REPETITIVE DNA AND NUCLEAR INTEGRANTS OF ORGANELLAR DNA SHAPE THE EVOLUTION OF COCCIDIAN GENOMES

by

SIVARANJANI NAMASIVAYAM

(Under the Direction of Jessica C. Kissinger)

ABSTRACT

The greatest diversity in eukaryotic genomes is observed in the deep branching protist phyla. The protist phylum Apicomplexa consists of at least 5,000 species of mostly obligate intracellular parasites, many of medical and veterinary importance. A number of apicomplexan genomes have been sequenced, providing us with a rich resource for comparative evolutionary studies. Apicomplexans have reductive genomes, yet, they show immense genome diversity and innovation. This diversity is captured in the parasites of the coccidian lineage. *Sarcocystis neurona* has the largest sequenced apicomplexan genome at >125 Mb, twice as large as the next largest genome from *Toxoplasma gondii*, the highly-successful zoonotic pathogen. We find that repeats are responsible for a large *S. neurona* genome, however the out-group parasite *Eimeria* while repetitive has a ~55 Mb genome. The in-group parasites, *T. gondii*, *Neospora caninum* and *Hammondia hammondi* are repeat-poor. Instead, they contain unprecedented levels of organellar DNA insertions; NUM/PTs. While repeats shape the genomes of the early-branching coccidians, our study suggests NUM/PTs to be the significant drivers of genome evolution in later-branching coccidians.

INDEX WORDS: Apicomplexa, evolution, NUMTs, NUPTs, *T. gondii*, *S. neurona*

REPETITIVE DNA AND NUCLEAR INTEGRANTS OF ORGANELLAR DNA SHAPE THE EVOLUTION OF COCCIDIAN GENOMES

by

SIVARANJANI NAMASIVAYAM

B. TECH, Vellore Institute of Technology, India, 2008

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2015

REPETITIVE DNA AND NUCLEAR INTEGRANTS OF ORGANELLAR DNA SHAPE THE EVOLUTION OF COCCIDIAN GENOMES

by

SIVARANJANI NAMASIVAYAM

Major Professor:	Jessica C. Kissinger
Committee:	Jeffrey L. Bennetzen
	Michael J. McEachern
	Boris Striepen
	Chung-Jui Tsai

Electronic Version Approved:

Suzanne Barbour
Dean of the Graduate School
The University of Georgia
August 2015

ACKNOWLEDGEMENTS

I would like to thank my advisor Jessie and my committee members Jeff, Mike, CJ and Boris for providing useful suggestions and feedback about my research. Committee meetings were always productive and encouraging. My committee helped me to think critically about my work, which has helped me grow as a scientist.

Jessie has been an excellent mentor. She has encouraged me to think independently and look at the big picture always keeping the question I am trying to answer in mind. She has sent me to a number of conferences. Presenting at these meetings and interacting with other scientists has helped me grow in confidence. I also want to thank her for greatly improving my scientific writing and presentation skills. She has always encouraged my interests and supported me to attend the Biology of Parasitism course. Attending this course has truly been a rewarding experience.

The Kissinger lab has been a positive environment to work in. Everyone is willing to help and share his or her knowledge. I must thank the past members of the lab from whose work and advice I have greatly benefitted. I also want to thank the Genetics Department and members of the CTEGD, particularly the Striepen lab. It has been a fruitful and enjoyable graduate school journey.

TABLE OF CONTENTS

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION AND LITERATURE REVIEW

THE APICOMPLEXA

The phylum Apicomplexa (supergroup Chromalveolata) consists of at least 5000 species of mostly obligate, intracellular protist parasites, many with significant medical and veterinary relevance (1, 2). The phylum was thought to consist of only parasites until the discovery of marine symbionts (3). Recently a number of apicomplexan parasites and symbionts of marine animals have been reported, suggesting there may be significantly more than 5000 species in this phylum (4, 5). Examples of apicomplexan-derived human diseases include malaria (caused by *Plasmodium*), toxoplasmosis (*Toxoplasma*) and cryptosporodosis (*Cryptosporidium*). Diseases caused by *Babesia*, *Theileria* and *Eimeria* have a great impact on the cattle and poultry industry. Given their significant importance to public health, it is not surprising that the genomes of a number of apicomplexans have been sequenced. The sequenced genomes, in addition to being important resources for diagnostics and the biology of disease, also serve in the study of genome evolution. The phylum Apicomplexa is ancient, estimated to have diverged between ~350-820 Mya (6, 7). It is difficult to date the age of the phylum since there are no fossil records for the apicomplexans. Apicomplexan genomes range from ~8.5 Mb in the genus *Theileria* to >125 Mb in the genus *Sarcocystis* (presented here) (Figure 1.1, Table 1.1) and are highly reduced even in comparison to genomes of other parasitic organisms (Figure 1.2). The genome organization and gene content also show variation across the phylum (Table 1.1). These genomes are characterized by genome rearrangement (8, 9), gene loss (10), and intracellular and lateral gene

transfers (11-16). The deep branching nature of this phylum, its diversity, its gene- and genome-level innovations combined with differential gene retention relative to the closest free-living ancestor (8, 9) makes the Apicomplexa an excellent system for the study of genome evolution.

Apicomplexans are mostly obligate intracellular parasites. They proliferate by invading a host cell, replicating and dividing inside that host cell, followed by lysis and egress from the host cell and reinvasion of new host cells. They can also differentiate into gametes or other developmental forms that invade distinct tissues (17-19). The parasites possess a number of distinct features and organelles that aid their parasitic lifestyle. The most distinguishing feature of this phylum is the apical complex and its collection of unique organelles (Figure 1.3). Rhoptries, micronemes and dense granules are organelles that hold important proteins that play a role in host cell invasion, attachement, motility and formation of the parasitophorous vacuole. (19-21). The apical complex also contains a cone-shaped structure called the conoid (22). The parasite is enclosed by a plasma membrane and an inner membrane complex that is associated with actin, myosin and microtubules among other structures and proteins (19, 23-25). Most apicomplexans also possess a mitochondria and a relic plastid called the apicoplast (26-28). The mitochondria and apicoplast contain their own genomes (discussed below).

PURPOSE OF THIS STUDY

The Coccidia are a sub-class of parasites in the Apicomplexa that affect the intestinal tracts of a wide range of animals (29). The coccidians have the largest observed genomes among the Apicomplexa (Table 1.1). Their genome sizes range from ~44 Mb in some *Eimeria* species to >125 Mb in *Sarcocystis neurona*. Transposable elements, which are important evolutionary forces in many eukaryotic genomes, are noticeably absent in most apicomplexans and only a few have been reported. Four families of retrotransposons were observed in *Ascogregarina*, a

gregarine that infects mosquito larvae and *Eimeria* (30, 31) and we have also identified them in *Sarcocystis*. In this dissertation I present my findings on factors that impact the evolution of coccidian genomes as represented by *Toxoplasma gondii* and *Sarcocystis neurona* with comparisons made to related coccidians *Neospora caninum*, *Hammondia hammondi* and *Eimeria tenella* as necessary. In this chapter, I review relevant information on the evolution of eukaryotic, apicomplexan and organellar genomes. The coccidian parasites are discussed in further detail.

EVOLUTION OF EUKARYOTIC GENOMES

With the advancement of sequencing technologies, there has been a steep increase in genome sequencing projects covering more phylogenetically distinct organisms. The availability of numerous genome sequences has paved the way for comparative genomics and a better understanding of eukaryotic genome evolution. The genome sizes of eukaryotes vary widely and do not correlate with organism size or complexity, called the C-value paradox (32). The gene content and genome architecture can vary even among closely related species. A number of forces have influenced eukaryotic genome evolution, including intracellular and lateral gene transfers, genome duplications and repetitive content.

The endosymbiotically acquired mitochondrial and plastid genomes have had a great impact on eukaryotic nuclear genomes. A larger number of genes from these organelles have been transferred to the nuclear genome through a process called intracellular gene transfer (33, 34) (discussed below). In yeast, nearly 75% of its genes are thought to be mitochondrial in origin (35). Lateral or horizontal gene transfers that are common in prokaryotes have also been reported in eukaryotes (36-39). Feschotte *et. al.,* have reported DNA type transposons (SPIN) to be involved in horizontal transfers in tetrapods and mammals (40). They have also demonstrated the role of host-parasite interactions in horizontal transfers (41). Horizontal gene transfers have also

been reported in the apicomplexans including the acquisition of a nucleotide biosynthesis salvage pathway protein, thymidine kinase, in *C. parvum* (14). The thymidine kinase was acquired as a result of horizontal gene transfer from a bacterium likely, α or γ-proteobacterium.

Genetic duplications in the form of entire genome duplication, segmental duplications or short tandem duplications are observed in almost all eukaryotes. Polyploid genomes are particularly frequent in flowering plants and these genome duplications partly contribute to their large genome sizes and provide a source of innovation. Paralogs and multi-gene families are also a result of smaller-scale duplications. The genes in duplicated regions may acquire new functions (neo-functionalization), split functions with its paralogs (sub-functionalization) or (most common) just accumulate mutations and become a pseudogene.

Mobile and/or repetitive sequences are one of the important driving forces in eukaryotic evolution. Repetitive sequences are ubiquitous in eukaryotic genomes. They can be simple tandem repeats, in the form of mini- and microsatellites, and/or dispersed repeats, which include paralogs, pseudogenes or the infamous transposable elements (32, 42). Transposable elements are further divided into LINEs, SINEs, LTR retrotransposons and DNA transposons (32, 42). Transposable elements are notorious for bringing about large-scale genome rearrangements, many by ectopic recombination, and are responsible for shaping the genomes of many eukaryotes. Often regarded as selfish DNA, many TEs have been co-opted by the resident nuclear genome for evolution of important gene regulatory functions, including epigenetic regulation in some plant genes (43).

Mitochondrial genome diversity in eukaryotes

The leading theory concerning the origin of mitochondria suggests a monophyletic event resulting from the endosymbiotic association of an alpha-proteobacterium (44). The most

concrete evidence in support of this theory came from sequencing of the alpha-proteobacterium *Rickettisia* and what is considered to be the most bacteria-like mitochondrial genome of the eukaryote, *Reclinomonas americana*. The mitochondrial (mt) genome of *R. americana* is found to have the largest number of conserved genes among the sequenced mitochondrial genomes, a total of 97, and it contains all mitochondrial protein coding genes found in all other sequenced organisms to date. In some organisms, the order of protein-coding genes appears to be retained in comparison to the bacterial ancestor and mitochondrial-specific deletions are thought to have occurred in the common mitochondrial ancestor before divergence, since, similar patterns of deletion have been observed. Mitochondrial genes have been found in the nuclear genome of some of the amitochondriate protists as well (44). This finding combined with the shared primitive nature of many of the protist mitochondrial genomes strongly supports a single origin for the mitochondrion.

Despite its single origin, mitochondrial genome structure and content varies hugely in plants, animals, fungi and protists. Until recently, it was believed that mitochondrial genome sequences occurred only as circular molecules, reminiscent of their bacterial ancestry. However, linear concatemers and monomers often with telomere-like terminal repeats have been observed (45, 46). Mitochondrial genomes with multiple circular molecules have been noted in the fungus *Spizellomyces (47)*. The kinetoplastid mitochondrial genome, called the kinetoplast, contains multiple gene-encoding maxi-circles and guide RNA encoding mini-circles that form a tight network (48). The mtDNA of the single-celled protist *Amoebidium parasiticum* is comprised of hundreds of distinct types of linear molecules with a common pattern of terminal repeats (45). How the replication and equal segregation occurs in such cases is not clear. A rolling-circle mechanism is the preferred mode of replication for lariat molecules (or linear concatamers). This

mechanism has been observed in yeast (49). Transitions in topology from linear to circular have been observed in yeasts, golden algae and fungi within short evolutionary times (45).

Huge variation is also seen in the gene content and size of mitochondrial genomes. Their size varies from just 6 Kb in *Plasmodium* to a few hundred Kb in some plants. However there is not always a correlation between size and content; plants have larger mitochondrial genomes because of larger introns, intergenic regions and repeats but they can have fewer genes when compared to mammals (45). Fused genes forming single contiguous transcripts have been observed in two amoebozoan protists, *Acanthomeba castellanni* and *Dictyostelium discoideum*. Split genes are also seen. The mitochondrial rRNA genes of *Plasmodium* show the highest degree of fragmentation ever observed (50).

Many organisms contain more than one mitochondrion organelle per cell and more than one copy of the mitochondrial genome inside each organelle. The presence of more than one genome copy provides scope for mutations to accumulate without necessarily causing deleterious effects on the cell, this in turn allows for more evolutionary 'experiments'. The same cannot be said for nuclear genomes; where only one copy is present per cell except in a polyploid. The rate of evolution is different in the mitochondrial and nuclear genomes. In plants, the mitochondrial genome evolves at a slower rate than the nuclear genome and in animals the mitochondrial genome evolves faster than the nuclear genome (45). In plants, the mitochondrial and nuclear genomes are open to the uptake of foreign DNA as in seen from the presence of repeats whereas, in animals, the mitochondrial genome appears to be closed to foreign DNA integration. The above factors and possibly others have contributed to highly-diversified mitochondrial genomes.

Mitochondrial DNA transfer to the nuclear genome

Gene loss or transfer of mitochondrial genetic material to the nuclear genome is an important phenomenon resulting in reduction of the mitochondrial genome size. Almost all eukaryotic nuclear genomes bear evidence of gene transfers from the mitochondrial genome. In most eukaryotes, the mitochondrial proteome does not correlate with its gene content because, many of the mitochondrial proteins are products of nuclear genes. While 'actual' nuclear genes code some of the mitochondrial proteins, in many cases the mitochondrial genes that have been transferred to the nuclear genome have evolved to target their protein back to the mitochondria. A few factors can be attributed to why a gene is lost from the mitochondria; (i) since the endosymbiont is no longer free living it probably lost many genes whose function is not required for the current lifestyle; (ii) the function is already coded by the nuclear genome and redundancy is not necessary and (iii) in the case of unique genes coded only by the mitochondria, like genes involved in electron transport and oxidative phosphorylation, have also been transferred to the nucleus; such transfers could have occurred to reduce redundancy in the genome maintenance machinery (51). In such cases the genes need to be targeted back to the mitochondria to function. A theoretical model suggests that the rate of transfer of mtDNA depends on the effective population size, intensity of intracellular competition and probability of paternal organelle transmission (52). In many eukaryotes, including animals, transfer of functional genes is rare or has ceased. However, flowering plants show evolutionarily recent transfers of certain mitochondrial genes (53).

Evolution of plastid genomes and plastid DNA transfer to the nucleus

The plastid is also widely believed to have a monophyletic origin (54) from an endosymbiosis between a photosynthetic bacterium and a non-photosynthetic host (55, 56).

Present day plastid genome sizes are typically 120-190 Kb (57). As for the mitochondrion, there is clear evidence of the transfer of genes from the plastid to the nucleus. The number of genes ranges from 15-209 genes (58-60), depending on the plant lineage. The photosynthetic bacterium which was engulfed to give rise to the plastid is expected to have contained ~3,200 genes, similar to the genome sizes of extant cyanobacteria (56). While many genes acquire transit peptides and are targeted back to the plastid, it is not always the case. For example, ~18% of the genes in *Arabidopsis thaliana* have a plastid origin but about half of these genes perform non-plastid functions (53, 61). The reasons attributed to the movement of mitochondrial genes to the nucleus and the mechanisms suggested for transfer and integration into the nuclear genome (discussed below) hold true in case of the plastids as well.

NUCLEAR INTEGRANTS OF MITOCHONDRIAL ORIGIN (NUMTS)

Transfer of random organellar DNA fragments (average size ~250 bp), as opposed to whole genes as described above, to the nuclear genome is still an active process in plants, animals, fungi and protists (62). These random fragments are called NUMTs (NUclear integrants of MiTochondrial origin). NUMTs are insertions of DNA sequences into the nuclear genome that were derived from mtDNA. A NUMT can originate from any region of the mitochondrial genome and can be of any size up to the size of the largest mtDNA molecule. A NUMT can arise as a result of independent insertion from the mitochondrion or from duplications in the nuclear genome following the intital insertion event. NUMTs are found in almost all eukaryotes. There is considerable variation in the nuclear NUMT density across eukaryotes ranging from no detected insertions in some protists and animal genomes to 0.2% in certain plants and fungi (63). The factors that govern these transfers are not clear. The honeybee genome is the current record holder for the largest number of insertions (1,380) seen in any metazoan (64). The human

genome is estimated to contain 452 NUMTs. MtDNA is routinely used for phylogenetic classification. NUMTs can pose a challenge for these studies, since NUMTs are derived from mt DNA and depending on sequence similarity to the donor mtDNA (NUMTs decay over time) they can interfere with PCR amplifications. NUMTs have caused such errors in studies of arthopods (65). In humans, NUMTs have been misclassified as mtDNA sequence variation falsely implicating the mtDNA sequence in disease. NUMTs have also justly been associated with human disease (63). NUMTs have also been suggested as a new source of DNA for novel exons and splice-sites (66).

The mechanism of NUMT integration was first demonstrated in yeast and subsequently proven in mammalian cells and plants. It was observed that fragments of mtDNA were transferred into the site of repair of a double-strand break. Double-strand break repair (DBSR) through the non-homologous end joining (NHEJ) pathway has been cited as the mechanism of integration (67-69). NHEJ repairs double-strand breaks with little or no sequence homology between the termini. NHEJ often results in the deletion of sequences between the DSB primarily due to nuclease activity. Deletion of functional sequences can be deleterious to the cell. It has been proposed that mitochondrial DNA fragments (NUMTs) act as filler DNA reducing the incidence of deletions (69). This has been shown in human cells - when a NUMT is involved in the repair, deletion is not detected. Overall, based on experimental evidence from different organisms, DSBR via NHEJ appears to be a highly plausible mechanism for integration of mitochondrial genes and fragments in the nuclear genome.

How the mitochondrial DNA enters the nucleus is not very clear. A few theories have been proposed; (i) Disruption of the mitochondrial membrane during autophagy by lysosomes or vacuoles, fusion or division during cell stress. It has been shown that the mitochondria of male

gametes degenerate, during male gametogenesis in plants and penetration of the sperm into the ovum in mammals. Also, mutation in the yeast mitochondria escape (YME) protein results in the increased release of mitochondrial DNA; (ii) Illegitimate uptake through nuclear import machinery or escape using the RNA export machinery; (iii) Direct contact between the mitochondria and nucleus; (iv) mtDNA enters along with some proteins (53), perhaps NHEJ proteins (70).

Nuclear integrants of Plastid origin (NUPTs)

Similar to the NUMTs, NUPTs (NUclear integrants of PlasTid origin) are observed in plant nuclear genomes. The plant genomes contain the largest insertions of organellar DNA; *Arabidopsis* contains a 620 Kb NUMT (71) and rice contains a 131 Kb NUPT (72). A large proportion of the NUPTs in rice are located in pericentromeric regions (53). The NUPT profiles in rice and *Arabidopsis* are similar, suggesting that there was an ancient influx of NUPTs that decayed over time (73). Transposable elements can contribute significantly to NUM/PT decay and fragmentation via integration into NUM/PTs (74). NUM/PTs show variation even among closely-related species (63), suggesting the evolution of these insertions is a dynamic process that continues to have the potential to impact nuclear genome evolution.

EVOLUTION OF THE APICOMPLEXAN GENOME

Model organisms and economically-relevant organisms have enjoyed the benefit of whole-genome sequencing, while most eukaryotic diversity is seen in deep-branching protists (75). Apicomplexans constitute an ancient protist phylum that falls under the kingdom of Chromalveolata along with the sister phyla of the dinoflagellates and ciliates among others (76). Due to their medical and veterinary importance, many apicomplexan genomes have been

sequenced (Table 1.1) (EuPathDB.org), providing us with a rich resource to study evolution in an early-branching protist phylum.

Apicomplexans have highly-reduced genomes, a consequence of their parasitic lifestyle (9, 77, 78) (Figure 1.2). In an extreme case of gene loss, the early-branching apicomplexan *Cryptospordium parvum* has lost its ability for *de novo* nucleotide synthesis and salvages nucleotides from the host. The gene content in the phylum ranges from ~4000 genes in *Babesia* and *Cryptosporidium* to ~9000 genes in *Toxoplasma*. Only ~1000 genes are conserved across the phylum (10) and many genus-specific genes, particularly in processes involved with host cell invasion and survival, have evolved (2).

One of the most striking differences in the apicomplexan genomes is a dearth of TEs. They have been reported in only a few species (31, 79, 80). The repeat content widely varies across the phylum. The *P. falciparum* genome is 81% AT and contains tandem repeats in its coding regions (81). Homopolymeric repeats are present in *Eimeria* proteins and chromosomes but *Toxoplasma* is repeat-poor (79, 82). Apicomplexan genomes also lack synteny across the phylum, only order/family- or genus-specific synteny is observed (8). Other eukaryotic genomes of comparable divergence time do show synteny. The kinetoplastid parasites, *Leishmania* and *Trypanasoma,* which diverged ~200-500 mya, still maintain ~70% of their genomes in syntenic blocks, although this is likely an effect of polycistronic transcription (8, 83, 84).

Organellar genome sequences in apicomplexans

Apicomplexans have a distinct evolutionary history. They contain two extra-chromosomal genomes acquired through two distinct endosymboitic events; the genomes of the mitochondria and apicoplast (a relic plastid) (Figure 1.5). An exception is the early branching apicomplexan parasite *Cryptosporidium* that lacks an apicoplast organelle and has a reduced

mitochondrion called the mitosome that lacks a genome. However, evidence of organellar genes in the nuclear genome is noted in this genus of parasites as well (12).

The four-membraned apicoplast, a non-photosynthetic plastid, arose from a secondary endosymbiosis event (85) where an eukaryote was engulfed by another eukaryote, likely a red algae (86, 87). The apicoplast genome sequence is available for a number of apicomplexans (88-90). The apicoplast genome is AT-rich, occurs as a 22-35 Kb genome, smaller than typical chloroplast genomes (usually 120-190 Kb) (57). There is clear evidence of intracellular gene transfer from the apicoplast genome to the nuclear genome and a number of the proteins encoded by these genes are targeted back to the apicoplast via organellar transit signal peptides (91, 92). Apicoplast gene content is highly conserved across the phylum (93). It codes for large and small ribosomal subunits, tRNA genes and a number RNA polymerase beta (*rpo*) genes (88-90). The *Plasmodium* and coccidian apicoplast genomes contain an inverted repeat of the tRNAs, LSU and SSU rRNAs which gives it a cruciform shaped structure. This inverted repeat is absent in piroplasm apicoplast genomes (93). The apicoplast is essential for survival (94, 95). In *Plasmodium*, treatments with drugs that target the apicoplast prevent the parasite from completing its intra-erthrocyctic lifecycle (96-99) and, in *Toxoplasma,* cause a 'delayed-death' phenotype (94, 100). Since the apicoplast is of chloroplast ancestry, it serves as an excellent drug target for use in animal hosts and is an area of intense research.

Gene loss and transfer is most extreme in the mitochondrial genomes of the apicomplexans. All sequenced apicomplexan mitochondrial genomes encode only three protein coding genes: cyctochrome oxidase subunits I and III (*cox1* and *cox3*) and cytochrome b (*cob*). No tRNAs are found in the genome sequence, rather, they are imported from the cytosol for use in mitochondrial translation (101). rRNA genes are present, but the level of fragmentation seen

in the rRNA genes is most unusual (Figure 1.6A). Recently, 27 out of the 34 rRNA fragments in *P. falciparum* have been associated to the large or small ribosomal units using the *E. coli rrnB* operon (102). While transcripts of these fragments have been observed, it is not clear how they form a functional ribosome. The ribosome is functional since these parasites are not obligate anaerobes and a functional mitochondrial ribosome has been observed in *Chlamydomonas* which also has fragmented rRNA genes (45), although not to the degree seen in the Apicomplexa. A sister clade to the Apicomplexa, the dinoflagellates, also share this feature of reduced coding content and rRNA fragmentation suggesting this is an ancient characteristic. However, the dinoflagellates' mtDNA also contain numerous repeats and ORFs, making their mitochondrial genomes larger than the apicomplexan mitochondrial genomes (103).

Typically, apicomplexans contain only one mitochondrion per cell, but each can contain more than one copy of the mitochondrial genome. Genome sizes range from 6-8 Kb. While all sequenced apicomplexan mtDNAs contain only three protein-coding genes and show a high level of rRNA fragmentation, the orientation and arrangement of these genes varies across the phylum, as does the genome architecture (Figure 1.6B). The mtDNA of *P. falciparum* has been well studied. It occurs as a linear concatemer with each repeating unit measuring ~6 Kb. The nuclear and mtDNA start replicating at the same time (28). The linear molecules form multiple recombinatory interactions with other molecules and the circular forms replicate through a rolling circle mechanism (as seen *in S. cerevisiae* and some bacteriophages and bacterial plasmids) resulting in linear concatemers (28, 104). *Theileria* encodes a 7.1 Kb linear monomer with telomere-like termini (105). Members of the genus *Babesia* have linear monomers with a dual flip-flop inversion system that results in four different sequences that are present in equal ratios in the mitochondrion (106). The mitochondrial genome of *Eimeria tenella*, a coccidian,

occurs as a linear concatemer of 6.2 Kb similar to that of *Plasmodium (106)*. The mitochondrial genome of *Toxoplasma gondii* is discussed in detail below.

The Coccidia

The Coccidia are a sub-class of parasites in the phylum Apicomplexa. They typically infect the intestinal tract and some, like *Toxoplasma gondii,* are highly promiscuous in their ability to infect any nucleated cell in warm-blooded animals (107, 108). Like most other apicomplexans, they undergo their sexual phase in a definitive host and an asexual phase in an intermediate host. The coccidians of the family Sarcocystidae, which include *Sarcocystis*, *Neospora*, *Toxoplasma* and *Hammondia,* have a two-host or heteroxenous lifecyle, where they undergo the sexual cycle in a single, definitive host and the asexual cycle in many possible intermediate hosts (29). The *Eimeria, Isospora* and *Cyclospora* genera, however, complete both their sexual and asexual phases within a single host. Over 1,200 species of *Eimeria* have been described (109) and these collectively infect many vertebrates and invertebrates, however, each species is highly host-specific and replicates only within the intestinal epithelial cells (29). *Toxoplasma* and *Neospora* can cross the epithelial barrier into the blood stream and migrate to other tissue sites. Also, both of these parasites show vertical transmission from the mother to fetus (110, 111). *Toxoplasma* has evolved a new mode of transmission (oral infectivity via the consumption of raw infected tissues) and has become a generalist. *Toxoplasma gondii* and *Sarcocystis neurona*, which are the focus of this dissertation, are discussed in detail below.

Genome sequences for the coccidians *Toxoplasma gondii*, *Hammondia hammondi* *Neospora caninum, Sarcococystis neurona* and *Eimeria* (many species) are publicly available (EuPathDB.org). Eight species of *Eimera* have been sequenced and, through a collaborative effort, we have published the first genome sequence of *S. neurona*. The Coccidia are a deep-

branching lineage with an estimated divergence time of ~500 Mya (112). Some coccidians show a wide host range while some are restricted to specific hosts. The Coccidia also have the largest genomes observed thus far among the apicomplexans. All of these factors make the study of the evolution of coccidian genomes intriguing. Some interesting findings have already been reported.

Eimerian chromosomes have a unique feature. They have a series of alternating repeat-rich and repeat-poor regions (82), which are observed in all sequenced *Eimiera* species. However, the repeat content varies between species and there is also variation in the level of synteny. *E. tenella* and *E. necatrix* show a high level of synteny, whereas *E. tenella*, *E. maxima* and *E. acervulina* show significant rearrangement (79). The genome sequences of *T. gondii* and *N. caninum* are highly syntenic (8, 30) and are not repeat rich. In terms of the gene repertoire, there are significant differences, especially with respect to genes associated with host-specificity and pathogenesis or secreted pathogenesis determinants (2, 29, 30). The surface antigen domain-containing proteins are among the most divergent. These proteins are involved in modulating host immunity and promoting attachment to the host cell (113, 114). *T. gondii* and *N. caninum* have 104 and 227 SAGs, respectively (30), whereas the numbers in *Eimeria* species vary from 16 to 172 (79). The rhoptry kinase proteins that are involved in immune evasion (115-117) are reduced in *Eimeria*. Only 28 *ropk* genes are reported in *E. tenella* (79), whereas 68 are reported in *T. gondii* and *N. caninum* (30, 118). Other secreted proteins involved in invasion are targeted to organelles called micronemes, which are involved in host cell attachment and mediating gliding motility (119) and dense granules, which are associated with the protective parasitophorous vacuole (120). Except for a few differences, protein family members localized to

these organelles are fairly conserved among the coccidians but they do vary in a few significant ways (117).

Toxoplasma gondii

Toxoplasma gondii is one of the most successful zoonotic parasites, essentially capable of infecting any warm-blooded animal. It is found in nearly a third of the human population. It is an opportunistic pathogen causing toxoplasmosis in immune-compromised and AIDS patients. It can be transmitted congenitally to the fetus if the mother becomes infected during pregnancy, resulting in abortion or blindness and death of the newborn (121).

Like most other apicomplexans, the life cycle of *T. gondii* is complex, oscillating between the asexual and sexual stages (Figure 1.7A). The definite hosts are the felids; the only organisms in which the sexual stage has been identified. All other warm-blooded animals are intermediate hosts, where asexual replication takes place. It has three invasive forms, tachyzoites, bradyzoites (tissue cysts) and sporozoites (from oocysts). All three forms are haploid, with the parasite remaining in a haploid state during most of its lifecycle. In the asexual cycle, tachyzoites invade the host cell, replicate, exit by host cell lysis and invade other cells. Tachyzoites can also differentiate into a slow-growing form resulting in tissue cysts called bradyzoites. When cats ingest the tissue cysts, tachyzoites or oocysts, the sexual cycle takes place resulting in the formation of the diploid oocysts. The oocysts are shed in cat feces. In the outside environment, the oocysts undergo meiosis to form haploid sporozoites that infect cells. Humans and other animals contract the infection through the oocysts or consumption of uncooked meat (via tissue cysts) (122).

Toxoplasma is considered a model organism for the study of apicomplexan biology. It can be continuously maintained in culture in the lab and the mouse animal model is also well-

established. A number of genetic manipulation techniques have been developed including efficient transfection and selection protocols, reverse and forward genetic approaches, reporter constructs and clonal parasite lines. The genome sequence and extensive expression data, particularly for three strains, ME49, GT1 and VEG, are available and well-annotated (123). Recently, 62 strains of *T. gondii* were sequenced, providing an excellent resource for population biology studies (117).

Toxoplasma can be propagated asexually through direct oral transmission, circumventing the sexual stage (112). This has resulted in a highly-clonal population in North America and Europe (124) and is a possible cause of its widespread occurrence. The three clonal lineages (types I, II and III) of North America and Europe differ by only a small percent at the DNA level for any given locus (124). The lineages arose from a few, recent genetic crosses and following a bottleneck have expanded their ranges in the last 10,000 years (112). In spite of the clonal population structure, these genotypes show remarkable differences in their virulence (125). Just one oocyst of the type I RH strain is capable of killing a mouse, where as the mouse can survive a dose of a million type III parasites (126, 127). Genetic mapping has identified the rhoptory kinases ROP18/ROP5 loci to be among the key determinants of acute virulence (116, 128).

Sequencing of strains from South America identified additional diversity in the *T. gondii* population structure (129, 130). While clonality did exist in the South American strains as well, it was different from the North American and European strains. Multi-locus mapping and RFLP analysis grouped them into 12 clades that correlated roughly with geographic segregation (131-133). Recently, we were part of a consortium that sequenced and analyzed 62 *T. gondii* strains from around the world (117). While a neighbor network using SNPs from these strains showed traces of the previously described clade structure, it did reveal higher than expected

levels of genetic recombination between lineages, thus showing the extent of gene flow (Figure 1.8A). Based on this analysis, the strains were grouped in to six clades and I have maintained this nomenclature in our analysis of the *T. gondii* strains in Chapter 3. However, there is clear evidence of local admixture. Strains from different clades share regions or haploblocks, suggesting shared ancestry, to such an extent that the clades are no longer as clearly defined (Figure 1.8B). These findings suggest that although sexual recombination is infrequent in the wild, it does happen and maintenance of these shared haploblocks further suggests that they may confer an advantage and drive local adaptations (117).

Our preliminary studies on the strains GT1, ME49 and VEG representing types I, II and III respectively have shown strain-specific differences in organellar DNA insertions (NUM/PTs). In Chapter 2, I have discussed our initial findings and the sheer number of NUM/PTs in the *T. gondii* genome. Chapter 3 discusses the contribution of these organellar insertions, particularly NUMTs, in the evolution of the *T. gondii* genome and the role they can play in strain diversification. It is interesting to consider the impact of NUMTs in *T. gondii* genome evolution in light of the recent findings on sexual recombination in a parasite that was previously thought to be extremely clonal.

Sarcocystis neurona

The genus *Sarcocystis* comprises a large number of species collectively capable of infecting a wide range of animals, including fish, birds, reptiles, humans and other mammals (134). Infections caused by *Sarcocystis* spp. are generally moderate and asymptomatic. However, *Sarcocystis neurona*, the species of interest here, causes a debilitating neurologic disease, equine protozoal myeloencephalitis, in horses (135). Symptoms include muscle atrophy, uncoordinated gait, facial nerve paralysis, head tilt and ataxia. This disease causes a serious economic burden

on the equine industry and without treatment the disease is fatal. Even with treatment the horses may not completely recover.

The definitive hosts are opossums (136) and *S. neurona* has a wide range of intermediate hosts including skunks, raccoons and armadillos (135, 137) (Figure 1.7B). Opossums become infected when they ingest the sarcocysts in the muscles of intermediate hosts. The parasites excyst from the sarcocysts and undergo sexual reproduction in the intestinal epithelium to form oocysts. The oocysts contain two sporocysts and each sporocyst contains four sporozoites. The oocysts usually rupture and the sporocysts are excreted in the feces of the opossums. When the sporocysts are ingested by an intermediate host through contaminated food or water, the asexual cycle takes place. Sporozoites from the sporocysts are released in the intestines and form schizonts or merozoites. The schizonts are the intracellular stage where the parasite replicates by endopolygeny, a mode of division that involves multiple rounds of DNA replication before cytokinesis (138). The daughter cells lyse the host cell and emerge as extracellular merozoites. The process continues and eventually merozoites form a sarcocyst and reside in the muscle tissues of intermediate hosts. When opossums scavenge the intermediate hosts, the cycle continues. In horses, the central nervous system is infected instead of muscles and sarcocysts are not typically formed and the parasite cannot be transmitted. Hence horses are considered aberrant or dead-end intermediate hosts (135).

In comparison to the other coccidians, *S. neurona* has some interesting differences in its biology. Rhoptries, an apical organelle important for invasion, are not formed in the merozoites. The protective parasitophorous vacuole, inside which the parasite resides during its intracellular stages, is also not formed. Further reasons that make *S. neurona* an important organism for evolutionary studies are discussed in Chapter 4.

THE MITOCHONDRIAL GENOME OF *TOXOPLASMA GONDII*

The mitochondrial genome of *T. gondii* has been an enigma. Attempts to isolate the mitochondrion organelle and identify the complete mitochondrial genome have been unsuccessful. The mitochondrial organelle occurs as an elongated S or lasso in the interphase and forms other polymorphic topologies during the cell-cycle, making physical isolation of the mitochondrion from other cellular fractions nearly impossible (139). Approaches to elucidate the mtDNA sequence using sequenced-based probes were unsuccessful due to the large number of mtDNA fragments (NUMTs) in the nuclear genome. The NUMTs interfere with signals from molecular-based mtDNA isolation methods. The presence of NUMTs in the *T. gondii* nuclear genome (initially referred to as 'REP' elements in this parasite) was first reported by Ossorio *et al.,* (140). The region of sequence downstream of the single copy nuclear gene hybridized to multiple regions on the nuclear genome resulting in a smeared signal in Southern hybridization assays. This region contained fragments of the *coxI* and *cob* genes typically found in the apicomplexan mitochondrial genomes. This finding and the inability to isolate the mitochondrion raised the question of whether a mitochondrion with it's own genome is still functional in *T. gondii*. Multiple lines of evidence suggest that it is. First, rhodamine 123 and MitoTracker dyes which are markers of a transmembrane potential indicate the mitochondrion is active (141). Second, a complete set of tRNAs are imported into the *T. gondii* mitochondrion (apicomplexan mt genomes do not code for tRNAs) (101). Third, *cob,* one of the three mitochondrial genes, has been characterized and mutations in the cob were found to be associated with atovaquone resistance (142). Fourth, NUMT sequences, even if combined together cannot encode for any of the cytochrome genes functionally and in their entirety, thus, the genes must be encoded in the mitochondrion (discussed below).

An ~8Kb assembled contig generated in the *T. gondii* ME49 genome project was found to contain sequences of mtDNA genes found in other apicomplexans (personal communication, David Roos). The field hoped this was the complete mitochondrial genome of *T. gondii*. However, attempts to reproduce and verify this sequence by several research groups, including the Roos group using a variety of methods were not successful. It appears to be an assembly artifact. However, the sequence information contained in this contig as well as individual Sanger and EST reads that matched portions of the 8 Kb contig and were typical of genes and rRNA of mitochondrial origin served as a template for PCR primer design. The idea was to use a PCR-based approach based on putative extant mtDNA sequences emerging from the genome project (as opposed to NUMTs which appear to be degenerate relative to the putative mtDNA) to elucidate the mtDNA sequence itself. We performed PCR under very stringent conditions using several combinations of primer pairs with DNA and RNA from mitochondrial-enriched cellular fractions as template. The results were puzzling. No amplicon was larger than ~3Kb. Primer pairs often generated multiple amplicons, that showed a high-level of sequence identity to each other in some regions but differed in others. The variety of results led the Kissinger laboratory to attempt PCR reactions with only a single primer. They too, generated amplicons (Figure 1.9). We did identify primers that avoided NUMTs and would amplify one (100-150 bp) region for each of the three cytochrome genes (Figure 1.9). MSA analysis of the amplicon sequences revealed 23 distinct sequence elements, which make up all of the observed sequence in each permuted amplicon. The elements are named from A to W (Figure 6.1, Appendix 1). The elements range in size from 40 bp to 1049 bp and none of them encode a gene in its entirety. Each element has a discrete sequence boundary that is observed regardless of the sequence element permutation pattern it is observed in. Some sequence elements are observed in

combination with certain elements more often than with others. However, each element, when it appears in a PCR product, always has the same length, sequence and boundaries irrespective of the elements that flank it. Each of the precise elements as well as arrangements of elements observed in the PCR amplicons are also represented in the unassembled Sanger reads and non-chromosomal contigs from the *T. gondii* ME49 genome project as well as in EST sequence reads.

The full-length coding sequences of the *coxI* and *cob* genes, identified via reverse transcriptase PCR, are publicly available (142, 143). However, we were never able to PCR any full-length cytochrome genes using genomic DNA as template. 13 out of the 23 sequence elements, when artificially assembled together, can encode each of the cytochrome genes (Figure 1.10, Table 1.3). In order to create the above-mentioned cytochrome gene assembly an element is either used in its entirety or in some cases only the beginning or the end of the of the element is used but this situation only occurs when the element is located at the beginning or end of the coding sequence. An element is never required to be broken such that the middle of the element is used. For example, in the assembly of *coxIII*, the final 381 bp of the gene is encoded by the first 381 bp of element M. However, the orientation of the element can change such that different strands are used to contribute to the coding sequence. For example, the entire 161 bp of element V is used in the assembly of the *coxI* gene from coding positions 11-171 and first 71 bp of element V codes for the first 71 bp of the *coxIII* gene but in the reverse orientation (Figure 1.10). Examination of unassembled Sanger and EST reads provided evidence that the elements do naturally occur in an order such that distinct reads can be assembled in to full-length coding sequences of the genes (Figure 1.10, grey lines). However, it must be noted that the reads also contain the same elements in other arrangements that would not facilitate this assembly. This

finding likely means the elements occur in a population of various arrangements in the mitochondrion. We were able to identify LSU and SSU rRNA fragments in these elements using sequence homology to the *P. falciparum* LSU and SSU rRNA fragments (Figure 6.1). The different amplicons and the various permutations of the elements seen in the PCR experiments and unassembled reads make it difficult to assemble a full-length, linear mtDNA sequence. It is not clear if such a molecule even physically exists in the mitochondrion, but it must be possible to create a full-length RNA since it is detected and because the mitochondrion is a functional organelle.

Attempts to identify the size and physical structure of the mtDNA molecule(s) have been far less successful. Southern hybridization of CHEF gels performed with DNA extracted from mitochondrial-enriched fractions and probed with a 1,021 bp *cob* probe produced a smear in the lane that ranged in size from 5 Kb - 23 Kb, similar to smears observed in *P. falciparum* but, the probe also hybridized to DNA remaining in the well. Physically, if the mtDNA genome exists as a linear concatemer of tandemly repeating units, restriction digestion with an enzyme that has a single cut site should result in one major band and a tailing-off smear of smaller fragments, as seen in *Plasmodium* (104) (Figure 1.11A). However, when *T. gondii* RH DNA was digested with *XhoI*, which uniquely cuts in the *cob* gene and probed with a *cox1* probe, the probe hybridized to multiple bands and a pattern as seen in the *P. falciparum* experiments was not seen (Figure 1.11B). The experiments were carried out under stringent conditions and the restriction site and probe used were located in a region within an element, thus rearrangement should not influence the resulting hybridization pattern. These experiments did not provide any clear indications of the topology of the *T. gondii* mitochondrial genome. Electron microscopy was attempted on DNA isolated from the mitochondrial-enriched fractions but the experiments were unsuccessful.

Thus, we currently do not know the physical structure of the mitochondrial genome or a definitive order for 23 sequence elements. It is possible the mt genome exists as a collection of permutation or recombination variants of these 23 elements.

Comparison of these permutation patterns observed in *T. gondii* to the non-chromosomal contigs of its close relative *N. caninum* revealed the same 23 sequence elements and highly-similar permutation patterns. We were able to identify all 23 elements with the same precise sequence boundaries, except for elements B, L, M, N, T and W (Table 1.4). Between 1- 5 bp at the beginning or end of these six elements could not be identified. Finding this evolutionary conservation provides further support for the unique nature of the *T. gondii* mt genome.

The NUMT fragments inserted in the *T. gondii* nuclear genome that are described in Chapter 2 are not to be confused with the 23 sequence elements of the mtDNA. NUMT fragments may or may not be insertions of an entire element or of one or more elements. A NUMT can be derived from any region of an element and the boundaries of a NUMT rarely match the boundaries of the element, NUMTs can begin and end at any sequence point. Perhaps the best visualization is one of randomly sheared mtDNA.

NUMTs can be distinguished from the mtDNA/elements in a variety of ways. mtDNA elements have distinct sizes, boundaries and a non-varying sequence. NUMTS vary greatly in size, they demonstrate varying levels of sequence decay with respect to the 23 sequence elements in unique ways and NUMTs do not have conserved sequence boundaries as is seen with the 23 sequence elements. NUMTs are flanked by non-mitochondrial (nuclear) sequences. It is impossible to assemble full-length coding sequences with an open reading frame for any of the cytochrome genes using only NUMTs, further providing proof that the cytochrome genes must be encoded by a mitochondrial genome. Whether the presence of abundant NUMTs and the

existence of the mtDNA as permutation variants of the 23 elements is an indication of *T. gondii* being in the process of losing its mt genome is up for debate. It is not unprecedented within the Apicomplexa. *Cryptosporidium* has lost its mtDNA and is an obligate anaerobe.

ORGANIZATION OF THIS DISSERTATION

In chapter 2, I present my findings on the unprecedented levels of integrated organellar DNA in the nuclear genome of *T. gondii* and *N. caninum.* Chapter 3 extends on these findings and assesses the role of mitochondrial DNA integrants as an important evolutionary force in *T. gondii*. The impact of this evolutionary force in *H. hammondi* and *N. caninum* is also discussed. I used 13 of the 62 newly-sequenced *T. gondii* strains for comparative analysis. I also performed some of the comparative analyses for the 62 genome project but the work is not presented in this thesis (117) (Chapter 6-2). I was able to use the genome resources and the data generated as part of the 62 genome project to investigate the intriguing evolutionary story of NUMTs in the *T. gondii* genome. Figure 1.8, which was generated through collaborative efforts at The University of Georgia, is critical for inferring results discussed in Chapter 3.

In chapter 4, I present the genome of *S. neurona* (SN3 strain). We were working on this genome in parallel with John Parkinson and Michael Grigg's group who were working on the *S. neurona* SN1 strain. Through collaborative efforts, we published the *S. neurona* SN1 manuscripts first (80) (Chapter 6-1), which focuses on the gene content of secretory pathogenesis determinants and metabolic pathways. The SN3 manuscript (Chapter 4) focuses on evolutionary aspects including genome size and AP2 transcription factor content. In chapter 5, I discuss plans to perform comparative transcriptomics and population studies in *S. neurona.* I also provide an overall discussion of the impact of these different evolutionary forces in shaping the evolution of coccidian genomes.

REFERENCES

1. Levine ND (1988) *The Protozoan Phylum Apicomplexa Vol II* (CRC Press) p 154.
2. Wasmuth J, Daub J, Peregr Alvarez JM, Finney CAM, & Parkinson J (2009) The origins of apicomplexan sequence innovation. *Genome research* 19(7):1202-1213.
3. Saffo MB, McCoy AM, Rieken C, & Slamovits CH (2010) Nephromyces, a beneficial apicomplexan symbiont in marine animals. *Proceedings of the National Academy of Sciences of the United States of America* 107(37):16190-16195.
4. Kirk NL*, et al.* (2013) Tracking transmission of apicomplexan symbionts in diverse Caribbean corals. *PloS one* 8(11):e80618.
5. Rueckert S, Wakeman KC, Jenke-Kodama H, & Leander BS (2015) Molecular systematics of marine gregarine apicomplexans from Pacific tunicates, with descriptions of five new species of Lankesteria. *International journal of systematic and evolutionary microbiology*.
6. Berney Cd & Pawlowski J (2006) A molecular time-scale for eukaryote evolution recalibrated with the continuous microfossil record. *Proceedings of the Royal Society B: Biological Sciences* 273(1596):1867-1872.
7. Okamoto N & McFadden GI (2008) The mother of all parasites. *Future Microbiol* 3:391-395.
8. DeBarry J & Kissinger J (2011) Jumbled Genomes: Missing Apicomplexan Synteny. *Molecular Biology and Evoluiton* 28.
9. Kissinger JC & DeBarry J (2011) Genome cartography: charting the apicomplexan genome. *Trends in parasitology* 27(8):345-354.
10. Kuo CH & Kissinger JC (2008) Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. *BMC evolutionary biology* 8:108.
11. Zhu G & Keithly JS (2002) Alpha-proteobacterial relationship of apicomplexan lactate and malate dehydrogenases. *The Journal of eukaryotic microbiology* 49(3):255-261.
12. Huang J, Mullapudi N, Sicheritz-Ponten T, & Kissinger JC (2004) A first glimpse into the pattern and scale of gene transfer in Apicomplexa. *Int J Parasitol* 34(3):265-274.
13. Huang J*, et al.* (2004) Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in *Cryptosporidium parvum*. *Genome biology* 5(11):R88.
14. Striepen B*, et al.* (2004) Gene transfer in the evolution of parasite nucleotide biosynthesis. *Proceedings of the National Academy of Sciences of the United States of America* 101(9):3154-3159.
15. Huang J & Kissinger J (2006) Horizontal and intracellular gene transfer in the Apicomplexa: The scope and functional consequences. *Genome Evolution in Eukaryotic Microbes*, eds Katz L & Bhattacharya D (Oxford University Press), p 256.
16. Nagamune K & Sibley LD (2006) Comparative genomic and phylogenetic analyses of calcium ATPases and calcium-regulated proteins in the apicomplexa. *Molecular biology and evolution* 23(8):1613-1627.
17. Aikawa M (1988) Morphological changes in erythrocytes induced by malarial parasites. *Biology of the cell / under the auspices of the European Cell Biology Organization* 64(2):173-181.
18. Dubey JP (1977) Persistence of *Toxoplasma gondii* in the tissues of chronically infected cats. *The Journal of parasitology* 63(1):156-157.

19. Morrissette NS & Sibley LD (2002) Cytoskeleton of apicomplexan parasites. *Microbiology and molecular biology reviews : MMBR* 66(1):21-38; table of contents.
20. Aikawa M, Komata Y, Asai T, & Midorikawa O (1977) Transmission and scanning electron microscopy of host cell entry by *Toxoplasma gondii*. *The American journal of pathology* 87(2):285-296.
21. Carruthers VB & Sibley LD (1999) Mobilization of intracellular calcium stimulates microneme discharge in *Toxoplasma gondii*. *Mol Microbiol* 31(2):421-428.
22. Nichols BA & Chiappino ML (1987) Cytoskeleton of *Toxoplasma gondii*. *The Journal of protozoology* 34(2):217-226.
23. Aikawa M (1967) Ultrastructure of the pellicular complex of *Plasmodium fallax*. *The Journal of cell biology* 35(1):103-113.
24. Vivier E & Petitprez A (1969) Observations ultrastructurales sur l'hematozoaire anthemosa garnhami et examen de criteres morphologiques utilisables pour la taxonomie chez les sporozoaires. *Protistologica* 3:363-379.
25. Francia ME & Striepen B (2014) Cell division in apicomplexan parasites. *Nature reviews. Microbiology* 12(2):125-136.
26. Kohler S*, et al.* (1997) A plastid of probable green algal origin in Apicomplexan parasites. *Science* 275(5305):1485-1489.
27. McFadden G, Waller RF, Reith ME, & Lang-Unnasch N (1997) Plastids in apicomplexan parasites. *Plant Systematics and Evolution* 11:Suppl. 261-287.
28. Wilson RJ & Williamson DH (1997) Extrachromosomal DNA in the Apicomplexa. *Microbiology and molecular biology reviews : MMBR* 61(1):1-16.
29. Cowper B, Matthews S, & Tomley F (2012) The molecular basis for the distinct host and tissue tropisms of coccidian parasites. *Molecular and biochemical parasitology* 186(1):1-10.
30. Reid AJ*, et al.* (2012) Comparative genomics of the apicomplexan parasites *Toxoplasma gondii* and *Neospora caninum*: Coccidia differing in host range and transmission strategy. *PLoS pathogens* 8(3):e1002567.
31. Templeton TJ*, et al.* (2010) A Genome-Sequence Survey for Ascogregarina taiwanensis Supports Evolutionary Affiliation but Metabolic Diversity between a Gregarine and Cryptosporidium. *Molecular biology and evolution* 27(2):235-248.
32. Richard GF, Kerrest A, & Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and molecular biology reviews : MMBR* 72(4):686-727.
33. Ellis J (1982) Promiscuous DNA--chloroplast genes inside plant mitochondria. *Nature* 299(5885):678-679.
34. Stern DB & Lonsdale DM (1982) Mitochondrial and chloroplast genomes of maize have a 12-kilobase DNA sequence in common. *Nature* 299(5885):698-702.
35. Esser C*, et al.* (2004) A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Molecular biology and evolution* 21(9):1643-1660.
36. Keeling PJ & Palmer JD (2008) Horizontal gene transfer in eukaryotic evolution. *Nature reviews. Genetics* 9(8):605-618.
37. Strese A, Backlund A, & Alsmark C (2014) A recently transferred cluster of bacterial genes in *Trichomonas vaginalis*--lateral gene transfer and the fate of acquired genes. *BMC evolutionary biology* 14:119.

38. Hirt RP, Alsmark C, & Embley TM (2015) Lateral gene transfers and the origins of the eukaryote proteome: a view from microbial parasites. *Current opinion in microbiology* 23:155-162.
39. Nyvltova E*, et al.* (2015) Lateral gene transfer and gene duplication played a key role in the evolution of *Mastigamoeba balamuthi* hydrogenosomes. *Molecular biology and evolution* 32(4):1039-1055.
40. Pace JK, 2nd, Gilbert C, Clark MS, & Feschotte C (2008) Repeated horizontal transfer of a DNA transposon in mammals and other tetrapods. *Proceedings of the National Academy of Sciences of the United States of America* 105(44):17023-17028.
41. Gilbert C, Schaack S, Pace JK, 2nd, Brindley PJ, & Feschotte C (2010) A role for host-parasite interactions in the horizontal transfer of transposons across phyla. *Nature* 464(7293):1347-1350.
42. McClintock B (1942) The fusion of broken ends of chromosomes following nuclear fusion. *Proc Assoc Am Physicians*:458-463.
43. Bennetzen JL & Wang H (2014) The contributions of transposable elements to the structure, function, and evolution of plant genomes. *Annual review of plant biology* 65:505-530.
44. Gray MW, Burger G, & Lang BF (1999) Mitochondrial evolution. *Science* 283(5407):1476-1481.
45. Burger G, Gray MW, & Lang BF (2003) Mitochondrial genomes: anything goes. *Trends in genetics : TIG* 19(12):709-716.
46. Lang BF, Gray MW, & Burger G (1999) Mitochondrial genome evolution and the origin of eukaryotes. *Annual review of genetics* 33:351-397.
47. Burger G & Lang BF (2003) Parallels in genome evolution in mitochondria and bacterial symbionts. *IUBMB Life* 55(4-5):205-212.
48. Morris JC*, et al.* (2001) Replication of kinetoplast DNA: an update for the new millennium. *Int J Parasitol* 31(5-6):453-458.
49. Maleszka R, Skelly PJ, & Clark-Walker GD (1991) Rolling circle replication of DNA in yeast mitochondria. *EMBO J* 10(12):3923-3929.
50. Feagin JE, Mericle BL, Werner E, & Morris M (1997) Identification of additional rRNA fragments encoded by the *Plasmodium falciparum* 6 kb element. *Nucleic Acids Res* 25(2):438-446.
51. Adams KL & Palmer JD (2003) Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Mol Phylogenet Evol* 29(3):380-395.
52. Yamauchi A (2005) Rate of gene transfer from mitochondria to nucleus: effects of cytoplasmic inheritance system and intensity of intracellular competition. *Genetics* 171(3):1387-1396.
53. Kleine T, Maier UG, & Leister D (2009) DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annual review of plant biology* 60:115-138.
54. Palmer JD (2003) The symbiotic birth and spread of plastids: how many times and whodunit? *Journal of Phycology* 39:4-12.
55. Howe CJ, Beanland TJ, Larkum AW, & Lockhart PJ (1992) Plastid origins. *Trends Ecol Evol* 7(11):378-383.
56. Howe CJ*, et al.* (2003) Evolution of the chloroplast genome. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 358(1429):99-106; discussion 106-107.

57. Sugiura M (1995) The chloroplast genome. *Essays in biochemistry* 30:49-57.
58. Sugiura M (1992) The chloroplast genome. *Plant Mol Biol* 19(1):149-168.
59. Glockner G, Rosenthal A, & Valentin K (2000) The structure and gene repertoire of an ancient red algal plastid genome. *Journal of molecular evolution* 51(4):382-390.
60. Keeling PJ (2008) Evolutionary biology: bridge over troublesome plastids. *Nature* 451(7181):896-897.
61. Martin W*, et al.* (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences of the United States of America* 99(19):12246-12251.
62. Leister D (2005) Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends in genetics : TIG* 21(12):655-663.
63. Hazkani-Covo E (2009) Mitochondrial insertions into primate nuclear genomes suggest the use of numts as a tool for phylogeny. *Molecular biology and evolution* 26(10):2175-2179.
64. Behura SK (2007) Analysis of nuclear copies of mitochondrial sequences in honeybee (Apis mellifera) genome. *Molecular biology and evolution* 24(7):1492-1505.
65. Song H, Buhay JE, Whiting MF, & Crandall KA (2008) Many species in one: DNA barcoding overestimates the number of species when nuclear mitochondrial pseudogenes are coamplified. *Proceedings of the National Academy of Sciences of the United States of America* 105(36):13486-13491.
66. Noutsos C, Kleine T, Armbruster U, DalCorso G, & Leister D (2007) Nuclear insertions of organellar DNA can create novel patches of functional exon sequences. *Trends in genetics : TIG* 23(12):597-601.
67. Ricchetti M, Fairhead C, & Dujon B (1999) Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* 402(6757):96-100.
68. Lin Y & Waldman AS (2001) Capture of DNA sequences at double-strand breaks in mammalian chromosomes. *Genetics* 158(4):1665-1674.
69. Hazkani-Covo E & Covo S (2008) Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS genetics* 4(10):e1000237.
70. Hazkani-Covo E, Zeller RM, & Martin W (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS genetics* 6(2):e1000834.
71. Stupar RM*, et al.* (2001) Complex mtDNA constitutes an approximate 620-kb insertion on Arabidopsis thaliana chromosome 2: implication of potential sequencing errors caused by large-unit repeats. *Proceedings of the National Academy of Sciences of the United States of America* 98(9):5099-5103.
72. Huang CY, Grunheit N, Ahmadinejad N, Timmis JN, & Martin W (2005) Mutational decay and age of chloroplast and mitochondrial genomes transferred recently to angiosperm nuclear chromosomes. *Plant physiology* 138(3):1723-1733.
73. Richly E & Leister D (2004) NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. *Molecular biology and evolution* 21(10):1972-1980.
74. Noutsos C, Richly E, & Leister D (2005) Generation and evolutionary fate of insertions of organelle DNA in the nuclear genomes of flowering plants. *Genome research* 15(5):616-628.
75. Baldauf SL (2003) The deep roots of eukaryotes. *Science* 300(5626):1703-1706.
76. Keeling PJ*, et al.* (2005) The tree of eukaryotes. *Trends Ecol Evol* 20(12):670-676.

77. Templeton TJ, *et al.* (2004) Comparative Analysis of Apicomplexa and Genomic Diversity in Eukaryotes. *Genome research* 14(9):1686-1695.
78. Keeling PJ & Fast NM (2002) Microsporidia: biology and evolution of highly reduced intracellular parasites. *Annual review of microbiology* 56:93-116.
79. Reid AJ, *et al.* (2014) Genomic analysis of the causative agents of coccidiosis in domestic chickens. *Genome research* 24(10):1676-1685.
80. Blazejewski T, *et al.* (2015) Systems-based analysis of the *Sarcocystis neurona* genome identifies pathways that contribute to a heteroxenous life cycle. *mBio* 6(1).
81. Gardner MJ, *et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419(6906):498-511.
82. Ling KH, *et al.* (2007) Sequencing and analysis of chromosome 1 of *Eimeria tenella* reveals a unique segmental organization. *Genome research* 17(3):311-319.
83. El-Sayed NM, *et al.* (2005) The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease. *Science* 309(5733):409-415.
84. El-Sayed NM, *et al.* (2005) Comparative genomics of trypanosomatid parasitic protozoa. *Science* 309(5733):404-409.
85. Roos DS, *et al.* (1999) Origin, targeting, and function of the apicomplexan plastid. *Current opinion in microbiology* 2(4):426-432.
86. Moore RB, *et al.* (2008) A photosynthetic alveolate closely related to apicomplexan parasites-original article. *Nature* 451(7181):959-963.
87. Janouskovec J, Horak A, Obornik M, Lukes J, & Keeling PJ (2010) A common red algal origin of the apicomplexan, dinoflagellate, and heterokont plastids. *Proceedings of the National Academy of Sciences of the United States of America* 107(24):10949-10954.
88. Wilson RJ, *et al.* (1996) Complete gene map of the plastid-like DNA of the malaria parasite *Plasmodium falciparum*. *J Mol Biol* 261(2):155-172.
89. Cai X, Fuller AL, McDougald LR, & Zhu G (2003) Apicoplast genome of the coccidian *Eimeria tenella*. *Gene* 321:39-46.
90. Brayton KA, *et al.* (2007) Genome sequence of *Babesia bovis* and comparative analysis of apicomplexan hemoprotozoa. *PLoS pathogens* 3(10):1401-1413.
91. Waller RF, *et al.* (1998) Nuclear-encoded proteins target to the plastid in *Toxoplasma gondii* and *Plasmodium falciparum*. *Proceedings of the National Academy of Sciences of the United States of America* 95(21):12352-12357.
92. Waller RF, Reed MB, Cowman AF, & McFadden GI (2000) Protein trafficking to the plastid of *Plasmodium falciparum* is via the secretory pathway. *Embo J* 19(8):1794-1802.
93. Sato S (2011) The apicomplexan plastid and its evolution. *Cellular and molecular life sciences : CMLS* 68(8):1285-1296.
94. Fichera ME & Roos DS (1997) A plastid organelle as a drug target in apicomplexan parasites. *Nature* 390(6658):407-409.
95. McConkey GA, Rogers MJ, & McCutchan TF (1997) Inhibition of *Plasmodium falciparum* protein synthesis. Targeting the plastid-like organelle with thiostrepton. *J Biol Chem* 272(4):2046-2049.
96. Dahl EL, *et al.* (2006) Tetracyclines specifically target the apicoplast of the malaria parasite *Plasmodium falciparum*. *Antimicrob Agents Chemother* 50(9):3124-3131.
97. Dahl EL & Rosenthal PJ (2007) Multiple antibiotics exert delayed effects against the *Plasmodium falciparum* apicoplast. *Antimicrob Agents Chemother* 51(10):3485-3490.

