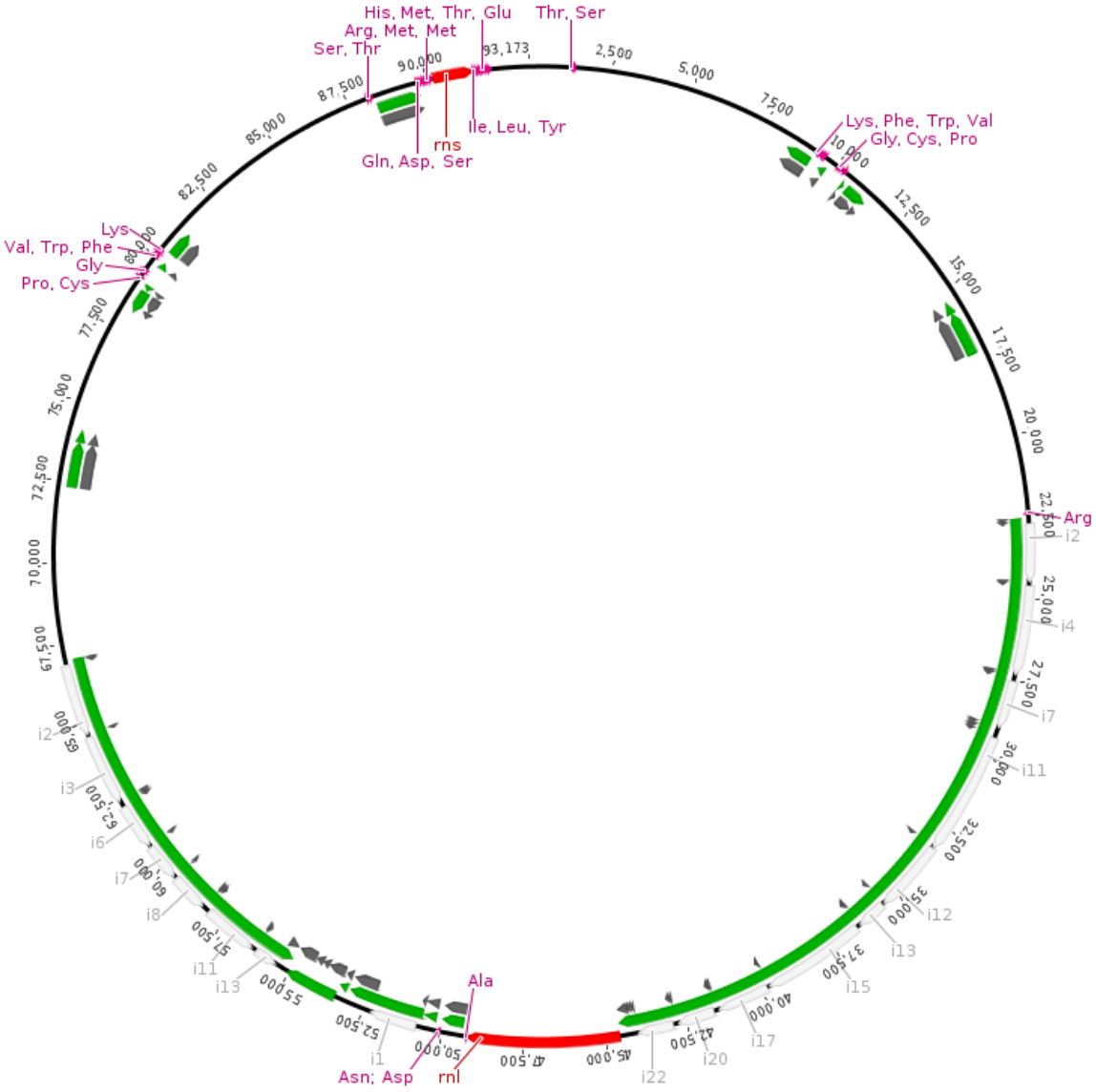

Figure S 60: Mitochondrial genome of *Metschnikowia* sp. UWOPS04-218.3. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

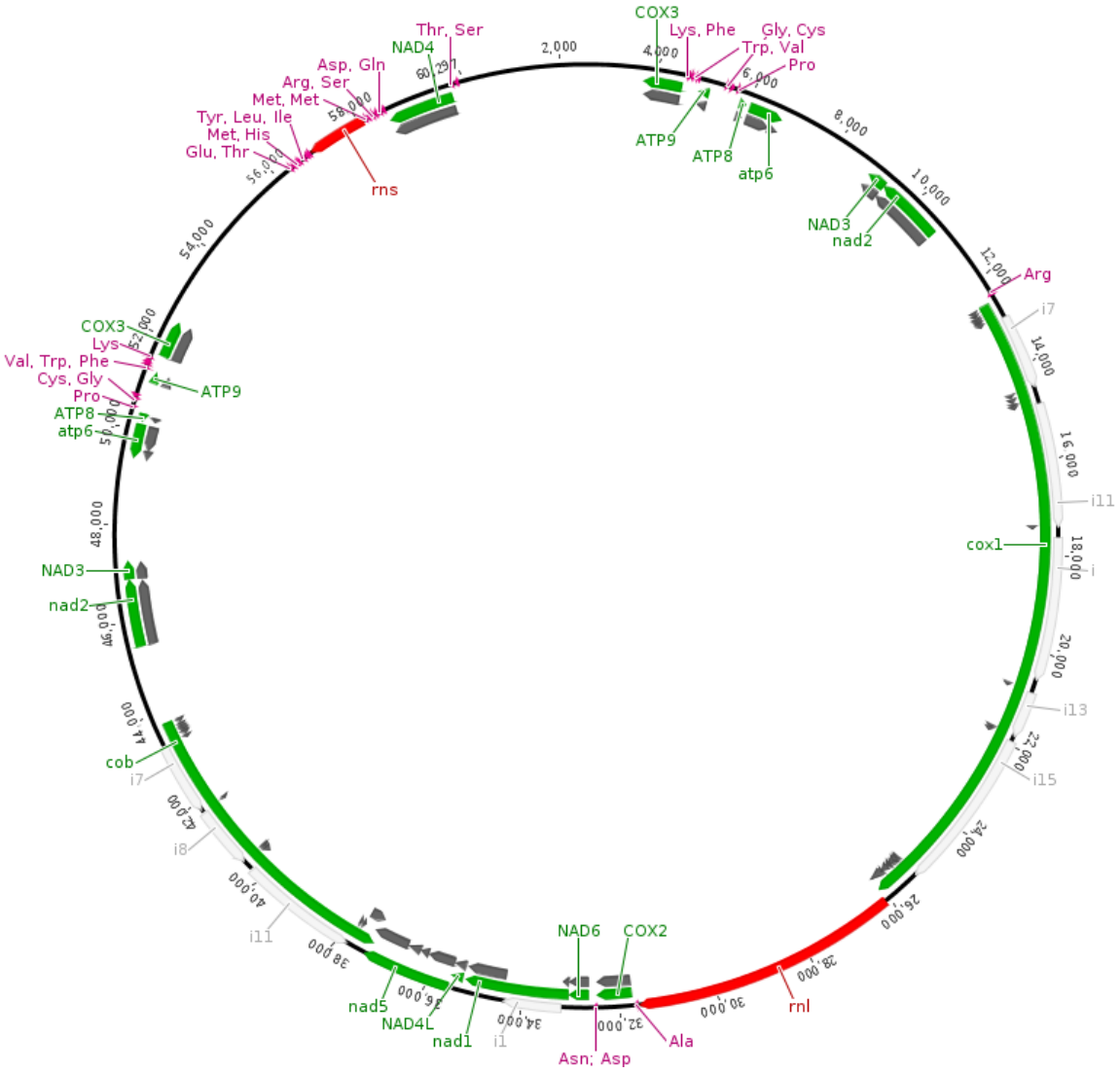

Figure S 61: Mitochondrial genome of *Metschnikowia* sp. UWOPS04-226.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

