ROLE OF REPETITIVE DNA IN APICOMPLEXAN GENOME EVOLUTION

by

ASSIATU B BARRIE

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

MASTER OF SCIENCE IN BIOLOGY

THE UNIVERSITY OF TEXAS AT ARLINGTON

May 2010

ACKNOWLEDGEMENTS

First and foremost I would like to thank my supervising advisor, Ellen J. Pritham, who gave me the opportunity to work in her lab since my undergraduate years, and was supportive of me when I decided to pursue my Masters. I have been presented with many opportune activities throughout my graduate career. Being able to travel and interact with scientist in the field of my research, has greatly enhanced my knowledge and made this an enlightening experience, and I thank Ellen, for providing me this opportunity.

I would like to thank my committee members, Cédric Feschotte who has been a constant source of knowledge and new ideas and Esther Betrán for her support and feedback throughout. Together, with my supervising advisor, I have been lucky to have continual support.

Financial support for this project comes from the National Institute of Health (NIH) R01 Research Grant, in collaboration with Cédric Feschotte, and Jessica Kissinger at the University of Georgia. Throughout my graduate studies, I was on a Research Assistantship (RA), so I could focus solely on my research project, and I am very grateful for that. In addition, I have had several monies from the Department of Biology, and Phi Sigma Biological Society to help with travel expenses.

I would also like to thank our collaborator, Jessica Kissinger at the University of Georgia for providing the mitochondrial genome for *Toxoplasma gondii* and *Neospora*

caninum. I am also thankful to Cheng Sun, my collaborator on this project. Cheng is one of the nicest people I have ever met, and he has been extremely helpful and made my work much easier since he embarked on this project with me.

Life as a graduate student is not easy, but because of the constant support from my fellow lab mates: Claudia Marquez, Jainy Thomas, Sarah Schaak, Komal Vadnagara, An La, Megha Bajaj, and Cheng Sun, I have been able to stay positive. These individuals have had to listen to multiple practice talks, and provided me with very helpful and positive criticism to help me do a much better job. And I am also thankful that I can also consider them my friends. I would also like to thank Clément Gilbert, for being a great friend and for keeping me motivated and sane, allowing me to stay focus on my goals and for being blatantly honest.

I would also like to thank the Mobile DNA Group and Genome Biology Group at UTA, for providing me the stage to present my work in a positive and encouraging atmosphere. I would also like to thank the UTA Biology Department Office Staff ladies for all their help.

During the beginning of my graduate career, I got separated from my sister, who is more like my mother, and it was difficult. But she has been continually supportive and encourages me to accomplish whatever I want to do in life. My nieces (Husiantou Diallo, Fatimah Bah) and my nephew (Alhassan Diallo) and my sister Salamata Bah keep inspiring me to be a positive role model. My family is very important to me, and my uncle, Hamidou Bah, has been extremely supportive of my choices to pursue my academic goals and I am very grateful for that. I thank my cousin and friend Assiatu

Barrie, for always being there, listening to my complaints and thinking that I can do whatever I set my mind to doing.

Even though, my father, the late Alhaji Ibrahim Bah, is no longer around to see the woman I am growing up to be, I know he will be proud of my accomplishment. I also know that my mother, Haja Salamata Bah is just as proud of me for following my dreams. My entire family in general, has always been there and they keep me pushing forward.

Again I thank my sister Isata Bah, and my ex-brother in-law, Alhaji Ibrahim Barrie, who changed my life significantly, and without their guidance and support during my childhood, I would probably have had a completely different life and might not have even had the opportunity to be here at all. So I would be eternally grateful to them for all that they have done for me.

April 2, 2010

ABSTRACT

ROLE OF REPETITIVE DNA IN APICOMPLEXAN

GENOME EVOLUTION

Assiatu B. Barrie, M.S.

The University of Texas at Arlington, 2010

Supervising Professor: Ellen J. Pritham

The Apicomplexa represent a phylum of obligate intracellular parasites that impart significant medical, veterinary, and socioeconomic burdens worldwide. The opportunistic, AIDS-associated, pathogen *Toxoplasma gondii* for instance, infects about one-third of the human population causing serious, life-threatening illness and birth defects in some. The genomes of fifteen apicomplexan species ranging in size from 8.3Mb to 64.0Mb have been sequenced and reveal a significant amount of plasticity in terms of size, AT-richness, introns and gene density. In many instances, genome size variation can be explained by differential expansion of repetitive DNA acquired through varying processes: intracellular organellar DNA transfer events, and proliferation of transposable elements (TEs). TEs make up the largest and most dynamic component of many multi-cellular and unicellular eukaryotic genomes. Moreover there is a positive

correlation between genome size variation and accumulation of TEs. In an effort to determine the source of genomic variation and genetic innovation in several apicomplexan parasites, we aim to explore their repetitive DNA content and the potential propagation of TEs within these organisms. We also seek to ascertain the extent of mitochondrial DNA transfer in *T. gondii* and in four other apicomplexans (*Babesia*, *Theileria*, *Neospora* and *Plasmodium)* in order to facilitate a better understanding of these insidious parasites. To this end we employed a comparative genomic approach using complementary bioinformatic tools: RepeatScout, RepeatMasker, and Blast to query and classify the repetitive DNA repertoire in these parasites. Interestingly, we find that TEs tend to be rare in the apicomplexans, with only two of the fifteen genomes harboring any identifiable mobile elements. We find that for most of the apicomplexans analyzed, the repetitive DNA is comprised of multi-gene families clustered within sub-telomeric and telomeric regions, and most of these repeats may be involved in generation of antigenic variation in these parasites. The intracellular parasitic lifestyles of these parasites may to some extent confer some protection to these organisms from the invasion of mobile elements. Concomitantly, we do find very high content of mtDNA-derived sequences within the *T. gondii* nuclear genome, referred to as numts. With numts occupying 1.88% of the *T. gondii* genome, *T. gondii* harbors the highest density of numts ever reported, nearly a 100 fold greater than that observed in the human genome. Comprehensive characterization of numts in *T. gondii* reveals that they originate from all regions of the mitochondrial genome and are distributed across all 14 chromosomes. Careful examination of numt flanking

regions show structural features suggesting that integration occurs at the DNA level during the repair of double-strand breaks by non-homologous end joining. Plotting the age distribution of the numts, we show the acquisition of DNA from mitochondria by *T. gondii* has been a continuous and probably still ongoing process, with integration events occurring ranging from 20 million years ago to less than 1 million year ago. In contrast to the *T. gondii*, the pattern of numt accumulation was strikingly different for the other apicomplexan genomes we analyzed, with a twofold difference between *Neospora* and *Toxoplasma* and very few to no numts detected in *Plasmodium*, *Theileria* and *Babesia*. These results, combined with lack of TEs within *T. gondii*, suggest that numts have had a considerable impact on the evolution of this parasite.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

Apicomplexa

The phylum Apicomplexa contains a diverse group of about 5,000 protozoan species, all of which are obligate intracellular parasites, i.e. they are completely reliant on their host to reproduce. Apicomplexan parasites are named for a group of unique structures located at the anterior point of their cell, known as the apical complex. This complex consists of the rhoptries, micronemes, conoid and apical polar rings: organelles that are crucial for parasite invasion and survival within the host cell (Figure 1.1) (BARTA 1989; WILSON and WILLIAMSON 1997). These parasites are capable of invading multiple cell types, and display complex life cycles, with some species capable of invading a wide host range (FRENAL and SOLDATI-FAVRE 2009). Many apicomplexans impart significant medical, veterinary and socioeconomic burdens worldwide. Most notorious are malarial parasites of the genus *Plasmodium* causing 1-5 million human deaths every year (CARLTON *et al.* 2001; GARDNER *et al.* 2002). Other apicomplexans of significant impact are the AIDS-associated pathogens, *Toxoplasma* and *Cryptosporidium*, and the parasite of veterinary importance, *Theileria*. *Toxoplasma* causes toxoplasmosis, a cosmopolitan disease that infects about one-third of the human population with severe effects in individuals with a weakened immune system (KIM and WEISS 2004). *Cryptosporidium*, a water-borne pathogen, causes cryptosporidiosis,

which is generally limited to diarrhea in the immunocompetent, but can be severe and sometimes fatal in the immunocompromised (ABRAHAMSEN *et al.* 2004). *Theileria*, a tick-borne pathogen, causes East Coast fever in cattle leading to 1 million cattle deaths annually in sub-Saharan Africa (GARDNER *et al.* 2005). The phylum also includes the gregarines, parasites that invade the guts of invertebrates like shrimp and cockroaches (MORRISSETTE and SIBLEY 2002). This group is thought to be the earliest lineage of the apicomplexans (Leander, Clopton, and Keeling 2003).

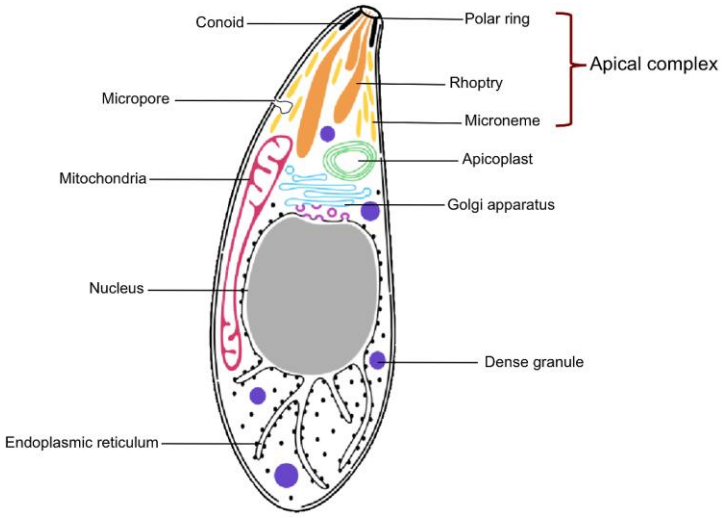


Figure 1.1 A typical apicomplexan cell.

Figure adapted from http://www2.bc.edu/~gubbelsj/Toxoplasma.html, modified from review by (AJIOKA *et al.* 2001).

Evolutionary History

The Alveolates

The Alveolates is comprised of three main groups: the dinoflagellates, the ciliates, and the apicomplexans (Figure 1.2). The dinoflagellates are free-living, unicellular eukaryotic protists, of which many are photosynthetic, and parasitic in

nature. The ciliates are free-living, non-photosynthetic unicellular protists, with relatively few parasitic species. The apicomplexans on the other hand, are non-photosynthetic, obligate parasites of animals including human. Members of the alveolates are divergent in form, but are thought to be united based on common ultra-structural and genetic similarities (GAJADHAR *et al.* 1991; LEANDER and KEELING 2003). These include the omnipresence of the alveoli (inner membranous sacs), microspores, similar extrusive organelles and a variety of molecular sequence characters (LEANDER *et al.* 2003; LEANDER and KEELING 2003). Even though the exact position of the alveolates within the eukaryotic tree of life is not well defined, the monophyley of this group is unequivocally supported (FAST *et al.* 2002; LEANDER and KEELING 2003).

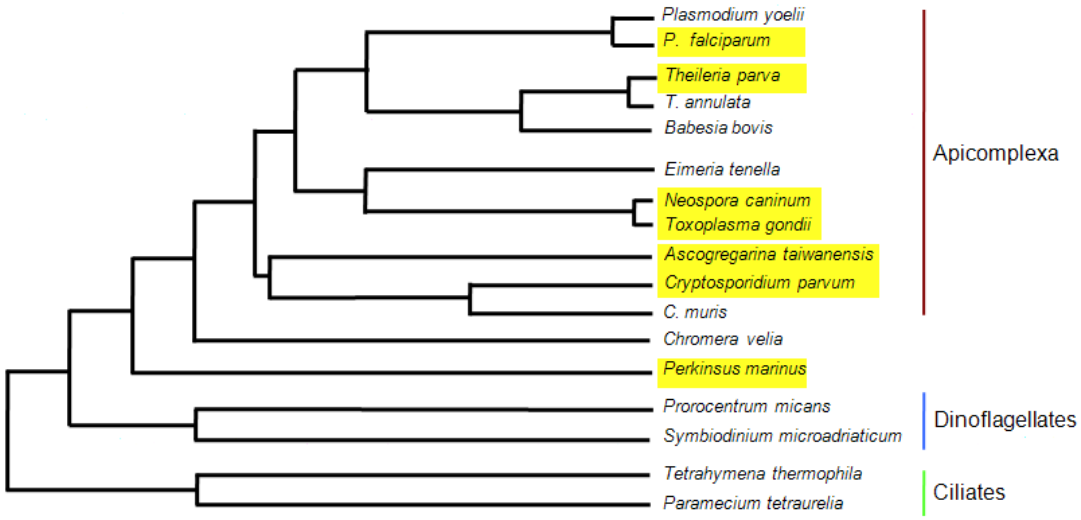


Figure 1.2 The Alveolates.

A consensus tree generated from several different phylogenetic[1] analyses depicting the relationship of the apicomplexa to the dinoflagellates, the ciliates, and the newly discovered organism, *Chromera velia*. Species of interest are highlighted in yellow.

[1] (LI *et al.* 2003; MOORE *et al.* 2008; TEMPLETON *et al.* 2010).

The apicomplexans and the dinoflagellates are thought to be more closely related to each other, than they are to the cilates (Figure 1.2). This is supported by multiple independent molecular phylogenies based on several ribosomal and nuclear genes (FAST *et al.* 2001; FAST *et al.* 2002). Concomitantly, molecular data suggests that the ancestor of the apicomplexans was most likely a dinoflagellate-like organism capable of parasitizing marine invertebrates (OBORNIK *et al.* 2009). The recent discovery of *Chromera velia*, an alga with a photosynthetic plastid with its similar origin as the apicomplexans, provides supporting evidence for this hypothesis (MOORE *et al.* 2008). Molecular phylogenetic studies of *C. velia* nuclear genes along with analysis of plastid genes reveals that this organism shares significant similarities with apicomplexans and dinoflagellates. *C. velia* is an excellent model to study the evolution of apicomplexans, as this organism represents the closest photosynthetic relatives of the this group (OKAMOTO and MCFADDEN 2008).

Plastid in the apicomplexa

Another interesting feature of the Apicomplexa phylum is the presence of a non-photosynthetic plastid known as the apicoplast: an organelle homologous to the chloroplast found in plants (BARTA 1989; WILSON and WILLIAMSON 1997). The apicoplast first noted in *P. knowlesi* was initially presumed to be of mitochondrial origin. Since the 1960's, the 35 kb circular organellar DNA of the apicoplast had been observed, measured, along with characterization of features typical of plastid genomes (inverted repeats of rRNA genes), but was mistaken for the mitochondria (ROOS *et al.* 1999; WALLER and JACKSON 2009; WILSON and WILLIAMSON 1997). However, the

lack of co-segregation with the mitochondria, and the subsequent isolation of a 6 kb molecule in *P. yoelii* finally revealed that the organelle was distinct from the mitochondria (WALLER and JACKSON 2009; WILSON and WILLIAMSON 1997).

Although the true beginnings of the apicoplast remain controversial, morphological studies, backed up with molecular phylogenetic analysis of apicoplast genes are indicative of a secondary endosymbiotic origin (ROOS *et al.* 1999). The secondary endosymbiosis most likely involved the uptake of a eukaryotic alga by a heterotroph, giving rise to the multi-membrane plastid we now recognize as the apicoplast (Figure 1.1) (CAVALIER-SMITH 1999).

The apicoplast genome is fairly reduced to ~ 35 kb, with remarkable conservation in genome organization observed within all apicomplexan species with current available apicoplast genome data. The exact function of this organelle is not quite clear, however this plastid is essential for the survival of these parasites and partakes in important metabolic pathways (FUNES *et al.* 2004). Studies in *Toxoplasma* show that a defect in the apicoplast can lead to delayed cell death of the parasite following the first replicative cycle (HE *et al.* 2001), indicative of an important role of this organelle. Pharmacological evidence also confirms the significance of this organelle. The apicoplast might serve as a therapeutic drug target with minimal damaging effects to the host as it metabolic pathways are different from those of the invaded host because of the prokaryotic origin (LIZUNDIA *et al.* 2009; MCFADDEN and ROOS 1999; ROOS *et al.* 1999).

The apicoplast is present in nearly all extant apicomplexan species, with the exception of *Cryptosporidium* spp, and the early branching lineage, the Gregarines. Recent sequencing and robust phylogenetic analysis of the gregarine, *Ascogregarina taiwanensis*, unites the gregarines and *Cryptosporidium* at the base of the apicomplexan clade (Figure 1.2) (TEMPLETON *et al.* 2010; ZHU *et al.* 2000). The apicoplast is not observed in both these hypothesized early diverging lineages, but present in the dinoflagellates, implying that the organelle was once present within the common ancestor of the apicomplexans but was subsequently lost in some lineages, although this idea is debated (TOSO and OMOTO 2007; ZHU *et al.* 2000).

Mitochondrial genome of the apicomplexa

The mitochondrion undoubtedly represents one of the most widespread and important organelles within eukaryotic cells. Mitochondria, proposed to have originated through primary endosymbiosis from α-proteobacterium specifically *Rickettsia*-like, carries out many essential metabolic processes and energy production for many eukaryotic organisms (DE SOUZA *et al.* 2009; GRAY *et al.* 1999; GRAY *et al.* 2001). The mitochondria contain its own self-synthesizing genome that varies significantly in organization and mode of gene expression (LANG *et al.* 1999). Many extant mitochondria contain dramatically fewer numbers of genes as compared to their free-living counterparts, with one to three orders of magnitude in difference. For instance, typically mitochondrial gene ranges from as few as five genes and up to 97 genes, as opposed to 834 protein-coding genes identified in the free-living α-proteobacterium, *R. prowazekii* (ADAMS and PALMER 2003; ANDERSSON *et al.* 2003;

WALLER and JACKSON 2009). To date, the apicomplexans are the current record holder for the smallest known mitochondrial genomes (WILSON and WILLIAMSON 1997).