98. Goodman CD, Su V, & McFadden GI (2007) The effects of anti-bacterials on the malaria parasite *Plasmodium falciparum*. *Molecular and biochemical parasitology* 152(2):181-191.
99. Ramya TN, Mishra S, Karmodiya K, Surolia N, & Surolia A (2007) Inhibitors of nonhousekeeping functions of the apicoplast defy delayed death in *Plasmodium falciparum*. *Antimicrob Agents Chemother* 51(1):307-316.
100. Fichera ME, Bhopale MK, & Roos DS (1995) In vitro assays elucidate peculiar kinetics of clindamycin action against *Toxoplasma gondii*. *Antimicrob Agents Chemother* 39(7):1530-1537.
101. Esseiva AC, Naguleswaran A, Hemphill A, & Schneider A (2004) Mitochondrial tRNA import in *Toxoplasma gondii*. *J Biol Chem* 279(41):42363-42368.
102. Feagin JE*, et al.* (2012) The fragmented mitochondrial ribosomal RNAs of *Plasmodium falciparum*. *PloS one* 7(6):e38320.
103. Waller RF & Jackson CJ (2009) Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. *BioEssays : news and reviews in molecular, cellular and developmental biology* 31(2):237-245.
104. Preiser PR*, et al.* (1996) Recombination associated with replication of malarial mitochondrial DNA. *Embo J* 15(3):684-693.
105. Kairo A, Fairlamb AH, Gobright E, & Nene V (1994) A 7.1 kb linear DNA molecule of *Theileria parva* has scrambled rDNA sequences and open reading frames for mitochondrially encoded proteins. *EMBO J* 13(4):898-905.
106. Hikosaka K*, et al.* (2011) Concatenated mitochondrial DNA of the coccidian parasite *Eimeria tenella*. *Mitochondrion* 11(2):273-278.
107. Wong SY & Remington JS (1993) Biology of *Toxoplasma gondii*. *Aids* 7(3):299-316.
108. Dubey JP (1998) Advances in the life cycle of *Toxoplasma gondii*. *International journal for parasitology* 28(7):1019-1024.
109. Chapman HD*, et al.* (2013) A selective review of advances in coccidiosis research. *Advances in parasitology* 83:93-171.
110. Wolf A, Cowen D, & Paige B (1939) Human Toxoplasmosis: Occurrence in Infants as an Encephalomyelitis Verification by Transmission to Animals. *Science* 89(2306):226-227.
111. Dubey JP (2008) The history of *Toxoplasma gondii*--the first 100 years. *The Journal of eukaryotic microbiology* 55(6):467-475.
112. Su C*, et al.* (2003) Recent expansion of *Toxoplasma* through enhanced oral transmission. *Science* 299(5605):414-416.
113. Saeij JP*, et al.* (2006) Polymorphic secreted kinases are key virulence factors in toxoplasmosis. *Science* 314(5806):1780-1783.
114. Kim SK & Boothroyd JC (2005) Stage-specific expression of surface antigens by *Toxoplasma gondii* as a mechanism to facilitate parasite persistence. *Journal of immunology* 174(12):8038-8048.
115. Saeij JP*, et al.* (2007) *Toxoplasma* co-opts host gene expression by injection of a polymorphic kinase homologue. *Nature* 445(7125):324-327.
116. Fentress SJ*, et al.* (2010) Phosphorylation of immunity-related GTPases by a *Toxoplasma gondii*-secreted kinase promotes macrophage survival and virulence. *Cell host & microbe* 8(6):484-495.
117. Lorenzi H. KA, Benke M.S. ,Namasivayam S.,Seshadri L.S.,Hadjithomas M.,Karamycheva S.,Pinney D.,Brunk B.,Ajioka J.W.,Ajzenberg D.,Boothroyd J.C.,Boyle

J.P.,Dardé M.L.,Dubey J.P.,Fritz H.M.,Gennari S.M., Gregory B.D.,Kim K.,Rosenthal B. M.,Saeij J., Su C.,White M.W.,Zhu X.Q,Howe D.K.,Grigg M.E.,Parkinson J.,Liu L.,Kissinger J.C.,Roos D.S.,Sibley L. D. (2015) Comparative sequence analysis of *Toxoplasma gondii* reveals local genomic admixture drives concerted expansion and diversification of secreted pathogenesis determinants. *In Review*.

118. Talevich E & Kannan N (2013) Structural and evolutionary adaptation of rhoptry kinases and pseudokinases, a family of coccidian virulence factors. *BMC evolutionary biology* 13:117.
119. Keeley A & Soldati D (2004) The glideosome: a molecular machine powering motility and host-cell invasion by Apicomplexa. *Trends in cell biology* 14(10):528-532.
120. Dunn JD, Ravindran S, Kim SK, & Boothroyd JC (2008) The *Toxoplasma gondii* dense granule protein GRA7 is phosphorylated upon invasion and forms an unexpected association with the rhoptry proteins ROP2 and ROP4. *Infection and immunity* 76(12):5853-5861.
121. Kim K & Weiss LM (2004) *Toxoplasma gondii*: the model apicomplexan. *International journal for parasitology* 34(3):423-432.
122. David S. Roos RGD, Naomi S. Morrissette, A. Lindsay C. Moulton (1995) *Molecular Tools for Genetic Dissection of the Protozoan ParasiteToxoplasma gondii* 1995 Ed.
123. Gajria B*, et al.* (2008) ToxoDB: an integrated *Toxoplasma gondii* database resource. *Nucleic Acids Res* 36(Database issue):D553-556.
124. Sibley LD & Ajioka JW (2008) Population structure of *Toxoplasma gondii*: clonal expansion driven by infrequent recombination and selective sweeps. *Annual review of microbiology* 62:329-351.
125. Howe DK, Honore S, Derouin F, & Sibley LD (1997) Determination of genotypes of *Toxoplasma gondii* strains isolated from patients with toxoplasmosis. *Journal of clinical microbiology* 35(6):1411-1414.
126. Sibley LD & Boothroyd JC (1992) Virulent strains of *Toxoplasma gondii* comprise a single clonal lineage. *Nature* 359(6390):82-85.
127. Howe DK & Sibley LD (1995) *Toxoplasma gondii* comprises three clonal lineages: correlation of parasite genotype with human disease. *The Journal of infectious diseases* 172(6):1561-1566.
128. Reese ML, Zeiner GM, Saeij JP, Boothroyd JC, & Boyle JP (2011) Polymorphic family of injected pseudokinases is paramount in *Toxoplasma* virulence. *Proceedings of the National Academy of Sciences of the United States of America* 108(23):9625-9630.
129. Khan A*, et al.* (2007) Recent transcontinental sweep of *Toxoplasma gondii* driven by a single monomorphic chromosome. *Proceedings of the National Academy of Sciences of the United States of America* 104(37):14872-14877.
130. Su C*, et al.* (2012) Globally diverse *Toxoplasma gondii* isolates comprise six major clades originating from a small number of distinct ancestral lineages. *Proceedings of the National Academy of Sciences of the United States of America* 109(15):5844-5849.
131. Khan A*, et al.* (2011) Genetic analyses of atypical *Toxoplasma gondii* strains reveal a fourth clonal lineage in North America. *International journal for parasitology* 41(6):645-655.
132. Ajioka JW & Morrissette NS (2009) A century of *Toxoplasma* research. *International journal for parasitology* 39(8):859-860.

133. Au KF, Underwood JG, Lee L, & Wong WH (2012) Improving PacBio long read accuracy by short read alignment. *PloS one* 7(10):e46679.
134. Tenter AM (1995) Current research on Sarcocystis species of domestic animals. *International journal for parasitology* 25(11):1311-1330.
135. Dubey JP*, et al.* (2001) A review of *Sarcocystis neurona* and equine protozoal myeloencephalitis (EPM). *Veterinary parasitology* 95(2-4):89-131.
136. Fenger CK*, et al.* (1995) Identification of opossums (Didelphis virginiana) as the putative definitive host of Sarcocystis neurona. *The Journal of parasitology* 81(6):916-919.
137. Cheadle MA*, et al.* (2001) The nine-banded armadillo (Dasypus novemcinctus) is an intermediate host for *Sarcocystis neurona*. *International journal for parasitology* 31(4):330-335.
138. Striepen B, Jordan CN, Reiff S, & van Dooren GG (2007) Building the perfect parasite: cell division in apicomplexa. *PLoS pathogens* 3(6):e78.
139. Nishi M, Hu K, Murray JM, & Roos DS (2008) Organellar dynamics during the cell cycle of *Toxoplasma gondii*. *Journal of cell science* 121(Pt 9):1559-1568.
140. Ossorio PN, Sibley LD, & Boothroyd JC (1991) Mitochondrial-like DNA sequences flanked by direct and inverted repeats in the nuclear genome of *Toxoplasma gondii*. *J Mol Biol* 222(3):525-536.
141. Weiss LM & Kim K (2007) *Toxoplasma gondii : the model apicomplexan : perspectives and methods* (Elsevier Academic Press, London ; Burlington, Mass.) pp xx, 777 p., [740] p. of plates.
142. McFadden DC, Tomavo S, Berry EA, & Boothroyd JC (2000) Characterization of cytochrome b from *Toxoplasma gondii* and Q(o) domain mutations as a mechanism of atovaquone-resistance. *Molecular and biochemical parasitology* 108(1):1-12.
143. Gjerde B (2013) Characterisation of full-length mitochondrial copies and partial nuclear copies (numts) of the cytochrome b and cytochrome c oxidase subunit I genes of *Toxoplasma gondii*, *Neospora caninum*, *Hammondia heydorni* and *Hammondia triffittae* (Apicomplexa: Sarcocystidae). *Parasitology research* 112(4):1493-1511.
144. Ajioka JW, Fitzpatrick JM, & Reitter CP (2001) *Toxoplasma gondii* genomics: shedding light on pathogenesis and chemotherapy. *Expert reviews in molecular medicine* 2001:1-19.
145. Jackson SP (2002) Sensing and repairing DNA double-strand breaks. *Carcinogenesis* 23(5):687-696.

Figures and table legends

Figure 1.1. Apicomplexan cladogram

Relationships among select apicomplexans for whom we have a genome sequence. The coccidian branch is highlighted in blue. Genome sizes obtained from EuPathDB.org release 24.

Figure 1.2. Comparison of genome sizes of select eukaryotes

Apicomplexan species are highlighted in red. Inset is a cladogram of select apicomplexan species. Reproduced from (9).

Figure 1.3. Ultrastructure of a *Toxoplasma gondii* tachyzoite

Morphology of a *Toxoplasma* tachyzoite parasite is shown. Tachyzoites are an invasive form of the asexual stage of *T. gondii*. The conoid is a defining feature of the apicomplexan cell and is thought to be associated with penetration of the host cell. Micronemes, rhoptries and dense granules are three major secretory organelles predominately found at the apical end. These organelles secrete proteins involved in motility, host cell attachement, invasion and formation of the parasitophorous vacuole. The four-membrane apicoplast and the mitochondrion are also indicated. Modified from (144)

Figure 1.4. Model of double-strand break repair via non-homologous end joining

Double-strand breaks (DSB) are induced by various exogenous and endogenous sources. NHEJ proteins can repair the DSB (proteins typically forming the NHEJ complex in vertebrates are shown here). One of the possible outcomes is the insertion of an mtDNA fragment. Modified from (62) and (145).

Figure 1.5. Schematic illustrating evolution of the apicomplexan cell

(a) Primary endosymbiosis of a cyanobacterium (green) by a eukaryotic cell to create a photosynthetic eukaryotic cell; **(b)** Transfer of genes from the chloroplast (green) and the

mitochondria (orange) to the nuclear genome (black) in an algal cell; **(c)** Secondary endosymbiosis of a red alga by another eukaryotic cell. Continued gene transfer from the algal nucleus and chloroplast to the new 'host' nuclear genome; **(d)** Loss of the algal nucleus to give rise to a modern apicomplexan cell; **(e)** Lateral gene transfer (LGT) from external sources (red) is occurring throughout evolution of the Apicomplexa. Gene transfer continues from organelles to the nuclear genome. Reproduced from (9)

Figure 1.6. Apicomplexan mitochondrial genomes

A. Schematic map of the *P. falciparum* mitochondrial genome. The ~6 Kb genome is shown with genes above and below representing the direction of transcription as indicated by the arrows. The white boxes indicate the protein coding genes. Blue boxes indicate SSU rRNA fragments, green boxes indicate LSU rRNA fragments and black represent the unassigned fragments. Abbreviations: *cox1* and *cox3*- cytochrome oxidase subunits I and III, *cob* –cytochrome b, UN-unidentified. Modified from (102) **B**. Mitochondrial genomes in apicomplexans. The apicomplexans show variation in size and topology of their mitochondrial genomes, even though the protein coding content and the fragmented nature of the rRNA is conserved.

Figure 1.7. Life cycles of *Toxoplasma gondii* and *Sarcocystis neurona*

A. *Toxoplasma gondii* **B.** *Sarcocystis neurona.* See text for explanation (Pages 16 and 18 respectively).

Figure 1.8. Population structure of *T. gondii*

A. Population genetic structure of *T. gondii*. Neighbor-net analysis based on genome-wide SNPs (802,764 common data points) from 62 isolates of *T. gondii*. Color wheel indicates major clades of *T. gondii*. Haplogroups are indicated by the circled number based on previous designations. Names written in pink denote the representative strains from each haplogroup. **B**. Chromosome

painting of 62 *Toxoplasma gondii* strains. Local admixture analyses were conducted on SNP blocks of size 1,000 on each of the 14 chromosomes. For each SNP block, local admixture was used to assign strains to a particular ancestral population. The shared inheritance of blocks across members reveals color patterns that extend vertically in the plot. For example, several pink regions show strong vertical patterns across multiple clades. The dominant color is not meant to imply origin. Reproduced from (117).

Figure 1.9. PCR analysis of *T. gondii* mtDNA

Genomic DNA from mitochondrial-enriched fragments was assayed with different primer pairs. The labels across the top indicate the primer pairs used. The primers are named with either the cytochrome gene or the element name. FWD and REV indicate the primer orientation with respect to the target cytochrome gene or the element sequence (Appendix 1). Primer sequences are provided in Table 1.2. In some cases, use of a single primer produced amplicons (lanes 8, 9). Often primer-pairs produced multiple amplicons (lanes 5-7). Cytochrome-specific primers were designed that avoid NUMTs but they amplify a small fragment (lanes 2-4). Size markers are as indicated.

Figure 1.10. *T. gondii* mitochondrial protein-coding genes can be assembled using mtDNA sequence elements

Each cytochrome coding region is represented as a colored box. The sequences of the cytochrome genes were determined using data available in GenBank and via TBLASTN to know apicomplexan cytochrome genes. Based on these sequences, the 23 mtDNA elements were artificially assembled as indicated by the thick black line below each colored box to prove that the coding capacity exists. The elements used to assemble each gene are marked on the thick black line. Numbers above the gene/colored box represent start/stop co-ordinates of the

corresponding element on the gene. In every case, the entire element was used. Sanger/EST reads are provided as evidence to suggest that the elements do occur in this order. The elements that comprise each read are indicated. Each Sanger/EST read is indicated as a grey line, with the parts that match the gene represented as a block grey line and the parts that do not match the gene as a dotted grey line. Only a few reads are indicated (many more exist); absence of reads spanning certain regions of a gene does not imply reads for that region are not available. Note, in some cases the reads are in the reverse orientation. GenBank IDs of the EST reads and a Sanger/genomic read that aligns with the corresponding EST read are provided in Table 1.3.

Figure 1.11. Structural characterization of the mitochodondrial genome via Southern analysis

Restriction digestion and Southern blot hybridization of *P. falciparum* (A) *and T. gondii* (B*)* mt genomes. **A.** All listed enzymes except *AvaII* and *AvaI,* cut the 6 Kb linear tandemly repeated mtDNA of the *P. falciparum* only once. *AvaI* failed to recognize one site as predicted from the sequence. Size markers are on the right. UR = unrestricted DNA; CCC = covalently closed circle, 6 Kb contour length. Reproduced from (104). **B.** Total DNA from *T. gondii* RH tachyzoites was digested with *XhoI* and hybridized to a *coxI* probe under very stringent conditions. The yellow arrows point to possible bands. Lanes: 1 = Undigested 2 µg genomic DNA; 2 = Digested 15 µg genomic DNA; 3 = Digested 5 µg genomic DNA; Digested 2 µg genomic DNA. Size markers are as indicated.

Table 1.1. Summary of sequenced apicomplexan genomes

Data obtained from EuPathDB.org release 24.

Table 1.2. Primers used in PCR analysis of the *T. gondii* mt genome

Primers correspond to the primers used in Figure 1.9. Specific primers or primer pairs are listed above each lane of the gel.

Table 1.3. Genomic and EST reads representing different arrangements of mtDNA elements

Data correspond to reads shown in Figure 1.10. Reads are indicated as grey lines in the figure. Each row represents a read in the order in which they appear in the figure. The mtDNA elements are annotated for each read and are indicated in order. Each element in present in its entirety in each read.

Table 1.4. Comparison of *T. gondii* and *N. caninum* mtDNA elements

BLASTN of the 23 *T. gondii* mtDNA elements against the *N. caninum* contig 8315 (downloaded from ToxoDB.org release 5) was used to identify the mtDNA in *N. caninum*. The start and stop co-ordinates and the percent identity of the identified *N. caninum* mtDNA to the corresponding *T. gondii* mtDNA element are indicated.

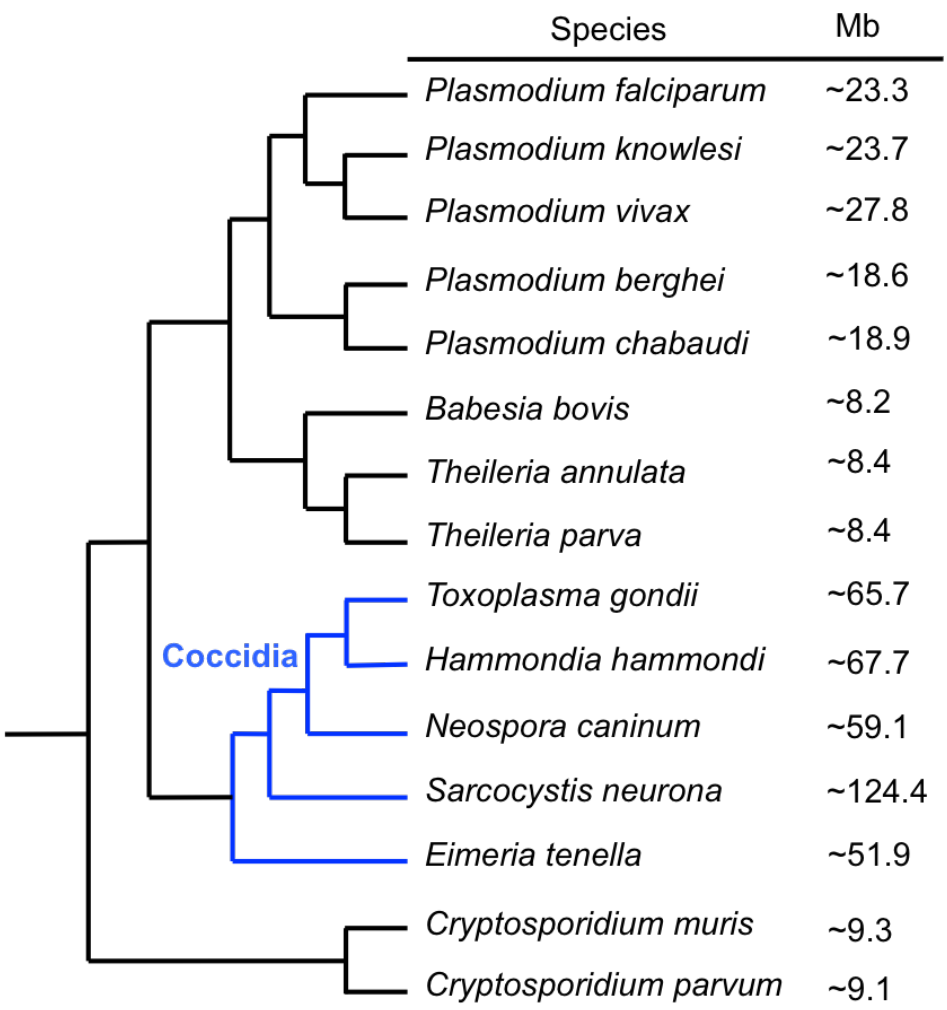
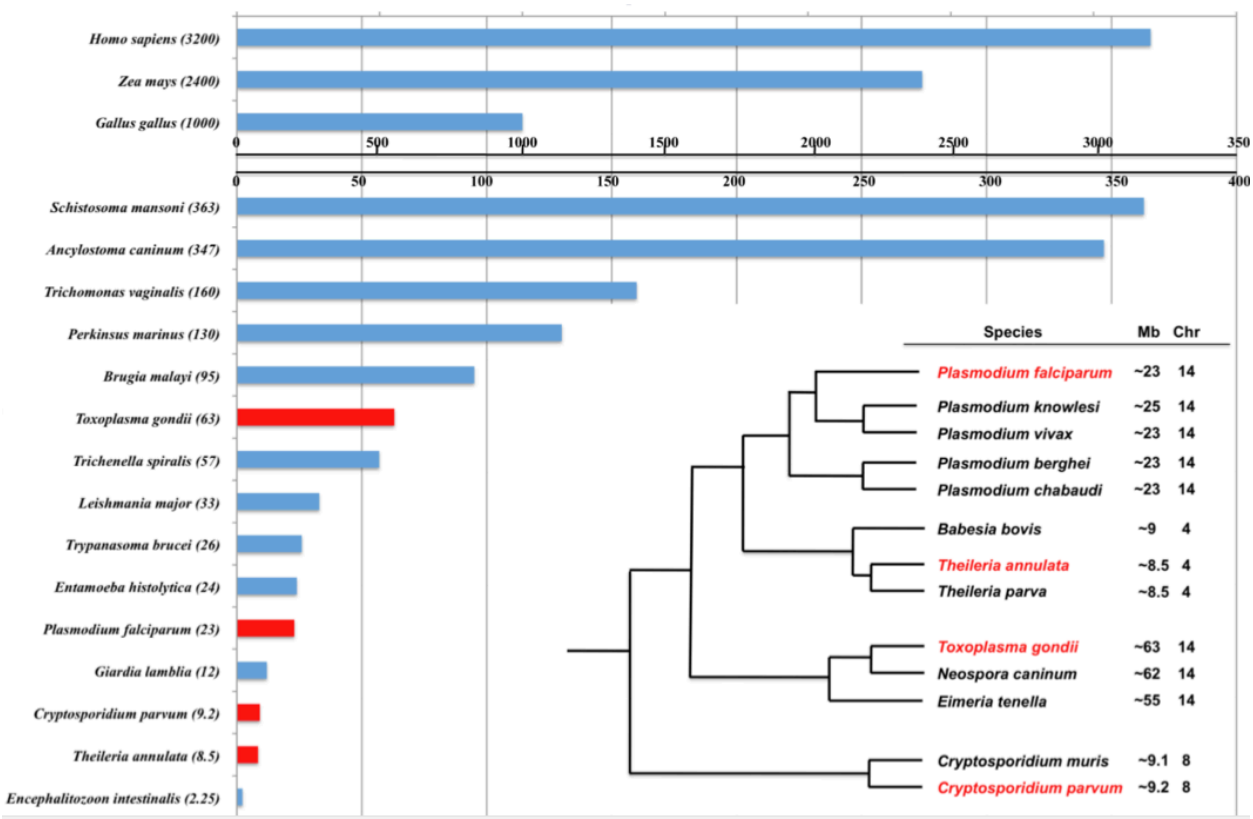
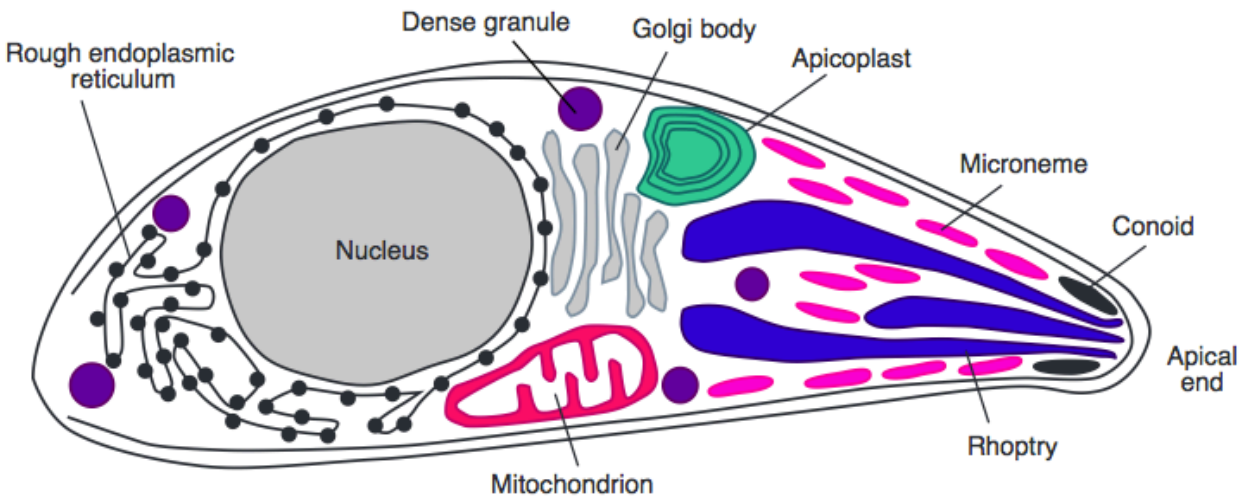


Figure 1.1. Apicomplexan cladogram



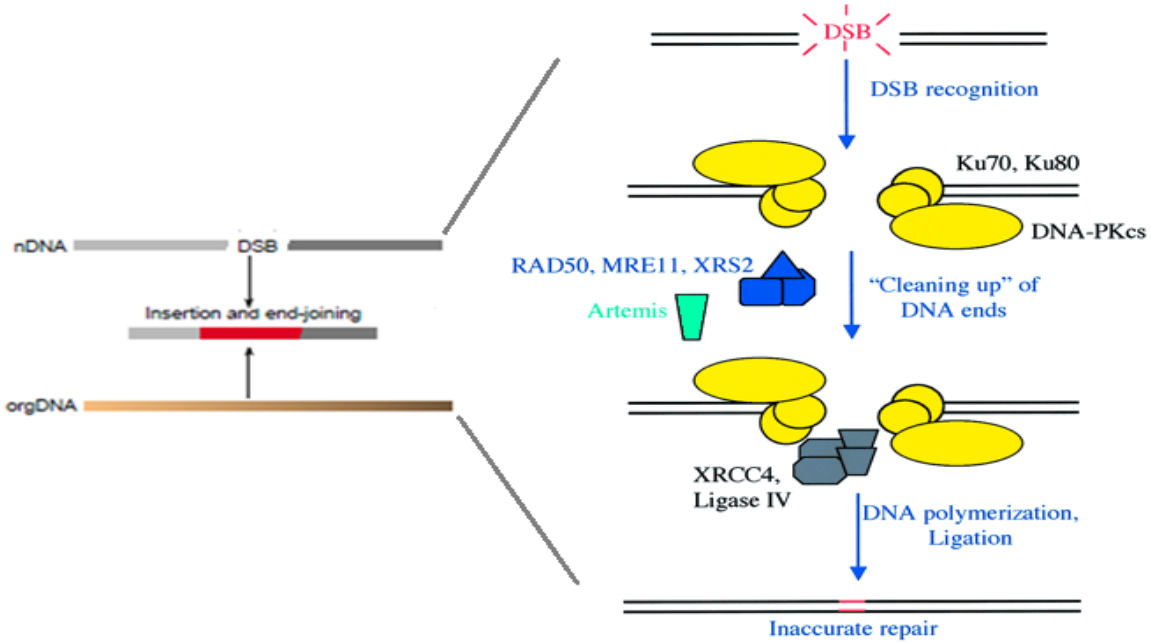
(Reproduced from Kissinger and DeBarry, Trends Parasitol. 2011 Aug; 27(8): 345–354.)

Figure 1.2. Comparison of genome sizes of select eukaryotes



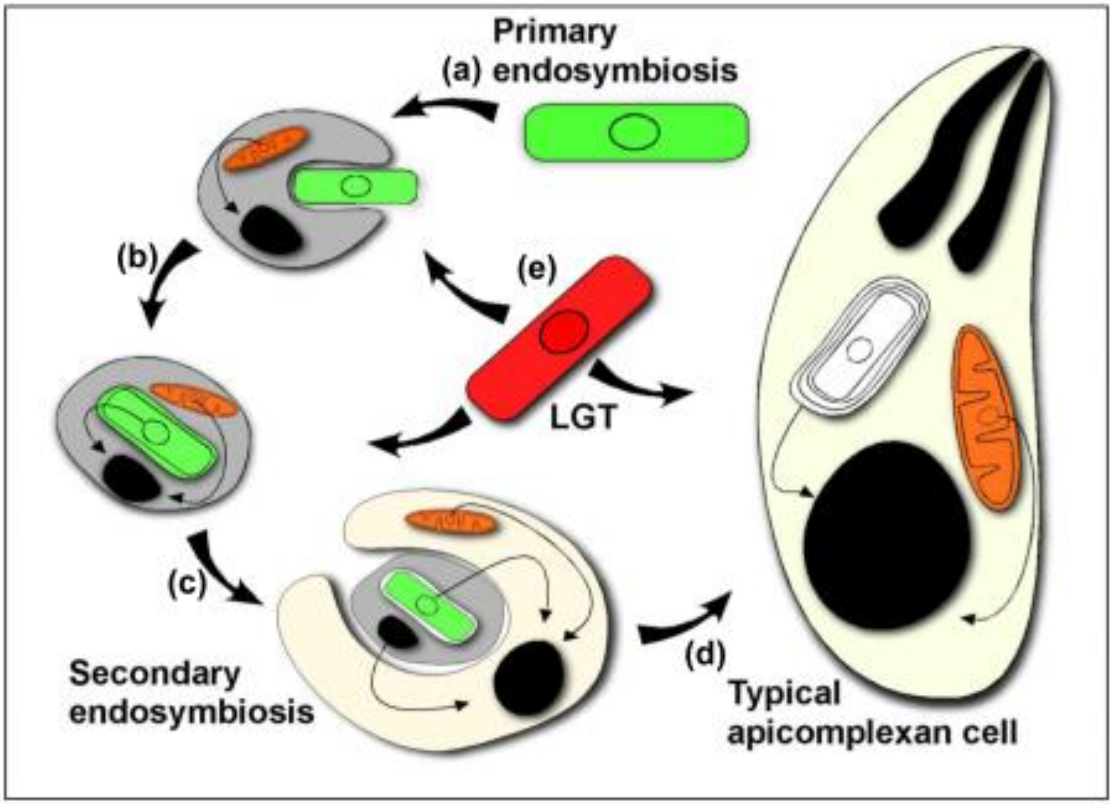
(Reproduced from Ajioka *et al.*, Expert Reviews in Molecular Medicine 2001, Cambridge University Press.)

Figure 1.3. Ultrastructure of a *Toxoplasma gondii* tachyzoite



Modifidied from Leister, Trends in Genetics, 2005, Pages 655-663 and Jackon, Carcinogenesis (2002) 23 (5): 687-696

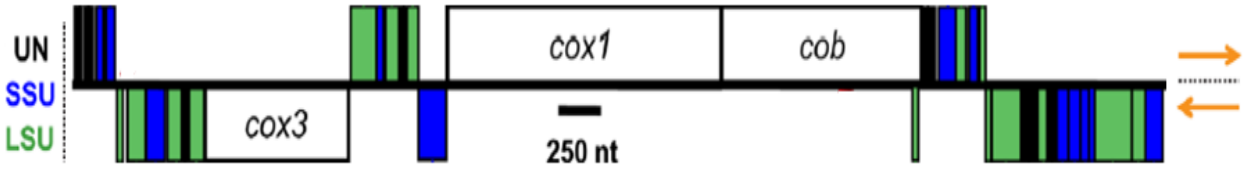
Figure 1.4. Model of double-strand break repair via non-homologous end joining



(Reproduced from Kissinger and DeBarry, Trends Parasitol. 2011 Aug; 27(8): 345–354.)

Figure 1.5. Schematic illustrating evolution of the apicomplexan cell

A



(Modified from Feagin JE*, et al.,* (2012), *PLoS One* 7(6):e38320)

B

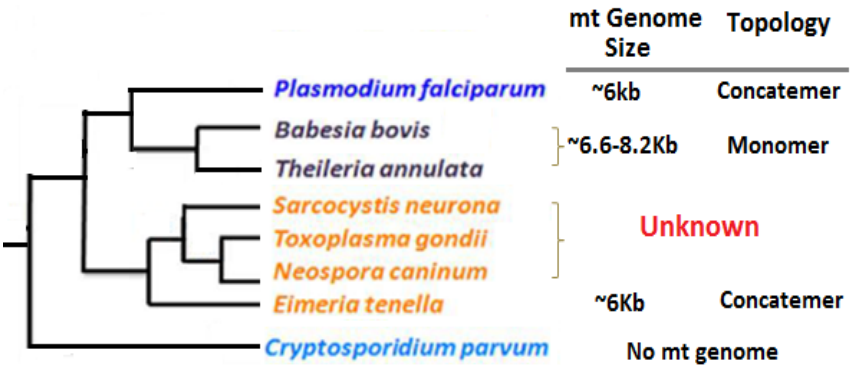
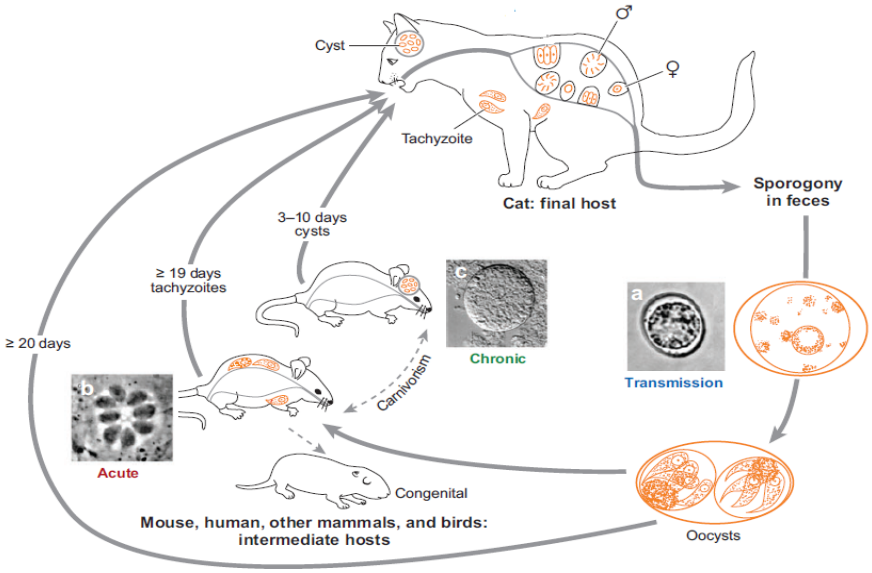


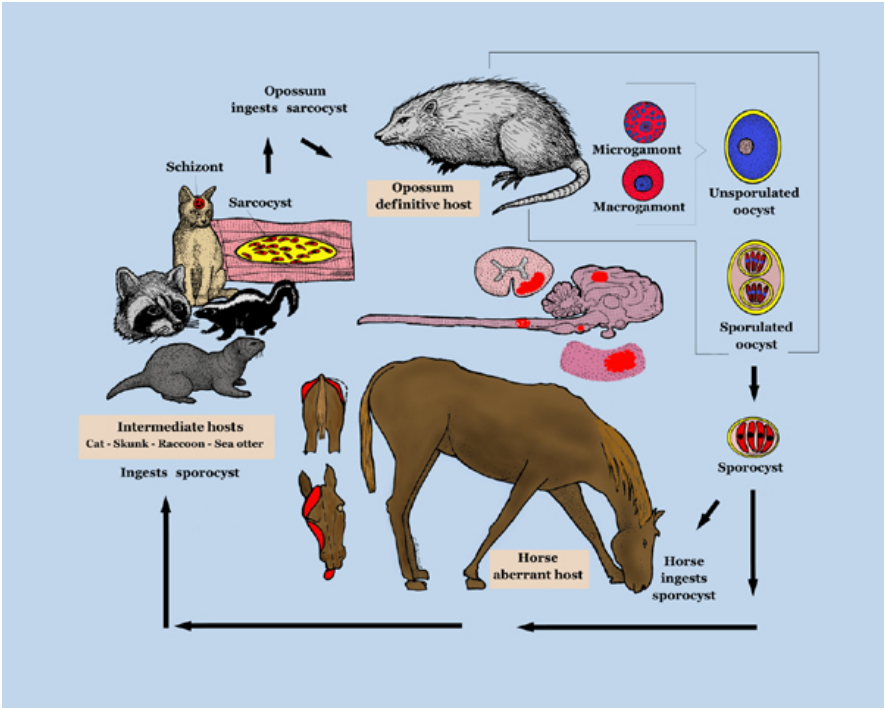
Figure 1.6. Apicomplexan mitochondrial genomes

A



(Reproduced from Sibley and Ajioka, Annu. Rev. Microbio. 2008.62:329-351)

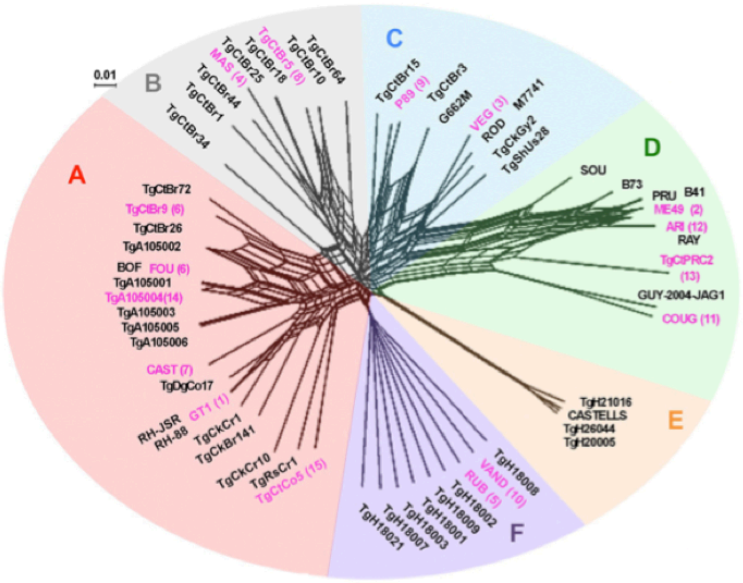
B



(Reproduced from http://www.ars.usda.gov)

Figure 1.7. Life cycles of *Toxoplasma gondii* and *Sarcocystis neurona*

A



B

(Reproduced from Lorenzi *et al*., In Review)

Figure 1.8. Population structure of *T. gondii*

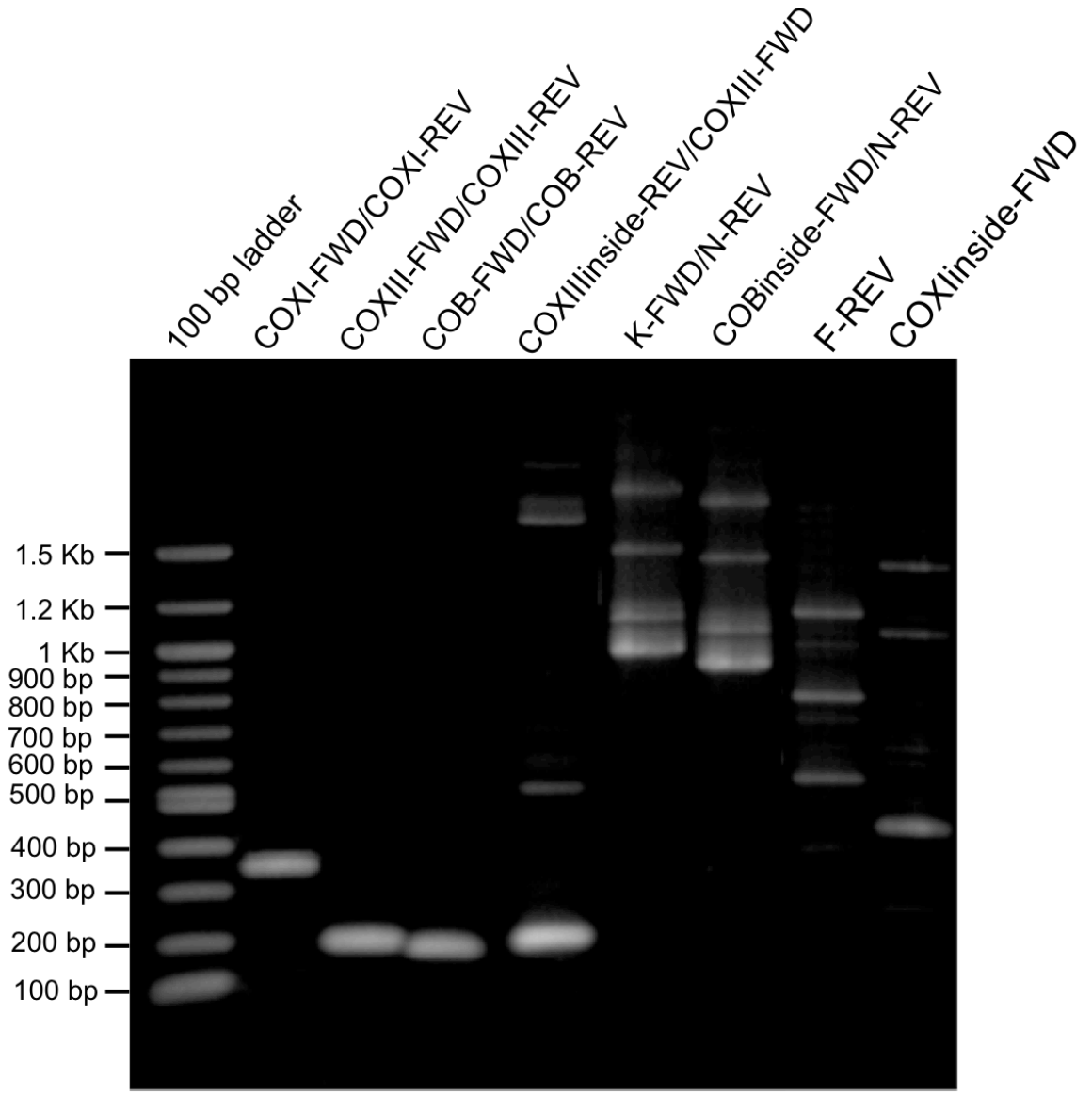


Figure 1.9. PCR analysis of *T. gondii* mtDNA

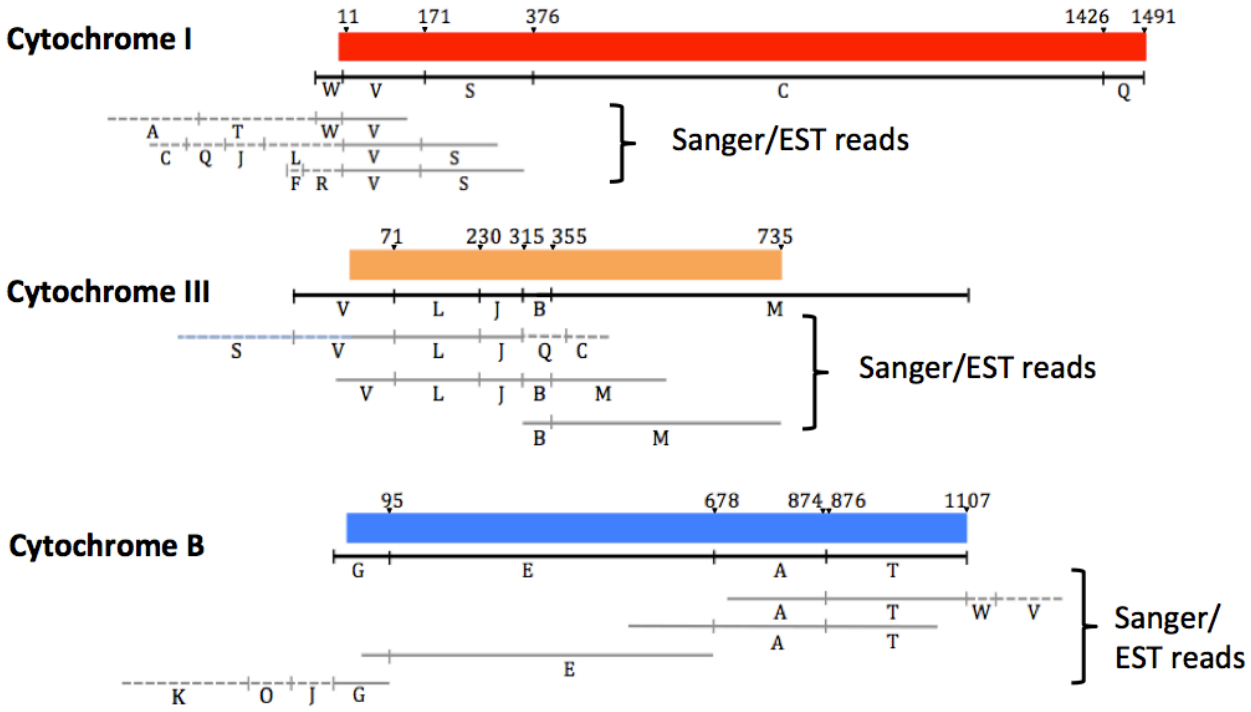
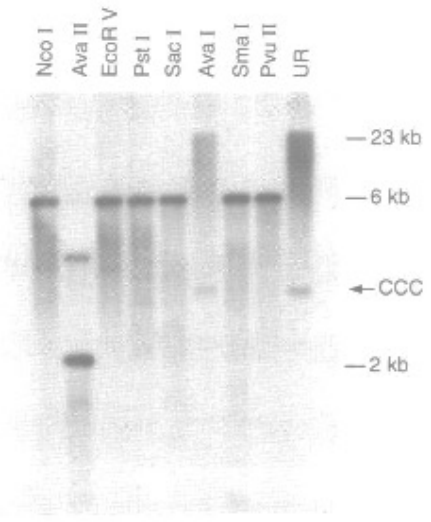


Figure 1.10. *T. gondii* mitochondrial protein-coding genes can be assembled using mtDNA sequence elements

A.



(Reproduced from Preiser *et al.*, The EMBO Journal, vol. 15 no. 3 pp. 684-693, 1996)

B.

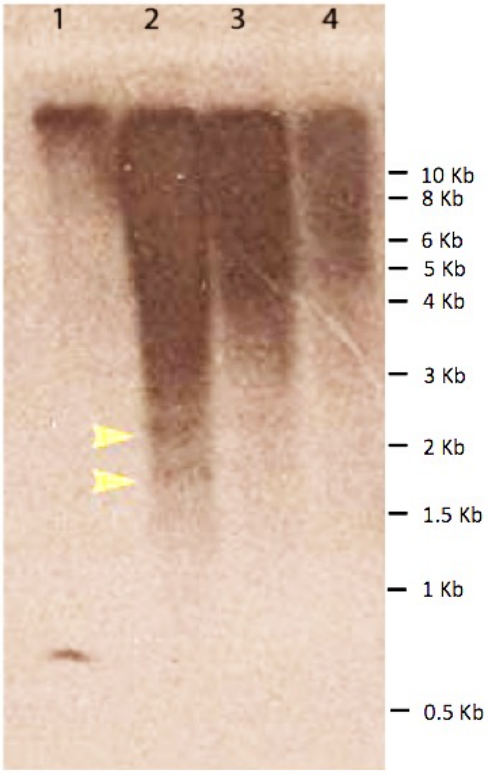


Figure 1.11. Structural characterization of the mitochodondrial genome via Southern analysis

Table 1.1. Summary of sequenced apicomplexan genomes

	Species	Chr	Genome size (Mb)	Annotated proteins
Haemospordia	*Plasmodium falciparum* 3D7	14	23.33	5,777
	Plasmodium reichenowi CDC		23.92	6,069
	Plasmodium chabaudi chabaudi		18.87	5,297
	Plasmodium berghei ANKA		18.56	5,164
	*Plasmodium yoeli*i yoelii		22.94	7,774
	Plasmodium cynomolgi B		26.18	5,776
	*Plasmodium knowles*i H		23.74	5,342
	Plasmodium vivax Sal-1		27.76	5,626
Piroplasma	*Babesia bigemina* BOND	4	13.84	5,125
	Babesia bovis T2Bo		8.18	3,781
	Babesia microti RI		6.39	3,554
	Theileria annulata Ankara		8.36	3,845
	Theileria equi WA		11.67	5,397
	Theileria orientalis Shintoku		9.01	4,058
	Theileria parva Muguga		8.35	4,167
Coccidia	*Eimeria acervulina* Houghton	14	45.83	7,105
	Eimeria brunetti Houghton		66.89	9,025
	Eimeria falciformis Bayer Haberkorn 1970		43.67	6,749
	Eimeria mitis Houghton		72.24	6,369
	Eimeria necatrix Houghton		55.01	8,864
	Eimeria praecox Houghton		60.08	8,025
	Eimeria tenella Houghton		51.86	8,634
	Sarcocystis neurona SN3	N/A	124.38	7,039
	Neospora caninum Liverpool	14	59.1	7,266
	Hammondia Hammondi H.H.34		67.7	8,176
	Toxoplasma gondii ME49		65.67	8,920
	Cryptosporidium hominis TU502	8	8.74	3,956
	Cryptosporidium muris RN66		9.25	3,980
	Cryptosporidium parvum IowaII		9.1	3,886
	Gregarina niphandrodes	N/A	14.01	6,606

Table 1.2. Primers used in PCR analysis of the *T. gondii* mt genome

Primer Name	**Sequence**
COXI-FWD	5'-TGA TTG GTT AAT TGG AGG ACT TGC TGT-3'
COXI-REV	5'-GTT TGA GAT ACA ACA CCA AAA GCA GG-3'
COXIII-FWD	5'-TCA TGT TAT TGT CGG TGC TAT CTT GG-3'
COXIII-REV	5'-GAT CAT TAT CCA CAC TGC TTC GAC GA-3'
COB-FWD	5'-CGT AGT AAC CTC CAA GTA GCC AAG G-3'
COB-REV	5'-AAC TAC CGC TTG GAT GTC TGG TTT AG-3'
COXIIIinside-REV	5'-ATC CAC ACT GCT TCG ACG AAA TGT AGA-3'
N-REV	5'-GAA GTT ATG GTT TTG GGC TCG TGA GT-3'
K-FWD	5'-TCT TTG CCT GGA GGT TTG TTA CGT T-3'
COBinside-FWD	5'-TTG ATA CCG CGC TTA AAG TTG CCT 3'
F-REV	5'-GGG GAC AAA AAG ACA TCA CGA T-3'
COXIinside-FWD	5'-ATG ATC CGC GAA CCA GAA CTG TAT AAC T-3'

Table 1.3. Genomic and EST reads representing different arrangements of mtDNA elements

Gene	Elements in EST read	GenBank Accession ID of EST	Trace Archive ID of a genomic read aligning with the EST read
COXI	A, T, W, V	CN617107.1	gnl|ti|2057042422
	C, Q, J, L, V, S	DK934897.1	gnl|ti|2056951560
	V, S, R, F	CV701032.1	gnl|ti|2057358855
COXIII	C, Q, J, L, V, S	DK934897.1	gnl|ti|2056951560
	V, L, J, B, M	CV654565.1	gnl|ti|2057374420
	B, M	DV110375.1	gnl|ti|2064971134
COB	A, T, W, V	CN617107.1	gnl|ti|2057042422
	A, T, E	CV654565.1	gnl|ti|2057374420
	G, E	CB384005.1	gnl|ti|2057332917
	K, O, J, G	CB752279.1	gnl|ti|2057396732

Table 1.4. Comparison of *T. gondii* and *N. caninum* mtDNA elements

Element	Length of element in *T. gondii*	Co-ordinates of elements in *N. caninum*		% Identity to *T. gondii* element
		Element Start	Element End	
A	196	1	196	95
B	40	1	37	92.5
C	1049	1	1049	90
D	82	1	82	100
E	583	1	583	97
F	179	1	179	98
G	100	1	100	97
H	447	1	446	98
I	204	1	204	97
J	85	1	85	98
K	445	1	445	98
L	159	1	158	98
M	754	3	754	96
N	166	1	166	99
O	86	1	86	98
P	184	1	184	99
Q	65	1	65	96
R	85	1	85	97
S	205	1	205	98
T	233	5	233	90.1
U	353	1	353	100
V	161	1	161	100
W	45	1	40	88.89

CHAPTER 2

EVOLUTIONARY FATE AND CONSEQUENCE OF >11,000 NUCLEAR-INTERGRATED ORGANELLAR DNAS IN THE ZOONOTIC PARASITE, *TOXOPLASMA GONDII*

Namasivayam, S.*, Sun, C. *, Barrie, A. B., Xiao, W., Hall, E. M., Oberstaller, J., Feschotte, C., Kissinger, J. C. and Pritham, E. J. To be submitted to PNAS

ABSTRACT

Toxoplasma gondii is a significant zoonotic pathogen that infects up to 50% of humans in some regions. This unicellular parasite contains three genome sequences: one nuclear (65 Mb); one plastid organellar, ptDNA (35 Kb); and one mitochondrial organellar, mtDNA, whose sequence has been elusive. Traditional mtDNA isolation methods have been hampered by a highly polymorphic mitochondrion, a degenerate genome and nuclear insertions of small mtDNA fragments, called (NUMTs). We employed techniques that allowed the identification of 23 sequence elements that constitute the mtDNA genome. Homology-based analyses of NUMT and NUPT (ptDNA insertions) revealed that insertion/deletion is a frequent and active process generating polymorphism between strains. NUMT accretion has generated 1.4% of the *T. gondii* ME49 genome. This is the highest percentage ever reported for a eukaryote. Our analyses indicate that within apicomplexans, NUM/PTs are found throughout the Coccidia. Comparisons with *Neospora caninum* (28 MY divergence) revealed that the movement and fixation of NUMTs predates the species split. Because most of the NUMT insertions reside within (~60%) or near (~23% <1 Kb away) genes, we hypothesize that they might impact gene expression. Candidate NUMTs were examined *in vitro* for gene regulatory effects. Results indicate that NUMTs carry, or evolve into, *cis*-elements that influence gene expression. These observations portray a role for organellar sequence insertion in shaping the genetic architecture of this important human pathogen. NUM/PTs are excellent candidates for contributions to strain-specific differences involved in adaptation, virulence and speciation.

INTRODUCTION

When contemplating genome evolution and the differences between strains and species, it is common practice to examine single nucleotide polymorphisms (SNPs), insertion/deletion events (indels), rearrangements, horizontal and intracellular gene transfers and transposable element (TE) insertion or excision. The insertion of random fragments of organellar DNA is not high on the list of features examined as a cause of phenotypic or genotypic diversity, but if you are studying a Coccidian parasite, it should be. The protist *Toxoplasma gondii* is a cosmopolitan apicomplexan parasite, capable of infecting nearly all warm-blooded animals including humans. *T. gondii* causes toxoplasmosis, a disease that infects as much as one third of the world's human population (1, 2). While the parasite rarely causes symptoms in healthy adults, it can lead to serious and even life-threatening illness in immunosuppressed or pregnant individuals (3, 4). Phylogenetic and population studies of *T. gondii* reveal an unusual population structure consisting of mostly clonal lineages that occasionally recombine (5) as well as the recent evolution of oral infectivity that permits the parasite to spread via consumption of tissue cyst forms in addition to traditional transmission via oocysts in the environment (6).

In eukaryotes, DNA of endosymbiotic organelles (mitochondria and chloroplasts) has been observed to be transferred to the nuclear genome (7, 8). During the early phase of organelle evolution, this process resulted in a massive relocation of organellar genes to the nuclear genome. As many as 75% of yeast nuclear genes are derived from the ancestral mitochondrial genome (9). In many eukaryotes, the transfer of functional genes appears rare or has ceased altogether (10-12) and almost all recent transfers of mitochondrial (mtDNA) or plastid (ptDNA) DNA to the nuclear genome are fragmental in nature and give rise to noncoding sequences, called NUMTs (nuclear integrants of mtDNA) and NUPTs (nuclear integrants of ptDNA).

Organellar-to-nuclear DNA transfer has been reported for many eukaryotes and a review based on 85 fully sequenced eukaryotic genomes reveals the significant driving force NUM/PTs provide for gene and genome innovation in eukaryotes (Table 2.1) (7, 13). In *Saccharomyces cerevisiae*, the non-homologous end-joining repair pathway, NHEJ, was shown to be the primary mechanism responsible for NUMT integration (14).

Apicomplexan genomes are streamlined, dynamic and rapidly evolving (15, 16). Transposable elements (TEs), which generally constitute the most significant proportion of repetitive DNA in genomes and are known to be powerful agents for generating genomic plasticity (17), are absent from most apicomplexans (18). Consequently, it is of interest to understand what factors contribute to the generation of genomic plasticity in these organisms. While organellar-to-nuclear DNA transfer represents a significant driving force for genome innovation in eukaryotes, it has not been well defined in apicomplexans, and little is known about the evolutionary fate and consequence of these transferred DNA sequences.

Most apicomplexans contain two organelles, a mitochondrion and an apicoplast that was acquired through the secondary endosymbiosis of an alga (19). The ptDNA sequence (35 Kb) is a circular/cruciform structure and is well conserved. The mtDNA sequences generated thus far are variable and occur as linear monomers (~6 Kb) or concatemers with, or without, inverted repeats depending on the species (20). Despite the topological differences, all characterized apicomplexan mtDNAs encode three and only three cytochrome genes: cytochrome oxidase subunits I and III (*cox1, coxIII)* and cytochrome b (*cob*) and also contain highly fragmented large and small subunit rRNA genes (LSU – SSUrRNA)(21, 22). Unlike the ptDNA sequence, the mtDNA sequence of *T. gondii* has remained elusive. Numerous NUMTs, called REP elements when discovered because of their abundance, were first reported by Ossorio *et al.,* (23). The high

NUMT density in *T. gondii* has interfered with molecular approaches to identify the mtDNA sequence. Physically, the mitochondrial organelle is highly polymorphic, making it hard to isolate (24). We used data mining and stringent molecular approaches on mitochondrial-enriched tachyzoite-stage cell fractions to identify a set of mtDNA sequence elements, which constitute the *T. gondii* mtDNA. Using the organellar DNA sequences, we performed a systematic examination of NUM/PTs in *T. gondii* to understand their evolutionary fate and contribution to genome innovation.

MATERIALS AND METHODS

Parasite culture and mitochondrial enriched fractions

Toxoplasma gondii parasites were propagated on human fibroblast reverse transcriptase (hTERT) cells grown in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% heat-inactivated Cosmic Calf Serum, 0.5% 10mg/ml pencillicin-streptomycin and 0.05% 10 mg/ml gentamycin. Harvested parasites were resuspended in lysis buffer and then passed through cotton mesh, 8 and 5 micron filters. The suspension was collected via centrifugation at 1000 g and fractured by pre-cooled silicon powder 3 or 4 times, 15 sec each time on ice. Fractured cell components were centrifuged at 50 g, 150 g (twice) and then 1000 g or 1200 g to collect mitochondrial-enriched fractions.