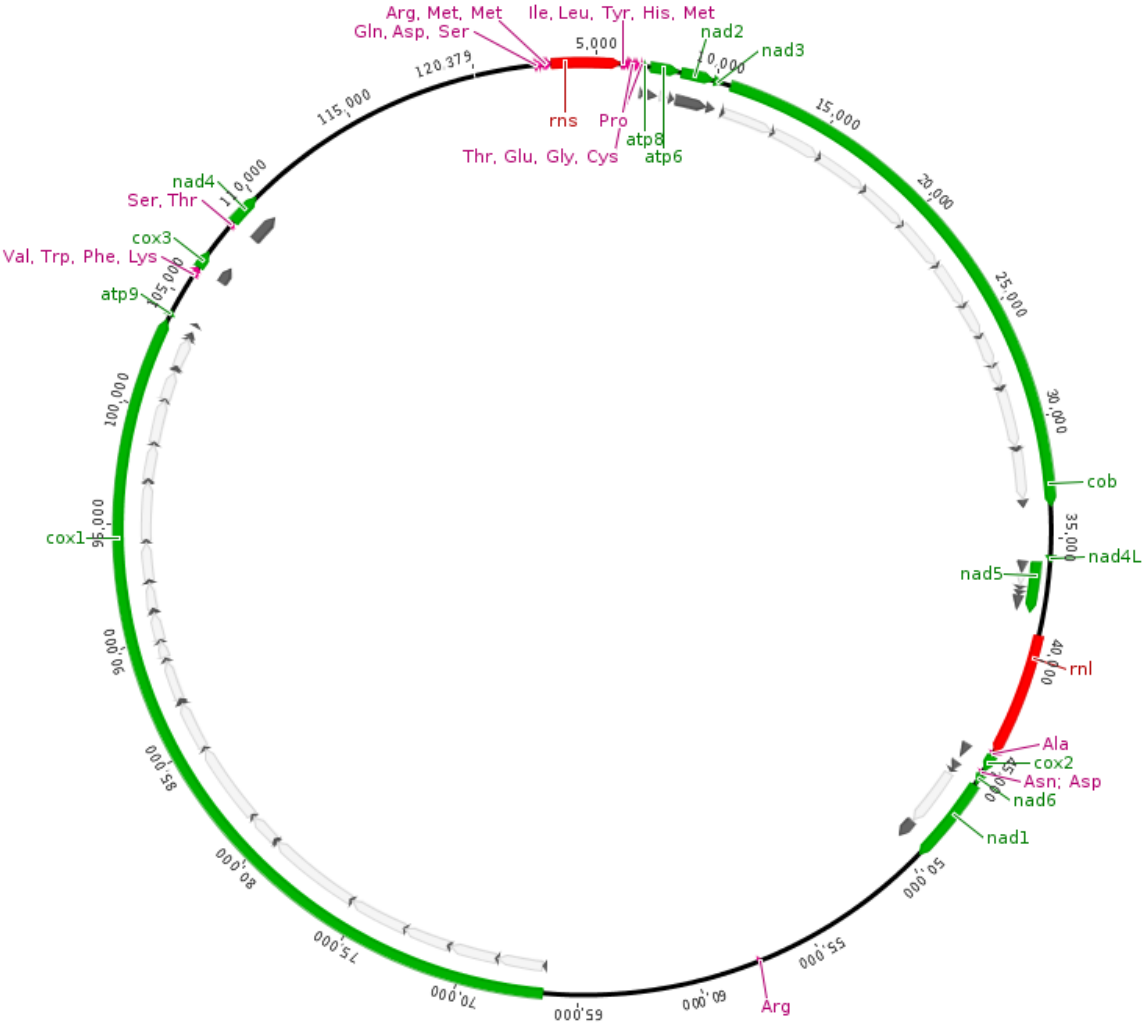

Figure S 62: Mitochondrial genome of *Metschnikowia proteae* EBD-T1Y1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

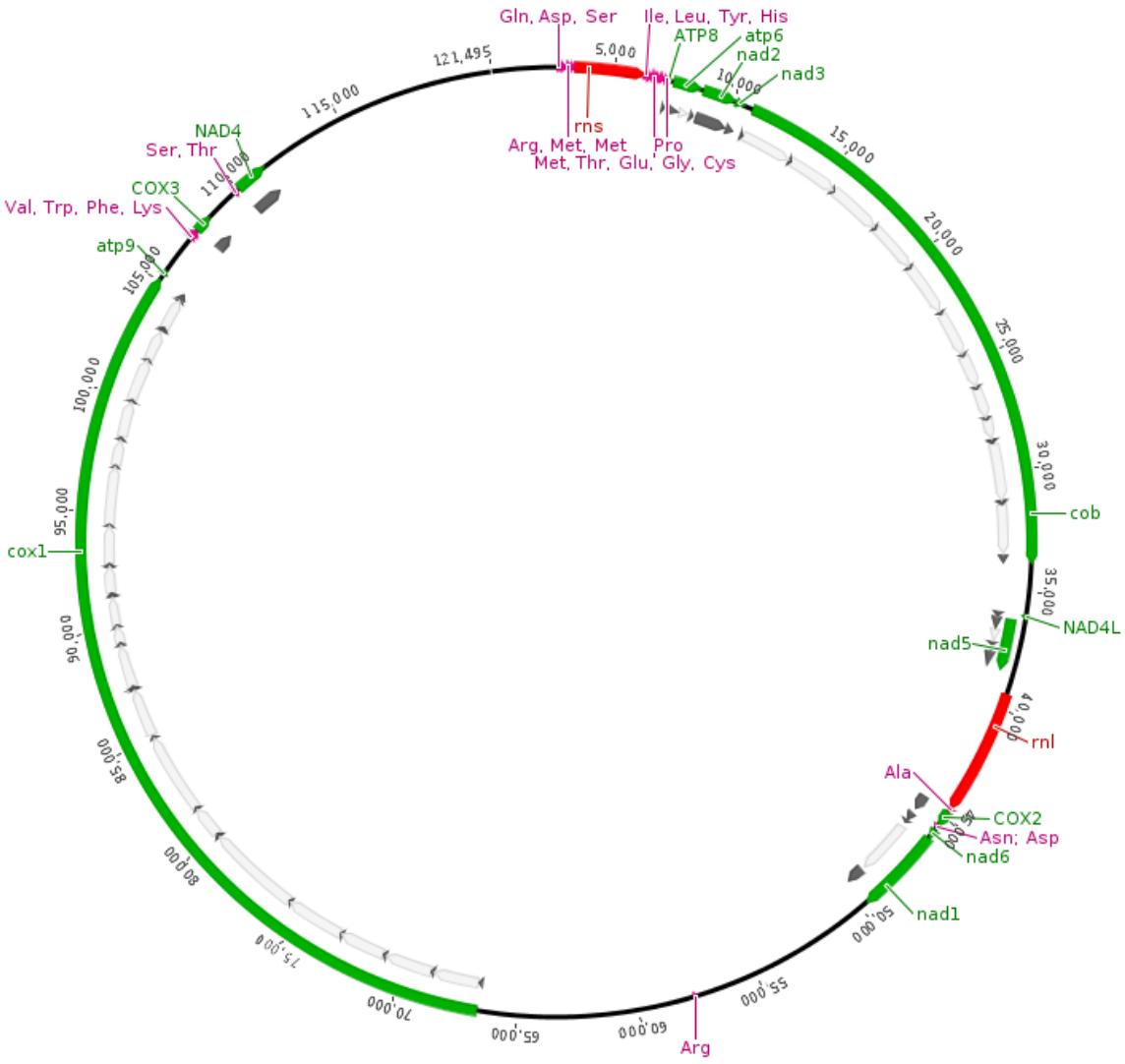

Figure S 63: Mitochondrial genome of *Metschnikowia proteae* EBD-A10Y1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

