Massive invasion of organellar DNA drives nuclear genome evolution in *Toxoplasma*

Sivaranjani Namasivayam[a,1], Cheng Sun[b,1] Assiatu B Bah[c], Jenna Oberstaller[d], Edwin Pierre-Louis[e], Ronald Drew Etheridge[e], Cedric Feschotte[f], Ellen J. Pritham[c,2] and Jessica C. Kissinger[g,2]

[a]Department of Genetics, University of Georgia, Athens, GA 30602, USA; Present address: Clinical Microbiome Unit, Laboratory of Host Immunity and Microbiome, NIAID, NIH, Bethesda, MD 20892, USA

[b]Department of Biology, University of Texas at Arlington, Arlington, TX 76019, USA; Present address: College of Life Sciences, Capital Normal University, Beijing 100048, China

[c]Department of Biology, University of Texas at Arlington, Arlington, TX 76019

[d]Department of Genetics, University of Georgia, Athens, GA 30602, USA; Present address: Department of Global Health, University of South Florida, Tampa, FL 33620, USA

[e]Department of Cellular Biology, Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, GA 30602, USA

[f]Department of Biology, University of Texas at Arlington, Arlington, TX 76019; Present address: Department of Molecular Biology and Genetics, Cornell University, Ithaca, NY 14853-2703, USA

[g]Department of Genetics, Institute of Bioinformatics, and Center for Tropical and Emerging Global Diseases, University of Georgia, Athens, GA 30602, USA

[1]S.N. and C.S contributed equally to this work
[2]To whom correspondence should be addressed

Email: jkissing@uga.edu

Author Contributions: EJP, JCK and CF designed and supervised the research; SN, CS, AB, JO, EPL and RDE performed the research; SN, CS, EJP, JCK, and RLP analyzed data; SN, CS, CF, JCK and EJP wrote the paper.

Competing Interest Statement: The authors declare no competing interests.

Keywords: Nuclear integrants of Mitochondrial DNA - NUMTs, Nuclear integrants of Plastid DNA - NUPTs, Nuclear DNA of organellar origin - NUOT, Apicomplexa, Coccidia, Non-Homologous End-Joining repair - NHEJ

This PDF file includes:

Main Text
Figures 1 to 5
Tables 1 to 2

Abstract

Toxoplasma gondii is a zoonotic protist pathogen that infects up to 1/3 of the human population. This apicomplexan parasite contains three genome sequences: nuclear (63 Mb); plastid organellar, ptDNA (35 kb); and mitochondrial organellar, mtDNA (5.9 kb of non-repetitive sequence). We find that the nuclear genome contains a significant amount of NUMTs (nuclear DNA of mitochondrial origin) and NUPTs (nuclear DNA of plastid origin) that are continuously acquired and represent a significant source of intraspecific genetic variation. NUOT (nuclear DNA of organellar origin) accretion has generated 1.6% of the extant *T. gondii* ME49 nuclear genome; the highest fraction ever reported in any organism. NUOTs are primarily found in organisms that retain the non-homologous end-joining repair pathway. Significant movement of organellar DNA was experimentally captured via amplicon sequencing of a CRISPR-induced double-strand break in non-homologous end-joining repair competent, but not *ku80* mutant, *Toxoplasma* parasites. Comparisons with *Neospora caninum*, a species that diverged from *Toxoplasma* ~28 MY ago, revealed that the movement and fixation of 5 NUMTs predates the split of the two genera. This unexpected level of NUMT conservation suggests evolutionary constraint for cellular function. Most NUMT insertions reside within (60%) or nearby genes (23% within 1.5 kb) and reporter assays indicate that some NUMTs have the ability to function as cis-regulatory elements modulating gene expression. Together these findings portray a role for organellar sequence insertion in dynamically shaping the genomic architecture and likely contributing to adaptation and phenotypic changes in this important human pathogen.

