MITOCHONDRIAL GENOME EVOLUTION IN THE DEEP-BRANCHING HETEROLOBOSEIDS AMOEBA 'BB2' AND *PHARYNGOMONAS KIRBYI*

by

Jiwon Yang

Submitted in partial fulfilment of the requirements
for the degree of Master of Science

at

Dalhousie University
Halifax, Nova Scotia
July 2016

Table of Contents

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

Studies of mitochondrial genomes from diverse eukaryotes provide insights into how mitochondria evolved. Previous research revealed several evolutionarily interesting mitochondrial genomes in the taxon Discoba (Excavata), including the most bacteria-like mitochondrial genomes of Jakobida and extensively fragmented mitochondrial genomes of Euglenozoa. I characterized the mitochondrial genomes of amoeba BB2 and *Pharyngomonas kirbyi,* which represent deep branches within Heterolobosea, the third main group of Discoba. Using phylogenomic analyses, I showed that BB2 and *P. kirbyi* are sister taxa at the very base of Heterolobosea. I assembled mitochondrial genomes of BB2 (119 kbp) and *P. kirbyi* (75 kbp) encoding 45 and 48 putative protein-coding genes, respectively. Interestingly, BB2 mitochondrial genes were extensively fragmented by frame-shifts. Comparative analysis of genomic and transcriptomic data revealed that mitochondrial transcripts of BB2 are heavily edited by mononucleotide insertions to produce functional RNAs. Bioinformatics analyses suggested that this RNA editing is very accurate and efficient, and possibly co-transcriptional.

LIST OF ABBREVIATIONS USED

A	Adenosine
AT	A+T
ATP	Adenosine triphosphate
ATCC	The american type culture collection
BIs	Bayesian inferences
bp	Base pair
BPP	Bayesian posterior probability
C	Cytidine
CNE	Constructive neutral evolution
DNA	Deoxyribonucleic acid
G	Guanosine
gRNA	Guide RNA
IR	Inverted repeat
kbp	Kilo base pair
LECA	Last eukaryotic common ancestor
LSU	Large subunit
Mbp	Mega base pair
MCMC	Markov chain Monte Carlo
ML	Maximum-likelihood
MLBP	ML-bootstrap
mRNA	Messenger RNA
mt	Mitochondrial

NTPs	Ribonucleoside triphosphates
ORFs	Open-reading frames
PPR	Pentatricopeptide repeat
RNA	Ribonucleic acid
RNAP	RNA polymerase
Rpo	RNA polymerase open promoter complex
rRNA	Ribosomal RNA
SSU	Small subunit
sp.	Species
Spc	Spectinomycin
tRNA	Transfer RNA
TUTase	Terminal uridylyl transferase
U	Uridine/ uracil
UFBoot	Ultrafast bootstrap
URFs	Unknown open reading frame

ACKNOWLEDGEMENTS

First of all, I am highly grateful to my supervisors, Dr. Andrew Roger and Dr. Alastair Simpson. Their expertise, support, understanding and generous guidance made it possible for me to successfully complete my research that was of great interest to me.

I would like to express my gratitude to Tommy Harding for showing great interest in my research, teaching me many lab techniques and providing valuable guidance throughout my time in the Roger lab. I also like to thank Dr. Ryoma Kamikawa for his kind advice regarding my research and for sharing his immense knowledge on the mitochondrial genomes. Many thanks to Michelle Leger, Laura Eme, Dayana Salas, Courtney Stairs, and Susan Sharpe for being highly supportive and creating a great place to work.

I would like to thank my supervisory committee, Dr. John Archibald and Dr. Claudio Slamovits, for insightful comments and encouragement. I also like to thank everyone in the Center for Comparative Genomics and Evolutionary Bioinformatics for the opportunity to hear inspiring talks. A special thanks goes to Dr. Mike Gray for sharing his knowledge on the mitochondrial RNA editing.

Finally, I would like to thank my parents and sisters for their unconditional love and support.

CHAPTER 1 Introduction

Mitochondria are crucial organelles that produce energy for survival and growth of eukaryotic cells. Over time, evolution gave rise to remarkable variation in mitochondrial genomes, and some mitochondria acquired new mechanisms involved in gene expression, such as RNA editing (Benne et al. 1986; Gray 2003). This chapter describes some evolutionarily interesting aspects of mitochondria; the origin and evolution of mitochondria and various examples of mitochondrial RNA editing that are phylogenetically and mechanistically unrelated.

1.1 Mitochondrial Evolution

It is now well recognized that mitochondria originated from an α-proteobacterium via endosymbiosis (Gray et al. 1999). Since this event, mitochondrial genomes have undergone various evolutionary changes, such as genome reduction during which genes were either transferred to the nucleus or completely lost (Adams and Palmer 2003). For example, the most bacteria-like (gene-rich) mitochondrial genome known to date, *Andalucia godoyi*, is highly reduced compared to known α-proteobacterial genomes including *Rickettsia prowazekii,* which has an atypically small genome for an α-proteobacterium (66 versus 834 protein-coding genes in *A.godoyi* and *R. prowazekii*, respectively) (Burger et al. 2013; Andersson et al. 1998). Mitochondrial genomes continued to change even after the diversification of major eukaryotic lineages, resulting in drastic variation in their genome organization, genome size and gene content in distant as well as in closely-related organisms (Gray et al. 1999; Kamikawa et al. 2014; Hajduk

et al. 1993; Marande and Burger 2007; Herman et al. 2013; Fu et al. 2010). The most extreme cases of genome reduction are in the mitochondrion-related organelles (e.g., mitosomes) of certain anaerobic eukaryotes, where complete genome loss has occurred concomitant with the loss of aerobic ATP synthesis in adaptation to their anaerobic environments (Tovar et al. 1999; Makiuchi and Nozaki 2014; Stairs et al. 2015). Recently, a complete loss of the mitochondrial organelle itself was reported for the anaerobic protistan eukaryote *Monocercomonoides* sp. (Karnkowska et al. 2016). Clearly, characterization and comparison of mitochondrial genomes and organelles from various diverse lineages of eukaryotes, particularly protists, is of key importance to understanding the evolutionary history of mitochondria.

Analyses of complete mitochondrial genome sequences revealed a great diversity in terms of structural complexity and size variation. In most organisms, the mitochondrial genome consists of a single circular-mapping molecule like typical bacterial genomes (Gray et al. 1998). In a few species, however, mitochondrial genomes exist as linear molecules. For example, the mitochondrial genomes of Medusozoa species, belonging to the phylum Cnidaria, contain one to several linear molecules (Kayal et al. 2012).

Even more complex mitochondrial genome structures were reported in the phylum Euglenozoa. Kinetoplastids possess two types of molecules: maxicircles encoding both protein-coding genes and guide RNAs (gRNAs), and minicircles encoding only gRNAs (the use of guide RNAs will be further discussed in the next section) (Hajduk et al. 1993). The mitochondrial genome of diplonemids also has unusual features such as gene fragmentation, with each non-overlapping piece encoded on a separate

circular chromosome: these are transcribed separately and assembled to an mRNA by trans-splicing (Kiethega et al. 2013; Marande and Burger 2007).

Mitochondrial genome sizes range from 6 kbp in parasites to more than 200 kbp in land plants (Anderson et al. 1981; Vaidya et al. 1989; Unseld et al. 1997; Sloan et al. 2012). So far, the smallest mitochondrial genome known is that of *Plasmodium falciparum* (the human malaria parasite), which is only 6 kbp (Vaidya et al. 1989). The largest mitochondrial genome known is the one of the flowering plant *Silene conica* that has a size of 11.3 Mbp and consists of 128 circular molecules (Sloan et al. 2012).

Mitochondrial genomes also vary according to gene content (Gray et al. 1998). In general, mitochondrial genomes contain approximately 40 to 50 genes (Burger et al. 2003). Two extreme examples for mitochondrial gene content are *Chromera velia* (a relative of the apicomplexan parasites) and the jakobid *Andalucia godoyi. Chromera* has the most reduced mitochondrial genome known, containing only two protein-coding genes (*cox1* and *cox3*), and the small subunit (SSU) and large subunit (LSU) ribosomal RNA (rRNA) genes (Flegontov et al. 2015). In contrast, *Andalucia godoyi* has the most gene-rich mitochondrial genome containing 100 genes (Burger et al. 2013). Other Jakobida species also have gene-rich mitochondrial genomes with 91-97 genes (Burger et al. 2013). The largest known mitochondrial genome, that of *Silene*, has only 25 protein-coding genes, 3 rRNAs (5S, 18S and 26S) and 2 transfer RNAs (tRNA) (Sloan et al. 2012). Although the *Andalucia* mitochondrial genome is more than 150 times smaller than the *Silene* mitochondrial genome, the former contains 4 times more genes; this is due to much smaller intergenic spaces in the *Andalucia* mitochondrial genome (<10%) compared to the *Silene* mitochondrial genome (99.3%) (Burger et al. 2013; Sloan et al.

2012). Along with the gene-richness, another interesting characteristic of jakobid mitochondrial genomes is that they have four genes for bacteria-like multi-subunit RNA polymerase (*rpo A-D)*, whereas the mitochondria of other eukaryotes transcribe their genes using nucleus-encoded single-subunit bacteriophage T3/T7-like RNA polymerases (Burger et al. 2013). Given that recent studies suggest that the root of the eukaryotes is not on the Jakobida branch (e.g., Derelle et al. 2015), it seems likely that LECA (the last eukaryotic common ancestor) had both types of RNA polymerase (Stechmann and Cavalier-Smith 2002).

Variation in gene content is observed even between close relatives. For example, the mitochondrial genome of the heteroloboseid *Acrasis kona* contains significantly fewer protein coding genes and transfer RNAs compared to its relatively close relatives in the genus *Naegleria* (26 versus 42 protein-coding genes and 11 versus 20 tRNAs) (Herman et al. 2013; Fu et al. 2014). Most of the genes missing in the *A. kona* mitochondrial genome have been identified in its nuclear genome, indicating that functional gene transfer from mitochondrial to nuclear genomes continued after the divergence of *A. kona* from its common ancestor with *Naegleria* (Fu et al. 2014).

1.2 RNA Editing in Mitochondria

As previously discussed, the evolution of mitochondrial genomes is highly dynamic. Another example of intriguing mitochondrial genome evolution is the independent development of RNA editing mechanisms in a diverse array of unrelated organisms. The term RNA editing was initially introduced to describe uridine (U) insertion and deletion from RNA transcribed from the mitochondrial genome of kinetoplastids, which was

discovered about 30 years ago (Benne et al. 1986; Horton and Landweber, 2002). Now, the term RNA editing is more generally used to describe targeted sequence alterations in RNAs that result in sequence differences between the mature RNAs and their DNA templates (Knoop 2011; Gray 2003). Two general types of RNA editing are nucleotide insertions/deletions (where nucleotides are added/removed from the DNA template), and substitutions (where nucleotides are replaced or modified), each of which can happen post-transcriptionally or co-transcriptionally (Horton and Landweber 2002; Cheng et al. 2001).

Table 1.1 Examples of mitochondrial RNA editing (see text for references).

Organism	Editing type	Edited RNAs
Kinetoplastids	U insertion/deletion	mRNAs
Myxomycota	Mono-/dinucleotide insertion, A deletion, substitution (C-to-U, U-to-G, C-to-G)	mRNAs, tRNAs, rRNAs
Land plants	Substitution (C-to-U, U-to-C)	mRNAs, tRNAs,
Dinoflagellates	N substitution	mRNAs, rRNAs

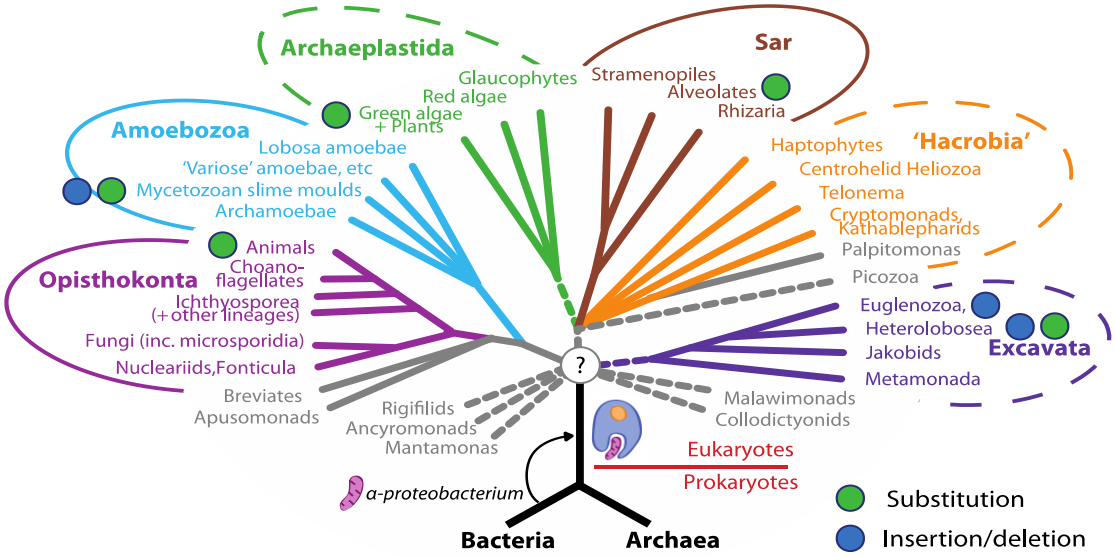


Figure 1.1 Tree of eukaryotes and the distribution of mitochondrial mRNA editing. Lineages where substitution and insertion/deletion-type RNA editing occur are indicated by green and blue circles, respectively (modified from Simpson AGB (unpublished)).

RNA editing has been reported in a broad diversity of organisms including plants, animals, protists and viruses (Knoop 2011) and it can affect the expression of mRNA, rRNA and tRNA encoded by mitochondrial, chloroplast and nuclear genomes (Gott and Emeson 2000). In mitochondria, various types of RNA editing are observed, including both insertion/deletion and nucleotide substitution, in widely distributed phylogenetic groups (Table 1.1, Figure 1.1).

U insertion/deletion editing in Kinetoplastids

RNA editing was first reported in the mitochondria of the kinetoplastid *Trypanosoma brucei* (Benne et al. 1986). In kinetoplastids, mitochondrial transcripts, encoded on maxicircles, are extensively altered by uridine (U) insertions and deletions, with the most extensive case of 553 insertions and 89 deletions in a single mRNA (Alfonzo et al. 1997). Editing involves small antisense RNAs (guide RNAs, gRNAs), encoded on both maxicircles and minicircles, and 20S multi-subunit protein complexes called 'editosomes' (Aphasizhev et al. 2003; Alfonzo et al. 1997; Knoop 2011).

For U insertion/deletion editing, the gRNA specifies the site of editing and the number of Us to be inserted by binding to the pre-edited mRNA by base-pairing. This is followed by three main biochemical activities: an endonuclease cleaves the pre-mRNA at the mismatch, a terminal uridylyl transferase (TUTase) adds U to the pre-mRNA for U insertion (or an exonuclease removes U from the pre-mRNA for U deletion), and a ligase rejoins the transcript ends. The fully edited mRNA is completely complementary to the gRNA in this region (Horton and Landweber 2002; Estevez et al. 1999; Madison-Antenucci et al. 2002; Knoop 2011).

Multiple editing mechanisms in Myxomycetes

Several types of editing are observed in the mitochondrial transcripts of the slime mould *Physarum polycephalum* and other myxomycete species (Traphagen et al. 2010). The mitochondrial transcripts of *Physarum* are extensively edited, with mRNAs being edited at approximately 4% of their nucleotides and tRNAs and rRNAs at approximately 2% of the nucleotides (over 1300 sites in total). The most frequently observed editing type is insertion of a non-encoded nucleotide, mostly C (94.2% of edits) and U (3.2%). Insertions of other nucleotides (2.0%) are also found (G, A, UC/CU, AA, UG/GU, UU, UA, GC/CG) (Bundschuh et al. 2011). In addition, rare cases of deletions and substitutions are reported (Bundschuh et al. 2011). Deletion of three adjacent As is observed in *Physarum* mitochondrial *nad2* mRNA. Four C-to-U substitutions are found in *cox1* mRNA, and two substitutions to G (U-to-G and C-to-G) are found at 5′ ends of tRNAs (Bundschuh et al. 2011).