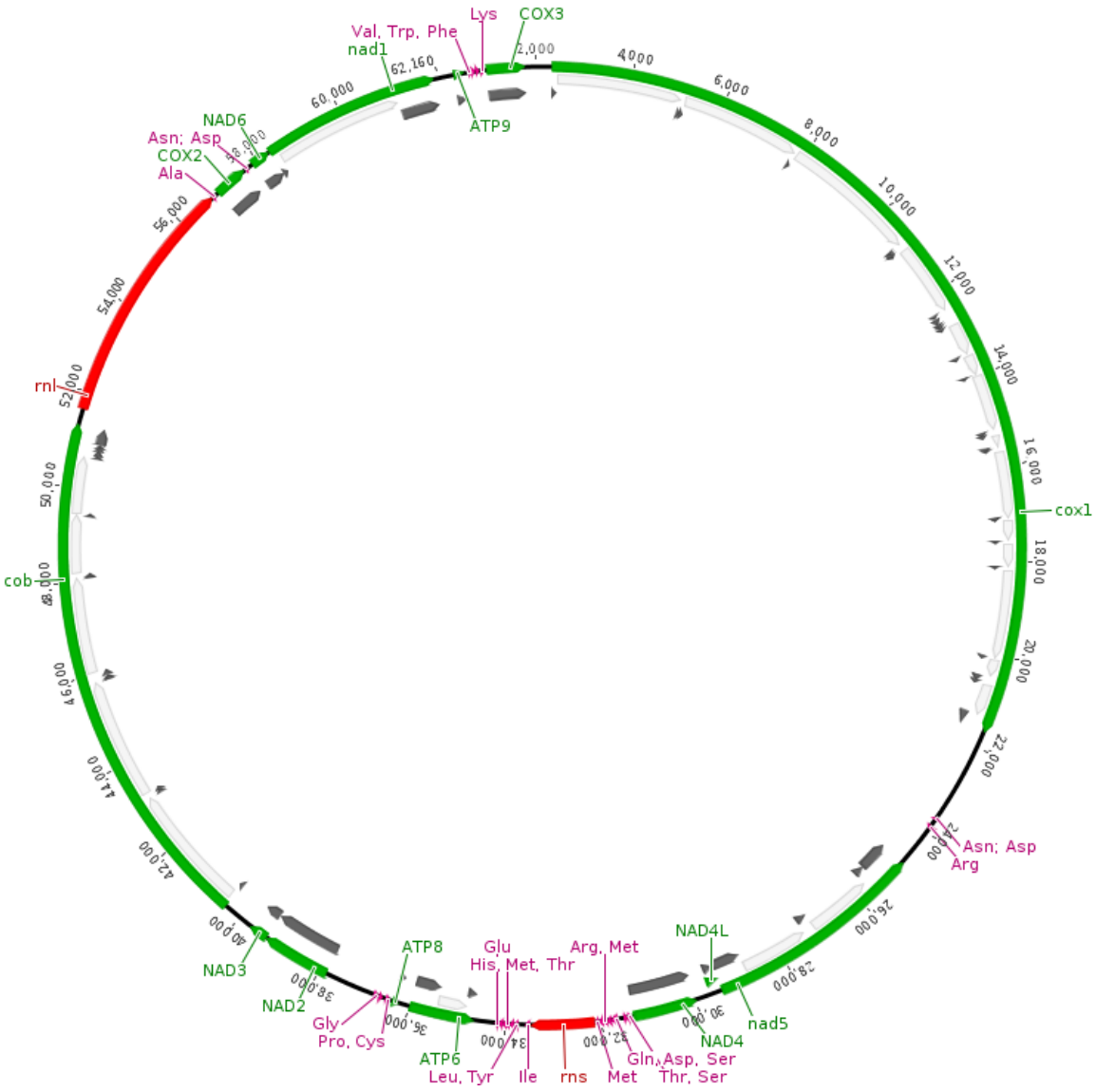

Figure S 64: Mitochondrial genome of *Metschnikowia santaceciliae* UWOPS01-517a1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

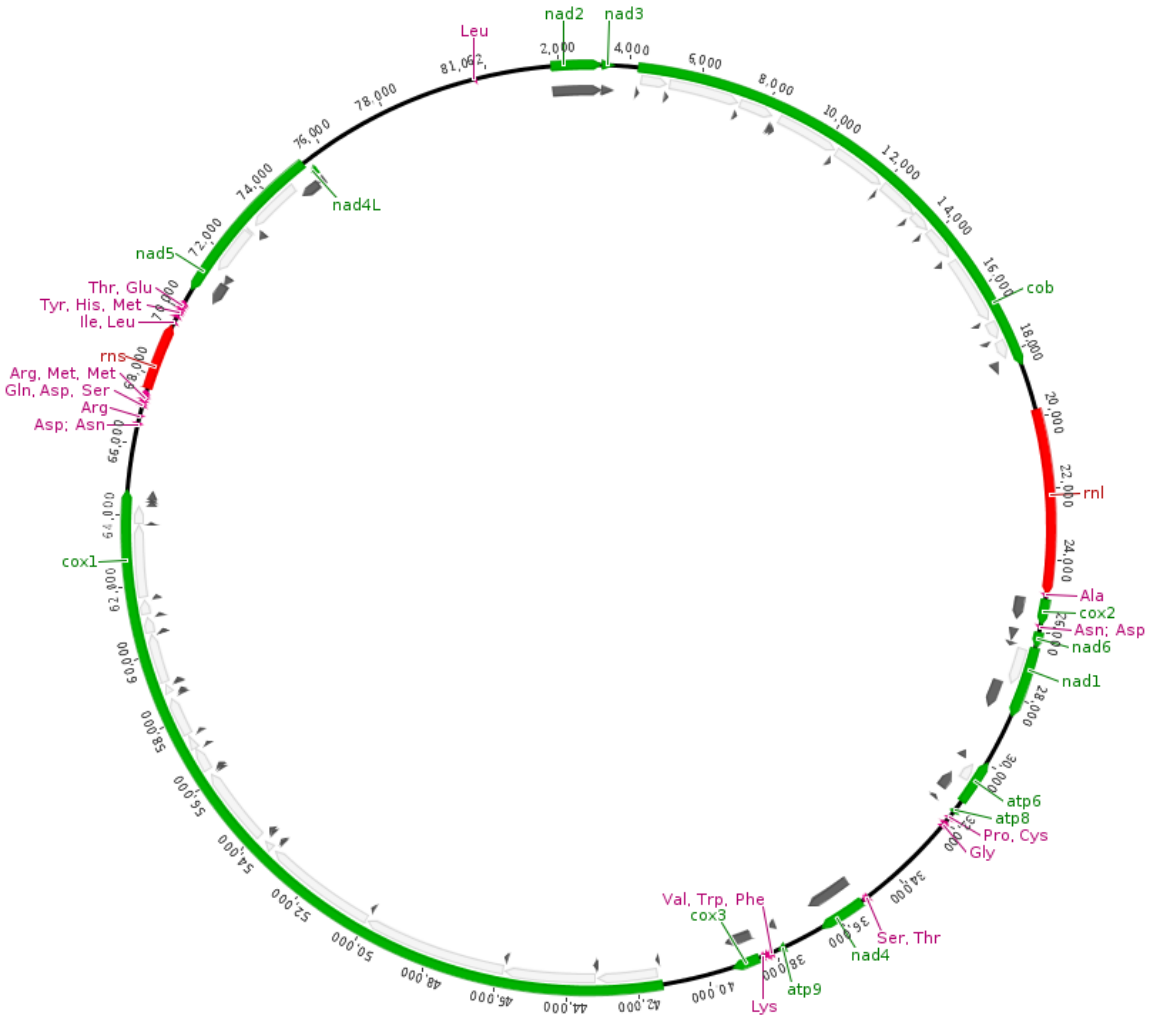

Figure S 65: Mitochondrial genome of *Metschnikowia santaceciliae* UWOPS01-142b1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