Data mining strategies, PCR, sequencing and annotation

Unassembled non-chromosomal contigs and ESTs of *T. gondii* (ToxoDB.org release 5.1) were mined using BLASTN (E-value 10^{-10}) and other known apicomplexan mitochondrial genomes for initial identification of the *T. gondii* mitochondrial DNA sequences. Primers were designed based on identified sequences and publicly available cDNA sequences. All primers had a Tm of higher than 57°C. The Platinum Taq Polymerase protocol was used for PCR with

mitochondrial enriched fractions as template. PCR products were gel extracted and sequenced. New primers were designed based on PCR sequencing data and the process was repeated until no new sequences were identified. The 23 mitochondrial sequences were determined manually based on their occurrence and arrangements in sequenced PCR products and unassembled genomic and EST reads obtained from NCBI dbEST (*Toxoplasma* EST project) and Trace Archives (Assembly ID: 2606).

Genes were predicted using TBLASTN and the NCBI ORF finder (genetic code - protozoan mitochondrion). ClustalX with default settings was used for multiple sequence alignment with other predicted apicomplexan mitochondrial protein sequences. RNA genes were identified by comparison with *Plasmodium falciparum* mitochondrial rRNA (22) via ClustalW alignments and conserved nucleotides. The identified RNA genes were folded manually and compared with the secondary structure of *E.coli* and *Plasmodium* rRNAs.

N. canium mtDNA sequences were identified computationally using the *T. gondii* mtDNA sequences from a non-chromososmal contig (contig 8315) in ToxoDB.org v5. The latest *N. caninum* assembly does not contain this contig.

Retrieval of genome sequences

Nuclear genome sequences for *Toxoplasma gondii* (version 2014-04-23) *Neospora caninum* (version 2011-08-12), *Hammondia hammondi* (version 2014-06-30*), Sarcocystis neurona* (version 2013-07-02*) and Eimeria tenella* (2013-11-05), including unassembled contigs, were downloaded from ToxoDB (http://toxodb.org/). *Plasmodium falciparum, Babesia bovis* and *Theileria annulata* were obtained from PlasmoDB.org and PiroplasmaDB.org respectively.

Identification of NUMTs and NUPTs

RepeatMasker (ver 4.0.5) (http://repeatmasker.org/) rather than the frequently used BLAST was used to identify NUM/PTs because it provides better accuracy than BLAST, handles many-to-one hits better by correctly avoiding overestimates and provides a detailed annotation of the identified repetitive sequences, which facilitates subsequent characterization. In the present study, mitochondrial and apicoplast genome sequences were used as the "repeat library" to mask the corresponding nuclear genomes after filtering out low complexity sequences. Other parameters were set at default values. For *T. gondii* and *N. caninum,* the identified species-specific 23 sequence elements as well as reconstructions of each of the cytochrome coding sequences were used to mask the genome sequences. *S. neurona.* and *H. hammondi* were masked with *T. gondii* mtDNA. Since only the 23 mtDNA elements and the cytochrome gene sequences are used to identify NUMTs, any NUMTs that originate from junctions that might be present in the actual mt genome sequence will be missed. *E. tenella* was masked with its published mtDNA (GenBank:AB564272.1). The *T. gondii* apicoplast sequence (ToxoDB.org release 11, tgme49_asmbl.1944) was used to mask *T. gondii*, *N. caninum and H. hammondi*. *S. neurona* and *E. tenella* were masked with their respective apicoplast sequences (*S.n.*- unpublished, *E.t.*- GenBank:AY217738.1) *Plasmodium, Babesia* and *Theileria* species-specific organellar sequences (obtained from EuPathDB.org) were used to screen their respective nuclear genome sequences.

Since non-contiguous DNA from organellar genomes can get inserted next to each other in a single insertion event, one caveat of the above approach is the possible overestimation of the number of insertion events. However, the identification of the amount of DNA inserted is not

affected. In addition, we did not count NUM/PTs shorter than 28 bp and overlapping insertions were only counted once.

Identification of NUMT location

Files containing the coordinates of coding regions, introns, exons, intergenic and 1 Kb upstream and downstream flanking regions were generated using the annotation data on ToxoDB.org (release 13.0). Based on coordinates, a custom script was used to identify a genomic location for each NUMT. NUMTs located in non-coding exons were classified as being present in the UTR. NUMTs spanning more than one genomic feature were classified separately.

Calculation of NUM/PT age

To estimate the timing of NUM/PT insertions, we employed a phylogeny-independent approach. As most NUM/PTs are pseudogenized on arrival into the nucleus (8) and therefore evolve at a neutral rate, the age of insertions can be roughly calculated by comparing the NUM/PT sequence divergence from the mtDNA or ptDNA and application of the neutral substitution rate of the species. Percent divergences reported by RepeatMasker were converted to nucleotide distance measures using the Jukes-Cantor formula to correct for multiple hits. The age of each NUM/PT was then calculated by dividing the nucleotide distance by the *T. gondii* neutral mutation rate of 2.12×10^{-8} substituions/base/million years (6).

Identification of segmental duplications and strain-specific NUM/PTs

Segmental duplications were identified by extracting NUM/PT sequences along with 150 bp of upstream and downstream flanking regions from the *T. gondii* ME49 nuclear genome. The extracted sequences were used as BLASTN queries against the *T. gondii* ME49 genome sequence. If a NUM/PT along with at least 100 bp of flanking sequence was present at least twice, it was counted as a segmental duplication.

RepeatMasker output was parsed to identify *Toxoplasma* strain-specific NUM/PTs. In short, NUM/PTs from each strain along with 200 bp of flanking sequence were used in a BLASTN search against the other genomes. NUM/PTs that appeared in one of the three genomes with gaps in the reciprocal locations in the two other genomes were identified. Candidate strain-specific NUM/PTs were verified using Mercator alignments in ToxoDB.org.

Identification of orthologous NUM/PTs.

Orthologous NUM/PTs between *T. gondii* and *N. caninum* were identified by extracting *T. gondii* ME49 NUM/PTs and 100 bp of flanking sequence and using them to search the *N. caninum* genome via BLASTN. Hits had to cover both the NUM/PT and >50 bp of flanking sequence to be candidate orthologs. Orthologous NUMTs were confirmed by evaluating their presence in all three *T. gondii* strains and *N. caninum* via whole genome alignments at ToxoDB.org. We extracted genic sequences near *T. gondii* and *N. caninum*, orthologous NUMTs, and aligned them with Vista plot (http://genome.lbl.gov/vista/index.shtml) to examine conservation.

Expression constructs and transient transfection

PCR primers containing the attB1 and attB2 sites were used to amplify the appropriate promoter and 5′-UTR regions from Type II genomic DNA for the two genes tested. A two-step overlap-extension PCR technique (Sambrook *et al*. 1989) was employed to delete the NUMT sequence from each promoter. The Gateway™ cloning system (Invitrogen) was used to clone the WT and deletion promoters. These two kinds of promoters were first cloned into pDONR221 via the BP reaction. Following sequencing verification, these promoter fragments were moved into a firefly luciferase-expressing vector (destination vectors) via the LR reaction. A constitutive *T. gondii* promoter (α-tubulin)-driven renilla luciferase-expressing construct was co-transfected

along with the experimental construct. Nucleotide positions in these deletion studies are referenced with respect to the start of translation (+1). Parasite culture and transient transfections were performed as described (25). In short, *T. gondii* RH tachyzoites were transiently transfected via electroporation. The cells were scraped and lysed 18-24 hours post-electroporation using passive lysis buffer (Promega, Madison, WI, USA). A dual luciferase assay was performed on the extract using the Promega Dual Luciferase kit. The different substrate requirements for firefly and renilla luciferase allowed us to assay reporter expression for each construct sequentially within the same extract. Reporter activity from the WT or mutagenized promoter was measured relative to the internal control, eliminating errors due to variation in parasite populations and individual transfections. Each electroporation experiment was conducted six times and luciferase assays were performed in duplicates for expression measurements. The unpaired Students *t*-test was used to calculate the statistically significant difference in expression levels between WT and mutagenized promoter activity; $p < 0.05$ was considered statistically significant.

RESULTS

The mtDNA of *T. gondii* encodes three proteins and many ribosomal RNA fragments

Efforts to clone the mtDNA of *T. gondii* by hybridization of genomic libraries, PCR from genomic DNA, reverse transcription from tachyzoite-stage RNA and screens of assembled *T. gondii* genome sequence were unsuccessful in the definitive identification of a linear genome sequence as is present in other apicomplexan species. Each approach did however yield portions of sequences in a variety of arrangements that encode genes found in other apicomplexan mtDNAs. We used these initial findings to design several rounds of increasingly-specific PCR primers and repeated analyses on DNA and RNA from mitochondrial-enriched tachyzoite

fractions as well as bioinformatic searches of individual unassembled EST and Sanger sequence reads generated as part of several different previous studies on *T. gondii*. This approach resulted in a collection of 23 sequence elements that constitute the nearly complete, if not complete, mtDNA of *T. gondii* (Table 2.2). We use the word "elements" because none of the sequences encodes a functional gene by itself and because these elements occur in dozens of different, non-random, permutations. However, together, these sequence elements encode *coxI, coxIII* and *cob* in their entirety (Figure 2.1) as well as 12 ribosomal RNA fragments, LSUA, LSUB, LSUD, LSUE, LSUF, RNA10 and SSUA, SSUB, SSUE, SSUF, RNA8 (22).

Thirteen sequence elements are required to construct the full coding regions of *coxI, coxIII* and *cob* (Figure 2.1). Multiple sequence alignments of the predicted protein sequences reveal significant conservation with other apicomplexan cytochrome sequences (Figure 6.2). RT-PCR as well as examination of individual EST reads present in the NCBI GenBank support the joining of these sequence elements to generate putatively functional products, however, at both the RNA and DNA level other physical arrangements exist (Figure 2.1, Table 2.3). There are 83 non-chromosomal *T. gondii* ME49 contigs that demonstrate >98% identity to the 23 sequence elements (Table 6.1) as well as a large number of individual Sanger sequence reads, ESTs and sequenced PCR products.

Given the observation of sequences of mitochondrial origin in the nuclear genome, it was necessary to demarcate with certainty mtDNA sequences from those encoded in the nuclear genome. This was accomplished in several ways. The assembled *T. gondii* ME49 chromosomes were searched for the presence of each of the 23 sequence elements. Nuclear insertions differ by being fragmental and displaying the hallmarks of sequence erosion (Table 2.2), a pattern consistent with DNA not constrained by natural selection. Sequences with perfect matches to the

identified sequence elements are most often obtained from enriched mitochondrial fractions and no sequenced PCR products from these fractions could be found in their entirety in the nuclear genome sequence. Finally, the mitochondrial insertions in the nuclear genome are present in multiple copies that display a range of mutations relative to the identified mtDNA sequence elements. All 23 elements were present in the nuclear genome (Table 2.2). Full-length perfect copies of 8 out of the 13 elements that make up the cytochrome genes can be reconstructed from the chromosomally-encoded copies. But it was not sufficient to reconstruct full-length coding sequences for *coxI, coxIII* or *cob* from these chromosomally-encoded elements.

NUMTs and NUPTs are identified in *T. gondii*

We employed the homology-based program RepeatMasker with a Smith-Waterman algorithm and the *T. gondii* mtDNA sequence elements and protein gene sequences or the ptDNA genome sequence as queries to mask regions of the nuclear genome derived from organellar DNA. To obtain reliable and conservative estimates, only chromosomal assemblies were queried for NUM/PTs. NUM/PTs in other species typically represent < 0.1% of total genomic DNA, with a few exceptions noted in some plant (0.26%) and yeast (0.29%) species (7). The fraction of the nuclear genome composed of NUM/PTs in *T. gondii* and related species represents the largest ever reported for any eukaryote (1.4% NUMTs and 0.18% NUPTs) and the NUMT insertions are nine times more frequent than is observed in the honeybee genome, the current record holder for metazoans, and over 80 times more frequent than human NUMTs (13)(Figure 2.2A, Table 2.1).

The start and end coordinates for all identified NUM/PTs were used to infer their size. Most (>90%) of the NUM/PTs are 50 bp to 200 bp in length, although a few larger NUM/PTs were identified (Figure 2.3). This size distribution is similar to that observed for yeast and rat

species, but is dramatically different from that of *Arabidopsis*, *Neurospora*, and *Ciona* (26). A single 3,369 bp NUMT that does not contain any complete coding sequences is present in the *T. gondii* ME49 genome sequence (Table 2.4). It is 99% identical to individual mtDNA sequence elements, suggesting it is a recent insertion. This 3,369 bp insertion could be identified in ME49 PacBio reads spanning this region (unpublished data). NUMTs originate from all 23 mtDNA sequences, including both coding and non-coding regions, however there are some differences in their frequency of occurrence. Sequence elements V, M, E, K and J are detected most often (Table 2.2). Portions of the element V (on opposite strands) are utilized to encode portions of *coxI* and *coxIII* respectively. Thus, we suspect it is present twice in the mtDNA both because of its higher frequency and because it is bounded by different elements in each case (Figure 2.1). To understand the insertion pattern of NUM/PTs in the *T. gondii* genome, we took a closer look at the distribution of NUM/PTs within and around genes. Table 2.5 shows the percentage of NUMTs identified in different parts of the *T. gondii* nuclear genome. We find that ~60% of the identified NUMTs are located within genic regions (introns and exons) with the majority in introns. In addition, we can see that ~23% of NUMTs are present within the 1 Kb flanking regions of genes. Overall, ~85% of *T. gondii* NUMTs reside within and around genes. This finding is not surprising given the highly compact nature of the *T. gondii* genome, 8,322 protein-encoding genes (8920 annotated genes overall) in 65 Mb. ~70% of the *T. gondii* genome is made of genes.

To determine if NUM/PTs are duplicated post-insertion, we reasoned that NUM/PT duplications would likely involve flanking regions of nuclear origin. Therefore, we considered NUM/PTs to be the product of a segmental duplication event if the NUM/PT along with100 bp of its flanking regions, were duplicated in the nuclear genome. Using this criterion, 105 NUMTs

and 13 NUPTs were identified as duplicated one or more times in the ME49 genome sequence (Table 6.2). Therefore, approximately, 0.7% of the 9,356 NUMTs were involved in post-insertion duplication events. The remaining NUMTs are likely independent insertions from the mtDNA in the mitochondrion.

NUMTs and NUPTs are acquired frequently and evolving rapidly

To determine the age distribution of the NUM/PTs, a phylogeny-independent molecular clock-based approach (see Methods) was employed. This analysis revealed a continuous influx of NUMT assimilations with an accumulation peak approximately 12.7 million years ago (Figure 2.4). While NUMT assimilation appears to be a continuing process, most insertions occurred prior to the last 1 million years, suggesting a recent slowing of the process. If the age of the NUM/PTs were calculated using a more recent estimate of a lower mutation rate (27), the NUM/PTs become approximately twice as old but the pattern of the age profile remained unaltered.

To examine the possibility of strain-specific insertions, two additional strains, GT1 and VEG, were examined. Our strategy (see methods) yielded a total of 32 strain-specific NUMTs: 12 in ME49, 13 in GT1 and 7 in VEG (Figure 2.2, Table 2.6). Strain specificity of NUMTs was confirmed by multiple sequence alignment (see Methods). In total, 26 of the 32 strain-specific NUM/PTs were found in or near annotated genes, with 16 detected in introns, 1 in 3′ UTR, and 9 within 2 Kb of annotated genic regions, suggesting that their presence has the potential to affect activity of the associated gene. The NUMT content is very similar in all three strains ranging from 1.41 % in VEG to 1.43% in both the ME49 and GT1 strains. One strain-specific NUPT was identified in ME49.

To ascertain if the strain-specific NUM/PTs arose via either a novel insertion or deletion event, we determined the precise boundaries by performing a three-way genome comparison of the chromosomal regions in question. A NUM/PT was considered a novel insertion if it was precisely missing from two of the three strains and displayed high sequence identity (>99%) when compared to the mtDNA sequence elements. If a NUM/PT insertion was only missing in one of the three strains and the NUMP/T in the 2 strains displayed high sequence divergence (>10%) when compared to the mt/ptDNA, then we concluded that this was a strain-specific deletion. Using these criteria, 5 isolate-specific NUM/PTs were confidently identified as novel insertions, and 13 were inferred to have been false positives due to deletion (most are partial deletions). An example of a strain-specific NUMT insertion and deletion are illustrated in Figure 2.5. Together these results indicate that the insertion/deletion of NUM/PTs is an ongoing process that contributes to the divergence of *T. gondii* strains.

Apicomplexan NUMTs and NUPTs are restricted to the Coccidia

To determine if the mtDNA sequence elements and their observed permutations are unique to *T. gondii*, we investigated its closest relative, *Neospora caninum*. We find that the *N. caninum* mtDNA also consists of 23 sequence elements that are highly similar to those in *T. gondii* (Table 2.2). *N. caninum* mtDNA elements are also observed to occur in permutation patterns. However, the NUMP/T content in *N. caninum* is lower than that observed in *T. gondii* at ~0.66% of the nuclear genome sequence (Table 2.1, Figure 2.2B). Additional coccidian genome sequences were examined and an interesting trend was observed: Genera closest to *Toxoplasma* like *Hammondia and Neospora* had high NUM/PT insertion percentages while more distant genera like *Sarcocystis* and *Eimeria* had significant NUM/PT content, but considerably less than *T. gondii* (Table 2.1).

To determine if NUM/PT accumulation is a common feature among apicomplexan species outside of the Coccidia, we measured the total NUM/PT content for three additional species with available mitochondrial and plastid genome sequences. These include *Plasmodium falciparum, Babesia bovis* and *Theileria parva* (Table 2.1). The results revealed a dramatic discordance with respect to NUM/PT content in these species with only five insertions identified in *P. falciparum* of which four most likely arose from post duplication events, since all five insertions are the same NUMT at the same level of decay with conserved flanking regions. The NUMTs account for 0.002% of nuclear genome in *P. falciparum* and NUPTs account for (0.02%). *B. bovis* and *T. parva* had no recognizable NUMTs and ~0.02% NUPT density (Table 2.1). Our results demonstrate that the propensity for NUM/PT accumulation is a coccidian, rather than an apicomplexan phenomenon.

Phylogenetically conserved NUMTs may have a function

We hypothesize that the density, lifespan and genic proximity of NUM/PTs might foster their occasional functionalization. Because NUMTs predominate, we searched for orthologous NUMT insertions between *T. gondii* and *N. caninum*. Since these species diverged ~28 million years ago (latest estimate of divergence time) (27) and given the observation that most of the NUM/PT insertions appear to be unique, detection of orthologous NUMT insertions might be indicative of functional constraint. In total, we identified five orthologous NUMTs (Table 2.7), with one located 0.5 Kb upstream, one located 0.7 Kb downstream and three located in the intron of nuclear genes.

In order to determine if these orthologous NUMTs show any signs of selective constraint, we extracted these genic sequences from *T. gondii* and *N. caninum*, respectively, and aligned them with VISTA plot. The NUMT residing in the upstream region of the Sm-like protein gene

(Figure 2.6A) and the NUMT located downstream of a hypothetical protein gene (Figure 2.6B), are more conserved than the other orthologous NUMTs on average and thus may be functionally constrained and might, potentially, exert some regulatory effect on the associated gene. The conserved orthologous NUMT upstream of the gene encoding Sm-like protein and a non-orthologus NUMT located in the putative promoter region of a myosin heavy chain gene were selected for a transient transfection assay. Both NUMTs have diverged by more than 30% from the mtDNA. The upstream regions of both of these genes show evidence of active transcription based on (H3K9ac) acetylation and (H3K4Me3) tri-methylation marks (ChIP data in ToxoDB.org). The marks are typically associated with promoter sequences. To assess the effect of removal of these elements on a luciferase reporter gene expression in *T. gondii*, PCR knockout expression constructs were generated (Table 2.8, Appendix 2) for each and compared to expression from their wild-type upstream, putative promoter regions. Deletion of the 70 bp NUMT upstream of the Sm-like protein significantly decreases promoter activity. Deletion of the 66 bp NUMT upstream of the myosin gene dramatically increases promoter activity (Figure 2.7). It is quite possible that the insertion or deletion of any sequence at this site can impact promoter activity because of the altered spacing of existing regulatory regions. When expression data for different strain become available, we will have the opportunity to look at the expression levels of genes whose promoters show a differential presence/absence of NUMTs, to address this question. Thus, these results suggest, but do not prove, that NUMTs can carry, or can evolve into, cis-elements capable of activating or repressing gene expression. Taken together, these two examples provide evidence that some NUMTs may have contributed to the emergence of new cis-regulatory elements in the *Toxoplasma* lineage.

DISCUSSION

Using exhaustive data mining strategies, we have identified the first complete sequence of the *T. gondii* mitochodondrial genome. It consists of 23 mtDNA sequence elements. Full-length cytochrome genes can be assembled using these 23 elements. The topology of the mt genome, however, remains elusive. Using the newly identified mt genome sequence we surveyed the *T. gondii* nuclear genome for NUMTs.

The *T. gondii* genome harbors the highest number of NUM/PTs insertions and content ever reported. This difference ranges from a 7-fold increase when compared to the mustard plant *A. thaliana* (0.256%) and the fungus *Ustilago maydis* (0.286%) to 22 - 40 times more NUMT density as compared to the honeybee (0.086%), and the protist *Phytophthora infestans* (0.046%) (13). While the NUPT in *T. gondii* content is considerably lower than the NUMTs, it is still the highest ever reported. This variation in the NUMT and NUPT content in *T. gondii* could be due to (i) the more conserved nature of the apicoplast genome versus the yet undefined and bizarre nature of the mitochondrial genome; (ii) the mitochondrion may be dispensable in certain life cycle stages and hence some of it's DNA may disintegrate or perhaps even circularize and be available for nuclear integration if it can escape the mitochondrion and avoid degredation; (iii) due to difference in the permeability or integrity of the organelles themselves and/or (iv) differences in selection pressure acting on the two types of organellar integrants. However the inability to assemble full-length cytochrome genes using the NUMTs suggests that these genes are still encoded in the mitochondrial genome.

Similar to the pattern observed in humans, the majority of the NUMP/Ts in *T. gondii* arose from independent insertions rather than post-insertion duplications. Ossorio *et al.,* reported the presence of inverted repeats flanking the REP elements (NUMTs). The inverted repeats are

mtDNA element J, which is part of the *coxIII* gene. We rarely find these inverted repeats flanking the NUMTs and element J does not make up a higher proportion of the NUMTs, suggesting it is unlikely that inverted repeats are part of the mechanism responsible for propagating NUMTs. Although, it is possible they may facilitate integration into the nuclear genome. Non-contiguous fragments from the organellar genome can insert at the same site (28). We find evidence of such insertions in *T. gondii* as well. Repeatmasking the *T. gondii* chromosomes with an assembled mtDNA sequence in addition to the mtDNA sequence elements identifies a higher NUMT content.

We have shown that the NUMT content even within closely related species varies significantly. These data are largely consistent with previous observations where a large variation in NUMT content has been described among insect species like *Drosophila melanogaster*, *Anopheles gambiae*, and *A. mellifera*, and even among mammals like human, mouse and rat (26, 29). This variation can be explained by two major factors: differences in the frequency at which species acquire and retain DNA from the mitochondria and the differential rates of NUMT removal within the nuclear genome (26). The frequency at which mtDNA is transferred to the nucleus can be influenced by a number of things, including the total number of mitochondria within a given cell, and the level of vulnerability to stressful agents that may damage the mitochondria thereby releasing mtDNA. However, given that apicomplexans generally contain only one mitochondrion (and apicoplast) per cell, this cannot sufficiently explain the observable differences in NUMT (and NUPT) accumulation.

Very few to no NUMP/Ts in the apicomplexans, *P. falciparum*, *T. parva* and *B. bovis*, demonstrates that the high number of insertions found in *T. gondii* and other coccidians is not a feature common to all apicomplexans. Molecular and bioinformatic studies performed in yeast,

tobacco and human have shown that integration of organellar DNA fragments occur during illegitimate repair of double-strand breaks (DSB) by non-homologous end joining (NHEJ) (30-32). Interestingly, we do not find homologs for the NHEJ proteins, KU70, KU80, DNA ligase IV-Xrcc4 proteins in *P. falciparum*, *T. parva* or *B. bovis* (Figure 2.8). These organisms predominantly employ homologous recombination (33). NHEJ proteins are present in *T. gondii* and other coccidians. Although the role of these proteins has not been validated in *T. gondii*, it has been experimentally demonstrated that NHEJ is preferentially used in DSB repair at significantly high frequencies (33). These findings suggest that the proficiency of the NHEJ pathway could explain the high NUM/PT content in just the coccidian lineage of the apicomplexans but doesn't explain the variation among the coccidians.

Our quest to identify orthologus NUM/PTs between *T. gondii* and *N. caninum* resulted in only 5 orthologous NUMTs and no orthologous NUPT. Of these, two NUMTs showed a higher level of conservation when compared to the surrounding sequence, suggesting they may have acquired some function leading to sequence conservation. Indeed, functional assays suggest they might act as cis-regulatory elements affecting promoter activity. Myosin heavy chain ATPase is a gene essential to parasite motility, division and penetration of host cells, and thus vital for parasitic virulence (34). The NUMT inserted within this gene dramatically decreases the gene promoter activity. This finding is intriguing, as this may suggest that the insertion of NUMTs could not only influence gene activity but can also affect the pathogenicity of these parasites.

Strain-specific NUMT differences in the three *T. gondii* strains examined suggest NUMTs could contribute to strain diversification and possibly to phenotype. Further characterization of NUMTs in more *T. gondii* strains will provide insights into their rate of turnover and the contribution of NUMTs to the evolution of *T. gondii.*

Acknowledgments

We would like to acknowledge Silvia Moreno, University of Georgia for providing mitochondrial-enriched fractions of *T. gondii*. This work was supported through NIH funding (NIH R01 AI068908) awarded to JCK and EP.

REFERENCES

1. Kim K & Weiss LM (2004) *Toxoplasma gondii*: the model apicomplexan. *International journal for parasitology* 34(3):423-432.
2. Carey KL, Westwood NJ, Mitchison TJ, & Ward GE (2004) A small-molecule approach to studying invasive mechanisms of *Toxoplasma gondii*. *Proceedings of the National Academy of Sciences of the United States of America* 101(19):7433-7438.
3. Belanger F, Derouin F, Grangeot-Keros L, & Meyer L (1999) Incidence and risk factors of toxoplasmosis in a cohort of human immunodeficiency virus-infected patients: 1988-1995. HEMOCO and SEROCO Study Groups. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* 28(3):575-581.
4. Gilbert RE*, et al.* (2001) Effect of prenatal treatment on mother to child transmission of *Toxoplasma gondii*: retrospective cohort study of 554 mother-child pairs in Lyon, France. *Int J Epidemiol* 30(6):1303-1308.
5. Su C*, et al.* (2012) Globally diverse *Toxoplasma gondii* isolates comprise six major clades originating from a small number of distinct ancestral lineages. *Proceedings of the National Academy of Sciences of the United States of America* 109(15):5844-5849.
6. Su C*, et al.* (2003) Recent expansion of *Toxoplasma* through enhanced oral transmission. *Science* 299(5605):414-416.
7. Kleine T, Maier UG, & Leister D (2009) DNA Transfer from Organelles to the Nucleus: The Idiosyncratic Genetics of Endosymbiosis. *Annual Review of Plant Biology* 60:115-138.
8. Leister D (2005) Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends in Genetics* 21(12):655-663.
9. Esser C*, et al.* (2004) A genome phylogeny for mitochondria among alpha-proteobacteria and a predominantly eubacterial ancestry of yeast nuclear genes. *Molecular biology and evolution* 21(9):1643-1660.
10. Boore JL (1999) Animal mitochondrial genomes. *Nucleic Acids Res* 27(8):1767-1780.
11. Martin W*, et al.* (2002) Evolutionary analysis of *Arabidopsis*, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences of the United States of America* 99(19):12246-12251.
12. Palmer JD (2003) The symbiotic birth and spread of plastids: how many times and whodunit? *Journal of Phycology* 39:4-12.
13. Hazkani-Covo E, Zeller RM, & Martin W (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS genetics* 6(2):e1000834.
14. Ricchetti M, Fairhead C, & Dujon B (1999) Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* 402(6757):96-100.
15. Wasmuth J, Daub J, Peregr √ ≠n-Alvarez JM, Finney CAM, & Parkinson J (2009) The origins of apicomplexan sequence innovation. *Genome research* 19(7):1202-1213.
16. DeBarry JD & Kissinger JC (2011) Jumbled genomes: missing Apicomplexan synteny. *Molecular biology and evolution* 28(10):2855-2871.
17. Feschotte C & Pritham EJ (2007) DNA transposons and the evolution of eukaryotic genomes. *Annual review of genetics* 41:331-368.

18. Templeton TJ*, et al.* (2010) A Genome-Sequence Survey for Ascogregarina taiwanensis Supports Evolutionary Affiliation but Metabolic Diversity between a Gregarine and Cryptosporidium. *Molecular biology and evolution* 27(2):235-248.
19. Roos DS*, et al.* (1999) Origin, targeting, and function of the apicomplexan plastid. *Current opinion in microbiology* 2(4):426-432.
20. Hikosaka K*, et al.* (2011) Concatenated mitochondrial DNA of the coccidian parasite Eimeria tenella. *Mitochondrion* 11(2):273-278.
21. Feagin JE, Mericle BL, Werner E, & Morris M (1997) Identification of additional rRNA fragments encoded by the *Plasmodium falciparum* 6 kb element. *Nucleic Acids Res* 25(2):438-446.
22. Feagin JE*, et al.* (2012) The fragmented mitochondrial ribosomal RNAs of *Plasmodium falciparum*. *PloS one* 7(6):e38320.
23. Ossorio PN, Sibley LD, & Boothroyd JC (1991) Mitochondrial-like DNA sequences flanked by direct and inverted repeats in the nuclear genome of *Toxoplasma gondii*. *J Mol Biol* 222(3):525-536.
24. Nishi M, Hu K, Murray JM, & Roos DS (2008) Organellar dynamics during the cell cycle of *Toxoplasma gondii*. *Journal of cell science* 121(Pt 9):1559-1568.
25. Mullapudi N, Joseph SJ, & Kissinger JC (2009) Identification and functional characterization of cis-regulatory elements in the apicomplexan parasite *Toxoplasma gondii*. *Genome biology* 10(4):R34.
26. Richly E & Leister D (2004) NUMTs in sequenced eukaryotic genomes. *Molecular biology and evolution* 21(6):1081-1084.
27. Reid AJ*, et al.* (2012) Comparative genomics of the apicomplexan parasites *Toxoplasma gondii* and *Neospora caninum*: Coccidia differing in host range and transmission strategy. *PLoS pathogens* 8(3):e1002567.
28. Kleine T, Maier UG, & Leister D (2009) DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annual review of plant biology* 60:115-138.
29. Pamilo P, Viljakainen L, & Vihavainen A (2007) Exceptionally high density of NUMTs in the honeybee genome. *Molecular biology and evolution* 24(6):1340-1346.
30. Blanchard JL & Schmidt GW (1996) Mitochondrial DNA migration events in yeast and humans: integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns. *Molecular biology and evolution* 13(6):893.
31. Hazkani-Covo E & Covo S (2008) Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS genetics* 4(10):e1000237.
32. Henze K & Martin W (2001) How do mitochondrial genes get into the nucleus? *Trends Genet* 17(7):383-387.
33. Fox BA, Ristuccia JG, Gigley JP, & Bzik DJ (2009) Efficient gene replacements in *Toxoplasma gondii* strains deficient for nonhomologous end joining. *Eukaryotic cell* 8(4):520-529.
34. Meissner M, Schluter D, & Soldati D (2002) Role of *Toxoplasma gondii* myosin A in powering parasite gliding and host cell invasion. *Science* 298(5594):837-840.
35. Richly E & Leister D (2004) NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. *Molecular biology and evolution* 21(10):1972-1980.

Figure and table legends

Figure 2.1. *T. gondii* mitochondrial protein-coding genes assembled using mtDNA sequence elements

Each cytochrome coding region is represented as a colored box. The sequences of the cytochrome genes were determined using data available in GenBank and via TBLASTN to know apicomplexan cytochrome genes. Based on these sequences, the 23 mtDNA elements were artificially assembled as indicated by the thick black line below each colored box to prove that the coding capacity exists. The elements used to assemble each gene are marked on the thick black line. Numbers above the gene/colored box represent start/stop co-ordinates of the corresponding element on the gene. In every case, the entire element was used. Sanger/EST reads are provided as evidence to suggest that the elements do occur in this order. The elements that comprise each read are indicated. Each Sanger/EST read is indicated as a grey line, with the parts that match the gene represented as a block grey line and the parts that do not match the gene as a dotted grey line. Only a few reads are indicated (many more exist); absence of reads spanning certain regions of a gene does not imply reads for that region are not available. Note, in some cases the reads are in the reverse orientation.GenBank IDs of the EST reads and a Sanger/genomic read that aligns with the corresponding EST read are provided in Table 2.2.

Figure 2.2. Distribution of NUM/PTs along *T. gondii* and *N. caninum* chromosomes

A. The yellow bands in the outer circle represent *T. gondii* chromosomes (1 tick on the yellow band= 100 Kb). **B.** The blue bands represent *N. caninum* chromosomes (1 tick on blue band = 100 Kb). The ticks interior to the chromosomes in A and B denote the location of the different features as described in the key. The figure was generated using Circos version 0.51.

Figure 2.3. Size and decay distribution of NUM/PTs

A. Distribution of NUM/PT length and **B.** Distribution of NUM/PT percent identity to the mt and ptDNA respectively. Values were calculated based on RepeatMasker results. The number of NUM/PTs in each bin is plotted.

Figure 2.4. Age distribution of NUM/PTs in *T. gondii*

The Age of NUM/PTs was calculated as described in Methods (Page 61).

Figure 2.5. Strain-specific NUMT insertion and deletion

A. A schematic of typical NUMT insertion and deletion patterns. The green box represents a NUMT and the black boxes represent 200 bp upstream and downstream flanking nuclear genome sequences. The blue broken line represents a gap. **B-C**. Example of a strain-specific insertion in TGGT1 (B) and strain-specific deletion in TGVEG (C). Multiple sequence alignments of two loci in the three *T. gondii* strains, GT1, VEG and ME49. Green nucleotides indicate NUMT sequence and black nucleotides indicate the flanking nuclear genome sequences.

Figure 2.6. VISTA plot of orthologous NUMTs

Green brackets indicate the position of NUMTs. Genomic features are as indicated in the key. The peaks represent the level of conservation. The grey arrow above each box indicates the orientation of the gene. **A.** The genic sequence of *T. gondii* TGME49_286560 along with its 2 Kb upstream sequence was aligned with its *N. caninum* counterpart. **B.** The genic sequence of *T. gondii* TGME49_260520 along with its 1kb downstream sequence is aligned with its *N. caninum* orthologous counterpart.

Figure 2.7. Experimental evidence for NUMT effects

The structure of WT and mutant promoter regions in which the NUMT is present, or has been removed. **A.** Sm-like protein gene (TGME49_286560) and **B.** Myosin heavy chain gene

(TGME49_254850) (Sequences available in Appendix 2). Nucleotide positions are referenced with respect to the start of translation (+1) of host gene and red line indicates the position of NUMT. **C-D.** Reporter assay results for the promoter of Sm-like protein gene and myosin heavy chain gene respectively. The graphs depict luciferase activity as ratios of firefly:renilla activity in relative luciferase units (RLU) from constructs containing either WT or mutagenized promoter as indicated. All luciferase readings are relative to an internal control (α-tubulin-renilla). Error bars represent standard error calculated across the mean of six independent electroporations.

Figure 2.8. Distribution of NHEJ pathway genes in select apicomplexans

The presence or absence of each gene is indicated as a '+' or '-' respectively. The coccidians are highlighted in blue.

Table 2.1. NUM/PTs in apicomplexan genomes and other eukaryotes

NUM/PTs in apicomplexan genomes were identified as described in methods (Page 59). NUM/PTs data for other eukaryotes were gathered from (7, 13) and (35) .

Table 2.2. Characteristics of the 23 mtDNA elements in *T. gondii* and *N. caninum*

The 23 mtDNA elements were used as a repeat library via RepeatMasker to identify NUMTs. A NUMT can be a fragment of an element. The statisitcs for NUMTs that arose from each mtDNA element are indicated as the number of NUMTs that arose from that element. The total number of base pairs and percentage of the *T. gondii* nuclear genome generated by each NUMT is indicated. The total number of NUMTs at different levels of identity to its mtDNA element is also indicated. If the entire element can be encoded with 100% identity using one or more NUMTs, it is indicated as 'Y' otherwise 'N'. The repeat sequence identified by Ossorio *et al.,* flanking the 'REP' element is highlighted in red.

Table 2.3. Genomic and EST reads representing different observed arrangements of mtDNA elements

Data correspond to the sequence reads shown in Figure 2.1. Reads are indicated as grey lines in the figure. Each row represents a read in the order in which they appear in Figure 2.1. The mtDNA elements are annotated for each read and are indicated in order. Each element in present in its entirety in every read.

Table 2.4. Annotation of the 3,369 bp strain-specific NUMT in ME49

T. gondii ME49 genome sequence was downloaded from ToxoDB.org release 13. The co-ordinates and divergence levels were obtained from RepeatMasker results. Note: some mtDNA elements are present in their entirety in the NUMT. The lengths of the elements are available in Table 2.2.

Table 2.5. Distribution of NUM/PTs in different *T. gondii* features

T. gondii ME49 annotations were downloaded from ToxoDB.org release 13. Intergenic regions include the 1 Kb upstream and downstream flanking regions.

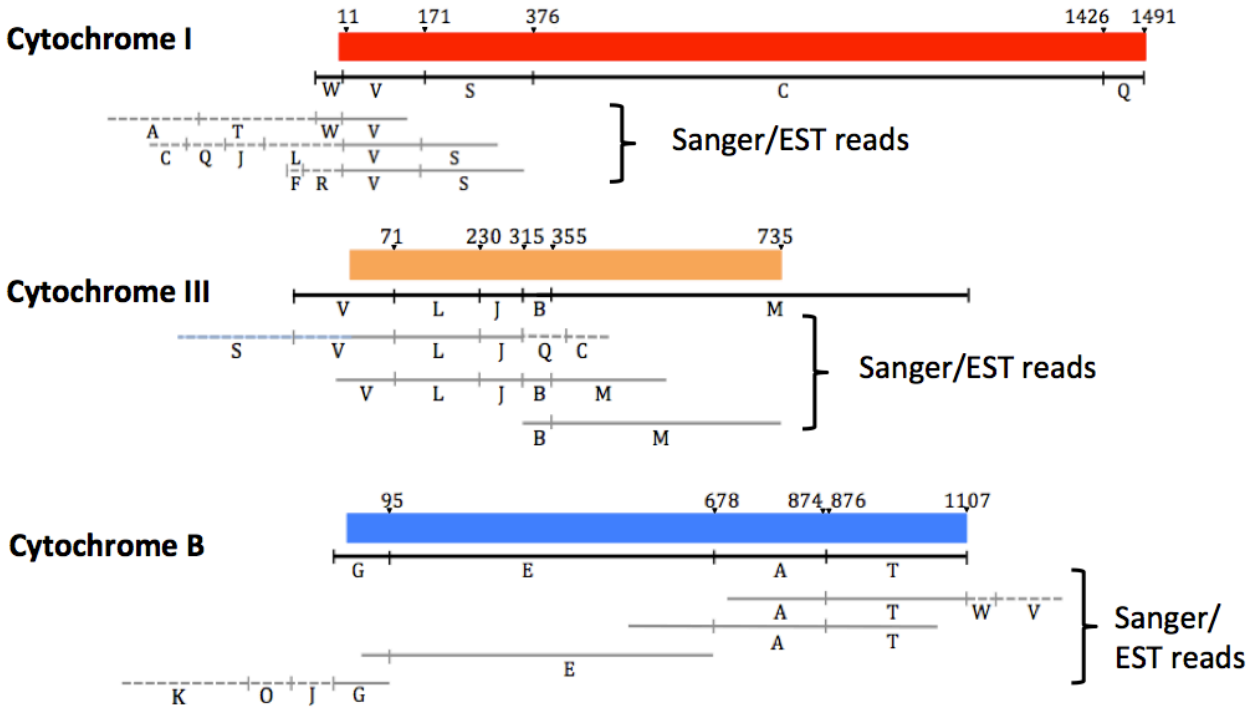
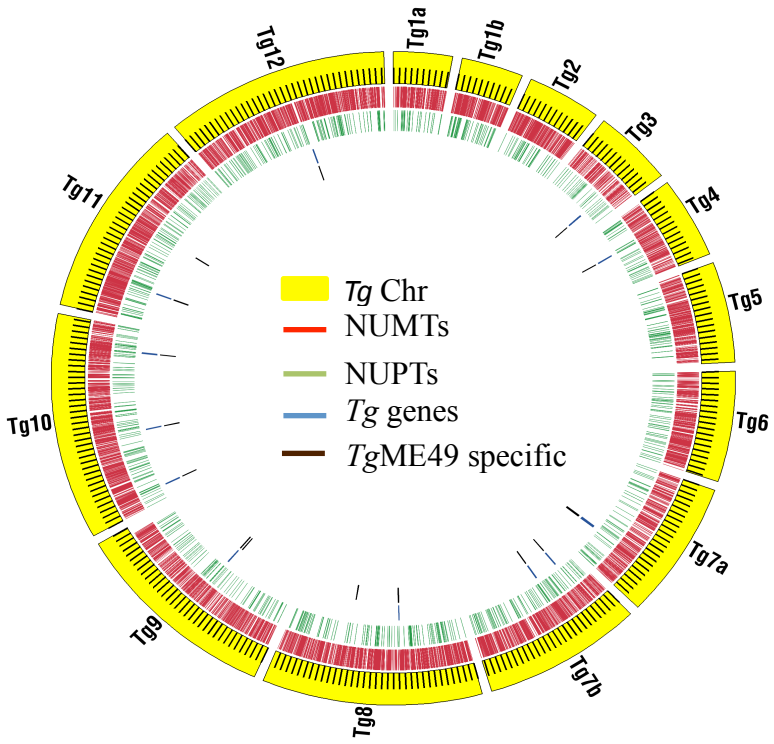


Figure 2.1. *T. gondii* mitochondrial protein-coding genes assembled using mtDNA sequence elements

A.



B.

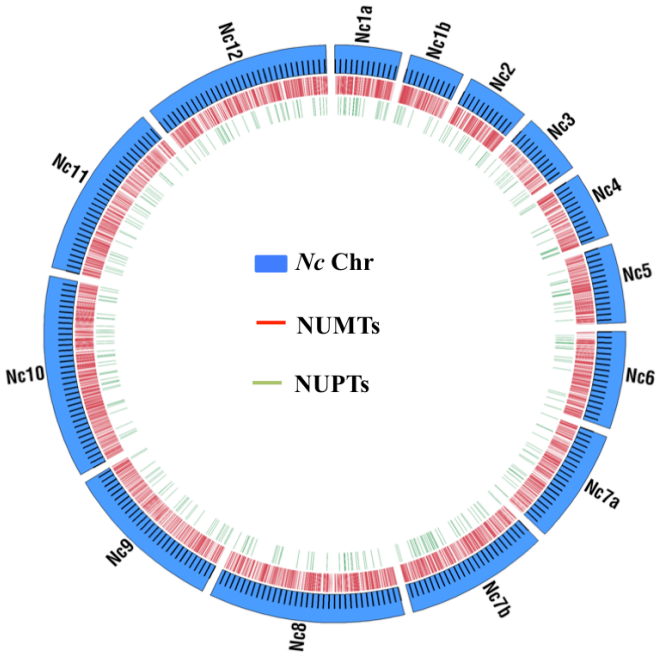
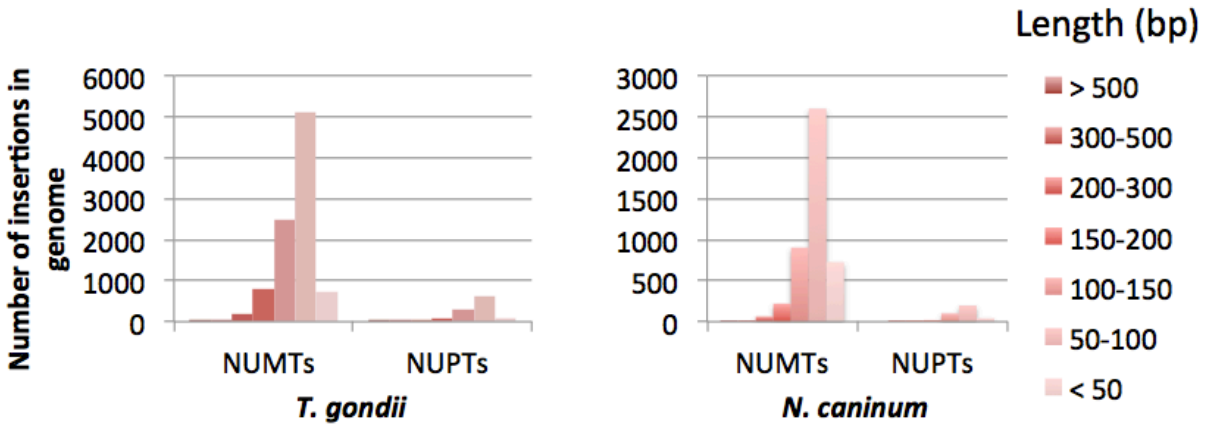


Figure 2.2. Distribution of NUM/PTs along *T. gondii* and *N. caninum* chromosomes

A.



B.

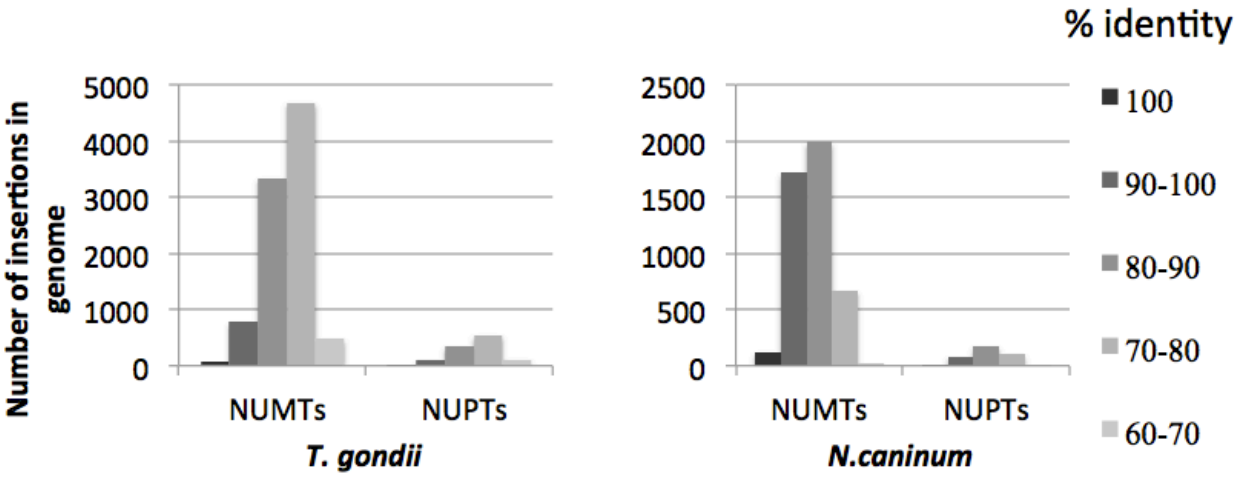


Figure 2.3. Size and decay distribution of NUM/PTs

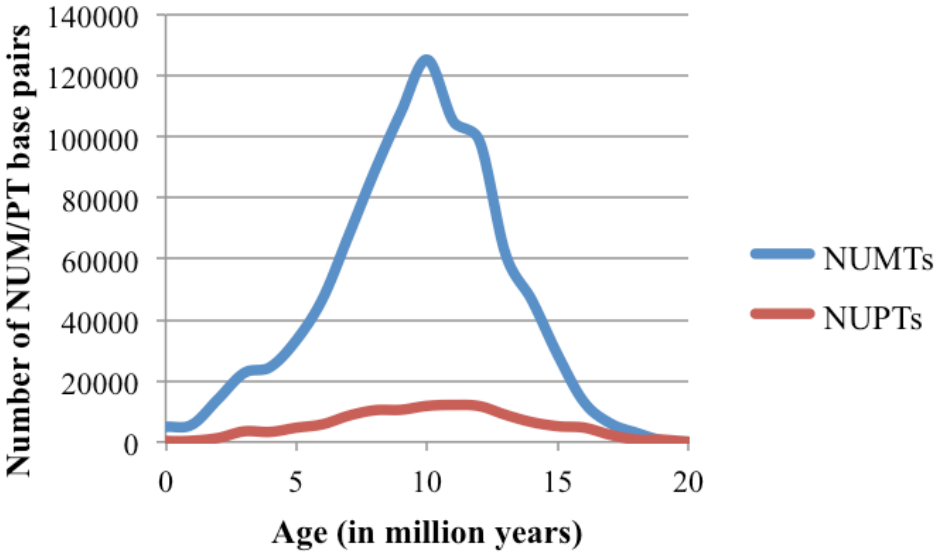
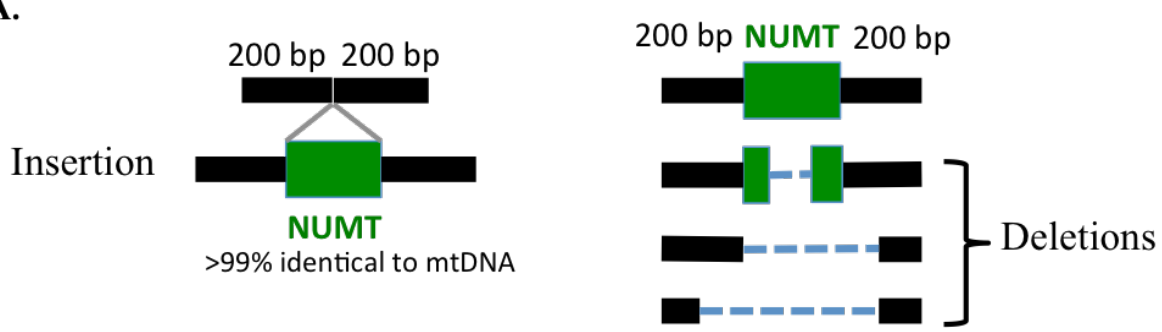


Figure 2.4. Age distribution of NUM/PTs in *T. gondii*

A.



B.

```
tgonGT1  699205 TGAAGCGAAGAAGAGAAATCGAAAAAAAAAACGCAAGAGAGTGTAGACATTTAGCATCTGTCGTTAACATATGAGGATAAA
tgonME49 706836 TGAAGCGAAGAAGAGAAATCGAAAAAAAA-CGCGAGAGAGT-------------------------------------
tgonVEG  719670 TGAAGCGAAGAAGAGAAATCGAAAAAAAA-CGCGAGAGAGT-------------------------------------

tgonGT1  699285 AGGCAACTTTAAGCGCGGTATCAATACCTGCAGGATTGCTAGAACCATTTAAATGTAAATAGAGAGAGTGTGCACGCCCG
tgonME49 706876 ------------------------------------------------------------GAGAGAGTGTGCACGCCCG
tgonVEG  719710 ------------------------------------------------------------GAGAGAGTGTGCACGCCCG

tgonGT1  699365 CTGTTGCTGCGTCTTTCTTC
tgonME49 706895 CTGTTGCTGCGTCTTTCTTC
tgonVEG  719729 CTGTTGCTGCGTCTTTCTTC
```

C.

```
tgonGT1  1045139 TGTGCGACAGGTCCGTCAGCACATGCCTCGTCAAAAGAAAGGGGATTTCAGTATCCTACGGCACTCAGATTCTTCCACAT
tgonME49 1102932 TGTGCGACAGGTCCGTCAGCACATGCCTCGTCAAAAGAAAGGGGATTTCAGTATCCTACGGCACTCAGATTCTTCCACAT
tgonVEG  1113047 TGTGCGACAGGTCCGTCAGCACATGCCTCGTCAAAAGAAAGGGGATTTCAGTATC----------------------

tgonGT1  1045219 CGGATTTGTTCTCGCGCAATACCTTGACTACTGTTATCATTCGCACTAACCACGGCAACCTTCCCCCTGTCAGATCTGAG
tgonME49 1103012 CGGATTTGTTCTCGCGCAATACCTTGACTACTGTTATCATTCGCACTAACCACGGCAACCTTCCCCCTGTCAGATCTGAG
tgonVEG  1113102 -----------TCGCGCAATACCTTGACTACTGTTATCATTCGCACTAACCACGGCAACCTT-CCCCTGTCAGATCTGAG

tgonGT1  1045299 AGGCCACACGATTC
tgonME49 1103092 AGGCCACACGATTC
tgonVEG  1113170 AGGCCACACGATTC
```

Figure 2.5. Strain-specific NUMT insertion and deletion

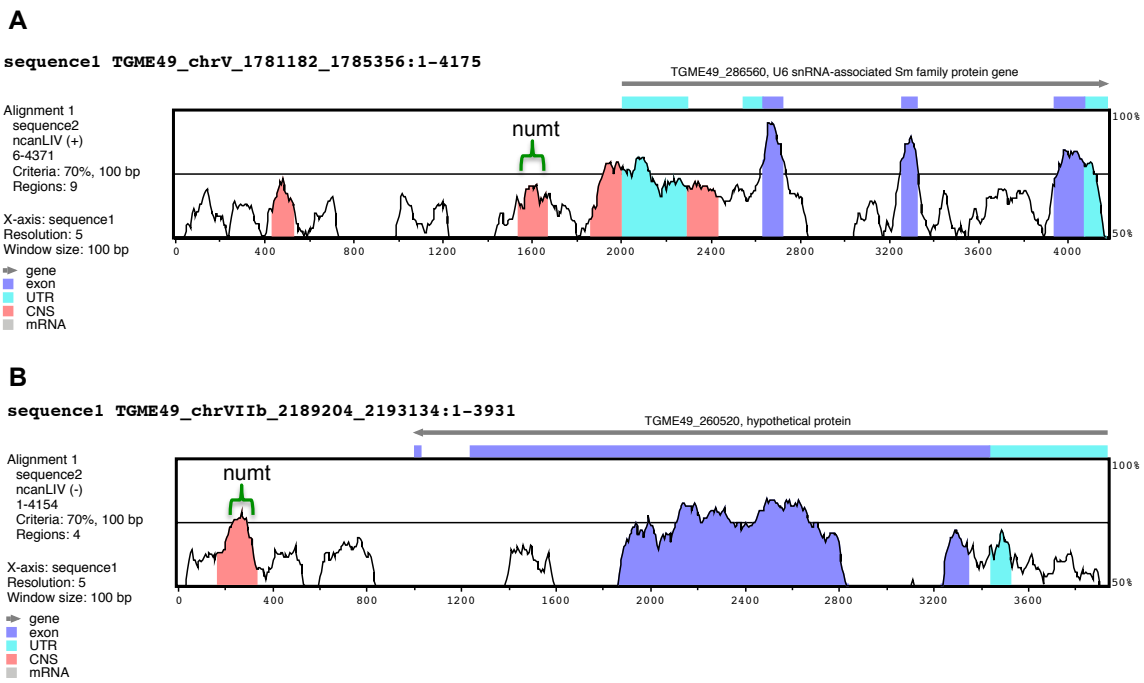


Figure 2.6. VISTA plot of orthologous NUMTs

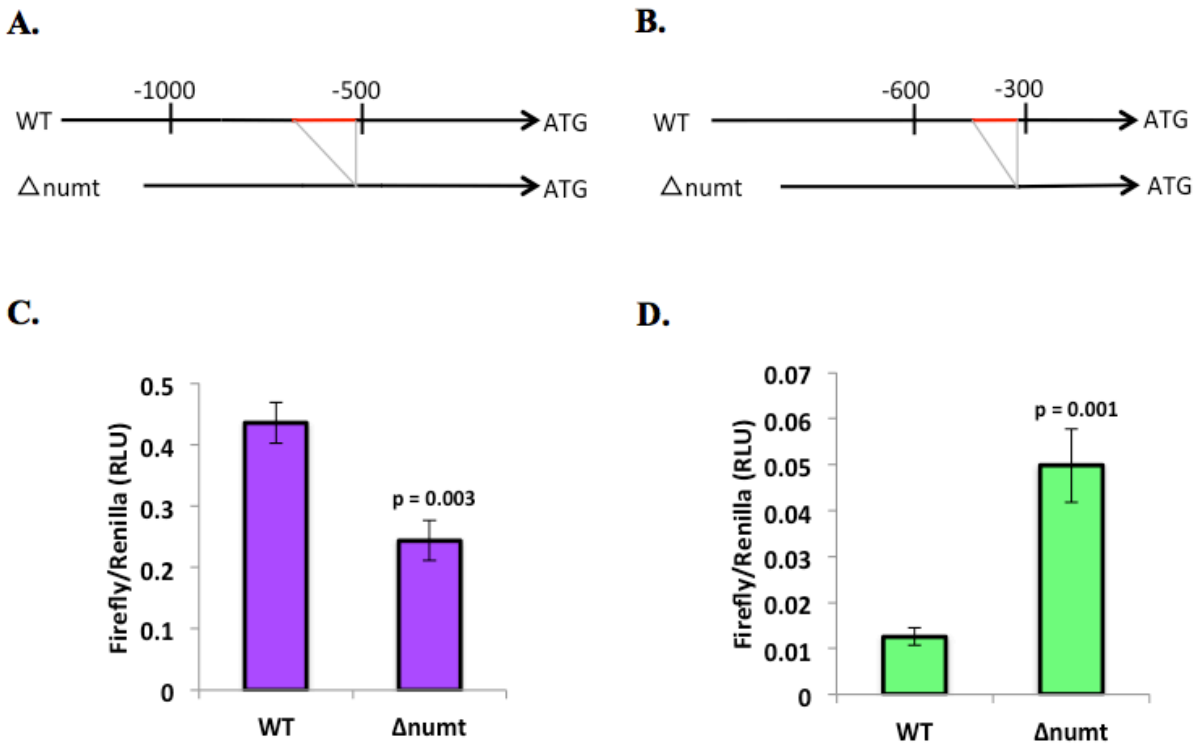


Figure 2.7. Experimental evidence for NUMT effects

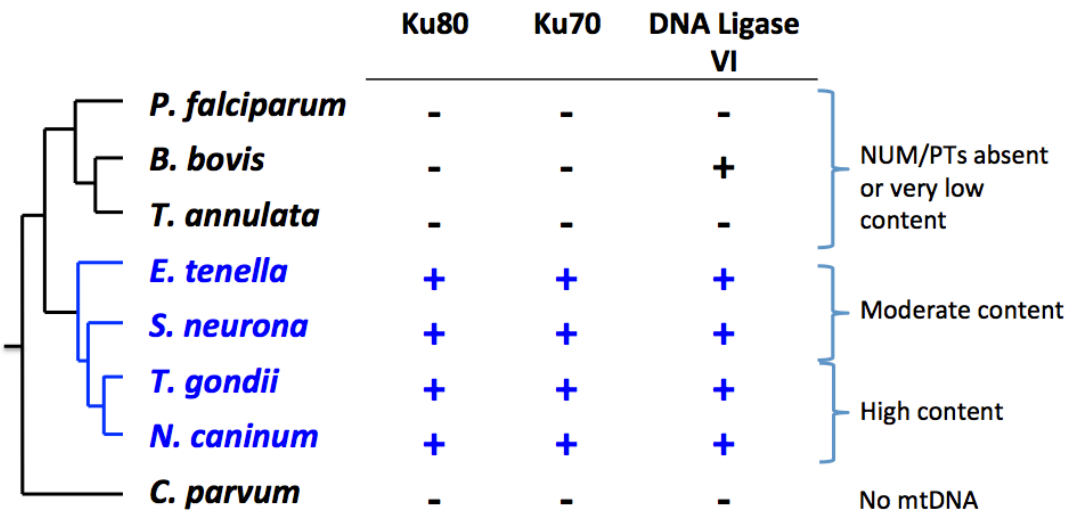


Figure 2.8. Distribution of NHEJ pathway genes in select apicomplexans

Table 2.1. NUM/PTs in apicomplexan genomes and other eukaryotes

Apicomplexan Species	NUMTs			NUPTs		
	Density(%)	Number of insertions	Number of base pairs	Density(%)	Number of insertions	Number of base pairs
T. gondii ME49	1.430	9,356	891,057	0.180	1,108	112,443
H. hammondi H.H.34	1.480	10,643	999,296	0.190	1,206	117,917
N. caninum NCLIV	0.660	4,534	380,139	0.060	373	34,236
S. neurona SN*3*	0.030	454	38,353	0.020	54	27,624
E. tenella Houghton	0.040	227	22,024	0.088	338	45,513
B. bovis	0.000	0	0	0.020	11	1,697
T. parva	0.000	0	0	0.010	7	1,036
P. falciparum	0.002	5	515	0.020	30	5,408
Other Eukaryotes						
A. thaliana	0.2564	820	305,602	0.03	301	50
O. sativa	0.1768	5,357	823,923	0.15	2090	782
H. sapiens	0.0087	871	263,478			
S. cerevisiae	0.0081	18	983			
D. melanogaster	0.0057	50	10,331			
M. musculus	0.0015	137	37,670			
C. elegans	0.0001	1	126			
Monodelphis domestica	0.0698	1859	2093,63			
A. mellifera	0.086	1,790	172,131			

Table 2.2. Characteristics of the 23 mtDNA elements in *T. gondii* and *N. caninum*

Element	*T. gondii* ME49								*N. caninum* LIV								Entire element found at 100% identity in nuclear genome?	
	Length of element	# of NUMTs	# of bp	% of genome	Number of insertions based on % identity to mtDNA[a]				Length of element	# of NUMTs	# of bp	% of genome	Number of insertions based on identity to mtDNA[a]					
					100	>=90	>=80	>=70					100	>=90	>=80	>=70	*Tg*	*Nc*
A	196	298	24284	0.04	1	42	158	293	196	161	12130	0.02	6	59	71	25	N	N
B	40	31	1213	0.002	4	15	31	31	37[b]	33	1215	0.002	4	21	8	0	Y	Y
C	1049	245	27866	0.04	2	26	84	222	1049	177	15239	0.03	0	51	94	29	N	N
D	82	181	11522	0.02	4	25	115	178	82	78	4686	0.01	3	45	27	3	N	N
E	633	711	74381	0.12	3	53	294	675	583	356	32391	0.06	15	134	147	57	N	N
F	179	308	25710	0.04	2	35	166	297	179	182	14144	0.02	4	71	88	19	Y	N
G	100	121	8651	0.01	5	19	78	116	100	60	3922	0.01	4	24	21	10	Y	Y
H	447	291	28642	0.05	1	36	123	274	447	162	13357	0.02	4	54	83	21	N	N
I	204	159	13361	0.02	1	21	92	151	204	71	5266	0.01	0	36	25	10	N	N
J	85	626	42357	0.07	5	68	386	624	85	303	18266	0.03	19	149	115	20	Y	Y
K	445	697	64267	0.1	6	69	334	669	445	292	24203	0.04	12	126	110	40	N	N
L	159	611	53087	0.08	3	45	289	589	158[b]	276	21213	0.04	13	109	109	44	Y	Y
M	754	891	86307	0.14	9	87	394	837	752[b]	404	37010	0.06	11	151	173	66	N	N
N	166	173	15509	0.02	0	14	78	170	166	123	9620	0.02	3	39	61	20	N	N
O	86	234	15933	0.03	3	33	137	234	86	130	8012	0.01	6	44	72	8	Y	Y
P	184	301	26267	0.04	0	43	159	281	184	146	11225	0.02	6	52	68	20	N	N
Q	65	62	3480	0.01	2	15	47	62	65	65	3468	0.01	2	36	24	3	Y	Y
R	85	161	10588	0.02	3	18	110	160	85	61	3786	0.01	1	31	24	5	Y	N

S	205	530	49216	0.08	4	56	234	496	205	293	22682	0.04	9	113	133	35	Y	Y
T	233	223	21821	0.03	2	23	99	211	229[b]	125	10436	0.02	4	46	64	9	N	N
U	353	610	56469	0.09	9	70	292	581	353	306	25204	0.04	6	140	123	37	N	N
V	161	961	83495	0.13	7	87	488	943	161	461	34894	0.06	16	188	199	58	Y	N
W	45	39	1636	0.002	3	13	39	39	40[b]	33	1376	0.002	3	18	12	0	Y	Y

[a] - Insertion may not be of the entire element.