Insertion editing in *Physarum* is efficient and accurate, with mis-editing occurring at a low rate (approximately 5%) (Byrne et al. 2002; Visomirski-Robic and Gott 1995). In an *in vitro* assay of isolated mitochondria, RNA transcripts were efficiently edited at insertion sites but unedited at substitution sites. This suggests that insertional and substitutional editing likely use different RNA editing mechanisms: Insertional editing occurs co-transcriptionally since the RNA transcripts are mostly fully edited, while substitutional editing happens post-transcriptionally (Visomirski-Robic and Gott 1995; Bundschuh et al. 2011).

In another study, a series of chimeric templates were generated through manipulation of the DNA template (mtDNA was digested using restriction enzymes that

cleaved at varying distances from the sites of C insertion, and were then ligated to one another or to exogenous DNA fragments) (Rhee et al. 2009). The run-on transcripts generated with these chimeric templates showed that the critical information for editing is encoded within 9-10 nucleotides downstream and upstream of the edit sites – the so-called 'critical region' (Rhee et al. 2009). Changes outside of this critical region did not affect the accuracy or efficiency of editing (Rhee et al. 2009). In addition, the editing

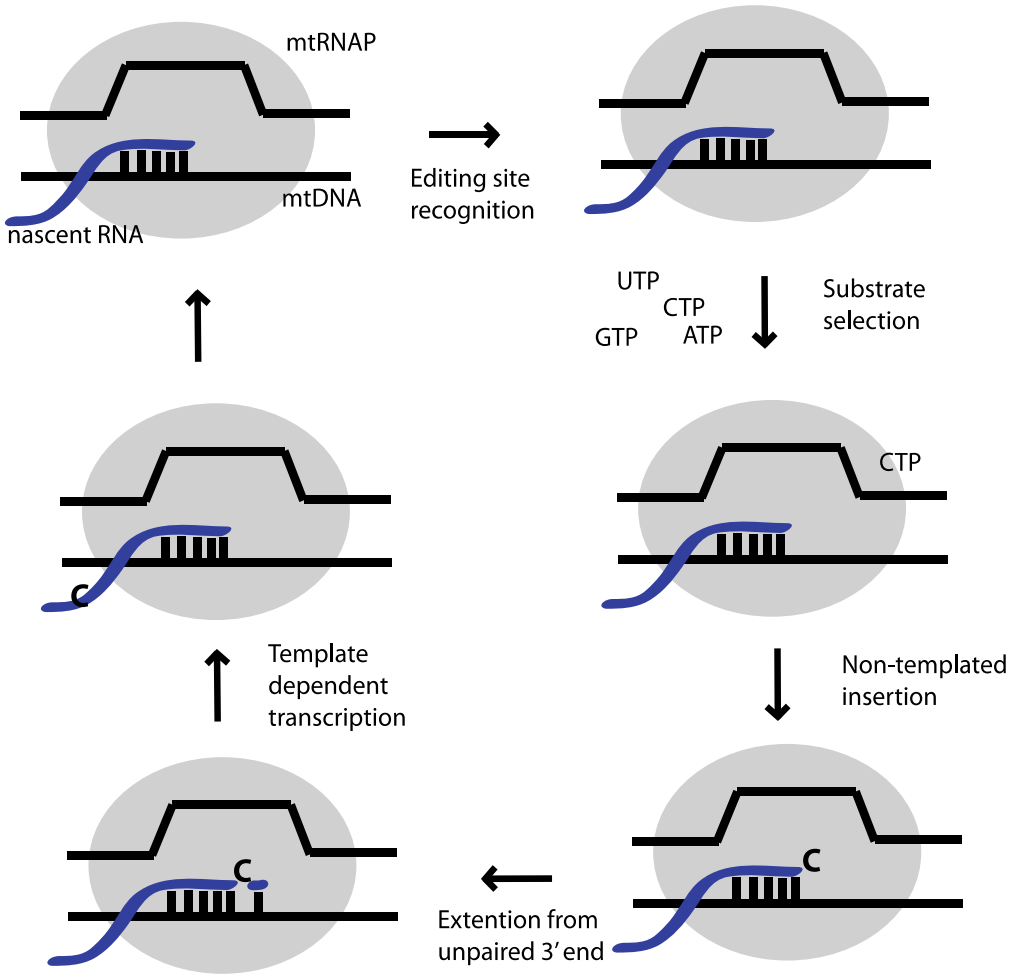


Figure 1.2 Co-transcriptional insertion-type editing in the mitochondria of *Physarum* (modified from Rhee et al. 2009).

pattern observed in run-on transcripts suggests that information 9 nucleotides upstream of the edit sites likely affects nucleotide selection and insertion, and information 9-10 nucleotides downstream likely affects editing site recognition and extension (Rhee et al. 2009). Rhee et al (2009) proposed a model for co-transcriptional insertion editing in *Physarum* mitochondria (Figure 1.2). Their model suggests that the template-dependent transcription by the mitochondrial RNA polymerase (mtRNAP) is interrupted as it reaches the 'critical region', and this interruption allows the selection and insertion of the non-templated nucleotide at the insertion site. The mechanisms for these steps are currently unknown. After insertion, the template-dependent transcription continues.

C-to-U and U-to-C editing in Plants

In plants, RNA editing post-transcriptionally modifies mitochondrial transcripts from C-to-U and U-to-C (Shikanai 2006). In *Arabidopsis thaliana*, 441 editing sites have been identified in mitochondrial transcripts, affecting 31 out of 34 protein-coding genes and a few tRNAs (Giege and Brennicke 1999). Most editing occurs at first or second codon positions, resulting in alteration in identity of the encoded amino acid (Gray and Covello 1993). Editing is addressed by a specific protein with pentatricopeptide repeat (PPR proteins) that binds upstream of editing sites and facilitates the access of an unknown enzyme for cytosine deamination in order to convert C to U.

This C-to-U or U-to-C editing is found in all reported mitochondrial genomes of land plants, however, it has not been detected in the green algae, the close relatives of land plants (Hiesel et al. 1994). This suggests that this editing likely emerged in a

common ancestor of land plants, after the separation from the green algal lineages (Gray 2003; Hiesel et al. 1994).

Mixed substitution editing in Dinoflagellates

RNA editing is found in most of known dinoflagellate mitochondria, but is absent in the other alveolates (e.g., ciliates and apicomplexans) and in the basal lineages of dinoflagellates (Waller and Jackson 2009). Dinoflagellates have highly reduced and fragmented mitochondrial genomes. In their mitochondria, both mRNAs and rRNAs are extensively edited by substitution (Lin et al. 2002; Jackson et al. 2007). In the dinoflagellate *Karlodinium micrum,* some transcripts such as cytochrome oxidase subunit 3 (*cox3*) are edited extremely heavily, at a rate of one substitution per 17 nucleotides (6% of the ORF) (Jackson et al. 2007). Different types of substitutions have been reported (A-to-G, C-to-U, U-to-C, G-to-A, U-to-G, G-to-U, U-to-A, A-to-C and G-to-C), with the most frequent being A-to-G changes (Jackson et al. 2007). Most substitutions were observed at first or second codon positions, leading to changes in encoded amino acids (Jackson et al. 2007). The RNA editing machinery is still unknown.

1.3 Evolutionary Origins of RNA Editing

Since RNA editing has multiple origins, the specifics of how each type evolved will be different, but some general principles may apply. Constructive neutral evolution (CNE) is one of the general models proposed to describe the origin of complex cellular processes such as RNA editing (Covello and Gray 1993; Gray 2012; Doolittle et al. 2011). According to CNE, the cell originally did not require RNA editing to produce functional

RNAs, and genes encoding RNAs with mutations that made large alterations in the protein (either frameshifts or missense mutations) would be removed by purifying selection. However, the presence of an RNA editing system, which emerged from the activity of pre-existing proteins in a neutral way (i.e., a side-reaction of a particular enzyme), could suppress the effect of the mutation and allowed mutations to accumulate. RNA editing could disappear without any harm to the cell if the number of mutations is low. If, on the other hand, the number of mutations were high, RNA editing would become an essential part of the gene expression pathway that could not be lost because it would be required for the generation of multiple functional RNAs (Gray 2012; Doolittle et al. 2011).

1.4 Aim of This Thesis

The overall aim of this thesis is to extend our knowledge of mitochondrial genome evolution. In the following chapter, I aim to determine if the mitochondrial genomes of amoeba 'BB2' and *Pharyngomonas kirbyi,* two deep-branching members of the taxon Heterolobosea, are as "ancestral" (gene-rich) as in Jakobida, to which they are related. Further, I aim to understand the features of a novel RNA editing system that I have discovered in the mitochondria of BB2.

CHAPTER 2 Mitochondrial Genome Evolution and RNA Editing in Deep-Branching heteroloboseids.

This chapter will be submitted for publication as Yang J, Harding T, Kamikawa R, Simpson AGB, Roger AJ (2016).

2.1 INTRODUCTION

Mitochondria are organelles that originated from α-proteobacteria by endosymbiosis and contain their own DNA and transcription/translation machinery (Gray et al. 1999). This ancient endosymbiosis was followed by massive gene loss in the ancestral mitochondrial (mt) genome by deletion of unnecessary genes and gene transfer from the mitochondrial genome to the nucleus (Gray et al. 1999). Mitochondrial gene content has continued to decrease in some eukaryote groups, giving rise to the huge variety among mitochondrial genomes we now observe (Lang et al. 1997; Kamikawa et al. 2016).

An intriguing process that has emerged several times in mitochondria is RNA editing. The phenomenon of RNA editing is defined as targeted sequence modifications to the RNAs that result in sequence differences between the transcriptome and the corresponding genomic sequences (Knoop 2010; Gray 2003). Two general types of RNA editing are known, namely nucleotide insertions/deletions and nucleotide substitutions, and both coding RNAs (mRNAs) and non-coding RNAs (e.g., rRNAs and tRNAs) can be affected (Horton and Ladweber 2002; Gray 2003; Gott and Emeson 2000). RNA editing has been found in a wide range of organisms (Knoop 2011, Chaterigner-Boutin and Small 2011).

Discoba (Excavata) is a major (‘kingdom-level’) group of protistan eukaryotes

that includes species with some of the most extraordinary mitochondrial genomes known (Simpson et al. 2006; Gray 2004; Hampl et al. 2009). Discoba comprises three main subgroups: Jakobida, Euglenozoa, and Heterolobosea, plus the isolated genus *Tsukubamonas* (Hampl et al. 2009; Kamikawa et al. 2014). Jakobida (e.g., *Reclinomonas, Andalucia*) have the most bacterial-like (ancestral) and gene-rich mitochondrial genomes known (Lang et al. 1997; Burger et al. 2013). For example, the mitochondrial genome of *Andalucia godoyi* contains 66 protein-coding and 34 structural RNA genes. Jakobid mitochondrial genome encode multiple subunits (almost always four) of bacteria-like RNA polymerase, whereas all other eukaryotes possess instead a nucleus-encoded, single-subunit enzyme homologous to bacteriophage RNA polymerases (Burger et al. 2013).

The mitochondria of Euglenozoa also have unusual features including gene fragmentation, and extensive RNA editing of mitochondrial transcripts (Flegontov et al. 2011). The mitochondrial genome of the model euglenid *Euglena gracilis* encodes fragmented ribosomal RNAs (Spencer and Gray 2011; Dobakova et al. 2015). The mitochondrial genome of diplonemids contains fragmented genes, with each non-overlapping piece encoded on a separate circular chromosome: these are transcribed separately and assembled to a mRNA by trans-splicing (Kiethega et al. 2013; Marande and Burger 2007). Most spectacularly, many pre-mRNAs of kinetoplastid mitochondria are massively edited post-transcriptionally by U-insertion/deletion to produce functional RNAs (Benne et al. 1986; Horton and Landweber 2002; Lukes et al. 2005). The U-insertion/deletion is mediated by small antisense RNAs (guide RNAs), which specify editing sites, and 20S multi-subunit protein complexes called ‘editosomes’ (Knoop 2011).

While the mitochondrial genomes of Jakobida and Euglenozoa have been intensively studied, only three mitochondrial genomes from the third main group of Discoba, Heterolobosea, have been characterized: those from *Naegleria gruberi* (Fritz-Laylin et al. 2011), *Naegleria fowleri* (Herman et al. 2013) and *Acrasis kona* (Fu et al. 2014). The *N. gruberi* and *N. fowleri* mitochondrial genomes have 42 protein-coding and 21 structural RNA genes. In contrast, the mitochondrial genome of *A. kona* contains only 26 protein-coding genes and 13 structural RNA genes. In addition, C-to-U type RNA editing was reported in mitochondrial transcripts of *N. gruberi* and *A. kona,* along with the presence of DYW-type pentatrico-peptide repeat (PPR) protein, previously only found in land plants (Knoop and Rüdinger 2010 Fu et al. 2014). In land plants, the PPR protein recognizes and binds to specific C residues for RNA editing (Yagi et al. 2013). Fu et al. (2014) suggested that these DYW-type PPR proteins in *N. gruberi* and *A. kona* were acquired by multiple independent lateral gene transfer events and were thus not ancestral to the Heterolobosea.

Here, we characterize the mitochondrial genomes of the undescribed amoeba 'BB2' and *Pharyngomonas kirbyi*, two putatively early-diverging species within Heterolobosea (Park and Simpson 2011; Harding et al. 2013). Our phylogenomic analyses demonstrated they form a single clade that emerges at the base of the Heterolobosea, and we describe the dynamics of mitochondrial genome evolution in this group in light of this newly resolved phylogeny. Unexpectedly, we found that a form of insertional RNA editing is very widespread in BB2 mitochondria, occurring at nearly 500 positions, affecting all protein-coding and rRNA genes, and half of the encoded tRNAs. This phenomenon is clearly different from the forms of mitochondrial RNA editing

documented previously in Discobid mitochondria, and, presumably, evolved independently.

2.2 MATERIAL AND METHODS

2.2.1 Transcriptomic sequencing of BB2

Amoeba BB2 strain PRA-19 was obtained from the American Type Culture Collection (ATCC) and grown at 42°C in ATCC medium 1034 (modified PYNFH medium: Bacto-peptone 10.0g/L, yeast extract 10.0 g/L, yeast nucleic acid 1.0 g/L, folic acid 15.0 mg/L, hemin 1.0 mg/L, fetal bovine serum 10%, KH_2PO_4 0.36 g/L, Na_2HPO_4 0.5 g/L). *Pharyngomonas kirbyi* strain AS12B was grown at 12.5% salt at 37°C as described in Harding et al., 2016. Total RNA was isolated from BB2 cells harvested using TRIzol (Rio et al. 2010) following the manufacturer's instructions (Ambion), and treated with Turbo DNAse (Ambion) to remove residual DNA.

For sequencing of the BB2 transcriptome, a cDNA library was constructed using the TruSeq RNA sample preparation kit version 2 (Illumina) and sequenced on a MiSeq platform, generating 19,206,268 150-bp paired-end reads. The high-quality reads were assembled using Trinity 2.0.2 (Grabherr et al. 2011), and open-reading frames (ORFs) were predicted using TransDecoder included in the Trinity package.

Mitochondrial ORFs missing from the assembled transcripts but present in genomic sequences (see below) were sequenced by RT-PCR. First-strand cDNA was synthesized from DNase-treated total RNA using a RevertAid H Minus First Strand cDNA Synthesis Kit (Thermo) with random hexamer primers, following the manufacturer's instructions (Thermo).

2.2.2 Phylogenetic Analysis

In order to clarify the phylogenetic position of our study organisms, we modified a curated 'phylogenomic' dataset containing 252 house-keeping proteins from a broad range of eukaryotes, described in Brown et al. (2012) and Harding et al. (2016). To the original dataset, we added sequences from the BB2 transcriptomic data, and well as a *Pharyngomonas kirbyi* transcriptome (data from Harding et al. 2016; GECH01000000), plus six other Excavates: *Spironucleus vortens, Tritrichomonas foetus, Diplonema papillatum, Leishmania major, Seculamonas ecuadoriensis, Jakoba bahamensis* (all available in GenBank (http://www.ncbi.nlm.nih.gov/, last accessed February 10, 2016) and *Stygiella incarcerata* (data from Leger et al. 2016).