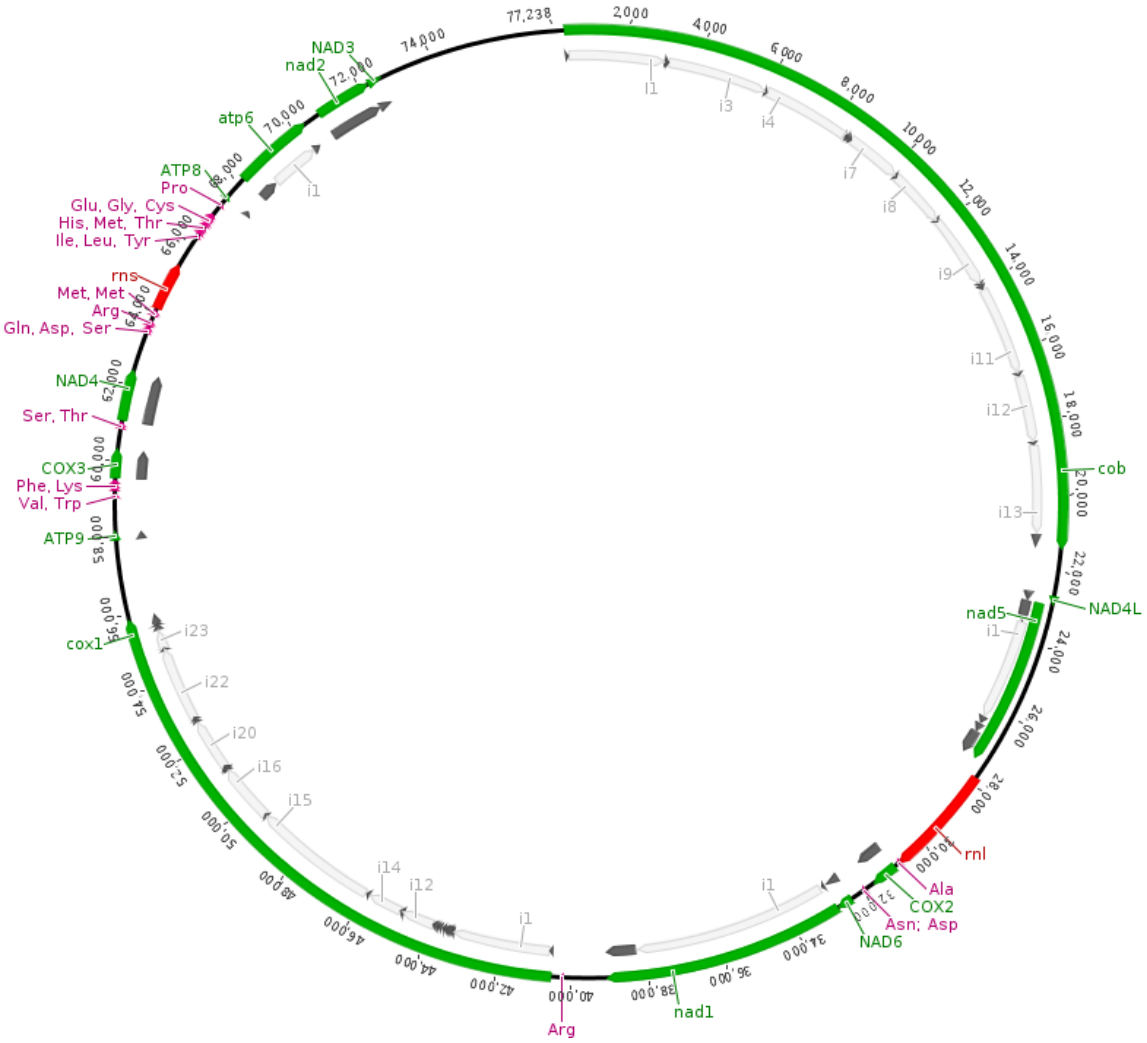

Figure S 66: Mitochondrial genome of *Metschnikowia shivogae* UWOPS04-310.1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

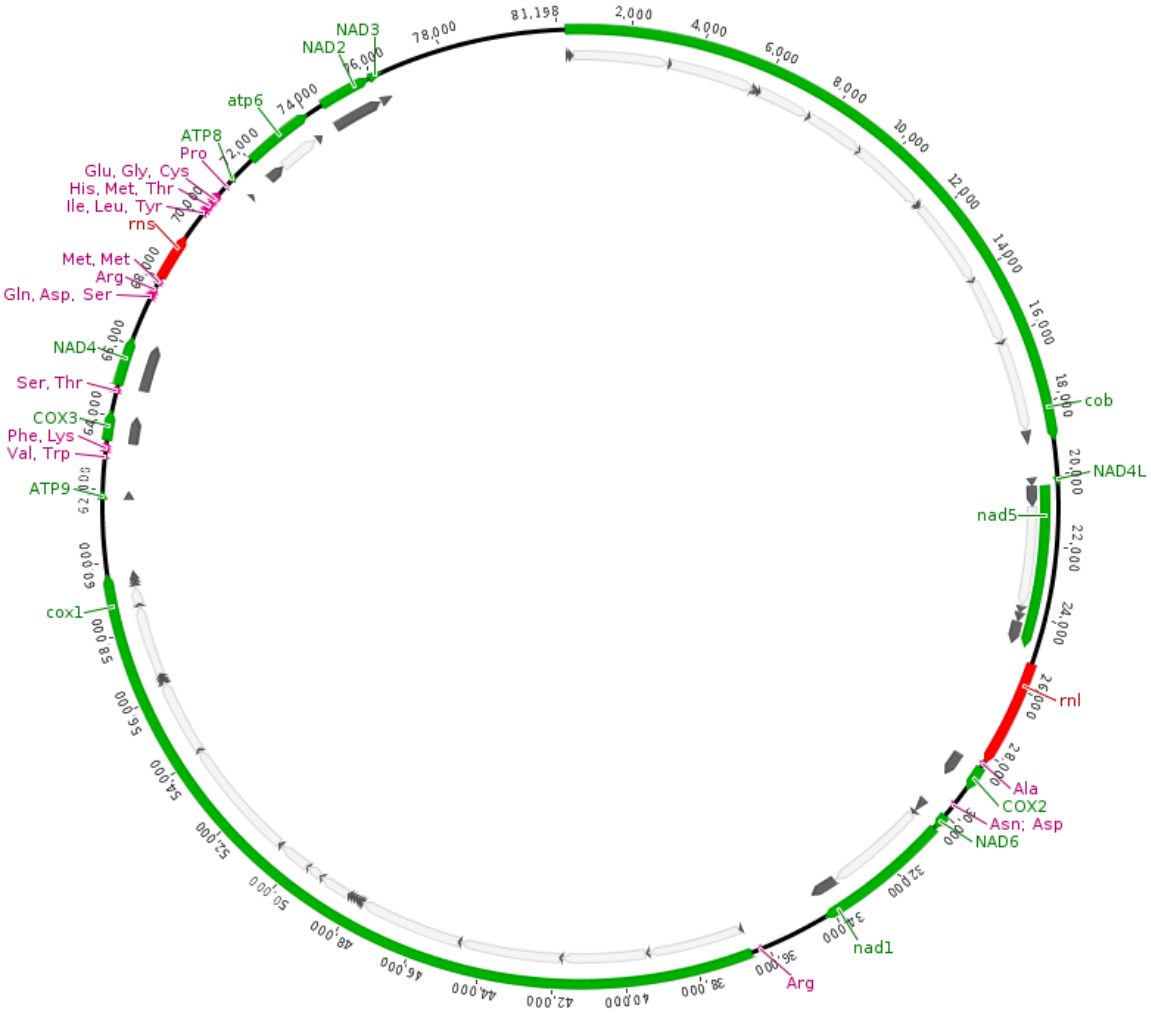

Figure S 67: Mitochondrial genome of *Metschnikowia shivogae* UWOPS07-203.2. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

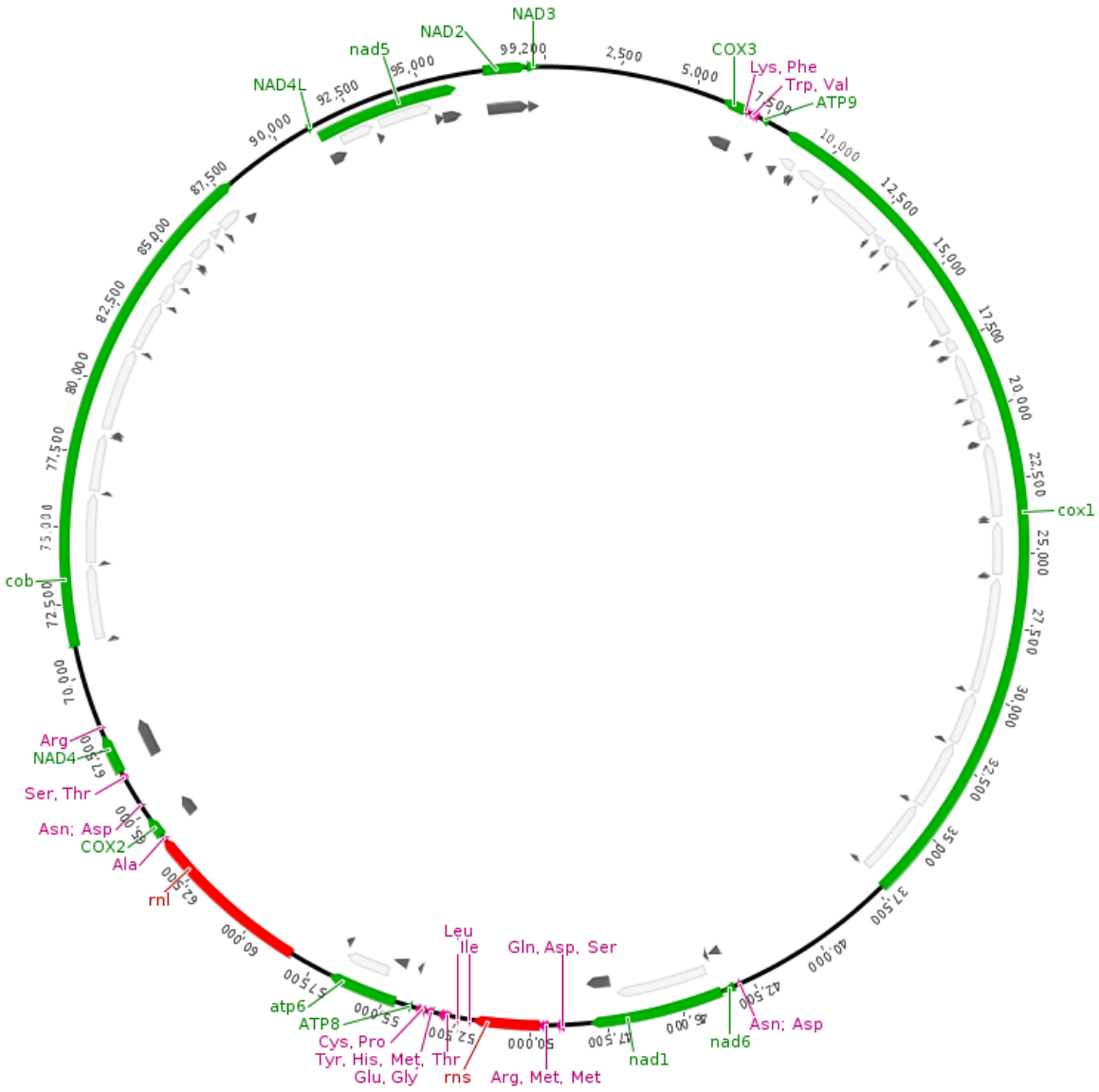

Figure S 68: Mitochondrial genome of *Metschnikowia similis* UWOPS03-158.2. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