[b] - the first or last X number of base pairs did not match with corresponding *T. gondii* element.

Table 2.3. Genomic and EST reads representing different observed arrangements of mtDNA elements

Gene	Elements in EST read	GenBank Accession ID of EST	Trace Archive ID of a genomic read aligning with the EST read
COXI	A, T, W, V	CN617107.1	gnl|ti|2057042422
	C, Q, J, L, V, S	DK934897.1	gnl|ti|2056951560
	V, S, R, F	CV701032.1	gnl|ti|2057358855
COXIII	C, Q, J, L, V, S	DK934897.1	gnl|ti|2056951560
	V, L, J, B, M	CV654565.1	gnl|ti|2057374420
	B, M	DV110375.1	gnl|ti|2064971134
COB	A, T, W, V	CN617107.1	gnl|ti|2057042422
	A, T, E	CV654565.1	gnl|ti|2057374420
	G, E	CB384005.1	gnl|ti|2057332917
	K, O, J, G	CB752279.1	gnl|ti|2057396732

Table 2.4. Annotation of the 3,369 bp strain-specific NUMT in ME49

Chromosome	Start	End	Divergence from mtDNA(%)	mtDNA gene/element	Start	End
TGME49_chrX	5732558	5732660	0	U	1	103
TGME49_chrX	5732662	5732907	1.6	coxIII	216	459
TGME49_chrX	5732891	5732976	0	O	1	86
TGME49_chrX	5732977	5733228	0	K	1	252
TGME49_chrX	5733227	5733447	0	cob	1	221
TGME49_chrX	5733451	5733767	0.3	coxIII	1	317
TGME49_chrX	5733697	5734062	0	coxI	11	376
TGME49_chrX	5734077	5734453	0.5	cob	20	396
TGME49_chrX	5734448	5734516	0	H	1	69
TGME49_chrX	5734467	5734581	0	coxI	1377	1491
TGME49_chrX	5734580	5734683	4.9	coxIII	216	317
TGME49_chrX	5734582	5734666	0	J	1	85
TGME49_chrX	5734667	5734752	0	O	1	86
TGME49_chrX	5734753	5734931	0	F	1	179
TGME49_chrX	5734932	5735016	0	R	1	85
TGME49_chrX	5735033	5735098	0	U	1	66
TGME49_chrX	5735100	5735284	0	coxIII	275	459
TGME49_chrX	5735285	5735354	0	M	457	526
TGME49_chrX	5735351	5735396	0	D	1	46
TGME49_chrX	5735397	5735508	0	K	334	445
TGME49_chrX	5735509	5735860	0	U	1	353
TGME49_chrX	5735862	5735926	0	coxIII	395	459

Table 2.5. Distribution of NUM/PTs in different *T. gondii* features

NUMT	Number of insertions	Number of bp	% of NUM/PT in each genomic feature	% of each feature in total genome
CDS	30	2,151	0.24	31.69
Intron	4,784	458,133	50.63	31.76
UTR	1,053	97,854	10.81	14.86
Within 1 kb upstream of gene	1,151	109,124	12.06	21.31
Within 1 kb downstream of gene	1,155	105,692	11.68	21.31
Intergenic	1,596	343,207	12.65	21.64
Overall*	9,356	891,057	1.43	100
NUPT				
CDS	4	430	0.38	31.69
Intron	724	74,995	66.70	31.76
UTR	95	8,824	7.85	14.86
Within 1 kb upstream of gene	80	7,168	6.37	21.31
Within 1 kb downstream of gene	103	9,875	8.78	21.31
Intergenic	86	10,448	9.29	21.64
Overall*	1,107	112,443	0.18	100

*NUMTs spanning more than one genomic feature is included only in 'Overall'

Table 2.6. Strain-specific NUMTs

Coordinates of NUMTs	Divergence (%)	Associated nuclear gene	Location in genes	Only present/ absent in	Insertion/ deletion
TGGT1_chrII:244462..244554	4.4	TGGT1_221330	intron	GT1	insertion
TGGT1_chrII:1102510..1102622	27.4	TGGT1_222410	1.8 Kb upstream	ME49	deletion
TGME49_chrIII:1730617..1730859	23.5	TGME49_254420	intron	ME49	not sure
TGME49_chrIV:1102974..1103065	20.6	TGME49_318770	1.3 Kb upstream	VEG	deletion
TGME49_chrIX:2923871..2923932	3.2	none	none	ME49	insertion
TGME49_chrIX:3047280..3047543	12.5	TGME49_289180	intron	VEG	deletion
TGGT1_chrVI:699247..699347	1	TGGT1_239480	intron	GT1	insertion
TGGT1_chrVIII:3463130..3463400	24.5	TGGT1_273478	intron	GT1	not sure
TGME49_chrVIII:4435556..4435633	20.8	none	none	ME49	not sure
TGGT1_chrVIIa:2687959..2688077	7.6	TGGT1_203135	intron	VEG	not sure
TGME49_chrVIIa:3250536..3250697	30.6	TGME49_202540	1.3 Kb upstream	VEG	deletion
TGGT1_chrVIIb:1000247..1000321	14.7	TGGT1_262780	173 bp downstream	ME49	deletion
TGGT1_chrVIIb:3183923..3184103	23.2	none	none	GT1	not sure
TGGT1_chrVIIb: 4560651..4560737	21.2	TGGT1_255890	intron	GT1	not sure
TGME49_chrX:475518..475568	23.5	TGME49_228230	3′ UTR	ME49	not sure
TGGT1_chrX:2610785..2610846	17.7	TGGT1_224900	591 bp upstream	ME49	deletion
TGGT1_chrX:2885685..2885842	24.1	TGGT1_224540B	intron	ME49	deletion
TGME49_chrX:3011523..3011876	16.9	TGME49_224510	257 bp downstream	VEG	deletion
TGME49_chrX:6399715..6399810	21.3	TGME49_014600	intron	VEG	deletion
TGGT1_chrX:6755042..6755100	25.4	TGGT1_215260	intron	GT1	not sure
TGGT1_chrXI:4963753..4963855	25.2	none	none	GT1	not sure
TGGT1_chrXII:287472..287613	22.7	TGGT1_299820	1.1 Kb upstream	ME49	not sure

TGGT1_chrXII:2346630..2346730	21.7	none	none	GT1	not sure
TGGT1_chrXII:2367574..2367660	25.3	TGGT1_245610	intron	GT1	not sure
TGME49_chrVIII:2496663..2496886	19.1	TGME49_233200	1.1 Kb upstream	GT1	deletion
TGME49_chrVIIa:3199077..3199253	22.5	TGME49_202580	intron	VEG	deletion
TGME49_chrVIIb:701168..701236	24.6	TGME49_263220	intron	ME49	not sure
TGME49_chrVIIb:1710367..1710510	25	TGME49_261490	intron	ME49	not sure
TGME49_chrXI:1310258..1310322	24.6	TGME49_310380	intron	GT1	deletion
TGME49_chrXI:3550511..3550993	19.4	TGME49_313620	400 bp upstream	GT1	deletion
TGME49_chrXII:3864844..3864916	0	TGME49_248450	intron	GT1	not sure

Table 2.7. Orthologous NUMTs in *T. gondii* and *N. caninum*

Coordinates of NUMT	**Associated nuclear gene**	**Relationship with associated gene**
TGME49_chrV: 1782800-1782918	TGME49_286560	0.9 kb of upstream
TGME49_chrVIIb: 2189430-2189499	TGME49_260520	0.7kb of downstream
TGME49_chrVIII: 1529286-1529361	TGME49_231770	intron
TGME49_chrIX: 3420407-3420485	TGME49_289740	intron
TGME49_chrX: 3664739-3664832	TGME49_223672	UTR

Table 2.8. Primer sequences used in promotor assays

Usage	**Sequence (5'-3')**
U6 snRNA-associated Sm family protein gene WT promoter (reverse)	GGGGACCACTTTGTACAAGAAAGCTGGGTCCATTATGTCTAACTCCAGGAGGT
U6 snRNA-associated Sm family protein gene WT promoter (forward)	GGGGACAAGTTTGTACAAAAAAGCAGGCTAATGTTGCCTACGGTTCTGACTA
delete NUMT from U6 snRNA-associated Sm family protein gene (reverse)	ATTATGTACCTCCTCTATTACCGCAGAGCATTAGCAGTATCAAATG
delete NUMT from U6 snRNA-associated Sm family protein gene (forward)	GGCAGTTGCATTCCAACATTTGATACTGCTAATGCTCTGCGGTAATAGA
myosin heavy chain gene WT promoter (reverse)	GGGGACCACTTTGTACAAGAAAGCTGGGTCCATTGTTTTCGAGCAGAGACATT
myosin heavy chain gene WT promoter (forward)	GGGGACAAGTTTGTACAAAAAAGCAGGCTAACACCAGTGCGGGAGGCGTTCTAG
delete NUMT from myosin heavy chain gene (reverse)	ACTAGGAATCCTCCGATACACCCAAATATAATTGAACACAGAGAA
delete NUMT from myosin heavy chain gene (forward)	TCCAGTTGAACTGCTCCTCTTTCTCTGTGTTCAATTATATTTGGGTGTA

CHAPTER 3

NUCLEAR SEQUENCES OF MITOCHONDRIAL ORIGIN GENERATE STRAIN-SPECIFIC DIFFERENCES IN THE GENOME OF *TOXOPLASMA GONDII*

Namasivayam, S. and Kissinger, J. C. To be submitted to PLoS Pathogens.

ABSTRACT

Nuclear sequences of mitochondrial origin (NUMTs) are found in most eukaryotes. They typically make up an insignificant proportion (< 0.1%) of the nuclear genome. NUMT differences do occur between closely related species like humans and chimpanzees. In humans, insertion of NUMTs has been implicated in a few diseases. We previously reported an unusually high NUMT content in *Toxoplasma gondii*, a very successful zoonotic pathogen. Nearly 1 Mb of its 65 Mb nuclear genome is made up of NUMTs and evidence suggests this acquisition of NUMTs was rapid and is an ongoing process. We examined NUMTs in 16 *T. gondii* strains to assess patterns of differential presence and absence. We identified 57 NUMTs that show differential presence across the strains. Comparisons to the population structure of the strains revealed that patterns of NUMTs gain and loss do not necessarily correlate with the recombination patterns observed in these strains. We inferred the age of each NUMT based on their divergence from the mitochondrial sequence and found that the closely-related avirulent parasite, *Hammondia hammondi* shows a similar NUMT age profile. However, *Neospora caninum*, a genus that diverged from a shared lineage ~28 Mya shows a younger age profile, indicating that factors governing its NUMT acquisition and deletion are different. The observation that NUMTs show differences across *T. gondii* strains that are believed to have diverged ~1 Mya, and significant differences in their profile in comparison to a closed related species, indicates that NUMTs could act as important drivers of genome evolution in this parasite, and may be a source of innovation that affects virulence.

INTRODUCTION

Toxoplasma gondii is present in a third of the human population and is one of the most successful zoonotic parasites. It belongs to the lineage of tissue-cyst forming coccidians in the phylum Apicomplexa (1). Most species of this phylum have a two-host lifecycle, a sexual phase in the definitive host and an asexual phase in an intermediate host. Parasites of the coccidan genus *Eimeria* can complete their life cycle in a single host and are restricted to the gut, whereas coccidians of the Sarcocystidae sub-group, *T. gondii*, *N. caninum, H. hammondi* and *S. neurona,* have a two-host lifecycle and can infect other tissues (2). *Toxoplasma.gondii* has excelled as a parasite for a number of reasons. It has a wide host range, essentially capable of infecting any warm-blooded animal (3). The sexual life cycle takes place only in felines. However, the parasite has evolved the ability to bypass the sexual phase. Consumption of tissue cysts and oocysts by intermediate hosts can lead to transmission (4, 5). This oral transmission capabilty is credited for the recent expansion of *T. gondii* (6).

Toxoplasma gondii is the only species in its genus and has an unusual clonal population structure. It was initially thought to consist primarily of three clonal genotypes I, II and III in North America (NA) and Europe (7, 8). The SNPs in these lineages are biallelic and indicative of relationships via a single, ancestral, cross (9). Later, sequencing of strains from South America (SA) revealed a more diverse population structure. Clonal haplotypes and bialleles in SA were different from the ones in NA (10). Notably, the chromosome Ia haplotype is monomorphic in NA strains and many SA strains and this near fixation of chromosome Ia is proposed to drive clonal expansion (10, 11). Based on RFLP and multi-locus genotyping, these strains were subsequently grouped into 12 clades that showed strong geographic segregation (12). It is predicted that the NA and SA lineages shared common ancestry a million years ago (10).

Recently, 62 strains from around the world were sequenced as part of a community led project to better understand the genetic variation of *T. gondii* (13). While a neighbor network using 802,764 SNP positions from all the strains supported the major clades previously defined, it revealed a higher than expected degree of gene flow among and between the strains (Figure 1.8A). There is clear evidence of local admixture with strains from different clades showing regions or haploblocks of shared ancestry to such an extent that the clades are no longer as clearly defined (Figure 1.8B). This finding suggested that, although sexual recombination is infrequent in the wild, it is possible that the maintenance of haploblocks can drive local adaptations (13). Indeed, the strains exhibit differences in pathogenesis and virulence. Type I strains and a number of SA strains are highly virulent while Type II and III strains show low virulence (7).

We previously reported that NUclear sequences of MiTochondrial origin (NUMTs) showed strain-specific difference in three *T. gondii* strains, ME49, GT1 and VEG. Polymorphisms in NUMTs have been reported in humans (14, 15). 12/40 NUMTs are variable (15) and 5 cases have been associated with disease (16). 80% of the NUMTs are conserved between human and chimpanzees, with a calcuated insertion rate of 6.7 and 11. 3 NUMTs/million years (my) in humans and chimpanzees, respectively (17). Insertion rates are much lower in *Drosophila* with an average of 1.26 insertions/my and the rate varies among the species. Differences in NUMT and NUPT (NUclear sequences of PlasTid origin) dynamics and turnover have also been reported in plants (18).

The NUMT (and NUPT) content in *T. gondii* is at least two orders of magnitude higher than in any species previously reported (Chapter 2). These insertions have the potential to act as important evolutionary forces in this parasite. In this study, we perform comparative analyses of

NUMTs among the coccidian species *N. caninum*, *H. hammondi* and *T. gondii* to better understand the NUMT profiles and to determine the rate of NUMT turnover. Although *T. gondii* and *N. caninum* are closely related species (~28 Mya) (19), *T. gondii* has twice as many NUMTs as *N. caninum* and only 5 ortholgous NUMTs were identified (Chapter 2). We extend our analyses of strain-specific NUMTs to 13 additional *T. gondii* strains to evaluate the role of NUMTs in strain diversification and genome evolution. Use of NUMTs as a phylogenetic tool has been suggested (17). However, with infrequent sexual recombination and local admixture seen in *T. gondii* strains, NUMTs are unlikely to serve that purpose in *T. gondii*. Finally, we provide experimental evidence for the high level of observed NUM/PT insertions in the *T. gondii* nuclear genome.

MATERIALS AND METHODS

Identification of NUMTs in *T. gondii* strains and *H. hammondi*

The *de novo* assembled genomes of 13 *T. gondii* strains were downloaded from ToxoDB.org (http://toxodb.org/) release 9: ARI, CAST, TgCtBr5, TgCtBr9, TgCtPRC2, COUG, TgCtCo5, FOU, GAB2-2007-GAL-DOM2, MAS, p89, RUB and VAND. These genome sequences are in scaffolds and not assembled into chromosomes. However, they are fairly complete in comparison to each other, ranging 61-65 Mb in size. Only genome sequences for *T. gondii* ME49, GT1 and VEG, previously analyzed for NUMTs (Chapter 2), are assembled into chromosomes and annotated. These three genomes were downloaded from ToxoDB release 12 to reflect the latest assemblies available at the time of preparation of this manuscript for a total of 16 genome sequences.

NUMTs were identified in the 13 new *T. gondii* strains using RepeastMasker (ver 4.0.5) (http://repeatmasker.org/) as described previously (Chapter 2). Briefly, the 23 mitochondrial DNA sequence elements and assembled cytochrome genes *coxI*, *coxIII* and *cob* from *T. gondii* were used as a custom library in RepeatMasker to identify the NUMTs present in each strain (Table 3.1). The search engine 'crossmatch' was used. In the case of the ME49, GT1 and VEG, only the chromosomes were used in the identification of NUMTs since the non-chromosomal contigs contain sequences from the mitochondrion itself. For the remaining 13 strains, all available contigs and scaffolds were used for RepeatMasking. Therefore the NUMT content reported from these strains is likely an overestimate since sequences from the mitochondrial genome itself will be counted. However, this overestimation does not affect the analyses reported here since orthologous NUMTs are identified based on location (see below).

The *de novo* assembled *H. hammondi* contigs were downloaded from ToxoDB.org release 24. Using the mtDNA sequences from *T. gondii*, we attempted to identify the 23 mtDNA elements and cytochrome genes. We were not able to identify full-length copies of all the elements or the cytochrome genes from the *H. hammondi* genome sequence. Three elements, K, M and U were not found. However, we were able to identify full-length copies of *coxI* and *cob* from other *Hammondia* species (20) (NCBI JX473251.1, JX473247.1). The *T. gondii* and *H. heydorni coxI* and *cob* genes were 92% and 95% identical at the nucleotide level and 96% and 99% identical at the amino acid level to each other, respectively. Since the identified mtDNA sequences in *Hammondia* are 92%-100% identical at the nucleotide level, *T. gondii* mtDNA sequences were used to identify NUMTs in *H. hammondi* (Table 3.1). It should be noted that the percent divergence between the mtDNA and the NUMT sequence in *H. hammondia* could be

slightly lower than reported in some cases since the *T. gondii* mtDNA was used to identify NUMTs in *H. hammondi.*

Calculation of mtDNA mutation rate

To determine a more accurate estimate of the age of the NUMTs, we calculated the mtDNA mutation rate using the previously identified mtDNA sequences from *N. canium* and *T. gondii.* Two mutation rates were calculated. One using only the three cytochrome gene sequences and another using all 23 mtDNA elements (Table 3.2). The sequences were concatenated into a single sequence for each organism and aligned using MUSCLE in MEGA 6 (21). Distance estimates were calculated using MEGA 6. Since NUMTs arise from all parts of the mitochondrial genome and the distance calculated using the 23 mtDNA sequences encompassed the entire mitochondrial genome, the distance calculated using all the elements was used to calculate a mutation rate. The mutation rate was estimated using

$$\mu = K/2T$$

where μ is the mutation rate or rate of nucleotide substitution, K is the genetic distance between the two sequences and T is the time of divergence between *N. caninum* and *T. gondii*. The mutation rate was calculated using two previous divergence time estimates: 12.7 Mya (6) and also a more recent estimate of 28 Mya (19).

***T. gondii* and *H. hammondi* divergence time estimate**

The intron mutation rate between *T. gondii* and *N. caninum* was calculated using the introns from the *ACT1*, *ATUB* and *MIC2* genes (6). We calculated the average genetic distance between *T. gondii* and *H. hammondi* using the introns of these genes (Gene IDs from ToxoDB.org release 24: *ACT1*-TGME49_209030/HHA_209030, *ATUB*-TGME49_316400/HHA_316400, *MIC2*-TGME49_201780/HHA_201780). The introns of each

gene were combined into a single sequence and aligned using CLUSTALW via MEGA6. The genetic distance for each set of introns was obtained using the Jukes-Cantor Model in MEGA 6 and averaged. The divergence time was estimated by substituting the intron mutation rate (μ) and the genetic distance (K) in the formula T=K/2μ.

Estimation of NUMT age

The relative date of insertion of a NUMT can be inferred from the NUMTs age. RepeatMasker reports a percent divergence between the NUMT and the corresponding mtDNA sequence. The percent divergences were converted into genetic distances using the Jukes-Cantor Model formula

$$K = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right)$$

where K is the genetic distance and p is the percent divergence between the mtDNA and NUMT. Percent divergence was calculated by dividing the number of substitutions between the two sequences by the sequence length. Using the genetic distance the age of each NUMT was calculated using

$$T = \frac{K}{\mu_{nu}+\mu_{mt}}$$

where T is the age of the NUMT, K is the genetic distance, μ_{nu} is the intron mutation rate and μ_{mt} is the mtDNA mutation rate. The intron mutation rate was used to calculate the age of the NUMTs since NUMTs are not expected to be under selective pressure and should have a neutral mutation rate similar to the introns. The intron mutation rate was previously calculated as 2.12×10^{-8} (6). Using the more recent estimate of 28 Mya as the divergence time between *T. gondii* and *N. caninum* and the intron genetic distance from Su *et al.,* (6), the intron mutation rate is 9.6×10^{-9}. Similarly the mtDNA mutation rate also varies depending on the time of divergence used

(see above, Table 3.2). The age of each NUMT was calculated based on both rates for comparison. The same rate was for used for all three species.

Identification of indels in orthologous sequences

Sequences of all coding regions, introns, exons, intergenic and 1 Kb regions flanking the genes were obtained from the *T. gondii* ME49 reference genome sequence. A BLASTN of each dataset was performed against the genomes of the remaining 15 assembled *T. gondii* strains using default parameters. The number of gaps in the subject and query of each alignment was obtained for all strains and datasets. Similarly, orthologous NUMTs were identified as described below and the number of gaps in each alignment was identified.

Identification of orthologous NUMTs

To identify ancestral NUMTs (present in all 16 strains), we performed a BLASTN search using each of the ME49 NUMTs, including 200 bp of flanking sequence as the query against each of the 15 strain genome sequences. Use of the NUMT sequences alone will not suffice since there are numerous independent insertions of the same mtDNA sequence in the genome at varying levels of decay. The 200 bp flanking sequence helps identify the orthologous location of each NUMT. The BLASTN results were filtered for an E-value = 0 and query coverage of ≥90%. Since 13 of the *T. gondii* strains have their sequences assembled only into contigs, some of the orthologous NUMTs were not identified because of missing ends.

Identification of NUMT presence/absence

We classified the NUMTs that are present in one or more strains but not in all strains as NUMTs that are differentially present or absent. If the NUMT was present in only one strain or absent from only one strain it was called a putative strain-specific insertion or deletion, respectively. In order to identify differentially present/absent NUMTs, we performed a BLASTN

of all the NUMTs including 200 bp flanking sequence from one strain against the genomes of all other strains in a pairwise fashion. BLASTN was performed using the following modified parameters: -q -5 -r 4 -D 0 -G 6 -E 5 -X 50 -F F -e 0.0000. For each pairwise comparison, the NUMT was considered strain-specific or differential present only when ≥150 bp of the upstream and downstream flanking sequence was present and the NUMT was missing at that same location in other genome sequences. If both flanking sequences were not present, the hit was not considered. The results from all pairwise comparisons were compared to identify and classify each NUMT as differentially present/absent or as a strain-specific insertion or deletion. An enrichment analysis of the genes present within 1 Kb upstream or downstream of differentially present NUMTs was performed using the GO enrichment analysis tool on ToxoDB.org.

We followed a similar approach to identify NUMTs that are present or absent between *T. gondii* and *H. hammondi*. *T. gondii* ME49 NUMTs were used as representatives. If a NUMT was present in *H. hammondi* and absent in *T. gondii* ME49, then it was considered specific to *H. hammondi*. Similarly if the NUMT was present in *T. gondii* ME49 and absent in *H. hammondi,* it was considered specific to *T. gondii*.

Calculation of NUMT insertion and deletion rates

The rate of NUMT insertion/deletion was calculated by dividing the number of insertions/deletions specific to that species or branch by the time of divergence or the total time on that branch (17, 22, 23). The NUMTs specific to *N. caninum* should not be present in the *T. gondii* and *H. hammondi* branch. In the case of *T. gondii,* only NUMTs that are present in *T. gondii* ME49 strain and absent in *H. hammondi* were considered. Likewise, *H. hammondi*-specific NUMTs would not be not present in *T. gondii* ME49. Insertion or deletion events were inferred based on the age of the NUMT and by looking for presence/absence between species.

CRISPR-Cas9 transfection assays

We experimentally tested if the non-homologous end joining (NHEJ) pathway is the mechanism of NUMT acquisition. The *T. gondii* strains RH (Type I) and the RH mutant TATiΔKu80 (referred here as ΔKu80) (24) (provided by B. Striepen, University of Georgia, USA) were cultured on human fibroblast reverse transcriptase (hTERT) cells grown in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% heat-inactivated Cosmic Calf Serum, 0.5% 10mg/ml pencillicin-streptomycin and 0.05% 10 mg/ml gentamycin. The CRISPR/Cas9 plasmid with a single guide RNA targeting the UPRT locus was generously provided by M. Cipriano/ B. Striepen (University of Georgia, USA).

T. gondii RH and ΔKu80 parasites were filtered, washed and resuspended in cytomix (supplemented with ATP and Glutamine) at a concentration of 3.3×10^7 parasites/ml. 10^7 parasites were electroporated with 8 μg of the CRISPR/Cas9 plasmid. Electroporation was done in a BTX electroporator using a 4 mm gap cuvette in a final volume of 400 μl. An hTERT confluent T25 flask was infected with 10^7 parasites and grown for 48 hours without drug selection. After 48 hours, 1ml of the lysed parasites were passed to another confluent T25 flasks and placed under 10 μM of FUDR drug selection to select for parasites that had a mutation in the UPRT gene. Parasites were harvested after 7 days of drug selection and DNA was extracted. Primers (Forward primer: 5′- GAGTTTGAACGCGTGTATCCCGCTTCATACG -3′; Reverse primer: 5′- GTAACAAAGTGGACAGCAGCCTCTTGGGTTG-3′) spanning a 450 bp region were used to PCR amplify the UPRT locus. The amplicons were TOPO cloned and transformed using the manufacturer's protocol. 1,152 clones from the RH strain and 576 clones from the ΔKu80 strain were screened for changes in amplicon length via PCR and gel electrophoresis. 10

clones showing amplicons of >500 bp were sequenced using the Sanger method (Macrogen USA).

RESULTS AND DISCUSSION

The age profile of NUMTs in *N. caninum* is different from *T. gondii* and *H. hammondi*

We used RepeatMasker to identify NUMTs in the nuclear genome of three coccidian species, *T. gondii*, *H. hammondi* and *N. caninum*. In our previous study, we reported *T. gondii* to have twice the number of NUMTs as *N. caninum* (Chapter 2). *H. hammondi* has a slightly greater NUMT content than *T. gondii* (Table 3.1). Only assembled chromosomes were used to identify NUMTs in *N. caninum* and *T. gondii*, since some of the unassembled contigs contain mtDNA sequences and would inflate the NUMT numbers. The *H. hammondi* genome assembly is currently in scaffolds. Although the *T. gondii* and *H. hammondi* genome sequences show a divergence of only 4.9% (25), we did not attempt to assemble the scaffolds into chromosomes using *T. gondii* chromosomes to avoid introducing any reference-based assembly biases especially with respect to the NUMTs. If the 3.14 Mb of unassembled *T. gondii* contigs were used, they contain approximately 128 Kb of identified NUMTs in varying levels of decay. Therefore, it is not possible to conclusively say if *H. hammondi* has a higher NUMT content than *T. gondii.*

To understand the evolutionary history of the NUMTs in this lineage, we estimated the time of insertion of each NUMT. Since the NUMTs are typically not functional in the nucleus, they are expected to evolve at a neutral rate. We used the intron mutation rate, mtDNA mutation rate and the percent divergence of the NUMT from the mtDNA sequence to calculate the age of each NUMT. Previously, 2.12×10^{-8} substitutions/base/million years was reported as the intron mutation rate (6). A more recent estimate suggests the divergence time between *T. gondii* and *N.*

caninum is 28 Mya (19) and not 12.7 Mya as previously reported, so the mutation rate would be 9.6×10^{-9} substitutions/base/million years (See Methods). We calculated the age of each NUMT using both these rates (Figure 3.1), assuming that the rates are similar in all three species. The age distribution of the NUMTs differs considerably between *N. caninum* and *T. gondii*/*H. hammondi*. The same general pattern of age distribution is observed for all three species with only a shift in the profile that correlates with the difference in the mutation rates employed. Therefore, similar inferences can be drawn in both cases. Since the 28 Mya estimate is based on a more comprehensive analysis, we will discuss the results calculated using the mutation rate of 9.6×10^{-9} substitutions/base/million years.

The NUMTs in N. caninum are younger

Overall nearly 90% of the NUMTs in *T. gondii* and *H. hammondi* and 98% of the NUMTs in *N. caninum* were acquired after the divergence of these two lineages ~28 Mya. Of the NUMTs that were likely present in the common ancestor, we were able to detect only 5 NUMTs that are shared between *T. gondii* and *N. caninum* (Chapter 2). It is possible that some orthologous NUMTs have subsequently diverged enough that they are no longer detected by our approach. However, 10% of the NUMTs in *T. gondii* and *H. hammondi* are estimated to be older than 28 million years (the estimated divergence time from *N. caninum*) compared to only 2% of the NUMTs in *N. caninum*, which likely means that both *T. gondii* and *H. hammondi* have retained more NUMTs than *N. caninum*.

N. caninum has a younger age distribution for its NUMTs. Nearly 70% of the NUMT sequences (268 Kb out of 385 Kb) were inserted within the last 15 million years. In comparison, 73% of the NUMT sequences in *T. gondii* (661 Kb out of 904 Kb) and *H. hammondi* (727 Kb out of 1.021 Mb) were inserted between 15 and 30 million years ago. This distribution suggests that

there was a rapid acquisition of NUMTs in *T. gondii* and *H. hammondi* in the 15- 25 million year time window followed by gradual decrease in that rate. Since the majority of the *N. caninum* NUMTs are younger, there was a recent, rapid, acquisition in this species. Taken together, these findings suggest that *N. caninum* has a higher rate of both acquiring and removing NUMTs than the other two species analyzed. It is also possible that *N. caninum* is just starting to acquire NUMTs like *T. gondii* did 15-30 million year ago. However, this theory does not explain the lack of ancestral NUMTs in the *N. caninum* genome. *T. gondii* and *H. hammondi* had a burst of NUMT acquisition and these NUMTs have been retained. The higher rate of retention could explain why both *T. gondii and H. hammondi* have the highest NUMT content ever reported. It is clear that the forces involved in acquisition and deletion of NUMTs in *N. caninum* and *T. gondii*/*H. hammondi* are different or are acting differently.

T. gondii NUMTs have a similar age spectra in different genomic locations

Nearly 90% of the NUMTs are present in introns and intergenic regions (Figure 3.2) and they each make up approximately 2.5% of the sequence present in these regions in the *T. gondii* genome. Only 0.01% of the coding region is comprised of NUMTs (Figure 3.2B). We observe a similar trend in *N. caninum*. Since UTRs are not annotated for most genes in *N. caninum*, we see a greater of proportion of the *N. caninum* NUMTs in the intergenic region (Figure 3.2B). It has been observed that recently transposed or younger TEs were present at similar densities in heterochromatic and euchromatic regions while older TEs were preferentially present in heterochromatin (26). It was proposed that TEs transpose randomly in to all regions and since heterochromatic regions have lower levels of recombination, TEs tend to get fixed there. We looked at the distribution of NUMTs in the different genomic features with respect to NUMT age and we do not see a bias in the location of the NUMTs (Figure 3.2C). The proportion of NUMTs

found in each genomic feature shows a fairly even distribution in each age bin (with the exception of CDS NUMTs in *N. caninum*, see below). This distribution profile mirrors the age profile of the NUMTs. Heterochormatic and euchromatic regions are not clearly defined in the *T. gondii* genome (27) like they are in plants and many other organisms. Chromatin immuno-precipitation studies reveal upstream flanking regions of most *T. gondii* genes to be sites of active transcription or open chromatin (ToxoDB.org). The NUMTs in the upstream flanking regions do not show an age bias. Recombination is considered an important factor for the preferential location of both older and younger TEs. While sexual recombination occurs in *T. gondii* (13), a sexual phase is not required for the parasite lifecycle and the parasite is haploid for the most of its life cycle (Figure 1.7A). Also, the *T. gondii* genome is much more compact in comparison to the genomes in which most TEs have been studied. These differences could explain the lack of age bias in the NUMT distributions.

Although 90% of the *N. caninum* NUMTs in coding regions are less than 10 million years old, they make up only 452 bp. It is possible that these NUMTs have not diverged because of selection pressure on the coding sequence. It is interesting that NUMTs present in coding regions are tolerated. One would expect that insertion into a coding sequence would disrupt the reading frame and thus would be deleterious to the gene. 30 insertions in *T. gondii* and 9 insertions *in N. caninum* are located in annotated coding regions (Table 3.5). However 23 out of the 30 are annotated as hypothetical in *T. gondii* and 5 out of the 9 in *N. caninum* are annotated as hypothetical. Only 10 genes (6 genes in *N. caninum* and 4 in *T. gondii*) have expression evidence suggesting that they are functional genes. It is interesting that insertions in some of these genes are predicted to cause a frame shift in the encoded protein sequence and these insertions are tolerated.

All three species show a high rate of NUMT insertion

We compared the *T. gondii* NUMTs to the genome of *H. hammondi* to identify orthologous NUMTs. A NUMT was called orthologous only if at least 90% of the NUMT and its 200 bp flanking region could be identified, contiguously, in the *H. hammondi* genome. NUMTs and flanking regions that fell at the end of contigs were not included. 7,109 NUMTs are orthologous between *T. gondii* and *H. hammondi* (Figure 3.3). 503 NUMTs that are orthologous between these species but not found in the *T. gondii* strains examined are not included in the *T. gondii* - *H. hammondi* orthologous NUMT dataset. These NUMTs could not be identified in some of the *T. gondii* strains because they were either deleted from these strains (discussed below) or present at the ends of contigs. These 7,109 include the 5 NUMTs shared with *N. caninum*. Using these orthologous NUMTs, we calculated insertion and deletion rates along the *N. caninum* and *T. gondii*/*H. hammondi* branches. To identify NUMT insertion and deletion events, we looked for alignment of the flanking sequences with a gap at the NUMT location. When the NUMT is present in only one of the species being compared, it's a called an insertion and if it absent in only one species it is called a deletion. However, due to the lack of conservation of NUMTs between *N. caninum* and *T. gondii/H. hammondi* this criterion is not very useful to distinguish insertion and deletion events. Also, it is not possible to use this criterion when comparing only two branches. Therefore we used the age of the NUMTs to distinguish insertion and deletion events.

599 of the 7,109 NUMTs on the *T. gondii*/*H. hammondi* branch are estimated to be older than 28 million years. Excluding these NUMTs and the 5 shared with *N. caninum*, we calculated the rate of insertion along *T. gondii/H. hammondi* branch as 232.5 NUMT insertions per million years. Similarly when considering all the *N. caninum* NUMTs younger than 28 million years, the

rate of insertion is 159.9 NUMTs per million years (Table 3.6, Figure 3.3). From the age distribution of the NUMTs (Figure 3.1), it can be predicted that the rate of insertion has not been uniform in either branch. These insertion rates are orders of magnitude higher than 1.26 NUMTs per million years reported for *Drosophila* (22) and 11.3 NUMTs per million years reported for chimpanzee (17) which has the highest insertion rate among investigated primates. It should be noted that it is difficult to estimate the exact number of insertion events in the coccidian genomes. In other studies, NUMTs within a certain distance (25 Kb in *Drosophila*) were predicted to have arisen from a single event (17, 22). These coccidian genomes are compact, measuring only 65 Mb. The NUMTs are fairly evenly distributed across the chromosomes (Chapter 2); hence we cannot use similar criteria to distinguish insertion events. Even if the rate of insertion is an overestimate, the number of base pairs inserted and retained is considerably higher (Table 3.6). An assembled full-length mitochondrial sequence is currently unavailable, so it is not possible to distinguish if nearby or adjacent NUMTs arose from single or multiple insertions events, although insertions from non-contiguous regions of the mtDNA have been described (28).

Using the same age criteria, the rate of deletion along the *N. caninum* branch is 21.3 deletions/million years and 2 deletions/million years along the *T. gondii/H. hammondi* branch (Table 3.6). These rates are likely incorrect. If *N. caninum* is indeed loosing NUMTs at a faster rate, quantifying the older NUMTs in *N. caninum* to estimate the deletion rate in the *T. gondii*/*H. hammondi* branch will underestimate the rate. Also, it is difficult to calculate a rate of deletions since it is hard to prove the reason for an absence.

Similar rates of insertion and deletion in T. gondii and H. hammondi

The high rate of insertion in the *T. gondii*/*H. hammondi* branch correlates with the rapid accumulation of NUMTs 15-25 million years ago. From the age profile of the NUMTs, it appears that the rate of insertion decreased over time. To test this hypothesis, we calculated the rate of insertion in *T. gondii* and *H. hammondi* after they diverged. Since this is a two-way comparison, we need to use the age of the NUMTs and the divergence time between the two species to distinguish insertion and deletion events. We estimated the time of divergence between *T. gondii* and *H. hammondi* as 6.1 mya. 56 NUMTs younger than 6.1 mya are present in *T. gondii* ME49 (and all 16 *T. gondii* strains) and absent at the orthologous location in *H. hammondi* (Table 3.6, Figure 3.3). We considered these as *T. gondii*-specific insertions, which gives a *T. gondii* specific insertion rate of 9.1 insertions/million years. Conversely, if a NUMT younger than 6.1 my is present in *H. hammondi* and absent in *T. gondii* ME49, it is considered an *H. hammondi*-specific insertion. *H. hammondi* has 62 specific insertions, leading to a calculated insertion rate of 10.1 insertions/million years. Although many *T. gondii* strains are available only the ME49 strain was used for this analysis to maintain a fair comparison with the one *H. hammondi* strain.

Similarly, we calculated the rate of deletion as 3.9 deletions/million years in *T. gondii* and 5.9 deletions/million years in *H. hammondi*. 24 NUMTs older than 6.1 my are present only in *H. hammondi* and absent at that location in *T. gondii* ME49, while 36 NUMTs older than 6.1 my are present in *T. gondii* ME49.

Depending on the divergence time, these rates will vary slightly. However, it is clear that the rate of insertion has decreased over time. The rate of deletion has remained low, suggestive of *T. gondii* and *H. hammondi's* ability to retain NUMTs.

SNPs and indels in T. gondii NUMTs are similar to the rest of the genome

In addition to insertion and deletion of NUMTs, we looked at other evolutionary forces acting on the NUMTs. Single-nucleotide polymorphisms (SNPs) and small insertions and deletions (indels) are important forces of sequence decay or mutation. Genomic sequence reads from the 62 strains were mapped to the reference *T. gondii* ME49 genome and 802,764 SNP positions were identified (13). We identified the genomic location and associated genomic features for each of these SNPs (Figure 3.4). As expected, introns had the highest percentage of SNPs and coding regions had the lowest SNP content. Except in the case of coding sequences, the SNP content in the NUMTs correlated with NUMT location. That is, NUMTs in introns had the highest SNP content. NUMTs in the CDS do not follow this trend; the percentage of SNPs in NUMTs in the CDS is higher than the overall percentage of SNPs in the CDS. The SNPs in NUMTs located in coding regions are all present in coding regions that are annotated as hypothetical proteins and these hypothetical proteins also show very little expression evidence (ToxoDB.org). It is possible these may not be real genes. Overall, the SNPs in these NUMTs do not necessarily occur at a rate significantly higher than the surrounding location.

We next looked at insertion/deletion (indels) events. The sequences of different genomic features from ME49 were aligned against the genomes of all assembled strains and the number of gaps in each pairwise alignment was determined. Except for a few strains that showed an insignificant number (< 0.001%) of indels with respect to ME49, indels were generally not detected in the NUMTs. Therefore, neither SNPs nor indels in NUMTs occur at a rate significantly different from the surrounding location. When NUMTs enter a locus, it appears that they are shaped by the evolutionary forces inherent to that region.

Differential presence and absence of NUMTs in *T. gondii* strains

We have previously shown that the acquisition and deletion of NUMTs in *T. gondii* is an ongoing process. The three genome sequences examined, from strains ME49, GT1 and VEG, revealed differential presence and absence of NUMTs (Chapter 2). Notably, ME49 has an ~3500 bp NUMT region that is nearly identical to the mtDNA sequence and is likely to be a recent insertion (and we did not find this NUMT in the other 13 strains analyzed here). We extended this analysis to 13 additional strains. All of the strains showed a similar NUMT content (Table 3.1). Only the ME49, GT1 and VEG genome sequences are assembled into chromosomes. The remaining genome sequences are in scaffolds or contigs. NUMTs that are present at the ends of scaffolds/contigs could not be included in these analyses since sufficient flanking regions were not present. We included NUMTs only if the flanking regions could be identified in all strains. Overall, we identified 8,271 ancestral NUMTs that are present in all 16 strains.

57 NUMTs showed differential presence or absence in the 16 strains (Table 3.7, Figure 3.5). These NUMTs are in varying states of decay, ranging from 0%-31% divergence from the mtDNA. No strain shows a preferential loss or gain/retention of NUMTs. MAS has the largest number of NUMTs at 36 out of 57 and VEG has lost the most NUMTs (33/57). MAS has 2 strain-specific NUMTs at >98% identity (NUMT 6, 7, Table 3.7, Figure 3.5) to the mtDNA. Cat_PRC2 has 1 strain-specific NUMT (NUMT 8) that has only 94% identity to the mtDNA. Although this NUMT is only present in Cat_PRC2, it has likely been lost in the other strains. All other strains have the first 10 bp of this NUMT and show a 5 bp micro-homology at the proposed site of deletion discussed below. Similarly, NUMT 56, which is present in 7 strains, is 100% identical to the mtDNA sequence, but it is also likely a loss in the other strains. ~50 bp of the upstream flanking sequence is absent in the strains that lack the NUMT and the flanking sites

show a 5 bp micro-homology. The micro-homology sequence is different in each case mentioned above. We observed this pattern of micro-homology at many NUMT sites, particularly at sites of deletion. If a NUMT has diverged by at least 2% from the mtDNA and/or if some of the flanking region was missing in the strains where the NUMT is absent, we considered it a deletion (Figure 3.6). The observed micro-homology regions span 2-6 bp and the site is not conserved across NUMT deletion sites. Non-homologous end joining (NHEJ) is the preferred mechanism of double-strand break repair in *T. gondii* and NHEJ can repair double-strand breaks with little or no sequence homology at the break sites..

Some NUMTs with differential presence/absence do not correlate with local ancestry

Between 2 to 15 strains share 54 out of the 57 differentially present NUMTs. The sites of insertion or deletion of these NUMTs, barring a few exceptions, are conserved among the strains. It is more likely that these NUMTs were inserted or deleted in an ancestor of these specific strains and that the strains which evolved from this ancestor inherited the presence/absence state. For example, NUMTs 19 and 43 (100% identical to the mtDNA) are present in only MAS and TgCATBr5, which clustered together as Clade B, and are therefore likely, specific insertions in the ancestor of Clade B (if we assume the strains in each clade arose from a common ancestor of that clade). (Refer to Figure 1.8 for the classification of clades. In Figure 3.5, the strains are grouped in to the same clades and color-coded appropriately as in Figure 1.8). However NUMTs 13, 36 and 38, are also present in two strains and are 100% identical to the mtDNA sequence but are found in strains present in different clades. It appears as if the strain in each clade gained these NUMTs independently. However, it seems unlikely these were independent insertions, since the site of insertion is conserved.

We hypothesized that the NUMTs which are differentially present/absent and are also shared by one or more strains belonging to different clades arose as a result of recombination. We examined data from (13) (in review, of which I am also an author) to determine the regions of local admixture, to see if genetic crosses could explain the observed pattern. The local admixture analysis used bins of 1000 SNPs to reveal regions of common ancestry among the strains, an effect brought about by sexual recombination (13) (Figure 1.8B). We identified the ancestral state of the locus for each of the 57 NUMTs by using the coordinates of the SNP bins. Each NUMT was then colored using the color of the major or dominant clade as depicted in Figure 1.8B. In short, Figure 3.5 is a snap shot of the loci of the 57 NUMTs from Figure 1.8B.

We re-examined the differential presence/absence of NUMTs in light of local ancestry. For example, the location of NUMT 27 in Chr VIII shows an ancestry pattern in agreement with its clade except for TgCATBr5 (clade B) which shares common ancestry with clade C. Presence/absence of the NUMT is conserved for each clade based on its ancestry pattern, except for TgCATBr5 which lacks the NUMT like the strains of clade C. Our hypothesis is true in this case and also for NUMTs 29, 30 and 38. However, it is not always true. NUMT 13 is present only in p89 and VAND but these regions do not share ancestry according to this analysis. NUMT 39 is present in 4 strains, 3 of which share common ancestry and one which does not.

A number of reasons could be responsible for the observed difference in the NUMT presence/absence and local ancestry profile. First, recombination can occur in spans of less than 1,000 SNPs. Second, recombination may have occurred with strains that are not analyzed in this study. Third, NUMTs may have been gained and lost independently. Fourth, the NUMT may have been lost after the recombination event (for example NUMTs 18 and 24). Fifth, the ancestral population may have been polymorphic for these NUMTs and when the population

diverged the NUMTs may have sorted differentially; a process called lineage-sorting. Although the third and fourth reasons are the easiest explanations to the observed patterns, these seem to be the least probable causes, because, the site of NUMT gain/loss is conserved. While it may be possible for the NUMTs to be deleted in the same fashion in different strains, unless there is a signature for such sites, it is unlikely the NUMTs were inserted at a particular site in multiple strains independently. We did not observe any such signatures or conserved sequence motifs at NUMT sites (Chapter 2).

NUMTs are differentially present/absent in genic regions

We did not observe a clade-specific or strain-specific pattern in differential presence or absence of NUMTs, most likely because of recombination. We looked at the location of these 57 differentially present NUMTs to understand any biological significance. 31 out of the 57 NUMTs are located in introns and 13 NUMTs are located in the flanking regions or UTR, the remaining 13 are located in intergenic regions outside of the 1 Kb upstream and downstream flanking regions (Table 3.7). None of the 57 NUMTs were located in a CDS. We conducted a GO enrichment analysis with the 53 genes whose introns or flanking regions contained NUMTs (Table 3.8). Genes involved in purine and pyrimidine metabolism are enriched. These pathways are core survival pathways and, interestingly, *T. gondii* acquires purines from its host via salvage pathways. These pathways are enriched even when conducting the analysis with just the genes that contain NUMTs in the UTR or flanking regions. It is possible that the presence of the NUMTs in these regions may affect transcription at the locus. However, transcriptome data are not currently available for many of the strains, so these findings warrant further exploration.

Non-homologous end joining is the likely mechanism of NUM/PT acquisition

It has been shown that organellar DNA fragments can become integrated into the nuclear genome during the error-prone double-strand break (DSB) mechanism, non-homologous end joining (NHEJ). NHEJ can repair DSBs without sequence homology but it often results in the deletion of a few nucleotides (29). This mechanism of integration was first demonstrated in *Saccharomyces cerevisiae* and subsequently proven in mammalian cells and plants (30-33). In *T. gondii*, it has been observed that DNA is readily integrated in to the nuclear genome at random sites via NHEJ. In fact, it was necessary to knock-out a critical NHEJ protein, Ku80, in order to increase the efficiency of homologous recombination in this species (34). The preferential use of NHEJ in *T. gondii* is, in itself, suggestive of NHEJ being the likely mechanism of NUM/PT acquisition. Also, interestingly, apicomplexans that lack the NHEJ machinery contain very few NUM/PTs (Table 2.1, Figure 2.8).

We used the recently-developed CRISPR-Cas9 system in *T. gondii* (35, 36) to experimentally test if NHEJ is the mechanism of integration. We disrupted the *uprt* locus in the RH strain and in an RH strain that was ΔKu80. The RH strain is located in Clade A and is highly virulent. The *uprt* locus was amplified and the amplicons were cloned. We screened 1,152 colonies from the RH transfection and 576 colonies from the ΔKu80 transfection. 16% (183) of the RH colonies had amplicons that were larger than expected. None of the ΔKu80 colonies showed larger amplicons, 2 colonies had smaller amplicons. We sequenced 1 RH (M11, Figure 3.7) colony that did not contain a larger amplicon and 10 of the 183 colonies (M1-10, Figure 3.7) that did contain large amplicons to identify the insert at the DSB sites. M11 had a 3 bp deletion and M3 had a 2 bp insertion. The rest of the colonies had insertions ranging from 44 bp to 161 bp. All colonies had inserted sequence originating from the transfected plasmid, which was

similar to what was observed previously (36). Only 2 colonies, M4 and M8 had a one bp deletion at the flank site, which is similar to an observation made in mammalian cells where strand deletion was rare during re-ligation of DSBs (31). We expect the majority of the inserts to be from plasmid since that it is the DNA readily available for repair. Deep amplicon sequencing needs to performed to identify any organellar DNA insertions because of intervening plasmid DNA and these experiments are underway. We are awaiting the sequence results.

These results, however, do inform us about the propensity with which DNA is becoming integrated at DSB sites. 1.5% of the colonies from the yeast experiments and 8.3% of the colonies from experiments in mammalian cells contained inserts (although these numbers are dependent on the number of DSBs induced). The number of insertions we observed in *T. gondii* is considerably higher and can explain why *T. gondii* has the highest NUM/PT content ever observed. Deletion of (functional) sequence at the DSB site can have a negative impact on the organism. Based on observations in yeast (33) and mammalian cells (29), it has been proposed that organellar DNA can act as filler DNA reducing the incidence of deletion. In the yeast experiments, where mtDNA was observed in DSBs, the mitochondria were found to be still functional (33). It is possible DSBs are unusually high in the *T. gondii* genome and in a genome that is haploid the majority of the time it is possible these organellar DNA sequences are inserted to mitigate deletion of sequences at the site of DSB repair, although insertion of organellar DNA will still cause mutations. Based on the age of the NUMTs, *N. caninum* has a larger number of recent insertions. It will be interesting to investigate if *N. caninum* has a higher insertion rate in similar experiments.

CONCLUSIONS

T. gondii is a notorious parasite that has evolved ways to become a generalist, capable of infecting any warm blooded animal. Here we provide evidence for another of its innovations, massive numbers of NUMT (and NUPT see Chapter 2) insertions. NUMTs are present at significantly higher numbers in *T. gondii*. They acquired these sequences at a rapid rate ~15 Mya and appear to have a high NUMT retention rate. While the rate of acquisition has slowed down, it is still high enough to generate differences across strains. The functional consequence of these differences needs to be explored, particularly through examination of expression data of genes in the vicinity of differentially present NUMTs. The NUMT age profile of the closely related species *N. caninum* is significantly different. NHEJ appears to the likely mechanism of NUMT acquisition in *T. gondii.* While the NHEJ pathway proteins are present in *N. caninum*, based on the age profile of the NUMTs, it appears that other forces may also be involved in creating a different profile. Our findings suggest that NUMTs play a significant role in the evolution of *T. gondii.*

Acknowledgements

We would like to thank Michael Cipriano and Boris Striepen, University of Georgia for graciously providing the CRISPR-Cas9 plasmids.

REFERENCES

1. Levine ND (1988) *The Protozoan Phylum Apicomplexa Vol II* (CRC Press) p 154.
2. Cowper B, Matthews S, & Tomley F (2012) The molecular basis for the distinct host and tissue tropisms of coccidian parasites. *Molecular and biochemical parasitology* 186(1):1-10.
3. Dubey JP (2008) The history of *Toxoplasma gondii*--the first 100 years. *The Journal of eukaryotic microbiology* 55(6):467-475.
4. Weinman D & Chandler AH (1954) Toxoplasmosis in swine and rodents; reciprocal oral infection and potential human hazard. *Proceedings of the Society for Experimental Biology and Medicine. Society for Experimental Biology and Medicine* 87(1):211-216.
5. Hutchison WM (1965) Experimental transmission of *Toxoplasma gondii*. *Nature* 206(987):961-962.
6. Su C*, et al.* (2003) Recent expansion of *Toxoplasma* through enhanced oral transmission. *Science* 299(5605):414-416.
7. Howe DK & Sibley LD (1995) *Toxoplasma gondii* comprises three clonal lineages: correlation of parasite genotype with human disease. *The Journal of infectious diseases* 172(6):1561-1566.
8. Sibley LD & Boothroyd JC (1992) Virulent strains of *Toxoplasma gondii* comprise a single clonal lineage. *Nature* 359(6390):82-85.
9. Sibley LD & Ajioka JW (2008) Population structure of *Toxoplasma gondii*: clonal expansion driven by infrequent recombination and selective sweeps. *Annual review of microbiology* 62:329-351.
10. Khan A*, et al.* (2007) Recent transcontinental sweep of *Toxoplasma gondii* driven by a single monomorphic chromosome. *Proceedings of the National Academy of Sciences of the United States of America* 104(37):14872-14877.
11. Khan A*, et al.* (2011) A monomorphic haplotype of chromosome Ia is associated with widespread success in clonal and nonclonal populations of *Toxoplasma gondii*. *mBio* 2(6):e00228-00211.
12. Su C*, et al.* (2012) Globally diverse *Toxoplasma gondii* isolates comprise six major clades originating from a small number of distinct ancestral lineages. *Proceedings of the National Academy of Sciences of the United States of America* 109(15):5844-5849.
13. Lorenzi H. KA, Benke M.S. ,Namasivayam S.,Seshadri L.S.,Hadjithomas M.,Karamycheva S.,Pinney D.,Brunk B.,Ajioka J.W.,Ajzenberg D.,Boothroyd J.C.,Boyle J.P.,Dardé M.L.,Dubey J.P.,Fritz H.M.,Gennari S.M., Gregory B.D.,Kim K.,Rosenthal B. M.,Saeij J., Su C.,White M.W.,Zhu X.Q,Howe D.K.,Grigg M.E.,Parkinson J.,Liu L.,Kissinger J.C.,Roos D.S.,Sibley L. D. (2015) Comparative sequence analysis of Toxoplasma gondii reveals local genomic admixture drives concerted expansion and diversification of secreted pathogenesis determinants. *In Review*.
14. Ricchetti M, Tekaia F, & Dujon B (2004) Continued colonization of the human genome by mitochondrial DNA. *PLoS biology* 2(9):E273.
15. Hazkani-Covo E, Zeller RM, & Martin W (2010) Molecular poltergeists: mitochondrial DNA copies (numts) in sequenced nuclear genomes. *PLoS genetics* 6(2):e1000834.
16. Chen JM, Chuzhanova N, Stenson PD, Ferec C, & Cooper DN (2005) Meta-analysis of gross insertions causing human genetic disease: novel mutational mechanisms and the role of replication slippage. *Human mutation* 25(2):207-221.