Orthologous protein sequences were aligned by MAFFT-linsi (Katoh et al. 2005), and any sites in the alignments with more than 40% gaps were masked using BMGE v1.1 (Criscuolo and Gribaldo, 2010). Single protein trees were generated using RAxML v7.2.6 with the PROTGAMMALG model, and manually examined to remove putative contaminants, paralogs or laterally transferred genes. Sequences of remaining proteins were concatenated into a super-matrix containing 67 taxa and 68,718 amino acid positions.

Maximum-likelihood (ML) trees were estimated using IQ-TREE (Nguyen et al, 2014) under the LG+C20+F+gamma model. Topological support was assessed by 1000 ultrafast bootstrap (UFBoot) replicates and the SH-like approximate likelihood ratio test with 1000 replicates. ML trees were also estimated using RAxML under the LG4X model (from 100 starting trees), with topological support assessed by 100 bootstrap replicates.

Bayesian inferences (BIs) were conducted using PhyloBayes-MPI v. 1.6.5 under

the CAT-Poisson model. Five independent Markov chain Monte Carlo (MCMC) chains were run for 5,000 generations, sampling every two generations. 500 generations were discarded as burn-in. Convergence was achieved for three of the chains, with the largest discrepancy observed across all bipartitions (maxdiff) less than 0.26.

2.2.3 Genomic DNA sequencing of BB2 and P.kirbyi

Total DNA was extracted from BB2 and *P. kirbyi* using a salt-based separation method (Aljanabi and Martinez, 1997). First, cells were disrupted by vortexing in lysis buffer (50 mM) and digested with proteinase K (0.2 mg/mL) and 0.01% sodium dodecyl sulfate. DNA was separated from the other organic phases by centrifugation in a 3 M NaCl solution and precipitated with 70% ethanol.

For sequencing of genomic DNA, DNA libraries were prepared using the Nextera XT DNA sample preparation kit (Illumina). Sequencing was done using the MiSeq platform, yielding 43,376,820 150-bp paired-end reads for BB2 and 31,606,349 250-bp paired-end reads for *P. kirbyi*. Reads were trimmed to remove adapter sequences and low-quality sequences using Trimmomatic-0.30 (Bolger et al. 2014). Reads generated from *P. kirbyi* were filtered to remove sequences derived from food prokaryotes as described in Harding et al. (2016).

Genomic contigs for BB2 and *P. kirbyi* were assembled with the *de novo* assemblers Ray v2.3.1 (Boisvert et al. 2010) and/or MIRA v4.9.5_2 (Chevreux et al. 2004). For *P. kirbyi*, a second round of decontamination was performed by using assembled contig sequences as queries in BLASTn searches against the NT database. Contigs >100 bp showing more than 90% identity to a prokaryotic sequence were discarded as potential contaminants.

2.2.4 Mitochondrial Genome Assembly and Annotation

The genomic contigs from both species were screened for regions homologous to the mitochondrial genomes of other heteroloboseids (*Naegleria gruberi* (NC_002573)*, Naegleria fowleri* (NC_021104)*, Acrasis kona* (NC_026286) as well as the jakobid *Reclinomonas americana* (NC_001823) and *Tsukubamonas globosa* (NC_023545)) using BLASTn and BLASTx. For BB2, eight contigs with sizes ranging from 6kb to 18kb were highly similar to mitochondrial-derived sequences (identities >25%). These contigs were linked together into a circular-mapping mitochondrial genome with a size of 119,312 bp after PCR amplification of 'bridging' fragments using the LongAmp Taq PCR kit (NEB) and combinations of specific primers (supplementary table S1, figure S1). These amplicons were Sanger-sequenced. For *P. kirbyi,* two contigs (sizes 55 kbp and 19 kbp) were linked into one linear 75,717 bp scaffold by the same approach (supplementary table S1, figure S1).

Annotation was done using the automated gene annotation tools, MFannot (http://megasun.bch.umontreal.ca/cgi-bin/mfannot/mfannotInterface.pl, last accessed April 26, 2016) and RNAweasel (http://megasun.bch. umontreal.ca/RNAweasel/, last accessed April 26, 2016), and BLASTp searches against the NR database (NCBI) with a E-value cutoff of $1e^{-10}$. Transfer RNA (tRNA) genes were confirmed using tRNAscan-SE v1.23 (Lowe and Eddy 1997).

For gene prediction in the BB2 mitochondrial genome, the assembled transcripts (from the transcriptome data) were aligned to the mitochondrial genome in order to compare the sequences. Mitochondrial ORFs missing from the assembled transcripts but present in genomic sequences were sequenced by RT-PCR. First-strand cDNA was

synthesized from DNase-treated total RNA using RevertAid H Minus First Strand cDNA Synthesis Kit (Thermo) with random hexamer primers, following the manufacturer's instructions (Thermo).

The secondary structure of tRNAs were predicted using tRNAscan-SE v1.23. The secondary structure of the SSU rRNA was predicted and adjusted manually according to the secondary structure conserved among bacteria, archaea, eukaryotes, chloroplasts and mitochondria (as complied by the Gutell lab at the Comparative RNA Web Site and Project; http://www.rna.icmb.utexas.edu, last accessed May 18, 2016), and generated using the xrna program (http://rna.ucsc.edu/rnacenter/xrna/, last accessed May 18, 2016).

Genome maps were illustrated using GenomeVx (Conant and Wolfe 2007) followed by manual adjustment. Mitochondrial gene content was compared amongst eukaryotes from diverse lineages by extending the analyses of Kamikawa et al. (2016).

2.2.5 Confirmation of RNA Editing in BB2 and Error-Rate Estimation

The BB2 mitochondrial genomic sequences contained highly fragmented genes with numerous apparent reading frame-shifts. Comparison of transcript sequences to the genome indicated the presence of non-encoded nucleotides (insertions) in the former. A subset of these RNA editing sites (50/475 sites) was confirmed by sequencing 8 distinct PCR products using cDNA as template (supplementary table S2).

To characterize the efficiency and fidelity of the RNA editing mechanism, transcriptomic-sequencing reads were aligned to the mature transcript sequences using BLASTn (NCBI). We used 10-nucleotide-long sliding windows along the transcript sequences to examine the frequency and types of mismatches (insertion, deletion or substitution in RNA-derived sequencing reads compared to the genomic sequence) near

editing sites and remote from editing sites. We then performed Z-tests to determine if the varying types of error rates were significantly different between editing and non-editing sites. The Z-scores for each error-type were calculated as follows:

$$Z = \frac{p_1 - p_0}{\sqrt{\frac{p(1-p)}{n_1} + \frac{p(1-p)}{n_0}}}$$

where p_1 is the frequency of the mismatches near editing sites, p_0 is the frequency of mismatches remote from editing sites, p is the frequency of mismatches for all sites, while n_1 and n_0 are the total numbers of nucleotides near editing sites and remote from editing sites, respectively. The p-value for the null hypothesis that p_1 and p_0 are equal was determined from the Z-score based on the standard Normal distribution.

2.3 RESULTS

2.3.1 Phylogenetic Positions of 'BB2' and P. kirbyi

Both ML and BI analyses showed that BB2 and *P. kirbyi* robustly group with other heterolobosеids with strong statistical support (100% ML-bootstrap (MLBP), 100% ML-ultrafast bootstrap (UFBoot) and Bayesian posterior probability (BPP) of 1; figure 2.1). Interestingly, BB2 and *P. kirbyi* were inferred to be sister taxa at the base of this group in both ML and BI analyses, with 100% MLBP, 100% UFBoot and BPP of 1.0 (figure 2.1). The remaining heterolobosеids (*Percolomonas cosmopolitus, N. gruberi, Sawyeria marylandensis* and *Stachyamoeba lipophora)* formed a well-supported group within the Heterolobosea clade to the exclusion of BB2 and *P. kirbyi* (100% MLBP, 100% UFBoot and BPP of 1.0). Therefore, our analyses clearly indicated that BB2 and *P. kirbyi* group together in a clade that emerges at the base of Heterolobosea.

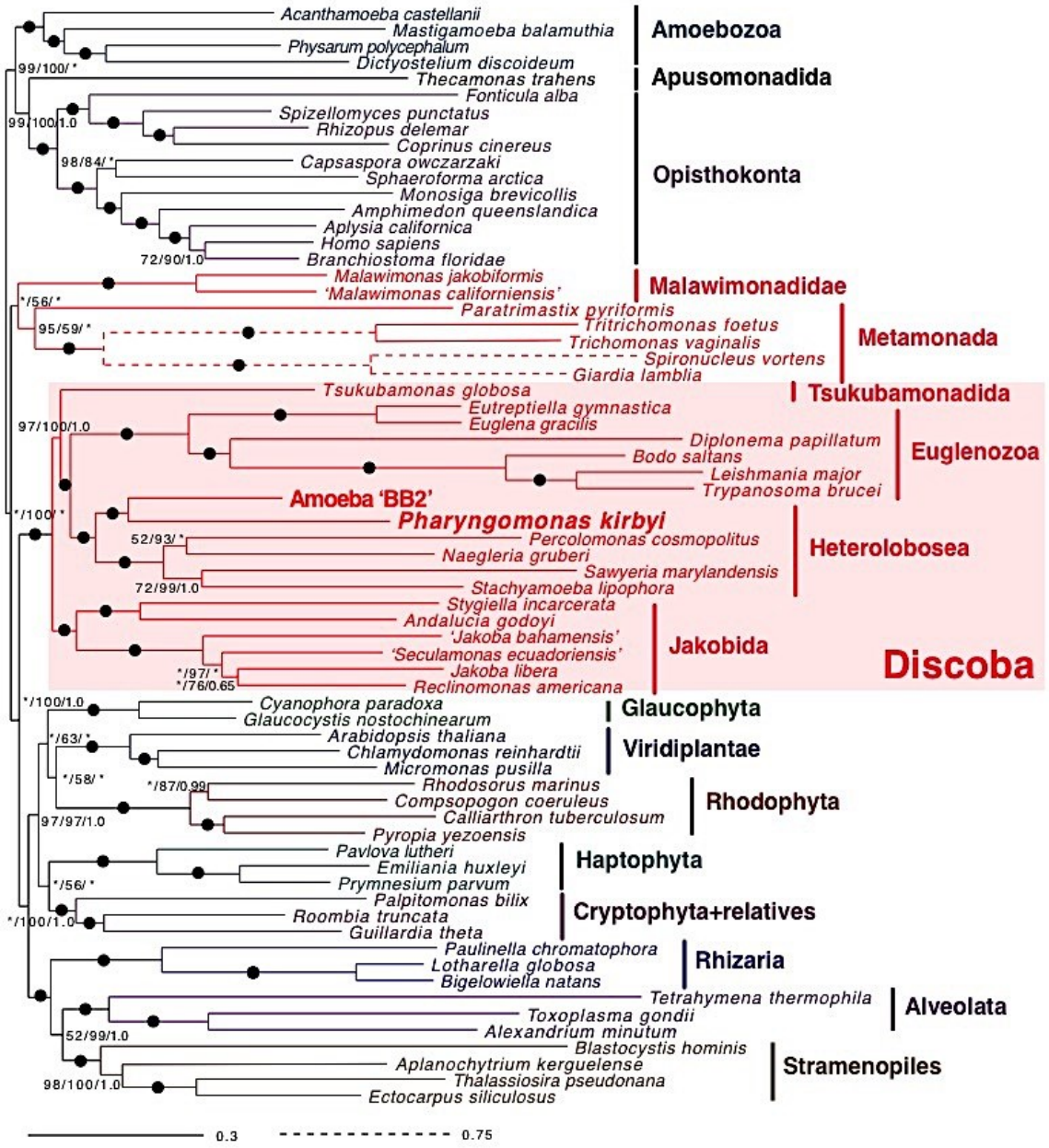


Figure 2.1 Phylogenetic tree estimated from the 252-protein dataset, inferred by IQ tree under the LG+C20+F+Gamma model with ML ultrafast bootstrap support (UFBoot). ML bootstrap support (MLBP) was also estimated by RAxML under the LG4X model, and BI posterior probabilities (BPP) were estimated by Phylobayes-MPI under the CAT-Poisson model. Support values are shown at each branch in the following order: MLBP, UFBoot and BPP. Black dots indicate 100% MLBP, 100% UFBoot and 1.0 BPP. Asterisks (*) indicate branches that were not recovered in the RAxML or Phylobayes-MPI analysis.

2.3.2 Mitochondrial Genome Overview

The mitochondrial genome of BB2 was assembled as a single circular-mapping molecule with a size of 119,312 bp (figure 2.2A). This mitochondrial genome contained a 49 kbp-long repeated region (figure 2.2A, highlighted in grey) with the two copies in an inverted orientation and situated opposite one another on the map (inverted repeat: IR). This organization is the simplest one among many possible organizations (multiple different linear or circular molecules, etc.) that are supported by sequencing of regions between the IR and non-repeated regions (shown in black in figure 2.2A). In addition, read mapping onto the mitochondrial genome revealed higher coverage (4-5 times on average) in the IR region compared to non-repeated regions (supplementary figure S2): This difference in coverage depth is consistent with the presence of multiple copies of the 49kbp-long region. However, the average coverage statistics in the IR region should be treated with caution since the read depth is extremely variable (not observed in the coverage analysis of *Pharyngomonas,* see supplementary figure S2); the reason for this variability is unclear. It is possible that the mitochondrial genome structure is more complex than depicted in figure 2.2A.

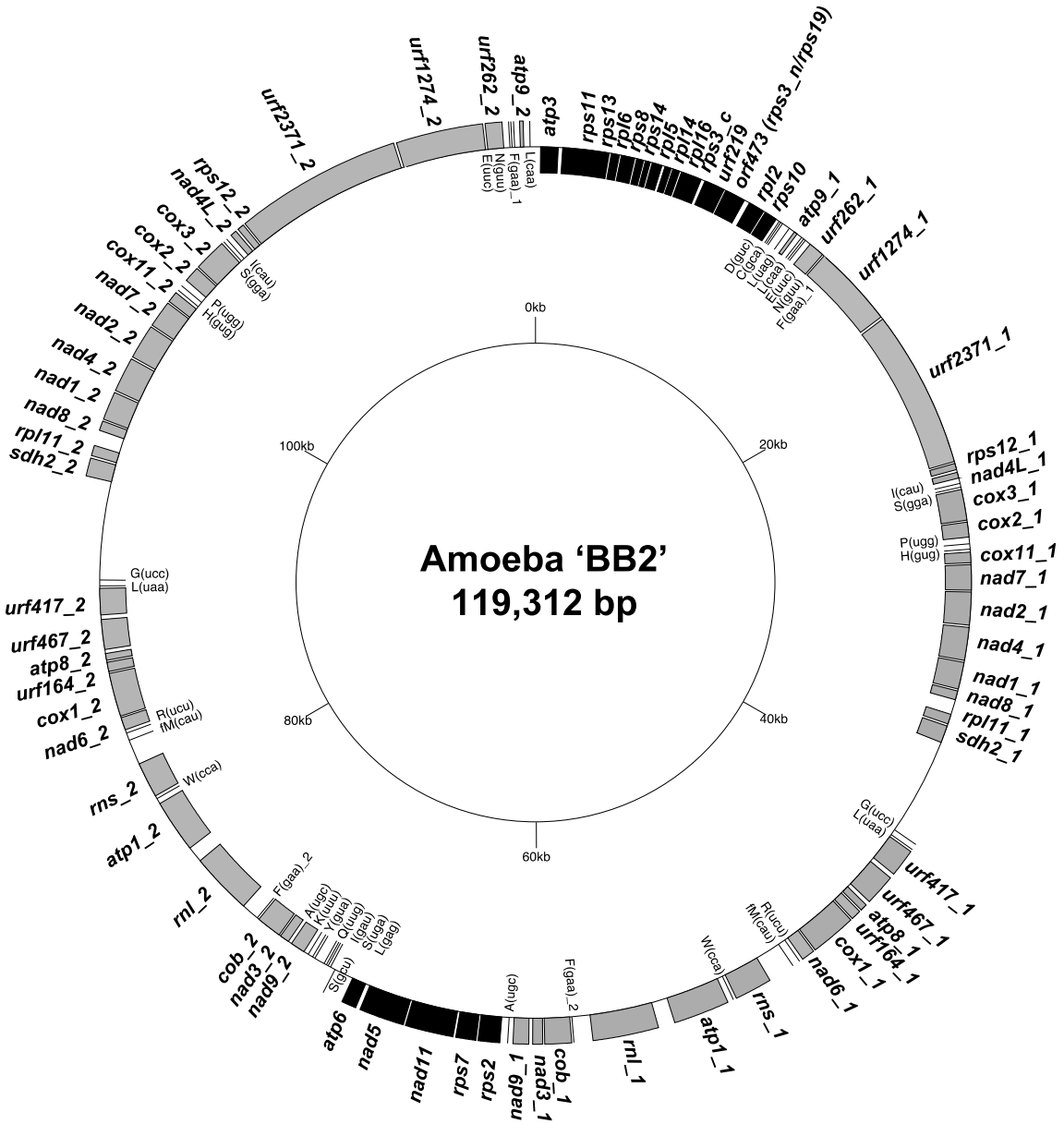
This genome conformation with such a large IR is somewhat unusual among mitochondrial genomes of protists, although mitochondrial genomes are extremely diverse in size and organization (Gray et al. 1997; Burger et al. 2003; Lavrov et al. 2012). A large IR (generally 20-30kb in length) is often observed in chloroplast genomes of most photosynthetic eukaryotes as a result of genome rearrangements (Palmer et al. 1982; Cosner et al. 1997). For the conformation shown in figure 2.2A, the BB2 mitochondrial genome is the largest mitochondrial genome known within Excavata. The overall A+T

(AT) content is 70.1%, the lowest among currently known mitochondrial genomes of heterolobosеids (*N. gruberi*: 77.8%, *N. fowleri*: 74.8% and *A. kona*: 83.3%).