The apicomplexan mitochondrial genome first discovered in *P. yoelii* and now in a few other apicomplexans is extremely reduced with a genome size of ~6 kb (WILSON and WILLIAMSON 1997). This is somewhat similar to what is observed in animals, which typically have small mitochondrial genomes of 15 kb to 20 kb, but drastically different from plants with mitochondrial genome sizes from 180 kb to 2400 kb (BOORE 1999; SCHUSTER and BRENNICKE 1994). The apicomplexan mitochondrial genome is extremely streamlined, encoding only three genes as opposed to about 37 in other eukaryotic organisms. These are the cytochrome oxidase c subunits (I & III), cytochrome b, and several ribosomal subunits (LSU, & SSU); however no tRNAs have been observed for the apicomplexans. The gene content within the apicomplexans is conserved but they differ significantly in their organization and structure (WILSON and WILLIAMSON 1997). For example, the mitochondrial genome of *Plasmodium* spp occurs as a circular and/or tandemly repeated linear unit, while the mitochondrial genome of *Theileria* and *Bebasia* is a 6.6 kb linear molecule flanked by terminal direct repeats (HIKOSAKA *et al.* 2009; MATHER and VAIDYA 2008). In *Toxoplasma,* the true characterization of the mitochondrial genome remains unclear, however recently the Kissinger Lab completed sequencing of the proposed mitochondrial genome. On the other hand, *Cryptosporidium* species harbor a mitochondrion-related organelle known as the mitosome that has completely lost its genome and the ability to produce ATP (SEEBER *et al.* 2008).

Together, analysis of the apicoplast and mitochondria of the apicomplexans provides insightful views into the biology and evolutionary origins of these parasites. Study of these endosymbionts also presents researchers with potential resources for developing much needed sustainable chemotherapeutics (FISHER *et al.* 2008).

Genome Biology

Nuclear genome organization

The malaria epidemic and a rise in mortality rates in both humans and animals due to apicomplexan parasites has greatly stimulated interest in obtaining genome sequence information for these organisms. There are several apicomplexan nuclear genomes that have been completely sequenced and made publicly available, while more ongoing projects are in the works (Table 1.1). The availability of genome sequences presents an opportunity to gain a better understanding of the genomic architecture of these parasites through comparative genomic studies (TAYLOR *et al.* 2007). This in turn advances our knowledge on the factors influencing virulence and pathogenicity in these organisms.

Table 1.1 Comparative genome statistics of current and ongoing apicomplexan genome sequencing projects.

Database	Parasite	Strain	Genome size (Mb)	Genome coverage	Number of chromosomes	G + C content (%)	Gene count	Gene density (bp)	Percent coding	% of genes w/ introns	Average gene length (bp)	Average length of intron (bp)
CryptoDB[1]	*C. hominis*	*TU502*	8.74	12x	8	31.7	3956	2293	69	5-20	1576	ND
	C. muris	*RN66*	9.21	ND	ND	ND	3980	ND	ND	ND	ND	ND
	C. parvum	*IOWA*	9.10	13x	8	30	3807	2305	74	5	1795	ND
PlasmoDB[2]	*P.berghei*	*ANKA*	18.00	4x	14	23.7	12345	1476	56.7	40	235.9	135.6
	P.chabaudi	*chabaudi*	18.80	4x	14	24.3	5144	1126	58.6	50	207.2	132
	P.falciparum	*3D7*	23.26	14.5x	14	19.4	5560	4338	52.6	53.9	2283	179
	P.gallinaceum		16.93	3x–underway	ND	ND	ND	ND	ND	ND	ND	ND
	P.knowlesi	*H*	23.46	8x	14	38.1	5161	4593	47.4	51.6	2180	224.4
	P.reichenowi		7.38	3x – partial	ND	ND	ND	ND	ND	ND	ND	ND
	P.vivax	*Sal_1*	27.01	10x	14	37.6	5507	4463	48.5	51.2	2164	192
	P.yoelii	*17XNL*	22.94	5x	14	22.6	7865	2556	50.6	54.2	1298	209
ToxoDB[3]	*N.caninum*	*NC Liverpool*	*62.48*	ND	14	54.84	*5761*	ND	ND	ND	ND	ND
	T.gondii[†]	*ME49*	*61.77*	10x		52.39	*8072*					
		VEG	*62.16*	ND	14	52.38	*7977*	8552	57.4	74.3	2431	523.5
		GT1	*60.80*	ND		52.37	*8229*					
EupathDB[4]	*T.annulata*	*Ankara*	*8.35*	8x	4	32.54	*3792*	ND	72.8	70.6	1606	69
	T.parva	*Muguga*	*8.35*	~8x	4	34.1	*4035*	2057	68.4	73.6	1407	94
	B. bovis	*T2B0*	*8.20*	Ongoing	4	41.8	*3671*	2228	70.2	61.5	1514	ND

Note ND = No data

[1] (ABRAHAMSEN *et al.* 2004; AURRECOECHEA *et al.* 2010; LAU 2009; XU *et al.* 2004).

[2] (AURRECOECHEA *et al.* 2010; CARLTON *et al.* 2005; CARLTON *et al.* 2008; HALL *et al.* 2005).

[3] (AURRECOECHEA *et al.* 2010; KHAN *et al.* 2006).

[†] GC content and gene count for *T. gondii* computed from available resources at EupathDB.org. Gene density, percent coding, average gene length and average intron length obtained from average computed for chrIa and chrIb from sequence data for *T. gondii RH* strain.

[4] (AURRECOECHEA *et al.* 2010; BRAYTON *et al.* 2007; GARDNER *et al.* 2005; PAIN *et al.* 2005).

The complete nuclear genome sequences for *P. falciparum*, *T. gondii*, *C. parvum* and *T. parva* are currently available in public databases. These projects show that the apicomplexan genomes are very small in size relative to other eukaryotes, ranging from 8.3 Mb (*T. parva*) to ~64 Mb (*T. gondii*). The estimated gene density and count varies between these parasites, although caution should be taken when comparing these genomes as there are many discrepancies in the annotated data sets (Table 1.1) (WAKAGURI *et al.* 2009). The number of introns, GC content and number of chromosomes also varies. *P. falciparum* and *T. gondii* contain similar chromosomal numbers, with a total of 14 chromosomes for both species, where as *Theileria* species only have four (ABRAHAMSEN *et al.* 2004; GARDNER *et al.* 2005; KHAN *et al.* 2006; KHAN *et al.* 2007). A striking difference observed between the *T. gondii* genome and other apicomplexans currently sequenced is the difference in genome size. The *T. gondii* genome at ~64 Mb is nearly three times larger than the *P. falciparum* genome and up to seven times larger than *C. parvum* and *T. parva* genomes (Table 1.1). This difference may be due in part to the level of gene density, and the number of introns present per gene (KHAN *et al.* 2007). *T. gondii* contains up to 5 introns per gene with larger average gene and intron length, as opposed to 1, 2 introns found in *P. falciparum* and *T. parva* (Table 1.1). Only a small fraction (5%) of the annotated genes in *C. parvum* harbors introns (ROY and PENNY 2007). Despite extreme genome compactness, the *C. parvum* genome contains up to three-quarters more genes than *P. falciparum* at 1.8 times more gene density (KEELING 2004). The *P. falciparum* genome is extremely AT rich with 19% GC content while the percent GC content range from 34% in *T.*

parva to 52% in *T. gondii* (Table 1.1) (ABRAHAMSEN *et al.* 2004; GARDNER *et al.* 2005; KHAN *et al.* 2006).

Genome sequencing data within the genus *Plasmodium* also shows that only 60% of the genes identified are shared by *Plasmodium* species. The remaining genes are unique to each species, and have been shown to be involved in virulence and host specificity (CARLTON *et al.* 2001). The three *Toxoplasma* genome sequences available in the databases represent the three clonal lineages (Type I-III). These strains differ dramatically in infectivity and virulence, with type I (GT1/RH), being extremely virulent, as compared to type II (ME49), and type III (VEG) (KIM and WEISS 2004). These genomic studies have provided profound insights about genome organization and the genetic architecture of these poorly understood pathogens. This builds the foundation for further studies to comprehend the physiological processes of the apicomplexans that with any luck will advance the current understanding of, and treatment associated with these deadly parasites.

Genome innovation

The increased emergence of drug resistance in the apicomplexans, exemplifies their efficiency in quickly adapting to their environmental niche. These parasites have developed complex strategic mechanisms to invade host cells, evade immunological defenses, and hijack necessary nutrients to facilitate their survival. This has made these parasites extremely successful (TOMLEY 2009). Therefore it is of acute interest to biologist to determine the processes governing genome innovation in these organisms. In many eukaryotes, repeats (repetitive DNA) within the nuclear genome provide a

source of genomic plasticity by providing the framework for genome rearrangement, generation and deletion of genes, gene shuffling, and modulation of gene expression. Repetitive DNA broadly classified into two major categories, tandem and interspersed, can be acquired from different sources (WICKSTEAD *et al.* 2003). This includes the acquisition of new genes from horizontal gene transfer, lateral gene transfer (Figure 1.3), and the proliferation of mobile genetic elements known as transposable elements. These factors will be discussed in context of apicomplexan genome innovation in the following sections.

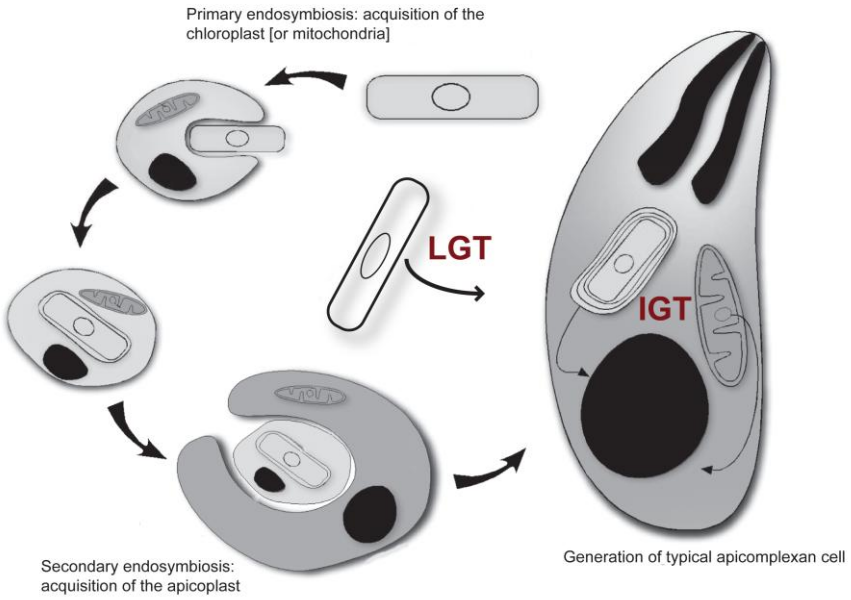


Figure 1.3 Evolution and genome innovation in an apicomplexan cell.

Endosymbioisis and gene transfer in the evolution of the apicomplexans. First, primary endosymbiosis of cyanobacterium by a eukaryotic cell giving rise to a photosynthetic algal cell. Subsequently, secondary endosymbiosis of the photosynthetic algal cell by another eukaryotic cell occurs. Following the endosymbiotic events, multiple waves of intracellular gene transfer (IGT) occur to the point the mitochondria (rectangular grey boxes) and nucleus (solid black circles) of the endosymbiont is completely lost finally giving rise to a typical apicomplexan cell. Within the apicomplexan cell IGT is still occurring and lateral gene transfer from other prokaryotic cell (white rectangular box) can also occur. Figure modified from original [1].

[1] (HUANG and KISSINGER 2006).

Lateral and intracellular gene transfer in the apicomplexans

The process of gene transfer either from genetically unrelated organisms referred to as horizontal/lateral gene (LGT) transfer, or the transfer of genes from endosymbionts known as intracellular gene transfer (IGT) has been implicated in the evolution of the apicomplexans (HUANG and KISSINGER 2006; HUANG *et al.* 2004b). IGT involving plastids and mitochondrial DNA is also documented for many other eukaryotic organisms (ANDERSSON 2005; MOURIER *et al.* 2001). Indeed current mitochondrial genomes, ranges between 180-2400 kb in plants to 15-20 kb in animals where as a typical α-proteobacterium genome is from 1 Mb to 9 Mb in size (ANDERSSON *et al.* 2003; KURLAND and ANDERSSON 2000). So where did all the DNA go? There is mounting evidence that much of the DNA has moved to the nucleus. Examination of 85 fully sequenced eukaryotic genomes shows the ongoing transfer of mitochondrial DNA (mtDNA) to the nucleus and the evolutionary impact of such processes. In plants, like *Arabidopsis thaliana*, transferred mtDNA and plastid DNA (ptDNA) accounts for 0.29% of the total nuclear genome, and in *Apis mellifera* (honeybee), 0.08% of the nuclear genome is derived from mtDNA (BEHURA 2007; HAZKANI-COVO *et al.* 2010; RICHLY and LEISTER 2004a). Although for most part, current relocation of ptDNA and mtDNA to the nuclear genome results in pseudogenization or non-functional genes, on occasion, functionalization of transferred organelle DNA has occurred. For instance, exonization of nuclear mtDNA and ptDNA (numts & nupts) has been documented for *Saccharomyces cerevisiae*, *Homo sapiens*, *A. thaliana* and *Oryza sativa* (NOUTSOS *et al.* 2007).

Genes of algal and cyanobacterial origin has also been documented within the apicomplexans' nuclear genomes. This indeed is not surprising given that the modern apicomplexan cell is a host to three distinct genomes: two acquired from endosymbiotic processes – the apicoplast and the mitochondria, and the third – the nuclear genome (Figure 1.1, Figure 1.3) (AJIOKA 2005; KEELING 2009; OBORNIK *et al.* 2009). Reduction of the *P. falciparum* apicoplast and identification of nuclear genes targeted back to the apicoplast suggest a unidirectional transfer of apicoplast genes to the nuclear genome and subsequent loss within the organelle (GARDNER *et al.* 2002). IGT within the apicomplexans is also exemplified by the detection of nuclear mitochondrial derived DNA commonly referred to as numts in *T. gondii* (OSSORIO *et al.* 1991). Furthermore, isolation of plastid-derived genes within the *C. parvum* nuclear genome provides evidence for loss of the apicoplast rather than complete lack of the organelle within this genus (HUANG *et al.* 2004a). The full impact of IGT in the evolutionary processes of these parasites cannot completely be appreciated. However, the isolation of genes of significant impact on these parasites survival, that are distinct from their host, is indeed important in both furthering our knowledge and aiding current and future studies of this phylum.

The extent of lateral gene transfer is not clear within the apicomplexans, and most documented cases of LGT involve *Cryptosporidium* spp. About 0.5 to 2.5 % of 9.1Mb genome of the *C. parvum* is hypothesized to have a prokaryotic origin. In 2004, Huang et al showed that most of these prokaryotic genes are associated with metabolism of nucleotides and amino acids, energy production and various other

biochemical processes that are essential to the survival of this organism (HUANG *et al.* 2004a)

The mechanism of gene transfer, either IGT or LGT, is not fully understood. In the case of IGT, repair process of double strand DNA breaks (DSBR) via non-homologous end joining (NHEJ) is proposed, following the breakdown of membrane compartments (BURMA *et al.* 2006; HAZKANI-COVO and COVO 2008). In this process, organellar DNA is used to 'patch' chromosomal breakpoints while requiring very little homology between the donor and target sites and on occasion generates small target duplications flanking the inserted mtDNA. This process has to some extent been experimentally used to explain numt accumulation in both the human and yeast genome (LEISTER 2005). The mechanism underlying gene transfer in the apicomplexans still remains to be determined. Despite that, gene transfer events via IGT and LGT have undeniably had considerable impact on genome innovation and evolution, evidenced by the successful parasitic lifestyles of species within the Apicomplexa phylum.

Transposable elements and apicomplexan genomes

Transposable elements (TEs) are mobile pieces of DNA capable of moving from one genomic location to another. They do so either by encoding the necessary proteins to mobilize their transposition or hijacking the proteins encoded by related elements (CRAIG *et al.* 2002). These elements are divided into two major classes based on their transposition mechanisms in integrating into the host genome (Figure 1.4). Class 1 elements or retrotransposons move via an RNA based mechanism, encoding the enzyme reverse transcriptase to reverse transcribe the RNA genome into DNA. Class 2

elements or DNA transposons move via a "cut and paste" process using the transposase enzyme (FESCHOTTE and PRITHAM 2007b; WICKER *et al.* 2007).

Since their first discovery in the 1950s by Barbara McClintock, TEs have been shown to constitute the most dynamic fraction of the genomes of many plants, animals, and fungi (PRITHAM 2009). They account for 3% of the yeast genome, 45% to 50% of both the mouse and human genomes, and 50% to 80% of the maize and barley genomes (LANDER *et al.* 2001; SANMIGUEL *et al.* 1996; VICIENT *et al.* 1999). TEs have been shown to be key players in genome evolution. They do so by contributing to the raw genetic material for the development of new genes, controlling the expression of neighboring genes, or providing the resources to establish regulatory networks within the genome (FESCHOTTE 2008). These elements are so common and found in such high copy numbers in many sequenced genomes within the databases that they are thought to be ubiquitous and integral components of most, if not all eukaryotic genomes (FESCHOTTE 2008; KIDWELL and LISCH 2001; KING 1992). Therefore it is extremely curious that of the fifteen complete apicomplexan genomes available, a few TEs have been clearly described for only one species, *Ascogregarina taiwanensis* (Figure 1.2) (TEMPLETON *et al.* 2010).