17. Hazkani-Covo E (2009) Mitochondrial insertions into primate nuclear genomes suggest the use of numts as a tool for phylogeny. *Molecular biology and evolution* 26(10):2175-2179.
18. Michalovova M, Vyskot B, & Kejnovsky E (2013) Analysis of plastid and mitochondrial DNA insertions in the nucleus (NUPTs and NUMTs) of six plant species: size, relative age and chromosomal localization. *Heredity* 111(4):314-320.
19. Reid AJ*, et al.* (2012) Comparative genomics of the apicomplexan parasites *Toxoplasma gondii* and *Neospora caninum*: Coccidia differing in host range and transmission strategy. *PLoS pathogens* 8(3):e1002567.
20. Gjerde B (2013) Characterisation of full-length mitochondrial copies and partial nuclear copies (numts) of the cytochrome b and cytochrome c oxidase subunit I genes of *Toxoplasma gondii*, *Neospora caninum*, *Hammondia heydorni* and *Hammondia triffittae* (Apicomplexa: Sarcocystidae). *Parasitology research* 112(4):1493-1511.
21. Tamura K, Stecher G, Peterson D, Filipski A, & Kumar S (2013) MEGA6: Molecular Evolutionary Genetics Analysis version 6.0. *Molecular biology and evolution* 30(12):2725-2729.
22. Rogers HH & Griffiths-Jones S (2012) Mitochondrial pseudogenes in the nuclear genomes of Drosophila. *PloS one* 7(3):e32593.
23. Bensasson D, Feldman MW, & Petrov DA (2003) Rates of DNA duplication and mitochondrial DNA insertion in the human genome. *Journal of molecular evolution* 57(3):343-354.
24. Sheiner L*, et al.* (2011) A systematic screen to discover and analyze apicoplast proteins identifies a conserved and essential protein import factor. *PLoS pathogens* 7(12):e1002392.
25. Walzer KA*, et al.* (2013) *Hammondia hammondi*, an avirulent relative of *Toxoplasma gondii*, has functional orthologs of known *T. gondii* virulence genes. *Proceedings of the National Academy of Sciences of the United States of America* 110(18):7446-7451.
26. Blumenstiel JP, Hartl DL, & Lozovsky ER (2002) Patterns of insertion and deletion in contrasting chromatin domains. *Molecular biology and evolution* 19(12):2211-2225.
27. Brooks CF*, et al.* (2011) *Toxoplasma gondii* sequesters centromeres to a specific nuclear region throughout the cell cycle. *Proceedings of the National Academy of Sciences of the United States of America* 108(9):3767-3772.
28. Kleine T, Maier UG, & Leister D (2009) DNA transfer from organelles to the nucleus: the idiosyncratic genetics of endosymbiosis. *Annual review of plant biology* 60:115-138.
29. Hazkani-Covo E & Covo S (2008) Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS genetics* 4(10):e1000237.
30. Ricchetti M, Fairhead C, & Dujon B (1999) Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* 402(6757):96-100.
31. Lin Y & Waldman AS (2001) Promiscuous patching of broken chromosomes in mammalian cells with extrachromosomal DNA. *Nucleic Acids Res* 29(19):3975-3981.
32. Lin Y & Waldman AS (2001) Capture of DNA sequences at double-strand breaks in mammalian chromosomes. *Genetics* 158(4):1665-1674.
33. Yu X & Gabriel A (1999) Patching broken chromosomes with extranuclear cellular DNA. *Mol Cell* 4(5):873-881.

34. Fox BA, Ristuccia JG, Gigley JP, & Bzik DJ (2009) Efficient gene replacements in *Toxoplasma gondii* strains deficient for nonhomologous end joining. *Eukaryotic cell* 8(4):520-529.
35. Shen B, Brown KM, Lee TD, & Sibley LD (2014) Efficient gene disruption in diverse strains of *Toxoplasma gondii* using CRISPR/CAS9. *mBio* 5(3):e01114-01114.
36. Sidik SM, Hackett CG, Tran F, Westwood NJ, & Lourido S (2014) Efficient genome engineering of *Toxoplasma gondii* using CRISPR/Cas9. *PloS one* 9(6):e100450.

Figures and table legends

Figure 3.1. Distribution of estimated NUMT insertion times

The insertion time of each NUMT was calculated based on the nucleotide distance from the mtDNA sequences. Y-axis represents the percentage of NUMT base pairs occupied. X-axis represents age in millions of years. Age was calculated using a mutation rate of **A.** 9.6×10^{-9} substitutions/base/million years. **B.** 2.18×10^{-8} substitutions/base/million years (See Methods).

Figure 3.2. Age distribution of NUMTs in different genomic locations

A. The percent of each genomic feature as annotated in the genome sequences of *T. gondii* ME49 and *N. caninum* LIV. Intergenic regions include the 1 Kb upstream flanking and downstream flanking regions. Data for flanking regions are also indicated separately. **B**. The distribution of NUMTs in each genomic feature is indicated as a percentage of NUMT sequence. Note, very few UTRs are annotated for *N. caninum*. **C-D**. The age of each NUMT was calculated using a mutation rate of 9.6×10^{-9} substitutions/base/million years. The percentage of NUMT base pairs from different genomic locations in each age bin is shown for *T. gondii* (C) and *N. caninum* (D)

Figure 3.3. Insertion and deletion of NUMTs in the Coccidia

Lineage-specific insertions and deletions are shown along the branches in green and red respectively. Insertions and deletions were determined based on age of the NUMT relative to the branch point. The total number of NUMTs along a branch irrespective of age is shown in brackets. Only 5 orthologous NUMTs could be identified between *N. caninum* and *T. gondii*/*H. hammondi*. The divergence times are as indicated. The tree is not drawn to scale.

Figure 3.4. Distribution of SNPs in *T. gondii* genome features and NUMTs

The number of SNP positions in the different genomic locations of the *T. gondii* genome was calculated using data from 62 strains (13). The percentage of SNPs in each genomic feature or NUMTs alone are shown for the entire genome.

Figure 3.5. NUMTs show differential presence and absence in *T. gondii* strains

Strains, arranged by clade, are shown on the y-axis. The clades are colored as shown in (13) and Figure 1.8A and B. The 14 chromosomes, divided by black lines, are represented on the x-axis. NUMTs that show differential presence/absence in the 16 *T. gondii* strains are numbered and ordered based on their location on the ME49 chromosomes. The coordinates of each NUMT plus 200 bp flanking region on the ME49 chromosomes are shown at the bottom. The 200 bp regions were included to provide a location for NUMTs missing in ME49. The percent divergence of the NUMT from the mtDNA sequence is indicated above the coordinates. In the matrix, 1 indicates presence of the NUMT and 0 indicates absence in a particular strain. Each cell in the matrix is colored based on local ancestry of that location identified from analyses conducted using blocks of 1,000 SNPs (13) and Figure 1.8A and B although it doesn't imply that location originated from that clade.

Figure 3.6. Micro-homology at NUMT deletion sites

Multiple sequence alignment of a 63 bp NUMT with its flanking sequences. Only a few representative strains were chosen for alignment. The NUMT sequence is highlighted with a green box. The flanking region sequences that align in all strains shown are colored in black. Regions of flanking sequences that are not present in all strains are not colored. Regions of micro-homology are marked with red brackets.

Figure 3.7. Insertions discovered at the site of CRISPR/Cas9-directed double-strand breaks

WT is the wild type sequence from the *T. gondii* GT1 genome of the CRISPR/Cas9 targeted *uprt* locus (GT1 genome sequence was used since an assembled RH genome sequence is not available. GT1 also belongs to Clade A and is highly virulent). The gap (-) in the sequence of WT indicates the site of double-strand break (DSB); the gap between the two nucleotides is just to aid alignment with the mutant sequences and does not indicate a gap or loss of nucleotides at that site. M1 to M11 are mutant sequences from clones generated through sub-cloning of PCR amplicons. The size of the insert (in base pairs) at the site of the DSB is indicated for each mutant.

Table 3.1. NUMT content in *T. gondii* strains and *H. hammondi*

For *T. gondii* strains ME49, GT1 and VEG, NUMT content is based on assembled chromosomes. For the remaining strains and *H. hammondi,* all available scaffolds/contigs were used.

Table 3.4. Age distribution of NUMTs in different genomic features

The number of NUMTs base pairs in each genomic location for each age bin is indicated for *T. gondii* and *N. caninum*. NUMTs that span more than one genomic feature are not included. Intergenic region includes the 1 Kb upstream flanking and downstream flanking regions.

Table 3.5. NUMTs in coding regions

Genes that contain NUMTs in their coding regions are listed for *T. gondii* ME49 and *N. caninum* LIV. Expression of the gene was determined based on transcriptome data available on ToxoDB.org release 24.

Table 3.6. NUMT insertion and deletion rates

Lineage-specific insertions and deletions were determined based on the age of each NUMT and its presence or absence in other branches.

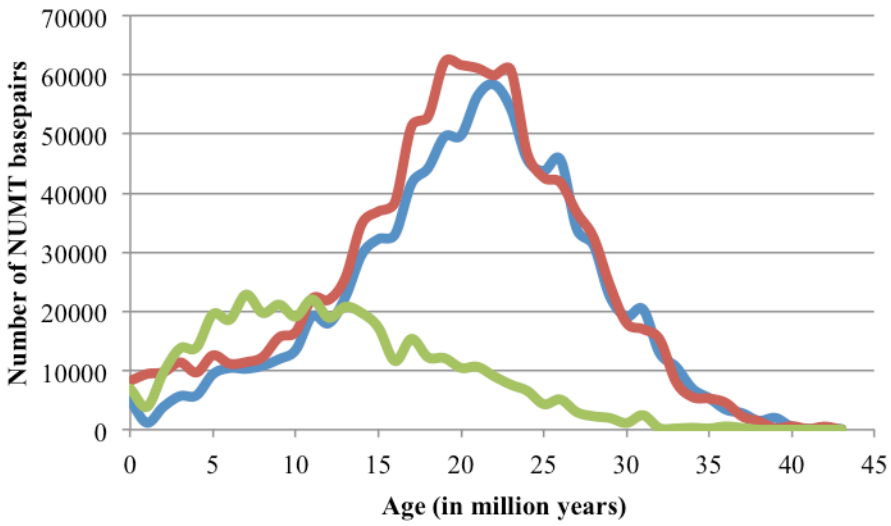
Table 3.7. Location of NUMTs displaying differential presence/absence in 16 *T. gondii* strains

Location of NUMT is with respect to the ME49 genome sequence. Genomic locations are indicated as CDS = C, UTR = U, Intron = I, Upstream flank = UF, downstream flank = DF, Intergenic = IG (intergenic does not include flanking regions). Coordinates for NUMT location include 200 bp upstream and downstream flanking regions of the NUMT. Presence =1, absence 0.

Table 3.8. Enrichment analyses of genes associated with NUMTs displaying differential presence/absence

Enrichment analysis was performed on ToxoDB.org with all ME49 genes as the background. Results were filtered for p-value of 0.05 (Bgd = Background).

A.



B.

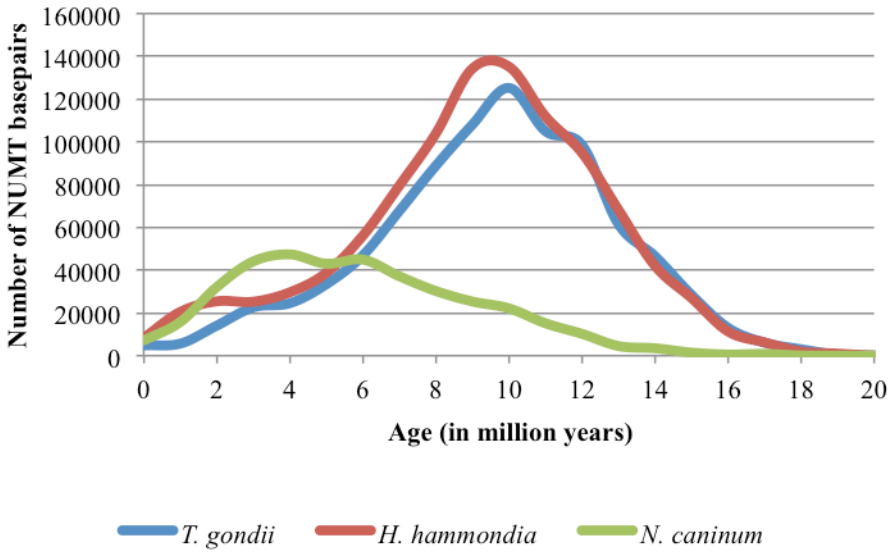


Figure 3.1. Distribution of estimated NUMT insertion times

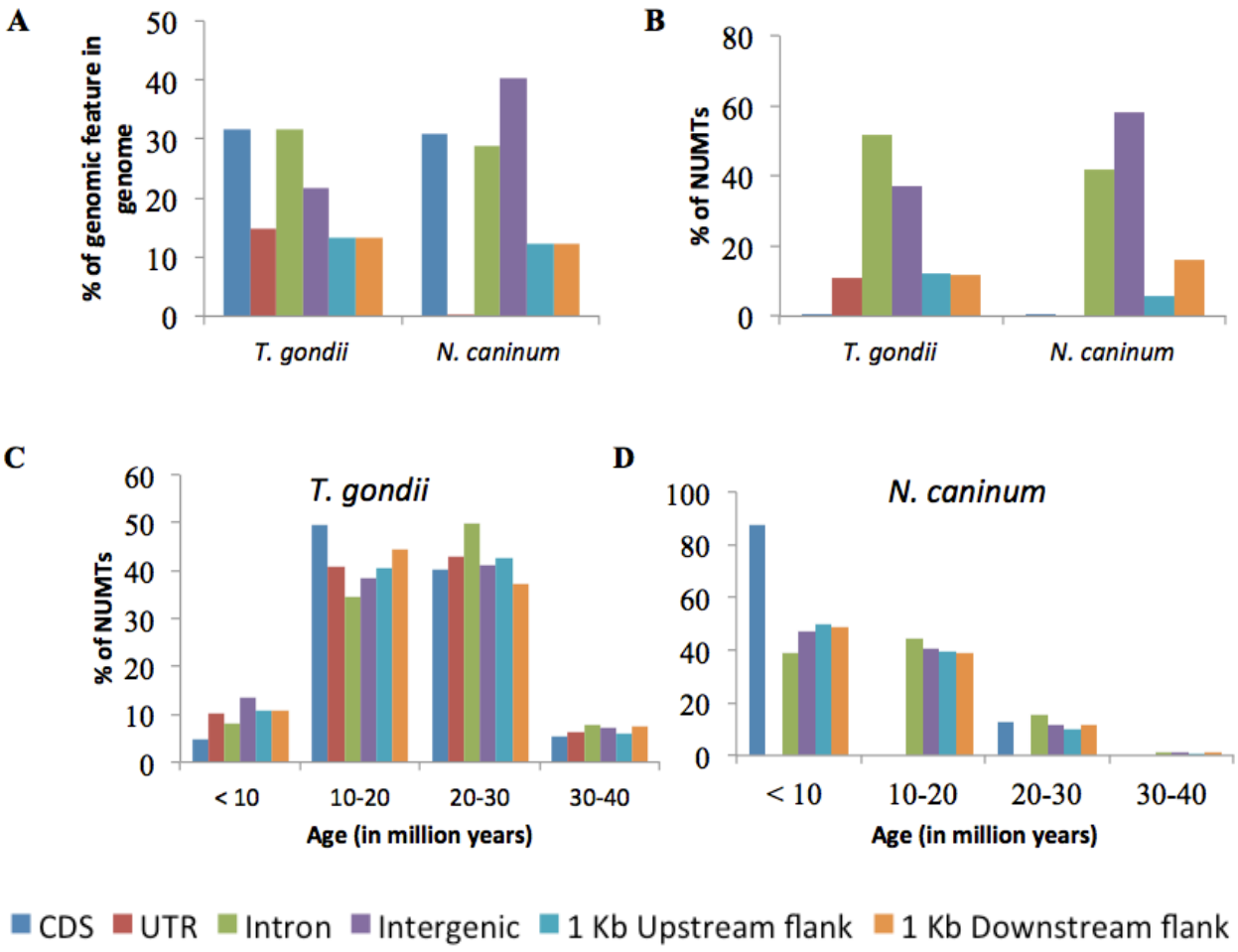


Figure 3.2. Age distribution of NUMTs in different genomic locations

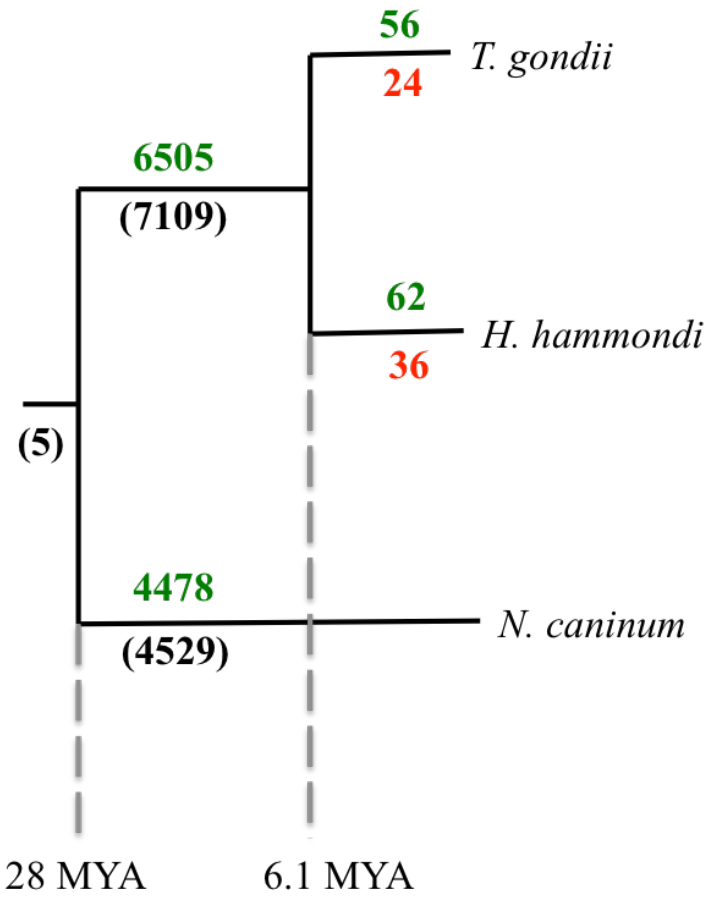


Figure 3.3. Insertion and deletion of NUMTs in the Coccidia

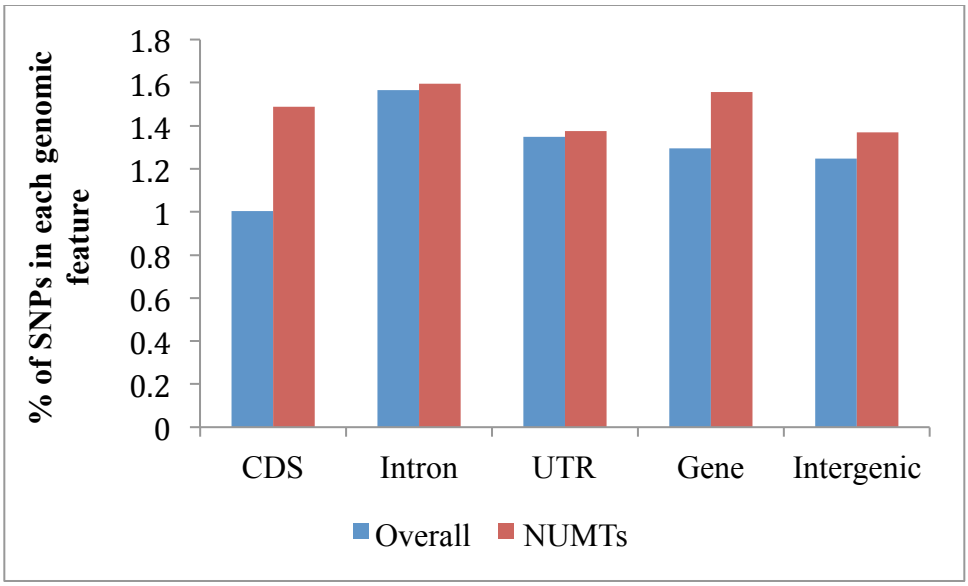


Figure 3.4. Distribution of SNPs in *T. gondii* genome features and NUMTs

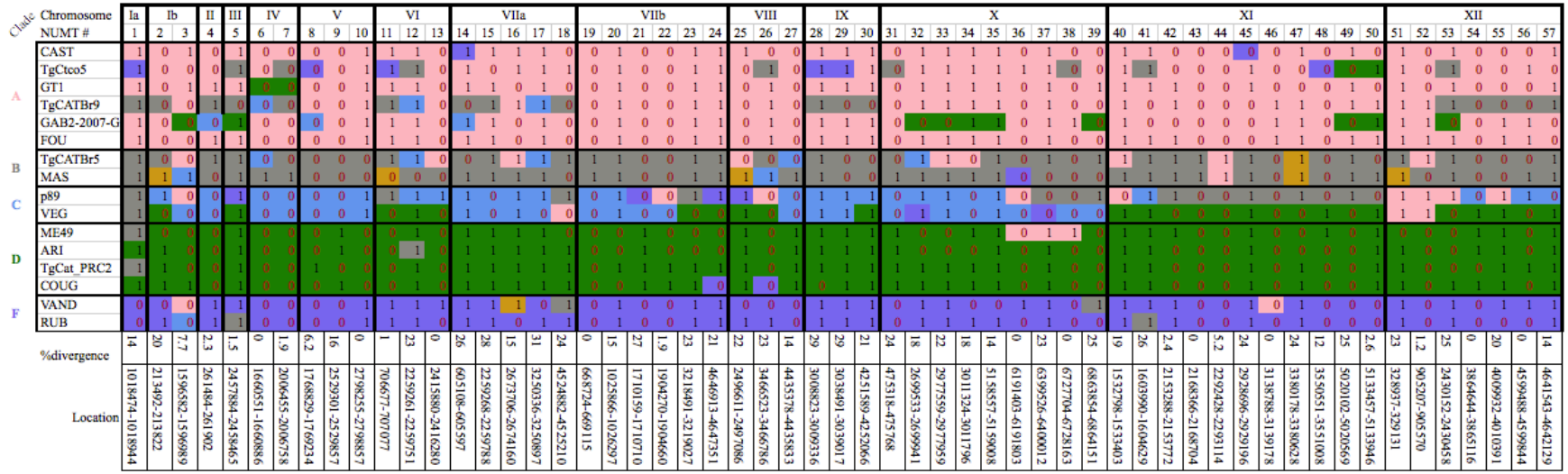


Figure 3.5. NUMTs show differential presence and absence in *T. gondii* strains

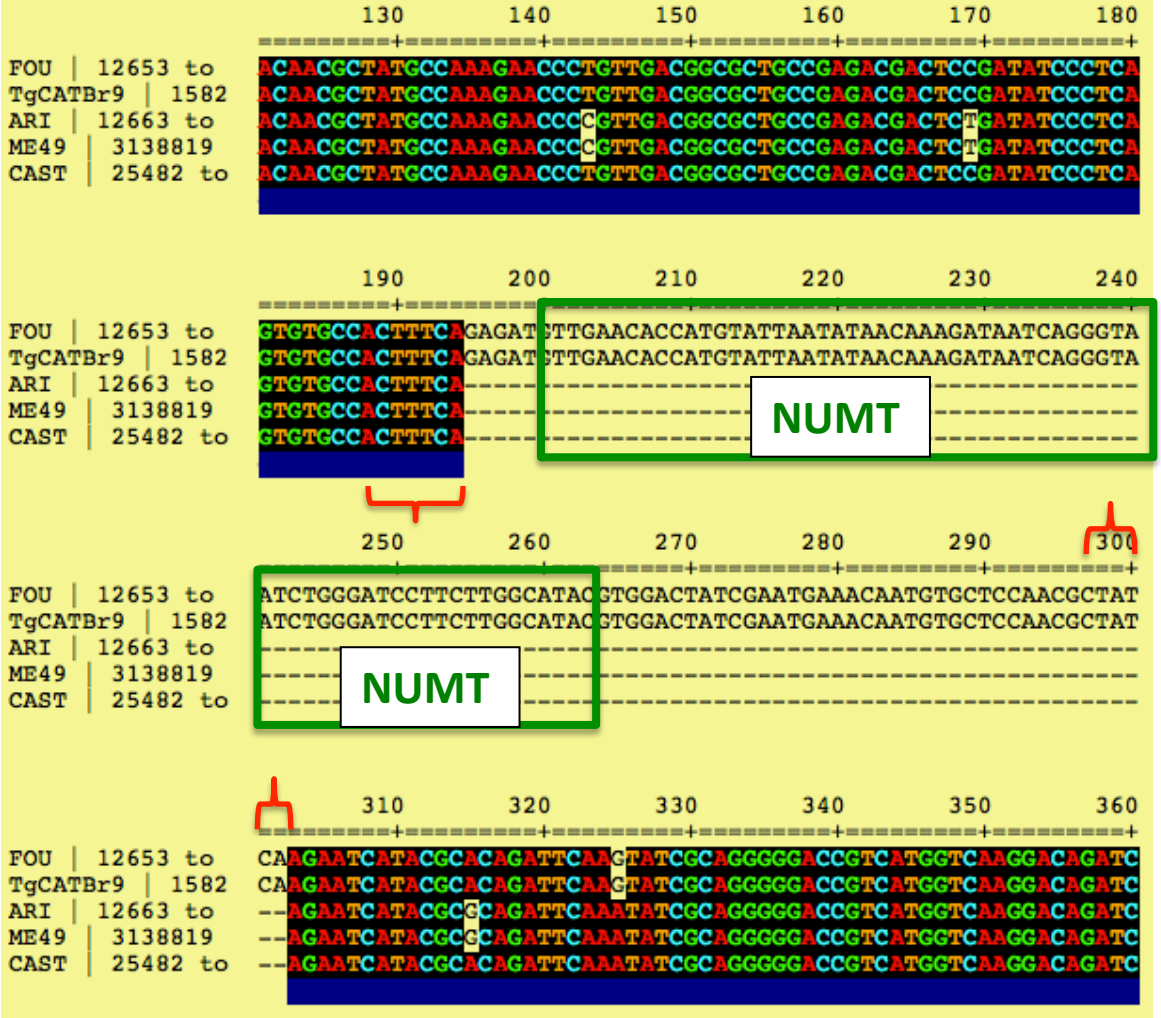


Figure 3.6. Micro-homology at NUMT deletion sites

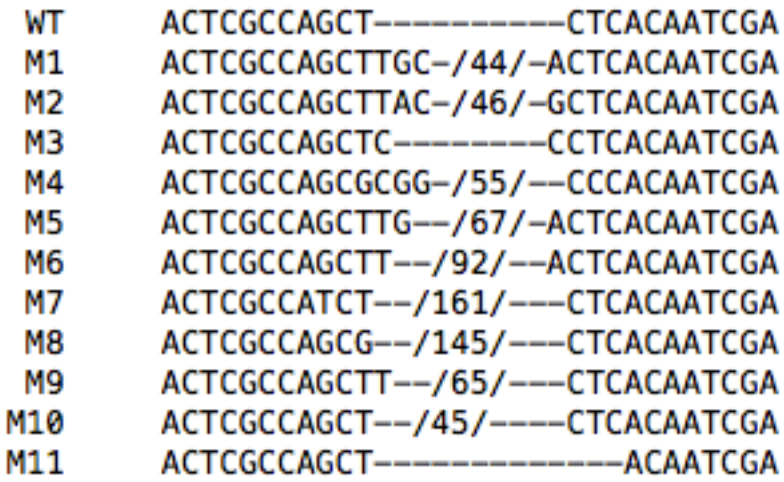


Figure 3.7. Insertions discovered at the site of CRISPR/Cas9-directed double-strand breaks

Table 3.1. NUMT content in *T. gondii* strains and *H. hammondi*

Species	Density (%)	Number of insertions	Number of basepairs
T. gondii CAST	1.39	9,380	876,895
T. gondii TgCtCo5	1.4	9,421	874,358
T. gondii GT1	1.44	9,184	874,525
T. gondii TgCATBr9	1.42	9,344	874,826
T. gondii GAB2	1.4	9,385	880,313
T. gondii FOU	1.41	9,325	871,351
T. gondii TgCATBr5	1.41	9,381	869,048
T. gondii MAS	1.42	9,329	874,616
T. gondii p89	1.41	9,326	875,326
T. gondii VEG	1.41	9,279	883,437
T. gondii ME49	1.43	9,356	891,057
T. gondii ARI	1.4	9,456	883,577
T. gondii TgCat_PRC2	1.4	9,495	880,946
T. gondii COUG	1.39	9,537	884,389
T. gondii VAND	1.41	9,342	879,564
T. gondii RUB	1.41	9,439	880,485
H. hammondi H.H.34	1.48	10,643	999,296

Table 3.2. Calculation of mtDNA mutation rate

	Average genetic distance	Divergence time	
		28 Mya	**12.7 Mya**
Cytochrome genes	0.056	1×10^{-9}	2.20×10^{-8}
23 mtDNA elements	0.174	3.11×10^{-9}	6.85×10^{-9}

Table 3.3. Calculation of divergence time between *T. gondii* and *H. hammondi*

Gene	**Average genetic distance**
ACT1	0.141
ATUB	0.114
MIC2	0.097
Mutation rate (μ)	**Divergence (Mya)**
9.6×10^{-9}	6.11
2.12×10^{-8}	2.77

Table 3.4. Age distribution of NUMTs in different genomic features

Age (in my)	CDS	UTR	Intron	Intergenic	Upflank	Downflank
T. gondii						
< 10	101	9979	36917	38689	11878	11481
10-20	1067	39859	157579	134872	44046	46874
20-30	868	41882	227630	132975	46464	39415
30-40	115	6134	35869	22517	6614	7810
< 43	0	0	138	234	122	112
Total (bp)	**2151**	**97854**	**458133**	**329287**	**109124**	**105692**
N. caninum						
< 10	452	0	61203	104833	10557	29335
10-20	0	0	70349	87868	8413	23656
20-30	65	0	24730	24972	2060	6857
30-40	0	0	2171	2237	180	619
Total (bp)	**517**	**0**	**158453**	**219910**	**21210**	**60467**

Table 3.5. NUMTs in coding regions

Chr	NUMT Start	NUMT End	NUMT length	Associated nuclear gene	Annotation	Expressed?
TGME49_chrVIIa	3381280	3381311	32	TGME49_202375	hypothetical protein	No
TGME49_chrVIIa	1530075	1530151	77	TGME49_205230	hypothetical protein	No
TGME49_chrIb	1167730	1167772	43	TGME49_209270	hypothetical protein	No
TGME49_chrIV	1878488	1878534	47	TGME49_211270	sushi domain (scr repeat) domain-containing protein	Yes
TGME49_chrV	1285277	1285398	122	TGME49_213748	hypothetical protein	No
TGME49_chrII	171240	171361	122	TGME49_221240	hypothetical protein	No
TGME49_chrII	208311	208354	44	TGME49_221280	hypothetical protein	No
TGME49_chrX	1040173	1040229	57	TGME49_227300	hypothetical protein	No
TGME49_chrX	668899	668958	60	TGME49_228020	hypothetical protein	No
TGME49_chrX	551216	551289	74	TGME49_228145	hypothetical protein	No
TGME49_chrX	4485492	4485561	70	TGME49_234590	hypothetical protein	No
TGME49_chrX	5734448	5734516	69	TGME49_237130	cytochrome b, putative	No
TGME49_chrX	5759259	5759305	47	TGME49_237160	hypothetical protein	No
TGME49_chrX	5819002	5819059	58	TGME49_237270	hypothetical protein	No
TGME49_chrX	5819129	5819195	67	TGME49_237270	hypothetical protein	No
TGME49_chrVI	2742158	2742213	56	TGME49_243590	endonuclease/exonuclease/phosphatase family protein	No
TGME49_chrXII	5515511	5515667	157	TGME49_251665	hypothetical protein	No
TGME49_chrXII	5515511	5515625	115	TGME49_251665	hypothetical protein	No
TGME49_chrIII	1241686	1241747	62	TGME49_253860	Tyrosine kinase-like (TKL) protein	No
TGME49_chrIII	2066729	2066814	86	TGME49_254950	RNA cap guanine-N2 methyltransferase	No
TGME49_chrVIII	6546639	6546685	47	TGME49_268320	hypothetical protein	No
TGME49_chrXII	6674483	6674543	61	TGME49_277490	hypothetical protein	Yes

TGME49_chrXII	6197131	6197194	64	TGME49_278250	hypothetical protein	No
TGME49_chrV	3012730	3012797	68	TGME49_283720	phosphotyrosyl phosphate activator (ptpa) protein	Yes
TGME49_chrIX	5764863	5764915	53	TGME49_306035	hypothetical protein	No
TGME49_chrIX	5964320	5964395	76	TGME49_306320	Myb family DNA-binding domain-containing protein	Yes
TGME49_chrXI	174789	174918	130	TGME49_306940	hypothetical protein	No
TGME49_chrXI	2244457	2244533	77	TGME49_311740	hypothetical protein	Yes
TGME49_chrXI	3071992	3072054	63	TGME49_313055	hypothetical protein	No
TGME49_chrIV	902820	902866	47	TGME49_319570	WD domain, G-beta repeat-containing protein	No
FR823380	1924459	1924499	41	NCLIV_002090	putative CAM kinase, CDPK family	Yes
FR823386	84702	84758	57	NCLIV_012340	conserved hypothetical protein	Yes
FR823388	3876726	3876790	65	NCLIV_023770	hypothetical protein	No
FR823389	102252	102307	56	NCLIV_023910	conserved hypothetical protein	No
FR823389	102314	102348	35	NCLIV_023910	conserved hypothetical protein	No
FR823390	6479369	6479427	59	NCLIV_038090	helicase conserved C-terminal domain-containing pr	Yes
FR823391	860464	860506	43	NCLIV_045810	Niemann-Pick type C1 disease protein/ patched like	Yes
FR823393	455258	455332	75	NCLIV_060880	conserved hypothetical protein	Yes
FR823393	3770820	3770905	86	NCLIV_064970	Phospholipid phospholipase C beta isoform,related	Yes

Table 3.6. NUMT insertion and deletion rates

Branch	Insertions	Insertion Rate (per m.y)	Deletions	Deletion Rate (per m.y.)
T. gondii	56 (5,200 bp)	9.1 (852 bp)	24 (2,161 bp)	3.9 (354 bp)
H.hammondi	62 (5,613 bp)	10.2 (920 bp)	36 (3,246 bp)	5.9 (532 bp)
T.gondii/H. hammondi	6,505 (607,705)	232.3 (21,704)	-	-
N. caninum	4,478 (378,064)	159.9 (13,502 bp)	-	-

Table 3.7. Location of NUMTs displaying differential presence/absence in 16 *T. gondii* strains

Numt #	Length (bp)	% Div	ME49 chr	Start	End	CAST	TgCTCo5	GT1	TgCATBr9	GAB2	FOU	TgCATBr5	MAS	p89	VEG	ME49	ARI	TgCat_PRC2	COUG	VAND	RUB	Location	Associated gene
1	61	14	chrIa	1018474	1018944	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	0	I	294630
2	73	20	chrIb	213492	213822	0	0	0	0	0	0	0	1	1	0	0	1	1	1	0	1	UF	207690
3	67	7.7	chrIb	1596582	1596989	1	0	1	0	0	0	0	1	0	0	0	0	0	1	0	0	I	209970
4	45	2.3	chrII	261484	261902	0	0	1	1	0	1	1	0	0	0	0	0	0	0	1	1	I	221330
5	67	1.5	chrIII	2457884	2458465	1	1	1	0	1	1	1	1	1	1	1	1	1	1	1	1	IG	
6	28	0	chrIV	1660551	1660886	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	E	211630
7	157	1.9	chrIV	2006455	2006758	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	DF	210960
8	63	6.2	chrV	1768829	1769234	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	E	286590
9	51	16	chrV	2529301	2529857	0	0	0	0	0	0	0	0	0	0	1	1	0	1	0	0	I	285210
10	52	0	chrV	2798255	2798857	1	1	1	1	1	1	0	0	1	1	0	0	0	0	1	1	I	284170
11	100	1	chrVI	706677	707077	1	1	1	1	1	1	1	0	1	0	0	0	0	0	1	1	I	239480

12	91	23.3	chrVI	2259261	2259751	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	1	E	242770
13	64	0	chrVI	2415880	2416280	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	UF	243140
14	88	26.1	chrVIIa	605108	605597	1	1	1	0	1	1	0	1	1	1	1	1	1	1	1	1	I	304650
15	133	27.8	chrVIIa	2259268	2259788	1	0	1	1	1	1	1	1	0	0	1	1	1	1	1	1	I	203760
16	54	14.6	chrVIIa	2673706	2674160	1	1	0	1	0	0	1	1	1	1	1	1	1	1	1	0	DF	203220
17	162	30.6	chrVIIa	3250336	3250897	1	1	1	1	1	1	1	1	1	0	1	1	1	1	0	1	E	202540
18	88	23.6	chrVIIa	4524882	4525210	0	1	0	0	0	0	1	0	1	0	0	1	1	1	1	1	IG	
19	59	0	chrVIIb	668724	669115	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	I	263270
20	75	14.7	chrVIIb	1025866	1026297	1	1	1	1	1	1	1	1	1	1	0	0	1	0	1	1	I	262780
21	140	26.5	chrVIIb	1710159	1710710	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	I	261490
22	53	1.9	chrVIIb	1904270	1904660	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0	IG	
23	181	23.2	chrVIIb	3218491	3219027	1	1	1	1	1	1	1	1	1	0	0	0	1	1	1	1	IG	
24	87	21.2	chrVIIb	4646913	4647351	1	1	1	1	1	1	1	1	1	0	0	0	1	0	1	1	I	255890
25	77	22.1	chrVII	2496611	2497086	1	0	0	0	0	0	0	1	1	1	1	1	1	1	0	1	IG	
26	99	23.2	chrVII	3466523	3466786	0	1	1	1	1	1	0	1	0	0	0	0	0	0	0	1	I	243482
27	56	14.3	chrVII	4435378	4435833	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	DF	271740

28	100	28.6	chrIX	3008823	3009336	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	1	I	289120
29	133	29.3	chrIX	3038491	3039017	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1	1	I	289170
30	73	20.6	chrIX	4251589	4252066	1	1	1	0	1	1	0	0	1	1	1	1	1	1	1	1	IG	
31	51	23.5	chrX	475318	475768	0	0	0	0	0	0	0	1	0	0	1	1	1	1	0	0	E	228230
32	162	17.7	chrX	2699533	2699941	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	UF	224900
33	81	21.9	chrX	2977559	2977959	1	1	1	1	0	1	1	1	1	1	0	0	1	1	1	1	I	224540
34	74	17.6	chrX	3011324	3011796	1	1	1	1	1	1	0	1	0	0	1	0	1	1	0	1	I	224510
35	52	13.5	chrX	5158557	5159008	1	1	1	1	1	1	1	1	1	1	1	1	1	1	0	1	I	235920
36	78	0	chrX	6191403	6191803	0	1	0	0	0	0	0	0	0	0	0	0	0	0	1	0	I	214260
37	84	23.2	chrX	6399526	6400012	1	1	1	1	1	1	1	0	0	0	1	1	1	1	1	1	I	214600
38	59	0	chrX	6727704	6728163	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0	0	UF	215060
39		25.4	chrX	6863854	6864151	0	0	1	0	0	0	1	0	1	0	0	0	0	0	1	0	I	215260
40	166	19.4	chrXI	1532798	1533403	1	1	1	1	1	1	1	1	0	1	1	1	1	1	1	1	DF	310710
41	239	25.8	chrXI	1603990	1604629	1	1	1	0	0	1	1	1	1	1	1	1	1	1	1	1	DF	310802
42	85	2.4	chrXI	2153288	2153772	0	0	1	1	0	1	1	1	1	0	1	0	0	0	1	1	IG	
43	48	0	chrXI	2168366	2168704	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	I	311625

44	57	5.2	chrXI	2292428	2293114	0	0	0	0	0	0	1	1	0	0	0	0	0	0	0	0	E	311800
45	102	24	chrXI	2928696	2929196	0	1	1	0	0	0	1	1	1	1	1	1	1	1	1	1	I	312820
46	37	0	chrXI	3138788	3139178	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	E	313150
47	122	24.1	chrXI	3380178	3380628	1	0	0	1	1	1	1	1	1	0	0	0	1	1	1	1	I	313430
48	58	12.1	chrXI	3550551	3551008	0	0	0	0	0	0	0	0	0	1	1	0	0	0	0	0	I	313630
49	103	25.2	chrXI	5020102	5020569	1	0	1	1	0	1	1	1	1	0	0	0	0	0	0	0	I	315845
50	38	2.6	chrXI	5133457	5133946	0	1	0	1	1	1	0	1	0	1	1	1	1	1	0	0	E	316100
51	142	22.7	chrXII	328937	329131	1	1	1	1	1	1	1	1	1	1	0	1	1	1	1	1	UF	299820
52	84	1.2	chrXII	905207	905570	0	0	0	1	1	1	1	0	1	1	0	0	0	0	0	0	I	219500
53	79	25	chrXII	2430152	2430458	1	1	1	1	0	1	1	1	1	0	0	0	0	1	1	1	I	245610
54	73	0	chrXII	3864644	3865116	0	0	0	0	0	0	0	0	0	1	1	1	0	1	0	0	I	248450
55	60	20	chrXII	4009932	4010391	0	0	0	0	1	1	0	0	1	1	1	1	1	1	1	0	UF	248580
56	35	0	chrXII	4599488	4599844	0	1	0	0	1	1	0	1	1	0	0	0	1	0	1	0	I	249530
57	41	13.7	chrXII	4641543	4642129	1	0	1	1	0	0	1	1	0	1	1	1	1	1	1	1	I	249560

Table 3.8. Enrichment analyses of genes associated with NUMTs displaying differential presence/absence

ID	Name	Bgd count	Result count	% of bgd	Fold enrichment	P-value	Benjamini	Bonferroni
GO:0003777	microtubule motor activity	30	3	10	12.09	0.002	0.037	0.062
GO:0003774	motor activity	41	3	7.3	8.84	0.005	0.037	0.142
GO:0016776	phosphotransferase activity, phosphate group as acceptor	15	2	13.3	16.12	0.008	0.037	0.221
GO:0016656	monodehydroascorbate reductase (NADH) activity	1	1	100	120.87	0.016	0.037	0.441
GO:0004087	carbamoyl-phosphate synthase (ammonia) activity	1	1	100	120.87	0.016	0.037	0.441
GO:0050521	alpha-glucan, water dikinase activity	1	1	100	120.87	0.016	0.037	0.441
GO:0033857	diphosphoinositol-pentakisphosphate kinase activity	1	1	100	120.87	0.016	0.037	0.441
GO:0016887	ATPase activity	188	5	2.7	3.21	0.019	0.037	0.509
GO:0008081	phosphoric diester hydrolase activity	25	2	8	9.67	0.020	0.037	0.543
GO:0005524	ATP binding	525	9	1.7	2.07	0.023	0.037	0.627
GO:0032559	adenyl ribonucleotide binding	526	9	1.7	2.07	0.023	0.037	0.634

GO:0030554	adenyl nucleotide binding	527	9	1.7	2.06	0.024	0.037	0.642
GO:0035639	purine ribonucleoside triphosphate binding	621	10	1.6	1.95	0.024	0.037	0.654
GO:0016781	phosphotransferase activity, paired acceptors	2	1	50	60.43	0.024	0.037	0.659
GO:0008889	glycerophosphodiester phosphodiesterase activity	2	1	50	60.43	0.024	0.037	0.659
GO:0032553	ribonucleotide binding	622	10	1.6	1.94	0.024	0.037	0.661
GO:0032555	purine ribonucleotide binding	622	10	1.6	1.94	0.024	0.037	0.661
GO:0017076	purine nucleotide binding	623	10	1.6	1.94	0.025	0.037	0.668
GO:0016787	hydrolase activity	740	11	1.5	1.8	0.030	0.040	0.816
GO:0008716	D-alanine-D-alanine ligase activity	3	1	33.3	40.29	0.032	0.040	0.876
GO:0008859	exoribonuclease II activity	3	1	33.3	40.29	0.032	0.040	0.876
GO:0008479	queuine tRNA-ribosyltransferase activity	3	1	33.3	40.29	0.032	0.040	0.876
GO:0000166	nucleotide binding	665	10	1.5	1.82	0.037	0.044	1.000
GO:0036094	small molecule binding	692	10	1.4	1.75	0.047	0.048	1.000
GO:0017111	nucleoside-triphosphatase activity	324	6	1.9	2.24	0.048	0.048	1.000

GO:0003993	acid phosphatase activity	5	1	20	24.17	0.048	0.048	1.000
GO:0016655	oxidoreductase activity, acting on NADH or NADPH, quinone or similar compound as acceptor	5	1	20	24.17	0.048	0.048	1.000

CHAPTER 4

INSIGHTS INTO THE ABNORMALLY LARGE GENOME OF THE APICOMPLEXAN PARASITE *SARCOCYSTIS NEURONA*

Namasivayam, S., Dangoudoubiyam, S., Jaromczyk, J., Yeargan, M., Gautam, A., Bullock, T., Mahmud, O., Korunes, K., Parkinson, J., Grigg, M., Schardl, C. L., Kissinger, J. C., and Howe, D. K. To be submitted to Microbial Genomics.

ABSTRACT

The coccidian parasite *Sarcocystis neurona* is a leading cause of infectious neurologic disease in horses and is a pathogen of marine animals. Evolutionarily, *S. neurona* is intermediate between the *Toxoplasma* and *Eimeria* lineages, ~125-200 MY from each. The *S. neurona* genome was sequenced using 454 pyrosequencing and Sanger sequencing of fosmid ends resulting in 172 scaffolds with an assembled genome size of 124 Mb. 454 and Illumina transcriptome data sets were generated for multiple time points. We predicted ~6,936 protein coding genes, slightly fewer than other coccidians. The most notable finding is the unusually large nuclear genome size, twice as large as other coccidian genomes and the largest among the sequenced apicomplexans. No new or overly expanded gene families were detected, but some gene families were reduced including the major apicomplexan transcription factor family, ApiAP2. *S. neurona* genes are similar in structure to *Toxoplasma gondii* genes with an average of five exons/gene but the introns average about thrice the size. Coding regions are also slightly larger. Repeats are partly responsible for these larger genomic features. *S. neurona* has approximately five times the repeat content of *T. gondii*. Nearly 40% of the genes are shared with *T. gondii* and *Neospora caninum*. Only 13% of the genes are shared with *Eimeria tenella.* Some synteny exists with *T. gondii*, but none with *E. tenella*.

INTRODUCTION

The Apicomplexa are a phylum of protozoan parasites with over 5,000 described species (1). Apicomplexan parasites are responsible for many devastating diseases, including malaria caused by *Plasmodium spp*, cryptosporidiosis caused by *Cryptosporidium spp.* and coccidiosis in poultry caused by *Eimeria spp*. Given their medical and veterinary importance, it is reasonable that the genomes of a number of these apicomplexan species have been sequenced. The apicomplexans have comparatively reduced genomes, reflective of their parasitic lifestyle (2-4). The genomes are organized into 4 – 14 chromosomes and range from only ~8.5 Mb in *Theileria* and 65 Mb in *Toxoplasma* to >125 Mb for *Sarcocystis neurona*, reported here (Figure 1.1, Table 1.1). The gene content also varies from ~4,000 genes in *Cryptospordium spp.* to ~9,000 genes in *Toxoplasma gondii*. The genomes of some *Plasmodium spp*. are >80% A-T rich (5) whereas others like *T. gondii,* have a G-C content of ~53%. Thus, there is a wide variation in genome size, content and composition across the phylum.

The parasites of the coccidian lineage have some of the largest genomes observed within the Apicomplexa (6, 7), averaging ~60 Mb. These parasites exhibit varying tissue and host tropisms (8). Unlike the "tissue-cyst forming" coccidians in the family Sarcocystidae, species of the genus *Eimeria* utilize only a single host in which they undergo both sexual and asexual phases of their lifecycle and they limit infection to gut tissues (8). Members of the family Sarcocystidae, however, including *Toxoplasma*, *Hammondia*, *Neospora* and *Sarcocystis* have a two-host life cycle which includes a wide range of intermediate hosts in which asexual replication occurs (8). Notably, *Toxoplasma gondii* can invade virtually any nucleated cell type in almost all warm-blooded animals and it has bypassed the need for a sexual stage via the evolution of oral infectivity (9). Genome sequences for *T. gondii*, *Hammondia hammondi,*

Neospora caninum and a number of *Eimeria* species have been published. We herein generate and provide (ToxoDB.org) the first publicly available genome sequence and annotation of *Sarcocystis neurona*.

Sarcocystis neurona belongs to the genus *Sarcocystis*, which consists of over 100 species collectively capable of infecting mammals, reptiles, birds and fish (10, 11). *S. neurona* causes the neurologic disease equine protozoal myeloencephalitis (EPM) in horses and is an emerging pathogen of marine animals (11). Its definitive host is the opossum and it has a wide range of intermediate hosts including skunks, racoons and armadillos (12, 13). Horses are an aberrant or dead-end intermediate host from which the parasite is generally not transmitted. Evolutionarily, *S. neurona* is roughly midway between *Eimeria tenella* and the *Toxoplasma, Hammondia* and *Neospora* grouping (9).

S. neurona exhibits some interesting biology. Unlike many apicomplexans, *S. neurona* does not form a parasitophorous vacuole, a protective interface, during the intracellular stage (11, 14). Instead it is in direct contact with the host cell cytoplasm. Rhoptry organelles, which are important for secreting proteins at the time of invasion (15, 16), are absent in the extracellular merozoite stage (17). During intracelluar development, the parasite divides by endopolygeny, in which the mother cell undergoes multiple rounds of DNA replication resulting in a polyploid nucleus before nuclear division and cytokinesis occurs (18). This mode of division is different from the endodyogeny that occurs in the asexual stages of *T. gondii* (19) or the schizogony that occurs in *Plasmodium* and *Eimeria* (20). Comparative analyses with genomic and transcriptomic sequences for *S. neurona* will greatly enhance veterinary, biological and evolutionary studies.

Early comparisons of propidium-iodide stained *S. neurona* merozoites to *T. gondii* tachyzoites using flow cytometry (21, 22), revealed a larger genome size for *S. neurona.* It was

estimated to be 105-108 Mb (Daniel Howe, personal communication). This is consistent with the genome size estimate of *Sarcocystis cruzi* (23). The estimated genome size is almost twice as large as that of *T. gondii*, and much larger than the genomes of most other apicomplexans (Figure 1.1, Table 1.1), and the karyotype is still unknown (discussed below).

Polyploidy and repetitive DNA (transposable elements in particular) are important contributors to genome size and evolution (24-27). The human genome is predicted to be 66%-69% repetitive (28). The unicellular eukaryote, *Amoeba dubia*, has 200 times more DNA content than the human genome because of the repetitive nature of its genome (27). TE content can range from only ~3% of some small plant genomes up to >85% of some large genomes, making genome size directly proportional to the TE content (29). Ploidy or whole genome duplication also contributes to genome size. The current *Sacchromyces cerevisiae* genome is proposed to have derived from a tetraploid genome (30). Polyploidy is well described in flowering plants (31).

In the case of unicellular parasites, the kinetoplastid *Trypansoma cruzi* genome has up to 30% repetitive DNA (32) and *Giardia lamblia* is polyploid (33). Two-thirds of the ~160 Mb *Trichomonas vaginalis* genome is made of repeats and transposable elements, including ~1000 copies of the first *mariner* element discovered outside animals. It also contains ~152 cases of potential prokaryote-to-eukaryote lateral gene transfers (34). Apicomplexan genomes are generally not very repetitive. *P. berghei* (35) contains only ~5% repetitive DNA. Transposable elements have been reported in only a few apicomplexans, including the coccidian *Eimeria* (7, 36). *Eimeria* contains long terminal repeat (LTR) retrotransposons from a group of chromoviruses, although these LTR elements do not appear to be active (7, 36). Unlike other studied apicomplexans, *Eimeria* chromosomes show a bipartite structure with repeat-rich and

repeat-poor regions (7, 36). The different *Eimeria* species also show variation in their genome sizes, ranging from 45 Mb – 72 Mb (7) (ToxoDB.org).

Apicomplexans also show lineage- and species-specific genes and an expansion of multi-gene families (37-39). The expanded gene-families encode proteins associated with host cell invasion and immune evasion. These include SAG, MIC, ROPs, RON and GRA (6, 40-42). For example, *T. gondii* and *N. caninum* contain 104 and 227 surface antigen domain (SAG) containing genes, respectively (6). The number of SAG genes varies across *Eimeria* species, ranging from 19 in *E. praecox* to 172 in *E. mitis* (7). In our published manuscript of the *S. neurona* SN1 strain draft genome, the gene content and metabolic pathways have been discussed in detail (43). In brief, we find that most of the invasion machinery is conserved across the Coccidia. However, many dense granule proteins and rhoptry kinases that are involved in altering host pathways are absent. Absence of some of the rhoptry kinases is not too surprising since *S. neurona* does not form the rhoptry organelle in its extracellular stage. Examination of metabolic pathways revealed the ability of *S. neurona* to use alternative metabolic pathways. In this manuscript, we provide an analysis of the *S. neurona* SN3 genome, providing insights into its abnormally large genome, and set the stage for population biology studies in this parasite.

MATERIALS AND METHODS

Parasite cultivation

S. neurona SN3 strain parasites isolated from an infected horse in Panama (44) were cultured in bovine turbinate cells and merozoites were harvested for genome sequencing. For transcriptome sequencing, extracellular merozoites and intracellular schizonts were harvested at 2 h, 8 h, 24 h, 53 h and 72 h time points after infection and total RNA was purified using Trizol

(Life Technologies, Grand Island, NY). Poly(A) RNA enrichment was performed using NucleoTrap mRNA (Macherey-Nagel, Bethlehem,PA).

Genome sequencing and assembly

3 Kb and 8 Kb paired-end libraries were prepared using the GS FLX titanium rapid library preparation method and sequenced on the GS FLX genome sequencer at the University of Kentucky's Advance Genetic Technologies Center. Paired-end sequencing of fosmid clones was performed using Sanger sequencing technology. A hybrid assembly of the 454 and Sanger reads was performed using Newbler 2.5.3 with the -large parameter turned off. All other parameters were default. UMD 3.1 assembly of *Bos taurus* nuclear and mitochondrial genomes and transcriptome sequences from UniGene were used as a vector screening file to screen out host cell contaminates. BLASTN searches were performed on the assembled scaffolds and contigs to further eliminate host cell contamination. 6 scaffolds were eliminated as bovine. Assembly statistics are summarized in Table 4.1. The genome sequence is available in GenBank (Accession number: JAQE00000000).

Transcriptome sequencing and assembly

454 sequencing and assembly

Available cDNA libraries (Howe lab) prepared from the merozoite stages of the SN3 and SN4 strains were used for EST sequencing. Recovered mRNA from different time points were used for 454 rapid cDNA library constructions and sequenced using the 454 FLX-titanium genome sequencer. Newbler v2.5.3 was used for assembly. The *Bos taurus* database, described above, was used to screen for contamination during and after assembly. Transcripts that hit a *Bos taurus* sequence at 80% coverage and did not hit the *S. neurona* genome at >50% coverage via BLASTN were removed.

Illumina sequencing and assembly

Initially, 100 bp paired-end unstranded Illumina sequencing was performed on merozoite mRNA. Subsequently, strand-specific sequencing was performed using mRNA from several schizont time points (2 separate runs, 2 and 8h time points were combined) and the merozoite time point (1 run). Library insert size was 150-300 bp. Sequencing was performed at the Duke University sequencing facility. Reads were pre-screened for removal of sequencing adaptors and indices. Reads were further screened using Trimmomatic (45) and only reads with a sliding window and an overall quality score of 30 were retained. Each data set was assembled using the reference-based TopHat-Cufflinks pipeline (46, 47). The assembled 172 *S. neurona* SN3 genome sequence scaffolds and 701 large contigs were used as the reference. Minimum and maximum intron sizes were adjusted to 30 and 5000 bp respectively. Distance between mate pairs was set to 50 for the un-stranded library and 100 for the stranded libraries to reflect the appropriate library insert size. Stringent mapping criteria were used such that a read was accepted only when both mate pairs mapped. For stranded libraries, library type –fr-secondstrand was used. Data from all the time points were merged in two ways. First, a combined assembly was generated by merging the raw reads. Second, assembled transcripts from each of the strand-specific data sets were merged using the Cuffmerge program from the Cufflinks suite of programs. The assemblies were assessed by comparison to the *T. gondii* representatives of conserved apicomplexan genes. Neither assembly was exceedingly better or worse than the other. For simplicity, Cuffmerge transcripts were used in the annotation pipeline.

A *de novo* assembly using the Trinity pipeline (48) was performed to account for any incompleteness in the reference genome. Assembled transcripts were screened for bovine contamination as described above. Trinity is known to over predict the number of transcripts

(48). The RSEM pipeline (49) was used to obtain expression (FPKM) values for the predicted transcripts and filter out transcripts expressed at a FPKM value of < 1. The number of transcripts from all transcriptome assemblies is summarized in Table 4.2.

Gene prediction and annotation

Gene predictions were attempted using *S. neurona*-trained GlimmerHMM, Genemark-ES, Augustus and SNAP algorithms (50-53). The annotation pipelines PASA, EVM and MAKER v2.31.7 (54, 55) were explored. The MAKER pipeline was determined to produce the better predictions (evaluated by comparison to the conserved apicomplexan gene set). The following were provided as input to MAKER; (i) transcriptome evidence: Cuffmerge transcripts of all strand-specific assemblies; (ii) protein evidence: the best Uniprot protein hit at each location of the *S. neurona* SN3 genome and the predicted proteins from *S. neurona* SN1 scaffold 1; (iii) SNAP trained with *S. neurona* SN1 scaffold 1 proteins and (iv) the existing model of Augustus trained with *T. gondii*. RepeatMasker was run as part of the MAKER pipeline with 'Alveolata' as model organism and all TE proteins as repeat library. All *S. neurona* SN1 proteins were included as evidence in a separate round of MAKER predictions. Both predictions were compared to the conserved apicomplexan orthologs, *T. gondii* proteins and manually annotated *S. neurona* genes. The predictions from the first iteration of MAKER run were better, except for ~300 genes that had better predictions in the second iteration of the MAKER run. These 300 genes were added to, or replaced, in the initial MAKER run. Retraining with MAKER did not improve the predictions. A local installation of WebApollo (56) was used to load all the evidence tracks and annotations for manual verification and curation as necessary. Approximately 200 genes had unsupported non-canonical splice junction predicted and were manually curated. The

BLAST2GO pipeline was used for functional annotation. tRNAscan-SE (57) was used to predict tRNA genes. Annotations have been submitted to ToxoDB.org (58).

Annotation of the apicoplast and mitochondrial genomes

The apicoplast genome sequence was identified and annotated using other apicomplexan apicoplast genome sequences. Scaffolds containing hits to known apicoplast genes were screened for a length of 25-35 Kbp. A single scaffold of 34,488 bp was identified. This scaffold consisted of two contigs of length 5,780 bp and 24,002 bp joined by a paired-end read creating a gap of 4,706 bp. The smaller contig had twice the read coverage when compared to the larger contig. Comparison to the *T. gondii* apicoplast genome sequence revealed it was part of an inverted repeat that was collapsed during assembly. tRNA content was predicted using tRNAscan-SE using default parameters and Mito/Chloroplast as source. Apicoplast-encoded protein sequences from *P. falciparum* (PlasmoDB.org (59), *T. gondii* (ToxoDB.org) and *E. tenella* (GenBank (NC_004823.1)) were used to predict and annotate *S. neurona* apicoplast proteins. Proteins sequences were manually curated to account for the use of the "UGA" codon as amino acid tryptophan rather than a stop codon (a common feature of apicoplast genome codon usage). LSU rRNA and SSU rRNA from *T. gondii* were used to annotate these ribosomal subunits.

The 23 mtDNA elements and cytochrome genes from *T. gondii* (unpublished, Chapter 1) and *E. tenella's* mitochondrial genome (AB564272.1) were used to annotate *S. neurona's* mitochondrial genome. Neither BLASTN nor TBLASTX found any significant mitochondrial hits in the assembled *S. neurona* genome sequence. BLASTN and TBLASTX searches against the Trinity assembled transcripts identified the three expected cytochrome genes. Only partial matches to the 23 mtDNA elements and the rRNA fragments could be identified even in the Trinity transcripts. RepeatMasker v4.0.5 (http://www.repeatmasker.org) with search engine 'crossmatch'

and the organellar sequences as libraries was used to identify NUMTs and NUPTs. Since only the cytochrome genes of the mitochondrial genome could be annotated, *E. tenella's* mitochondrial genome and the *T. gondii* mtDNA elements were also used as input to RepeatMasker.

Identification of repeats

RepeatModeler v1.0.8 (http://www.repeatmasker.org) and RepeatMasker v4.05 were used to identify repeats. RepeatModeler was run with default parameters for *de novo* identification of repeat families. Families with 15 or more members were used as input for RepeatMasking the genomes.