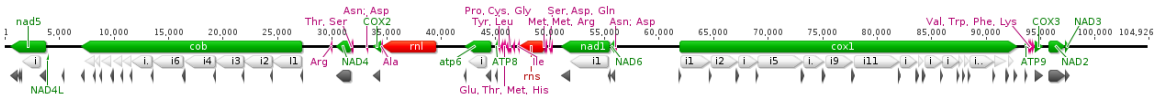

Figure S 69: Mitochondrial genome of *Metschnikowia similis* UWOPS03-133.4. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

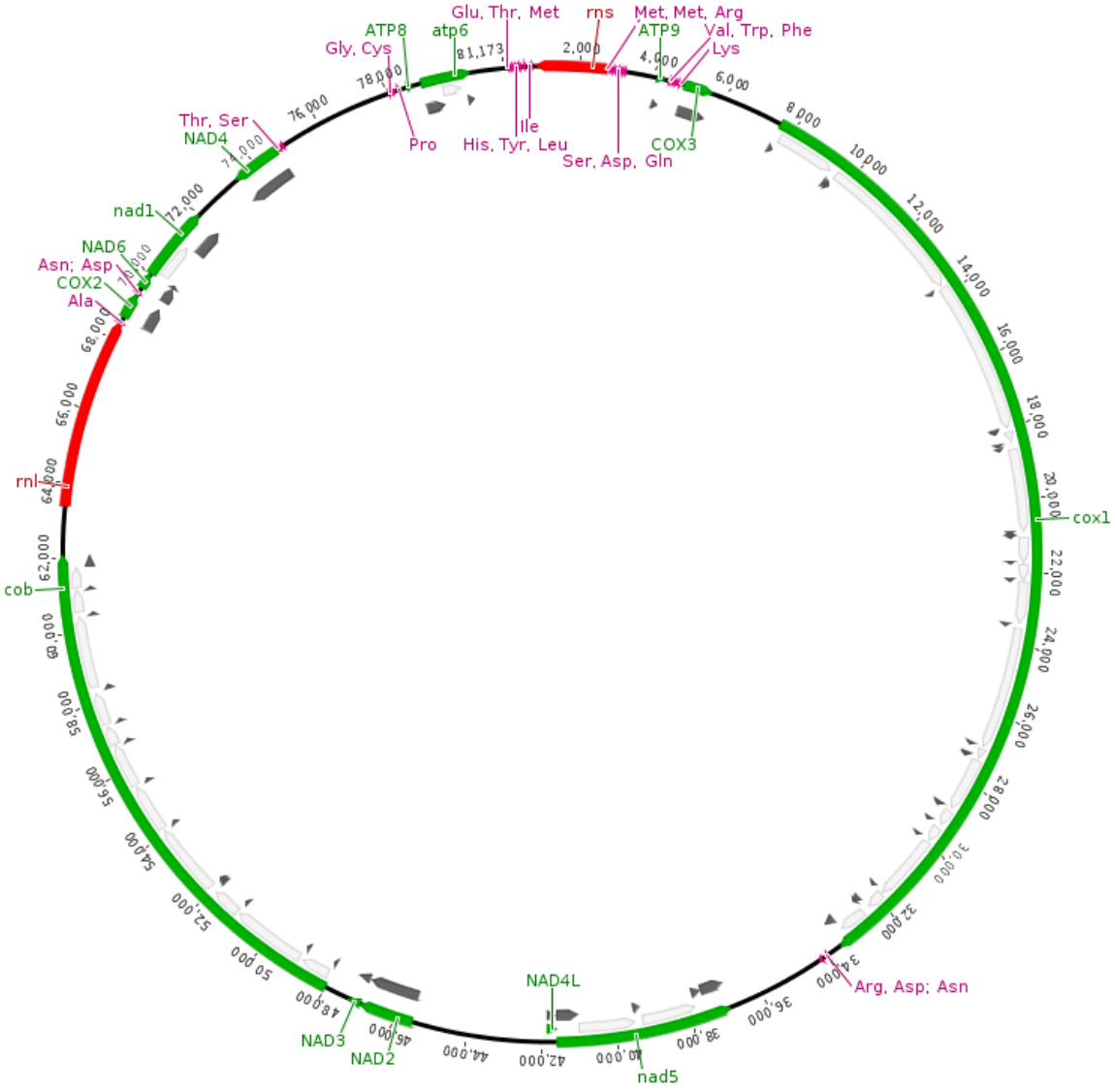

Figure S 70: Mitochondrial genome of *Metschnikowia* sp. UWOPS01-655c1. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

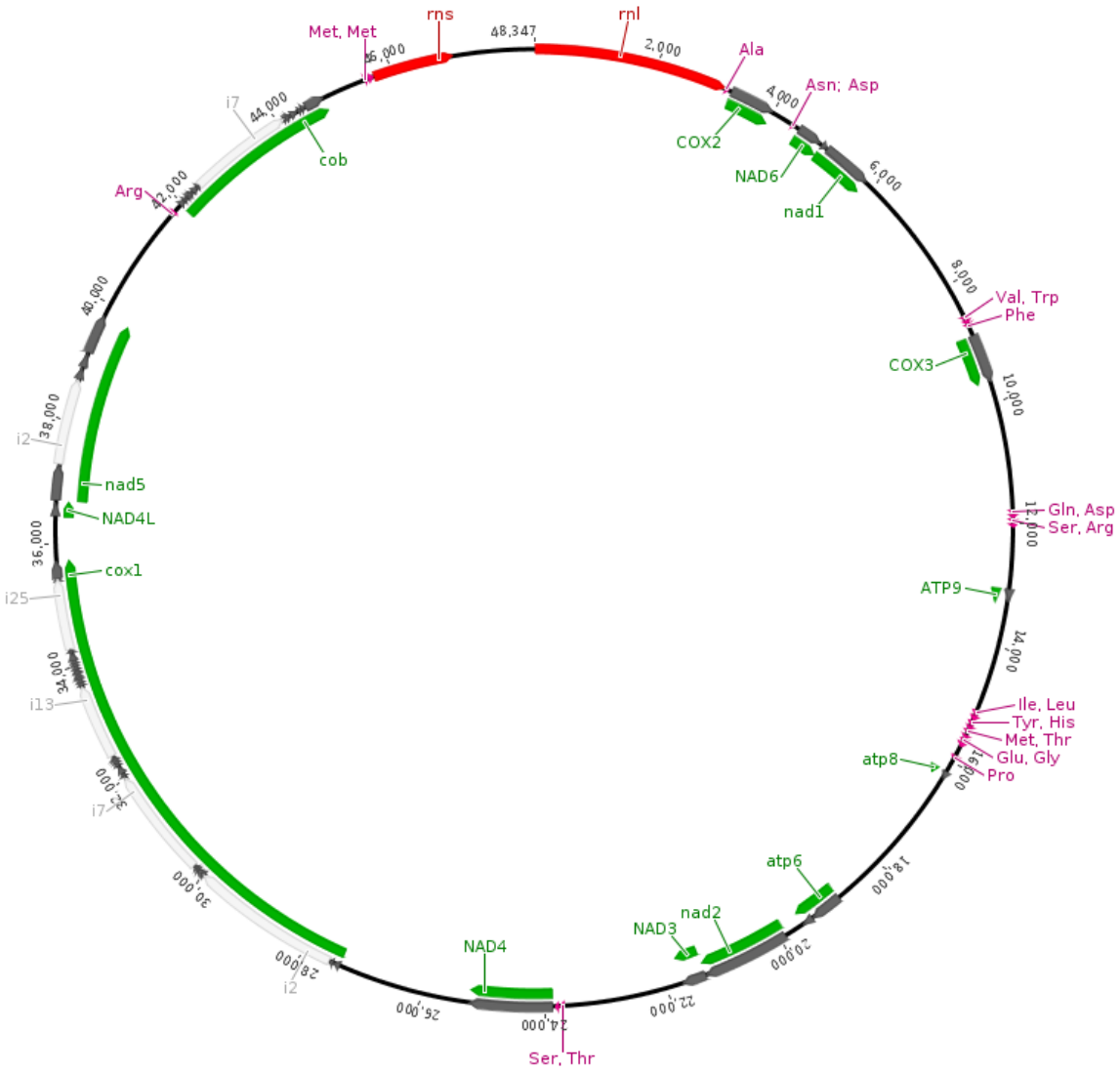

Figure S 71: Mitochondrial genome of *Metschnikowia torresii* CBS5152. Regions covered by green, black, grey, red and purple bars represent gene, exon, intron, rRNA and tRNA loci, respectively.