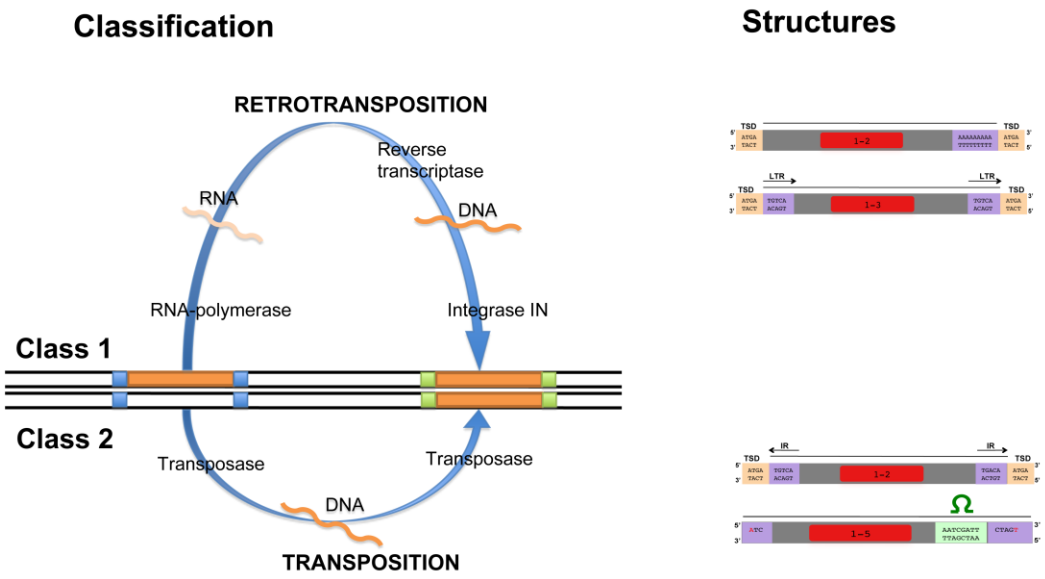


Figure 1.4 Classification of transposable elements.

Classification of transposable elements based on their transposition intermediates and typical structural features of each class. General structural features associated with TEs: red boxes show the open reading frames encoding the proteins for transposition. TSD= target site duplication; LTR= long terminal repeat; IR= inverted repeat; green loop= palindrome.

Interestingly, very little is known about the repetitive DNA content of the apicomplexan genomes. In many other organisms, TEs dominate a substantial fraction of the repetitive DNA within the genome. Within the apicomplexans very few repeat families have been identified. Those reported include the mitochondrial-like REP family in *Toxoplasma* (OSSORIO *et al.* 1991), the sub-telomeric associated repeats in *Plasmodium* (TARE 1-6 in *P. falciparum*) (FIGUEIREDO *et al.* 2000), and one report of TEs in several *Plasmodium* species (DURAND *et al.* 2006). As previously noted, several class 1 elements have also been identified in *A. taiwanensis* (TEMPLETON *et al.* 2010).

However, none of these elements have been extensively analyzed in light of their evolutionary impact on the apicomplexans.

Aims of this study

This study aims to examine the repetitive DNA content of several unicellular eukaryotes with the intent to identify TEs and document their distribution in the eukaryotic taxa. In many genomes a large fraction of the repetitive DNA is derived from TEs. Therefore understanding the nature of the repetitive DNA complement will aid in identifying TEs, if they are truly present. To this end, I analyzed the genomes of five apicomplexan species along with three other unicellular organisms, using two complementary methods. The first method is a *de novo* repeat identification program called RepeatScout that identifies repetitive DNA in a genome based solely on the intrinsic repetitive nature of the sequence (PRICE *et al.* 2005). The second method relies on homology-based tools and sequence analysis to classify the identified repeats. Chapter 2 provides details on the methods employed and the resulting data.

An additional aspect of this study focuses on studying the genome biology of the apicomplexans, using the model apicomplexan parasite, *T. gondii*. This part of the study was done in collaboration with a postdoc in our lab, Cheng Sun. Taking advantage of the newly sequenced *T. gondii* mitochondrial DNA, we performed a systematic examination of mitochondrial derived DNA or numts in *T. gondii* to determine their contribution to genome innovation. Our findings suggest that the accumulations of numts are actively shaping the genetic architecture of this important

human parasite. The details and resulting data for this analysis are documented in Chapter 3.

CHAPTER 2

ARE TRANSPOSABLE ELEMENTS UBIQUITOUS IN EUKARYOTES?

Abstract

The genomes of eukaryotic organisms are heavily populated by repetitive mobile elements referred to as transposable elements (TEs). Despite being thought of as 'junk-DNA', these elements have had significant impact on the evolutionary trajectories of their hosts. Availability of genome sequencing data provides us with a unique opportunity to explore the breadth of TE distribution in some unicellular eukaryotes. To this end, the genomes of six Chromalveolate species, one Plantae, and one Unikont were investigated. Complementary strategies using, *de novo* tools (RepeatScout) along with homology-based strategies were employed. Here we show that TEs are not present in all eukaryotes and we find that a common feature of TE-free eukaryotes is that they are obligate intracellular parasites. Consequentially our results reveal that unicellular eukaryotic parasites are capable of generating dynamic genomes without the involvement of TEs.

Introduction

Transposable elements (TEs) commonly referred to as 'jumping genes' have the ability to move from one genomic locus to another. Typically viewed as 'parasitic DNA', these self-replicating elements are capable of reaching very high copy numbers

within an invaded host genome. This is due to their intrinsic ability to transpose at a frequency that surpasses the replication of the host genome. The dynamics and mechanisms by which TEs increase in copy number is associated with whether they replicate via RNA, class 1 (retrotransposons) or a DNA intermediate (class 2 DNA transposons) (BIÉMONT and VIEIRA 2006; FESCHOTTE and PRITHAM 2007b; WICKER *et al.* 2007). The reproductive mode of the host is also influential in the accumulation and maintenance of these genome parasites. Sexual reproduction facilitates the spread and persistence of TEs within a population by eliminating some of the negative mutational loads that accumulates due to proliferation of these elements, thus allowing TEs to outcompete host genes (ARKHIPOVA and MESELSON 2000; HICKEY 1982). However, in asexually reproducing organisms, uncontrollable transposition events can lead to rapid extinction of a population, therefore transposition is hypothesized to be under tight control by selection or these elements are completely purged from the genomes (ARKHIPOVA and MESELSON 2005; DOLGIN and CHARLESWORTH 2006).

The life history and ecological demographic of organisms may also affect the spread and proliferation of mobile elements. Parasitic organisms, mainly obligate intracellular parasites are characterized by genome miniaturization due to massive gene loss and elimination of non-coding DNA (CAVALIER-SMITH 2005; KEELING and SLAMOVITS 2004; LYNCH and CONERY 2003). As TEs mainly substantiate the non-coding regions of a given host genome, there may be a strong selective constraints against the accumulation and transposition of mobile elements especially within intracellular parasites. This may be due to the natural selective restraint placed on these

organisms to maintain a balance between their genome and cell volume size, which are known to be positively correlated (GREGORY 2005; PRITHAM 2009; WICKSTEAD *et al.* 2003). This phenomenon of extreme genome reduction and elimination of mobile elements has been observed for several obligate pathogenic bacteria (CASADEVALL 2008). Evidently, the success or lack thereof of TEs within a population or an individual genome, is govern by a complex interplay between natural selective forces, population dynamics of TEs, host defensive mechanisms, and a plethora of intrinsic and extrinsic features that have yet to be completely understood.

The emergence of vast amount of genomic data has yielded unprecedented insights about the tremendous impact transposable elements can exert on host genomes. As was suggested by Barbara McClintock, the first discoverer of these elements over 50 years ago, TEs are capable of drastically reshaping the genomes they inhabit (OLIVER and GREENE 2009). TEs populate a substantial fraction of many eukaryotic genomes, ranging from 3% in yeast, 45% in humans and over 77% in some plants (BIÉMONT and VIEIRA 2006; FESCHOTTE and PRITHAM 2007b; WICKER *et al.* 2007). Based on their natural mutagenic role, and their ability to facilitate ectopic recombination, these elements represent a suitable force for generating genomic plasticity. Indeed, TE driven recombination events, gene duplications, and generation of novel genic sequences has been documented for many organisms (VOLFF 2006). TEs are also hypothesized to have had preponderant contributions to eukaryotic evolutionary complexity (BOWEN and JORDAN 2002). TEs are implicated in the emergence of the immune system in jawed vertebrates, and have been shown to be essentially involved in human placental

morphogenesis (BOHNE *et al.* 2008). Having been discovered in most of the genomes present in public databases, TEs are said to be ubiquitously distributed across eukaryotic taxa, however very little is known about the TEs distribution in unicellular eukaryotic genomes.

Unicellular eukaryotic parasites are characterized by dynamic genomes in concert with having to constantly accommodate to their ever-changing environments imposed by parasite-host interactions (BAKER 1994). Given the dynamic nature of these unicellular eukaryotic genomes and knowing that the rapid turnover of TEs serves as excellent tools for generating genomic plasticity, it is interesting to note that the genome papers for several unicellular eukaryotes have failed to report any TEs (ABRAHAMSEN *et al.* 2004; BRAYTON *et al.* 2007; GARDNER *et al.* 2005; GARDNER *et al.* 2002; KATINKA *et al.* 2001; XU *et al.* 2004). One such representative group with reported dearth of TEs is within the phylum Apicomplexa. This phylum contains a diverse group of 5,000 protozoan parasites, including the causative agent of malaria, *Plasmodium* and others with tremendous medical and socioeconomic impact worldwide (MORRISON 2009). The genomes of fourteen apicomplexan species ranging in size from 8.3 Mb to 64.0 Mb have been sequenced and reveal a significant amount of genome plasticity in terms of size, AT-richness, introns and gene density. Interestingly, of the fourteen complete apicomplexan genomes available, only a handful of TEs have been described for one species, *Ascogregarina taiwanesis* (TEMPLETON *et al.* 2010). It is unclear whether TEs are truly missing within these genomes, or that TEs have not been detected due to fact

that the search of TEs involves using sequence homology to the distantly related TEs families currently present in the databases.

The focus of this study is to determine the extent of TE distribution covering a wide taxonomic diversity of unicellular eukaryotes, including six Chromalveolates, one Unikont and one Plantae. To this end, the genomes of five apicomplexan species, *Perkinus marinus*, one cyanidiophyte, and one microsporidian were investigated, using two complementary methods. The *de novo* repeat identification program called RepeatScout along with homology-based tools was used to explore and classify the repetitive DNA in these organisms.

Detecting TEs in newly sequenced genomes

Repeat identification

TEs make up a large proportion of the repetitive DNA of many genomes, therefore understanding the nature of the repetitive DNA complement will aid in identifying TEs if they are present. RepeatScout was used to determine the repetitive DNA content for the apicomplexans, *Plasmodium falciparum* (22.9Mb), *Cryptosporidium parvum* (9.1Mb), *Theileria parva* (8.3Mb), and *Toxoplasma gondii* (61.7Mb). These organisms represent some of the most successful parasites in the world, having developed a myriad of strategies to facilitate their survival within their invaded hosts. Diverse features enigmatic of having plastic genomes characterize these organisms: like genome size, genome organization and GC content. Since TEs may not have been identified in these organisms as a result of methodology limitations, rigorous scrutiny of repetitive DNA content may facilitate the detection of TEs. As a positive

control, the genome of the apicomplexan, *A. taiwanesis* (~20Mb) was also assessed, along with the red alga *Cyanidoschyzon merolae* (16.5Mb), and the genome of the Unikont, *E. cuniculi* was used as a negative control having been reported to harbor no TEs. The genome of the Chromoalveolate, *Perkinsus marinus* (80Mb) was also examined in this study. *Perkinsus marinus* is a prevalent pathogen of oysters, capable of causing massive mortalities in oyster populations. This organism belongs to the Alveolate group and shares close evolutionary relationship to the apicomplexan, and was once placed within the phylum Apicomplexa (LEANDER and KEELING 2004). The species chosen in this study represents unicellular eukaryotes from different branches of the eukaryotic tree of life, allowing for comparative studies of TE accumulation within unicellular eukaryotes from diverse lineages.

The program RepeatScout utilized in this study, works by employing a word/seed extension approach to generate a library of consensus sequences representative of the repeat families within a given genome. Repeated sequences are identified based solely on their intrinsic repetitive nature (PRICE *et al.* 2005). The advantage of using this method is that it eliminates the limitations associated with relying only on traditional homology based tools. Homology based methods detect TEs based on significant level of sequence similarity to previously characterized TE families within the databases. This is a severely limiting factor as newly sequenced genomes may be distantly related to the organisms from which TEs in the databases originates. This method also hinders the identification of new TE families (FESCHOTTE and PRITHAM 2007a). On the other hand, *de novo* tools are not equipped to classify the

repeats identified, so in conjunction to RepeatScout, homology based programs are used to determine the identity of the repeat families.

Repeat classification

Traditionally, the annotation of repeats requires manual comparative sequence analysis to identify diagnostic features of TEs. These include identification of terminal repeats, open reading frames homologous to TE derived proteins, and target site modifications associated with each type of TE family (FESCHOTTE and PRITHAM 2007a; WICKER *et al.* 2007). In addition to this manual analysis, programs such as Repclass(FESCHOTTE *et al.* 2009) can be used to automate several of the steps above, thereby expediting the classification process.

The RepeatScout generated consensus library contains not just TEs (if present), but also a number of other repeats, including satellites, gene families, segmental duplications, tandem and low-complexity repeats. Therefore to facilitate a quick yet efficient method to classify the repeat families, we devised a set of strategies to be used in conjunction with Repclass. This involves the identification of discrete repeat boundaries, followed by manual characterization of the boundaries to assess for the presence of TE-related features, along with investigation for the presence of open reading frames homologous to known TE families. Before these steps are applied, first the RepeatScout libraries were filtered for tandem and low complexity repeats using Tandem Repeat Finder (BENSON 1999) and nseg (WOOTTON and FEDERHEN 1996). Second, in order to eliminate redundancies of repeat families within the libraries, Sequencher (http://www.genecodes.com) was used to assemble repeat consensuses that

overlap over 100-bp with ≥ 90% identity at the nucleotide level. Subsequently, Repclass was run on the refined repeat libraries to automate classification. Repeats classified by Repclass based on homology to previously annotated TEs were confirmed by using the repeat as a query to perform BlastX searches against the NCBI databases (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

A known limiting factor of RepeatScout is the inaccurate or incomplete definition of the repeat ends. Therefore following Repclass, repeats not classified base on the homology module and exceeding 200- bp in length were selected and used as queries to mine individual copies from the corresponding genome in order to determine defined repeat ends. Figure 2.1A (i – iv) illustrates the distribution pattern of the RepeatScout generated repeat families within the genomes. All repeats matching the distribution pattern from [Figure 2.1A (i – iii)] were subjected to further analysis to determine the presence of terminal inverted repeats, long terminal repeats, polypurinic tracts, target modifications (target site duplications) [Figure 2.1B]: all representative features of known TEs. Gene families were also filtered out using the unclassified repeats as queries in BlastX searches against the Uniprot/Swiss protein databases (http://www.uniprot.org).

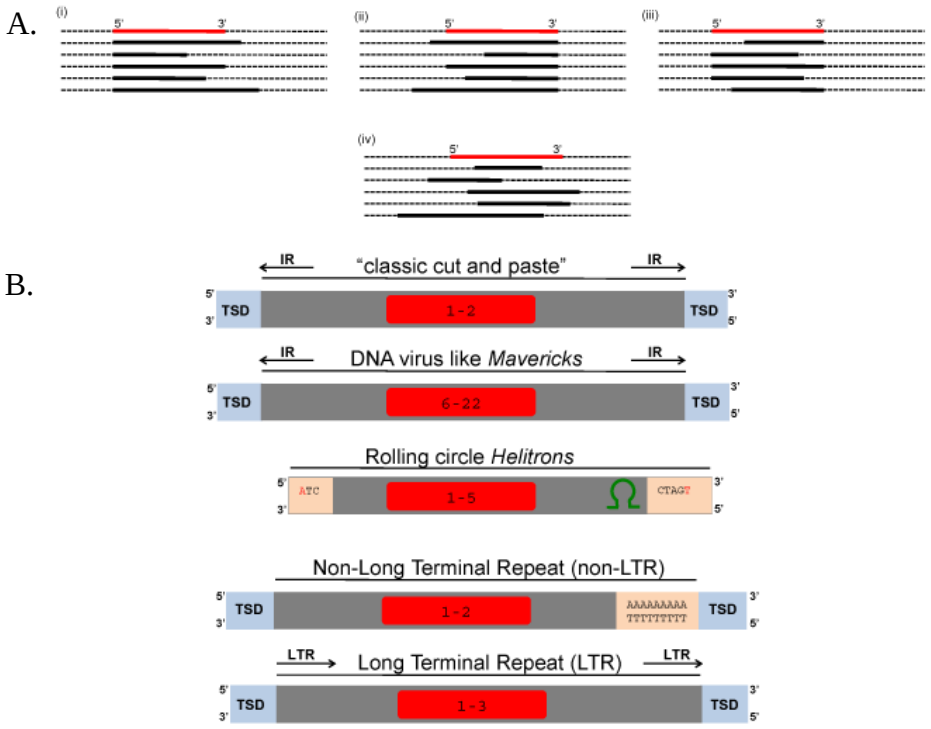


Figure 2.1 Illustration of transposable element associated features and steps for defining a repeat family as a transposable element (TE).