In total, the mitochondrial genome of BB2 is made up of 81.2% coding sequence and contains 38 protein-coding genes, 2 rRNA genes, 24 tRNA genes and 7 URFs (for these counts the IR region is only considered once). Interestingly, the *rps3* gene was split into two ORFs designated as rps3_c (homologous to the C-terminus of *rps3*) and *ORF473*. ORF473 in turn includes both rps3_n (homologous to N-terminus of *rps3*) and a full-length *rps19*. Genes are tightly packed, and many of them are partially overlapping, such as *rps8-rps14*, *rpl14-rpl16*, *rps3_c-URF219*, *URF2371-URF1274* and *cox1-nad6*. No introns were detected.

The mitochondrial genome of *P. kirbyi* was assembled into a single linear-mapping molecule with a size of 75,717 bp (figure 2.2B). Attempts at "closing" the genome into a circular map were unsuccessful. Because of this, the completeness of the *P. kirbyi* mitochondrial genome is unproven, but it is likely complete (or near-complete) because the demonstrated gene content is very similar to that of BB2 (see below). The overall AT content is 87.5%, the highest known within Heterolobosea. The mitochondrial genome is gene-dense, with 92.0% of the sequence in coding regions, and it contains 38 protein-coding genes, small and large rRNA genes, 23 tRNA genes and 10 URFs. There are many partially overlapping ORFs including *cox3-URF129, URF129-rps10, rps19-rpl2, rps19-rps3, rpl16-rpl14, rpl5-rps8, rpl6-rps13, rps13-rps11, nad8-URF141, URF141-URF168, URF168-atp8, atp8-nad9, URF640-nad6, URF820-rps2*. As for BB2, no introns were detected in the *P. kirbyi* mitochondrial genome.

(A)


Amoeba 'BB2'
119,312 bp
0kb
20kb
40kb
60kb
80kb
100kb
atp3
rps11
rps13
rpl6
rps8
rpl5
rpl14
rpl16
rps3_c
urf219
orf473 (rps3_n/rps19)
rpl2
rps10
atp9_1
urf262_1
urf1274_1
urf2371_1
rps12_1
nad4L_1
cox3_1
cox2_1
cox11_1
nad7_1
nad2_1
nad4_1
nad1_1
nad8_1
rpl11_1
sdh2_1
urf417_1
urf467_1
atp8_1
urf164_1
cox1_1
nad6_1
rns_1
atp1_1
rnl_1
cob_1
nad3_1
rps7
rps2
nad11
nad5
atp6
cob_2
nad3_2
nad9_2
rnl_2
atp1_2
rns_2
nad6_2
cox1_2
urf164_2
atp8_2
urf467_2
urf417_2
sdh2_2
rpl11_2
nad8_2
nad1_2
nad4_2
nad2_2
nad7_2
cox11_2
cox2_2
cox3_2
nad4L_2
rps12_2
urf2371_2
urf1274_2
urf262_2
atp9_2
L(caa)
F(gaa)_1
N(guu)
E(uuc)
D(guc)
C(gca)
L(uag)
L(caa)
E(uuc)
N(guu)
F(gaa)_1
I(cau)
S(gga)
P(ugg)
H(gug)
G(ucc)
L(uaa)
R(ucu)
fM(cau)
W(cca)
F(gaa)_2
A(ugc)
X(uuu)
T(gua)
Q(uug)
I(gau)
S(uga)
L(gag)
S(gcu)
A(ugc)
F(gaa)_2
W(cca)
R(ucu)
fM(cau)
G(ucc)
L(uaa)
I(cau)
S(gga)
P(ugg)
H(gug)

(B)

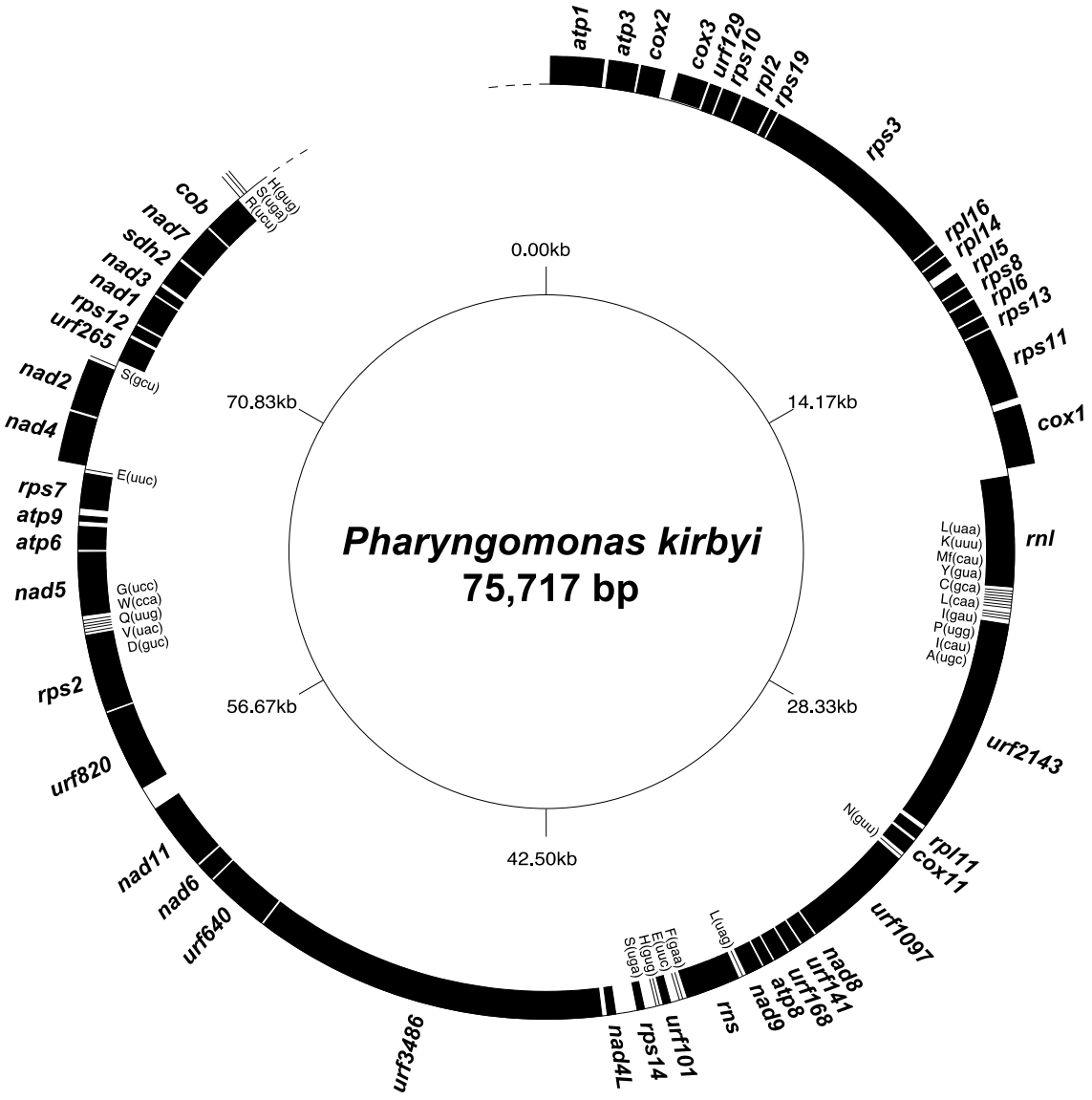


Figure 2.2 The mitochondrial genome maps of (A) amoeba 'BB2' and (B) ***Pharyngomonas kirbyi*****.** Genes encoding proteins and ribosomal RNAs are shown in boxes, Transfer RNA genes are shown by lines. Boxes on the outside of the circle represent RNAs encoded on one (positive) strand, boxes inside are RNAs encoded on another (negative) strand. The duplicated 'IR' regions are highlighted in grey. Single copy regions are shown in black.

2.3.3 Mitochondrial Gene Content and Synteny

The gene contents of these mitochondrial genomes were compared to eukaryotes from diverse lineages (figure 2.3). BB2 and *P. kirbyi* mitochondrial genomes appear to have

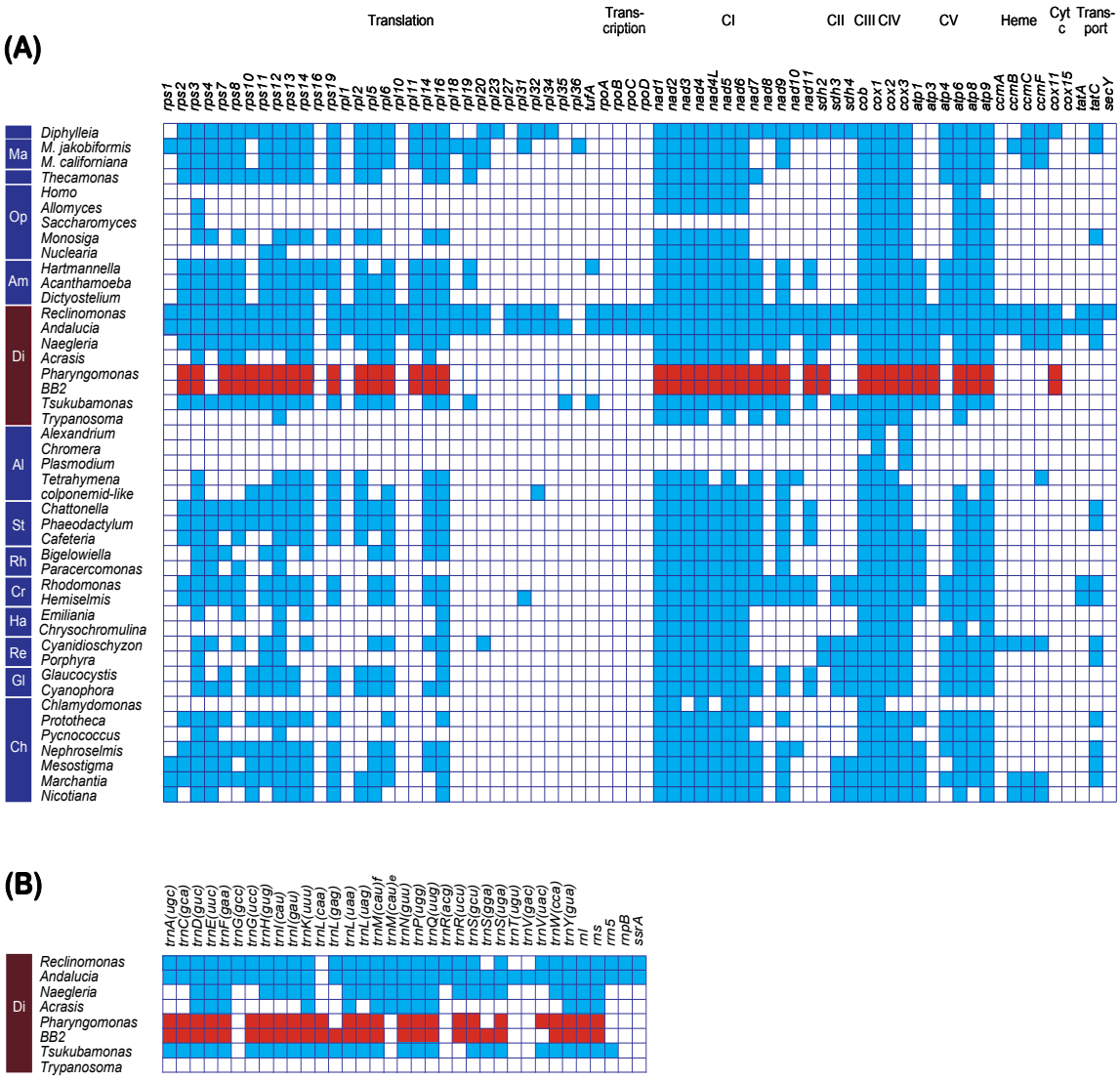


Figure 2.3 Presence and absence of (A) mitochondrial protein coding genes among various eukaryotes and (B) transfer RNA genes among Discoba. Ma: *Malawimonas,* Op: Opisthokonta, Am: Amoebozoa, Di: Discoba, Al: Alveolata, St: Stramenopiles, Rh: Rhizaria, Cr: Cryptophyceae, Ha: Haptophyta, Re: Red algae, Gl: Glaucophyta, Ch:Chloroplastida, CI-CV: electron transport chain complex I-V (following Kamikawa et al. 2016).

extremely similar gene contents, although the lack of certainty regarding completeness of the *P. kirbyi* mitochondrial genome prevents definitive conclusions. Their mitochondrial genomes encoded ribosomal proteins (*rps2, 3, 7, 8,10-14, 19, rpl2, 5, 6, 11, 14, 16*), components of electron chain transport complexes I (*nad1-4, 4L, 5-9, 11*), II (*sdh2*), III (*cob*), IV (*cox1-3)*, and V (*atp1, 3, 6, 8, 9*), and a cytochrome c oxidase assembling protein (*cox11*) (figure 2.3). Although BB2 and *P. kirbyi* are deep-branching heteroloboseids, their gene content represented a subset of that of the mitochondrial genome of *Naegleria.* The *Naegleria* mitochondrial genomes encode three additional protein-coding genes: two cytochrome c maturase subunits (*ccmC, ccmF*) and twin arginine translocase (*tatC*) (figure 2.3). The mitochondrial gene contents of BB2 and *P. kirbyi* also represented a subset of that of jakobids. These observations showed that BB2 and *P. kirbyi* do not have unusually "ancestral" (i.e., gene-rich) mitochondrial genomes, relative to other members of Discoba.

Gene order comparison among the representive mitochondrial genomes from Discoba (*N. fowleri,* BB2, *P. kirbyi*, *A. godoyi*, *T. globosa*) showed vestiges of a highly conserved ribosomal gene cluster that is similar to the three contiguous S10, spectinomycin (Spc), and Alpha operons of their close bacterial relative *Rickettsia* (supplementary figure S3). This ribosomal gene cluster was not detected in the mitochondrial genome of *A. kona.* The mitochondrial genome of BB2 and *P. kirbyi* also showed two other pairs of genes in the same order (*nad2-nad4* and *cox2-cox3*).

2.3.4 RNA Editing in Mitochondria of Amoeba 'BB2'

By comparing RNA transcripts to genomic DNA sequences, we identified some 475 sites where RNA transcripts contained single nucleotide insertions relative to the

mitochondrial genome sequence. No deletions or substitutions were observed. About sixty of these insertion sites were confirmed by sequencing PCR products from both genomic and cDNA (supplementary figure S4). Mononucleotide insertions were detected in all 44 protein-coding ORFs, SSU and LSU rRNAs, and 12 out of 24 tRNAs (table 2.1, see supplementary table S3 for numbers and types of nucleotide inserted for each class of RNA). The most frequently inserted nucleotide by far was guanylate (83.5%), followed by adenylate (7.8%), cytidylate (5.5%) and uridylate (3.2%) (table 2.1).

Table 2.1 Number and type of insertions found in edited mitochondrial transcripts of amoeba BB2.