Significance Statement

This study reveals how DNA located in cellular compartments called organelles can be transferred to the nucleus of the cell and inserted into the nuclear genome of apicomplexan parasite *Toxoplasma*. Insertions alter the DNA sequence and may lead to significant changes in how genes function. Unexpectedly, we found that the human protist pathogen, *Toxoplasma gondii* and closely-related species have the largest observed organellar genome fragment content (>11,000 insertion comprising over 1 Mb of DNA) inserted into their nuclear genome sequence despite their compact 65 Mb nuclear genome. Insertions are occurring at a rate that makes them a significant mutational force that deserves further investigation when examining causes of adaptation and virulence of these parasites.

Introduction

Toxoplasma gondii is a cosmopolitan apicomplexan protist parasite (Fig. 1A), capable of infecting nearly all warm-blooded animals including humans (1, 2). *T. gondii* causes toxoplasmosis, and the parasite infects as much as one third of the world's human population with individual countries showing infection rates ranging from 10 to 80% (3). While the parasite rarely causes symptoms in healthy adults, it can lead to serious and even life-threatening illness in immunosuppressed or pregnant individuals (4, 5). Phylogenetic and population studies of *T. gondii* reveal an unusual population structure consisting of mostly clonal lineages that occasionally recombine (6-9) as well as the recent evolution of oral infectivity that permits the parasite to spread asexually via consumption of tissue cyst forms in addition to traditional transmission of sexually derived oocysts in the environment via a fecal/oral route (10). While it is clear that *Toxoplasma* and other coccidian parasites evolve and adapt rapidly, little is known about the source(s) of genomic variation driving adaptation in *Toxoplasma*. The ability of *Toxoplasma* to propagate asexually, combined with the observed low levels of sexual recombination in existing data and the lack of detectable mobile elements create real challenges for *Toxoplasma* with respect to sources of genomic variation. When contemplating genome evolution and the differences between strains and species it is common practice to examine single nucleotide polymorphisms (SNPs), insertion/deletion events (indels), rearrangements, horizontal and intracellular gene transfers and transposable element (TE) insertion or movement. The insertion of random fragments of organellar DNA is not high on the list of features examined as a major cause of phenotypic or genotypic diversity, but if the organism is a tissue coccidian like *Toxoplasma*, perhaps it should be.

In eukaryotes, DNA of endosymbiotic organelles (mitochondria and chloroplasts) has been observed to be transferred to the nuclear genome (11, 12). During the early phase of organelle evolution, this process resulted in a massive relocation of entire organellar genes to the nuclear genome (13). In many eukaryotes, the transfer of functional organellar genes appears rare or has ceased altogether (14, 15), and almost all recent transfers of mitochondrial (mtDNA) or plastid (ptDNA) DNA to the nuclear genome are gene or genome fragments that give rise to noncoding sequences, called NUMTs (nuclear integrants of mtDNA) and NUPTs (nuclear integrants of ptDNA). Organellar-to-nuclear DNA transfer has been reported for 85 fully sequenced eukaryotic genomes and fully characterized in 23 mammals (16). The results are indicative of the significant driving force NUOTs (nuclear DNA of organellar origin) provide for gene and genome innovation in eukaryotes (11) including humans (17) and other primates (18, 19). In *Saccharomyces cerevisiae*, 16 NUMTs were unexpectedly recovered in a double-strand break experiment (20) and the non-homologous end-joining repair pathway, NHEJ, was shown to be the mechanism responsible for NUMT integration (20, 21). Mechanistically, reactive oxygen species (ROS) can generate double-stranded mtDNA breaks (DSBs) and mtDNA fragments (22) and there are hypotheses and some evidence for how organellar DNA, in particular mtDNA may enter the host nucleus and be used to patch nuclear DSBs (23).

Nuclear genome sequences in the Apicomplexa are highly streamlined, ranging from 8 to 130 Mb, dynamic in structure and rapidly evolving (24, 25). Transposable elements (TEs), which generally constitute the most significant proportion of repetitive DNA in genomes and are known to be powerful agents for generating genomic plasticity (26-28), are absent from most apicomplexans examined thus far (29) excepting the genera *Ascogregarina*, *Eimeria* and *Sarcocystis* (30, 31). Consequently, it is of interest to understand what factors contribute to the generation of genomic plasticity in apicomplexan parasites. While organellar-to-nuclear DNA transfer represents a significant driving force for genome innovation in eukaryotes, it has not been examined in apicomplexans (outside of nuclear encoded, plastid or mitochondrial targeted proteins) (32, 33), and little is known about the evolutionary fate and consequence of these transferred DNA sequences.