A. The distribution pattern of individual copies of a given RepeatScout consensus repeats (red lines). Black lines show the individual copies of a given repeat consensus generated by RepeatScout. The dash lines represent the non-homologous flanks of the repeat family. A. For (i) only the 5' end of the repeat family is defined, (ii) only the 3' end is defined, (iii) both ends are defined, and (iv) none of the ends are defined. All four repeat distribution scenarios were subjected to homology searches for putative open reading frames (ORFs – red rectangle in B.) encoding protein(s) involved in transposition of a TE. However only repeat families matching (A i – iii) were assessed for structural features typical of TEs. B. Structural features associated with the two major classes of TEs. IR= inverted repeat; TSD= target site duplication. A repeat is said to be a TE if (1) discrete boundaries are found followed by identification of the any of the structural features shown in b, or (2) open reading frames matching previously annotated repeat families is observed.

Paucity of TEs in several unicellular eukaryotes

The number of repeat families within the filtered RepeatScout libraries varies significantly between species ranging from only 12 in *C. parvum* (9.1Mb), to over

10,000 in *P. marinus* (86.1Mb) (Table 2.1). Overall, there seems to be no clear correlation between the number of repeat families identified and genome size (Table 2.1a). For instance 49 repeat families were identified in the 2.9 Mb genome of *E. cuniculi*, whereas only 12 repeats were observed in *C. parvum* (9.1Mb). To avoid misclassification due to contamination of genomic sequences, where possible, only chromosomal assemblies were subjected to the repeat annotation. Following initial classification with Repclass, 0 to < 0.1% of the RepeatScout libraries were classified for the apicomplexans, *C. parvum*, *P. falciparum*, *T. parva* and *T. gondii*. The majority of the repeats classified by Repclass based on homology were ribosomal genes and satellite DNA (data not shown). Manual inspection of all the repeats classified by Repclass using the structural modules and applying the steps illustrated in Figure 2.1 failed to show any TEs. The remainder of the unclassified repeats were then subjected to searches against the Uniprot/Swiss databases which revealed that a significant portion of the libraries were derived from known annotated gene families, satellites, telomeric and subtelomeric repeats and mitochondrial DNA (data not shown).

Applying a similar process to the *E. cuniculi* RepeatScout library resulted in a larger portion of the library being classified by Repclass. However of the five repeats reported as TE based on homology, none were confirmed using Blastx searches. Manual inspection of all the repeats for defined boundaries showed extensive overlapping sequences with no distinct ends. Having met none of the criteria outlined in Figure 2.1, we failed to report any TEs within this genome, thereby corroborating previous reports from the published genome paper (KATINKA *et al.* 2001).

On the other hand, we were able to confirm the annotation of TEs within *A. taiwanensis*, *C. merolae*, and *P. marinus* (NOZAKI *et al.* 2007; TEMPLETON *et al.* 2010) (EP -unpublished data). BlastX searches confirmed all repeats classified as TEs using the protein homology module in Repclass. The identities of the repeats were further confirmed by querying the genomes with the classified repeats and assessing for defined ends. The resulting class of TEs identified within these genomes reveals some interesting patterns. For example, the *A. taiwanensis* harbors only a few elements: four LTR retrotransposons, one non-LTR element and no DNA transposons. The *C. merolae* genome contains only a few inactivated DNA transposons (open reading frames with multiple stop codons) for which we could only identify very degenerate terminal inverted repeats for one family. The one non-LTR family identified within the RepeatScout library was also inactivate, with no positive characterization of the ends. However, diverse families of both DNA and RNA elements populate the *P. marinus* genome. The Repclass classified TEs were mined from the *P. marinus* genome and we were able to clearly identify features associated with each TE family (clear boundaries, ORF, and target duplications flanking elements termini).

Interestingly, the class of TEs (LTR and non-LTR retrotransposons) observed in *A. taiwanensis*, and *C. merolae* closely resembles the pattern observed in several other unicellular eukaryotes like *Entamoeba histolytica*, *E. dispar*, the trypanosomatid genomes *Trypanosoma cruzi*, and *T. brucei*, and in *the Leishmania major* genome, where only or mostly class 1 retrotransposons have been described (see Table 2.2) (BRINGAUD *et al.* 2008; PRITHAM *et al.* 2005). Several of the LTR retrotransposons

isolated both in *A. taiwanensis* and *P. marinus* show signs of recent activity, with 100 % identity between the long terminal repeats and intact open reading frames, suggesting recent acquisition of these elements by these parasites. Taken together, these data show differential accumulation of TEs in several unicellular eukaryotes, with no TEs detected in the genomes of all the obligate intracellular parasites assessed in this study.

Table 2.1 Identification and classification of repetitive DNA in several unicellular eukaryotes

Organism	Supergroup	Env. niche	Sexual reproduction	Host (s)	Genome size (Mb)	Sequencing cover-age	# Of contigs/ chr.	# Of repeat families identified	TE families classified	(%) of library classified by REPCLASS
Ec	Unikont	Strict ICP	Unclear	Mammals	2.9	3X	11	49	None	42.9
Cm	Plantae	Extremophile	Unclear	N/A	16.5	11X - complete	20	299	Non-LTRs & DNA	21.2
Pm	Chromoalveolate	Facultative ICP	Unclear	Oysters	86.05	Ongoing	Ongoing	10,330	Non-LTRs, LTRs, & DNA	21.8
At	Chromalveolate	Intracellular parasite	Yes	Mosquitoes & sand flies	6.18/(20)[a]	Ongoing	3,434	157	LTRs & Non_LTR	30.6
Cp	Chromoalveolate	Strict ICP	Yes	Mammals	9.1	13X	8	12	None	0
Pf	Chromoalveolate	Strict ICP	Yes	Mosquito & human	22.9	14.5X	14	114	None	<0.1
Tp	Chromoalveolate	Strict ICP	Yes	Tick, & bovine	8.3	8X	4	88	None	<0.1
Tg	Chromoalveolate	Strict ICP	Yes	Feline, & warm-blooded animals	61.7	12X	14	419	None	<0.1

NOTE. – Ec= Encephalitozoon cuniculi; Cm = Cyanidoschyzon merolae; Pm = Perkinsus marinus; At= Ascogregarina taiwanensis; Cp = Cryptosporidium parvum; Pf = Plasmodium falciparum; Tp = Theileria parva; Tg = Toxoplasma gondii. ICP = intracellular parasite

[a] Number in parenthesis is the minimal estimation of true genome size (Templeton 2010)

Table 2.2 Transposable element composition in various unicellular eukaryotes

Organism	Phylum	Life history	Sexual reproduction	Host (s)	Genome size (Mb)	Sequencing cover-age	# Of contigs/scaffolds/chromosomes	# Of repeat families identified	TE families classified/identified in genome	Citations
Ca	Ascomycete	Int. parasite	Yes	Human	14.3	10.9X	8	37*	Retrotransposons, & DNA	(FESCHOTTE *et al.* 2009)
Um	Basidiomycete	Int. parasite	Yes	Maize, teosinte	19.7	10X	274	25*	Retrotransposons, & DNA	
Ro	Zygomycete	Unclear	Yes	Human	45.3	12X	389	496*	Retrotransposons, & DNA	
Pg	Basidiomycete	Int. parasite	Yes	Cereal crops	81.5	7.8X	4,557	2,085*	Retrotransposons, & DNA	
Eh	Archamoebae	Ext. parasite	Unclear	Human	23.8	~12.5X	888	N/A	Retrotransposons	(LOFTUS *et al.* 2005) (PRITHAM *et al.* 2005)
Tc	Kinetoplastid	Int. parasite	Unclear	Mammals	34	~2X	5,489	N/A	Retrotransposons	(BRINGAUD *et al.* 2008)
Tb	Kinetoplastid	Ext./int parasite	Unclear	Mammals, Tsetse fly	26	~3.53	11chr/30contigs	NA	Retrotransposons	
Lm	Kinetoplastid	Int. parasite	Unclear	Sand flies, human	32.8	~5X	36	NA	Retrotransposons	
Pi	Oomycetes	Free-living pathogen	Yes	Tomato, papas, and potato	240	~7.6X	4,921	NA	Retrotransposons, & DNA	(HAAS *et al.* 2009)
Tv	Trichomonad	Ext. parasite	Unclear	Human	160	~7.2X	17,290	N/A	DNA transposons	(PRITHAM 2009)

NOTE - Definition - intracellular parasite refers to organisms whose reproductive cycle is restricted within another organism.
Ext = extracellular; Int. = intracellular; N/A = not applicable
* Number of repeat families assessed from Feschotte *et al*. study 2009.
Ca = *Candida albicans; Um = Ustilago maydis; Ro = Rhizopus oryzae; Pg = Puccinia graminis; Eh = Entamoeba histolytica; Tc = Trypanosoma cruzi; Tb = T. brucei; Lm = Leishmania major; Pi = Phytophthora infestans; Tv = Trichomonas vaginalis*

Factors contributing to dearth of TEs in several unicellular eukaryotes

Invasion and proliferation of TEs within a given genome is governed by a complex interplay between both the invading parasite (TE) and the targeted host. The environmental niche occupied by the target organism may play a very important role in modulating TE invasions (PRITHAM 2009). Interestingly, the organisms assessed in this study are either parasitic in nature, or inhabit extreme environmental niches. Being residents of the cellular environment of other organisms might actually confer some protection from invasive TEs. This may occur as a result of the intracellular environments preventing these organisms from coming into contact with external vectors like viruses, which can facilitate the horizontal transmission of TEs between species (PRITHAM 2009). This trend could potentially explain the lack of TEs in *C. parvum*, *E. cuniculi*, *P. falciparum*, *T. gondii*, and *T. parva*, all of which are strict obligate intracellular parasites. Although the ancestor of these species might have harbored TEs, the large-scale genome reduction that is associated with conversion to becoming obligate intracellular dwellers might have lead to the initial elimination of TEs. The more these parasites become reliant on their host, then the more important it becomes to maintain TE ridden genomes which explain the persistence of TE-free genomes. The finding of diverse population of TEs in four different *Entamoeba* species that are either free living or extracellular parasites, and the lack of TEs in the unicellular obligate microsporidian parasite, *E. cuniculi* is indicative of such a process (KATINKA *et al.* 2001; PRITHAM *et al.* 2005). In contrast, the genome of the facultative intracellular parasite, *P. marinus* do contain TEs, suggesting that the extent of genome reduction and

therefore TE elimination may rely on various other factors that could be related to level of dependency of the parasite to the invaded host.

An additional and related factor that may have led to the maintenance of TE depleted genomes could be attributed to the location of the intracellular environment inhabited by the pathogens. There are major two types of intracellular locations parasites can inhabit: vesicular compartments including phagosomes or parasite induce vacuoles, and non-vesicular compartments or non-enclosed cytoplasmic location within the host cell (CASADEVALL 2008). Parasites like *T. gondii* generate parasitophorous vacuoles on entering the host cell to protect from degradation by the host (PLATTNER and SOLDATI-FAVRE 2008). Coincidentally this parasite induced environmental niche may impose further protective barrier from the invasion by TEs. A similar vacuole is also established on *P. falciparum* and *C. parvum* host cell invasion; however *T. parva* does not follow this trend and freely inhabit the cytoplasm of the host cell (PLATTNER and SOLDATI-FAVRE 2008).

In summary, we have demonstrated the lack of TEs in several unicellular eukaryotic parasites that are characterized by very dynamic genomes. We observe that one common recurring feature of the TE-free eukaryotes is that they are obligate intracellular parasites, with fairly reduced genomes. Further examination of the increasingly available genomic sequences of intracellular unicellular parasites may provide fundamental insights at the dynamism involving accumulation, and persistence of TEs within unicellular eukaryotic organisms.

Concluding remarks

A comparative survey of the repetitive DNA complement within several unicellular eukaryotes reveals that TEs are not everywhere. Now of burning interest, is deciphering the factors influencing TE diversity and composition in eukaryotes. With the availability and analysis of many more unicellular eukaryotic genomes, we may be able tease apart or possibly answer some of the puzzling questions as to why some organisms are successful in maintaining TE-free genomes while TEs dominate the DNA content of others. Unicellular parasites provide the most suitable systems for answering these questions. These parasites are generally endowed with fairly manageable genomes where thorough analysis of the complete repetitive DNA complement is feasible. This makes it possible to confirm the presence/absence of TEs, given our current understanding of the characteristics of TEs.

CHAPTER 3

EVOLUTIONARY FATE AND CONSEQUENCE OF NUCLEAR MITOCHONDRIAL DNAS IN THE HUMAN PARASITE, *TOXOPLASMA GONDII*

Abstract

The *Apicomplexa* is a protozoan phylum of around 5000 species, most of which are medically or veterinary important parasites. While organelle-to-nuclear DNA transfer represents a significant driving force for genome innovation in eukaryotes, it has not been well defined in apicomplexans. Taking advantage of our recently sequenced mitochondrial DNA, we performed a systematic examination of nuclear mitochondrial DNAs (numts) in the model apicomplexan, *T. gondii*. We found that the genome harbors the largest fraction of numts ever reported, at 1.88% density. Most of the numts arose from independent insertion events rather than post-integration duplications involving segmental duplications. Although the three prevalent *T. gondii* isolates share a very recent ancestor, isolate-specific numts can still be found, suggesting that numt acquisition/deletion is an ongoing process that contributes to the divergence of the isolates. Through comparative analysis of numts between *T. gondii* and its cousin, *N. caninum*, we found that numts have a much higher retention rate in *T. gondii* than in *N. caninum*. Considering their genome organization, a high fraction of the numts was found residing within or near *T. gondii* genes, which might foster their occasional

functionalization. Bioinformatic method was employed to identify potentially functional numts. Numts found residing upstream of host genes were subjected to experimental validation. Results from reporter assay indicate that these numts carry *cis*-elements that can regulate gene expression. Together our results suggest that sequences of mitochondrial origin accumulating in the genome of *T. gondii* are actively shaping the genetic architecture of this important human parasite.

Introduction

The *Apicomplexa* is a protozoan phylum of around 5000 species, all of which are obligate intracellular parasites (LEVINE 1988). Many apicomplexans impart significant medical, veterinary and socioeconomic burdens worldwide. *Plasmodium* causes malaria, which deprives 1-5 million human lives every year (CARLTON *et al.* 2001; GARDNER *et al.* 2002). *Toxoplasma* causes toxoplasmosis, a cosmopolitan disease that infects as much as one third of the world's human population (KIM and WEISS 2004). While the parasite rarely causes symptoms in healthy adults, infection is of greater concern in immunosuppressed or pregnant patients (BELANGER *et al.* 1999; GILBERT *et al.* 2001). Cryptosporidium, a water-borne pathogen, causes cryptosporidiosis, which is generally limited to diarrhea in the immunocompetent, but can be severe and sometimes fatal in the immunocompromised (ABRAHAMSEN *et al.* 2004). *Theileria*, a tick-borne pathogen, causes East Coast fever in cattle leading to 1 million cattle deaths annually in sub-Saharan Africa (GARDNER *et al.* 2005).

Current treatments for apicomplexans are threatened by the emergence of drug resistance, and there is an urgent need to develop novel, sustainable therapies (TOMLEY

2009). All good therapeutic targets have one feature in common: the target molecule or pathway in the parasite is sufficiently distinct from similar molecules or pathways in the host so that therapeutic compounds can be distinguished from the host. However, apicomplexan parasites, like us, are eukaryotic organisms. Thus, there are fewer novel targets available for therapeutics. Consequently, it is exceptionally important to understand the fundamental mechanism underlying genome innovation in apicomplexan parasites.

In eukaryotes, DNA of endosymbionts (mitochondria and chloroplasts) is transferred into the nucleus (KLEINE *et al.* 2009; LEISTER 2005). There are basically two kind of such DNA transfer: the ancient relocation of entire plastid and mitochondrial genes to the nucleus and the ongoing generation of nuclear mitochondrial DNAs (numts) and nuclear plastid DNAs (nupts). The process of organelle-to- nucleus DNA transfer now reported for 85 fully sequenced eukaryotic genomes has proven to be a significant driving force for gene and genome innovation in eukaryotes (KLEINE *et al.* 2009). Apicomplexans also have an evolutionary history involving endosymbiosis, therefore it may not come as a surprise that the ancient relocation of entire plastid and mitochondrial genes to the nucleus have been reported for this group (HUANG *et al.* 2004a; OSSORIO *et al.* 1991). However, to our knowledge, no systematic study has been reported about the ongoing generation of numts and nupts in apicomplexans, and little is known about the evolutionary fate and consequence of these transferred organelle fragments in the host genome.

T. gondii is the best model system to study the biology of the Apicomplexa (KIM and WEISS 2004). Taking advantage of the newly sequenced *T. gondii* mitochondrial DNA, we performed a systematic examination of numts in *T. gondii* to determine their contribution to genome innovation. We found that the genome harbors the largest fraction of numts ever reported. Also, we found that the accumulation and loss of numts is an ongoing process in *T. gondii* that contributes to the divergence of its isolates. In addition, numts residing upstream of host genes were found containing *cis*-elements that can up or down-regulate gene expression. Together our results suggest that rather than just junk DNA, sequences of mitochondrial origin accumulating in the genome of *T. gondii* are actively shaping the genetic architecture of this important human parasite.

Results

Identification of numts in *T. gondii*

To determine the region of the nuclear genome that is derived from mitochondrial DNA we used the program RepeatMasker (http://www.repeatmasker.org/), which employs a Smith-Waterman algorithm. This program delineates the beginning and end of each region in the genome that is homologous to the mitochondrial genome. The output of RepeatMasker provides a detailed annotation of the identified numts, which facilitates downstream analyses. Using this program we identified ~10,000 *numt* insertions in each of the genomes of the three *T. gondii* strains, which equal more than 1 Mb of mitochondrial DNA in each nuclear genome (Table 3.1). The numt composition is very similar in all three strains ranging from 1.87 % in VEG to 1.88% in ME49 and GT1 strains. To obtain reliable

and conservative estimates and to exclude the possibility that some contigs actually represent the mitochondrial genome itself, we filtered out contigs that failed to map onto chromosomes, and only chromosomal assemblies were queried for numts. Numts typically represent < 0.1% of total genomic DNA, with a few exceptions noted in some plant (0.26%) and yeast (0.29%) species. The numt density presented here for the *T. gondii* genome constitutes the largest ever reported for any eukaryote at 10 times more than the honeybee genome, which is the current record holder for metazoans, and over a 100 times more than the human genome (HAZKANI-COVO *et al.* 2010).