Comparative analyses

Comparison between S. neurona strains SN3 and SN1

Whole genome alignments of the SN3 genome with itself and with the genome of *S. neurona* SN1 strain were generated using MUMmer (60) with default parameters. SN3 vs SN1 alignments were filtered to remove hits to repetitive regions. The show-snps program was used for a quick survey of SNPs between the two genomes. BLASTP was used to compare the annotated proteins of the 2 strains. Hits were filtered for identity of 90%, E-value of 1e-10 and coverage of at least 50%. These identified hits were used to identify syntenic blocks (described below).

ORTHOMCL clustering

Annotated proteins from the *E. tenella* Houghton strain (version 2013-11-05), *T. gondii* ME49 strain (version 2013-04-23), *N. caninum* LIV strain (version 2011-08-12) and *H. hammondi* H.H34 strain (version 2014-09-03) were downloaded from ToxoDBv10. Orthologous groups were defined using WUBLAST v2.26 (http://blast.wustl.edu) and OrthoMCL (61). An

all-by-all BLASTP was performed with the following parameters: E=1e-30 B=1000000 V=1000000 hspmax=0. OrthoMCL uses Markov clustering to identify orthologous clusters from the BLASTP output.

Synteny and visualization

The OrthoMCL output was formatted to represent pairwise matches in each orthologous cluster. MCSCAN v0.8 (62) was used to generate syntenic blocks between all combinations of genomes as described in (3). For a region to be considered syntenic between two genomes, a minimum of 3 orthologous genes are required to be in the same order and orientation. An intergenic space of up to 5 Kb was allowed. Circos v0.51 (63) was used for the visualization of the syntenic blocks. The MCSCAN output was formatted appropriately for input to Circos.

Detection of antisense transcripts

For each Cufflinks transcriptome assembly, a BLASTN analysis with self and between time points was performed. Hits were filtered for > 98% identity and an E-value of at least 1e-10. A transcript was called an antisense of another transcript if (i) the two transcripts are from the same genomic locus but on opposite strands and (ii) the shorter transcript overlaps the longer transcript by at least 20%. FPKM (expression) values of the transcripts were obtained from the Cufflinks output.

Identification of ApiAP2 DNA-binding domains

A multiple sequence alignment of ApiAP2 proteins was obtained from ((64), personal communication)). The alignment includes ApiAP2 proteins from *P. falciparum, P. vivax, T. gondii, C. parvum* and *C. hominis*. HMMER v3.1b1 (65) was used to generate the ApiAP2 HMM. The HMM built using ApiAP2 domains was more sensitive at detecting apicomplexan AP2 domains than the generic AP2 HMM from PFAM (64). This ApiAP2 HMM was used to

identify ApiAP2 proteins and domains in the proteomes of *S. neurona*, *E.tenella* and *H. hammondi. T. gondii* and *N. caninum* proteomes were used as control. In Oberstallar *et. al*, hits were filtered for an E-value of 1e-10. However, all predicted domains with an E-value greater than 1e-10 have been manually validated as ApiAP2 proteins in *T. gondii* (Michael White, ToxoDB.org). Therefore an E-value cut-off was not used in this study.

RESULTS AND DISCUSSION

S. neurona has an unususally large genome

Apicomplexan genomes are much smaller when compared to most other unicellular eukaryotes (3). Apicomplexan genome sizes (until this study) range from 8 – 65 Mb and are organized in to 4 – 14 chromosomes. The genome size and the karyotype of *S. neurona* were unknown prior to this study. To determine the karyotype, we attempted Contour-clamped homogeneous electric field (CHEF) electrophoresis, extensively, as described in (66). These conditions were not able not resolve the chromosomes. The DNA was retained in the wells suggesting that the *S. neurona* chromosomes are too large to enter the gel body.

Whole genome sequencing using 454 and Sanger sequencing technology generated ~125 Mb of assembled genome sequence. The assembly software estimates a genome size of 151.2 Mb based on the peak intensity of the read alignment histogram. Thus, the *S. neurona* genome is roughly twice as large as that of *T. gondii*, the second largest genome among the sequenced apicomplexans. The genome assembled into 172 scaffolds and 701 contigs greater than 500 bp in length (Table 4.1). The largest *S. neurona* scaffold is 11.3 Mb, which is larger than the entire genome of some apicomplexans like *Cryptosporodium* and *Babesia.*

Given that the *S. neurona* genome is twice as large as that of *T. gondii*, we suspected whole genome duplication. We addressed this question in two ways. First, the genome was

aligned to itself to identify duplicated regions (Figure 4.1); we did not detect any evidence for whole genome duplication. Second, we looked at the copy number of genes that exists as single copy orthologs in all apicomplexans. *S. neurona* orthologs of these genes occur as single copy genes in *S. neurona* as well. Taken together, these results suggest that genome duplication was not responsible for the large genome size.

NUMTs and NUPTs do not contribute to the genome size

We have previously shown that NUM/PTs contribute up to 1 Mb of the *T. gondii* genome. To assess the contribution of these organellar insertions to the genome size of *S. neruona*, we first identified and annotated the organellar genomes. The apicoplast genome had assembled in to a 35 Kb scaffold and was annotated as described in Methods.

An assembled mitochondrial genome could not be identified in the assembled scaffolds or contigs. BLASTN and TBLASTX searches using either *T. gondii* or *E. tenella* mt DNA sequences and cytochromes gene sequences did not yield any significant hits. Next, we searched the *de novo* based transcriptome assemblies (see Methods). From the trinity assembly, we identified one transcript cluster with 16 alternative splice forms that contained incomplete *coxI*, *coxIII* and *cob* transcripts (Figure 4.2) and parts of the *T. gondii* mtDNA elements. The absence of an assembled mitochondrial contig and the presence of the mtDNA elements in the transcriptome, similar to what is observed in *T. gondii* suggest the *S. neurona* mitochondrial genome maybe similar to that of *T. gondii*.

The *S. neurona* apicoplast genome was used to search the nuclear genome for NUPTs. The mitochondrial genome of *E. tenella*, the 23 mtDNA sequence elements and cytochrome genes of *T. gondii* and the identified cytochrome genes from *S. neurona* were used to identify NUMTs. NUMTs and NUPTs make up only 0.03% and 0.02% of the nuclear genome

respectively. The NUM/PT content in *S. neurona* is similar to that of *E. tenella* and considerably lower than *T. gondii* or *N. caninum* (Table 4.3). The non-homologous end joining (NHEJ) pathway proteins that are considered to be responsible for NUMT acquisition are present in *S. neurona*. Based on observations in *T. gondii* and its close relatives, one would have expected the NUM/PTs to play a significant role in the genome size of *S. neurona* but that is not the case.

Intron length is an important factor contributing to a larger genome

The genome was annotated with the aid of transcriptome data, proteins from UniProt and *ab initio* gene predictors as described in Methods. We identified 6,936 protein-coding genes, 104 tRNA genes and 3 partial rRNA genes in *S. neurona*. The lack of rRNA genes is due to a gap in the assembly similar to what is observed in *T. gondii* (Unpublished) and is believed to represent the site of a large cluster of rRNA genes. The number of annotated protein-encoding genes is similar to the other coccidians (Table 4.4), although this is likely a slight underestimate (see below). 87% of the predicted genes have transcript evidence, but 20-30% of the reference-based transcriptome assembly does not match a predicted gene. While it is possible some of the transcripts are read-through transcripts, a number of these transcripts match genes in the other coccidians (compare Table 4.5 and Table 4.6). We did not find any unusually expanded gene families among the predicted genes or the transcriptome data. Therefore, the gene content itself is not a significant contributor to a larger genome sequence.

We compared the genes of *S. neurona*, *T. gondii* and *E. tenella.* While the number of exons/gene was conserved between the three species, gene sequences were much larger in *S. neurona* (Table 4.4). Introns in *S. neurona* average 1,356 bp in length, which is almost 3X longer than the introns in *T. gondii* (Figure 4.3A). *S. neurona* transcripts are also over twice as long as *T. gondii* transcripts suggesting UTRs and and in some cases coding regions are also longer.

Based on the available UTR annotations, UTRs in *S. neurona* are 3 times longer. The intergenic space between genes is also much greater in *S. neurona* when compared to *T. gondii* (Figure 4.3B).

Interestingly, the predicted *S. neurona* proteins are also longer (Table 4.7). Examination of some orthologous proteins via multiple sequence alignment revealed insertions in coding exons (Figure 4.4). These insertions are supported by expression evidence and are often repeats of amino acids (discussed below). Proteomics data will need to be examined to confirm that these insertions are indeed real. Similar types of insertions are observed in *P. falciparum* and *E. tenella* but the location of these insertions is not conserved

Repeats are responsible for larger introns and proteins

To identify the source of the comparatively larger genomic features in *S. neurona* we studied their composition and looked for repetitive elements. Repeats, in particular TEs are important contributors to genome size in many eukaryotes. However, apicomplexan genomes generally have a low repeat content and a dearth of TEs, except for the gregarine, *Ascogregarina* (67) and the genus *Eimeria* whose genomes contain the chromovirus family of retrotransposons and a large number of various other repeats (7).

RepeatModeler identified 112 repeat families in *S. neurona* that masked 22.96% (28.4 Mb) of the genome (Table 4.8). 91 families were identified in *T. gondii* that masked 4.8% of the genome and we identified 141 families in *E. tenella* that masked 28% of the genome. The average length of a repeat in *S. neurona* is 148 bp in comparison to 95 bp in *T. gondii*. The CACTA-Mirage-Chapaev (CMC) family of Type II DNA transposons make up ~10% (12.8 Mb) of the *S. neurona* genome. However, no active transposase gene was detected suggesting these DNA transposons are ancient and are no longer active (43). Simple repeats make up 5% of the

genome, with $(AT)_n$ and $(CAG)_n$ being the most common. Repeat finding and repeat masking programs tend to not identify repeats of a very low copy number; the RepeatModeler program identified 112 repeat families with 15 or more copies. It is possible *S. neurona* has a higher repeat content.

Nearly 40% of the introns consist of repeats (Figure 4.3C). 70% of the CMC family DNA elements are present in the introns. 30% of the coding content is repetitive with 10% being simple repeats. These repeats likely account for the larger coding sequences in *S. neurona* (Figure 4.4). Homopolymeric amino acid repeats (HAARs) have been reported in *Eimeria* proteins (7). $(GCA)_n$ and $(CTG)_n$ are the most common short tandem repeats (STRs) in *S. neurona* making up 1.72 % and 1.56% of the coding sequence respectively. Different STRs are found in *S. neurona* proteins, unlike *Eimeria* proteins which primarily contain $(CAG)_n$. Interestingly, the most commonly found STRs are not conserved with *Eimeria* or *Toxoplasma* (7). However, this higher repeat content in *S. neurona* only partially accounts for the 2X larger genome.

The S. neurona genome is likely even larger

We were able to identify only 965 of the 1088 core apicomplexan orthologous proteins in the *S. neurona* genome sequence. While it is possible *S. neurona* is missing some of these genes, it is more likely that the genome sequence is incomplete. Some of the genes may be located in contig gaps and scaffold breaks. We used the *de novo* assembled transcriptome to further assess genome completeness. 20% of the transcripts from the *de novo* assembly did not map to the genome sequences. Of these, only 3% of the transcripts had hits to the NCBI protein database, which is expected since many of the coccidians proteins are lineage-specific (6). However the *de*

novo based Trinity assembler is also known to over predict transcripts and many of these transcripts may not be real (48).

Other types of artifacts may also be present. The *T. gondii* ME49 reference genome is missing copy number variants (CNVs) of some of its virulence genes. We now know that several gene and segmental duplications have been collapsed in the reference genome assembly (42). The rRNA gene cluster is missing from the genomes of both species (Unpublished). The Newbler assembler estimates a genome size of 151.2 Mb. Taken together; these pieces of evidence suggest that the *S. neurona* genome could be larger than the currently existing 125 Mb of sequence.

Comparative studies with other coccidians

The S. neurona SN3 and SN1 genomes are 99% identical

A recent study that compared 62 strains of *T. gondii* identified > 800,000 SNPS in the 65 Mb genome (42). More importantly, there were diversifications and copy number variations in important virulence factors. To look for such differences in *S. neurona*, we compared the SN3 and SN1 strains (43). The SN1 genome assembled into 116 scaffolds with a genome size of 127 Mb. We find that the two genome sequences are > 99% identical. Only 292 SNPs were identified between SN1 and SN3 (Michael Grigg, personal communication). This number is astounding small when compared to the > 800,000 SNPs across *T. gondii* strains.

Similar to our findings in the SN3 genome, the SN1 genome also has longer introns and a higher repeat content predicted. However, there are differences in the genome assemblies (Figure 4.5A). For example, the 9 Mb (and largest) scaffold of SN1 does not align in entirety to the 11.3 Mb scaffold of SN3. These differences in the assembly are further highlighted when visualizing the synteny between the two genomes (Figure 4.5B). It is not yet clear if these

breaks represent genuine rearrangements between the two genomes or if they are simply assembly artifacts. To address this question, we are currently using the PacBio SMRT sequencing technology to generate long reads that should span across these putative break points or assembly artifacts.

A similar number of genes were predicted in both strains. However only 65% of the genes are conserved between the two annotations. A greater percentage of the SN3 predictions match the transcriptome data and other coccidian proteins (Table 4.6, Table 4.7, Table 4.9). The SN3 proteins are also longer than the SN1 proteins (Table 4.7). Many of these differences in the genes and protein lengths between the two strains are likely a reflection of annotation differences and since the SN3 genes and proteins were annotated using RNA evidence we have used the SN3 predictions for comparative studies. An important biological difference is variation in the number of surface antigen domain-containing proteins. SAGs are important for cell adhesion and immune evasion. Using custom HMMs (68) 24 SAGs genes, were identified in SN1. We manually identified only 14 SAGs in SN3.

Only 36% of annotated genes are conserved across the coccidians

A large proportion of the apicomplexan genes are lineage- or even species-specific. We have identified only ~1000 genes as conserved across the entire phylum. While there is some gene innovation, this high level of specificity is due to lineage- or species-specific retention relative to the last free-living ancestor, since hits outside the apicomplexan have been identified (38). 38% (2,626) of the 6,941 predicted *S. neurona* SN3 proteins are classified as hypothetical. Using BLASTP we find 54% (3,769) of the proteins have hits to apicomplexan proteins, with coccidian proteins being the top hit for 53% (3,687) of the proteins. With a BLASTP cut-off of 40% identity, 50% coverage and an E-value of 1e-6, we find a few more hits to *N. caninum* than

to *T. gondii* (Table 4.5). Only 13% (908) of the proteins have matches in *E. tenella*. OrthoMCL analyses reveal a similar trend but identified more orthologs due to the more sensitive algorithm it uses (Figure 4.6B). Based on OrthoMCL analyses 36% of the genes were conserved across the phylum. Using the orthologs found in all 5 sequenced coccidian genomes, we identified and plotted synteny (Figure 4.6B) *T. gondii*, *H. hammondi* and *N. caninum* genomes are highly syntenic. There is synteny, albeit with significant rearrangement, between these genomes and *S. neurona*. No synteny is detected with *E. tenella*, in spite of >2500 groups of orthologous genes. Synteny is not conserved across the phylum, but is generally conserved with in a lineage (3). Here we see a lack of synteny within the lineage, which is likely because greater evolutionary time (>200MY divergence), the abnormally large *S. neurona* genome size and the repeat-rich and repeat-poor characteristic of the *Eimeria* chromosomes.

Of the 2,604 *S. neruona*-specific ortholog groups identified through OrthoMCL, 5 groups contain 2-copy paralogs. The remaining groups are *S. neurona*-specific single-copy genes. An enrichment analysis of the *S. neurona*-specific genes did not reveal any biological process that could be specific to *S. neurona* (Figure 4.7). No expanded *S. neurona*-specific gene families were identified; again highlighting that gene content itself is not responsible for the larger genome. In fact, some known virulence factors are greatly reduced in *S. neurona*. We were able to identify only 14 SAG/SRS genes. In comparison, *T. gondii*, *N. caninum* and *E. tenella* contain 104, 227 and 89 SAG proteins respectively (6, 7). It is possible the SAG domains in *S. neurona* are very different from the other coccidians. Using custom-built HMMs, only 24 SAG/SRS genes were identified in the SN1 strain. Homologs of other genes involved in virulence, namely those targeted to the dense granules, rhoptries and micronemes also show slight variation (43).

Gene Prediction

S. neurona encodes fewer ApiAP2 DNA-binding domains than T. gondii

Apicomplexans show tight transcriptional regulation (69). However, there was a dearth of recognizable transcriptional enhancers until the discovery of the apicomplexan AP2, ApiAP2, transcription factor family (70). These ApiAP2 transcription factors have now been implicated in many, if not most, regulatory roles including the intraerythrocytic developmental cycle (71) and the gametocyte differentiation cascade (72) in *Plasmodium* parasites, and cell cycle regulation in *T. gondii* (73). Evolutionarily, ApiAp2 protein domains cluster distinctly from the AP2 domains found in the plants, stramenophiles, ciliates and dinoflagellates (64). It is thought that there were a few progenitor domains in the perkinsid/apicomplexan ancestor and the ApiAP2 domains expanded independently in each lineage after the split. Within the apicomplexans, some ApiAP2 domains are conserved across the phylum but lineage-specific amplifications have occurred, often within the same protein (some proteins contain 9 AP2 domains). In the *Plasmodium* and coccidian lineages, between 50-80% of the ApiAP2 domains are thought be lineage-specific (64).

We surveyed the *S. neurona* proteome for ApiAp2 domains (the rest of the protein is not conserved even within the same species) using the ApiAP2 HMM (64). We identified 41 ApiAP2 proteins with 75 domains (Table 4.10). The number of annotated ApiAP2 proteins and domains in *S. neurona* is lower than in *T. gondii* but similar to *E. tenella*. We found putative orthologs for 35 out of 41 SnApiAP2 domains in *T. gondii*. It is difficult to identify AP2 orthologs since only the AP2 domains in the proteins are conserved and the rest of the protein is divergent. Except for *E. tenella*, the other four coccidians contain a 5-domain AP2 protein. *T. gondii* and *H. hammondi* each contain a 9-domain AP2 protein and *N. caninum* and *S. neurona*

contain an 8 domain protein. Such a large multi-domain protein is absent in *E. tenella*. These findings are, of course, dependent on the quality of the annotation.

To assess the role of the SnApiAP2 domain-containing proteins in stage-specific gene regulation, we looked for the *SnAP2* proteins in the stage-specific transcriptome data sets (Figure 4.8). 61 of the 68 TgApiAP2's were expressed in the tachyzoite stage (6). Similarly, we found 40 out of the 41 SnApiAP2's to be expressed in the schizont-merozoite stages. Unlike *Plasmodium* and *Cryptosporodium* species that show tight regulation and a cascade of ApiAP2 expression, the coccidians appear to express the majority of their ApiAP2's at all time points (at least in the asexual developmental stages examined).

Two SnApiAP2s show expression in the merozoite stage and three SnApiAP2 show expression at the 72h schizont time point. However, the expression values of these SnApiAP2s at their respective time points are not considerably higher and corresponding *T. gondii* orthologs do not show any variation in the expression between intracellular and extracellular tachyzoite stages. Since ApiAP2 proteins are expressed at very low levels, further exploration is required to validate the role of these proteins in stage-specific expression.

Evidence of antisense expression

Non-coding RNAs have been reported in many eukaryotes. Naturally occurring antisense transcripts (NATs) have been in implicated in the regulation of gene expression. NATs are complementary or antisense to a sense transcript. With the advent of strand-specific sequencing antisense transcripts are becoming easier to detect. While some of the detected antisense expression can be from read-through transcription and is just noise, antisense transcripts have been reported to play a role in tissue-specific and developmental gene regulation in *Toxoplasma* (74).

Antisense transcription has been reported in *P. falciparum (75)* and examination of strand-specific data suggested antisense transcripts could regulate different developmental stages (76, 77). When we used the non-strand-specific transcriptome data for gene prediction, genes were predicted on the forward and reverse strand at the same locus (Figure 4.9A). This suggested the presence of antisense expression in *S. neurona*. We examined the strand-specific data for antisense transcripts. Approximately 4% of the gene loci showed antisense expression (Figure 4.9B) and 50% of these loci show more antisense expression than sense expression (Figure 4.9C). 24% of the genes were reported to show antisense expression in *P. falciparum* (77). We only examined the assembled transcripts to detect antisense expression (as described in Methods), mapping the raw reads to the genes could increase the number of loci showing antisense expression. The number of antisense transcripts at each time point varies and the majority of these antisense transcripts are unique to that time point (Figure 4.10). This finding points to a role in developmental gene regulation. Next-gen sequencing is not devoid of noise. The reason for a greater percentage of antisense transcripts in the 53h schizont time point could be a factor of the sequencing depth. Therefore experimental validation is necessary before drawing definitive conclusions.

CONCLUSIONS

We have made publicly available the first annotated genome sequence of the horse parasite *Sarcocystis neurona.* To aid gene annotation, we generated transcriptome datasets from multiple time points that are currently being explored to understand differential gene regulation in distinct lifecycle stages. The most significant finding from the genome sequence is its abnormally large size of ~125 Mb. The genome doesn't show evidence of genome duplications or any expanded gene families; only ~36% of the genes are conserved with the other coccidians.

We did find a repeat content of ~23%, which is comparatively high in this phylum. ~11% of the repeats are TEs, although they do not appear to be active now, they could have, in the past, contributed to the genome rearrangement we observe in comparison to *T. gondii* and *Eimeria.* The repeat content alone, does not explain why the *S. neurona* genome is twice as large as *T. gondii*. With the availability of another *S. neurona* genome (SN1) and new *S. neurona* strains currently being sequenced, we look forward to exploring its population biology. It will be interesting to compare these findings to the recently completed analyses of genetic variation in the *T. gondii* strains (42).

REFERENCES

1. Levine ND (1988) *The Protozoan Phylum Apicomplexa Vol II* (CRC Press) p 154.
2. Templeton TJ*, et al.* (2004) Comparative Analysis of Apicomplexa and Genomic Diversity in Eukaryotes. *Genome research* 14(9):1686-1695.
3. DeBarry J & Kissinger J (2011) Jumbled Genomes: Missing Apicomplexan Synteny. *Molecular Biology and Evoluiton* 28.
4. Keeling PJ & Fast NM (2002) Microsporidia: biology and evolution of highly reduced intracellular parasites. *Annual review of microbiology* 56:93-116.
5. Gardner MJ*, et al.* (2002) Genome sequence of the human malaria parasite *Plasmodium falciparum*. *Nature* 419(6906):498-511.
6. Reid AJ*, et al.* (2012) Comparative genomics of the apicomplexan parasites *Toxoplasma gondii* and *Neospora caninum*: Coccidia differing in host range and transmission strategy. *PLoS pathogens* 8(3):e1002567.
7. Reid AJ*, et al.* (2014) Genomic analysis of the causative agents of coccidiosis in domestic chickens. *Genome research* 24(10):1676-1685.
8. Cowper B, Matthews S, & Tomley F (2012) The molecular basis for the distinct host and tissue tropisms of coccidian parasites. *Molecular and biochemical parasitology* 186(1):1-10.
9. Su C*, et al.* (2003) Recent expansion of *Toxoplasma* through enhanced oral transmission. *Science* 299(5605):414-416.
10. Tenter AM (1995) Current research on *Sarcocystis* species of domestic animals. *International journal for parasitology* 25(11):1311-1330.
11. Dubey JP*, et al.* (2001) A review of *Sarcocystis neurona* and equine protozoal myeloencephalitis (EPM). *Veterinary parasitology* 95(2-4):89-131.
12. Cheadle MA*, et al.* (2001) The nine-banded armadillo (Dasypus novemcinctus) is an intermediate host for *Sarcocystis neurona*. *International journal for parasitology* 31(4):330-335.
13. Dubey JP, Speer CA, Munday BL, & Lipscomb TP (1989) Ovine sporozoan encephalomyelitis linked to *Sarcocystis* infection. *Veterinary parasitology* 34(1-2):159-163.
14. Cheadle MA, Dame JB, & Greiner EC (2001) Sporocyst size of isolates of *Sarcocystis* shed by the Virginia opossum (Didelphis virginiana). *Veterinary parasitology* 95(2-4):305-311.
15. Carruthers VB & Sibley LD (1997) Sequential protein secretion from three distinct organelles of *Toxoplasma gondii* accompanies invasion of human fibroblasts. *Eur J Cell Biol* 73(2):114-123.
16. Sam-Yellowe TY (1996) Rhoptry organelles of the apicomplexa: Their role in host cell invasion and intracellular survival. *Parasitology today* 12(8):308-316.
17. Speer CA & Dubey JP (2001) Ultrastructure of schizonts and merozoites of *Sarcocystis neurona*. *Veterinary parasitology* 95(2-4):263-271.
18. Vaishnava S*, et al.* (2005) Plastid segregation and cell division in the apicomplexan parasite *Sarcocystis neurona*. *Journal of cell science* 118(Pt 15):3397-3407.
19. Sheffield HG & Melton ML (1968) The fine structure and reproduction of *Toxoplasma gondii*. *The Journal of parasitology* 54(2):209-226.

20. Striepen B, Jordan CN, Reiff S, & van Dooren GG (2007) Building the perfect parasite: cell division in apicomplexa. *PLoS pathogens* 3(6):e78.
21. Radke JR*, et al.* (2001) Defining the cell cycle for the tachyzoite stage of *Toxoplasma gondii*. *Molecular and biochemical parasitology* 115(2):165-175.
22. Radke JR & White MW (1998) A cell cycle model for the tachyzoite of *Toxoplasma gondii* using the Herpes simplex virus thymidine kinase. *Molecular and biochemical parasitology* 94(2):237-247.
23. Cornelissen AW, Overdulve JP, & van der Ploeg M (1984) Determination of nuclear DNA of five eucoccidian parasites, Isospora, *Toxoplasma gondii*, *Sarcocystis cruzi*, *Eimeria tenella*, *E. acervulina* and *Plasmodium berghei*, with special reference to gamontogenesis and meiosis in I. (T.) gondii. *Parasitology* 88 (Pt 3):531-553.
24. McClintock B (1942) The fusion of broken ends of chromosomes following nuclear fusion. *Proc Assoc Am Physicians*:458-463.
25. Kidwell MG (2002) Transposable elements and the evolution of genome size in eukaryotes. *Genetica* 115(1):49-63.
26. Bennetzen JL (2000) Transposable element contributions to plant gene and genome evolution. *Plant Mol Biol* 42(1):251-269.
27. Richard GF, Kerrest A, & Dujon B (2008) Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. *Microbiology and molecular biology reviews : MMBR* 72(4):686-727.
28. de Koning AP, Gu W, Castoe TA, Batzer MA, & Pollock DD (2011) Repetitive elements may comprise over two-thirds of the human genome. *PLoS genetics* 7(12):e1002384.
29. Lee SI & Kim NS (2014) Transposable elements and genome size variations in plants. *Genomics & informatics* 12(3):87-97.
30. Wolfe KH & Shields DC (1997) Molecular evidence for an ancient duplication of the entire yeast genome. *Nature* 387(6634):708-713.
31. Wendel JF (2000) Genome evolution in polyploids. *Plant Mol Biol* 42(1):225-249.
32. Castro C, Craig SP, & Castaneda M (1981) Genome organization and ploidy number in Trypanosoma cruzi. *Molecular and biochemical parasitology* 4(5-6):273-282.
33. Adam RD (2000) The *Giardia lamblia* genome. *International journal for parasitology* 30(4):475-484.
34. Carlton JM*, et al.* (2007) Draft genome sequence of the sexually transmitted pathogen Trichomonas vaginalis. *Science* 315(5809):207-212.
35. Ponzi M*, et al.* (1990) Generation of chromosome size polymorphism during in vivo mitotic multiplication of *Plasmodium berghei* involves both loss and addition of subtelomeric repeat sequences. *Molecular and biochemical parasitology* 41(1):73-82.
36. Ling KH*, et al.* (2007) Sequencing and analysis of chromosome 1 of *Eimeria tenella* reveals a unique segmental organization. *Genome research* 17(3):311-319.
37. Pain A*, et al.* (2005) Genome of the host-cell transforming parasite *Theileria annulata* compared with *T. parva*. *Science* 309(5731):131-133.
38. Kuo CH & Kissinger JC (2008) Consistent and contrasting properties of lineage-specific genes in the apicomplexan parasites *Plasmodium* and *Theileria*. *BMC evolutionary biology* 8:108.
39. Weir W*, et al.* (2009) Highly syntenic and yet divergent: A tale of two *Theilerias*. *Infection, Genetics and Evolution* 9(4):453-461.

40. Sibley LD (2011) Invasion and intracellular survival by protozoan parasites. *Immunological reviews* 240(1):72-91.
41. Besteiro S, Dubremetz JF, & Lebrun M (2011) The moving junction of apicomplexan parasites: a key structure for invasion. *Cellular microbiology* 13(6):797-805.
42. Lorenzi H. KA, Benke M.S. ,Namasivayam S.,Seshadri L.S.,Hadjithomas M.,Karamycheva S.,Pinney D.,Brunk B.,Ajioka J.W.,Ajzenberg D.,Boothroyd J.C.,Boyle J.P.,Dardé M.L.,Dubey J.P.,Fritz H.M.,Gennari S.M., Gregory B.D.,Kim K.,Rosenthal B. M.,Saeij J., Su C.,White M.W.,Zhu X.Q,Howe D.K.,Grigg M.E.,Parkinson J.,Liu L.,Kissinger J.C.,Roos D.S.,Sibley L. D. (2015) Comparative sequence analysis of *Toxoplasma gondii* reveals local genomic admixture drives concerted expansion and diversification of secreted pathogenesis determinants. *In Review*.
43. Blazejewski T*, et al.* (2015) Systems-based analysis of the *Sarcocystis neurona* genome identifies pathways that contribute to a heteroxenous life cycle. *mBio* 6(1).
44. Granstrom DE, Alvarez O, Jr., Dubey JP, Comer PF, & Williams NM (1992) Equine protozoal myelitis in Panamanian horses and isolation of *Sarcocystis neurona*. *The Journal of parasitology* 78(5):909-912.
45. Bolger AM, Lohse M, & Usadel B (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 30(15):2114-2120.
46. Trapnell C, Pachter L, & Salzberg SL (2009) TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* 25(9):1105-1111.
47. Trapnell C*, et al.* (2012) Differential gene and transcript expression analysis of RNA-seq experiments with TopHat and Cufflinks. *Nat Protoc* 7(3):562-578.
48. Grabherr MG*, et al.* (2011) Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat Biotechnol* 29(7):644-652.
49. Haas BJ*, et al.* (2013) De novo transcript sequence reconstruction from RNA-seq using the Trinity platform for reference generation and analysis. *Nat Protoc* 8(8):1494-1512.
50. Majoros WH, Pertea M, & Salzberg SL (2004) TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. *Bioinformatics* 20(16):2878-2879.
51. Besemer J, Lomsadze A, & Borodovsky M (2001) GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. *Nucleic Acids Res* 29(12):2607-2618.
52. Stanke M, Steinkamp R, Waack S, & Morgenstern B (2004) AUGUSTUS: a web server for gene finding in eukaryotes. *Nucleic Acids Res* 32(Web Server issue):W309-312.
53. Korf I (2004) Gene finding in novel genomes. *BMC bioinformatics* 5:59.
54. Holt C & Yandell M (2011) MAKER2: an annotation pipeline and genome-database management tool for second-generation genome projects. *BMC bioinformatics* 12:491.
55. Haas BJ*, et al.* (2008) Automated eukaryotic gene structure annotation using EVidenceModeler and the Program to Assemble Spliced Alignments. *Genome biology* 9(1):R7.
56. Lee E*, et al.* (2013) Web Apollo: a web-based genomic annotation editing platform. *Genome biology* 14(8):R93.
57. Schattner P, Brooks AN, & Lowe TM (2005) The tRNAscan-SE, snoscan and snoGPS web servers for the detection of tRNAs and snoRNAs. *Nucleic Acids Res* 33(Web Server issue):W686-689.
58. Gajria B*, et al.* (2008) ToxoDB: an integrated *Toxoplasma gondii* database resource. *Nucleic Acids Res* 36(Database issue):D553-556.

59. Bahl A, *et al.* (2003) PlasmoDB: the *Plasmodium* genome resource. A database integrating experimental and computational data. *Nucleic Acids Res* 31(1):212-215.
60. Kurtz S, *et al.* (2004) Versatile and open software for comparing large genomes. *Genome biology* 5(2):R12.
61. Chen F, Mackey AJ, Stoeckert CJ, Jr., & Roos DS (2006) OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups. *Nucleic Acids Res* 34(Database issue):D363-368.
62. Wang Y, *et al.* (2012) MCScanX: a toolkit for detection and evolutionary analysis of gene synteny and collinearity. *Nucleic Acids Res* 40(7):e49.
63. Krzywinski M, *et al.* (2009) Circos: An information aesthetic for comparative genomics. *Genome research*.
64. Oberstaller J, Pumpalova Y, Schieler A, Llinas M, & Kissinger JC (2014) The Cryptosporidium parvum ApiAP2 gene family: insights into the evolution of apicomplexan AP2 regulatory systems. *Nucleic Acids Res* 42(13):8271-8284.
65. Eddy SR (2001) HMMER: Pofile hidden Markov models for biological sequence analysis.
66. Blake DP, Oakes R, & Smith AL (2011) A genetic linkage map for the apicomplexan protozoan parasite Eimeria maxima and comparison with Eimeria tenella. *International journal for parasitology* 41(2):263-270.
67. Templeton TJ, *et al.* (2010) A Genome-Sequence Survey for Ascogregarina taiwanensis Supports Evolutionary Affiliation but Metabolic Diversity between a Gregarine and Cryptosporidium. *Molecular biology and evolution* 27(2):235-248.
68. Wasmuth JD, *et al.* (2012) Integrated bioinformatic and targeted deletion analyses of the SRS gene superfamily identify SRS29C as a negative regulator of Toxoplasma virulence. *mBio* 3(6).
69. Llinas M & DeRisi JL (2004) Pernicious plans revealed: *Plasmodium falciparum* genome wide expression analysis. *Current opinion in microbiology* 7(4):382-387.
70. Balaji S, Babu MM, Iyer LM, & Aravind L (2005) Discovery of the principal specific transcription factors of Apicomplexa and their implication for the evolution of the AP2-integrase DNA binding domains. *Nucleic Acids Res* 33(13):3994-4006.
71. Painter HJ, Campbell TL, & Llinas M (2011) The Apicomplexan AP2 family: integral factors regulating *Plasmodium* development. *Molecular and biochemical parasitology* 176(1):1-7.
72. Kafsack BF, *et al.* (2014) A transcriptional switch underlies commitment to sexual development in malaria parasites. *Nature* 507(7491):248-252.
73. Behnke MS, *et al.* (2010) Coordinated progression through two subtranscriptomes underlies the tachyzoite cycle of *Toxoplasma gondii*. *PloS one* 5(8):e12354.
74. Wery M, Kwapisz M, & Morillon A (2011) Noncoding RNAs in gene regulation. *Wiley interdisciplinary reviews. Systems biology and medicine* 3(6):728-738.
75. Gunasekera AM, *et al.* (2004) Widespread distribution of antisense transcripts in the Plasmodium falciparum genome. *Molecular and biochemical parasitology* 136(1):35-42.
76. Lopez-Barragan MJ, *et al.* (2011) Directional gene expression and antisense transcripts in sexual and asexual stages of *Plasmodium falciparum*. *BMC genomics* 12:587.
77. Siegel TN, *et al.* (2014) Strand-specific RNA-Seq reveals widespread and developmentally regulated transcription of natural antisense transcripts in *Plasmodium falciparum*. *BMC genomics* 15:150.

Figure and table legends

Figure 4.1. Nucleotide dotplot of the largest *S. neurona* SN3 scaffolds against themselves

The 30 largest *S. neurona* SN3 scaffolds were aligned were to self at the nucleotide level using NUCmer.

Figure 4.2. Multiple sequence alignment of the three cytochrome protein sequences

The three mitochondrial-encoded proteins, COXI, COB and COXIII were annotated using *S. neurona* transcriptome data and aligned to orthologous *T. gondii* and *E. tenella* protein sequences.

Figure 4.3. Factors contributing to a larger *S. neuorna* genome

A. A schematic representation of an orthologous gene found in *T. gondii* and *S. neurona*. Boxes indicate exons and are conserved in size between the two species. Numbers above and below the lines connecting the boxes indicate intron lengths (in bp) of *S. neurona* and *T. gondii* respectively. Terminal exons are not shown. **B.** Syntenic region between *S. neurona* and *T. gondii*. Blue boxes = genes on the forward strand. Red boxes = genes on the reverse stand. Numbers above and below the connecting lines indicate intergenic distance (in bp) in *S. neurona* and *T. gondii* respectively. **C.** Graph of repeat abundance in each genomic feature and the total percent of the genomic features in the *S. neurona* genome.

Figure 4.4. *S. neurona* orthologs show insertions in their coding sequence

Multiple sequence alignment of three orthologous proteins from *S. neurona*, *N. caninum* and *T. gondii*. Examples of insertion at the **A.** beginning **B.** middle and **C.** end of the protein sequence.

Figure 4.5. Synteny between SN1 and SN3 genomes and their predicted proteins

A. Nucleotide alignment between the largest fifteen SN3 scaffolds and largest thirty SN1 scaffolds. **B.** Circos plot illustrates synteny between the SN3 and SN1 scaffolds. The outer circle

represents the largest SN3 and SN1 scaffolds labeled with strain name and scaffold number. The colored bands link the syntenic blocks between the two strains. The figure was drawn with SN3 scaffolds as the reference.

Figure 4.6. Comparison of the coccidian gene distribution and synteny

A. Venn diagram of ortholog clusters generated using OrthoMCL. For an ortholog cluster to be shared by all four species, the cluster must contain at least one gene from each species and so on. **B.** Circos plot of synteny among the coccidian genomes. The outer circle represents the chromosomes/scaffolds, colored and labeled by species. Tick marks on the scaffolds/chromosmes represent 1 Mb. All 14 chromosomes are shown for *T. gondii, H. hammondi* and *N. caninum*. The 14 largest scaffolds are shown for *S. neurona* and *E. tenella.* The colored bands link syntenic regions. The figure is drawn with *T. gondii, H. hammondi* and *E. tenella* as references in that order. Syntenic blocks were generated using a total of 12,798 genes in ortholog clusters shared by all five species.

Figure 4.7. BLAST2GO and enrichment analyses of SN3 genes

A. Distribution of annotated molecular functions (at level 5). The BLAST2GO pipeline was used for the functional annotations of SN3 gene predictions. **B.** An enrichment analysis of genes unique to SN3 was performed using all SN3 genes as reference and a p-value of 0.01. The graph shows enriched molecular functions.

Figure 4.8. Stage-specific expression of *SnAP2* genes

Strand-specific transcriptome data from different time points were used to identify stage-specific expression of the annotated SnApiAP2 proteins. Figure generated using http://bioinformatics.psb.ugent.be/webtools/Venn/.

Figure 4.9. Strand-specific transcriptome data reveal antisense expression

A. The presence of antisense transcripts in the transcriptome results in an increased number of gene models by gene predictors. The transcript (pink), on the forward strand is supported by a *T. gondii* protein (red) suggesting it is the sense transcript (FPKM: 3.01). Two antisense transcripts occur at the same locus (FPKM: 3.69 and 4.4). The presence of these antisense transcripts results in genes predicted on both strands at this locus. **B.** Distribution of the number of antisense transcripts based on overlap with sense transcript. **C.** Ratio of expression FPKM values between sense and antisense transcript at each locus.

Figure 4.10. Role of antisense transcripts in stage-specific regulation

The total number of antisense transcripts at each time point was identified based on a minimum of 20% overlap with the sense transcript. The antisense transcript is unique to that time point if it is observed only at that time point.

Table 4.1. *S. neurona* SN3 genome assembly statistics

Statistics gathered from Newbler v2.5 assembly report.

Table 4.2. Summary of transcriptome data sets and assemblies

454 datasets were assembled with Newbler v2.5 and Illumina datasets were assembled with TopHat-Cufflinks and Trinity pipelines. *De novo* assembled data sets were filtered for bovine contaminants (See Methods).

Table 4.3. NUMTs and NUPTs in the Coccidia

Where available**,** the organellar genome of each species was used as the library to RepeatMask its nuclear genome sequence. For *H. hammondi,* mt and ptDNA sequences from *T. gondii* were used for masking and in the case of *S. neruona,* mtDNA sequences from *T. gondii* and *E. tenella* were used.

Table 4.4. Comparison of *S. neurona* genome with *T. gondii* and *E. tenella*

The genome and annotation data for *T. gondii* and *E. tenella* were downloaded from ToxoDB.org release 24.

Table 4.5. Comparisons to *S. neurona* SN3 gene predictions

A BLASTP analysis was performed with SN3 proteins as the database and each of the reported coccidian protein datasets as queries. For SN1, BLAST hits were filtered for an identity of at least 90% and an E-value of 1e-10. For the other coccidian species, the BLAST hits were filtered for an identity of at least 40% and E-value of 1e-6. *T. gondii*, *H. hammondi*, *N. caninum* and *E. tenella* protein sequences were downloaded from ToxoDB.org release 24. SN1 proteins were obtained pre-publication from collaborators John Parkinson and Micheal Grigg (43).

Table 4.6. Comparisons to *S. neurona* SN3 transcriptome

The cuffmerge assemblies of all Illumina strand-specific datasets were used as the subjects in a BLAST search against each data set shown. In the case of SN3 and SN1, predicted transcripts were used as the query in a BLASTN search and filtered for an identity of at least 98% and 90% respectively and an E-value of 1e-10. For the other coccidians, a TBLASTN was used with predicted proteins as query and filtered for an identity of at least 40% and an E-value of 1e-6.

Table 4.7. Average protein lengths of 676 1:1 orthologs shared across all apicomplexans

Orthologs were identified using BLASTP searches, only single-copy orthologs were used to calculate average length.

Table 4.8. Summary of *S. neurona* repeat content

RepeatModeler was used to *de novo* identify and classify repeat families. The identified repeat families were used as a library with RepeatMasker to obtain repeat content statistics.

Table 4.9. Comparisons to *S. neurona* SN1 gene predictions

A BLASTP analysis was performed with SN3 proteins as the database and the each of the reported coccidian protein datasets as queries. Results were filtered as described for Table 4.5.

Table 4.10. Number of ApiAP2 proteins and domains in the Coccidia

Custom-made HMMs were run on each coccidian protein dataset. Custom HMMs were built using ApiAP2 alignments from (64).

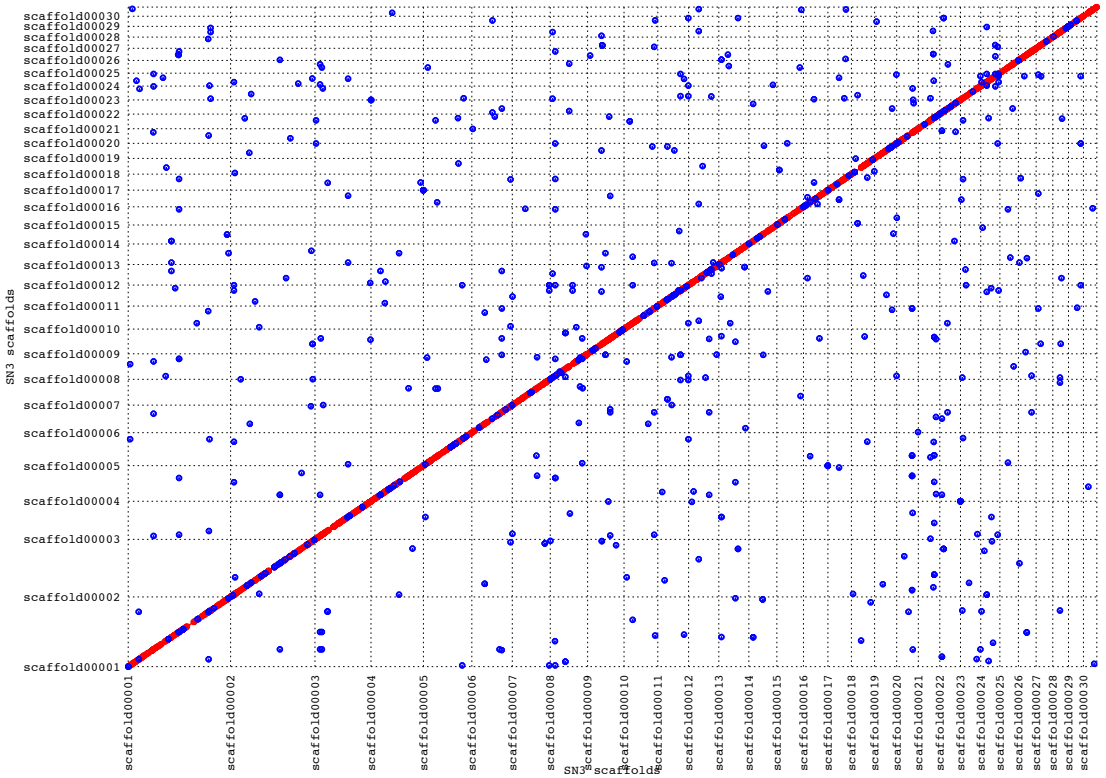
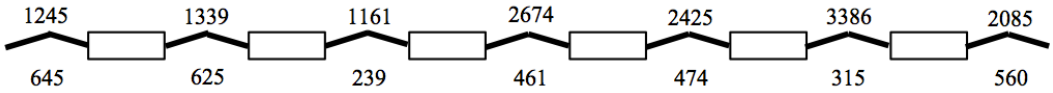


Figure 4.1. Nucleotide dotplot of the largest *S. neurona* SN3 scaffolds against themselves

Figure 4.2. Multiple sequence alignment of the three cytochrome protein sequences

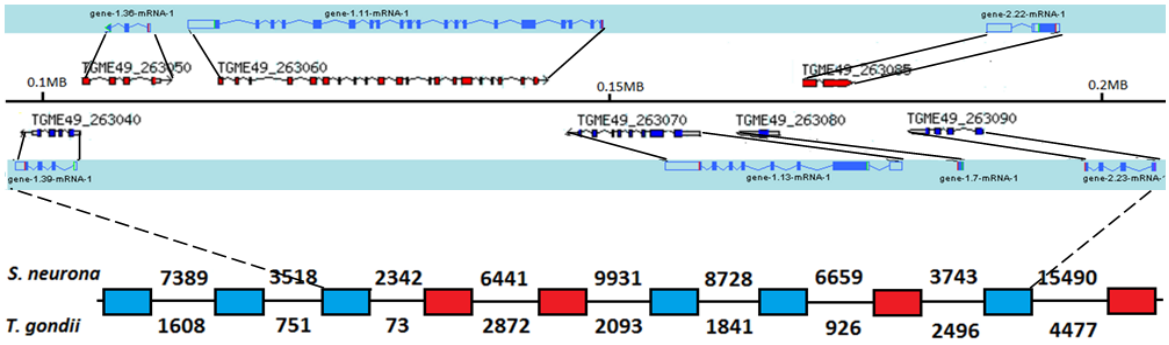
A.

SN3_0330055



TGME49_260370

B.



C.

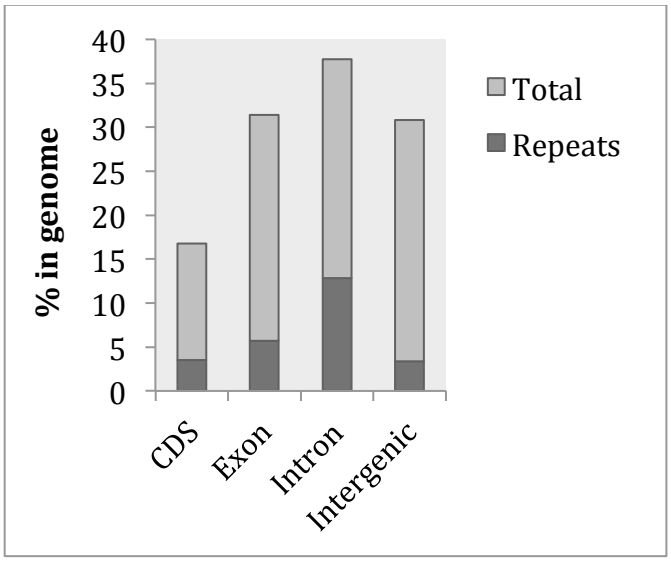
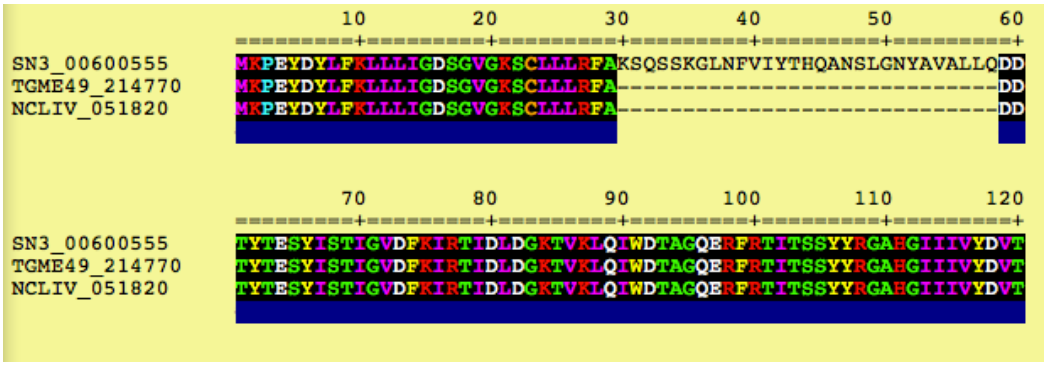
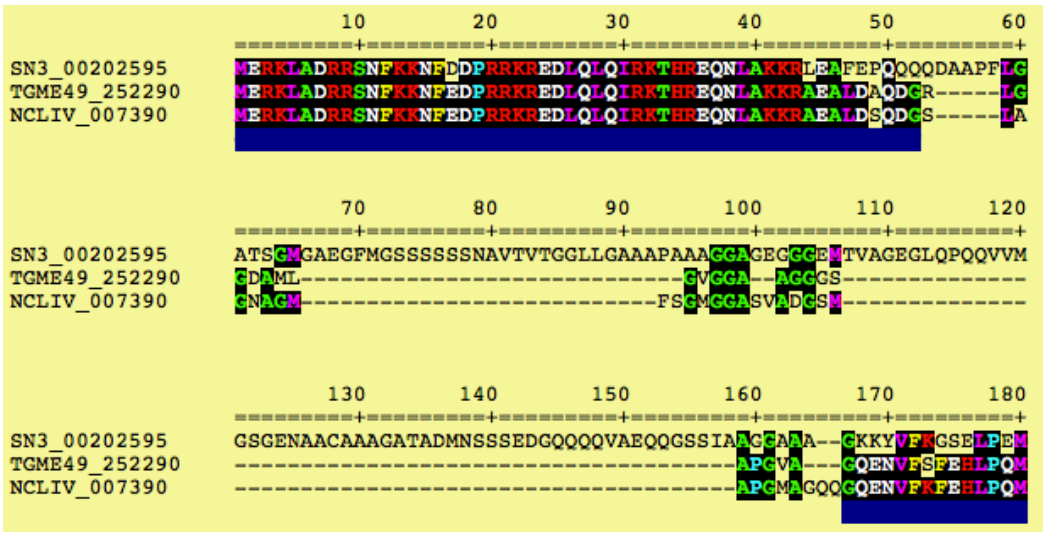


Figure 4.3. Factors contributing to a larger *S. neuorna* genome

A.



B.



C.

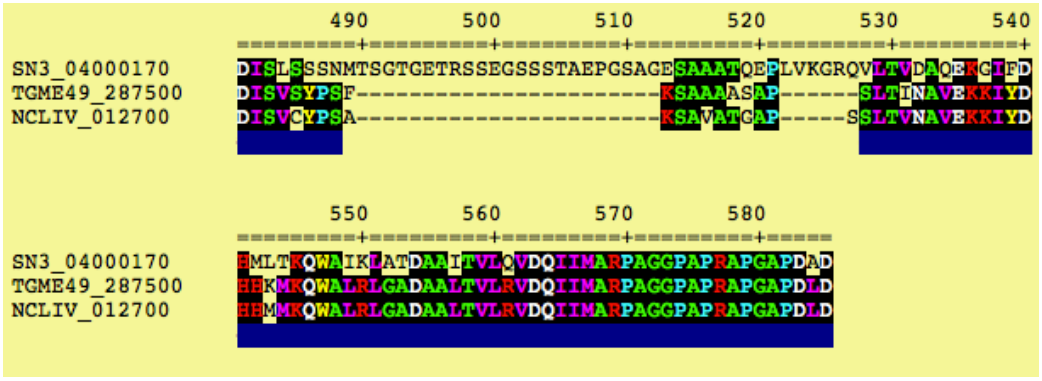
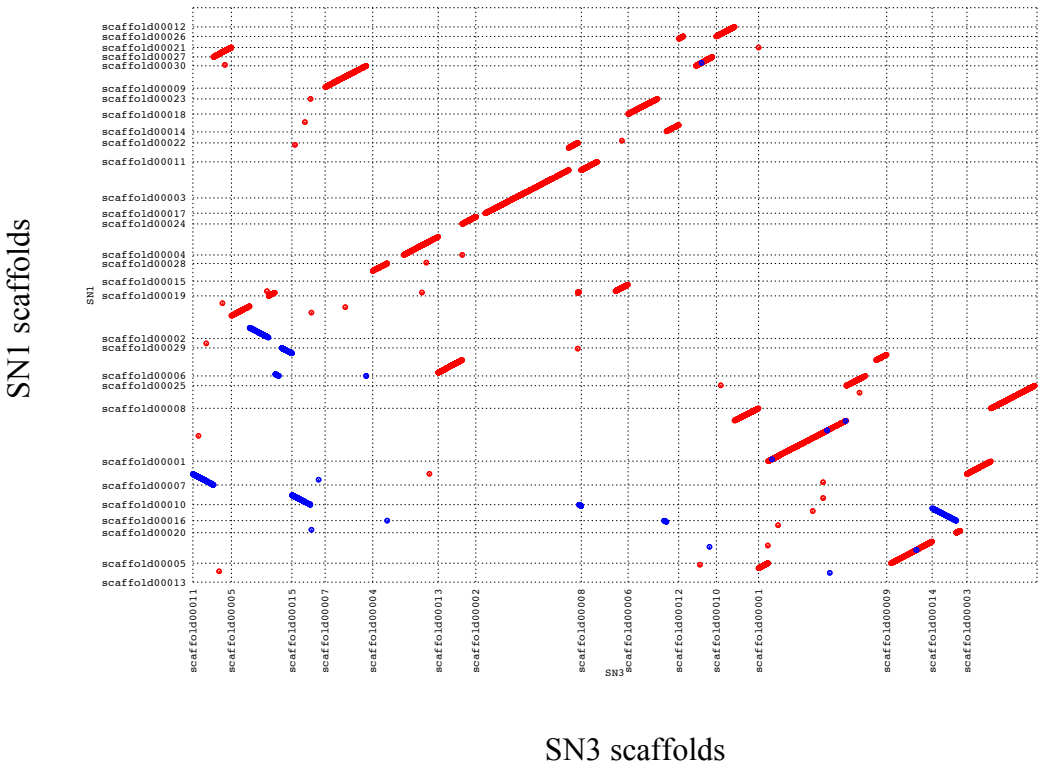


Figure 4.4. *S. neurona* orthologs show insertions in their coding sequence

A.



B.

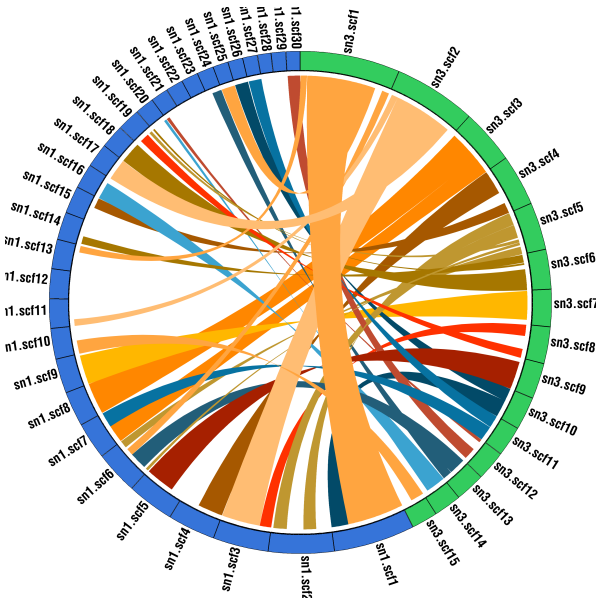
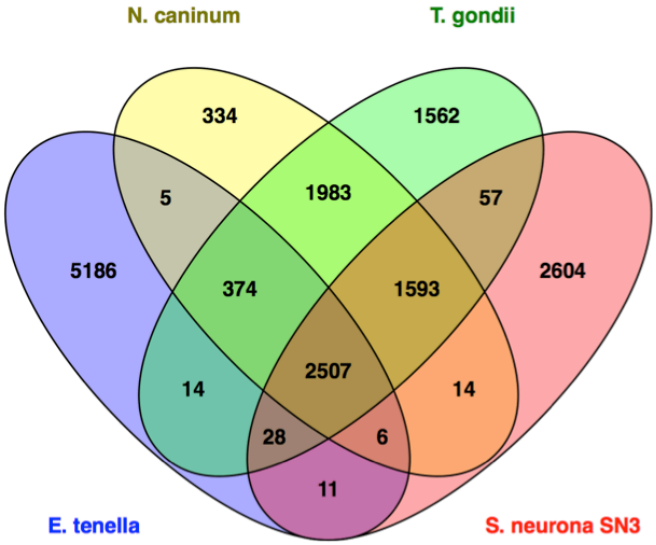


Figure 4.5. Synteny between SN1 and SN3 genomes and their predicted proteins

A.



B.

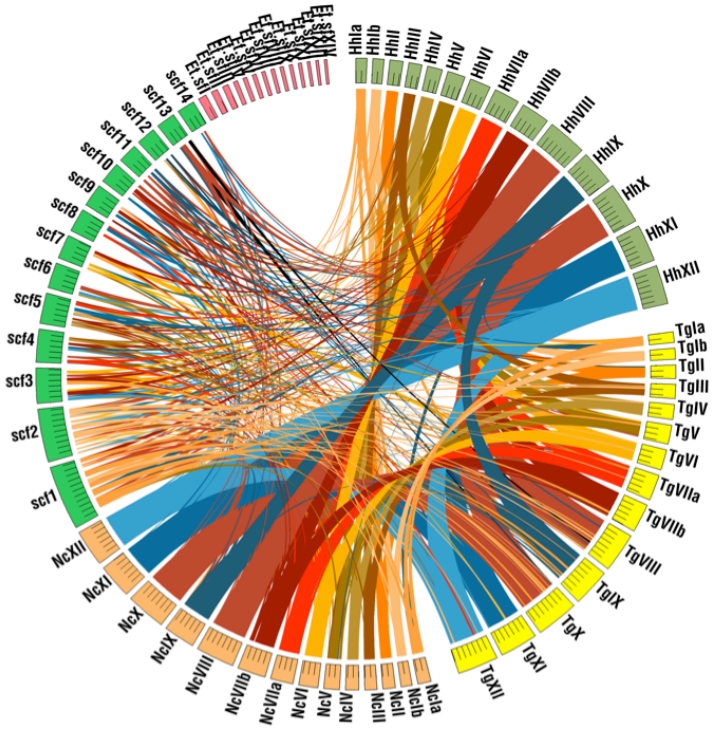
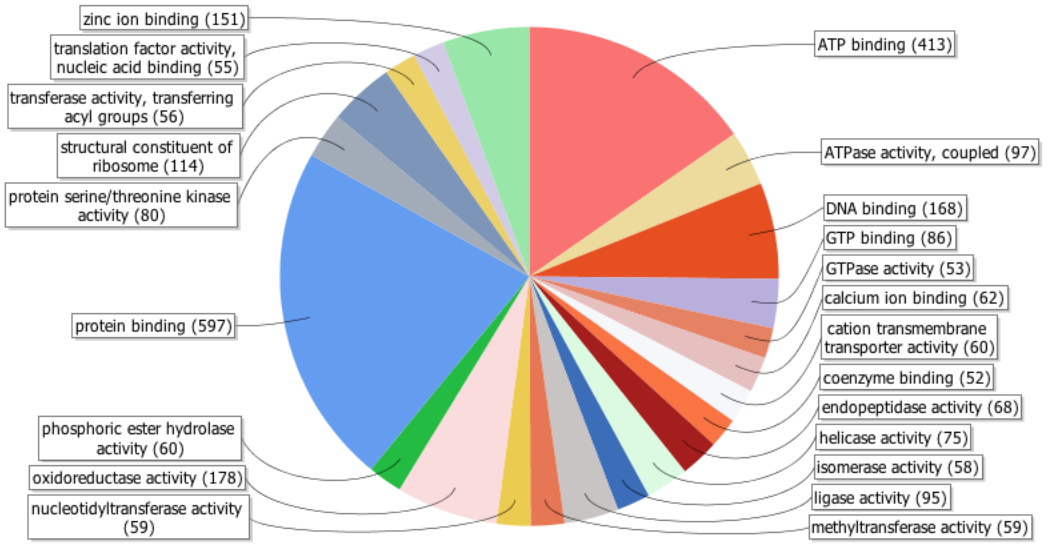


Figure 4.6. Comparison of the coccidian gene distribution and synteny

A.