Appendix B: Additional supplementary figures and tables

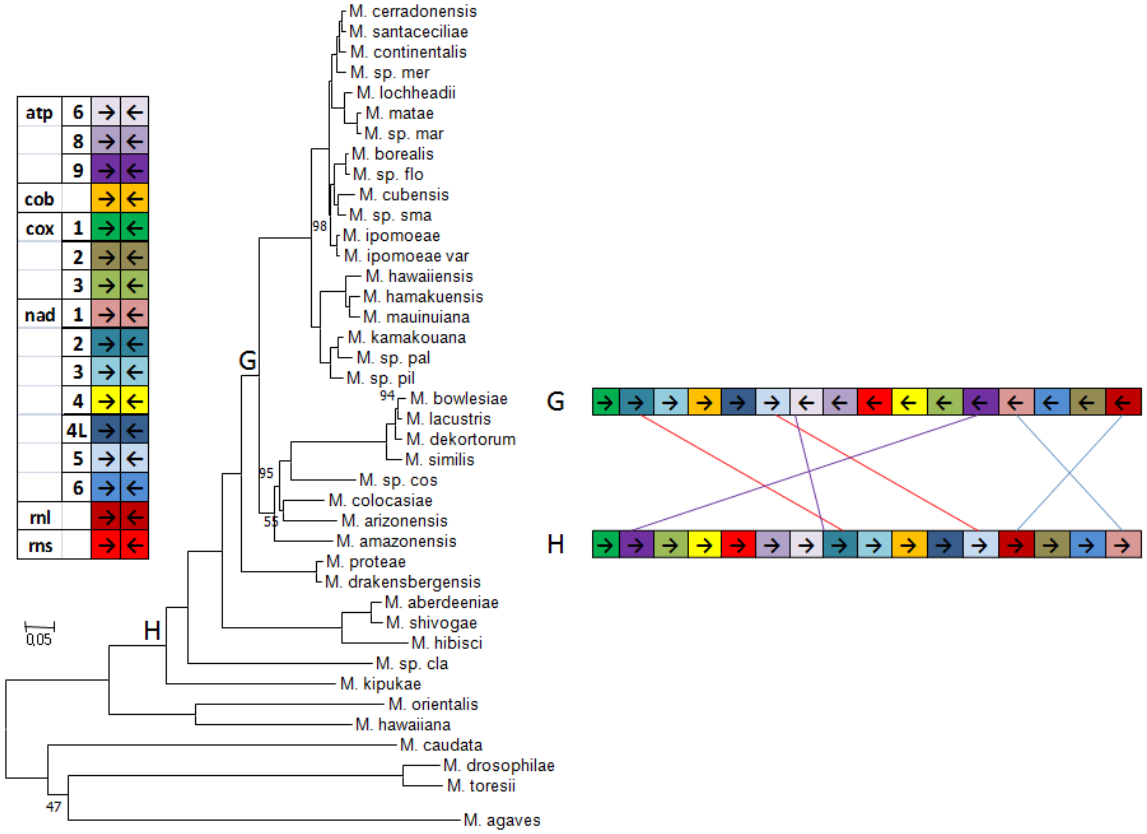

Figure S 72: Predicted gene order rearrangement between the reconstructed gene orders of the ancestor of the Arizonensis and continental subclades (G) and the ancestor of the early-emerging large-spored species (H). The figure was modified from Figure 3.

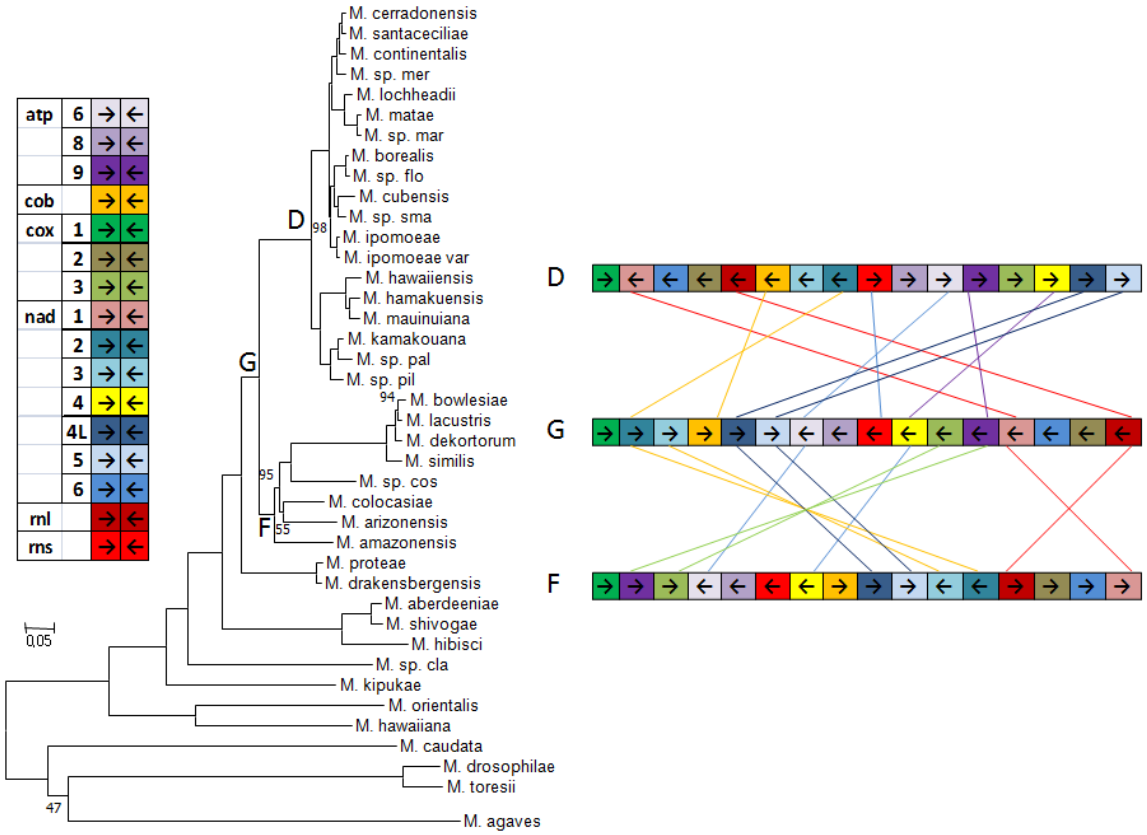

Figure S 73: Predicted gene order rearrangement from the common ancestor of the Arizonesis (F), and the *sensu stricto* (D) subclades to the common ancestor of early-emerging species (G). The figure was modified from Figure 3.