Table 3.1 The amount of identified numts in three *Toxoplasma.gondii* strains

Strain	Genome size (Mbps)[1]	MtDNA genome size (kb)	Hits[2]	Base pair occupied[2]	Percentage of genome (%)
GT1	60.8	7	9,827	1,139,048	1.88
ME49	61.8	7	9,958	1,159,821	1.87
VEG	62.2	7	10,074	1,163,635	1.87

To differentiate between multiple independent mitochondrial insertions and segmental duplications as mechanisms for the accumulation and dispersals of numts, we reasoned that duplications would frequently involve flanking regions of the host genome that were not of mitochondrial origin. Consequently, if a numt was duplicated in the nuclear genome, the numt along with its flanking region will likely be involved in the process. A numt was considered the product of duplication if 100bp of the flanking sequence was duplicated. Using this criterion, we found that 172 numts were the product of segmental duplication within the ME49 strain (Table S1). As there are a total of 9958 numts in this genome, approximately (172/9958) 1.7% of the numts were formed by segmental duplicative process, thereby revealing that most of the current numts in *T. gondii* arose from independent insertions of mitochondrial sequences rather than post-insertion duplications, a pattern similarly observed for human numts (BENSASSON *et al.* 2003).

[1] Genome size data obtained from the number of bases of genome masked in RepeatMasker
[2] Number of insertions and base pair composition based on RepeatMasker

To determine the mitochondrial origin of the numts, we used both RepeatMasker and BLASTN to identify homologous copies of numts in the *T. gondii* mitochondrial genome. Results revealed that numts originate from all regions of the ~7kb mitochondrial genome including both coding and non-coding regions (Table S2).

Size and chromosome distribution of numts

The RepeatMasker output provides the start and end coordinates for every numt identified. The coordinates were used to infer the size of each of the numts. The sizes were plotted to determine the range and frequency of occurrence, which are comparable for all three *T. gondii* strains (Figure 3.1). Over 90% of the numts fall within a100bp to 200bp range, although a few larger numts were also found. This size distribution is similar to that observed for yeast and rat species, but is dramatically different from that of *Arabidopsis*, *Neurospora*, and *Ciona* (RICHLY and LEISTER 2004a).

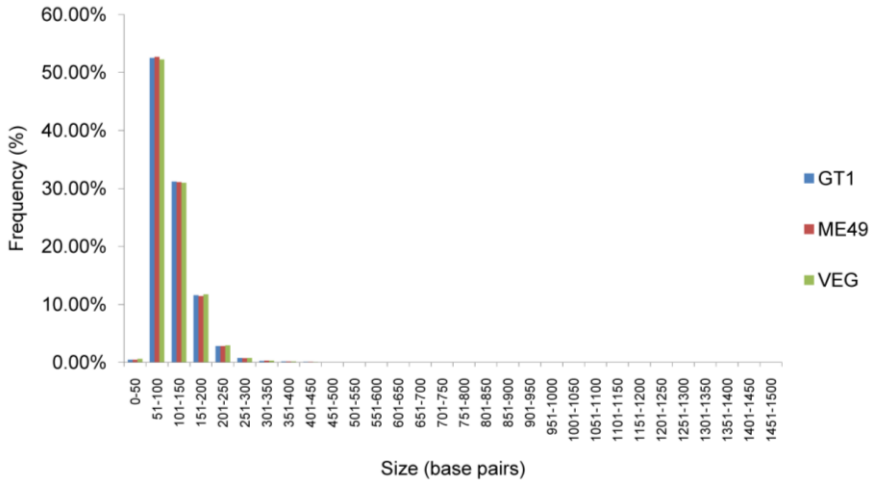
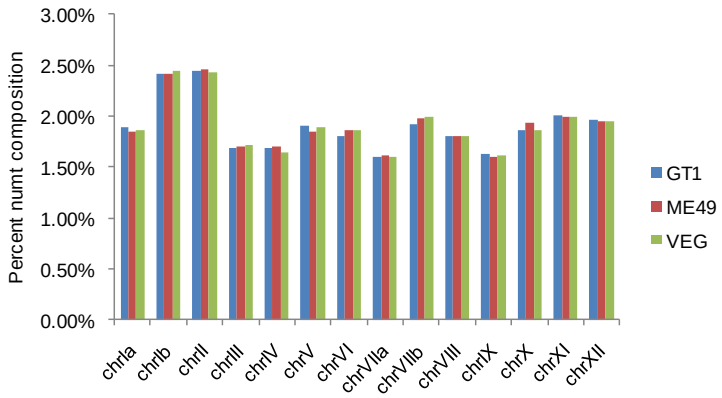


Figure 3.1 Size distribution of numts in the three *T. gondii* strains.

Plotted is the size distribution of numts in all three strains of *T. gondii* strains and percent frequency of occurrence.

The distribution of numts across chromosomes was very similar for each *T. gondii* strain (Figure 3.2A). The proportion of numts per chromosome is fairly constant except for chrIb and chrII, which harbor a slightly higher fraction of numts. However this difference cannot be explained by any observable variation in gene density between the chromosomes, as gene density is fairly equivalent on all chromosomes. We see much a higher proportion of numts (Figure 3.2B blue bars) clustering across the chromosomes when compared to gene distribution (Figure 3.2B orange bars).

A.



B.

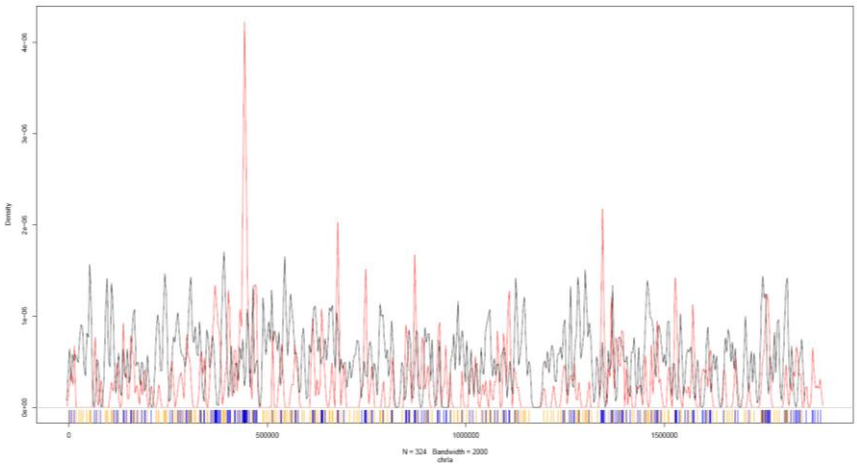


Figure 3.2 Distribution feature of numts on chromosomes.

A. Distribution of numts across all 14 chromosomes of the *T. gondii* strains. B. Density distribution of numts on chr1a ME49 strain. Blue bars represents numts, orange bars represents genes. Red peaks represent histogram of numt density and black peaks represent histogram of gene density.

Identification and analysis of strain-specific numts

In order to determine if any numts were unique to a single strain, first the numt sequence along with its flanking was retrieved and use as a query in the other two strains. The three-way chromosomal alignment available for the strains from ToxoDB was then used to confirm the absence or fragmentation of the numts at a specific chromosomal location within the other strains. This strategy yielded a total of 38 strain-

specific numts: 12 in ME49, 17 in GT1 and 9 in VEG (Table S3). In total, 31 of the 38 strain-specific numts were found within annotated genes, with 10 detected 2kb upstream or 1kb downstream of host genes and 20 within introns, suggesting that their presence might influence gene activity.

To ascertain if the strain-specific numts arose via either a novel insertion or deletion, we determined the precise boundaries by doing a three-way genome comparison of the chromosomal region in question. The numt was considered a novel insertion if it was precisely missing from two of the three strains and displayed high sequence identity (>98%) when compared to the mitochondrial DNA. If the numt insertion was only missing in one of the three strains and lacked precise boundaries then we concluded that this isolate-specific numt arose as a result of a deletion event. Using such criteria, four isolate-specific numts were confidently identified as novel insertions, and 17 were inferred to have occurred due to deletions in the other strains. An example of these two kinds of strain-specific numts is illustrated in Figure 3.3. Together these results points to numts contributing to strain diversification through differential deletions and insertions.

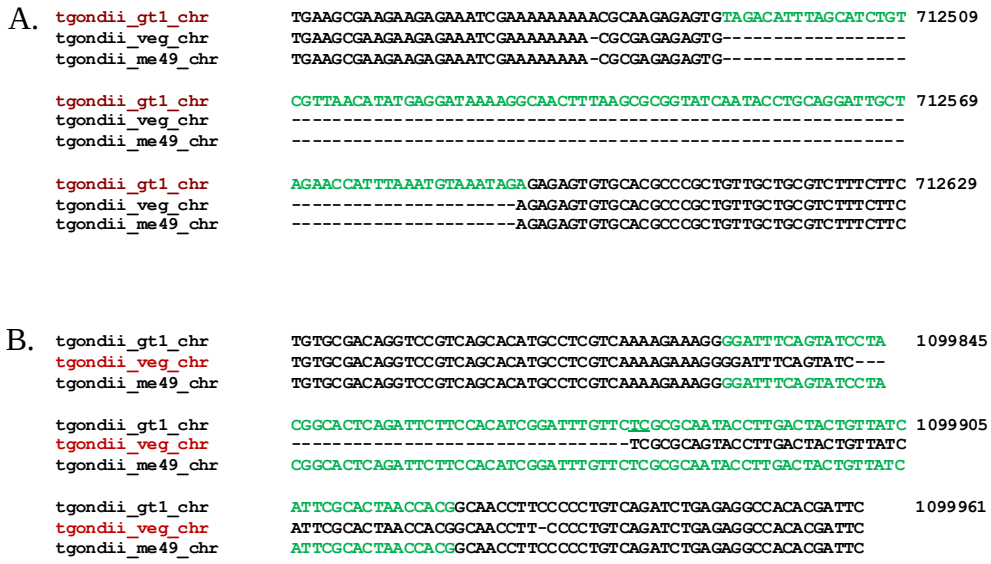


Figure 3.3 Examples of strain-specific numts.

Sequences in green indicate numts. A. Strain-specific numt arising from novel insertion in one strain highlighted in red. B. Strain-specific numt created by deletion.

Comparative analysis of numts between *T. gondii* and *Neospora caninum*

In an endeavor to understand some of the factors that contribute to the acquisition and retention of numts, we did a comparative analysis of *T. gondii* numts to the closely related species, *Neospora caninum*. To this end the genome of *N. caninum* was RepeatMasked with its mitochondrial genome. This analysis revealed that although this genome contains a large fraction of numts (0.78%), its content is less than half of that of *T. gondii*. We analyzed the size and chromosomal distribution of *Neospora* numts and found a similar profile as that of *T.gondii* (Figure S1).

To estimate the timing of insertion we employed a phylogenetically independent approach. As most numts are pseudogenized on arrival into the nucleus (LEISTER 2005) and therefore evolve at a neutral rate, the age of insertions can be calculated by

comparing the numt sequence divergence from the mtDNA and applying the neutral substitution rate of the species. Since the mtDNA is also subjected to mutation during its evolutionary timeframe, we also took into account the mutation rate of mtDNA when computing the age of numts. The mitochondrial mutation rate was calculated as described in Methods, and a mutation rate of 1.22×10^{-9} per site per year was obtained. This value is significantly lower than the reported nuclear neutral mutation rate of 2.12×10^{-8} (SU *et al.* 2003), suggesting that the *T. gondii* mtDNA evolves much slower than its nuclear counterpart, a pattern similarly noted for plants but different from animals (WOLFE *et al.* 1987; ZISCHLER *et al.* 1995). Percent divergences of each numt from mtDNA reported by RepeatMasker were converted to nucleotide distance measures using the Jukes-Cantor formula to correct for multiple hits. The age of each numt was then calculated by dividing the nucleotide distance by the sum of the neutral mutation rate and the mutation rate of mtDNA.

The age distribution profile of *T. gondii* and *N. caninum* varies drastically (Figure 3.4). For *T. gondii*, although a few insertions have occurred recently, the majority of the numt insertions took place around 12 million years ago (Mya). The age profile for *N. caninum* reveals a much younger distribution suggesting a more recent acquisition of the *numts* (between 3-6 Mya). Considering only the numts that accumulated before the split of these two genera (the split time indicated by arrow in figure 3 at ~12.7My), we observed that there are many more numts in *T. gondii* than in *N. caninum*. As the two genera share the same ancestor, and assuming that the neutral mutation rate is equivalent within the two generas, then the content of numts before

their split is expected to be the same. Therefore one possible explanation for the different amount of numts in these two genera may be that *T. gondii* has retained many more numts as compared to *N. caninum*. The higher retention rate of numts in *T. gondii* may to some extent explain why this organism contains the largest fraction of numts ever reported.

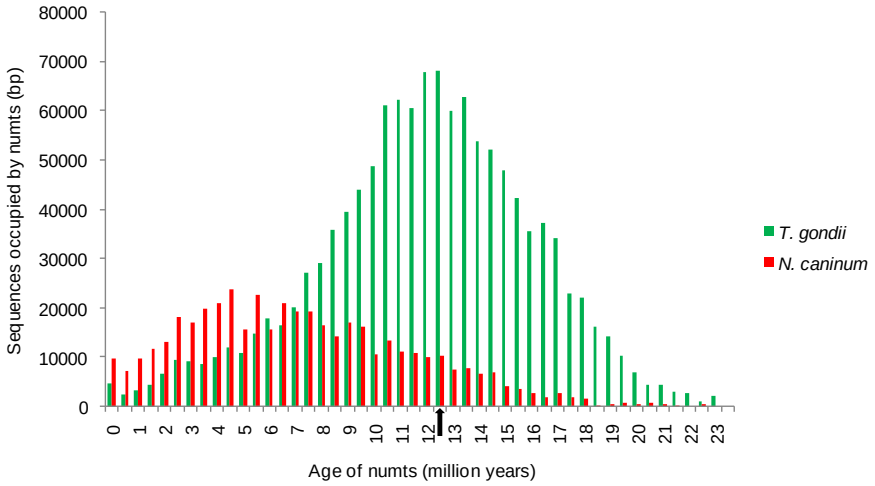


Figure 3.4 Age profile of *T. gondii* and *N.caninum* numts.

Age distribution profile of numts in *T. gondii* and *N. caninum*. Green bars: numts in *T. gondii* and red bars represents numts in *N. caninum*. Black arrow shows the split between *T. gondii* and *N. caninum* hypothesized to have occurred ~12.7Mya. Most of the numts are not shared by the two generas even prior to their divergence.

Insertion of numts in nuclear genes

To understand the insertion pattern of numts in the *T. gondii* genome, we took a closer look at the distribution of numts within and around genes. To do this, a library was generated consisting of *T. gondii_ME49* annotated genes along with their flanking regions and RepeatMasker was used to mask this library. Interestingly, we find that 54% of the numts were within genic regions, and the majority of these (5,000 numts)

were in introns (Figure 3.5). Overall we find 5580 numts located in genes and 1529 residing in 1 kb flanking regions. As there are about 7800 annotated genes in *T.gondii*, this rate translates to about one numt per gene within the *T. gondii* genome. This pattern is particularly high, surpassing that of the flowering plant *Arabidopsis thaliana*, where approximately 25% of nuclear organelle DNAs are located within genes (RICHLY and LEISTER 2004a) (RICHLY and LEISTER 2004b). However, this trend may just be reflective of the compact nature of the *T. gondii* genome, with ~ 64% coding capacity.

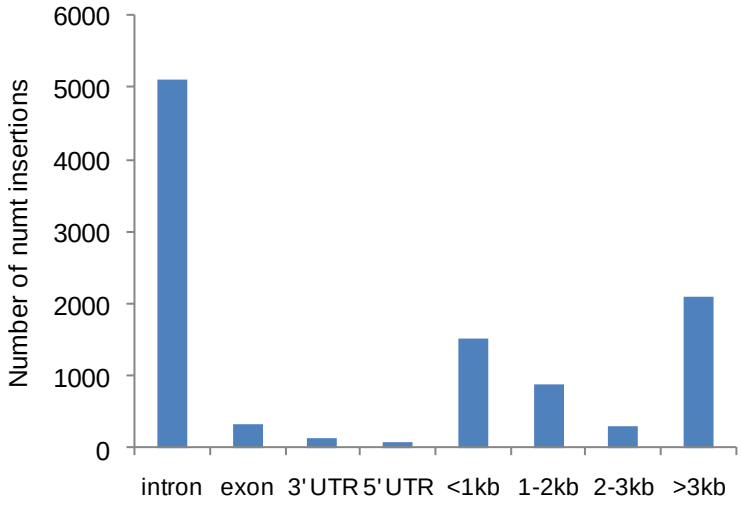


Figure 3.5 Distribution profiles of numts in or next to genes in *T. gondii.*

The number of numt insertions found within or next to genes in *T. gondii*. UTR= untranslated region. A large portion of numts found in introns, and the flanking regions of genes ranging from less than 1 kilo-base pairs (kb) to over 3 kb.

Identification of functional numts in *T. gondii*

Given the observed insertional bias of numts we attempt to gain further insights into their potential impact on genome evolution. We hypothesize that the density, lifespan and genic proximity of numts might foster their occasional functionalization.