Most apicomplexans contain two organelles, a mitochondrion and a plastid (pt) called the apicoplast that was acquired through the secondary endosymbiosis of a red alga (33-36). The ptDNA sequence (35 kb) is a circular/cruciform structure and is well conserved across apicomplexans (37). The mtDNA sequences generated thus far are variable and occur as linear monomers (~6 kb) or concatemers with, or without, inverted repeats depending on the species (38, 39). Despite the genome topological differences, the apicomplexan mtDNA consistently encodes three cytochrome genes: cytochrome oxidase subunits I and III (*coxI, coxIII)* and cytochrome b (*cob*) and highly fragmented large and small subunit rRNA genes (LSU – SSUrRNA) (38, 39). Unlike the ptDNA sequence, the mtDNA sequence of *T. gondii* remained elusive until recently (38). It is structurally divergent and consists of 21 unique sequence blocks (SBs) that represent highly fragmented protein coding and rRNA genes (38). These sequence blocks can be highly redundant, as was observed in dinoflagellates (40, 41). The presence of NUMTs in *Toxoplasma* was first reported more than three decades ago as REP elements (42), but these genetic elements have not been studied since. Thanks to the availability of nearly complete genome sequences for numerous *Toxoplasma* strains and organellar DNA sequences, we performed a systematic examination of NUOTs in *T. gondii* and its relatives to shed light on their evolutionary dynamics and impact on genome innovation.

Results

Exceptional density of NUMTs and NUPTs in *T. gondii*

To identify insertions of organellar DNA fragments in the nuclear genome of *T. gondii* ME49 we applied the homology-based program Repeat Masker with organellar genomes as libraries. Only the 14 nuclear chromosomal assemblies and not all available contigs were queried for organellar insertions to obtain reliable and conservative estimates. The 21 conserved mtDNA sequence blocks (38) as well as full-length sequences of the mitochondrial protein-coding genes, *coxI*, *coxIII* and *cob* were included in the repeat library to minimize overestimation of the number of

NUMTs. Additionally, the 35 kb apicoplast genome sequence of *T. gondii* was utilized as the repeat library to identify NUPTs. These analyses revealed that 1.43% (895,312 bp) and 0.18% (112,433 bp) of the nuclear genome of *T. gondii* is composed of NUMTs and NUPTs respectively (Table 1, Fig. 1B, Dataset S1). To our knowledge, this percentage of nuclear genome occupied by NUOTs in *T. gondii* represents the largest ever reported for any eukaryote (Table 1). Indeed, the NUMT insertions in *T. gondii* are almost 6 times higher than the current overall record holder, the pathogenic fungus *Ustilago maydis* (0.286%) (43) (Table 1).

Most (>90%) of the NUOTs identified in *T. gondii* are 50 bp to 200 bp in length and display less than 90% sequence identity to their respective organellar DNA sequences indicating most insertions have accumulated mutations after nuclear acquisition (Fig. 1C). The largest NUMT is a single 3,369 bp sequence on (Chr 10) that does not contain any complete coding sequences but retains 99% nucleotide identity to mtDNA sequence blocks suggesting it is a recent insertion (Table S1). NUMTs derive from all 21 mtDNA sequence blocks including coding and non-coding regions (Table S2). However, there are some differences in their frequency of occurrence with NUMTs derived from sequence blocks M, E, V and K are the most enriched each comprising at least 0.1 % of the nuclear genome (Table S2). Similarly, NUPTs arise from nearly all regions of the apicoplast genome, with the ~2400 bp region between coordinates 9600 to 12000 of the apicoplast genome being the least represented region in NUPTs (Dataset S1).

We next examined the location and distribution of NUOTs in the *T. gondii* nuclear genome. Approximately 60% of the identified NUOTs map to genic regions (introns and exons) with the vast majority located in introns and the rest almost exclusively in UTRs (Table 2). The other 40% of NUOTs are located in intergenic regions, and ~23% occur within the 1 kb flanking region of genes. Thus, ~85% of *T. gondii* NUOTs reside within or near genes. This distribution is not unexpected given the highly compact nature of the *T. gondii* genome where genes occupy ~70% of the nuclear genomic space .