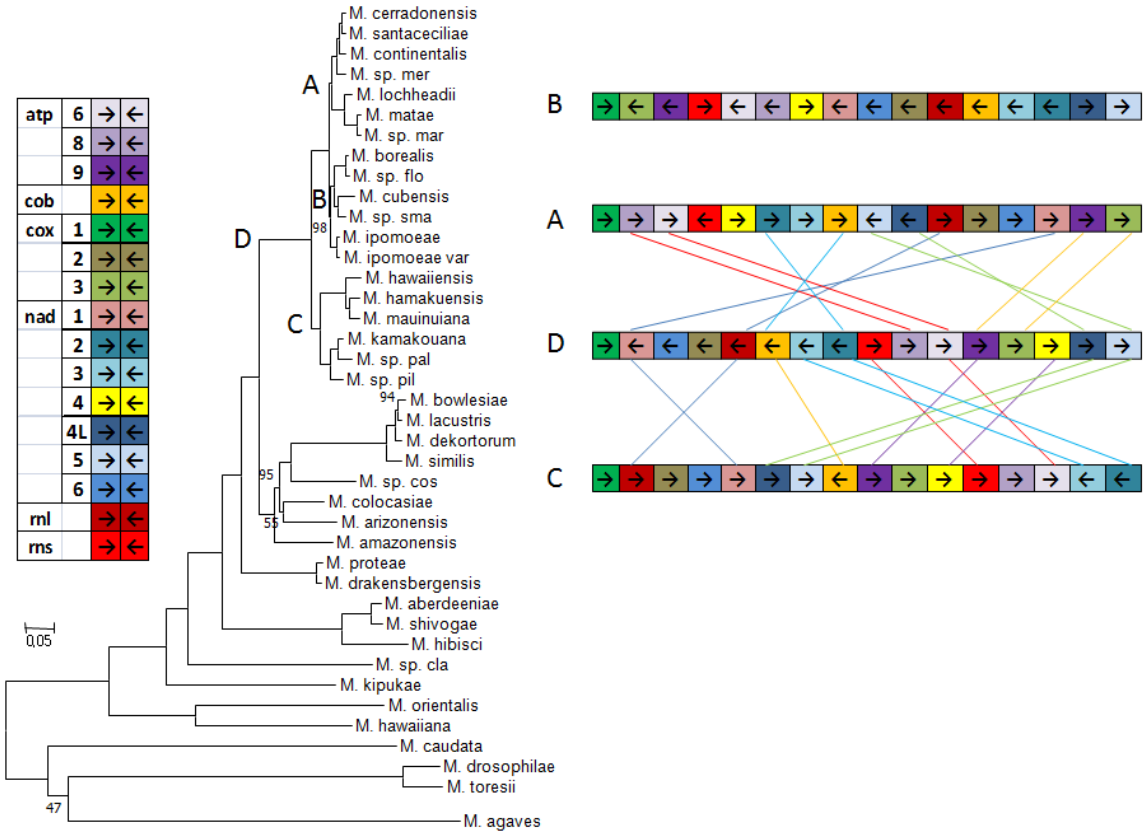

Figure S 74: Predicted gene order rearrangement from the ancestor of the *sensu stricto* (D) to the Hawaiian (C) and the upper branch of the Continental subclades (A). The figure was modified from Figure 3.

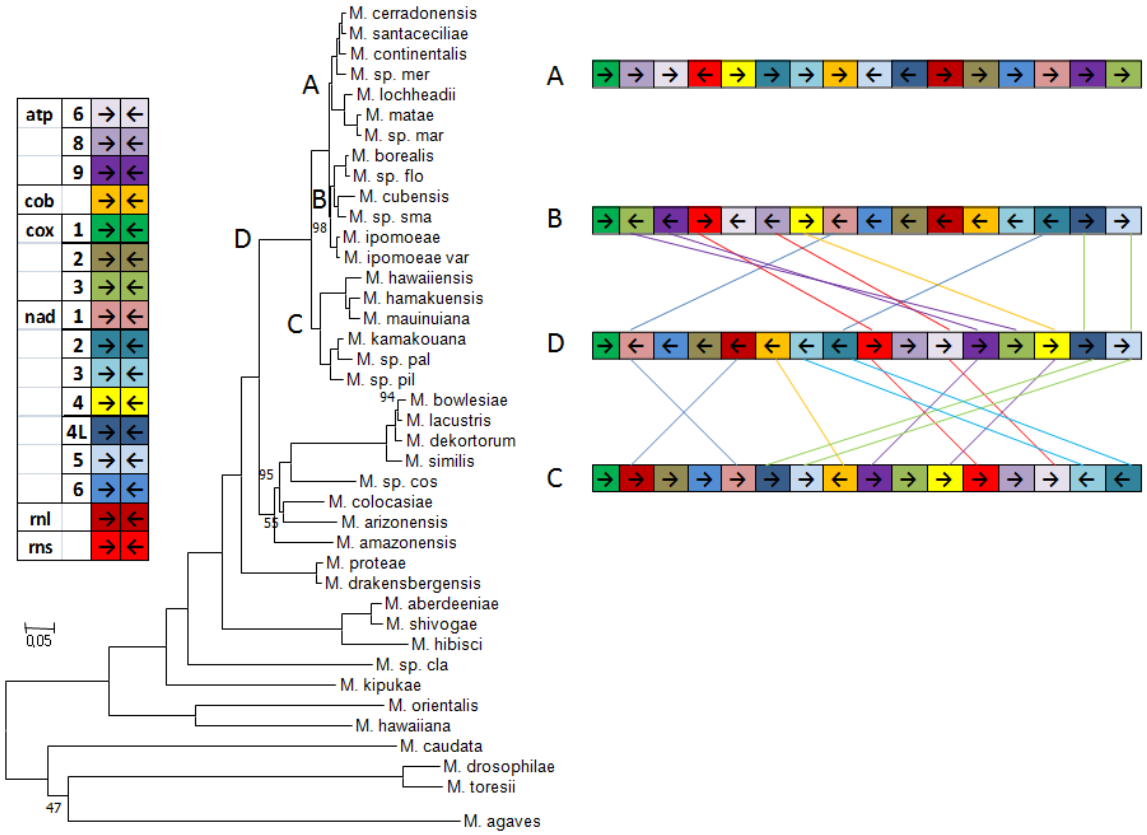

Figure S 75: Predicted gene order rearrangement from the ancestor of the *sensu stricto* (D) to the Hawaiian (C) and the lower branch of the Continental subclades (B). The figure was modified from Figure 3.

Table S 1: Primer sequences for PCR analysis.

Code	Direction	Sequence
cer+	f1	AGAACCAAAAGAAGAACGACGA
	r1	ACGATGATGAGTAGAAGCACCT
	f2	TCCTGCAGAAAGAGGTACATCT
	r2	TGAAGCTAATTCACGACGACT
cer-	f1	CGTCTTCCATTAAGAATACGTGATCT
	r1	ACGACCACGAAAATCCATTCT
cla+	f1	TCCCGAAGAGGATCATCCCC
	r1	CCGCGGGGATAATGCCATAA
col-	f1	GGTTTCGCTTTAGCAGAGGC
	r1	GCCTCCGGCGGGATTATTAA
	f2	TCCTATTAGGAACAGCCCCT
	r2	ACAAGATTCCCCATCGGCAA
col+	f1	TCTTAGATTGGATAACAGGGGGT
	r1	AACGGCCAACATCGCTATCA
con-	f1	CCCCCAGAAACCCCCATTTT
	r1	GACGGTTGTAGTCCCAAGCA
	f2	AGAACAATTGTATGTGGCGCA
	r2	AGAGGGCAAGCATTGATCAGA
	f3	AGGTGTACTCTAATCGCCAGG
	r3	TCAATGGTTAGAACAGACGCCT

Table S 2: Percent guanine-cytosine content, overall genome size (nt) and intergenic and intronic regions (%) of mitochondrial genomes of 71 *Metschnikowia* strains studied.

Metschnikowia strains	Code	%GC	Size	Intergenic	Intronic
M. aberdeeniae	abe+	28.4	70460	18.0	56.9
M. aberdeeniae	abe-	28.7	69221	19.7	54.8
M. agaves	aga+	24.7	61425	47.2	23.1
M. agaves	aga-	24.6	61641	47.6	22.8
M. amazonensis	ama+	23.3	132631	35.6	49.0
M. amazonensis	ama-	23.2	132165	35.4	49.2
M. arizonensis	ari+	23.3	187475	59.3	29.2
M. arizonensis	ari-	23.1	187024	59.4	29.2
M. borealis	bor+	23.6	96203	15.0	64.4
M. borealis	bor-	23.7	96587	14.9	64.3
M. bowlesiae	bow+	28.5	93037	27.1	52.1
M. bowlesiae	bow-a	28.2	93065	31.5	46.8
M. bowlesiae	bow-b	28.4	89822	28.4	50.7
M. caudata	cau+	24.3	24050	19.9	7.5
M. caudata	cau-	25.3	24995	19.6	10.5
M. cerradonensis	cer+	25.6	92261	24.2	54.3
M. cerradonensis	cer-	25.6	97238	27.5	52.0
Metschnikowia sp. M2Y3	cla+	24.4	52064	23.0	43.1
M. colocasiae	col+	26.1	62609	33.1	36.9
M. colocasiae	col-	25.4	64648	29.7	41.3
M. continentalis	con+	25.5	79731	19.8	55.3
M. continentalis	con-	25.8	79140	19.0	55.8
Metschnikowia sp. 03-147.1	cos-	26.2	63872	17.9	52.6
M. cubensis	cub+	23.6	116310	22.3	59.4
M. cubensis	cub-	23.6	116305	22.3	59.4
M. dekortorum	dek+	28.9	76385	24.9	51.7
M. dekortorum	dek-	28.9	79835	24.1	50.5
M. dekortorum	dekY	28.7	80983	26.5	49.6
M. drakensbergensis	dra+	22.7	109055	28.8	53.2
M. drakensbergensis	dra-	22.8	113475	27.7	55.0
M. drosophilae	dro+	23.6	33514	16.9	29.0
M. drosophilae	dro-	23.6	33571	17.1	29.0
Metschnikowia sp. 13-106.1	flo+	23.0	105181	13.8	66.7
M. hamakuensis	ham+	24.4	43542	14.9	53.5
M. hamakuensis	ham-	24.5	44350	14.0	55.2
M. hawaiiana	han+	29.4	35585	50.6	0.0