Type	# genes edited/ total # of genes	G	A	C	U	Total
Protein-coding ORFs	44/44	351	36	21	13	421
tRNA	12/24	10	1	1	1	13
rRNA	2/2	36	0	4	1	41
Total		**397 (83.5%)**	**37 (7.8%)**	**26 (5.5%)**	**15 (3.2%)**	**475**

Strikingly, all the inserted nucleotides were located next to one or more nucleotides with the same identity, encoded by the mtDNA. This characteristic made the precise localization of editing sites ambiguous (i.e., we could not determine whether nucleotides were inserted before or after the nucleotide already specified by the mitochondrial genome). Insertion-type RNA editing observed in protein-coding regions fixes apparent gene fragmentation and frame-shifts in the mtDNA, resulting in one continuous ORF for each transcribed gene (except for the split *rps3* gene; see above).

tRNA editing

The secondary structures predicted using mtDNA sequences showed the general cloverleaf structure of tRNAs but some seemingly lacked normally well-conserved features. The highly conserved GUUC motif was missing in the inferred mitochondrial tRNAs for C(gca), D(guc), F(gaa), G(ucc), I(cau), I(gau), L(uaa), P(ugg) and Y(gua). In the tRNA for K(uuu), only 6 bp were found in the acceptor stem, instead of the typical 7 bp. In addition, tRNAs for R(ucu) and S(guc) were not initially identified from the mtDNA.

These tRNAs were further investigated by comparing mtDNA and the transcriptome data, and the secondary structures were predicted again using the edited transcript sequences (figure 2.4). The tRNAs for C(gca), D(guc), F(gaa), G(ucc), I(gau), L(uaa), P(ugg), Y(gua) each had a single guanosine insertion in the GTψC stem, creating a G-C base pair and restoring the highly conserved GUUC motif. The tRNA for I(cau) had a single uridine insertion in the GTψC loop, which generated the GUUC motif. The tRNA for K(uuu) had a guanosine insertion in the variable loop and a cytidine insertion in the 3′ end portion of the acceptor stem, the latter of which created a G-C base pair and led to a typical 7 bp acceptor stem. Most interestingly, tRNAs for R(ucu) and S(gcu) were identified from their edited transcripts; these genes were initially unrecognizable from analysis of the mtDNA alone. The tRNA for R(ucu) had a guanosine insertion in the anticodon stem, which created a G-C base pair and a typical 4-5 bp-long anticodon stem (an unusually short 3 bp-long stem was implied by the mtDNA sequence alone). In the tRNA for S(gcu), a single adenosine was inserted at either the anticodon stem or loop, creating a typical 7 nucleotide loop (whereas the mtDNA implies an aberrant 6 nucleotide anticodon loop).

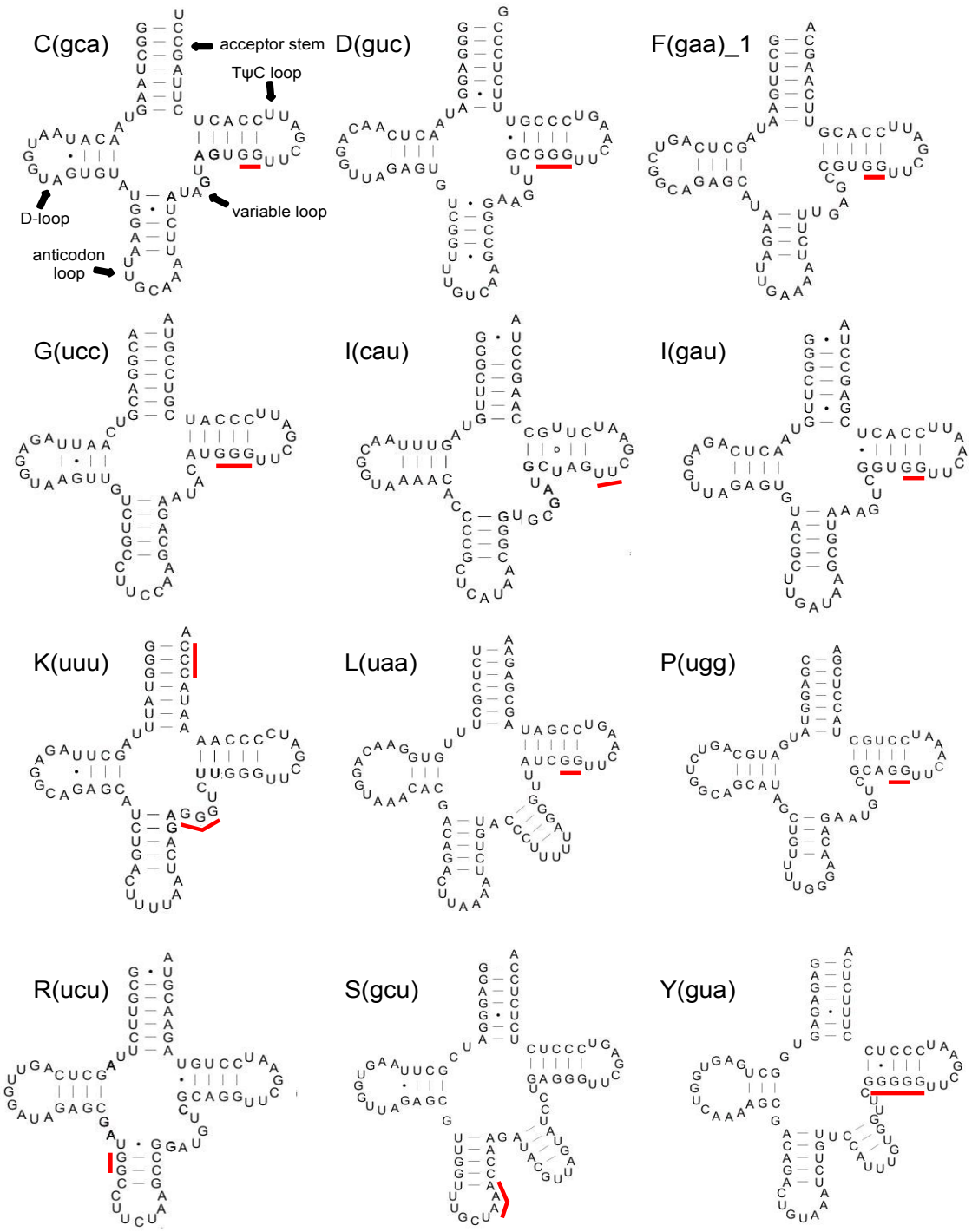


Figure 2.4 Secondary structures of mitochondrial tRNAs of amoeba BB2. Regions including an editing site are highlighted with red lines. For each highlighted region, only one nucleotide is inserted (the exact insertion sites are unknown since nucleotides are inserted next to encoded nucleotide(s) with the same identity).

SSU rRNA editing

The SSU rRNA of BB2 mitochondria is 1,665 nucleotides in length in mature form. Maturation requires editing at some 19 sites with single nucleotide insertions (16 guanylate, 1 uridylate and 2 cytidylate). Editing sites were distributed over the entire length of the rRNA. After inferring the secondary structure of the SSU rRNA, we determined that the localization of editing sites is not limited to any particular features of the predicted structure (figure 2.5). We found most insertions (11/19) in base-paired regions (stems) and 2 insertions that were unambiguously within loop structures. The location of other insertions (6/19) were uncertain, as the run of the same nucleotides (one of them being the nucleotide added by editing) were localized in regions covering both loop and stem structures. If we assume that RNA editing is directional (nucleotides are always inserted either before or after the identical encoded nucleotide), then we infer that 13 edits are in stems and 6 in loop structures (edits after), or 14 in stems and 5 in loop structures (edits before). In the overall SSU rRNA, about twice as many nucleotides form the stem regions (1,026/1,665 nucleotides) compared to loop regions (629/1,665 nucleotides); these proportions are similar to the inferred proportions of edits, and Z-tests indicated that the difference was not significant ($p = 0.47$ for after edits, or 0.18 for before edits).

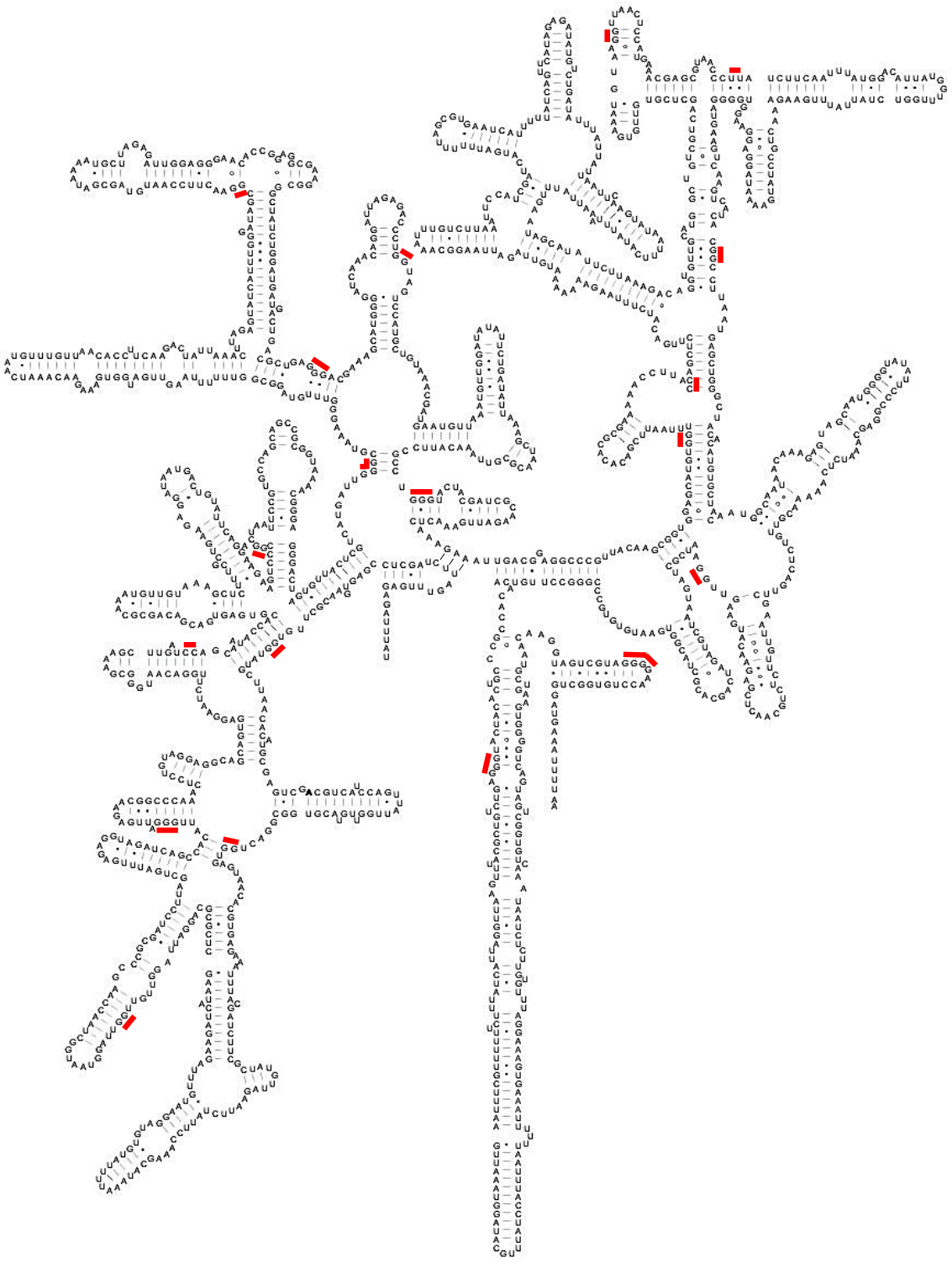


Figure 2.5 Predicted secondary structure of BB2 mitochondrial SSU rRNA. Regions including an editing site are highlighted with red lines. For each highlighted region, only one nucleotide is inserted (the exact insertion sites are unknown since nucleotides are inserted next to the encoded nucleotide(s) with the same identity).

2.3.5 Sequence Conservation around Editing Sites

We also examined sequence conservation near editing sites in order to identify any motifs that may be used as a localization signal for RNA editing. For the analysis, we collected 311 60-nucleotide-long sequences near editing sites (29 nucleotides before and after; any sequences with editing sites next to more than one previously encoded nucleotide were omitted), and 328 randomly selected 60-nt-long sequences as a negative control, and for each set generated a sequence LOGO. Apart from the enrichment of guanosine at editing sites (discussed previously), general patterns in sequence LOGOs were similar between the control and test sets (supplementary figure S5). For this analysis, we only considered insertions that happened next to a single identical nucleotide, thus the flanking nucleotides (nucleotides right before and after editing sites) were constrained to be low in G (and rich in less commonly inserted nucleotides) as an artifact of this site selection. No motifs were apparent when the sequence length was extended to 120 nucleotides (not shown).

2.3.6 Accuracy of Editing Mechanism

RNA sequence reads were compared with consensus transcript sequences to examine the accuracy of the editing mechanism. The overall 'apparent error rate' was 0.06% (4187 mismatches over 7,399,800 nucleotides) for non-editing windows and 0.4% (2489/ 635,920) for editing windows (figure 2.6). Inside non-editing windows, deletion, substitution and insertion-type errors were 0.01%, 0.04% and 0.01% respectively, At editing sites, apparent deletion-type errors were 30 times more common (0.34%), while substitution-type errors (0.04%) and insertion-type errors (0.01%) were similar to the error rates at non-editing windows.

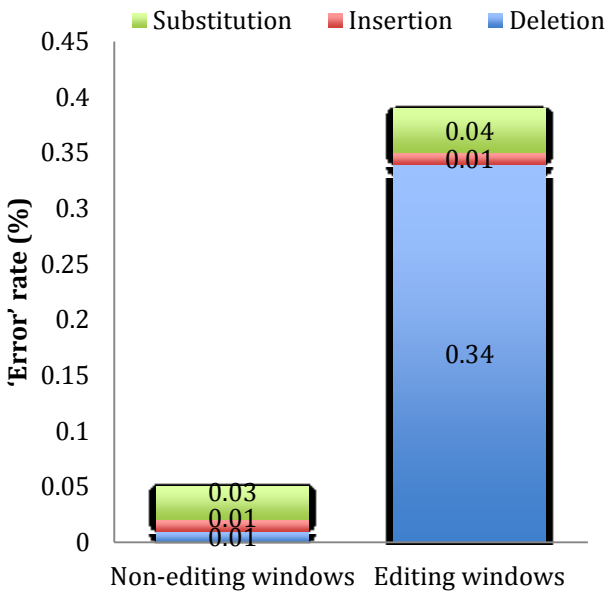


Figure 2.6 Apparent error rate estimated by comparing individual sequencing reads with consensus transcript sequences for the non-editing windows (10-nt windows that exclude editing sites) and the editing windows (10-nt windows around editing sites). Three types of apparent error rates are shown in percentages (%): substitution-type (green), insertion-type (red) and deletion-type (blue). The excess of apparent deletion errors at editing sites includes yet-to-be edited sites, which would not be actual errors.

Z-tests indicated that only the rate of deletion-type error was different between editing and non-editing sites ($p<0.00001$). The excess apparent deletion errors presumably represent instances where RNA editing has not happened (rather than transcription errors) and suggest that a (small) proportion of transcripts were not edited, or not edited yet, when RNA was extracted. If so, then most of these may not actually be errors, rather they represent immature pre-edited transcripts.

2.4 DISCUSSION

2.4.1 Mitochondria Genome Evolution in Discoba

In previously reported phylogenetic analyses based on the 18S rRNA gene alone, the relative phylogenetic positions of BB2 and *P. kirbyi* were ambiguous (Harding et al. 2013). A BI analysis showed BB2 as the deepest-branching member of Heterolobosea, while, BB2 and *P. kirbyi* were sister taxa at the base of Heterolobosea in an ML analysis, but with negligible statistical support in each case (Harding et al. 2013). Our phylogenomic analyses of 252 protein-coding genes have resolved this uncertainty, clearly showing that BB2 and *P. kirbyi* are sister taxa at the base of Heterolobosea.