B.

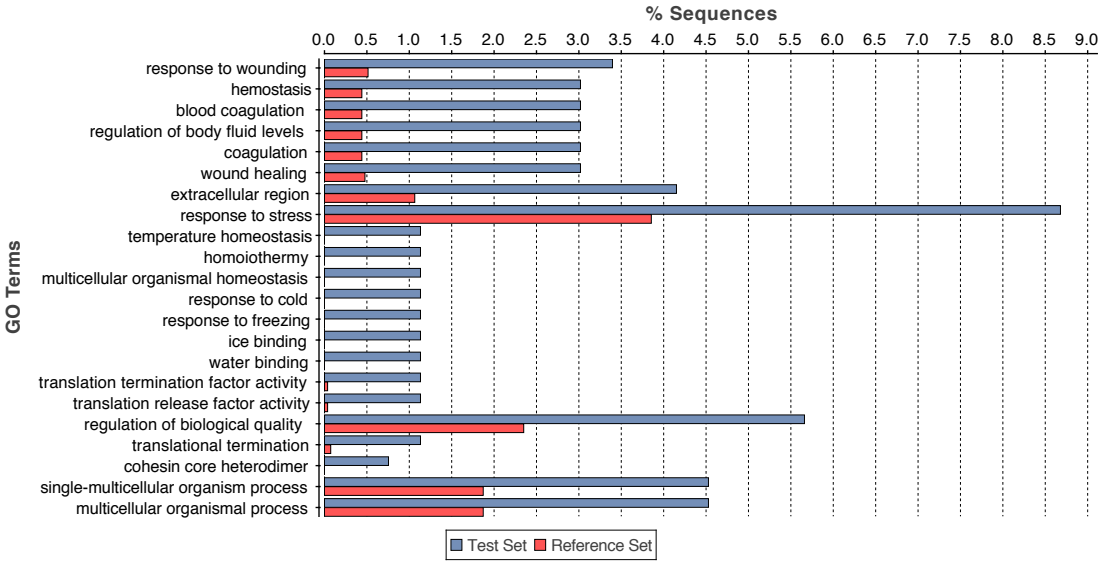


Figure 4.7. BLAST2GO and enrichment analyses of SN3 genes

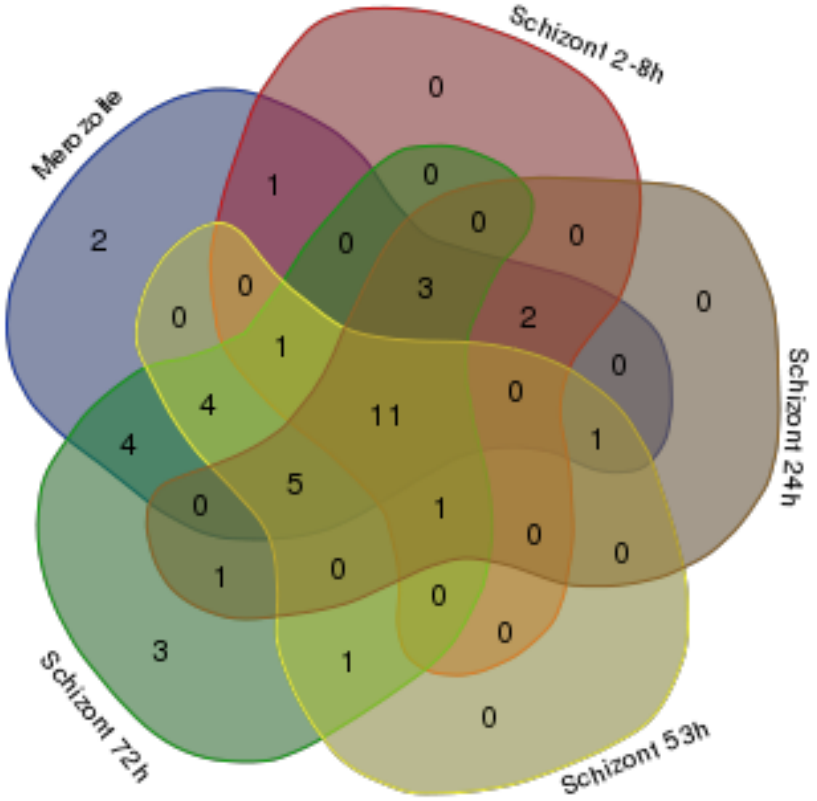
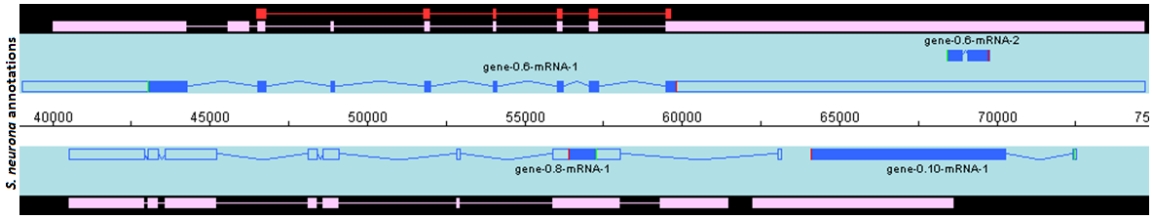
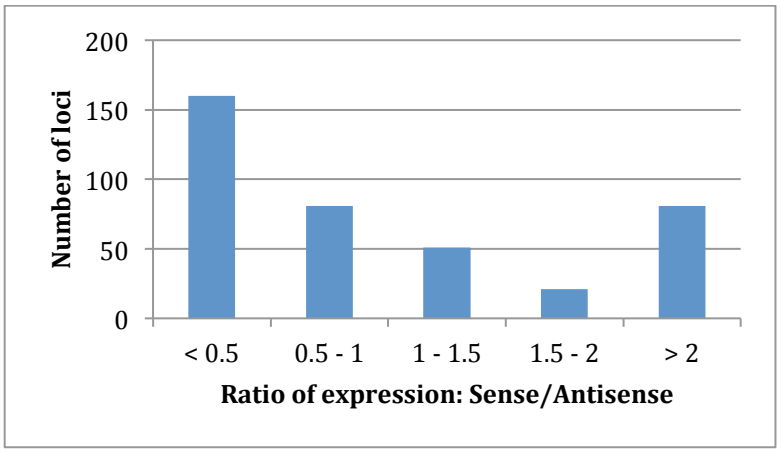


Figure 4.8. Stage-specific expression of *SnAP2* genes

A.



B.



C.

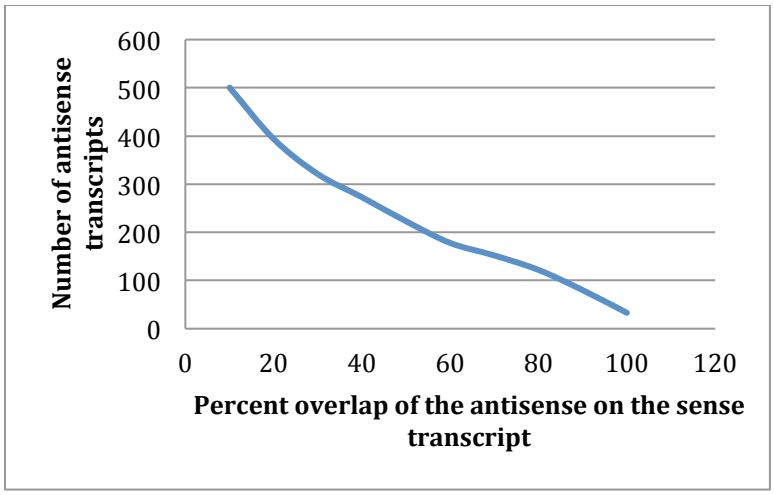


Figure 4.9. Strand-specific transcriptome data reveal antisense expression

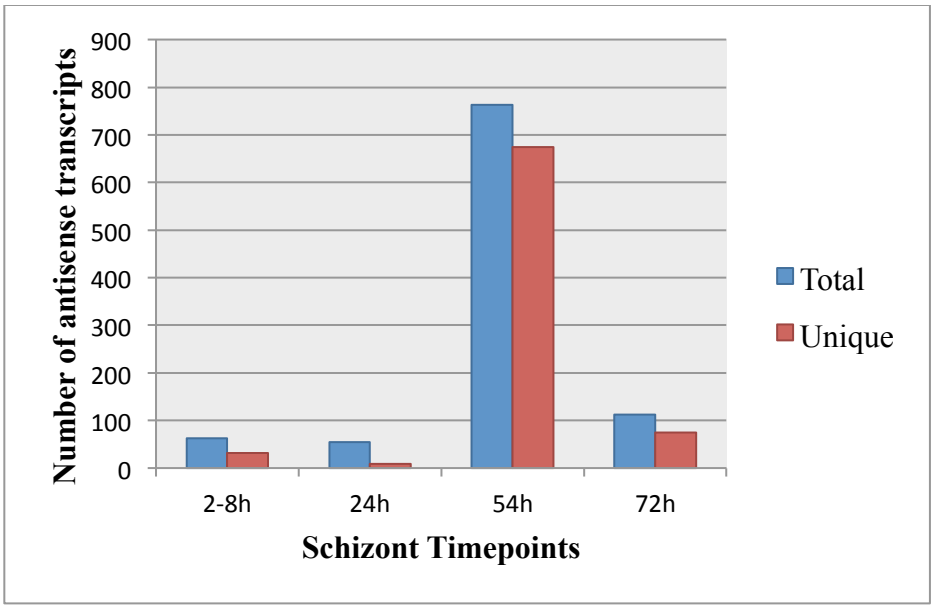


Figure 4.10. Role of antisense transcripts in stage-specific regulation

Table 4.1. *S. neurona* SN3 genome assembly statistics

	S. neurona **SN3**
Genome size (Mb)	124.4
Estimated genome size (Mb)	151.2
N50 Scaffold size (Mb)	3.4
Largest scaffold (Mb)	11.36
Number of scaffolds	172
Number of contigs > 500 bp	701
Number of contigs < 500bp	1,725
G-C content (%)	51.41

Table 4.2. Summary of transcriptome data sets and assemblies

Sequencing plaform	Transcriptome timepoint	Assembler	Number of transcripts
454	Early to mid schizont	Newbler	9,778
	Late schizont		14,757
	Merozoite		57,417
	Merozoite (SN4 - EST)		1,527
	Combined assembly		61326 (28,419 clusters+ 32,907 singletons)
Illmina (Non stranded-specific library)	Merozoite	Tophat-Cufflinks	12,607
		Trinity - after filtering	24,489
Illumina (Strand-specific library) -Run1	Schizont 2-8 hours	Tophat-Cufflinks	16,944
	Schizont 24 hours		17,092
	Schizont 53 hours		18,038
	Schizont 72 hours		16,957
	Combined assembly	Trinity - after filtering	16,530 23,927
Illumina (Strand-specific library) -Run2	Schizont 2-8 hours	Tophat-Cufflinks	13,563
	Schizont 24 hours		13,404
	Schizont 53 hours		14,242
	Schizont 72 hours		13,555
	Merozoite		14,869
	Merozoite		14,962
	Schizont Merozoite	**Cuffmerge - all strand-specific Cufflinks transcripts**	**23,961**

Table 4.3. NUMTs and NUPTs in the Coccidia

	NUMTs			NUPTs		
	Density (%)	**Number of insertions**	**Number of base pairs**	**Density (%)**	**Number of insertions**	**Number of base pairs**
T. gondii ME49	1.43	9,356	891,057	0.18	1,108	112,443
H. hammondi H.H.34	1.57	10,501	989,826	0.19	1,206	117,917
N. caninum NCLIV	0.66	4,534	380,139	0.06	373	34,236
S. neurona SN3	0.03	454	38,353	0.02	54	27,624
E. tenella Houghton	0.04	227	22,024	0.09	338	45,513

Table 4.4. Comparison of *S. neurona* genome with *T. gondii* and *E. tenella*

	S. neurona **SN3**	***T. gondii*** **ME49**	***E. tenella*** **Houghton**
Genome size (Mb)	124	63	52
Chromosomes	Unknown	14	14
GC content (%)	51	52	58
Repeat content (%)	23	5	28
Protein coding genes	6,936	8,322	8,597
tRNA genes	104	174	33
Average length (bp)			
Coding exon	523	419	338
Exon	893	606	338
Intron	1,356	498	238
Transcript	5,647	3,530	1,715
Gene	12,731	5,569	2,496
Protein	1,001	801	505
Intergenic	5,531	1,817	3,027

Table 4.5. Comparisons to *S. neurona* SN3 gene predictions

Query coverage (%)	**SN1 annotations**	***T. gondii***	***N. caninum***	***H. hammondi***	***E. tenella***
100	1,076	180	195	201	123
90	2,336	676	756	725	539
80	3,239	1,016	1,122	1,079	793
70	3,865	1,352	1,470	1,401	1,030
60	4,297	1,714	1,798	1,744	1,268
50	4,648	2,092	2,170	2,126	1,492

Table 4.6. Comparisons to *S. neurona* SN3 transcriptome

Query coverage (%)	**SN3 annotations**	**SN1 annotations**	***T. gondii***	***N. caninum***	***H. hammondi***	***E. tenella***
100	1,105	1,218	192	227	221	138
90	3,329	2,789	770	856	829	603
80	4,649	4,072	1,110	1,240	1,189	892
70	5,367	4,973	1,483	1,599	1,539	1,140
60	5,763	5,458	1,851	1,929	1,903	1,393
50	6,019	5,791	2,226	2,312	2,305	1,634

Table 4.7. Average protein lengths of 676 1:1 orthologs shared across all apicomplexans

Species	**Average protein length (bp)**
P. falciparum	720
P. vivax	708
B. bovis	546
E. tenella	542
N. caninum	740
T. gondii	764
S. neurona SN3	945
S. neurona SN1	751
C. parvum	580

Table 4.8. Summary of *S. neurona* repeat content

Repeat Class		Count	Bp Masked	% masked
DNA	CMC-EnSpm	66,750	12,824,492	10.36
	MULE-MuDR	4,333	434,883	0.35
	Novosib	4,432	656,078	0.53
	PIF-Harbinger	791	88,074	0.07
LINE	Jockey	6,348	951,899	0.77
	L1-Tx1	174	23,321	0.02
	L2	558	38,893	0.03
	R1	6,799	848,130	0.69
LTR	Copia	764	85,327	0.07
	Gypsy	16,090	3,641,347	2.94
Low complexity		7,173	498,940	0.40
Simple repeat		95,576	6,364,230	5.14
Unknown		10,239	1,949,347	1.58
Total		**220,027**	**28,404,961**	**22.96**

Table 4.9. Comparisons to *S. neurona* SN1 gene predictions

Query coverage (%)	***T. gondii***	***N. caninum***	***H. hammondi***	***E. tenella***
100	117	119	125	87
90	486	541	517	399
80	807	884	856	654
70	1,108	1,195	1,140	877
60	1,454	1,523	1,488	1,116
50	1,849	1,885	1,893	1,332

Table 4.10. Number of ApiAP2 proteins and domains in the Coccidia

	Number of ApiAp2 proteins	**Number of ApiAp2 domains**
E. tenella	49	77
S. neurona	41	75
N. caninum	63	109
H. hammondi	70	117
T. gondii	65	112

CHAPTER 5

DISCUSSION AND FUTURE DIRECTIONS

Most of the currently identified apicomplexans are obligate intracellular parasites and have complex life cycles involving a sexual and asexual phase. This deep-branching phylum is apt for studies of eukaryotic diversity. The gene repertoire varies widely across the phylum. Fewer than 30% of the genes have clear orthologs that are present across the entire phylum. Their genomes are reductive in nature and show immense lineage- and species-specific gene retention. These differences in gene content are expected, given the diverse array of hosts and environmental niches they infect and reside in. Their genomic landscape and architecture also vary. The majority of the genes in *Cryposporidium* lack introns and the genes are tightly packed (1); whereas introns and intergenic regions in *S. neurona* are comparable to those observed in some free-living or higher eukaryotes. Apicomplexan genomes show extensive rearrangement, as synteny is not conserved across the phylum. Interestingly, one of the main drivers of genome rearrangement in most species, TEs, have been reported in only a few apicomplexan genera.

The focus of this dissertation is the evolution of genomes in the coccidian lineage. The genomes of five coccidian genera, *Eimeria*, *Sarcocystis*, *Neospora*, *Toxoplasma* and *Hammondia*, are now available. The parasites in this lineage can collectively infect essentially all vertebrates. These parasites are similar, yet different. *T. gondii* in particular has evolved as a generalist, whereas the *Eimeria* species are host-specific. All sequenced coccidian genera have a different definitive host, except *Toxoplasma* and *Hammondia,* which share the felines as the definitive host. *S. neurona* does not form rhoptries or a parasitophorosous vacuole, some of the

defining characteristics of apicomplexans. The mitochondrial genomes of the coccidians may not be conserved with other apicomplexans. The mt genomes of a number of *Eimeria* species have been characterized as linear concatemers. The *T. gondii* mt genome is an enigma. Attempts to isolate the mt genome over the years have been unsuccessful. I have reported here (Chapter 1) the most comphrensive study of the *T. gondii* mt genome. Our examination of the genomic and transcriptome data suggests that the mt genomes of *N. caninum*, *H. hammondi* and *S. neurona* might be similar to that of *T. gondii*.

At the nuclear genome level, synteny is missing even within the coccidian lineage. *T. gondii, H. hammondi* and *N. caninum* are highly sytenic. Synteny starts to break down with *S. neurona* and is completely absent with *Eimeria*. The difference in genomic composition could be responsible for this observation. *Eimeria* genome sequences have a bipartite chromosome structure of repeat-rich and repeat-poor regions. The *S. neurona* genome does not show this bipartite feature but it is repeat-rich, leading primarily to abnormally large introns. Among the coccidians, TEs have been reported only in *Eimeria* and *S. neurona,* although they no longer appear to be active. In *Eimeria*, the chromovirus family of DNA transposons is prevalent whereas CMC-like and Gypsy elements are common in *S. neurona*. The predicted *Eimeria* proteins have a high percentage of homopolymeric amino acid repeats yet; overall, they are smaller than the predicted *S. neurona* proteins, which also contain repeats. In spite of the high repeat content observed in Eimerian genomes, their genomes sizes are only 45- 57 Mb in size, whereas *S. neurona* has an abnormally large 125 Mb genome. The comparatively high repeat content in *S. neurona* only partially accounts for the large genome size. We did not identify any other significant factors. In fact, some of the gene families are reduced in *S. neurona* is comparison to the other coccidians. As part of the genome project, we generated transcriptome

data from multiple time points, *in vitro,* representing the asexual cycle. Comparative studies from these time points are currently underway to better understand the nuances of differential gene regulation in this parasite and how it compares to other coccidians.

T. gondii, *H. hammondi* and *N. canium* genomes are repeat poor. Instead, their genomes have an unusual phenomenon. *T. gondii* and *H. hammondi* have unprecedented levels of organaller DNA (NUM/PT) insertions. *N. caninum* has only half the percentage of NUM/PTs but is still has a higher level than other eukaryotes. In contrast, *Eimeria* and *S. neurona* contain only a few NUM/PTs. If NHEJ is the mechanism of NUM/PT acquisition, why is this phenomenon restricted to only *T. gondii*, *H. hammondi* and *N. caninum*? Perhaps, the bizarre nature of the *T. gondii* mt genome is a factor. Our basic experiments do show that *T. gondii* integrates DNA at artificially induced breaks at a higher frequency than other organisms. It is not clear why. Does the NHEJ pathway function differently in *T. gondii*? We already know the NHEJ pathway is used preferentially relative to homologous recombination.

Another striking finding is the difference in the age profile of the NUMTs found in *N. caninum* with respect to *T. gondii*/*H. hammondi*. A large proportion of the NUMTs in *N. caninum* are young, having been acquired in the last 5-10 million years. *T. gondii* and *H. hammondi* saw a large influx of NUMTs around 15 million years ago and NUMT acquisition has dropped rather steeply over the last 10 million years. It is possible that there is a rapid turnover of NUMTs in *N.caninum* in comparison to *T. gondii* or *H. hammondi*. If we calculate the age of the NUMTs, assuming that the mutation rate is uniform in all three species and the insertions have an intron mutation rate, then the results suggest that the ancestor of *T. gondii*/*H. hammondi* and *N. caninum* did not contain a lot of NUMTs and the common ancestor of *T. gondii* and *H. hammondi* acquired NUMTs rapidly. This reasoning is supported by the lack of orthologous

NUMTs between *T. gondii* and *N. caninum*. However, if the rate of turnover is indeed high in *N. caninum*, it is possible *N. caninum* has simply just lost its ancestral NUMTs. It will be interesting to conduct artificially induced DSB experiment in *N. caninum* (as described above for *T. gondii*) to determine if *N. caninum* inserts DNA at a rate higher than *T. gondii*.

Whatever may be the reasons responsible for the differences in the NUMT profile, we have evidence to suggest NUMTs are important forces of genome evolution in these parasites. We see differential presence/absence of NUMTs in the *T. gondii* strains. The presence/absence patterns do not necessarily follow the local ancestry patterns identified by Lorenzi *et al.,* (2). Lorenzi *et al.,* had used bins of 1000 SNPs to determine local ancestry. It is very likely that the recombination in *T. gondii* can occur as events smaller than 1000 SNPs, which would explain why see different patterns in the differential presence/absence of NUMTs. We can repeat the analyses performed by Lorenzi *et al.,* with a smaller SNP bin size and evaluate the effect on ancestry and NUMT patterns. Or, it is possible the insertion and deletion of NUMTs occurs independently in the different strains. The first theory seems more plausible because the insertion and/or deletion sites are conserved across strains. The other possibility is that recombination occurred with strains not included in this study. Examination of additional *T. gondii* strains will address this question. GO enrichment analysis using genes in or near the differential NUMTs identified an enrichment of genes associated with pyrimidine metabolism and the purine salvage pathway (3). It should be noted most of these NUMTs are present in introns, although a few are present in UTRs and upstream flanking regions. Transcriptome data from these strains needs to be examined to see if the presence of the NUMT affects the expression in any way. This analysis would be important to determine the extent to which NUMTs can affect biological function.

The quality of the genome assemblies does have an impact on these findings. We know a problem of 'pseudo-diploidy' exists in the *T. gondii* ME49 assembly. This means repetitive regions in the reference genome sequence appear to have been collapsed in the assembly. Some of these collapsed regions have been identified as a result of new genome sequence reads being mapped to this reference match in greater numbers, indicating copy number variation (CNV) with respect to the reference sequence. The *rop5* locus and the loci of a number of secreted pathogenesis determinants are examples (2). We have also identified the missing rDNA cluster in the reference genome sequence and experimentally verified the copy number (data not shown in this dissertation). The collapsed regions could explain the high number of SNP positions identified, but they were filtered quite stringently to try and avoid these regions. Experiments are underway to understand the population biology of *S. neurona*. Preliminary analyses of genome sequences from additional strains of *S. neurona* and *N. caninum* revealed very few SNPs (personal communication, Micheal Grigg). It is possible that the population history of *T. gondii* is very different from the other coccidians. Efforts are currently underway to fix the *T. gondii* ME49 reference genome assembly to resolve the compressions. These compression artifacts in the *T. gondii* reference genome, analyzed here, may affect the number of NUM/PTs identified, but it does not change our observation that NUMP/Ts are important drivers of evolution in these parasites.

REFERENCES

1. Roy SW & Penny D (2007) Widespread intron loss suggests retrotransposon activity in ancient apicomplexans. *Molecular biology and evolution* 24(9):1926-1933.
2. Lorenzi H. KA, Benke M.S. ,Namasivayam S.,Seshadri L.S.,Hadjithomas M.,Karamycheva S.,Pinney D.,Brunk B.,Ajioka J.W.,Ajzenberg D.,Boothroyd J.C.,Boyle J.P.,Dardé M.L.,Dubey J.P.,Fritz H.M.,Gennari S.M., Gregory B.D.,Kim K.,Rosenthal B. M.,Saeij J., Su C.,White M.W.,Zhu X.Q,Howe D.K.,Grigg M.E.,Parkinson J.,Liu L.,Kissinger J.C.,Roos D.S.,Sibley L. D. (2015) Comparative sequence analysis of *Toxoplasma gondii* reveals local genomic admixture drives concerted expansion and diversification of secreted pathogenesis determinants. *In Review*.
3. Hyde JE (2008) Fine targeting of purine salvage in *Cryptosporidium* parasites. *Trends in parasitology* 2008 Aug;24(8):336-9.

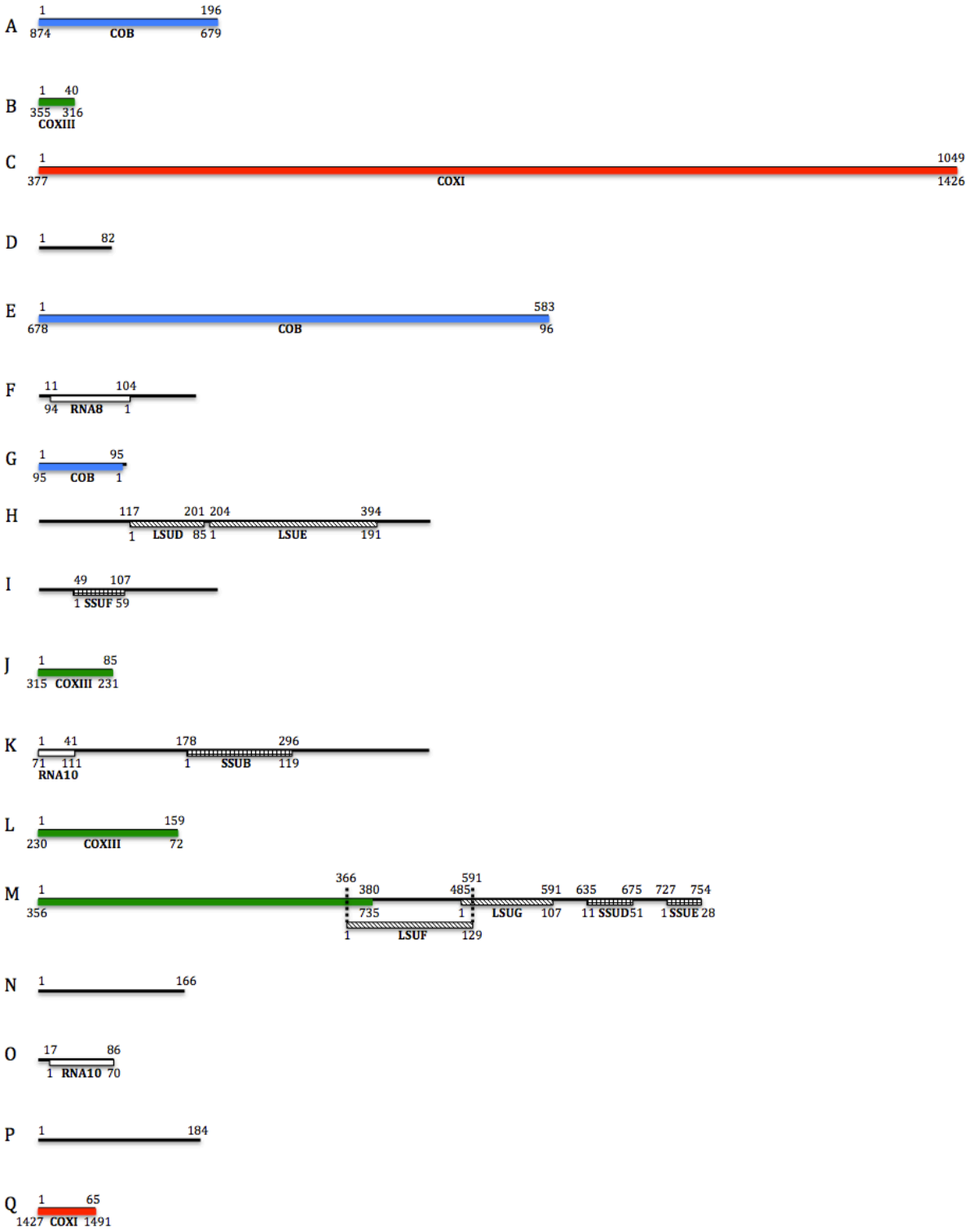
CHAPTER 6

APPENDICES

List of co-authored publications

1. Blazejewski T, Nursimulu N, Pszenny V, Dangoudoubiyam S, **Namasivayam S**, Chiasson MA, Chessman, Tonkin M, Swapna LS, Hung SS, Bridgers J, Ricklefs SM, Boulanger MJ, Dubey JP, Porcella SF, Kissinger JC, Howe DK, Grigg ME, Parkinson J. 2015. Systems-based analysis of the *Sarcocystis neurona* genome identifies pathways that contribute to a heteroxenous life cycle. mBio 6(1):e02445-14. doi:10.1128/mBio.02445-14
2. Lorenzi H. KA, Benke M.S. ,**Namasivayam S**.,Seshadri L.S.,Hadjithomas M.,Karamycheva S.,Pinney D.,Brunk B.,Ajioka J.W.,Ajzenberg D.,Boothroyd J.C.,Boyle J.P.,Dardé M.L.,Dubey J.P.,Fritz H.M.,Gennari S.M., Gregory B.D.,Kim K.,Rosenthal B. M.,Saeij J., Su C.,White M.W.,Zhu X.Q,Howe D.K.,Grigg M.E.,Parkinson J.,Liu L.,Kissinger J.C.,Roos D.S., Sibley L. D. (2015) Comparative sequence analysis of *Toxoplasma gondii* reveals local genomic admixture drives concerted expansion and diversification of secreted pathogenesis determinants. *In Revision, Nature Communications*.

Additional tables, figures and sequence data



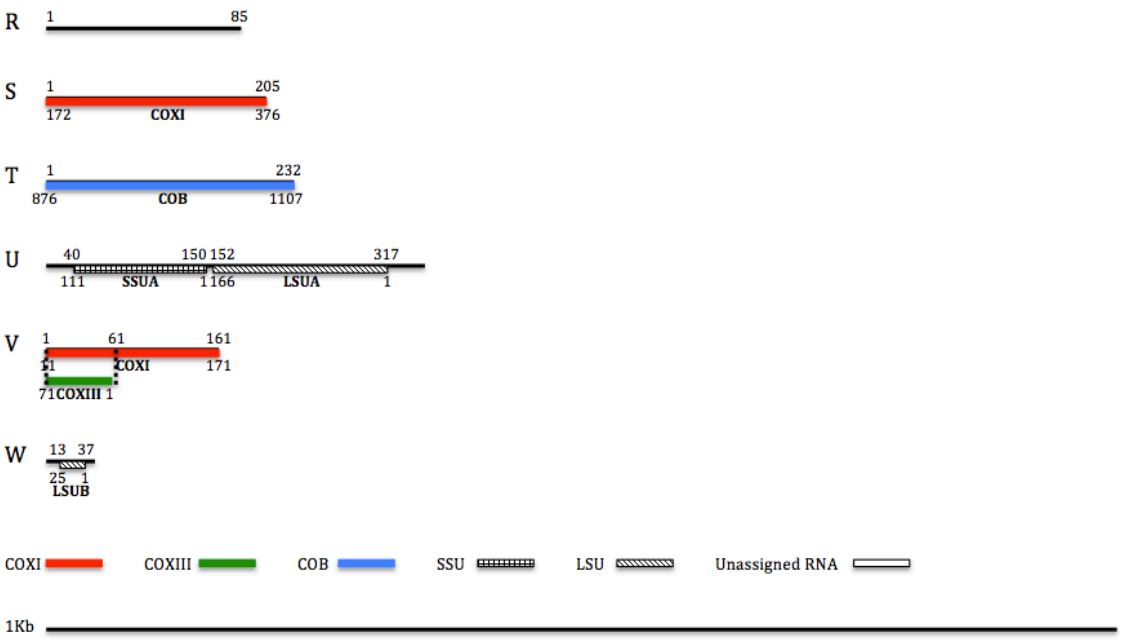
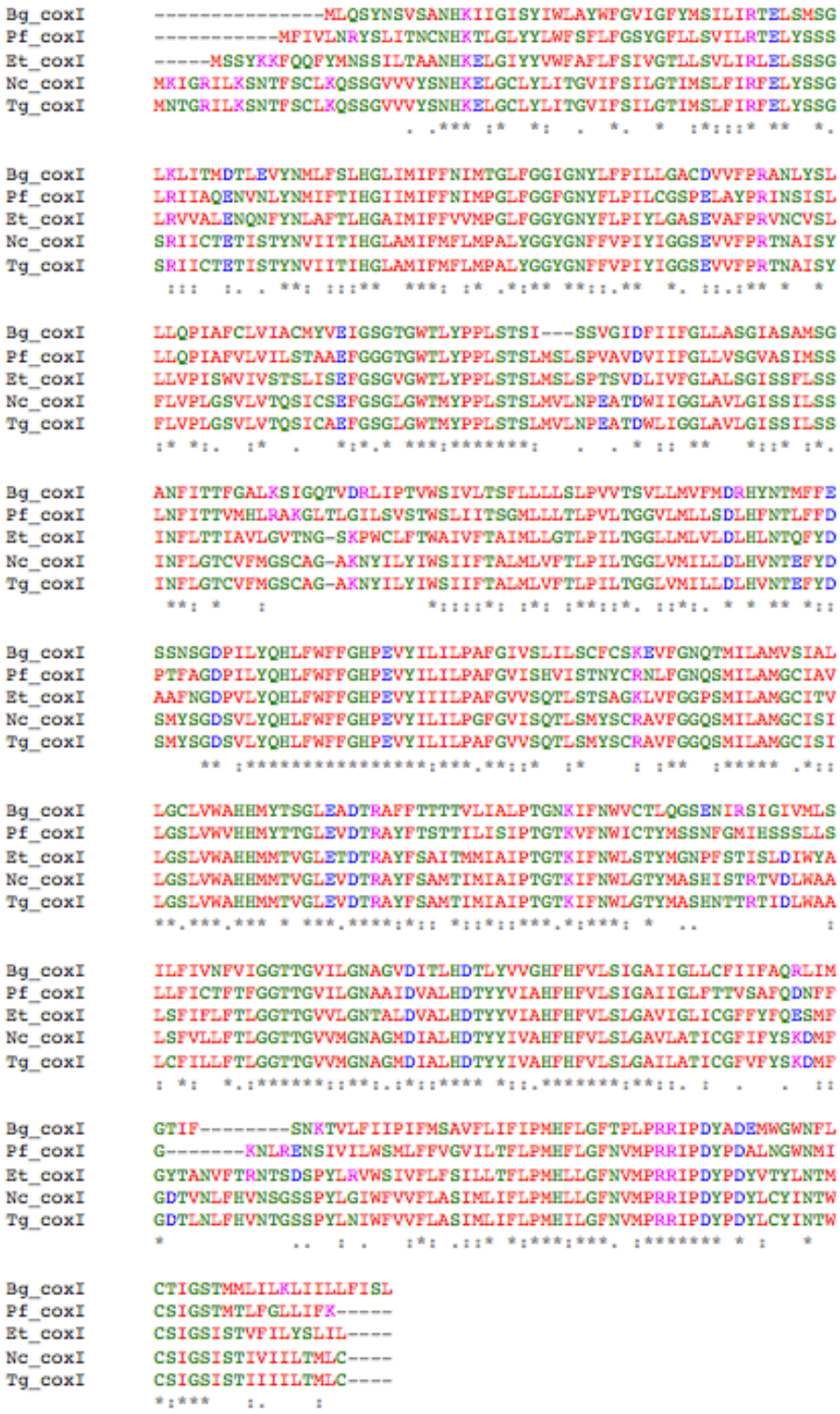


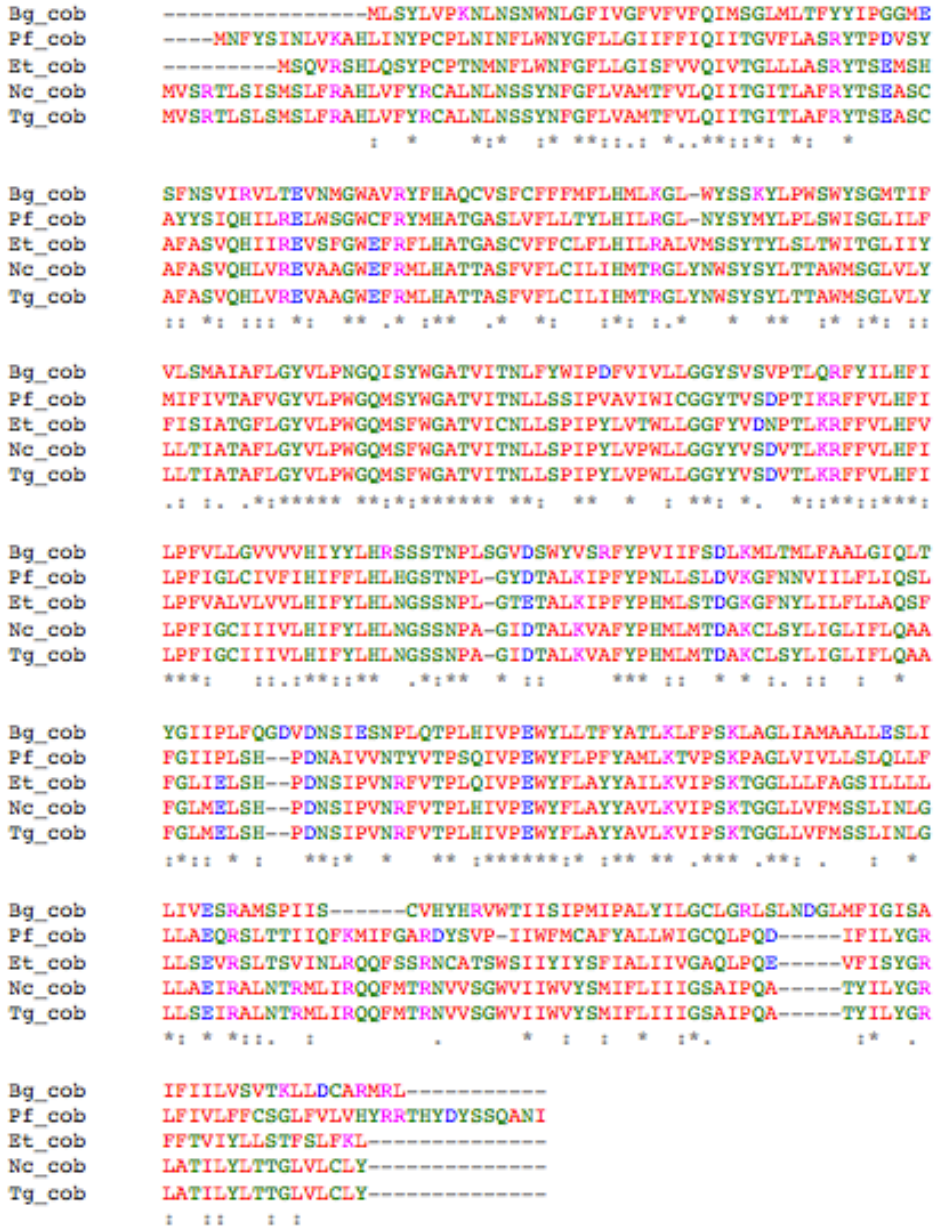
Figure 6.1. Annotation of the 23 mtDNA elements from *T. gondii*

23 sequence elements that make up the *T. gondii* mtDNA were identified. The three cytochrome genes and some of the LSU and SSU rRNA features are annotated on these elements. It is possible to assemble the cytochrome genes from these elements without fragmenting any element.

A.


Bg_coxI ----------MLQSYNSVSANHKIIGISYIWLAYWFGVIGFYMSILIRTELSMSG
Pf_coxI ----------MFIVLNRYSLITNCNHKTLGLYYLWFSFLFGSYGFLLSVILRTELYSSS
Et_coxI -----MSSYKKFQQFYMNSSILTAANHKELGIYYVWFAFLFSIVGTLLSVLIRLELSSSG
Nc_coxI MKIGRILKSNTFSCLKQSSGVVVYSNHKELGCLYLITGVIFSILGTIMSLFIRFELYSSG
Tg_coxI MNTGRILKSNTFSCLKQSSGVVVYSNHKELGCLYLITGVIFSILGTIMSLFIRFELYSSG
LKLITMDTLEVYNMLFSLHGLIMIFFNIMTGLFGGIGNYLFPILLGACDVVFPRANLYSL
Pf_coxI LRIIAQENVNLYNMIFTIHGIIMIFFNIMPGLFGGFGNYFLPILCGSPELAYPRINSISL
Et_coxI LRVVALENQNFYNLAFTLHGAIMIFFVVMPGLFGGYGNYFLPIYLGASEVAFPRVNCVSL
Nc_coxI SRIICTETISTYNVIITIHGLAMIFMFLMPALYGGYGNFFVPIYIGGSEVVFPRTNAISY
Tg_coxI SRIICTETISTYNVIITIHGLAMIFMFLMPALYGGYGNFFVPIYIGGSEVVFPRTNAISY
LLQPIAFCLVIACMYVEIGSGTGWTLYPPLSTSI---SSVGIDFIIFGLLASGIASAMSG
Pf_coxI LLQPIAFVLVILSTAAEFGGGTGWTLYPPLSTSLMSLSPVAVDVIIFGLLVSGVASIMSS
Et_coxI LLVPISWVIVSTSLISEFGSGVGWTLYPPLSTSLMSLSPTSVDLIVFGLALSGISSFLSS
Nc_coxI FLVPLGSVLVTQSICSEFGSGLGWTMYPPLSTSLMVLNPEATDWIIGGLAVLGISSILSS
Tg_coxI FLVPLGSVLVTQSICAEFGSGLGWTMYPPLSTSLMVLNPEATDWLIGGLAVLGISSILSS
ANFITTFGALKSIGQTVDRLIPTVWSIVLTSFLLLLSLPVVTSVLLMVFMDRHYNTMFFE
Pf_coxI LNFITTVMHLRAKGLTLGILSVSTWSLIITSGMLLLTLPVLTGGVLMLLSDLHFNTLFFD
Et_coxI INFLTTIAVLGVTNG-SKPWCLFTWAIVFTAIMLLGTLPILTGGLLMLVLDLHLNTQFYD
Nc_coxI INFLGTCVFMGSCAG-AKNYILYIWSIIFTALMLVFTLPILTGGLVMILLDLHVNTEFYD
Tg_coxI INFLGTCVFMGSCAG-AKNYILYIWSIIFTALMLVFTLPILTGGLVMILLDLHVNTEFYD
SSNSGDPILYQHLFWFFGHPEVYILILPAFGIVSLILSCFCSKEVFGNQTMILAMVSIAL
Pf_coxI PTFAGDPILYQHLFWFFGHPEVYILILPAFGVISHVISTNYCRNLFGNQSMILAMGCIAV
Et_coxI AAFNGDPVLYQHLFWFFGHPEVYIIILPAFGVVSQTLSTSAGKLVFGGPSMILAMGCITV
Nc_coxI SMYSGDSVLYQHLFWFFGHPEVYILILPGFGVISQTLSMYSCRAVFGGQSMILAMGCISI
Tg_coxI SMYSGDSVLYQHLFWFFGHPEVYILILPAFGVVSQTLSMYSCRAVFGGQSMILAMGCISI
LGCLVWAHHMYTSGLEADTRAFFTTTTVLIALPTGNKIFNWVCTLQGSENIRSIGIVMLS
Pf_coxI LGSLVWVHHMYTTGLEVDTRAYFTSTTILISIPTGTKVFNWICTYMSSNFGMIHSSSLLS
Et_coxI LGSLVWAHHMMTVGLETDTRAYFSAITMMIAIPTGTKIFNWLSTYMGNPFSTISLDIWYA
Nc_coxI LGSLVWAHHMMTVGLEVDTRAYFSAMTIMIAIPTGTKIFNWLGTYMASHISTRTVDLWAA
Tg_coxI LGSLVWAHHMMTVGLEVDTRAYFSAMTIMIAIPTGTKIFNWLGTYMASHNTTRTIDLWAA
ILFIVNFVIGGTTGVILGNAGVDITLHDTLYVVGHFHFVLSIGAIIGLLCFIIFAQRLIM
Pf_coxI LLFICTFTFGGTTGVILGNAAIDVALHDTYYVIAHFHFVLSIGAIIGLFTTVSAFQDNFF
Et_coxI LSFIFLFTLGGTTGVVLGNTALDVALHDTYYVIAHFHFVLSLGAVIGLICGFFYFQESMF
Nc_coxI LSFVLLFTLGGTTGVVMGNAGMDIALHDTYYIVAHFHFVLSLGAVLATICGFIFYSKDMF
Tg_coxI LCFILLFTLGGTTGVVMGNAGMDIALHDTYYIVAHFHFVLSLGAILATICGFVFYSKDMF
GTIF--------SNKTVLFIIPIFMSAVFLIFIPMHFLGFTPLPRRIPDYADEMWGWNFL
Pf_coxI G--------KNLRENSIVILWSMLFFVGVILTFLPMHFLGFNVMPRRIPDYPDALNGWNMI
Et_coxI GYTANVFTRNTSDSPYLRVWSIVFLFSILLTFLPMHLLGFNVMPRRIPDYPDYVTYLNTM
Nc_coxI GDTVNLFHVNSGSSPYLGIWFVVFLASIMLIFLPMHLLGFNVMPRRIPDYPDYLCYINTW
Tg_coxI GDTLNLFHVNTGSSPYLNIWFVVFLASIMLIFLPMHILGFNVMPRRIPDYPDYLCYINTW
CTIGSTMMLILKLIILLFISL
Pf_coxI CSIGSTMTLFGLLIFK-----
Et_coxI CSIGSISTVFILYSLIL----
Nc_coxI CSIGSISTIVIILTMLC----
Tg_coxI CSIGSISTIIIILTMLC----

B.


Bg_cob ----------MLSYLVPKNLNSNWNLGFIVGFVFVFQIMSGLMLTFYYIPGGME
Pf_cob ----MNFYSINLVKAHLINYPCPLNINFLWNYGFLLGIIFFIQIITGVFLASRYTPDVSY
Et_cob ---------MSQVRSHLQSYPCPTNMNFLWNFGFLLGISFVVQIVTGLLLASRYTSEMSH
Nc_cob MVSRTLSISMSLFRAHLVFYRCALNLNSSYNFGFLVAMTFVLQIITGITLAFRYTSEASC
Tg_cob MVSRTLSLSMSLFRAHLVFYRCALNLNSSYNFGFLVAMTFVLQIITGITLAFRYTSEASC
SFNSVIRVLTEVNMGWAVRYFHAQCVSFCFFFMFLHMLKGL-WYSSKYLPWSWYSGMTIF
AYYSIQHILRELWSGWCFRYMHATGASLVFLLTYLHILRGL-NYSYMYLPLSWISGLILF
AFASVQHIIREVSFGWEFRFLHATGASCVFFCLFLHILRALVMSSYTYLSLTWITGLIIY
AFASVQHLVREVAAGWEFRMLHATTASFVFLCILIHMTRGLYNWSYSYLTTAWMSGLVLY
AFASVQHLVREVAAGWEFRMLHATTASFVFLCILIHMTRGLYNWSYSYLTTAWMSGLVLY
VLSMAIAFLGYVLPNGQISYWGATVITNLFYWIPDFVIVLLGGYSVSVPTLQRFYILHFI
MIFIVTAFVGYVLPWGQMSYWGATVITNLLSSIPVAVIWICGGYTVSDPTIKRFFVLHFI
FISIATGFLGYVLPWGQMSFWGATVICNLLSPIPYLVTWLLGGFYVDNPTLKRFFVLHFV
LLTIATAFLGYVLPWGQMSFWGATVITNLLSPIPYLVPWLLGGYYVSDVTLKRFFVLHFI
LLTIATAFLGYVLPWGQMSFWGATVITNLLSPIPYLVPWLLGGYYVSDVTLKRFFVLHFI
LPFVLLGVVVVHIYYLHRSSSTNPLSGVDSWYVSRFYPVIIFSDLKMLTMLFAALGIQLT
LPFIGLCIVFIHIFFLHLHGSTNPL-GYDTALKIPFYPNLLSLDVKGFNNVIILFLIQSL
LPFVALVLVVLHIFYLHLNGSSNPL-GTETALKIPFYPHMLSTDGKGFNYLILFLLAQSF
LPFIGCIIIVLHIFYLHLNGSSNPA-GIDTALKVAFYPHMLMTDAKCLSYLIGLIFLQAA
LPFIGCIIIVLHIFYLHLNGSSNPA-GIDTALKVAFYPHMLMTDAKCLSYLIGLIFLQAA
YGIIPLFQGDVDNSIESNPLQTPLHIVPEWYLLTFYATLKLFPSKLAGLIAMAALLESLI
FGIIPLSH--PDNAIVVNTYVTPSQIVPEWYFLPFYAMLKTVPSKPAGLVIVLLSLQLLF
FGLIELSH--PDNSIPVNRFVTPLQIVPEWYFLAYYAILKVIPSKTGGLLLFAGSILLLL
FGLMELSH--PDNSIPVNRFVTPLHIVPEWYFLAYYAVLKVIPSKTGGLLVFMSSLINLG
FGLMELSH--PDNSIPVNRFVTPLHIVPEWYFLAYYAVLKVIPSKTGGLLVFMSSLINLG
LIVESRAMSPIIS------CVHYHRVWTIISIPMIPALYILGCLGRLSLNDGLMFIGISA
LLAEQRSLTTIIQFKMIFGARDYSVP-IIWFMCAFYALLWIGCQLPQD-----IFILYGR
LLSEVRSLTSVINLRQQFSSRNCATSWSIIYIYSFIALIIVGAQLPQE-----VFISYGR
LLAEIRALNTRMLIRQQFMTRNVVSGWVIIWVYSMIFLIIIGSAIPQA-----TYILYGR
LLSEIRALNTRMLIRQQFMTRNVVSGWVIIWVYSMIFLIIIGSAIPQA-----TYILYGR
IFIILVSVTKLLDCARMRL----------
LFIVLFFCSGLFVLVHYRRTHYDYSSQANI
FFTVIYLLSTFSLFKL--------------
LATILYLTTGLVLCLY--------------
LATILYLTTGLVLCLY--------------

C.

Figure 6.2. Multiple Sequence Alignment of cytochrome proteins

Protein sequences were aligned using Clustal Omega. *T. gondii* and *N. caninum* sequences were assembled using the 23 mtDNA elements present in each species. Sequences for the other species were downloaded from GenBank. GenBank Ids for COXI, COB and COXIII for each species: *P. falciparum*: AIJ28956.1, AIJ28888.1, AIJ28955.1. *B. gibsoni*: BAL73004.1, BAL73003.1, BAL73002.1. *E. tenella*: BAJ25753.1, BAJ25752.1, BAJ25754.1.