Therefore to determine if these numts are functional, we first searched for orthologous numts between *T. gondii* and *N. caninum*. Since the divergence time between these two genomes occurred ~12.7 million years ago, shared orthologous numts is indicative of conservation due to a potential functional constraint. Using one of the Perl scripts described in the Methods, we isolated 12 putative shared orthologs. Among the 12 putative orthologous numts, 10 were discarded because they were found in unassembled contigs making it impossible for us to determine their specific chromosomal locus. The remaining two orthologs were derived from two discontinuous mtDNA regions and were found occurring in tandem on chromosome V both in *T. gondii* and *N. caninum*. These two numts dubbed OrA and OrB, resides 1 kb upstream an RNA metabolite-related Sm-like gene.

In order to determine if the orthologous numts show any signs of selective constraint, we extracted 2kb upstream sequences following the start codon (ATG) for each orthologous numt in *T. gondii* and *N. caninum*, along with their genic sequences, and aligned them using Vista plot (http://genome.lbl.gov/vista/index.shtml). The results (Figure 3.6) show that the level of sequence conservation of OrA and OrB is higher than that of most of the introns and of upstream sequences, consistent with functional constraint. Given their proximity to the Sm-like gene, these numts could possibly exert some regulatory control on the gene.

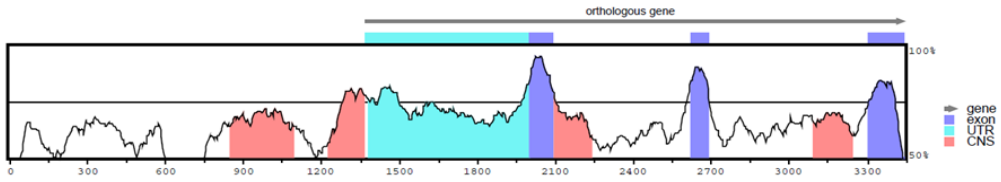


Figure 3.6 Vista plot of orthologous numt and gene present in *T. gondii* and *N.caninum.* The numt is found within 900 bp to 1200 bp and in close proximity to a gene. The level of conservation of the numt is very similar to the level of conservation of the CNS shown in pink.

To identify the number of numts with potential cis-regulatory effect on *T. gondii* genes, we further queried for numts inserted into less than 2 kb upstream from the start codon and less than 500 bp from the transcriptional start site (Table SX - not shown). Over 2000 numts were found indicating that a high portion of numts might be serving as cis-elements in the *T. gondii* genome; however, this needs further exploration.

Experimental validation of functional numts

The functionality of the numts identified upstream of the Sm-like gene and another numt found upstream of a myosin heavy chain gene were subjected to a functional assay to verify/identify their roles. PCR knockout constructs were generated and compared to the wild-type promoters. We made three different deletions for the promoter of Sm-like protein gene designated as ΔOrA, ΔOrB and ΔOrAB respectively (Figure 3.7A), and one mutant (Δnumt) for the myosin associated numt (Figure 3.7B). The WT and mutant promoters were then cloned upstream of a reporter gene and transiently expressed in *T. gondii* cells. The reporter assay results for Sm-like protein gene (Figure 3.7C) reveals that the deletion of either ΔOrA or ΔOrB significantly decreases promoter activity. This result suggests that both numts may contain cis-

elements capable of activating the reporter gene expression; however there may be a slight counteractive effect between the two numts, as deletion of both is slightly less disruptive of promoter activity. On the other hand, deletion of the numt upstream of the myosin gene dramatically increases promoter activity (Figure 3.7D). Taken together, these two examples provide evidence that some numts have contributed to the emergence of new cis-regulatory elements in the *Toxoplasma* lineage.

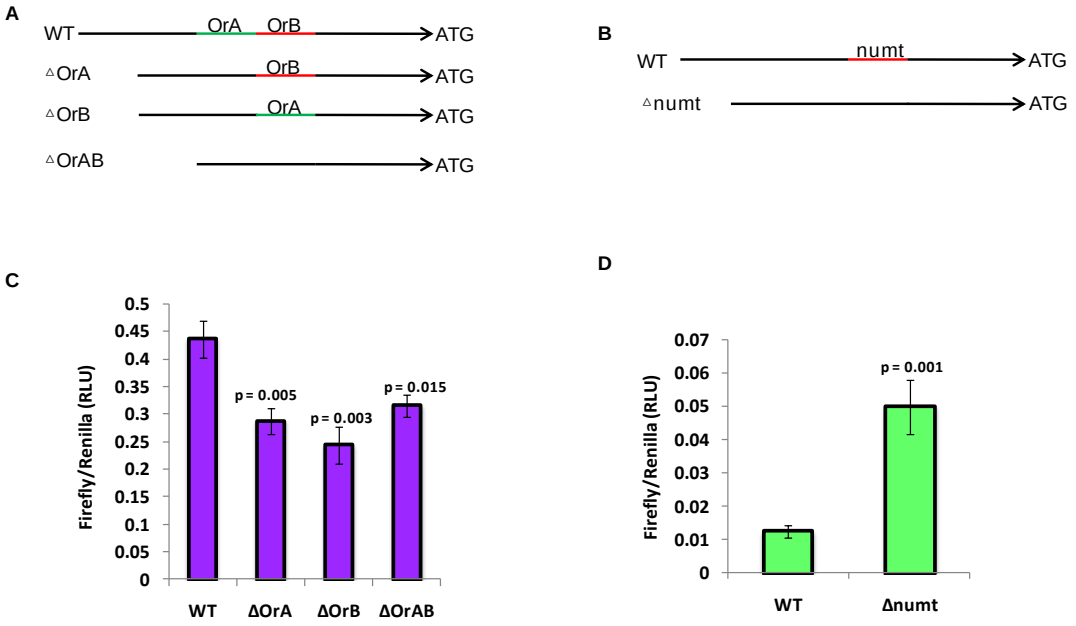


Figure 3.7 Experimental validations of functional numts.

A. The structure of wild type (WT) and mutant promoters for Sm-like gene. The numts are indicated by green and black lines. ΔOrA only the OrA numt is deleted, ΔOrB, the OrB numt is deleted, ΔOrAB, both numts are deleted. B The structure of the WT and mutant promoters for the myosin heavy chain gene. C Reporter assay for the Sm-like gene with p values shown above bars. D. Reporter assay for the myosin heavy chain gene.

The presence of numts in other apicomplexan species

To determine if numt accumulation is a common feature among apicomplexan species, we measured the total numt content for three additional species with available mitochondrial genome sequences. These include *Plasmodium falciparum*, *Babesia bovis* and *Theileria parva* (Table 3.2). The results revealed a dramatic discordance with numt content in these species with only five insertions identified in *P. falicparum* of which four most likely arose from post duplication events. The numts amounts to < 1 kb of mitochondrial DNA in *P. falciparum*. On the other hand, the *B. bovis* and *T. parva* had no recognizable mitochondrial insertions (Table 3.2). Conclusively, our results show that the propensity for numt accumulation differs significantly within the apicomplexan phylum.

Table 3.2 The amount mtDNA transferred to the nuclear genome of several apicomplexan species

Species	Genome size (Mbps)[1]	MtDNA genome size (kb)	Hits[2]	Base pair occupied[2]	Percentage of genome (%)
Plasmodium	22.9	6	5	515	0
Babesia	8.1	6.6	0	0	0
Theileria	8.3	7.5	0	0	0
Neospora	56.8	10	3,680	405,548	0.78
Toxoplasma (GT1)	60.8	7	9,827	1,139,048	1.88
Toxoplasma (ME49)	61.8	7	9,958	1,159,821	1.87
Toxoplasma (VEG)	62.2	7	10,074	1,163,635	1.87

[1] Genome size data obtained from the number of bases of genome used in RepeatMasker.
[2] Number of insertions and base pair composition based on RepeatMasker.

Repair pathway genes found in these species

Molecular and bioinformatic studies performed in yeast, tobacco and human have shown that integration of numts occurs at the DNA level during illegitimate repair of double-strand breaks (DSBR) by nonhomologous end-joining (NHEJ) (Blanchard and Schmidt 1996; Hazkani-Covo and Covo 2008; Henze and Martin 2001). However the NHEJ pathway, which is preferentially used in many eukaryotes for DSBR, is in contrast absent in many protozoan parasites (Fox et al. 2009a). For instance, *Plasmodium* species have been documented to lack identifiable key proteins like KU70, KU80, DNA ligase IV – Xrcc4 complex, proteins essential to NHEJ, and instead predominantly employ homologous recombination (Fox et al. 2009b). In order to evaluate the possible role of NHEJ in apicomplexan numt variation, we queried for NHEJ-related proteins in these genomes. Searches for the KU70, KU80, DNA ligase IV – Xrcc4 proteins in T. parva, B. bovis, and P. falciparum revealed no identifiable homologs (Table 3.3). However these proteins appear to be present in *T. gondii* and *N. caninum*, both of which harbor a higher proportion of numts. Although the role of these proteins has not been validated in *N. canium*, in *T. gondii* it has been experimentally demonstrated that NHEJ is preferentially used in DSBR at significantly high frequencies (Fox et al. 2009b). Consequently, these findings suggest that the proficiency of NHEJ pathway could possibly explain the numt density variation found between *T. gondii*, *P. falciparum*, *T.parva*, and *B. bovis*.

Table 3.3 Proteins involved in NHEJ pathway

Protein involved in non-homologous recombination		Organisms						
Human Homologs	Function/activity	TgME49	TgVEG	TgGT1	Nc	Bb	Tp	Pf
Ku70	DNA end binding	+	+	+	+	-	-	-
Ku86	DNA end binding	+	+	+	+	-	-	-
DNA ligase IV	Break joining	+	+	+	+	+	-	+
XRCC6		+	+	+	+	-	-	-

Note: Tg, *Toxoplasma gondii*; Bb, *Babesia bovis*; Tp, *Theileria parva*; Pf, *Plasmodium falciparum*; Nc, *Neospora caninum*.
Note: the proteins, function of proteins obtained from review by Aravind, L et al. 1999.
+ & - denotes presence absence for the genes in the apicomplexan species (Tg, Pf, Nc) based on blastp at EupathDB.org and tblastn for Tp & Bb at NCBI (http://blast.ncbi.nlm.nih.gov/Blast.cgi).

Discussion

Factors underlying differences in numt density within apicomplexan genomes

The *T. gondii* genome harbors the highest percentage of numts ever reported with numt density exceeding all previously documented cases in other eukaryotes. This difference ranges from a 7 fold increase when compared to the mustard plant *A. thaliana* (0.256% numt) and the fungus *Ustilago maydis* (0.286% numt) to 22 - 40 times more numt density as compared to the honeybee (0.086%), and the protist *Phytophthora infestans* (0.046% numt) (HAZKANI-COVO *et al.* 2010). In addition, the numt content found in the sister species *N. caninum* (0.78%) also surpasses previous cases, albeit to a much lesser extent. In this study we detected very few to no numts in the apicomplexans, *P. falciparum*, *T. parva* and *B. bovis*, suggesting that the high proportion of numts found in *T. gondii* and *N. caninum* is not a feature common to all apicomplexans.

Interestingly we have demonstrated that numt density even within closely related species varies significantly. This data is largely consistent with previous observation where a large variation in numt content has been described between insect species like *Drosophila melanogaster* (0.0057% numts), *Anopheles gambiae* (0 %), and *A. mellifera* (0.0861% numts), and even between mammals like human, mouse and rat (PAMILO *et al.* 2007; RICHLY and LEISTER 2004a). 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 (RICHLY and LEISTER 2004a). 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 per cell, this cannot sufficiently explain the observable differences in apicomplexans' numt accumulation (SEEBER *et al.* 2008).

Our analysis also shows that there are no detectable proteins associated with the non-homologous end-joining pathway (NHEJ) in the apicomplexans with little to no numts. This finding is significant in that it could potentially explain the differences we observed. Numts are thought to be integrated passively into the nuclear genome as fillers during repair of double strand breaks the NHEJ repair process (HAZKANI-COVO *et al.* 2010; KLEINE *et al.* 2009). If the pathway is not dominantly utilized by the host, then despite availability of mtDNA, very little to no integration into the nuclear genome can occur.

Furthermore, the comparative age distribution of numts in *T. gondii* and *N.caninum* suggests that a higher retention rate of mtDNA has occurred for *Toxoplasma* as compared to *Neospora*. However whether this is due to faster deletion of numts in *N. caninum* we cannot positively determine, as we know very little about the rate of DNA loss in these two generas.

Numts contribute to genome innovation in *T.gondii*

Here we have clearly demonstrated that more than half of the numts in *T. gondii* inserts in close proximity to genic regions. This finding is consistent with what is observed for human numts which are noted to preferentially insert within predicted genes (RICCHETTI *et al.* 1999). Our quest for orthologous numts between *T. gondii* and *N.caninum* yielded two unambiguous numts with higher levels of conservation when compared to other portions of their proximal genes suggesting to us that they might serve some functional role. We further investigated the role numts may have had on gene regulation by functional assays and show that numts are capable of dramatically affecting the promoter activity of the genes they are in close proximity to by acting most likely as cis-regulatory elements. To our knowledge this is the first study to experimentally establish that numts can drastically influence gene control. Among the targeted genes, one of intriguing interest is the myosin heavy chain ATpase, a gene essentially involved in parasite motility, division and penetration of host cells and parasitic virulence (MEISSNER *et al.* 2002). The numt inserted within this gene dramatically increases the gene promoter activity. This finding is exceptionally intriguing, as this may suggest that numts could not only influence gene activity but can also contribute to the pathogenicity of these parasites. In summary we have shown that numts have had very important roles in the evolution of the human parasite, *T. gondii* having successful invaded a large fraction of this parasites compact genome and possibly contributing cis-regulatory elements that can remodel nuclear genes. Future

characterizations of numts within this parasite will shed further insights as to the evolutionary consequences numts have had on apicomplexan evolution.

Materials and Methods

Retrieval of genome sequences

The nuclear genomic sequences of *Toxoplasma gondii* and *Neospora caninum* were downloaded from the ToxoDB (http://toxodb.org/toxo/) release 6.0. The Kissinger Lab at the University of Georgia provided mitochondrial genome data for *Toxoplasma gondii* and *Neospora caninum*.

Identification of numts

RepeatMasker (http://repeatmasker.org/) is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences. Sequences comparisons in RepeatMasker are performed by the program cross_match, an efficient implementation of the Smith-Waterman-Gotoh algorithm. Numts were identified using locally installed RepeatMasker version open-3.2.7. Mitochondrial genomes were used as the repeat library to mask the corresponding nuclear genomes with filtering out low complexity sequences and other parameters as default.

Identification of numts involved in segmental duplication

Based on the output of RepeatMasker, which contains pertinent nuclear positional information, the numt sequences along with 150bp of flanking regions were extracted from the *T. gondii* ME49 genome. These sequences were then subsequently used as BLAST queries back against *T. gondii* ME49. If one numt along with at least 100 bp of flanking regions was duplicated in the nuclear genome for at least twice, then

it was counted as being part of a segmental duplication. This process was automatically performed using a Perl script.

Analysis of chromosomal distribution of numts

Using R, freely available statistical software (http://www.r-project.org/), the numt density per chromosome was plotted for the *T. gondii ME49* strain. The start position of numt insertion was provided as input data in this computation, and the density was plotted in 2 kb to 10 kb windows taking the size of the chromosome into account. To assess relationship of numt distribution in comparison to genes, the gene start positions were also downloaded from ToxoDB and plotted in similar fashion as the numt density.

Identification of strain-specific numts

A Perl script was used to parse the output of RepeatMasker to identify the strain specific numts. In short, the sequences of numts from one of these three genomes, along with their 100 bp of flanking sequences were used as queries to BLAST against the two other genomes. Numts that appeared in one of the three genomes with gaps in the reciprocal locations in the two other genomes were identified. Candidate strain-specific numts were verified using Multiple Sequence Alignment (alignments made with Mercator) obtained from *Toxoplasma* genome browser (www.toxodb.org).

Calculation of mutation rate of mtDNA

First, Clustalw was used to align the mtDNA of *T. gondii* with that of *N. caninum*. Then, MEGA4.0 was used to calculate the pairwise genetic distance using the Jukes-Cantor model (Tamura, 2007). After such calculation, the genetic distance of

mtDNA between these two genera was 0.031. Finally, the mutation rate for *T. gondii* mtDNA was estimated by:

$$r = K/2T$$

where r is the rate of nucleotide substitution, K is the average genetic distance, and T is the divergence time between *T. gondii* and *N. caninum,* which was estimated to be 12.7 My (ADAMS and PALMER 2003).

Identification of functional numts

Using a Perl script, the sequences of *T.gondii* numts, along with their 100 bp of flanking sequences were extracted and used as queries to BLAST against the genomic sequences of *N. caninum*. Only homologous hits that cover both the numts and more than 50 bp of flanking sequences were taken as candidate orthologous. Orthologous numts were confirmed by evaluating the presence of the numts in all three strains of *T.* gondii and *N. caninum* using the whole genome alignments available at ToxoDB (www.toxodb.org). In order to identify possible functional numts based on their proximity to genes, scripts were written in Perl to extract the upstream sequence of all the annotated *T.gondii ME49* genes. RepeatMasker was then used to find numts in the 2 kb upstream regions of these sequences. (All the scripts used in this study are available on request).

Molecular techniques

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 each gene tested and a two-step overlap-extension PCR technique (SAMBROOK *et al.* 1989) was

employed to delete the numt sequence from each tested promoter. The Gateway™ cloning system (Invitrogen) was used to clone the WT and mutant promoters. These two kinds of promoters were first cloned into pDONR221 via the BP reaction. Following sequencing verification, promoter fragments were moved into a firefly luciferase-expressing vector (destination vectors) via the LR reaction. As an internal control, a constitutive promoter (*T.gondii* α-tubulin promoter)-driven renilla luciferase-expressing construct (α-tub-renilla) was co-transfected along with the experimental construct. Nucleotide positions in these deletion studies are referenced to with respect to the start of translation (+1).