Amoeba BB2 and *P. kirbyi* probably have identical mitochondrial gene content, at least for protein-coding genes of known function, and this represents a subset of the gene complement of *Naegleria* species. This suggests that BB2 and *P. kirbyi* do not have unusually "ancestral" (gene-rich) mitochondrial genomes, relative to other Discoba, despite their phylogenetic position as the deepest branch within Heterolobosea. The presence of three extra genes encoded on *Naegleria* mitochondrial genomes (*ccmC, ccmF* and *tatC*), but not on the mtDNA of BB2, *P. kirbyi* or *A. kona* is likely the result of parallel gene losses; one set of events in the common ancestor of *P. kirbyi* and BB2 and a second in the lineage leading to *A. kona*. In addition, both BB2 and *P.kirbyi* have the nucleus-encoded phage-type RNA polymerase (with a sequence highly similar to that of *Naegleria*; data not shown) instead of the bacteria-like mitochondrial RNA polymerase present in Jakobids. Since Jakobida does not seem to be the earliest branching eukaryotic lineage (Derelle et al. 2015), it seems likely that the last eukaryote common ancestor (LECA) had both types of RNA polymerase, and the bacterial one was lost in

Heterolobosea after the split from Jakobida (Stechmann and Cavalier-Smith 2002).

2.4.2 RNA editing in BB2

The mitochondrial transcripts of BB2 require RNA editing to produce functional RNAs. In BB2 transcripts, mononucleotides are inserted next to one or more nucleotide of the same identity encoded by mtDNA by a very accurate and efficient RNA editing mechanism (only 0.4% apparent error rate). A somewhat similar RNA editing system occurs in paramyxoviruses; the mature P protein mRNA of paramyxoviruses is produced after insertion of one or more additional G residues next to an encoded G residue (Jacques et al. 1994). This happens by co-transcriptional polymerase 'stuttering' at a homo-polymer tract (A_nG_n) (Jacques et al. 1994). Although RNA editing in BB2 adds one additional nucleotide next to an identical encoded nucleotide, it is unlikely to use the same mechanism as paramyxoviruses because BB2 pre-mRNA did not have homo-polymer tracts around editing sites, and the error rate of RNA editing is much lower for BB2 (see below).

Among the known RNA editing systems, RNA editing in *Physarum* shares some similar characteristics in that all four types of nucleotides are inserted, and all types of RNA (mRNA, rRNA and tRNA) are edited (Antes et al. 1998, Bundschuh et al, 2011, Mahendran et al. 1994). However, unlike BB2 in which all nucleotides are inserted next to an identical encoded nucleotide, this is true of only some edits in *Physarum.* In addition, some dinucleotide insertions are observed in *Physarum* whereas only mononucleotide insertions occur in BB2.

In kinetoplastid, each U insertion/deletion site is specified by a gRNA and editing is post-transcriptional. However, it is unlikely that BB2 mitochondria use the same RNA

editing mechanism as kinetoplastids, based on our findings. For instance, we could not detect any anti-sense transcripts (e.g., gRNA) in our RNA-Seq data that could be used as a form of template for nucleotide insertion (data not shown).

The polymerase stuttering mechanism of paramyxoviruses results in a high error rate (about 50%; Jacques et al. 1994). Editing in kinetoplastid mitochondria is also relatively error-prone, and a proportion of kinetoplastid RNAs found to be unedited, misedited and partially edited in RT-PCR assays (Abraham et al. 1988; Visomirski-Robic and Gott 1995). Recently, David et al. (2015) reported that approximately 50% of reads are mis-edited or not fully edited in mitochondria of the kinetoplastid *Perkinsela*. On the other hand, the co-transcriptional insertion editing of *Physarum* is very accurate, with only 5% mis-edited in RNAs synthesized by partially purified mtTECs (Byrne et al. 2002, Visomirski-Robic and Gott 1995). Although approaches used to estimate error rate vary between organisms, the extremely low error rate in BB2 (0.4%) suggests that the RNA editing mechanism in BB2 is unusually accurate and efficient. Also, the low proportion of unedited transcripts suggests that RNA editing probably takes place during transcription, or very soon after.

In BB2, RNA editing seems to be essential for generating functional tRNAs since RNA editing creates conserved features such as the GUUC motif (which is needed for tRNA recognition), the 7bp acceptor stem and a 7 nucleotide anticodon loop. Similar types of RNA editing are also found in some Myxomycota (*Physarum polycephalum* and *Didymium nigripes*), where a single nucleotide is inserted in the GTψC loop or stem, anticodon stem, DHU stem or acceptor stem (Antes et al. 1998). However, it is unknown how editing sites are specified in Myxomycota. There is also tRNA editing in

Acanthamoeba, but this is different in that it uses base-pairing in the stems as the template for editing (Lonergan and Gray, 1993); the mechanism for BB2 must be different since the editing sites in BB2 transcripts are not always in base-paired regions (either in tRNAs or in the SSU rRNA).

RNA editing machinery in BB2 mitochondria is currently unclear. Therefore, further studies will be needed to determine where precisely the insertions happen (i.e., before or after the identical encoded nucleotide), how editing sites are specified and if RNA editing in BB2 is truly co-transcriptional. To resolve these questions many molecular biology experiments must follow, including the development of *in vitro* assays using isolated mitochondria, similar to experiments which showed that RNA editing in *Physarum* is co-transcriptional (Visomirski-Robic and Gott 1995; Cheng et al. 2001).

2.4.3 Phylogenetic context

Mono-nucleotide insertion-type RNA editing in mitochondria of BB2 likely arose in the lineage leading to BB2 after the split from *Pharyngomonas*, as this type of RNA editing is not observed in *Pharyngomonas* or any other heteroloboseid known to date. Within Heterolobosea, *A. kona* and *N. gruberi* are known to undergo C-to-U RNA editing of their mitochondrial transcripts (Knoop and Rudinger, 2010; Fu et al. 2014), however, we found no evidence of C-to-U RNA editing or any other type of substitution editing in BB2 and *P.kirbyi.* This type of RNA editing is distinct from what we observed in BB2 mitochondria, where insertion-type RNA editing takes place in all protein-coding ORFs, rRNAs and many tRNAs. In the mitochondria of *A. kona* and *N .gruberi,* only protein-coding ORFs undergo substitution-type RNA editing. In addition, just two editing sites are identified in *N. gruberi* and six sites in *A. kona*, compared to 475 sites in BB2.

Therefore, it is unlikely that RNA editing arose in the common ancestor of all Heterolobosea, rather these appear to be two different phenomena that have evolved independently.

CHAPTER 3 Final Conclusion

The main goal of this study was to explore the diversity and evolutionary dynamics of mitochondrial genomes in newly-discovered protists. In the previous chapters, I discussed two main questions: are the mitochondrial genomes of BB2 and *P. kirbyi* as ancestral (i.e., gene-rich) as those of Jakobida, and what are the features of RNA editing in the mitochondrial genome of BB2? Here, I will discuss the implications of the findings of Chapter 2.

Gene content and mitochondrial genome structure

Many studies of mitochondrial genome evolution have found that the great diversity in mitochondrial genomes results from a continuous and/or parallel reduction of mitochondrial gene content in different lineages (Burger et al. 2013, Kamikawa et al. 2014). My analysis of the mitochondrial genome of BB2 and *P. kirbyi* showed that their mitochondrial genomes are not as gene rich or 'ancestral' in appearance as those of Jakobida despite their deepest-branching position within Heterolobosea. Previously there were only three mitochondrial genomes available within Heterolobosea, that of two *Naegleria* species and *Acrasis kona.* With only three genomes known in this lineage, it was difficult to address important evolutionary questions, such as the time frame for gene loss or gene transfer events. My results have almost doubled the number of mitochondrial genomes known within the Heterolobosea, allowing a better comparative analysis of gene contents and genome structure. However, this sampling of mitochondrial genomes within the Heterolobosea is still relatively sparse when compared with the vast ecological and

phylogenetic diversity of this protistan group (Pawlowski 2014). More mitochondrial genomes from Heterolobosea should be characterized to better understand the dynamics of mitochondrial genome evolution within the group and pinpoint where, in their phylogenetic history, events of gene transfer to the nucleus have occurred.

RNA editing in BB2

RNA editing in BB2 mitochondria is the first case of insertion-type RNA editing in Heterolobosea, expanding our understanding of both the diversity and evolution of mitochondrial RNA editing. The next step is to determine the molecular mechanisms underpinning the insertion-type RNA editing in BB2 mitochondria. Some *in vitro* experiments may help us reveal the hidden signals (which were not detectable using bioinformatics) for editing site specification and identification. To do this, it would be necessary to establish a system where we can produce the transcripts *in vitro* using a specific template we provide. One approach would be to purify mitochondria by differential centrifugation, and use this crude mitochondrial extract to initiate transcription on a DNA template we provide (Visomirski-Robic and Gott 1995). We could then synthesize run-on transcripts in the presence of radiolabelled ribonucleoside triphosphates (NTPs) to test the efficiency and accuracy of RNA editing in BB2 (Visomirski-Robic and Gott 1995). We could also generate run-on transcripts using chimeric DNA templates, which are cleaved at a varying distance from the editing site, to see where the critical information for editing is hidden (Byrne and Gott 2002). In addition, the editing machinery should be studied to find out which editing factors (i.e., proteins and small RNAs) are involved and how the enzymes function with or without nucleic

acid templates to accomplish accurate insertion editing. Small RNA sequencing could be one of possible approaches to detect a potential editing factor (i.e., small RNAs), which can serve as a template for RNA editing.

In addition, this finding of phylogenetically isolated RNA editing system suggests that RNA editing arose independently in the BB2 lineage since no other heteroloboseid studied to date, including its sister lineage *P. kirbyi,* has this type of editing of their mitochondrial transcripts. Once the RNA editing mechanism for BB2 is better characterized, it would be interesting to see how the mechanism for RNA editing in this organism differs from, or is similar to, other RNA editing mechanisms of other organisms. For example, the highly accurate mechanism of the BB2 RNA editing system may provide useful insights into understanding other RNA editing mechanisms such as that of *Physarum* whose RNA editing patterns seems to be somewhat similar to BB2. This would provide clues as to whether this type of insertional RNA editing was truly acquired *de novo* independently in the different lineages or if several of these systems share some common molecular features or enzymes predisposing them to the repeated emergence of RNA editing in each of the species where it is observed.

REFERENCES

Abraham JM, Feagin JE, Stuart K. 1988. Characterization of cytochrome c oxidase III transcripts that are edited only in the 3′ region. *Cell* 55, 267-272.

Adams KL, Palmer JD. 2003. Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Mol. Phylogene. Evol.* 29, 380-395.

Alfonzo JD, Thiemann O, Simpson L. 1997. The mechanism of U insertion/deletion RNA editing in kinetoplastid mitochondria. *Nucleic Acids Research* 25(19), 3751-3759.

Aljanabi SM, Martinez I. 1997. Universal and rapid salt-extraction of high quality genomic DNA for PCR-based techniques. *Nucleic Acids Res.* 25, 4692–4693.

Anderson S, Bankier AT, Barrell BG, De Bruijn MHL, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F, Schreier PH, Smith AJH, Staden R, Young IG. 1981. Sequence and organization of the human mitochondrial genome. *Nature* 290, 457-465.

Anderson SG, Zomorodipour A, Andersson JO, Sicheritz-Ponten T, Alsmark UC, Podowski RM, Naslund AK, Eriksson AS, Winkler HH, Kurland CG. 1998. The genome sequence of *Rickettsia prowazekii* and the origin of mitochondria. *Nature* 396(6706), 133-40.

Antes T, Costandy H, Mahendran R, Spottswood M, Miller D. 1998. Insertional editing of mitochondrial tRNAs of *Physarum polycephalum* and *Didymium nigripes*.. Mol. Cell. Biol. 18(12), 7521-7527.

Aphasizhev R, Aphasizheva I, Nelson RE, Gao G, Simpson AM, Kang X, Falick AM, Sbicego S, Simpson L. 2003. Isolation of a U-insertion/deletion editing complex from *Leishmania tarentolae* mitochondria. *EMBO J* 22:913–924.

Benne R, Van Den Burg J, Brakenhoff JPJ, Sloof P, Van Boom JH, Tromp MC. 1986. Major transcript of the frameshifted *cox11* gene from trypanosome mitochondria contains four nucleotides that are not encoded in the DNA. *Cell* 46, 819-826.

Brown MW, Kolisko M, Silberman JD, Roger AJ. 2012. Aggregative multicellularity evolved independently in the eukaryotic supergroup Rhizaria. *Curr Biol*. 22, 1123-127.

Burger G, Gray MW, Forget L, Lang BF. 2013. Strikingly bacteria-like and gene-rich mitochondrial genomes throughout jakobid protists. *Genome Biol Evol.* 5, 418-438.

Burger G, Forget L, Zhu Y, Gray MW, Lang BF. 2002. Unique mitochondrial genome architecture in unicellular relatives of animals. *PNAS.* 100(3), 892-897.

Bundschuh R, Altmuller J, Becker C, Nurnberg P, Gott JM. 2011. Complete characterization of the edited transcriptome of the mitochondrion of *Physarum*

polycephalum using deep sequencing of RNA. *Nucleic Acids Research.* 39(14), 6044-55.

Boisvert S, Laviolette F, Corbeil J, 2010. Ray: simultaneous assembly of reads from a mix of high-throughput sequencing technologies. *J Comp Biol.* 17(11), 1519 – 1533.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: A flexible trimmer for Illumina Sequence Data. *Bioinformatics*, 30(15), 2114-2120.

Brossard N, Delage E, Littlejohn TG, Plante I, Rioux P, Saint-Louis D, Zhu Y, Burger G. 1998. Genome structure and gene content in protest mitochondrial DNAs. *Nucleic Acids Research.* 26(4), 865-878.

Byrne EM, Stout A, Gott JM. 2002. Editing site recognition and nucleotide insertion are separable processes in *Physarum* mitohcondria. EMBO J. 21(22), 6154-6161.

Byrne EM, Gott JM. 2002. Cotranscriptional editing of *Physarum* mitochondrial RNA requires local features of the native template. *RNA* 8, 1174-1185.

Cavalier-Smith T, Nikolaev S. 2008. The zooflagellates Stephanopogon and Percolomonas are a clade (class Percolatea: Phylum Percolozoa). *J Eukaryot Microbiol.* 55:501–9.

Chaterigner-Boutin A, Small I. 2011. Organellar RNA editing. *John Wiley & Sons, Ltd. WIREs RNA* 2, 493-506

Cheng YW, Visomirski-Robic LM, Gott JM. 2001. Non-template addition of nucleotides to the 3′ end of nascent RNA during RNA editing in *Physarum*. *EMBO J.* 20, 1405-1414.

Chevreux B, Pfisterer T, Drescher B, Driesel AJ, Muller WEG, et al. 2004. Using the miraEST assembler for reliable and automated mRNA transcript assembly and SNP detection in sequenced ESTs. *Genome Res*. 14, 1147-59.

Chaterigner-Boutin A, Small I. 2011. Organellar RNA editing. *Wiley Interdiscip Rev RNA*, 2(4), 493-506.

Conant, G. C., and K. H. Wolfe. 2008. GenomeVx: Simple web-based creation of editable circular chromosome maps, *Bioinformatics,* 24(6), 861-2.

Conser ME, Jansen R, Palmer JD, Downie SR. 1997. The highly rearranged chloroplast genome of *Trachelium caeruleum* (Campanulaceae): multiple inversions, inverted repeat expansion and contraction, transposition, insertions/deletions, and several repeat families.

Covello PS, Gray MW. 1993. On the evolution of RNA editing. *Trend Genet.* 9(8), 265-8.

Criscuolo A, Gribaldo S. 2010. BMGE (Block Mapping and Gathering with Entropy): a new software for selection of phylogenetic informative regions from multiple sequence

alignments. *BMC Evol. Biol.* 10, 210.

David V, Flegontov P, Gerasimov E, Tanifuii G, Hashimi H, et al. 2015. Gene loss and error-prone RNA editing in the mitochondrion of *Perkinsela*, and endosymbiotic kinetoplastid. *Mbio*. 6(6), e01498-15.

Derelle R, Torruella G, Klimes V, Brinkmann H, Kim E, Vlcek C, Lang BF, Elias M. 2015. Bacterial proteins pinpoint a single eukaryotic root. *PNAS*. 112(7), E693-9.

Dobakova E, Flegontov P, Skalicky T, Lukes J. 2015. Unexpectedly Streamlined Mitochondrial Genome of the Euglenozoan *Euglena gracilis*. *Genome Biol Evol*. 7(12), 3358-67.