M. hawaiiensis	haw+	24.6	55026	22.2	43.2
M. hawaiiensis	haw-	24.9	51759	23.4	39.8
M. hibisci	hib+	24.1	67292	14.8	56.8
M. hibisci	hib-	24.0	66556	16.1	55.7
M. ipomoeae	ipo+	25.0	76936	24.5	49.9
M. ipomoeae	ipo-a	25.0	67552	15.3	55.4
M. ipomoeae	ipov	24.7	77864	14.8	60.0
M. kamakouana	kam+	26.0	52527	18.6	46.9
M. kamakouana	kam-	26.2	54921	16.0	49.6
M. kipukae	kip-	26.1	33454	22.5	24.6
M. lacustris	lac+	28.8	82226	24.1	51.3
M. lacustris	lac-	29.0	83905	23.2	52.6
M. lacustris	lacb	28.5	99039	26.6	52.8
M. lochheadii	loc+	25.7	89727	19.2	58.7
M. lochheadii	loc-	25.8	76321	20.4	53.6
M. matae var. *maris*	mar-	25.2	75407	21.1	53.7
M. matae var. *matae*	mat+	25.2	85788	21.4	55.4
M. matae var. *matae*	mat-	25.6	82136	22.8	52.9
M. mauinuiana	mau+	23.8	55476	21.7	53.0
M. mauinuiana	mau-	24.3	56148	20.0	55.1
Metschnikowia sp. 00-154.1	mer-	25.5	85847	25.9	51.1
Metschnikowia orientalis	ori+	25.9	56695	11.0	56.6
Metschnikowia orientalis	ori-	26.3	47030	14.0	47.1
Metschnikowia sp. 04-218.3	pal+	20.2	93173	39.1	35.3
Metschnikowia sp. 04-226.1	pil-	26.1	60297	28.1	33.5
M. proteae	pro+	22.4	120379	30.8	52.9
M. proteae	pro-	22.4	121495	30.5	53.3
M. santaceciliae	sce+	25.4	62160	12.8	55.4
M. santaceciliae	sce-	25.2	81062	25.0	50.6
M. shivogae	shi+	28.8	77238	20.4	56.6
M. shivogae	shi-	28.7	81198	20.6	57.6
M. similis	sim+	28.5	99200	26.2	53.1
M. similis	sim-	27.9	104926	27.2	53.8
Metschnikowia sp. 01-655c1	sma-	25.3	81173	22.3	53.5
M. torresii	tor-	24.8	48347	42.1	21.1

Table S 3: Number of introns (with or without ORFs) within *cox1* and *cob* genes.

***Metschnikowia* strains**	**Code**	***cox1***	***cob***
M. aberdeeniae	abe+	8	7
M. aberdeeniae	abe-	7	8
M. agaves	aga+	7	2
M. agaves	aga-	7	2
M. amazonensis	ama+	18	13
M. amazonensis	ama-	18	13
M. arizonensis	ari+	15	6
M. arizonensis	ari-	15	6
M. borealis	bor+	18	6
M. borealis	bor-	18	6
M. bowlesiae	bow+	16	10
M. bowlesiae	bow-a	15	10
M. bowlesiae	bow-b	15	10
M. caudata	cau+	1	0
M. caudata	cau-	1	0
M. cerradonensis	cer+	18	11
M. cerradonensis	cer-	18	11
Metschnikowia sp. M2Y3	cla+	8	3
M. colocasiae	col+	11	0
M. colocasiae	col-	11	6
M. continentalis	con+	17	10
M. continentalis	con-	16	11
Metschnikowia sp. 03-147.1	cos-	11	6
M. cubensis	cub+	17	8
M. cubensis	cub-	17	8
M. dekortorum	dek+	14	10
M. dekortorum	dek-	14	10
M. dekortorum	dekY	14	10
M. drakensbergensis	dra+	17	9
M. drakensbergensis	dra-	17	11
M. drosophilae	dro+	3	2
M. drosophilae	dro-	3	2
Metschnikowia sp. 13-106.1	flo+	20	9
M. hamakuensis	ham+	5	3
M. hamakuensis	ham-	5	3
M. hawaiiana	han+	0	0
M. hawaiiensis	haw+	8	5
M. hawaiiensis	haw-	7	5

M. hibisci	hib+	9	7
M. hibisci	hib-	9	6
M. ipomoeae	ipo+	16	8
M. ipomoeae	ipo-a	15	10
M. ipomoeae	ipov	17	11
M. kamakouana	kam+	8	5
M. kamakouana	kam-	9	6
M. kipukae	kip-	5	1
M. lacustris	lac+	11	10
M. lacustris	lac-	14	10
M. lacustris	lacb	15	10
M. lochheadii	loc+	14	9
M. lochheadii	loc-	16	10
M. matae var. *maris*	mar-	16	10
M. matae var. *matae*	mat+	18	10
M. matae var. *matae*	mat-	17	10
M. mauinuiana	mau+	6	4
M. mauinuiana	mau-	7	3
Metschnikowia sp. 00-154.1	mer-	16	11
Metschnikowia orientalis	ori+	11	6
Metschnikowia orientalis	ori-	7	4
Metschnikowia sp. 04-218.3	pal+	10	7
Metschnikowia sp. 04-226.1	pil-	5	3
M. proteae	pro+	18	11
M. proteae	pro-	19	11
M. santaceciliae	sce+	14	5
M. santaceciliae	sce-	15	11
M. shivogae	shi+	9	9
M. shivogae	shi-	11	8
M. similis	sim+	17	10
M. similis	sim-	16	11
Metschnikowia sp. 01-655c1	sma-	16	11
M. torresii	tor-	4	1

Curriculum Vitae

Name: Dong Kyung Lee

Post-secondary Education and Degrees:
The University of Western Ontario
London, Ontario, Canada
2013-2017 B.A.

Honours and Awards:
Province of Ontario Graduate Scholarship
2019

Related Work Experience
Research Assistant
The University of Western Ontario
2017

Teaching Assistant
The University of Western Ontario
2018-2021

Publications:

Lee, DK. Hsiang, T and Lachance, MA. 2018. *Metschnikowia* mating genomics. Antonie van Leeuwenhoek 111: 1935-1953

Gordon, Z. Soltysiak, MPM. Leichthammer, C. Therrien, JA. Meaney, RS. Lauzon, C. Adams, M. Lee, DK. Janakirama, P. Lachance, MA and Karas, BJ. 2019. Development of a transformation method for *Metschnikowia borealis* and other CUG-Serine yeasts. Genes 10: 78.

Santos, ARO. Lee, DK. Ferreira, AG. Carmo, MC. Rondelli, VM. Barros, KO. Hsiang, T. Rosa, CA and Lachance, MA. 2020. The yeast community of *Conotelus* sp. (*Coleoptera*: *Nitidulidae*) in Brazilian passionfruit flowers (*Passiflora edulis*) and description of *Metschnikowia amazonensis* sp. nov., a large-spored clade yeast. Yeast 37: 253-260.

Lee, DK. Santos, ARO. Hsiang, T. Rosa, CA and Lachance, MA. 2020. Catching speciation in the act-act 2: *Metschnikowia lacustris* sp. Nov., a sister species to *Metschnikowia dekortorum*. Antonie van Leeuwenhoek 113: 753-762.

Lee, DK. Hsiang, T. Lachance, MA and Smith, DR. 2020. The strange mitochondrial genomes of *Metschnikowia* yeasts. Curr Biol 30: 800-801.

Lachance, MA. Lee, DK and Hsiang, T. 2020. Delineating yeast species with genome average nucleotide identity: a calibration of ANI with haplontic, heterothallic *Metschnikowia* species. Antonie van Leeuwenhoek 113: 2097-2106.