Parasite culture and transient transfections

Parasite culture and transient transfections were performed as described (Mullapudi *et al*. 2009). In short, *T. gondii* RH tachyzoites were cultured in human foreskin fibroblasts. Transient transfection was performed via electroporation, using freshly lysed parasites, needle-passaged and filtered through a 3-micron filter and resuspended in cytomix. Post-electroporation, the parasites were allowed to rest for 15 minutes in the cuvette and then transferred to T25 tissue culture flasks. Then, 18-24 hours post-electroporation, the cells were scraped and lysed using passive lysis buffer (Promega, Madison, WI, USA) and a dual luciferase assay was performed with the extract using the Promega Dual Luciferase kit. Briefly, the different substrate requirements for each enzyme, firefly luciferase 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

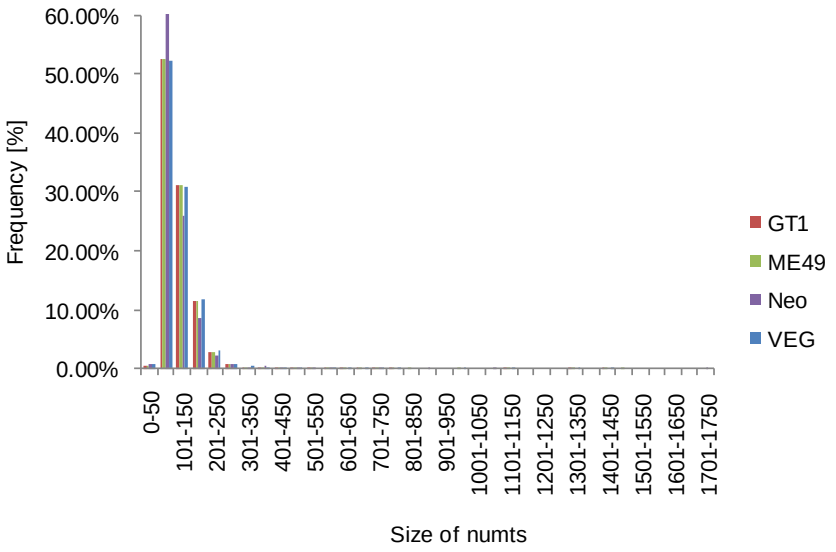
internal control, eliminating errors due to variation in parasite populations and individual transfections. Enzyme activity was measured using a dual luciferase-ready luminometer. Each electroporation experiment was performed in triplicate and luciferase assays were performed in duplicates for expression 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.

APPENDIX A

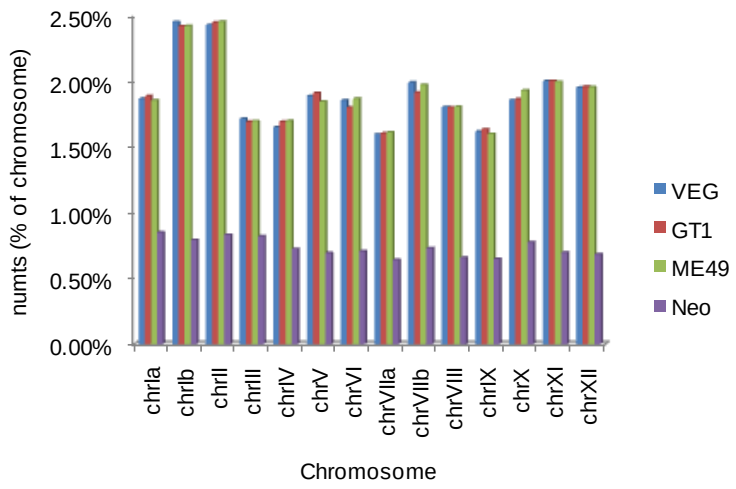
SUPPORTING INFORMATION - CHAPTER 3

Supplemental Figure(s)

A.



B.



Supplemental Figure 1 The distribution feature of *N. caninum* numts.

A. The size distribution of numts in *Neospora* as compared to the three strains of *T. gondii*. B. The numt distribution across all fourteen chromosomes of *Neospora* compared to *T. gondii* strains.

Supplemental Table(s)

Supplemental Table 1 Segmentally duplicated numts

duplicated_numts_ME49.xlsx

Table S1: Segmentally duplicated numts. The name of one numt indicates its chromosome position and its copy number in the nuclear genome. The sequence presented here includes the numt sequence along with its 150 bp of flanking regions.

Supplemental Table 2 Mitochondrial origin of numts in *T. gondii* strains

Tgondii_numts_final_output.xls

Supplemental Table 3 Summary of strain-specific numts and their distributional relationship to genes

Genomic position	Age	Score	Associated gene	Position in genes
TGME49_chrVIII:2494796..2495019	19.1	593	TGME49_033200	1.1kb upstream
TGME49_chrXII:807354..807681	21.2	748	TGME49_019590	1.6kb upstream
TGGT1_chrII:1024425..1024537	27.4	292	TGGT1_066220	1.8 kb upstream
TGGT1_chrXII:374756..374897	22.7	506	TGGT1_095780	1104 upstream
TGME49_chrVIIa:3214064..3214225	30.6	411	TGME49_002540	1298bp upstream
TGME49_chrIV:1080445..1080536	20.6	416	TGME49_118770	1347bp upstream
TGGT1_chrVIIb:1029868..1029942	14.7	412	TGGT1_007200	173bp downstream
TGME49_chrX:2951331..2951684	16.9	610	TGME49_024510	257 upstream
TGME49_chrX:422129..422179	23.5	268	TGME49_028230	3’ UTR
TGME49_chrXI:3542124..3542606	19.4	309	TGME49_113620	400bp upstream
TGGT1_chrX:2643367..2643428	17.7	350	TGGT1_079810	591bp upstream

TGGT1_chrII:167249..167341	4.4	640	TGGT1_065180	intron
TGME49_chrIII:1725825..1725973	23.5	574	TGME49_054420	intron
TGME49_chrIII:278871..278971	23.2	349	TGME49_075690	intron
TGME49_chrIX:3139296..3139559	12.5	1749	TGME49_089180	intron
TGGT1_chrVI:712492..712592	1	849	TGGT1_051390	intron
TGGT1_chrVIII:3465971..3466241	24.5	1050	TGGT1_112560	intron
TGME49_chrVIII:4960100..4960137	0	309	TGME49_070840	intron
TGGT1_chrVIIa:2770053..2770171	7.6	899	TGGT1_017730	intron
TGGT1_chrVIIb: 4631514..4631600	21.2	376	TGGT1_014330	intron
TGGT1_chrX:2918359..2918516	24.1	626	TGGT1_079280	intron
TGME49_chrX:6339113..6339208	21.3	289	TGME49_014600	intron
TGGT1_chrX:6789740..6789798	25.4	238	TGGT1_126000	intron
TGGT1_chrXII:2476569..2476655	25.3	292	TGGT1_027460	intron
TGME49_chrIX:5347444..5347902	25	286	TGME49_105150	intron
TGME49_chrVIIa:3162998..3163174	22.5	351	TGME49_002580	intron
TGME49_chrVIIb:694301..694369	24.6	315	TGME49_063220	intron
TGME49_chrVIIb:1702125..1702268	25	448	TGME49_061490	intron
TGME49_chrVIIb:4781476..4781610	28.8	340	TGME49_055400	intron
TGME49_chrXI:1301594..1301658	24.6	269	TGME49_110380	intron
TGME49_chrXII:3846605..3846677	0	621	TGME49_048450	intron
TGME49_chrIX:3015887..3015948	3.2	476	no	no
TGME49_chrVIII:4434179..4434256	20.8	367	no	no
TGGT1_chrVIIb:3221065..3221245	23.2	716	no	no
TGGT1_chrVIIb:3671637..3671700	12.7	364	no	no
TGGT1_chrXI:4632126..4632228	25.2	472	no	no
TGGT1_chrXII:2455829..2455929	21.7	310	no	no
TGME49_chrX:5672042..5675411	1.7	10000	no	no

REFERENCES

ABRAHAMSEN, M. S., T. J. TEMPLETON, S. ENOMOTO, J. E. ABRAHANTE, G. ZHU *et al.*, 2004 Complete genome sequence of the apicomplexan, Cryptosporidium parvum. Science **304:** 441-445.

ADAMS, K. L., and J. D. PALMER, 2003 Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. Mol Phylogenet Evol **29:** 380-395.

AJIOKA, J. W., BROOKE-POWELL, E.T., WAN, K, 2005 The nuclear genome of apicomplexan parasites, pp. in *Encyclopedia of Genetics, Genomics, Proteomics and Bioinformatics*.

AJIOKA, J. W., J. M. FITZPATRICK and C. P. RIETTER, 2001 Ultrastructure of a *Toxoplasma gondi*i tachyzoite, pp. in *Expert Reviews in Molecular Medicine*, edited by FIG001JAC. Cambridge University Press Cambridge.

ANDERSSON, J. O., 2005 Lateral gene transfer in eukaryotes. Cellular and Molecular Life Sciences **62:** 1182-1197.

ANDERSSON, S. G., O. KARLBERG, B. CANBACK and C. G. KURLAND, 2003 On the origin of mitochondria: a genomics perspective. Philos Trans R Soc Lond B Biol Sci **358:** 165-177; discussion 177-169.

ARKHIPOVA, I., and M. MESELSON, 2000 Transposable elements in sexual and ancient asexual taxa. Proc Natl Acad Sci U S A **97:** 14473-14477.

ARKHIPOVA, I., and M. MESELSON, 2005 Deleterious transposable elements and the extinction of asexuals. Bioessays **27:** 76-85.

AURRECOECHEA, C., J. BRESTELLI, B. P. BRUNK, S. FISCHER, B. GAJRIA *et al.*, 2010 EuPathDB: a portal to eukaryotic pathogen databases. Nucleic Acids Res **38:** D415-419.

BAKER, J. R., 1994 The origins of parasitism in the protists. Int J Parasitol **24:** 1131-1137.

BARTA, J. R., 1989 Phylogenetic analysis of the class Sporozoea (phylum Apicomplexa Levine, 1970): evidence for the independent evolution of heteroxenous life cycles. Journal of Parasitology **75:** 195-206.

BEHURA, S. K., 2007 Analysis of nuclear copies of mitochondrial sequences in honeybee (Apis mellifera) genome. Mol Biol Evol **24:** 1492-1505.

BELANGER, F., F. DEROUIN, L. GRANGEOT-KEROS and L. MEYER, 1999 Incidence and risk factors of toxoplasmosis in a cohort of human immunodeficiency virus-infected patients: 1988-1995. HEMOCO and SEROCO Study Groups. Clin Infect Dis **28:** 575-581.

BENSASSON, D., M. W. FELDMAN and D. A. PETROV, 2003 Rates of DNA duplication and mitochondrial DNA insertion in the human genome. J Mol Evol **57:** 343-354.

BENSON, G., 1999 Tandem repeats finder: a program to analyze DNA sequences. Nucleic Acids Res **27:** 573-580.

BIÉMONT, C., and C. VIEIRA, 2006 Genetics: Junk DNA as an evolutionary force. October **443:** 521-524.

BLANCHARD, J. L., and G. W. SCHMIDT, 1996 Mitochondrial DNA migration events in yeast and humans: integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns. Mol Biol Evol **13:** 893.

BOHNE, A., F. BRUNET, D. GALIANA-ARNOUX, C. SCHULTHEIS and J. N. VOLFF, 2008 Transposable elements as drivers of genomic and biological diversity in vertebrates. Chromosome Res **16:** 203-215.

BOORE, J. L., 1999 Animal mitochondrial genomes. Nucleic Acids Research **27:** 1767-1780.

BOWEN, N. J., and I. K. JORDAN, 2002 Transposable elements and the evolution of eukaryotic complexity. Current issues in molecular biology **4:** 65-76.

BRAYTON, K. A., A. O. LAU, D. R. HERNDON, L. HANNICK, L. S. KAPPMEYER *et al.*, 2007 Genome sequence of Babesia bovis and comparative analysis of apicomplexan hemoprotozoa. PLoS Pathog **3:** 1401-1413.

BRINGAUD, F., E. GHEDIN, N. M. EL-SAYED and B. PAPADOPOULOU, 2008 Role of transposable elements in trypanosomatids. Microbes Infect **10:** 575-581.

BURMA, S., B. P. CHEN and D. J. CHEN, 2006 Role of non-homologous end joining (NHEJ) in maintaining genomic integrity. DNA Repair (Amst) **5:** 1042-1048.

CARLTON, J., J. SILVA and N. HALL, 2005 The genome of model malaria parasites, and comparative genomics. Curr Issues Mol Biol **7:** 23-37.

CARLTON, J. M., J. H. ADAMS, J. C. SILVA, S. L. BIDWELL, H. LORENZI *et al.*, 2008 Comparative genomics of the neglected human malaria parasite Plasmodium vivax. Nature **455:** 757-763.

CARLTON, J. M. R., R. MULLER, C. A. YOWELL, M. R. FLUEGGE, K. A. STURROCK *et al.*, 2001 Profiling the malaria genome: a gene survey of three species of malaria parasite with comparison to other apicomplexan species. Molecular and Biochemical Parasitology **118:** 201-210.

CASADEVALL, A., 2008 Evolution of intracellular pathogens. Annu Rev Microbiol **62:** 19-33.

CAVALIER-SMITH, T., 1999 Principles of protein and lipid targeting in secondary symbiogenesis: Euglenoid, dinoflagellate, and sporozoan plastid origins and the eukaryote family tree. Journal of Eukaryotic Microbiology **46:** 347-366.

CAVALIER-SMITH, T., 2005 Economy, Speed and Size Matter: Evolutionary Forces Driving Nuclear Genome Miniaturization and Expansion. Annals of Botany**:** 147-175.

CRAIG, N., R. CRAIGIE, M. GELLERT and A. LAMBOWITZ (Editors), 2002 *Mobile DNA II*. American Society for Microbiology Press, Washington (DC).

DE SOUZA, W., M. ATTIAS and J. C. RODRIGUES, 2009 Particularities of mitochondrial structure in parasitic protists (Apicomplexa and Kinetoplastida). Int J Biochem Cell Biol **41:** 2069-2080.

DOLGIN, E. S., and B. CHARLESWORTH, 2006 The fate of transposable elements in asexual populations. Genetics **174:** 817-827.

DURAND, P. M., A. J. OELOFSE and T. L. COETZER, 2006 An analysis of mobile genetic elements in three Plasmodium species and their potential impact on the nucleotide composition of the P. falciparum genome. Bmc Genomics **7:** 282.

FAST, N. M., J. C. KISSINGER, D. S. ROOS and P. J. KEELING, 2001 Nuclear-encoded, plastid-targeted genes suggest a single common origin for apicomplexan and dinoflagellate plastids. Molecular Biology and Evolution **18:** 418-426.

FAST, N. M., L. XUE, S. BINGHAM and P. J. KEELING, 2002 Re-examining alveolate evolution using multiple protein molecular phylogenies. J Eukaryot Microbiol **49:** 30-37.

FESCHOTTE, C., 2008 Transposable elements and the evolution of regulatory networks. Nat Rev Genet **9:** 397-405.

FESCHOTTE, C., U. KESWANI, N. RANGANATHA, M. L. GUIBOTSY and D. LEVIN, 2009 Exploring repetitive DNA landscapes using REPCLASS, a tool that automates the classification of transposable elements in eukaryotic genomes. Genome Biology and Evolution Advance Access.

FESCHOTTE, C., and E. J. PRITHAM, 2007a Computational Analysis and Paleogenomics of Interspersed Repeats in Eukaryotes, pp. in *Computational Genomics: Current Methods*, edited by N. STOJANOVIC. Horizon Scientific Press.

FESCHOTTE, C., and E. J. PRITHAM, 2007b DNA transposons and the evolution of eukaryotic genomes. Annu Rev Genet **41:** 331-368.

FIGUEIREDO, L. M., L. A. PIRRIT and A. SCHERF, 2000 Genomic organisation and chromatin structure of Plasmodium falciparum chromosome ends. Mol Biochem Parasitol **106:** 169-174.

FISHER, N., P. G. BRAY, S. A. WARD and G. A. BIAGINI, 2008 Malaria-parasite mitochondrial dehydrogenases as drug targets: too early to write the obituary. Trends Parasitol **24:** 9-10.

FOX, B. A., J. G. RISTUCCIA and D. J. BZIK, 2009a Genetic identification of essential indels and domains in carbamoyl phosphate synthetase II of Toxoplasma gondii. Int J Parasitol **39:** 533-539.

FOX, B. A., J. G. RISTUCCIA, J. P. GIGLEY and D. J. BZIK, 2009b Efficient gene replacements in Toxoplasma gondii strains deficient for nonhomologous end joining. Eukaryot Cell **8:** 520-529.

FRENAL, K., and D. SOLDATI-FAVRE, 2009 Role of the parasite and host cytoskeleton in apicomplexa parasitism. Cell Host Microbe **5:** 602-611.

FUNES, S., A. REYES-PRIETO, X. PEREZ-MARTINEZ and D. GONZALEZ-HALPHEN, 2004 On the evolutionary origins of apicoplasts: revisiting the rhodophyte vs. chlorophyte controversy. Microbes and Infection **6:** 305-311.

GAJADHAR, A. A., W. C. MARQUARDT, R. HALL, J. GUNDERSON, E. V. ARIZTIA-CARMONA *et al.*, 1991 Ribosomal RNA sequences of Sarcocystis muris, Theileria annulata and Crypthecodinium cohnii reveal evolutionary relationships among apicomplexans, dinoflagellates, and ciliates. Mol Biochem Parasitol **45:** 147-154.