Doolittle WF, Lukes J, Archibald JM, Keeling PJ, Gray MW. 2011. Comment on "Does constructive neutral evolution play an important role in the origin of cellular complexity?". *Bioessays* 33, 427-9.

Estevez AM, Simpson L. 1999. Uridine insertion/deletion RNA editing in trypanosome mitochondria—a review. *Gene* 240:247-260

Fu C, Sheikh S, Miao W, Andersson SG, Baldauf SL. 2014. Missing genes, multiple ORFs, and C-to-U type RNA editng in *Acrasis kona* (Heterolobosea, Excavata) mitochondrial DNA. *Genome Biol. Evol*. 6(9), 2240-2257.

Fritz-Laylin LK, et al. 2010. The genome of *Naegleria gruberi* illuminates early eukaryotic versatility. *Cell* 140, 631-642.

Flegontov P, Gray MW, Burger G, Lukes J. 2011. Gene fragmentation: a key to mitochondrial genome evolution in Euglenozoa? *Curr Genet*. 57, 225–232.

Gott JM, Emeson RB. 2000. Functions and mechanisms of RNA editing. *Annu Rev Ganet*, 34, 499-531.

Gott JM. Somerlot BH, Gray MW. 2010. Two forms of RNA editing are required for tRNA maturation in *Physarum* mitochondria. *RNA* 16, 482-488.

Grabherr MG, Hass BJ, Yassour M, Levin JZ, Thompson DA, et al. 2011. Trinity: reconstructing a full-length transcriptome without a genome from RNA-Seq data. *Nat Biotechnol.* 29(7), 644-652.

Gray MW, Covello PS. 1993. RNA editing in plant mitochondria and chloroplasts. *FASEB J.* 7, 64–71.

Gray MW. Lang BF, Cedergren R, GoldingGB, Lemieux C, Sankoff D, Turmel M, Brossard N, Delage E, Littlejohn TG, Plante I, Rioux P, Saint-Louis D, Zhu Y, Burger G.

1998. Genome structure and gene content in protest mitochondrial DNAs. *Nucleic Acids Research.* 26(4), 865-878.

Gray MW, Burger G, Lang BF. 1999. Mitochondrial evolution. *Science* 283, 1476-1481.

Gray MW. 2003. Diversity and evolution of mitochondrial RNA editing systems. *IUBMB Life* 55, 227–233.

Gray MW, Lang BF, Cedergren R, GoldingGB, Lemieux C, et al. 2004. Mitochondria of protists. *Annu Rev Genet.* 38, 477-524.

Gray MW. 2012. Evolutionary origin of RNA editing. *Biochemistry* 51, 5235-42.

Hajduk SL, Harris ME, Pollard VW. 1993. RNA editing in kinetoplastid mitochondria. *FASEB J* 7(1), 54-63.

Hampl V, Hug LA, Leigh JW, Dacks JB, Lang BF, Simpson AGB, Roger AJ. 2009. Phylogenomic analyses support the monophyly of Excavata and resolve relationships among eukaryotic "supergroups". *PNAS.* 106, 3859-3864.

Harding T, Brown MW, Plotnikov A, Selivanova E, Park JS, et al. 2013. Amoeba stages in the deepest branching heteroloboseans, including *Pharyngomonas*: evolutionary and systematic implications. *Protist* 164, 272–286.

Harding T, Brown MW, Simpson AGB, Roger AJ. 2016. Osmoadaptative strategy and its molecular signature in obligately halophilic heterotrophic protists. *Genome Biol Evol*. doi: 10.1093/gbe/evw152.

Herman EK, Greninger AL, Visvesvara GS, Marciano-Cabral R, Dacks JB, Chiu CY. 2013. The mitochondrial genome and a 60-kb nuclear DNA segment from *Naegleria fowleri*, the causative agent of primary amoebic meningoencephalitis. *J Eukaryot Microbiol*. 60, 179-191.

Hieesel R, Combettes B, Brennicke A. 1994. Evidence for RNA editing in mitochondria of all major groups of land plants except the Bryophyta. *PNAS.* 91, 629-633.

Horton TL, Landweber LF. 2002. Rewriting the information in DNA: RNA editing in kinetoplastids and myxomycetes. *Curr Opin Microbiol.* 5, 620-626.

Jacques JP, Hausmann S, Kolakofsky D. 1994. Paramyxovirus mRNA editing leads to G deletions as well as insertions. *EMBO J.* 13(22), 5496-5503.

Jackson CJ, Norman JE, Schnare MN, Gray MW, Keeling PJ, Waller RF. 2007. Broad genomic and transcriptional analysis reveals a highly derived genome in dinoflagellate mitochondria. *BMC biol.* 5, 41.

Kamikawa R, Kolisko M, Nishimura Y, Yabuki A, Brown MW, et al. 2014. Gene content evolution in Discobid mitochondria deduced from the phylogenetic position and complete mitochondria; genome of *Tsukubamonas globosa*. *Genome Biol Evol.* 6(2), 306-315.

Kamikawa R, Shiratori T, Ishida K, Miyashita H, Roger AJ. 2016. Group II intron-mediated trans-splicing in the gene-rich mitochondrial genome of an enigmatic eukaryote, *Diphylleia rotans*. *Genome Biol Evol.* 8(2), 458-66.

Karnkowska A, Vacek V, Zubacova Z, Treitli SC, Petrzelkova R, Eme L, Novák L, Žárský V, Barlow LD, Herman EK, Soukal P, Hroudová M, Doležal P, Stairs CW, Roger AJ, Eliáš M, Dacks JB, Vlček C, Hampl V. 2016. A Eukaryote without a Mitochondrial Organelle. *Current Bio.* 26 (10), 1274-1284.

Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. *Nucleic Acids Res.* 33, 511-518.

Kayal E, Bentlage B, Collins AG, Kayal M, Pirro S, Lavrov DV. 2012. Evolution of linear mitochondrial genomes in medusozoan cnidarians. *Genome Biol. Evol.* 4(1), 1-12.

Kiethega GN, Yan Y, Turcotte M, Burger G. 2013. RNA-level unscrambling of fragmented genes in Diplonema mitochondria. *RNA Biol.* 10(2), 301-313.

Knoop V, Rudinger M. 2010. DYW-type PPR proteins in a heterolobosean protist: plant RNA editing factors involved in an ancient horizontal gene transfer? *FEBS Lett*. 584, 4287–4291.

Knoop V. 2011. When you can't trust the DNA: RNA editing changes transcript sequences. *Cell Mol Life Sci.* 68, 567–58.

Kuroiwa T. 2001. The complete DNA sequence of the mitochondrial genome of *Physarum polycephalum*. *Mol Gen Genet.* 264, 539-545.

Lang BF, Burger G, O'Kelly CJ, Cedergren R, Golding GB, et al. 1997. An ancestral mitochondrial DNA resembling a eubacterial genome in miniature. *Nature* 387:493-497.

Leger M, Eme L, Hug L, Roger AJ. 2016. Novel hydrogenosomes in the microaerophilic jakobid *Stygiella incarcerate. Mol Biol Evol.* doi:10.1093/molbev/msw103.

Lonergan, K., and M. W. Gray. 1993. Editing of transfer RNAs in *Acanthamoeba castellanii* mitochondria. *Science* 259, 812–816.

Lin S, Zhang H, Spencer DF, Norman JE, Gray MW. 2002. Widespread and extensive editing of mitochondrial mRNAs in dinoflagellates. *J Mol Biol.* 330, 727-739.

Lartillot N, Lepage T, Blanquart S. 2009. PHYLOBAYES 3: a Bayesian software package for phylogenetic reconstruction and molecular dating. *Bioinformatics* 25, 2286 – 2288.

Lowe TM, Eddy SR. 1997. tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res.* 25, 955–964.

Lukes J, Hashimi H, Zıkova A. 2005. Unexplained complexity of the mitochondrial genome and transcriptome in kinetoplastid flagellates. *Curr Genet.* 48, 277–299.

Madison-Antenucci S, Grams J, Hajduk SL. 2002. Editing machines: the complexities of trypanosome RNA editing. *Cell* 108, 435-438

Mahendran R, Spottswood MS, Ghate A, Ling M, Jeng K, Miller DL. 1994. Editing of the mitochondrial small subunit rRNA in *Physarum polycephalum*. *EMBO J*. 13(1), 232-240.

Makiuchi T, Nozaki T. 2014. Highly divergent mitochondrion-related organelles in anaerobi parasitic protozoa. *Sci Direct.* 100, 3-17.

Marande W, Lukes J, Burger G. 2005. Unique mitochondrial genome structure in diplonemids the sister group of kinetoplastids. *Eukaryot Cell.* 4(6), 1137-1146.

Nguyen LT, Schmidt HA, Haeseler AV, Minh BQ. 2014. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol.* 32(1), 268-74.

Panek T, Silberman JD, Yubuki N, Leander BS, Cepicka I. 2012. Diversity, evolution and molecular systematics of the *Psalteriomonadidae*, the main lineage of anaerobic/microaerophilic heteroloboseans (Excavata: Discoba). *Protist* 163, 807–831.

Park JS, Simpson AGB. 2011. Characterization of *Pharyngomonas kirbyi* (= Macropharyngomonas halophila" nomen nudum), a very deep-branching, obligately halophilic heterolobosean flagellate. *Protist* 162 (5), 691–709.

Palmer JD, Thompson WF. 1982. Chloroplast DNA rearrangements are more frequent when a large inverted repeat sequence is lost. *Cell Biol*. 29(2), 537-550.

Pawlowski J. 2014. Protist Evolution and Phylogeny. John Wiley & Sons, Ltd: Chichester. doi: 10.1002/9780470015902.a0001935.pub2.

Price DH, Gray MW. 1999. A novel nucleotide incorporation activity implicated in the editing of mitochondrial transfer RNAs in *Acanthamoeba castellanii. RNA* 5, 302-317.

Rhee AC, Somerlot BH, Parimi N, Gott JM. 2009. Distinct roles for sequences upstream and downstream from Physarum editing sites. *RNA* 15, 1753-1765.

Rio DC, Ares MJ, Hannon GJ, Nilsen TW. 2010. Purification of RNA using TRIsol (TRI reagent). *Cold Spring Harb Protoc.* 6, pdb.prot5439.

Shikanai T. 2006. RNA editing in plant organelles: machinery, physiological function and evolution. *Cell Mol Life Sci.* 63, 698-708.

Simpson AGB, Inagaki Y, Roger AJ. 2006. Comprehensive multigene phylogenies of excavate protists reveal the evolutionary positions of "primitive" eukaryotes. *Mol. Biol. Evol.* 23, 615-625.

Sloan DB, Alverson AJ, Chuckalovcak JP, Wu M, McCauley DE, et al. 2012. Rapid evolution of enormous, multichromosomal genomes in flowering plant mitochondria with exceptionally high mutation rates. *PLOS Biol.* 10, e1001241.

Spencer DF, Gray MW. 2011. Ribosomal RNA genes in *Euglena gracilis* mitochondrial DNA: fragmented genes in a seemingly fragmented genome. *Mol Genet Genomics* 285, 19-31.

Stairs CW, Leger MM, Roger AJ. 2015. Diversity and origins of anaerobic metabolism in mitochondria and related organelles. *Philos Trans R Soc Lond B Biol Sci.* 370(1678), 20140326

Stechmann A, Cavalier-Smith T. 2002. Rooting the eukaryote tree by using a derived gene fusion. *Science* 297, 89–91.

Takano T, Abe T, Sakurai R, Moriyama Y, Miyazawa Y, et al. 2001. The complete DNA sequence of the mitochondrial genome of *Physarum polycephalum*. *Mol Gen Genet.* 264, 539-545.

Tovar J, Fischer A, Clark CG. 1999. The mitosome, a novel organelle related to mitochondria in the amitochondrial parasite Entamoeba histolytica. *Mol Microbio.* 32 (5), 1013–21.

Traphagen SJ, Dimarco MJ, Silliker ME. 2010. RNA editing of 10 *Didymium iridis* mitochondrial genes and comparison with the homologous genes in *Physarum polycephalum*. *RNA* 16, 828–838.

Unseld M, Marienfeld JR, Brandt P, Brennicke A. 1997. The mitochondrial genome of *Arabidopsis thaliana* contains 57 genes in 366 924 nucleotides. *Nature Genet.* 15, 57–61.

Vaidya AB, Akella R, Suplick K. 1989. Sequences similar to genes for two mitochondrial proteins and portions of ribosomal RNA in tandemly arrayed 6-kilobase-pair DNA of a malaria parasite. *Mol Biochem Parasitol.* 35, 97-107

Visomirski-Robic LM, Gott JM. 1997. Insertional editing of nascent mitochondrial RNAs in *Physarum*. *PNAS.* 94, 4324-4329.

Visomirski-Robic LM, Gott JM. 1995. Accurate and efficient insertional RNA editing in isolated *Physarum* mitochondria. *RNA* 1, 681-691.

Yagi Y, Tachikawa M, Noguchi H, Satoh S, Obokata J, Nakamura T. 2013. Pentatricopeptide repeat proteins involved in plant organellar RNA editing. *RNA Biol*. 10 (9), 1419-25.

Yarza P, Yilmaz P, Pruesse E, Glochner FO, Ludwig W, et al. 2014. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews.* 12, 635-645.

Yubuki N, Leander BS. 2008. Ultrastructure and molecular phylogeny of *Stephanopogon minuta*: an enigmatic microeukaryote from marine interstitial environments. *Eur J Protistol.* 44, 241–253.

Zehrmann A, Verbitskiy D, Hartel B, Brennicke A, Takanaka M. 2011. PPR proteins network as site-specific RNA editing factors in plant organelles. *RNA Biol*. 8, 67-70.

APPENDIX A Supplementary Material for Chapter 2

Table S1. Primers used to link contigs.

Name	Primer (5′ – 3′)
B1	GAGGTGCCCTCTTCCTTTCT
B2	TGGTAACTTCCAAATCACCTCT
B3	TTTACGTTTGTTAGAGCGAAGTC
B4	GGGACTTGAACCGCAACTTC
B5	CGTTCATGCAATGAAAACAGAAGC
B6	CGGCTGATTACACCTCCGTG
P1	AGTCTCTCTAGCCGGTATTAGT
P2	TGCCAAGACGAATACCTGACT
P3	GTGACGTACGAAGTATGTTCCCT
P4	GTACTCCAGTGGCTCCTACG
P5	GGACATCATAGAAACATAATCTAACCTATC
P6	CATCAAATCTAAATCACTATACTCATATG

Table S2. Primers used to verify RNA editing sites in amoeba BB2

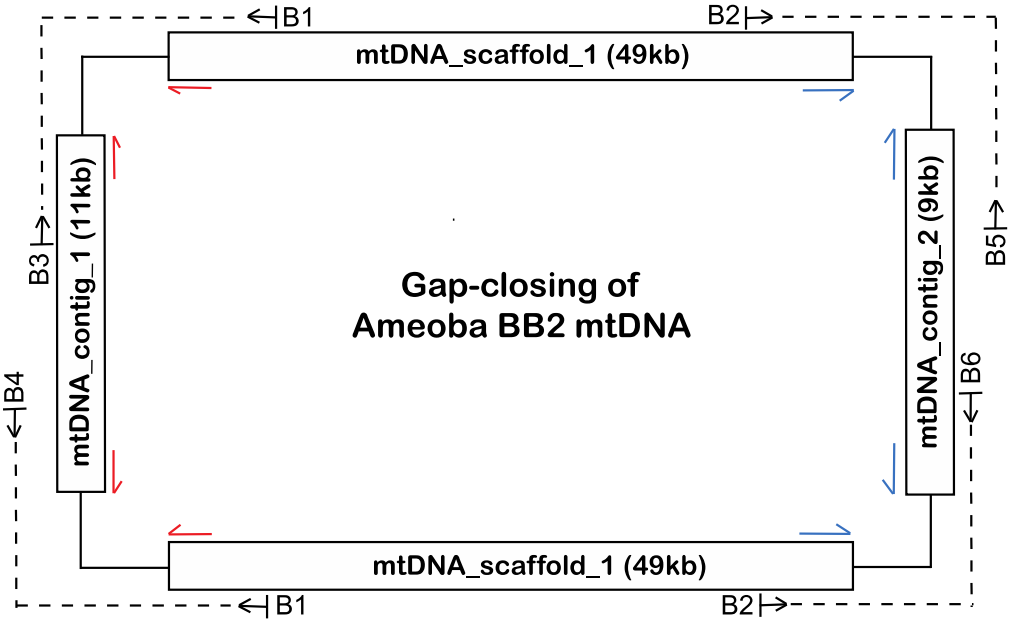
Name	Forward primer (5′-3′)	Reverse primer (5′-3′)	Position on mtDNA
B1	GCCAGGAAGTGGTGGTATTCT	CCGCTATTCGACGCAAAAACA	30511-31294
Br5	AGTGCGACTAGTGCTGAACC	GATGCGATTGTTGAAGGCCC	33489-34977
Br7	CGGCTACCGTCTAAACGAGG	GCAACTTCTTTGCAGTTTCCTG	35051-35692
Br9	GTTCTCAAGCGTTGGACCGT	AAGTGAAAATGCAAAATCGAAGTCA	36580-37216
Br13	TCTTTATGTTCATTGCGCTATTTGT	ACGTTTTCCGTTCTTAAAAGTCCA	38703-40109
Br15	TGGACTTTTAAGAACGGAAAACG	TCAAACATTCATTACACACAGCTT	40086-41253
Br21	AGGTTTGCCAAAGACGGTGA	TTCCACATACGTGGGTTCGG	55272-55645
Br23	GTCCATGTAGGTACGCCGAA	AACGCCCTTCGCTATAAGCA	61595-62301

Table S3. The number of editing sites for each RNA by the types of nucleotides.