To determine if NUOTs arose due to independent insertion events or were segmentally duplicated post-insertion, we examined the flanking regions of NUOTs as duplication would likely involve parts of the flanking nuclear sequence along with the NUOTs. We considered NUOTs to be the product of a nuclear segmental duplication event if the NUOT along with 100 bp of its flanking regions was detected more than once in the nuclear genome. Using this criterion, 105 NUMTs and 13 NUPTs were identified as segmentally duplicated two to ten times in the ME49 nuclear genome sequence (Table S3). Thus, less than 1% of the NUOTs were duplicated post-insertion suggesting that most insertions were independent insertion events of purely organellar DNA. Together, these findings demonstrate that the large proportion of NUOTs observed in the nuclear genome of *T. gondii* arose primarily through independent insertion events that are generally <200 bp in length and display varying levels of sequence degeneration relative to their organellar DNA counterparts.

Strain-specific NUMTs indicate recent insertions and frequent turnover

To determine the age distribution of the NUOTs, age-related decay was determined using $T = K/2\mu$ where μ is the intron mutation rate for *T. gondii* (10) and K, the genetic distance, was determined with the Jukes-Cantor model based on the difference between the NUOT and organellar sequences (see Methods). This analysis revealed a continuous influx of NUOT assimilations with an accumulation peak approximately 7 million years ago (Fig. 1D). While NUOT assimilation appears to be a continual process, most insertions occurred prior to the last 1 million years. Nevertheless, the presence of the NUOTs that are 100% identical to the organellar DNA indicate that the process is ongoing.

To identify recent insertions or deletions we compared the mtDNA insertions in the nuclear genome of different *T. gondii* strains. We restricted this investigation of strain-specific insertions and deletions to NUMTs, since the number of NUMTs is comparatively much higher than NUPTs. We began this analysis by comparing the NUMTs identified in *T. gondii* ME49 to those identified in strains, GT1 and VEG, both of which have chromosomal level nuclear genome assemblies. We identified a total of 32 strain-specific NUMTs: 12 in ME49, 13 in GT1 and 7 in VEG (Table S4). A few of these strain-specific NUMTs were experimentally verified via PCR (Fig. S1, Table S5). 26 of the 32 strain-specific NUMTs were found in or near annotated genes, with 16

detected in introns, 1 in 3′ UTR, and 9 within 2 kb of annotated genic regions, suggesting that their presence may have the potential to affect host gene activity (Table S4).

To ascertain if the strain-specific NUMTs arose via either a novel insertion or deletion event, we determined the precise sequence boundaries and performed a three-way genome comparison of the chromosomal regions in question. A NUMT was considered a novel insertion if it was precisely missing from two of the three strains and displayed high sequence identity (>95%) when compared to the mtDNA sequence (Fig. 2A). If a NUMT was missing in only one of the three strains and the NUMT in the 2 strains displayed high sequence divergence (>10%) when compared to the mtDNA, then we concluded that this NUMT was most likely deleted in the third strain (Fig. 2B). Using such criteria, 3 NUMTs were confidently identified as novel insertions, and 13 were inferred to be deletions and many of these are partial deletions (Table S4). These results suggest that NUMTs are dynamic and can exist as novel insertions and can be fully or partially deleted leading to strain differences in *Toxoplasma*.

To further explore the differential presence or absence of NUMTs in *T. gondii* we extended this analysis to 13 additional strains with assembled, but less complete genome sequences. All of the strains displayed a similar NUMT content (Table S6). Overall, we identified 8,271 'ancestral' NUMTs that are present in all 16 strains and 56 NUMTs that show differential presence or absence among these strains (Fig. 2C, Table S7). These differentially present NUMTs are in varying states of decay, ranging from 0%-31% divergence from the mtDNA sequence. No strain shows a preferential loss or gain/retention of NUMTs. MAS has the largest number of differential NUMTs at 36 out of 56 and VEG the fewest (25/56) (Table S7).