GARDNER, M. J., R. BISHOP, T. SHAH, E. P. DE VILLIERS, J. M. CARLTON *et al.*, 2005 Genome sequence of Theileria parva, a bovine pathogen that transforms lymphocytes. Science **309:** 134-137.

GARDNER, M. J., N. HALL, E. FUNG, O. WHITE, M. BERRIMAN *et al.*, 2002 Genome sequence of the human malaria parasite Plasmodium falciparum. Nature **419:** 498-511.

GILBERT, R. E., L. GRAS, M. WALLON, F. PEYRON, A. E. ADES *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:** 1303-1308.

GRAY, M. W., G. BURGER and B. F. LANG, 1999 Mitochondrial evolution. Science **283:** 1476-1481.

GRAY, M. W., G. BURGER and B. F. LANG, 2001 The origin and early evolution of mitochondria. Genome Biol **2:** REVIEWS1018.

GREGORY, T. R., 2005 Synergy between sequence and size in large-scale genomics. Nat Rev Genet **6:** 699-708.

HAAS, B. J., S. KAMOUN, M. C. ZODY, R. H. JIANG, R. E. HANDSAKER *et al.*, 2009 Genome sequence and analysis of the Irish potato famine pathogen Phytophthora infestans. Nature **461:** 393-398.

HALL, N., M. KARRAS, J. D. RAINE, J. M. CARLTON, T. W. KOOIJ *et al.*, 2005 A comprehensive survey of the Plasmodium life cycle by genomic, transcriptomic, and proteomic analyses. Science **307:** 82-86.

HAZKANI-COVO, E., and S. COVO, 2008 Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. PLoS Genet **4:** e1000237.

HAZKANI-COVO, E., R. M. ZELLER and W. MARTIN, 2010 Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes. PLoS Genet **6:** e1000834.

HE, C. Y., M. K. SHAW, C. H. PLETCHER, B. STRIEPEN, L. G. TILNEY *et al.*, 2001 A plastid segregation defect in the protozoan parasite Toxoplasma gondii. Embo Journal **20:** 330-339.

HENZE, K., and W. MARTIN, 2001 How do mitochondrial genes get into the nucleus? Trends Genet **17:** 383-387.

HICKEY, D. A., 1982 Selfish DNA: a sexually-transmitted nuclear parasite. Genetics **101:** 519-531.

HIKOSAKA, K., Y. I. WATANABE, N. TSUJI, K. KITA, H. KISHINE *et al.*, 2009 Divergence of the mitochondrial genome structure in the apicomplexan parasites, Babesia and Theileria. Mol Biol Evol.

HUANG, J., and J. C. KISSINGER, 2006 Horizontal and intracellular gene transfer in the Apicomplexa: The scope and functional consequences, pp. 123-136 in *Genomics and Evolution of Microbial Eukaryotes*, edited by L. A. KATZ and D. BHATTACHARYA. Oxford University Press, New York.

HUANG, J., N. MULLAPUDI, C. A. LANCTO, M. SCOTT, M. S. ABRAHAMSEN *et al.*, 2004a Phylogenomic evidence supports past endosymbiosis, intracellular and horizontal gene transfer in Cryptosporidium parvum. Genome Biol **5:** R88.

HUANG, J. L., N. MULLAPUDI, T. SICHERITZ-PONTEN and J. C. KISSINGER, 2004b A first glimpse into the pattern and scale of gene transfer in the Apicomplexa. International Journal for Parasitology **34:** 265-274.

KATINKA, M. D., S. DUPRAT, E. CORNILLOT, G. METENIER, F. THOMARAT *et al.*, 2001 Genome sequence and gene compaction of the eukaryote parasite Encephalitozoon cuniculi. Nature **414:** 450-453.

KEELING, P. J., 2004 Reduction and compaction in the genome of the apicomplexan parasite Cryptosporidium parvum. Dev Cell **6:** 614-616.

KEELING, P. J., 2009 Chromalveolates and the Evolution of Plastids by Secondary Endosymbiosis. Journal of Eukaryotic Microbiology **56:** 1-8.

KEELING, P. J., and C. H. SLAMOVITS, 2004 Simplicity and complexity of microsporidian genomes. Eukaryot Cell **3:** 1363-1369.

KHAN, A., U. BOHME, K. A. KELLY, E. ADLEM, K. BROOKS *et al.*, 2006 Common inheritance of chromosome Ia associated with clonal expansion of Toxoplasma gondii. Genome Research **16:** 1119-1125.

KHAN, A., S. TAYLOR, C. SU, D. L. SIBLEY, I. PAULSEN *et al.*, 2007 Genetics and Genome Organization of *Toxoplasma gondii* in *Toxoplasma, Molecular and Cellular Biology*, edited by J. W. AJIOKA and D. SOLDATI. Horizon Bioscience, Norfolk, UK.

KIDWELL, M. G., and D. R. LISCH, 2001 Perspective: Transposable elements, parasitic DNA, and genome evolution. Evolution **55:** 1-24.

KIM, K., and L. M. WEISS, 2004 Toxoplasma gondii: the model apicomplexan. Int J Parasitol **34:** 423-432.

KING, C. C., 1992 Modular transposition and the dynamical structure of eukaryote regulatory evolution. Genetica **86:** 127-142.

KLEINE, T., U. G. MAIER and D. LEISTER, 2009 DNA Transfer from Organelles to the Nucleus: The Idiosyncratic Genetics of Endosymbiosis. Annual Review of Plant Biology **60:** 115-138.

KURLAND, C. G., and S. G. ANDERSSON, 2000 Origin and evolution of the mitochondrial proteome. Microbiol Mol Biol Rev **64:** 786-820.

LANDER, E. S., L. M. LINTON, B. BIRREN, C. NUSBAUM, M. C. ZODY *et al.*, 2001 Initial sequencing and analysis of the human genome. Nature **409:** 860-921.

LANG, B. F., M. W. GRAY and G. BURGER, 1999 Mitochondrial genome evolution and the origin of eukaryotes. Annu Rev Genet **33:** 351-397.

LAU, A. O. T., 2009 An overview of the Babesia, Plasmodium and Theileria genomes: A comparative perspective. Molecular & Biochemical Parasitology **164:** 1-8.

LEANDER, B. S., R. CLOPTON and P. KEELING, 2003 Phylogeny of gregarines (Apicomplexa) as inferred from small-subunit rDNA and beta-tubulin. Int J Syst Evol Microbiol **53:** 345-354.

LEANDER, B. S., and P. J. KEELING, 2003 Morphostasis in alveolate evolution. Trends in Ecology & Evolution **18:** 395-402.

LEANDER, B. S., and P. J. KEELING, 2004 Early evolutionary history of dinoflagellates and apicomplexans (Alveolata) as inferred from hsp90 and actin phylogenies. Journal of Phycology **40:** 341-350.

LEISTER, D., 2005 Origin, evolution and genetic effects of nuclear insertions of organelle DNA. Trends in Genetics **21:** 655-663.
LEVINE, N. D., 1988 Progress in taxonomy of the Apicomplexan protozoa. J Protozool **35:** 518-520.
LI, L., B. P. BRUNK, J. C. KISSINGER, D. PAPE, K. L. TANG *et al.*, 2003 Gene discovery in the Apicomplexa as revealed by EST sequencing and assembly of a comparative gene database. Genome Research **13:** 443-454.
LIZUNDIA, R., D. WERLING, G. LANGSLEY and S. A. RALPH, 2009 Theileria Apicoplast as a Target for Chemotherapy. Antimicrobial Agents and Chemotherapy **53:** 1213-1217.
LOFTUS, B., I. ANDERSON, R. DAVIES, U. C. ALSMARK, J. SAMUELSON *et al.*, 2005 The genome of the protist parasite Entamoeba histolytica. Nature **433:** 865-868.
LYNCH, M., and J. S. CONERY, 2003 The origins of genome complexity. Science **302:** 1401-1404.
MATHER, M. W., and A. B. VAIDYA, 2008 Mitochondria in malaria and related parasites: ancient, diverse and streamlined. J Bioenerg Biomembr **40:** 425-433.
MCFADDEN, G. I., and D. S. ROOS, 1999 Apicomplexan plastids as drug targets. Trends in Microbiology **7:** 328-333.
MEISSNER, M., D. SCHLUTER and D. SOLDATI, 2002 Role of Toxoplasma gondii myosin A in powering parasite gliding and host cell invasion. Science **298:** 837-840.
MOORE, R. B., M. OBORNIK, J. JANOUSKOVEC, T. CHRUDIMSKY, M. VANCOVA *et al.*, 2008 A photosynthetic alveolate closely related to apicomplexan parasites. Nature **451:** 959-963.
MORRISON, D. A., 2009 Evolution of the Apicomplexa: where are we now? Trends in Parasitology **25:** 375-382.
MORRISSETTE, N. S., and L. D. SIBLEY, 2002 Cytoskeleton of apicomplexan parasites. Microbiol Mol Biol Rev **66:** 21-38.
MOURIER, T., A. J. HANSEN, E. WILLERSLEV and P. ARCTANDER, 2001 The Human Genome Project reveals a continuous transfer of large mitochondrial fragments to the nucleus. Mol Biol Evol **18:** 1833-1837.
NOUTSOS, C., T. KLEINE, U. ARMBRUSTER, G. DALCORSO and D. LEISTER, 2007 Nuclear insertions of organellar DNA can create novel patches of functional exon sequences. Trends in Genetics **23:** 597-601.
NOZAKI, H., H. TAKANO, O. MISUMI, K. TERASAWA, M. MATSUZAKI *et al.*, 2007 A 100%-complete sequence reveals unusually simple genomic features in the hot-spring red alga Cyanidioschyzon merolae. Bmc Biology **5**.
OBORNIK, M., J. JANOUSKOVEC, T. CHRUDIMSKY and J. LUKES, 2009 Evolution of the apicoplast and its hosts: From heterotrophy to autotrophy and back again. International Journal for Parasitology **39:** 1-12.
OKAMOTO, N., and G. I. MCFADDEN, 2008 The mother of all parasites. Future Microbiology **3:** 391-395.
OLIVER, K. R., and W. K. GREENE, 2009 Transposable elements: powerful facilitators of evolution. Bioessays **31:** 703-714.

OSSORIO, P. N., L. D. SIBLEY and J. C. BOOTHROYD, 1991 Mitochondrial-like DNA sequences flanked by direct and inverted repeats in the nuclear genome of Toxoplasma gondii. J Mol Biol **222:** 525-536.

PAIN, A., L. CROSSMAN and J. PARKHILL, 2005 Comparative apicomplexan genomics. Nat Rev Microbiol **3:** 454-455.

PAMILO, P., L. VILJAKAINEN and A. VIHAVAINEN, 2007 Exceptionally high density of NUMTs in the honeybee genome. Mol Biol Evol **24:** 1340-1346.

PLATTNER, F., and D. SOLDATI-FAVRE, 2008 Hijacking of host cellular functions by the Apicomplexa. Annu Rev Microbiol **62:** 471-487.

PRICE, A. L., N. C. JONES and P. A. PEVZNER, 2005 De novo identification of repeat families in large genomes. Bioinformatics **21 Suppl 1:** i351-358.

PRITHAM, E. J., 2009 Transposable elements and factors influencing their success in eukaryotes. J Hered **100:** 648-655.

PRITHAM, E. J., C. FESCHOTTE and S. R. WESSLER, 2005 Unexpected diversity and differential success of DNA transposons in four species of entamoeba protozoans. Mol Biol Evol **22:** 1751-1763.

RICCHETTI, M., C. FAIRHEAD and B. DUJON, 1999 Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. Nature **402:** 96-100.

RICHLY, E., and D. LEISTER, 2004a NUMTs in sequenced eukaryotic genomes. Mol Biol Evol **21:** 1081-1084.

RICHLY, E., and D. LEISTER, 2004b NUPTs in sequenced eukaryotes and their genomic organization in relation to NUMTs. Mol Biol Evol **21:** 1972-1980.

ROOS, D. S., M. J. CRAWFORD, R. G. K. DONALD, J. C. KISSINGER, L. J. KLIMCZAK *et al.*, 1999 Origin, targeting, and function of the apicomplexan plastid. Current Opinion in Microbiology **2:** 426-432.

ROY, S. W., and D. PENNY, 2007 Widespread intron loss suggests retrotransposon activity in ancient apicomplexans. Mol Biol Evol **24:** 1926-1933.

SAMBROOK, J., E. F. FRITSCH and T. MANIATIS, 1989 *Molecular cloning: a laboratory manual*. Cold Spring Harbor Laboratory Press, New York

SANMIGUEL, P., A. TIKHONOV, Y. K. JIN, N. MOTCHOULSKAIA, D. ZAKHAROV *et al.*, 1996 Nested retrotransposons in the intergenic regions of the maize genome. Science **274:** 765-768.

SCHUSTER, W., and A. BRENNICKE, 1994 The Plant Mitochondrial Genome: Physical Structure, Information Content, RNA Editing, and Gene Migration to the Nucleus. Annual Review of Plant Physiology and Plant Molecular Biology **45:** 61-78.

SEEBER, F., J. LIMENITAKIS and D. SOLDATI-FAVRE, 2008 Apicomplexan mitochondrial metabolism: a story of gains, losses and retentions. Trends Parasitol **24:** 468-478.

SU, C., D. EVANS, R. H. COLE, J. C. KISSINGER, J. W. AJIOKA *et al.*, 2003 Recent expansion of Toxoplasma through enhanced oral transmission. Science **299:** 414-416.

TAYLOR, S., A. KHAN, C. SU and D. L. SIBLEY, 2007 Pathogencity and Virulence in *Toxoplasma gondii* in *Toxoplasma Molecular and Cellular Biology*, edited by J. W. AJIOKA and D. SOLDATI. Horizon Bioscience, Norfolk, UK.

TEMPLETON, T. J., S. ENOMOTO, W. J. CHEN, C. G. HUANG, C. A. LANCTO *et al.*, 2010 A Genome-Sequence Survey for Ascogregarina taiwanensis Supports Evolutionary Affiliation but Metabolic Diversity between a Gregarine and Cryptosporidium. Mol Biol Evol **27:** 235-248.

TOMLEY, F., 2009 Apicomplexan biology in the post-genomic era: perspectives from the European COST Action 857. Int J Parasitol **39:** 133-134.

TOSO, M. A., and C. K. OMOTO, 2007 Gregarina niphandrodes may lack both a plastid genome and organelle. Journal of Eukaryotic Microbiology **54:** 66-72.

VICIENT, C. M., A. SUONIEMI, K. ANAMTHAWAT-JONSSON, J. TANSKANEN, A. BEHARAV *et al.*, 1999 Retrotransposon BARE-1 and Its Role in Genome Evolution in the Genus Hordeum. Plant Cell **11:** 1769-1784.

VOLFF, J. N., 2006 Turning junk into gold: domestication of transposable elements and the creation of new genes in eukaryotes. Bioessays **28:** 913-922.

WAKAGURI, H., Y. SUZUKI, M. SASAKI, S. SUGANO and J. WATANABE, 2009 Inconsistencies of genome annotations in apicomplexan parasites revealed by 5'-end-one-pass and full-length sequences of oligo-capped cDNAs. BMC Genomics **10:** 312.

WALLER, R. F., and C. J. JACKSON, 2009 Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. Bioessays **31:** 237-245.

WICKER, T., F. SABOT, A. HUA-VAN, J. L. BENNETZEN, P. CAPY *et al.*, 2007 A unified classification system for eukaryotic transposable elements. Nat Rev Genet **8:** 973-982.

WICKSTEAD, B., K. ERSFELD and K. GULL, 2003 Repetitive elements in genomes of parasitic protozoa. Microbiol Mol Biol Rev **67:** 360-375.

WILSON, R. J. M., and D. H. WILLIAMSON, 1997 Extrachromosomal DNA in the apicomplexa. Microbiology and Molecular Biology Reviews **61:** 1-&.

WOLFE, K. H., W. H. LI and P. M. SHARP, 1987 Rates of nucleotide substitution vary greatly among plant mitochondrial, chloroplast, and nuclear DNAs. Proc Natl Acad Sci U S A **84:** 9054-9058.

WOOTTON, J. C., and S. FEDERHEN, 1996 Analysis of compositionally biased regions in sequence databases. Methods Enzymol **266:** 554-571.

XU, P., G. WIDMER, Y. WANG, L. S. OZAKI, J. M. ALVES *et al.*, 2004 The genome of Cryptosporidium hominis. Nature **431:** 1107-1112.

ZHU, G., M. J. MARCHEWKA and J. S. KEITHLY, 2000 Cryptosporidium parvum appears to lack a plastid genome. Microbiology-Uk **146:** 315-321.

ZISCHLER, H., H. GEISERT, A. VON HAESELER and S. PAABO, 1995 A nuclear 'fossil' of the mitochondrial D-loop and the origin of modern humans. Nature **378:** 489-492.

BIOGRAPHICAL INFORMATION

Assiatu B. Barrie was born in Pita, Guinea to Alhaji Ibrahim Bah and Haja Salamata Bah. She immigrated to the United States of America in April 1999. She graduated from Trinity High School, and went ahead to pursue her Bachelor's of Science at the University of Texas, Arlington. As a first generation graduate, she completed her Bachelor's with honors, with the aim of becoming a physician. Her exposure to research at UTA during her undergraduate years motivated her to pursue a Masters degree in Genomics, studying transposable elements in unicellular parasites. Throughout her graduate career, she has been a recipient of several awards, including the Outstanding Graduate Research Award from the Department of Biology, and the William L, and Martha Hughes Biology Award. She has also demonstrated leadership qualities by serving as secretary for the Graduate Student Senate and the Phi Sigma Biological Sciences Honor Society. She was conferred Master's of Science in Biology as of May 2010. Her future plans include getting into Medical School and returning to her roots to help make a difference.