Name	Length (bp)	A	U	G	C	Total
atp1	2442	5	0	13	1	19
atp3	861	4	0	6	1	11
atp6	756	1	0	4	1	6
atp8	381	0	0	1	0	1
atp9	231	0	0	2	0	2
cob	1203	0	1	9	0	10
cox1	2016	1	1	23	2	27
cox11	546	0	0	1	0	1
cox2	744	0	0	7	1	8
cox3	1455	0	0	11	0	11
nad1	1281	0	0	11	1	12
nad11	2112	0	0	14	3	17
nad2	1533	0	3	8	0	11
nad3	501	0	0	3	0	3
nad4	1551	0	0	12	0	12
nad4L	345	0	0	3	0	3
nad5	1986	1	0	16	1	18
nad6	642	0	0	6	1	7
nad7	1203	0	0	12	2	14
nad8	483	1	0	4	1	6
nad9	735	2	0	2	0	4
rnl	2840	0	0	20	2	22
rns	1656	0	1	16	2	19
rpl11	516	1	0	2	0	3
rpl14	369	1	0	4	0	5
rpl16	396	0	0	3	0	3
rpl2	807	0	0	11	0	11
rpl5	525	0	0	5	0	5
rpl6	687	0	0	7	0	7
rps10	648	0	0	2	0	2
rps11	2139	0	0	14	0	14
rps12	363	1	0	3	0	4
rps13	444	0	0	3	0	3
rps14	306	1	0	3	0	4
rps2	993	0	0	5	0	5
rps3_c	1008	0	0	5	0	5
rps3_n/rps19	1419	1	0	10	1	12
rps7	930	0	0	7	0	7
rps8	396	0	1	2	0	3
sdh2	843	0	0	6	1	7
trnC(gca)	71	0	0	1	0	1
trnD(guc)	74	0	0	1	0	1

trnF(gaa)_1	72	0	0	1	0	1
trnG(ucc)	71	0	0	1	0	1
trnI(cau)	72	0	1	0	0	1
trnI(gau)	73	0	0	1	0	1
trnK(uuu)	72	0	0	1	1	2
trnL(uaa)	84	0	0	1	0	1
trnP(ugg)	74	0	0	1	0	1
trnR(ucu)	74	0	0	1	0	1
trnS(gcu)	85	1	0	0	0	1
trnY(gua)	83	0	0	1	0	1
URF1274	3825	3	1	22	0	26
URF164	495	0	0	4	0	4
URF219	660	1	0	1	0	2
URF2371	7116	8	4	49	3	64
URF262	786	1	1	3	1	6
URF417	1254	3	1	4	0	8
URF467	1404	0	0	8	0	8
Total		37	15	397	26	475

(A)



(B)

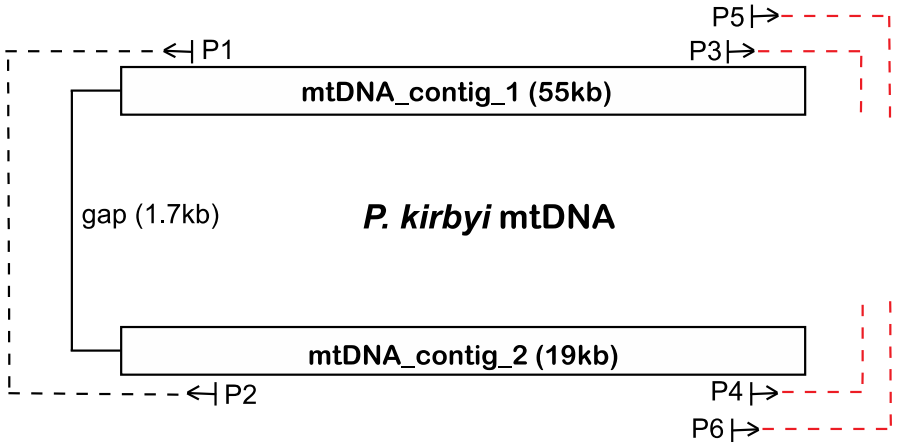


Figure S1. (A) A schematic map of gap-closing of amoeba BB2 based on long-range PCR. Red (about 80bp-long) and blue (about 200bp-long) arrows are repeated regions between and within contig and scaffolds. (B) A schematic map of gap-closing of *P. kirbyi* mtDNA based on long-range PCR. The set of PCR reactions that did not yield products is highlighted in red dashed lines.

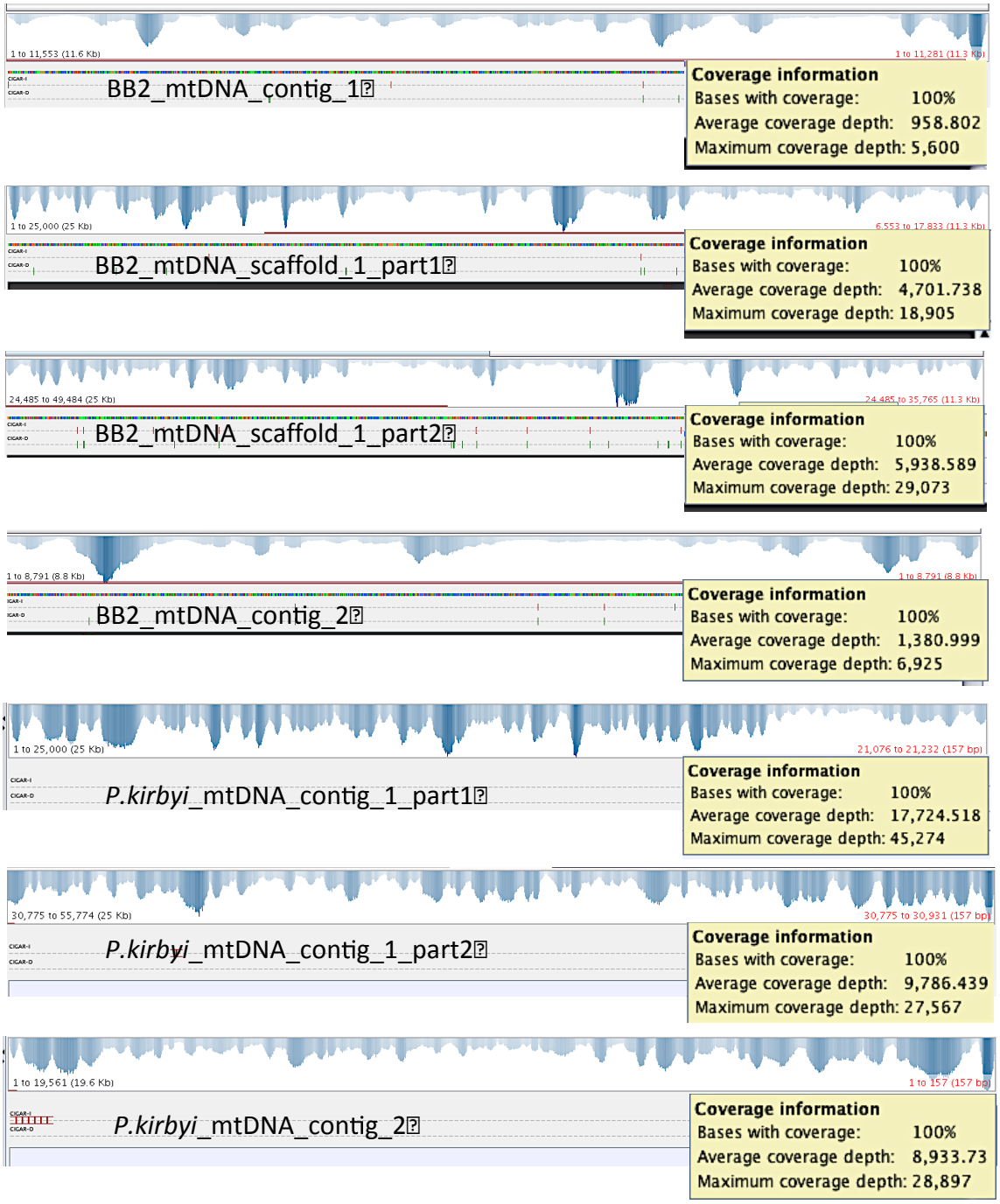


Figure S2. Assembly coverage plots for the mitochondrial genome of BB2 and *P. kirbyi* by contigs (shown in Figure S1).

	S10													Spc
Rickettsia	rps10	– – –	rpl3	rpl4	rpl23	rpl2	rps19	rpl22	– – –	rps3	rpl16	rpl29	rps17	rpl14
Andalucia	rps10	– – –	– – –	– – –	– – –	rpl2	rps19	– – –	– – –	rps3	rpl16	– – –	– – –	rpl14
Tsukubamonas	– – –	– – –	– – –	– – –	– – –	rpl2	rps19	– – –	– – –	rps3	rpl16	– – –	– – –	rpl14
Amoeba BB2	rps10	– – –	– – –	– – –	– – –	rpl2	rps19	– – –	urf219	rps3	rpl16	– – –	– – –	rpl14
Pharyngomonas	rps10	– – –	– – –	– – –	– – –	rpl2	rps19	– – –	– – –	rps3	rpl16	– – –	– – –	rpl14
Naegleria	rps10	rpl11	– – –	– – –	– – –	rpl2	rps19	– – –	– – –	rps3	rpl16	– – –	– – –	rpl14

	Spc										Alpha			
Rickettsia	rpl24	rpl5	rps14	rps8	rpl6	rpl18	rps5	rpl30	rpl15	secY	adk	rps13	rps11	rpoA
Andalucia	– – –	rpl5	rps14	rps8	rpl6	rpl18	– – –	– – –	– – –	– – –	– – –	rps13	rps11	rpoA
Tsukubamonas	– – –	rpl5	rps14	rps8	rpl6	– – –	– – –	– – –	– – –	– – –	– – –	rps13	rps11	– – –
Amoeba BB2	– – –	rpl5	rps14	rps8	rpl6	– – –	– – –	– – –	– – –	– – –	– – –	rps13	rps11	– – –
Pharyngomonas	– – –	rpl5	rps14	rps8	rpl6	– – –	– – –	– – –	– – –	– – –	– – –	rps13	rps11	– – –
Naegleria	– – –	rpl5	rps14	rps8	rpl6	– – –	– – –	– – –	– – –	– – –	– – –	rps11	rps13	– – –

Figure S3. Gene order comparison of Discoba mtDNA and α-proteobacteria *Rickettsia prowazekii* in the three contiguous ribosomal protein operons (S10-Spc-Alpha). The gene split into two different orf is highlighted with red color, and genes in reversed order are shown in green.

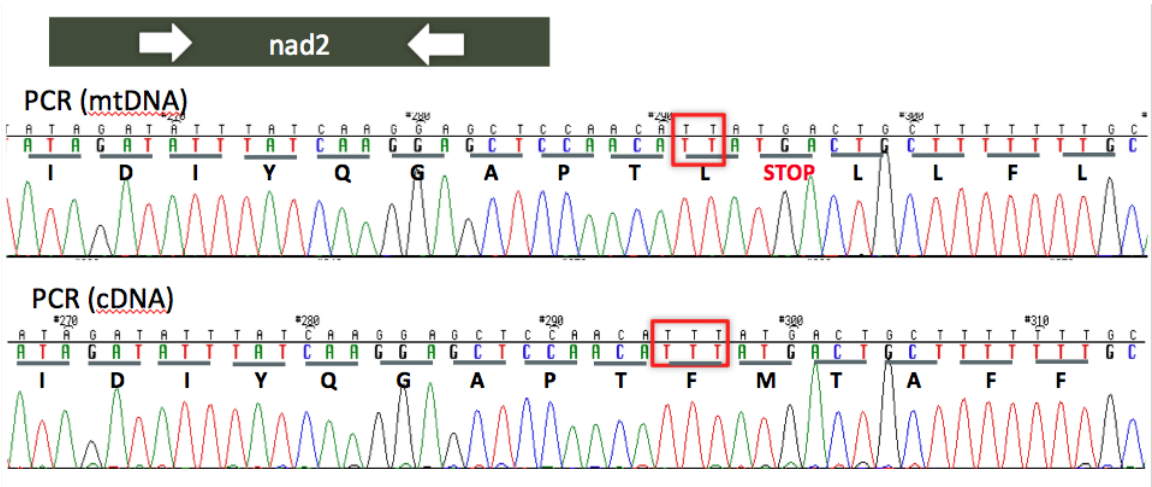


Figure S4. A chromatogram of PCR products from both mtDNA and cDNA of ORF homologous to *nad2* gene. Amino acid sequences are written below nucleotide sequences. Location of the exact insertion site is ambiguous due to uracil insertion next to encoded uracils, therefore, this site is indicated by red boxes.

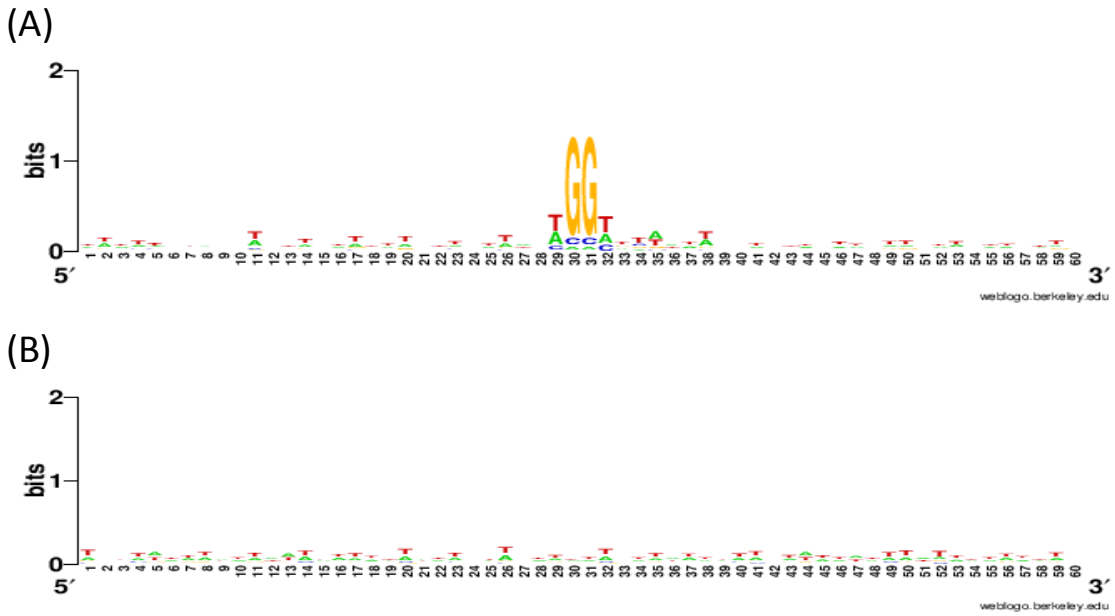


Figure S5. Sequence LOGO of (A) 311 60-nucleotide-long sequences near editing sites and (B) 328 random 60-nt-long sequences.