Table 6.1. 83 non-chromosomal contigs showing >98% identity to the mtDNA elements

TgME49 Contig	Start	End	Divergence from mtDNA (%)	mtDNA element/gene	Start	End
ABPA02001282	1	423	0.5	K	22	445
ABPA02001282	424	505	0	D	1	82
ABPA02001282	506	666	0	coxI	11	171
ABPA02001282	596	1330	0	coxIII	1	735
ABPA02001286	1	82	0	coxI	90	171
ABPA02001286	83	164	0	D	1	82
ABPA02001286	165	609	0	K	1	445
ABPA02001286	610	695	0	O	1	86
ABPA02001286	696	780	0	J	1	85
ABPA02001286	786	988	0	cob	1	203
ABPA02001287	1	51	0	N	1	52
ABPA02001287	52	480	0.7	cob	679	1107
ABPA02001287	482	526	0	W	1	45
ABPA02001287	517	1093	0.7	coxI	1	576
ABPA02001290	1	677	0	cob	199	875
ABPA02001290	677	860	0	P	1	184
ABPA02001290	861	916	1.8	H	392	447
ABPA02001296	1	203	0.5	I	1	203
ABPA02001296	204	648	0	K	1	445
ABPA02001296	649	734	0	O	1	86
ABPA02001296	735	888	0	coxIII	231	385
ABPA02001309	1	298	2	M	458	754
ABPA02001309	299	477	0	F	1	179
ABPA02001309	478	563	0	O	1	86
ABPA02001309	564	792	0	coxIII	231	459
ABPA02001309	794	1144	0	U	1	353
ABPA02001638	1	85	0	coxIII	1	85
ABPA02001638	15	175	0	coxI	11	171
ABPA02001638	176	257	0	D	1	82
ABPA02001638	258	369	0	K	334	445
ABPA02001638	370	722	0	U	1	353
ABPA02001638	724	1183	0	coxIII	1	459
ABPA02001639	1	351	0	U	1	351
ABPA02001639	353	811	0	coxIII	1	459
ABPA02001639	741	901	0	coxI	11	171

ABPA02001639	902	983	0	D	1	82
ABPA02001639	984	1095	0	K	334	445
ABPA02001639	1096	1398	0	U	51	353
ABPA02001640	1	330	0	coxIII	1	327
ABPA02001640	260	420	0	coxI	11	171
ABPA02001640	421	502	0	D	1	82
ABPA02001640	503	614	0	K	334	445
ABPA02001640	615	967	0	U	1	353
ABPA02001640	969	1043	0	coxIII	384	459
ABPA02001669	1	442	0	M	313	754
ABPA02001669	443	621	0	F	1	179
ABPA02001669	622	707	0	O	1	86
ABPA02001669	708	1212	0	coxIII	231	735
ABPA02001669	833	1397	0	M	1	565
ABPA02001670	1	139	0	F	1	138
ABPA02001670	140	225	0	O	1	86
ABPA02001670	226	730	0	coxIII	231	735
ABPA02001670	351	1104	0	M	1	754
ABPA02001670	1105	1153	0	F	131	179
ABPA02001671	13	766	0	M	1	754
ABPA02001671	767	857	0	F	89	179
ABPA02001671	858	1023	0	N	1	166
ABPA02001671	1024	1174	0.7	cob	679	815
ABPA02001672	1	349	0	coxI	28	376
ABPA02001672	350	434	0	R	1	85
ABPA02001672	435	525	0	F	89	179
ABPA02001672	526	691	0	N	1	166
ABPA02001672	692	888	0	cob	679	875
ABPA02001672	888	977	0	P	95	184
ABPA02001673	1	43	0	R	1	44
ABPA02001673	44	134	0	F	89	179
ABPA02001673	135	300	0	N	1	166
ABPA02001673	301	497	0	cob	679	875
ABPA02001673	497	680	0	P	1	184
ABPA02001673	681	884	0	I	1	204
ABPA02001673	885	996	0.9	K	334	445
ABPA02001674	96	261	0	N	1	166
ABPA02001674	262	458	0	cob	679	875
ABPA02001674	458	641	0	P	1	184
ABPA02001674	642	845	0	I	1	204
ABPA02001674	846	957	0	K	334	445
ABPA02001674	958	1193	0	U	118	353

ABPA02001675	1	134	0	cob	742	875
ABPA02001675	134	317	0	P	1	184
ABPA02001675	318	521	0	I	1	204
ABPA02001675	522	633	0	K	334	445
ABPA02001675	634	986	0	U	1	353
ABPA02001675	988	1160	0	coxIII	287	459
ABPA02001701	1	287	0	coxI	11	297
ABPA02001701	217	533	0.3	coxIII	1	317
ABPA02001701	532	646	0	coxI	1377	1491
ABPA02001701	597	989	0.5	H	1	393
ABPA02001702	3	113	0	coxI	11	121
ABPA02001702	43	359	0.3	coxIII	1	317
ABPA02001702	358	472	0	coxI	1377	1491
ABPA02001702	423	869	0.5	H	1	447
ABPA02001702	870	1053	0	P	1	184
ABPA02001702	1056	1112	0	cob	815	871
ABPA02001703	1	74	0	J	1	74
ABPA02001703	75	189	0	coxI	1377	1491
ABPA02001703	140	586	0.5	H	1	447
ABPA02001703	587	770	0	P	1	184
ABPA02001703	770	998	0.5	cob	656	875
ABPA02001704	3	63	0	coxI	1377	1437
ABPA02001704	14	460	0.5	H	1	447
ABPA02001704	461	644	0	P	1	184
ABPA02001704	644	841	0	cob	679	875
ABPA02001704	842	955	3.5	N	1	113
ABPA02001728	1	605	0	M	1	605
ABPA02001728	226	730	0	coxIII	231	735
ABPA02001728	731	816	0	O	1	86
ABPA02001728	817	1261	0	K	1	445
ABPA02001728	1264	1320	0	D	3	59
ABPA02001804	1	246	0	U	109	353
ABPA02001804	247	358	0	K	334	445
ABPA02001804	359	440	0	D	1	82
ABPA02001804	441	601	0	coxI	11	171
ABPA02001804	531	847	0.3	coxIII	1	317
ABPA02001804	846	1012	0	coxI	1324	1491
ABPA02001805	1	108	0.9	coxI	11	118
ABPA02001805	38	354	0.6	coxIII	1	317
ABPA02001805	353	1172	0.1	coxI	672	1491
ABPA02001806	1	862	0.5	coxI	11	871
ABPA02001806	792	1223	0	coxIII	1	432

ABPA02001807	1	157	0.6	coxI	11	166
ABPA02001807	87	821	0	coxIII	1	735
ABPA02001807	442	1194	0	M	1	754
ABPA02001809	1	306	0.3	K	139	445
ABPA02001809	307	388	0	D	1	82
ABPA02001809	389	549	0	coxI	11	171
ABPA02001809	479	795	0.3	coxIII	1	317
ABPA02001809	794	903	0.9	coxI	1383	1491
ABPA02001815	1	782	0	cob	1	782
ABPA02001815	786	976	0.5	coxIII	127	317
ABPA02001816	3	662	0	cob	1	660
ABPA02001816	666	933	0.8	coxIII	46	317
ABPA02001817	1	387	0.5	cob	1	387
ABPA02001817	391	707	0.3	coxIII	1	317
ABPA02001817	637	797	0	coxI	11	171
ABPA02001817	798	879	0	D	1	82
ABPA02001817	880	973	2.1	K	351	445
ABPA02001821	1	357	0.8	H	90	447
ABPA02001821	358	541	0	P	1	184
ABPA02001821	541	1129	0.5	cob	289	875
ABPA02001827	1	85	0	coxIII	375	459
ABPA02001827	87	439	0	U	1	353
ABPA02001827	440	551	0	K	334	445
ABPA02001827	552	633	0	D	1	82
ABPA02001827	634	804	0	coxI	1	171
ABPA02001827	795	839	0	W	1	45
ABPA02001827	840	1001	0	T	73	233
ABPA02001833	1	149	0	coxI	228	376
ABPA02001833	150	234	0	R	1	85
ABPA02001833	235	325	0	F	89	179
ABPA02001833	326	491	0	N	1	166
ABPA02001833	492	920	0	cob	679	1107
ABPA02001833	922	966	0	W	1	45
ABPA02001833	957	1148	0	coxI	1	192
ABPA02001841	14	99	0	O	1	86
ABPA02001841	100	184	0	J	1	85
ABPA02001841	185	979	0	coxI	697	1491
ABPA02001843	3	608	0.2	cob	1	606
ABPA02001843	614	698	0	J	1	85
ABPA02001843	699	784	0	O	1	86
ABPA02001843	785	927	0	F	1	142
ABPA02001844	16	180	0	N	1	166

ABPA02001844	181	377	0	cob	679	875
ABPA02001844	377	560	0	P	1	184
ABPA02001844	561	869	1	H	139	447
ABPA02001847	1	165	0	F	14	179
ABPA02001847	166	250	0	R	1	85
ABPA02001847	251	616	0	coxI	11	376
ABPA02001847	546	1004	0	coxIII	1	459
ABPA02001847	1006	1032	0	U	1	27
ABPA02001860	25	137	0	K	334	445
ABPA02001860	138	490	0	U	1	353
ABPA02001860	492	720	0	coxIII	231	459
ABPA02001860	721	806	0	O	1	86
ABPA02001860	807	930	0	F	1	124
ABPA02001861	6	360	0.6	U	1	353
ABPA02001861	362	590	0	coxIII	231	459
ABPA02001861	591	676	0	O	1	86
ABPA02001861	677	855	0	F	1	179
ABPA02001861	856	940	0	R	1	85
ABPA02001861	941	1246	0	coxI	72	376
ABPA02001862	1	134	0	coxIII	231	364
ABPA02001862	135	220	0	O	1	86
ABPA02001862	221	399	0	F	1	179
ABPA02001862	400	484	0	R	1	85
ABPA02001862	485	850	0	coxI	11	376
ABPA02001862	780	975	0	coxIII	1	196
ABPA02001866	3	142	0	I	1	140
ABPA02001866	143	254	0	K	334	445
ABPA02001866	255	607	0	U	1	353
ABPA02001866	609	837	0	coxIII	231	459
ABPA02001866	838	923	0	O	1	86
ABPA02001866	927	975	0	K	2	50
ABPA02001873	1	134	0	P	50	184
ABPA02001873	134	330	0	cob	679	875
ABPA02001873	331	496	0	N	1	166
ABPA02001873	497	587	0	F	89	179
ABPA02001873	588	976	0	M	366	754
ABPA02001877	1	443	0.5	K	1	443
ABPA02001877	444	529	0	O	1	86
ABPA02001877	530	758	0	coxIII	231	459
ABPA02001877	760	1112	0	U	1	353
ABPA02001877	1113	1224	0.9	K	334	445
ABPA02001877	1225	1277	1.9	I	1	53

ABPA02001886	1	157	1.3	P	1	158
ABPA02001886	158	604	0.5	H	1	447
ABPA02001886	555	669	0	coxI	1377	1491
ABPA02001886	670	754	0	J	1	85
ABPA02001886	755	840	0	O	1	86
ABPA02001886	841	981	0	F	1	141
tgme49_asmbl.1156	1	576	0	M	178	754
tgme49_asmbl.1156	577	667	0	F	89	179
tgme49_asmbl.1156	668	833	0	N	1	166
tgme49_asmbl.1156	834	1013	0.6	cob	679	858
tgme49_asmbl.1157	1	38	0	M	718	754
tgme49_asmbl.1157	39	132	0	F	89	179
tgme49_asmbl.1157	133	298	0	N	1	166
tgme49_asmbl.1157	299	727	0	cob	679	1107
tgme49_asmbl.1157	729	773	0	W	1	45
tgme49_asmbl.1157	764	1002	0	coxI	1	239
tgme49_asmbl.1158	1	252	0	cob	856	1107
tgme49_asmbl.1158	254	298	0	W	1	45
tgme49_asmbl.1158	289	664	0	coxI	1	376
tgme49_asmbl.1158	665	749	0	R	1	85
tgme49_asmbl.1158	750	840	0	F	89	179
tgme49_asmbl.1158	841	1006	0	N	1	166
tgme49_asmbl.1158	1007	1434	1.2	cob	679	1107
tgme49_asmbl.1158	1436	1480	0	W	1	45
tgme49_asmbl.1158	1471	1846	0	coxI	1	376
tgme49_asmbl.1158	1847	1931	0	R	1	85
tgme49_asmbl.1158	1932	2028	2.1	F	81	179
tgme49_asmbl.1159	1	206	1	coxI	171	376
tgme49_asmbl.1159	207	291	0	R	1	85
tgme49_asmbl.1159	292	382	0	F	89	179
tgme49_asmbl.1159	383	548	0	N	1	166
tgme49_asmbl.1159	549	745	0	cob	679	875
tgme49_asmbl.1159	745	928	0	P	1	184
tgme49_asmbl.1159	929	1130	0.5	I	3	204
tgme49_asmbl.1159	1133	1185	0	K	393	445
tgme49_asmbl.1169	1	110	0	I	1	110
tgme49_asmbl.1169	111	555	0	K	1	445
tgme49_asmbl.1169	556	641	0	O	1	86
tgme49_asmbl.1169	642	726	0	J	1	85
tgme49_asmbl.1169	732	1059	0.3	cob	1	325
tgme49_asmbl.1177	1	225	0	coxIII	231	454
tgme49_asmbl.1177	226	311	0	O	1	86

tgme49_asmbl.1177	312	756	0	K	1	445
tgme49_asmbl.1177	757	838	0	D	1	82
tgme49_asmbl.1177	839	999	0	coxI	11	171
tgme49_asmbl.1177	929	1245	0.3	coxIII	1	317
tgme49_asmbl.1177	1244	1358	0	coxI	1377	1491
tgme49_asmbl.1177	1309	1667	0.3	H	1	361
tgme49_asmbl.1181	1	193	0	U	161	353
tgme49_asmbl.1181	194	305	0	K	334	445
tgme49_asmbl.1181	306	387	0	D	1	82
tgme49_asmbl.1181	388	548	0	coxI	11	171
tgme49_asmbl.1181	478	794	0.3	coxIII	1	317
tgme49_asmbl.1181	798	1025	0.4	cob	1	228
tgme49_asmbl.1192	1	532	0	coxI	11	541
tgme49_asmbl.1192	462	1196	0	coxIII	1	735
tgme49_asmbl.1192	817	1245	0	M	1	429
tgme49_asmbl.1193	3	89	0	O	1	86
tgme49_asmbl.1193	90	534	0	K	1	445
tgme49_asmbl.1193	535	616	0	D	1	82
tgme49_asmbl.1193	617	777	0	coxI	11	171
tgme49_asmbl.1193	707	1003	0	coxIII	1	297
tgme49_asmbl.1199	1	796	0	coxI	1	794
tgme49_asmbl.1199	787	831	0	W	1	45
tgme49_asmbl.1199	833	1939	0.6	cob	1	1107
tgme49_asmbl.1199	1945	2029	0	J	1	85
tgme49_asmbl.1199	2030	2115	0	O	1	86
tgme49_asmbl.1199	2116	2439	0.3	K	1	324
tgme49_asmbl.1200	1	176	0.6	coxI	1	176
tgme49_asmbl.1200	167	211	0	W	1	45
tgme49_asmbl.1200	213	1319	0	cob	1	1107
tgme49_asmbl.1200	1323	1639	0.9	coxIII	1	317
tgme49_asmbl.1200	1569	1700	1.5	coxI	11	142
tgme49_asmbl.1201	1	104	1	P	81	184
tgme49_asmbl.1201	104	674	0	cob	305	875
tgme49_asmbl.1201	660	769	0.9	coxI	11	120
tgme49_asmbl.1201	699	1016	0.3	coxIII	1	317
tgme49_asmbl.1201	1015	1594	0	coxI	912	1491
tgme49_asmbl.1206	1	30	0	W	16	45
tgme49_asmbl.1206	21	191	0	coxI	1	171
tgme49_asmbl.1206	192	273	0	D	1	82
tgme49_asmbl.1206	274	718	0	K	1	445
tgme49_asmbl.1206	719	804	0	O	1	86
tgme49_asmbl.1206	805	889	0	J	1	85

tgme49_asmbl.1206	890	1004	0	coxI	1377	1491
tgme49_asmbl.1206	955	1016	0	H	1	62
tgme49_asmbl.1207	2	395	0	K	1	394
tgme49_asmbl.1207	396	481	0	O	1	86
tgme49_asmbl.1207	482	566	0	J	1	85
tgme49_asmbl.1207	567	1189	0	coxI	869	1491
tgme49_asmbl.1216	3	244	0	U	1	239
tgme49_asmbl.1216	246	474	0	coxIII	231	459
tgme49_asmbl.1216	475	560	0	O	1	86
tgme49_asmbl.1216	561	733	0	K	1	173
tgme49_asmbl.1216	717	779	3.2	F	89	151
tgme49_asmbl.1216	780	945	0	N	1	166
tgme49_asmbl.1216	946	1076	0	cob	679	808
tgme49_asmbl.1219	7	92	0	O	1	86
tgme49_asmbl.1219	93	271	0	F	1	179
tgme49_asmbl.1219	269	419	0.7	H	297	447
tgme49_asmbl.1219	420	603	0	P	1	184
tgme49_asmbl.1219	603	799	0	cob	679	875
tgme49_asmbl.1219	800	965	0	N	1	166
tgme49_asmbl.1219	966	1056	0	F	89	179
tgme49_asmbl.1219	1057	1347	1.7	M	462	754
tgme49_asmbl.1225	1	356	0.6	H	1	356
tgme49_asmbl.1225	307	421	0	coxI	1377	1491
tgme49_asmbl.1225	422	506	0	J	1	85
tgme49_asmbl.1225	507	592	0	O	1	86
tgme49_asmbl.1225	593	771	0	F	1	179
tgme49_asmbl.1225	772	856	0	R	1	85
tgme49_asmbl.1225	857	1161	0	coxI	72	376
tgme49_asmbl.1875	1	186	0	I	19	204
tgme49_asmbl.1875	187	370	0	P	1	184
tgme49_asmbl.1875	370	1244	0.1	cob	1	875
tgme49_asmbl.1875	1250	1334	0	J	1	85
tgme49_asmbl.1875	1335	1420	0	O	1	86
tgme49_asmbl.1875	1421	1599	0	F	1	179
tgme49_asmbl.1875	1600	1684	0	R	1	85
tgme49_asmbl.1875	1685	2060	0	coxI	1	376
tgme49_asmbl.1875	2051	2095	0	W	1	45
tgme49_asmbl.1875	2096	2251	0	T	78	233
tgme49_asmbl.1876	1	138	0	coxIII	1	139
tgme49_asmbl.1876	68	228	0	coxI	11	171
tgme49_asmbl.1876	229	310	0	D	1	82
tgme49_asmbl.1876	311	422	0	K	334	445

tgme49_asmbl.1876	423	775	0	U	1	353
tgme49_asmbl.1876	777	1005	0	coxIII	231	459
tgme49_asmbl.1879	4	448	0	M	310	754
tgme49_asmbl.1879	449	539	0	F	89	179
tgme49_asmbl.1879	540	705	0	N	1	166
tgme49_asmbl.1879	706	1134	0	cob	679	1107
tgme49_asmbl.1879	1136	1180	0	W	1	45
tgme49_asmbl.1879	1171	1460	0	coxI	1	292
tgme49_asmbl.1880	1	517	0	coxI	975	1491
tgme49_asmbl.1880	518	602	0	J	1	85
tgme49_asmbl.1880	603	688	0	O	1	86
tgme49_asmbl.1880	689	867	0	F	1	179
tgme49_asmbl.1880	868	952	0	R	1	85
tgme49_asmbl.1880	953	1095	0	coxI	230	376
tgme49_asmbl.1881	1	46	0	coxIII	231	276
tgme49_asmbl.1881	47	132	0	O	1	86
tgme49_asmbl.1881	133	576	0.5	K	1	445
tgme49_asmbl.1881	577	780	0	I	1	204
tgme49_asmbl.1881	781	964	0	P	1	184
tgme49_asmbl.1881	964	1160	0	cob	679	875
tgme49_asmbl.1882	1	49	0	coxIII	231	279
tgme49_asmbl.1882	50	135	0	O	1	86
tgme49_asmbl.1882	136	580	0.7	K	1	445
tgme49_asmbl.1882	581	662	0	D	1	82
tgme49_asmbl.1882	663	833	0	coxI	1	171
tgme49_asmbl.1882	824	868	0	W	1	45
tgme49_asmbl.1882	870	1298	0.2	cob	679	1107
tgme49_asmbl.1882	1299	1464	0	N	1	166
tgme49_asmbl.1882	1465	1555	0	F	89	179
tgme49_asmbl.1882	1556	1640	0	R	1	85
tgme49_asmbl.1882	1641	2006	0	coxI	11	376
tgme49_asmbl.1882	1936	2253	0.3	coxIII	1	317
tgme49_asmbl.1882	2252	2304	0	coxI	1437	1491
tgme49_asmbl.1927	1	411	0	coxIII	48	459
tgme49_asmbl.1927	413	765	0	U	1	353
tgme49_asmbl.1927	766	877	0	K	334	445
tgme49_asmbl.1927	878	1004	0	I	1	127
tgme49_asmbl.28	1	545	0	coxI	946	1491
tgme49_asmbl.28	544	860	0.3	coxIII	1	317
tgme49_asmbl.28	790	1155	0	coxI	11	376
tgme49_asmbl.29	11	176	0	N	1	166
tgme49_asmbl.29	177	373	0	cob	679	875

tgme49_asmbl.29	373	556	0	P	1	184
tgme49_asmbl.29	557	760	0	I	1	204
tgme49_asmbl.29	761	1205	0	K	1	445
tgme49_asmbl.29	1206	1291	0	O	1	86
tgme49_asmbl.29	1292	1367	0	J	5	85
tgme49_asmbl.29	1378	2248	0	cob	4	875
tgme49_asmbl.29	2248	2431	0	P	1	184
tgme49_asmbl.29	2432	2617	0	I	19	204
tgme49_asmbl.33	1	1462	0.5	coxI	32	1491
tgme49_asmbl.33	1463	1547	0	J	1	85
tgme49_asmbl.33	1548	1633	0	O	1	86
tgme49_asmbl.33	1634	2078	0	K	1	445
tgme49_asmbl.33	2079	2282	0	I	1	204
tgme49_asmbl.33	2283	2361	0	P	1	80
tgme49_asmbl.35	2	545	0	cob	563	1107
tgme49_asmbl.35	547	591	0	W	1	45
tgme49_asmbl.35	582	752	0	coxI	1	171
tgme49_asmbl.35	753	834	0	D	1	82
tgme49_asmbl.35	835	946	0	K	334	445
tgme49_asmbl.35	947	1298	1.1	U	1	353
tgme49_asmbl.35	1300	1758	0.2	coxIII	1	459
tgme49_asmbl.35	1688	2033	0	coxI	11	358
tgme49_asmbl.38	1	478	0.4	M	278	754
tgme49_asmbl.38	479	657	0	F	1	179
tgme49_asmbl.38	658	743	0	O	1	86
tgme49_asmbl.38	744	828	0	J	1	85
tgme49_asmbl.38	829	1273	0	coxI	1047	1491
tgme49_asmbl.39	26	1074	0	cob	1	1049
tgme49_asmbl.40	1	422	0.7	cob	1	423
tgme49_asmbl.40	426	742	0.3	coxIII	1	317
tgme49_asmbl.40	672	1037	0	coxI	11	376
tgme49_asmbl.40	1038	1098	0	R	25	85
tgme49_asmbl.42	19	523	0	coxIII	231	735
tgme49_asmbl.42	524	609	0	O	1	86
tgme49_asmbl.42	610	788	0	F	1	179
tgme49_asmbl.42	789	873	0	R	1	85
tgme49_asmbl.42	874	1109	0	coxI	141	376
tgme49_asmbl.44	5	181	1.7	coxI	200	376
tgme49_asmbl.44	182	266	0	R	1	85
tgme49_asmbl.44	267	357	0	F	89	179
tgme49_asmbl.44	358	523	0	N	1	166
tgme49_asmbl.44	524	720	0	cob	679	875

tgme49_asmbl.44	720	903	0	P	1	184
tgme49_asmbl.44	904	1130	0.4	H	220	447
tgme49_asmbl.45	1	52	0	P	133	184
tgme49_asmbl.45	52	421	0	cob	506	875
tgme49_asmbl.45	409	494	0	P	1	86
tgme49_asmbl.45	495	941	0.5	H	1	447
tgme49_asmbl.45	892	1006	0	coxI	1377	1491
tgme49_asmbl.45	1007	1091	0	J	1	85
tgme49_asmbl.45	1092	1177	0	O	1	86
tgme49_asmbl.45	1178	1538	0.3	K	1	361
tgme49_asmbl.46	1	372	0	cob	679	1050
tgme49_asmbl.46	373	538	0	N	1	166
tgme49_asmbl.46	539	629	0	F	89	179
tgme49_asmbl.46	630	714	0	R	1	85
tgme49_asmbl.46	715	1039	1.9	coxI	49	376
tgme49_asmbl.47	1	398	0.2	M	357	754
tgme49_asmbl.47	399	577	0	F	1	179
tgme49_asmbl.47	578	663	0	O	1	86
tgme49_asmbl.47	664	748	0	J	1	85
tgme49_asmbl.47	754	1359	0.2	cob	1	605
tgme49_asmbl.50	1	478	0	cob	398	875
tgme49_asmbl.50	478	661	0	P	1	184
tgme49_asmbl.50	662	865	0	I	1	204
tgme49_asmbl.50	866	977	0	K	334	445
tgme49_asmbl.51	3	259	0	M	498	754
tgme49_asmbl.51	260	350	0	F	89	179
tgme49_asmbl.51	351	516	0	N	1	166
tgme49_asmbl.51	517	713	0	cob	679	875
tgme49_asmbl.51	713	896	0	P	1	184
tgme49_asmbl.51	897	1018	0	I	84	204
tgme49_asmbl.52	1	74	0	F	1	74
tgme49_asmbl.52	75	160	0	O	1	86
tgme49_asmbl.52	161	389	0	coxIII	231	459
tgme49_asmbl.52	391	743	0	U	1	353
tgme49_asmbl.52	744	855	0	K	334	445
tgme49_asmbl.52	856	937	0	D	1	82
tgme49_asmbl.52	938	1065	0	coxI	44	171
tgme49_asmbl.52	12440	12730	0.3	K	158	445
tgme49_asmbl.52	12731	12934	0	I	1	204
tgme49_asmbl.52	12935	13118	0	P	1	184
tgme49_asmbl.52	13118	13645	0.4	cob	346	875
tgme49_asmbl.54	1	324	0.6	M	430	754

tgme49_asmbl.54	325	503	0	F	1	179
tgme49_asmbl.54	504	589	0	O	1	86
tgme49_asmbl.54	590	674	0	J	1	85
tgme49_asmbl.54	675	789	0	coxI	1377	1491
tgme49_asmbl.54	740	1032	0.3	H	1	294

Table 6.2. NUM/PTs that show segmental duplication

NUMTs				NUPTs			
Chr	**Start**	**End**	**# of paralogs**	**Chr**	**Start**	**End**	**# of paralogs**
chrII	5099	5205	4	chrII	13814	13914	3
chrII	12415	12541	10	chrIX	4970045	4970847	2
chrII	2236834	2236938	10	chrIX	5043622	5044424	2
chrII	2248185	2248306	5	chrIa	15741	15841	3
chrII	2254771	2254900	10	chrIa	1573069	1573330	3
chrII	2282076	2282144	4	chrVIIb	3733927	3733976	2
chrII	2284315	2284383	4	chrVIIb	3768369	3768418	2
chrII	2297290	2297357	4	chrX	5380837	5380969	2
chrIII	2298160	2298288	5	chrX	5384909	5384985	2
chrIII	2311257	2311387	6	chrX	5385366	5385498	2
chrIII	2317828	2317934	9	chrXII	39449	39549	3
chrIII	2320846	2320956	5	chrXII	5075497	5075568	2
chrIII	2370825	2370882	2	chrXII	5082862	5082947	2
chrIII	2370882	2370943	2				
chrIII	2393190	2393267	2				
chrIII	2465531	2465659	6				
chrIII	2474207	2474318	5				
chrIV	2255933	2256057	4				
chrIX	60905	61059	2				
chrIX	63920	64026	2				
chrIX	68225	68293	2				
chrIX	123864	123919	2				
chrIX	125504	125559	2				
chrIX	139089	139138	2				
chrIX	139140	139238	2				
chrIX	139933	139982	2				
chrIX	139984	140082	2				
chrIX	1205849	1205934	2				
chrIX	1205938	1206058	2				
chrIX	1206053	1206114	2				
chrIX	1215388	1215485	2				
chrIX	1215489	1215609	2				
chrIX	1215604	1215662	2				
chrIX	5043630	5043703	3				
chrIX	6271844	6271919	2				

chrIX	6271919	6271982	2
chrIX	6295817	6295894	2
chrIX	6312190	6312336	2
chrIX	6323557	6323643	4
chrIa	7040	7146	4
chrIa	14339	14468	10
chrIa	1573249	1573322	3
chrIb	1938127	1938216	3
chrIb	1942780	1942869	3
chrIb	1949100	1949189	3
chrV	4968	5026	3
chrV	5039	5130	2
chrV	13285	13376	2
chrV	92024	92125	4
chrV	113590	113715	3
chrV	509754	509812	4
chrV	509920	509997	4
chrV	1505158	1505214	2
chrV	1505226	1505333	2
chrV	1505382	1505487	3
chrV	3188040	3188109	2
chrV	3200304	3200423	2
chrV	3239267	3239336	2
chrV	3270617	3270756	2
chrVI	1701827	1701893	2
chrVI	3649599	3649705	4
chrVIIa	1527747	1527808	2
chrVIIa	1528989	1529050	2
chrVIIb	64268	64353	2
chrVIIb	67562	67630	2
chrVIIb	3750077	3750190	2
chrVIIb	3750271	3750393	2
chrVIIb	3750867	3751284	2
chrVIIb	3778763	3778876	2
chrVIIb	3778960	3779082	2
chrVIIb	3779556	3779973	2
chrX	5029	5140	6
chrX	8053	8120	10
chrX	56325	56431	8
chrX	63034	63093	5
chrX	140461	140567	4

chrX	140630	140698	8
chrX	143057	143119	2
chrX	4765609	4765669	2
chrX	4776491	4776552	2
chrX	4996027	4996088	2
chrX	5004771	5004832	2
chrX	7326929	7326985	3
chrX	7331267	7331323	3
chrX	7335390	7335442	3
chrXI	6008	6118	6
chrXI	8972	9078	10
chrXI	49359	49417	2
chrXI	51449	51507	2
chrXI	5597962	5598065	3
chrXI	5619174	5619266	2
chrXI	5619340	5619443	3
chrXII	30743	30849	4
chrXII	38049	38175	10
chrXII	565829	565894	2
chrXII	570379	570430	2
chrXII	1401437	1401499	2
chrXII	5688402	5688464	3
chrXII	5688465	5688515	3
chrXII	5693067	5693129	3
chrXII	5693130	5693180	3
chrXII	5695122	5695184	3
chrXII	5695185	5695235	3
chrXII	7055332	7055442	6
chrXII	7078152	7078309	3

Appendix 1. *T. gondii* mtDNA element and annotated cytochrome gene sequences

>A
ATGACATAAATACTAACAAACCACCGGTTTTGGATGGAATTACTTTTAACACCGCATAATATGC
TAGAAAATACCATTCAGGTACGATATGAAGTGGTGTTACGAACCGGTTGACTGGTATTGAGTTA
TCTGGATGTGATAGTTCCATCAAACCAAAAGCCGCTTGTAAGAAGATTAATCCAATTAAATAGG
ATAG
>B
CCACTATACTTAATGCACTTAACATGATGGTCATGAAGAG
>C
GTTCTGTGTTAGTAACTCAAAGTATTTGTGCTGAATTTGGTAGTGGTCTTGGTTGGACAATGTA
CCCTCCATTAAGTACTAGCTTGATGGTTTTAAATCCTGAGGCTACTGATTGGTTAATTGGAGGA
CTTGCTGTTCTTGGTATCAGTAGTATCCTTAGTTCTATTAACTTCCTTGGTACATGTGTCTTTA
TGGGATCTTGTGCAGGAGCAAAAAATTATATTTTATATATTTGGTCTATTATTTTTACAGCTCT
TATGCTAGTATTTACACTACCTATTTTAACAGGTGGACTAGTTATGATCTTATTAGATTTACAT
GTAAATACAGAATTTTATGATTCTATGTATTCTGGTGATAGTGTCTTATATCAACATCTATTCT
GGTTTTTGGACATCCAGAAGTATATATTCTAATTCTACCTGCTTTTGGTGTTGTATCTCAAACT
TTATCTATGTATTCATGTCGAGCGGTCTTCGGTGGTCAATCTATGATCTTAGCTATGGGTTGTA
TTTCTATTCTAGGTTCTCTAGTATGGGCACATCACATGATGACTGTCGGTCTAGAAGTAGATAC
ACGAGCTTACTTCTCAGCTATGACAATTATGATTGCTATTCCTACAGGTACTAAAATCTTTAAT
TGGTTAGGTACCTATATGGCTAGTCATAATACTACACGAACAATAGATTTATGGGCTGCCTTAT
GCTTTATTCTTCTATTTACTCTAGGTGGTACTACAGGTGTAGTTATGGGTAATGCAGGTATGGA
TATTGCACTACACGATACATATTATATTGTTGCACATTTCCATTTTGTATTATCTCTGGGAGCT
ATCTTAGCAACTATATGTGGATTTGTCTTCTATAGTAAAGATATGTTCGGAGATACTTTAAATC
TATTCCATGTGAATACAGGTTCATCACCTTATTTAAATATTTGGTTTGTTGTATTCTTGGCTAG
TATTATGTTAATCTTCTTACCTATGCATATATTAGGTTTCAACGTTATGCCAAGAAGGATCCCA
GATTACCCTGATTATCTTTGTTATA
>D
TTAAGGAGTCCTTTGTTTACAGCTTGTACCGTTACATTCAGGAGCATACCGTTATATTCGATGA
TCTTATGTGTTCACATTA
>E
ACATTTAGCATCTGTCATTAACATATGAGGATAAAAGGCAACTTTAAGCGCGGTATCAATACCT
GCAGGATTGCTAGAACCATTTAAATGTAAATAGAAGATATGTAAAACTATTATAATACAACCAA
TAAAAGGCAAGATAAAATGTAATACAAAGAATCGTTTTAATGTTACATCAGATACGTAGTAACC
TCCAAGTAGCCAAGGTACCAAATATGGTATTGGAGAAAGGAGGTTTGTAATGACTGTAGCACCC
CAGAAACTCATTTGTCCCCATGGTAGTACATATCCGAGGAAGGCAGTGGCTATAGTAAGTAAAT
ATAAAACTAAACCAGACATCCAAGCGGTAGTTAAATAACTATAACTCCAGTTATACAATCCTCG
AGTCATGTGAATTAAAATACACAAGAATACAAAAGAGGCTGTTGTAGCATGCAACATCCTAAAT
TCCCATCCTGCTGCTACTTCTCTAACTAGATGTTGAACGCTAGCAAATGCACAAGATGCTTCAG
AAGTATATCTAAACGCTAGAGTGATACCTGTAATTATTTGGAGTACAAAGGTCATTGCAACTAA
GAAACCA
>F
TGGTGACTTATATTATGATTCTTATTAATGGGAGCACAGTTCCCTGGGTATCCAATCCAGTGCT
CTGCCTTGGGCATTGAAACTAACCCACAGTTCAACCCTGTATTATTTAACTCAGTTAATTAGTA
TTAGGTTTGATATTCTGTACCAACCGATCTAAAATCGTGATGTCTTTTTGT
>G
AAATTATAAGATGAATTTAGATTTAGAGCACACCGATAAAAGACAAGGTGTGCCCGGAATAGAC
TCATAGATAGACTGAGTGTTCTCGAAACCATGCTAA

>H
TATAACAAAGATAATCAGGGTAATCTGGGATCCTTCTTGGCATAACGTTGTGTAGTTAGATGTT
GCGTACATTCCTTAATATCTGGAACGATAGATTATGATTAGGTAGTGGAACAAGGAGAGCGTCT
GTTGTACATCAACACTAGATACCGCTAATACCACTGTAGAGGTATAGTATATAATCCCTGCCCG
GTGCAGTAACATGTAAACGGCGGCTGTATTATGACGGTCCAAAGGTAGGAAAATCCTTGTCGGG
TAATTATCGTCGTGCGTGAAAGTTGGTCCGACTCTTCACATGTCGTTTATCTAAAACTTTTTGA
AATAGAATTATCTTTGAATATGAGGATCGAGATGGCCCGACGGTTAGACCCTGAGCACCTTTAC
ATCCCTTAAATCATAAACAGGATCAAATCTTCCTTGAGCGACTCGACAGGCACTAAAGATAGC
>I
CAGAACTGTAGTTTATCGGGACTAAAGTCAGCATAATCAATTAAAAGGTTTGTTCAGCCACTGG
TTCACCATCAACTACCTTGTTTCGACTTCGTACCGACTGTGTTATTGTAGCACATAGATTACCC
TTCTAGAAGGGAATTGTTCCCAAACAACCGGATCGTGTTGGCTAGGTGAACTAATCACGTTTCA
TAAATACAATCA
>J
CACAATAGAACTTGGATCCGGTAAACAAAGACCTTCAAGATCTAAACCAGTAGTCCAACTCGTA
GTATATACTCCCCAGAAAAAG
>K
CTGAGTACGTAAGGAAAAGGAAAGGTTAACCGCTATTTAAACACAACAGTTACCGTAGCTGTAG
ATGAATGCTAATTATAGAGTATATCTCCTATGACACTGCATAACATATAAATGCTCCTTCCGCC
ATTCGTTGACTGTGTTTACCACGGGGAATTAGAACAGAATATCAAGTTCTTTGCCTGGAGGTTT
GTTACGTTCCGTACAGTTGTAGGTAAAAGGTATGTTAGAGACTTAGACTAGCGTTGGAGCACAT
TGTTTCATTCGATAGTCCACGCTCAATCTTACCATACATAATACTTTTATGATCCCAGGCTGGT
TTAATAAGTCAAAGTTTAGCCGGGAAGTTAGCGTCTAACATATATAACCGAGTATCAACTTAGA
TGCACAGATGGACATAATTAATCCTTGTACGGTTTGTACCTACTTGACTCCTCAGTTTAAG
>L
CTTATAAATAATCCTGTCTCAGAGATGATTACTCCCAGTACGATGGTACTGATCATACTAGCAT
CTGAGTAGTAGTTTTCTCTCGCTGTTAATACGAGTGATAACAGTAATCCATATATTACGCCTAG
AACATAACCGATGTGGAATAACCTTAATGTC
>M
TATCCAGCGTATATTTGAAAAACCAACATTTATATACAAGCTGTACAAATATCATGATATTCAC
TTTGGTAGTCTCCTTCCTAATGTTAGTCTGTACGGAATACTTAGGTCTATCTATTTATATTAAC
GATAATGGATTTGGTAATGGTCTATTTATACTTACTGGTATACATTTCAGTCATGTTATTGTCG
GTGCTATCTTGGGTTTCTTTAATCAGGGTATGTATAGCTCTCTAGTTACATATTTACCAGTAAA
CTGCATAACTTTGAGTAAATGCAAAGGTACATTATGTAAAATCTTCTCAGAACCATTTACAATC
TTATATCTACATTTCGTCGAAGCAGTGTGGATAATGATCCACGTTACATTCTATCTCTAAATTA
TATAACGGTCGTAAGGTACGCCGGGGATAACAGGTCAGATAATATTGGGAGTTCTAATCCTCGG
ATTGTATCAGCACCTCCATGTCGGCTCATTACTCCCTTGTTATTGAACAAGATTCAGTTAGGAA
CGATAGTTCACCGTCAGATGTAATACGTGAGCTGGGTTAAGAACGTCTGGAGACAGTTTGTTCC
CTATCTACCATTTAAAATTGTATAAATATATACAGCAGAACTACAATGTTCACGCTCATGGATT
CAGTGTCCAGGACTACCTGGCGCTTAATAACGATTCCGTCTTCCAGCTTCCAATTAAACTACAT
ATTAAAATGAGCGCATGTAAACTAGTCTTAAACACACCGCTCGTCACGTA
>N
CTGTGATGTTAAGGAGCATAGGAAATGCTACACACCTTAATTGGTAATGCATTGATATAATTAT
CGTACAAGCACTCAGCTAGTTATTGAGAGACACTCACGAGCCCAAAACCATAACTTCGTAATCT
CAATGGTTTGCGATAGAACCATAACACAATGATATTTA
>O
GTAGTTTATATAAACCTAGGAATCCCATTTTATAATGTGTAACAGGGAGTCTAGCTTCAGTTGT
TATCTGATTGGTATTGCATGCC

>P
GTGAAAGCTCTTTTGATTTCCATGAACGGAGTTACATATTAGATTATCTTCGCTCCCATGGTCT
GTAGTAAGTTAACATTAAAACGTATCCAGTGTAAAGTTTAAACGTAATGCAGCTTAATCTTCTA
TGTTGTTATGTTAATCAAATAAGTTTCATCGTTATTGATATTTCATTGACATGTTG
>Q
TTAATACATGGTGTTCAATTGGTTCTATTTCTACAATAATTATCATCTTAACTATGCTCTGCTA
A
>R
CGTTTATATAAATCTAGTAATGCTTGTCAAGTTCCTTGTATCTATTTAGCTCACTGCGTACTTA
GGATCAGACCAAATGAGTTCA
>S
AGTTCTGGTTCGCGGATCATTTGTACAGAGACAATATCTACTTATAATGTGATAATTACAATAC
ATGGTCTAGCTATGATCTTTATGTTCTTAATGCCGGCTTTGTACGGAGGATATGGTAACTTCTT
TGTACCAATCTATATTGGTGGTTCGGAAGTCGTTTTCCCAAGAACTAACGCGATCTCCTATTTC
CTAGTACCATTAG
>T
TCTGATTAATTTAGGTTTACTTTCTGAGATTCGAGCTTTAAATACCCGAATGTTAATTCGTCAA
CAGTTCATGACTCGAAATGTAGTCAGTGGATGGGTGATTATTTGGGTATATAGTATGATCTTCT
TGATTATTATAGGTAGTGCTATTCCACAAGCAACATACATCTTATATGGTAGATTAGCTACTAT
CTTATACCTTACTACCGGATTGGTTCTATGCTTATACTAAA
>U
CGGATCGGATTCTTGTTGGCCTGGCACCTGTTTAGTAACTGGATGAACGCTTTTTACGCCTGGT
ATGTATGGATAATACTCGACTCTTCTATAGTTTAACCGCTACTGCTGGGACTGTATATTATGTA
CTTACGGTAGTACTATCAAGCCTCTTCTTCCAAATAGATTTCATGGAAAACCTAAAATTCGCAT
GTTTGATTGACATTTAGCCGCTAATATACAATCATCCAAGATATATTTATCTATCGCAGGTTCG
GTCTAATGTCCCGTTATACTATATAGATCACATGGCTTCTGGTACTTTGAGATCATACTAACGG
CGAGAAGGGAAGTGTGTTTCAAAGAAAAGGGAT
>V
GTAGGATATTGAAATCCAACACTTTTAGCTGTCTTAAGCAGTCCAGTGGGGTGGTGGTGTACAG
CAATCATAAAGAACTTGGTTGTCTGTATCTCATAACTGGAGTCATATTCAGTATCCTAGGTACT
ATAATGTCTCTGTTTATTCGATTTGAGTTATAC
>W
ACAACATTGTTAATGACTACAGCTTCCAAGCAAACATGAATACCG
>COXI
ATGAATACCGGTAGGATATTGAAATCCAACACTTTTAGCTGTCTTAAGCAGTCCAGTGGGGTGG
TGGTGTACAGCAATCATAAAGAACTTGGTTGTCTGTATCTCATAACTGGAGTCATATTCAGTAT
CCTAGGTACTATAATGTCTCTGTTTATTCGATTTGAGTTATACAGTTCTGGTTCGCGGATCATT
TGTACAGAGACAATATCTACTTATAATGTGATAATTACAATACATGGTCTAGCTATGATCTTTA
TGTTCTTAATGCCGGCTTTGTACGGAGGATATGGTAACTTCTTTGTACCAATCTATATTGGTGG
TTCGGAAGTCGTTTTCCCAAGAACTAACGCGATCTCCTATTTCCTAGTACCATTAGGTTCTGTG
TTAGTAACTCAAAGTATTTGTGCTGAATTTGGTAGTGGTCTTGGTTGGACAATGTACCCTCCAT
TAAGTACTAGCTTGATGGTTTTAAATCCTGAGGCTACTGATTGGTTAATTGGAGGACTTGCTGT
TCTTGGTATCAGTAGTATCCTTAGTTCTATTAACTTCCTTGGTACATGTGTCTTTATGGGATCT
TGTGCAGGAGCAAAAAATTATATTTTATATATTTGGTCTATTATTTTTACAGCTCTTATGCTAG
TATTTACACTACCTATTTTAACAGGTGGACTAGTTATGATCTTATTAGATTTACATGTAAATAC
AGAATTTTATGATTCTATGTATTCTGGTGATAGTGTCTTATATCAACATCTATTCTGGTTTTTT
GGACATCCAGAAGTATATATTCTAATTCTACCTGCTTTTGGTGTTGTATCTCAAACTTTATCTA
TGTATTCATGTCGAGCGGTCTTCGGTGGTCAATCTATGATCTTAGCTATGGGTTGTATTTCTAT

TCTAGGTTCTCTAGTATGGGCACATCACATGATGACTGTCGGTCTAGAAGTAGATACACGAGCT
TACTTCTCAGCTATGACAATTATGATTGCTATTCCTACAGGTACTAAAATCTTTAATTGGTTAG
GTACCTATATGGCTAGTCATAATACTACACGAACAATAGATTTATGGGCTGCCTTATGCTTTAT
TCTTCTATTTACTCTAGGTGGTACTACAGGTGTAGTTATGGGTAATGCAGGTATGGATATTGCA
CTACACGATACATATTATATTGTTGCACATTTCCATTTTGTATTATCTCTGGGAGCTATCTTAG
CAACTATATGTGGATTTGTCTTCTATAGTAAAGATATGTTCGGAGATACTTTAAATCTATTCCA
TGTGAATACAGGTTCATCACCTTATTTAAATATTTGGTTTGTTGTATTCTTGGCTAGTATTATG
TTAATCTTCTTACCTATGCATATATTAGGTTTCAACGTTATGCCAAGAAGGATCCCAGATTACC
CTGATTATCTTTGTTATATTAATACATGGTGTTCAATTGGTTCTATTTCTACAATAATTATCAT
CTTAACTATGCTCTGCTAA
>COXIII
ATGATTGCTGTACACCACCACCCCACTGGACTGCTTAAGACAGCTAAAAGTGTTGGATTTCAAT
ATCCTACGACATTAAGGTTATTCCACATCGGTTATGTTCTAGGCGTAATATATGGATTACTGTT
ATCACTCGTATTAACAGCGAGAGAAAACTACTACTCAGATGCTAGTATGATCAGTACCATCGTA
CTGGGAGTAATCATCTCTGAGACAGGATTATTTATAAGCTTTTTCTGGGGAGTATATACTACGA
GTTGGACTACTGGTTTAGATCTTGAAGGTCTTTGTTTACCGGATCCAAGTTCTATTGTGCTCTT
CATGACCATCATGTTAAGTGCATTAAGTATAGTGGTATCCAGCGTATATTTGAAAAACCAACAT
TTATATACAAGCTGTACAAATATCATGATATTCACTTTGGTAGTCTCCTTCCTAATGTTAGTCT
GTACGGAATACTTAGGTCTATCTATTTATATTAACGATAATGGATTTGGTAATGGTCTATTTAT
ACTTACTGGTATACATTTCAGTCATGTTATTGTCGGTGCTATCTTGGGTTTCTTTAATCAGGGT
ATGTATAGCTCTCTAGTTACATATTTACCAGTAAACTGCATAACTTTGAGTAAATGCAAAGGTA
CATTATGTAAAATCTTCTCAGAACCATTTACAATCTTATATCTACATTTCGTCGAAGCAGTGTG
GATAATGATCCACGTTACATTCTATCTCTAA
>COB
ATGGTTTCGAGAACACTCAGTCTATCTATGAGTCTATTCCGGGCACACCTTGTCTTTTATCGGT
GTGCTCTAAATCTAAATTCATCTTATAATTTTGGTTTCTTAGTTGCAATGACCTTTGTACTCCA
AATAATTACAGGTATCACTCTAGCGTTTAGATATACTTCTGAAGCATCTTGTGCATTTGCTAGC
GTTCAACATCTAGTTAGAGAAGTAGCAGCAGGATGGGAATTTAGGATGTTGCATGCTACAACAG
CCTCTTTTGTATTCTTGTGTATTTTAATTCACATGACTCGAGGATTGTATAACTGGAGTTATAG
TTATTTAACTACCGCTTGGATGTCTGGTTTAGTTTTATATTTACTTACTATAGCCACTGCCTTC
CTCGGATATGTACTACCATGGGGACAAATGAGTTTCTGGGGTGCTACAGTCATTACAAACCTCC
TTTCTCCAATACCATATTTGGTACCTTGGCTACTTGGAGGTTACTACGTATCTGATGTAACATT
AAAACGATTCTTTGTATTACATTTTATCTTGCCTTTTATTGGTTGTATTATAATAGTTTTACAT
ATCTTCTATTTACATTTAAATGGTTCTAGCAATCCTGCAGGTATTGATACCGCGCTTAAAGTTG
CCTTTTATCCTCATATGTTAATGACAGATGCTAAATGTCTATCCTATTTAATTGGATTAATCTT
CTTACAAGCGGCTTTTGGTTTGATGGAACTATCACATCCAGATAACTCAATACCAGTCAACCGG
TTCGTAACACCACTTCATATCGTACCTGAATGGTATTTTCTAGCATATTATGCGGTGTTAAAAG
TAATTCCATCCAAAACCGGTGGTTTGTTAGTATTTATGTCATCTCTGATTAATTTAGGTTTACT
TTCTGAGATTCGAGCTTTAAATACCCGAATGTTAATTCGTCAACAGTTCATGACTCGAAATGTA
GTCAGTGGATGGGTGATTATTTGGGTATATAGTATGATCTTCTTGATTATTATAGGTAGTGCTA
TTCCACAAGCAACATACATCTTATATGGTAGATTAGCTACTATCTTATACCTTACTACCGGATT
GGTTCTATGCTTATACTAA

Appendix 2. Sequences of wild type (WT) and Deletion Promoter and UTR constructs (up to ATG)

> WT-1 (WT promoter of the U6 snRNA-associated Sm family protein gene)

CCTACGGTTCTGACTACCTTGGACGTTTTTCTTAGATGAATCTTATCAGTCACAGTAGCTTCTG
TTTCCGTAGCGGCAAGTACATCCGACACTGTGGAGGTATTTACAAAGTTTACAGTTTGCGGCAA
TCGCAACATGCGCCTTACAACCCATTTCCTGGTCTCATCTTTCTCTGTGGAGTTGTCCTTAGGT
GCGCAGCATTTTCGGAGCTTGTTTTGACTCTGCGTGTCCCATGGAGCAACAAGTAACACGTCGA
TCTTCCATCATAAATGGTGAACTGGAGCGTCATCATTGAATCAGCAGCAGTTTTAGTGAAATCA
TTGTGTTATATGCATCATATGGTGTAAAGGAGGAGCATCAGTGCGGCAGTTGGCTGAATGGCGT
TTTTCATCGTCCCCATGCTACAGTTCGGGTGTTGGCATCCACTCCAGACCCACCTATGCCGCAC
CACGGAATTGATCGATCGAAGTTGTGGAGCAAGATACACATCTCAACAAAGACCTTCACTGTCG
GTATATAAATGGATCCATTGGTAATTTTGCGCTTTGAAAAGCGTTTCTATGTGACAGCGTTACA
CACGTACCCAACCCCTTCTTGACGTGAGCCAGATCTCCCAGCTTCTGGAGTCTCAGGGCTGCCA
CCGCGCTCCAAGTCAGGGGCTTAGTTCTTGACCATACCTTTCTATCATCCAAATCTGATCGACT
GGTATGGCCAGAAGGGCCCTTCTCTAAGCGTTGAAGGGCACAACAGAAACGAGTGCTGAGGGCA
TTTGACTAGCATCCACCCACAGCAATATTCTTCACAGTCACTGACTGACTCTCATCCATCGCAT
GTGCATTATTTTCTAGAATGCGTGTCACACATACCACCCATTTGCGTATCGTATCAACGCAGCT
CCGTGTCGCGTACGTTCCTCACTGGCCAGGACAGTACATTGTGAATAAGTGGCAGAGCGAAGGA
AGCGGCAGTTGCATTCCAACATTTGATACTGCTAATGGCAGCTAAATGCCACTAAAGCGAGGGA
ATTCTCGGTGTTCCCGTGTTCCCTGAAAGGTATTTGGAAGACTCTGCGGTAATAGAGGAGGTAC
ATAATGTACACTCGCAGTGGGAGGTCATAAGTGTTTAAAAACGAAAAAAAGTGGTGAAAATTAA
AGTTAGCGTCTAATATATATCCAGTATCCACAGGGATTAGTTGTTAAAATTCCAATTTGGTGGA
AGTTTGAAGTTTAGTTCACAATCGACTGCTAGTATGTCTCTGCCAGCTGTAGATCTGGCTGCGA
GGAAGACTTGCATGCACATGCGTAATGTTTCATATTCAAATCGGTTGTGGGTTAGCATCGGCCT
TCAACCTGACTCCCCTGCCTCGTTTTCCCTATAAGAATGACAGCTTTGTGGTTTTCATTGTTTA
ATACGTTGCTGACAACTGTGTGGTACCGCTCAATGTGGAACAGACGAGCTCTGCTTTGCTTTTC
TTCACAAGGTCCACTGATTGCTCCTACCGCGCTGCGCAGCGGACAACCGTAGATCTCCACAATA
AGACAGTGTCATTTTTTCCCGAAAACAGACTCAGCTGCTGCTCAGCATTCGTTTGTTTCTAGGC
GGCACACAGGGGCGAGTGACACCTGTGATGCCAAGTATTCTGTCAGAACCGTAGACTTGCCTAC
GGTAAGTCGAAAGACCGCCGCTCGACACAACGTAACCAGCAACTCACTGCTTTTCCGGAAGCTC
TATTTTCGCCGGTACATGCAAAGCCTTTGATATACATTTTTCACCAGTTTTGGGGGGGCGCTGC
TGATGTAGACTCGCTTCTTTCTGGCATCCGGGGCAGAGTTGCAATTCTGCGGCGGCGGTGGCGG
TTTGGGCGCCTTCCTTGTTTCACGCTTCGTCGACTGGCTCGTGGACGTGTCCTGATTACAGGCA
GTAGCTCAAAAGCGCAATCACAAAACAGAAGGTTTTTTAATCTTCTCGTCTTATTGGCTTCCGT
CGCAACCTCCTGGAGTTAGACATAATG

> ΔNUMT-1 (NUMT was deleted from the WT promoter of the U6 snRNA-associated Sm family protein gene)

CCTACGGTTCTGACTACCTTGGACGTTTTTCTTAGATGAATCTTATCAGTCACAGTAGCTTCTG
TTTCCGTAGCGGCAAGTACATCCGACACTGTGGAGGTATTTACAAAGTTTACAGTTTGCGGCAA
TCGCAACATGCGCCTTACAACCCATTTCCTGGTCTCATCTTTCTCTGTGGAGTTGTCCTTAGGT
GCGCAGCATTTTCGGAGCTTGTTTTGACTCTGCGTGTCCCATGGAGCAACAAGTAACACGTCGA
TCTTCCATCATAAATGGTGAACTGGAGCGTCATCATTGAATCAGCAGCAGTTTTAGTGAAATCA
TTGTGTTATATGCATCATATGGTGTAAAGGAGGAGCATCAGTGCGGCAGTTGGCTGAATGGCGT
TTTTCATCGTCCCCATGCTACAGTTCGGGTGTTGGCATCCACTCCAGACCCACCTATGCCGCAC

CACGGAATTGATCGATCGAAGTTGTGGAGCAAGATACACATCTCAACAAAGACCTTCACTGTCG
GTATATAAATGGATCCATTGGTAATTTTGCGCTTTGAAAAGCGTTTCTATGTGACAGCGTTACA
CACGTACCCAACCCCTTCTTGACGTGAGCCAGATCTCCCAGCTTCTGGAGTCTCAGGGCTGCCA
CCGCGCTCCAAGTCAGGGGCTTAGTTCTTGACCATACCTTTCTATCATCCAAATCTGATCGACT
GGTATGGCCAGAAGGGCCCTTCTCTAAGCGTTGAAGGGCACAACAGAAACGAGTGCTGAGGGCA
TTTGACTAGCATCCACCCACAGCAATATTCTTCACAGTCACTGACTGACTCTCATCCATCGCAT
GTGCATTATTTTCTAGAATGCGTGTCACACATACCACCCATTTGCGTATCGTATCAACGCAGCT
CCGTGTCGCGTACGTTCCTCACTGGCCAGGACAGTACATTGTGAATAAGTGGCAGAGCGAAGGA
AGCGGCAGTTGCATTCCAACATTTGATACTGCTAATGCTCTGCGGTAATAGAGGAGGTACATAA
TGTACACTCGCAGTGGGAGGTCATAAGTGTTTAAAAACGAAAAAAAGTGGTGAAAATTAAAGTT
AGCGTCTAATATATATCCAGTATCCACAGGGATTAGTTGTTAAAATTCCAATTTGGTGGAAGTT
TGAAGTTTAGTTCACAATCGACTGCTAGTATGTCTCTGCCAGCTGTAGATCTGGCTGCGAGGAA
GACTTGCATGCACATGCGTAATGTTTCATATTCAAATCGGTTGTGGGTTAGCATCGGCCTTCAA
CCTGACTCCCCTGCCTCGTTTTCCCTATAAGAATGACAGCTTTGTGGTTTTCATTGTTTAATAC
GTTGCTGACAACTGTGTGGTACCGCTCAATGTGGAACAGACGAGCTCTGCTTTGCTTTTCTTCA
CAAGGTCCACTGATTGCTCCTACCGCGCTGCGCAGCGGACAACCGTAGATCTCCACAATAAGAC
AGTGTCATTTTTTCCCGAAAACAGACTCAGCTGCTGCTCAGCATTCGTTTGTTTCTAGGCGGCA
CACAGGGGCGAGTGACACCTGTGATGCCAAGTATTCTGTCAGAACCGTAGACTTGCCTACGGTA
AGTCGAAAGACCGCCGCTCGACACAACGTAACCAGCAACTCACTGCTTTTCCGGAAGCTCTATT
TTCGCCGGTACATGCAAAGCCTTTGATATACATTTTTCACCAGTTTTGGGGGGGCGCTGCTGAT
GTAGACTCGCTTCTTTCTGGCATCCGGGGCAGAGTTGCAATTCTGCGGCGGCGGTGGCGGTTTG
GGCGCCTTCCTTGTTTCACGCTTCGTCGACTGGCTCGTGGACGTGTCCTGATTACAGGCAGTAG
CTCAAAAGCGCAATCACAAAACAGAAGGTTTTTTAATCTTCTCGTCTTATTGGCTTCCGTCGCA
ACCTCCTGGAGTTAGACATAATG

> WT-2 (WT promoter of the myosin heavy chain gene)
CACCAGTGCGGGAGGCGTTCTAGCGGCCTCGGTGTATGTCTTGTCGACACACGGGAGGGAAATG
GAGGTCGAGAAGTTCCAGAAACGAGGAACTTTTCTTGAAATCGAGTTCGACTCAAAGCAGGCTA
CCTTTCCATTGTCGGTTTCGCATTTTACCAGAGAGTTGCCGTGTCGCGGAACAGAGTCACCTCG
AGTGAGCGTCGCGACAATCGGCACGCGACACGGGCGCGAGAGACGCGTGCATGACCGTGGATTT
CTTTAGAAAAGCGAACCTTTGAATGGTGAAGCTGCAAGCGGCCACGGAGGAAAGTTGCGGGGTT
CCTTCTTGACTTGTTGGTGTATATTGGAGAGGACATTCGAGGAAAAACAACGAAATGCCAAGAC
AAGACAGAGTCGGAGAGACAGCCACAGATGCAAGAGCCCGAAACTTTGTGTGGCTTTTTGCATG
CATTTGAAGGATTAAAGCCGCGGGGTGGAACTGGACGATTTTGACGATGGGAGTCGGTTTCCTT
CCCTCAGGATCTTACCAGTTTTGCCACCTCCACCTTAAACTTCTCCTCTGTCTCCCCTCCTTGC
CGCTTGTCTGGCTAGTCGACGAGAAACAGAGTTCTTCCGTGTACGTACGTACACTCAATTGGAA
CTTGTCGCAGGGTGAGGCCTCGCGCTTTTTTGGAGGCCAGGCAGTCCTGCAGAATCAGTCGTGT
AATGAGAAAAAGCCGATCCACTTTTCTTGCGAAGTGCTACTCGAGTTAAAGTTCTACAGCGCAC
TCGTTTTCCTTTGTTGTCTGCCCAAGAAAAAACTGGAAACACGGTGACAGAGCTGATGGGCAAT
ACGTCTTTTCACCCGGGTTGAGAGAGTCGGGAAACAGCAGACGCATCAACTACATTCATTGGCT
TTGAGTGAAGAAGGATCCGGCAGTTTTGCGTAGTGAGAGAACGTGCAACGATCTAAATTTGAGG
AAGCCGTCGAATTTGAGGGAGCCCAGTGAAAAAAACAGACACCAGCATCACTCGTCAGCCTCGT
CTGTGCCGAGGTGAAAACAGAAAGTTCGCGGAGACAACGGAATTTCGAGGTCTACAAGGTAGGA
GGGACAGTTCTTCGAAAGCTAGCTGGGTTTTCGGTCTTAACGGAACAGTGTCGAGCTCCCAGTT
GCAAGCGGGAGAAAACCCTCCGGCACAGCAGGTGTGGGAAATCAAGAAAGCTGCTCCCACCGAG
GAAACTAAAGTACGGGAGTCGTTGGCGTGCTGGCCATGTGCGGTGTCTGCACCGATTCCATTCC
TTTGGTCGAAGATGTTCCGCGAGGAAAAGTGGCAGCCAGACAGTGTGGGTGCTGAACGGACGCG

TTGAAGAGCGCGATCAGCGATCCTCTGTTCAGTGGATTTCAGTGACAGGGGCAAGAAGGAAGGA
ACCAGAACTGTCGGAGATCGGGGTCCAGTTGAACTGCTCCTCTTTCTCTGTGTTCAACAACCCT
GTATCATTTAACTCAGTTCTCTAGCAGTAGGTCTGGTATTCTGTACTAACCTATCTGAATTATA
TTTGGGTGTATCGGAGGATTCCTAGTTAGTGGACCAGCTGGTATAAAATATCACACCTAACACT
CATTAATGAGTTAAGTAGTCGACATATTTCCCTGTCGCACTTCTTGGGTTTGACTGTTCGGTGG
GTTGTTCACGGTGGTATTTGCGTTTTCGCGGTGCAATATTTCCGAAGGAATGTGGAGTCCAGGA
AAAGCCACAATCGGTTCATCGAACTGCACAGCCGTACCCTAGACTGAGACACTCCAATAGAAAC
CGCTTTTTCGCTTTCGTCGAAAAATCAAATCTGCTGCGTAGCCACACCGGTGTAACTCTGGGTA
GAACACATACATTTTCTCGTCTGTGTTTCAGCAACCTAATTGATACGTAGATTTGAGTGATTCC
TCGCACACCGCCCTTGTCACCGAAAAAGTTGGTGCCATTTCCAATTTTCTAAATCGTACACCAA
TGTCTCTGCTCGAAAACAATG

> ΔNUMT-2 (NUMT was deleted from the WT promoter of the myosin heavy chain gene)
CACCAGTGCGGGAGGCGTTCTAGCGGCCTCGGTGTATGTCTTGTCGACACACGGGAGGGAAATG
GAGGTCGAGAAGTTCCAGAAACGAGGAACTTTTCTTGAAATCGAGTTCGACTCAAAGCAGGCTA
CCTTTCCATTGTCGGTTTCGCATTTTACCAGAGAGTTGCCGTGTCGCGGAACAGAGTCACCTCG
AGTGAGCGTCGCGACAATCGGCACGCGACACGGGCGCGAGAGACGCGTGCATGACCGTGGATTT
CTTTAGAAAAGCGAACCTTTGAATGGTGAAGCTGCAAGCGGCCACGGAGGAAAGTTGCGGGGTT
CCTTCTTGACTTGTTGGTGTATATTGGAGAGGACATTCGAGGAAAAACAACGAAATGCCAAGAC
AAGACAGAGTCGGAGAGACAGCCACAGATGCAAGAGCCCGAAACTTTGTGTGGCTTTTTGCATG
CATTTGAAGGATTAAAGCCGCGGGGTGGAACTGGACGATTTTGACGATGGGAGTCGGTTTCCTT
CCCTCAGGATCTTACCAGTTTTGCCACCTCCACCTTAAACTTCTCCTCTGTCTCCCCTCCTTGC
CGCTTGTCTGGCTAGTCGACGAGAAACAGAGTTCTTCCGTGTACGTACGTACACTCAATTGGAA
CTTGTCGCAGGGTGAGGCCTCGCGCTTTTTTGGAGGCCAGGCAGTCCTGCAGAATCAGTCGTGT
AATGAGAAAAAGCCGATCCACTTTTCTTGCGAAGTGCTACTCGAGTTAAAGTTCTACAGCGCAC
TCGTTTTCCTTTGTTGTCTGCCCAAGAAAAAACTGGAAACACGGTGACAGAGCTGATGGGCAAT
ACGTCTTTTCACCCGGGTTGAGAGAGTCGGGAAACAGCAGACGCATCAACTACATTCATTGGCT
TTGAGTGAAGAAGGATCCGGCAGTTTTGCGTAGTGAGAGAACGTGCAACGATCTAAATTTGAGG
AAGCCGTCGAATTTGAGGGAGCCCAGTGAAAAAAACAGACACCAGCATCACTCGTCAGCCTCGT
CTGTGCCGAGGTGAAAACAGAAAGTTCGCGGAGACAACGGAATTTCGAGGTCTACAAGGTAGGA
GGGACAGTTCTTCGAAAGCTAGCTGGGTTTTCGGTCTTAACGGAACAGTGTCGAGCTCCCAGTT
GCAAGCGGGAGAAAACCCTCCGGCACAGCAGGTGTGGGAAATCAAGAAAGCTGCTCCCACCGAG
GAAACTAAAGTACGGGAGTCGTTGGCGTGCTGGCCATGTGCGGTGTCTGCACCGATTCCATTCC
TTTGGTCGAAGATGTTCCGCGAGGAAAAGTGGCAGCCAGACAGTGTGGGTGCTGAACGGACGCG
TTGAAGAGCGCGATCAGCGATCCTCTGTTCAGTGGATTTCAGTGACAGGGGCAAGAAGGAAGGA
ACCAGAACTGTCGGAGATCGGGGTCCAGTTGAACTGCTCCTCTTTCTCTGTGTTCAAAATTATA
TTTGGGTGTATCGGAGGATTCCTAGTTAGTGGACCAGCTGGTATAAAATATCACACCTAACACT
CATTAATGAGTTAAGTAGTCGACATATTTCCCTGTCGCACTTCTTGGGTTTGACTGTTCGGTGG
GTTGTTCACGGTGGTATTTGCGTTTTCGCGGTGCAATATTTCCGAAGGAATGTGGAGTCCAGGA
AAAGCCACAATCGGTTCATCGAACTGCACAGCCGTACCCTAGACTGAGACACTCCAATAGAAAC
CGCTTTTTCGCTTTCGTCGAAAAATCAAATCTGCTGCGTAGCCACACCGGTGTAACTCTGGGTA
GAACACATACATTTTCTCGTCTGTGTTTCAGCAACCTAATTGATACGTAGATTTGAGTGATTCC
TCGCACACCGCCCTTGTCACCGAAAAAGTTGGTGCCATTTCCAATTTTCTAAATCGTACACCAA
TGTCTCTGCTCGAAAACAATG