The sites of insertion or deletion for the 56 differentially present NUMTs are mostly conserved among the strains. Thus, it is likely that these NUMTs were inserted in an ancestor and strains which evolved from this ancestor inherited the NUMT. For example, NUMTs 18 and 42 (100% identical to the mtDNA) are present only in strains MAS and TgCATBr5 and these strains are clustered together as Clade B (Fig. 2C) (7). NUMTs also appear to be inherited via recombination with a strain that does not contain a given NUMT. For example, NUMTs 13, 35 and 37 that display 100% identity to the mtDNA are each present in only two strains (p89 and VAND, TgCtCo5 and VAND and GAB2-2007 and ME49 respectively) suggesting inheritance of the NUMT via recombination, or, less likely, independent insertion of the same NUMT at the same location in two different strains. These pairs of strains are not grouped in the same clade and other clade members do not contain these NUMTs (Fig. 2C). Genome analysis of *T. gondii* has demonstrated that the parasites undergo sexual recombination at a higher frequency than previously thought (7). Our finding that the presence or absence of NUMTs does not follow clade or local ancestry (based on available data) is consistent with this observation.

Finally, to examine potential functional consequences of the 56 differentially present NUMTs across the 16 *T. gondii* strains, we identified their exact insertion location with respect to genes. 8 are located in exons and 31 are located in introns. 11 NUMTs are located in gene flanking regions or UTR, the remaining 6 are located in intergenic regions outside of the 1 kb upstream and downstream flanking regions (Table S7). None are located in a protein coding region. We conducted a gene-ontology (GO) enrichment analysis with the 50 genes whose exons, introns or flanking regions contained NUMTs (Table S8). Genes involved in microtubule-based processes and movement (10 genes) and nucleotide metabolic or biosynthetic processes (10 genes) are enriched 5-20 fold (Table S8). These pathways and processes are core survival and virulence pathways as *T. gondii* acquires purines from its host via salvage rather than *de novo* synthesis (44). It is possible that NUMTs in these regions may affect transcription at these loci. However, transcriptome data are only available for 3 of the 16 strains analyzed here and only for tachyzoites, thus these findings warrant future exploration.

NUOTs are drivers of genome evolution in coccidian parasites

To determine if the high NUOT density is a phenomenon specific to *T. gondii*, we investigated the nuclear genome sequences of its closest coccidian relatives *Hammondia hammondi* and *Neospora caninum* (Fig. 1A). *H. hammondi* contains NUMT and NUPT densities of 1.48% and 0.19% respectively, a level higher than that observed in *T. gondii*. However, because the *H. hammondi* genome sequence assembly is highly fragmented and not assembled into chromosomes, the NUOT densities are only estimates. The NUOT content in *N. caninum's*

nuclear genome is substantially lower than *T. gondii* at 0.66% for NUMTs and 0.06% for NUPTs, however, these densities are still higher than those observed in other eukaryotes (Fig. 3A, Table 1). By contrast, the distantly related coccidian species, *Sarcocystis neurona* and *Eimeria tenella* display greatly reduced NUMT densities at 0.03 and 0.04% respectively, relative to *T. gondii* and *N. caninum* but their NUPT densities, especially in the case of *E. tenella*, are only slightly reduced relative to *T. gondii* (Table 1). Examination of the apicomplexan nuclear genome sequences of species outside of the Coccidia revealed a paucity of NUOTs. *Plasmodium falciparum* 3D7 contains only five NUMTs whereas we were unable to detect any such insertions in *Babesia bovis* and *Theliera parva* (Table 1). Thus, an unusually high NUOT density is a feature of coccidian parasites and is more pronounced in species closely related to *T. gondii*.

Next, we examined the characteristics of the NUOTs in *N. caninum* and *H. hammondi*. Similar to *T. gondii*, we found that most NUMTs are 50-200 bp in length in both species (Fig. 3B, Fig. S2A). Greater than 90% of the NUOTs display less than 80% identity to the mt/ptDNA in *H. hammondi* as is seen in *T. gondii*. Interestingly, the NUOTs in *N. caninum* appear younger (Fig. 3C, Fig. S2B) suggesting either more recent insertions, a different mutation rate or a more rapid turnover rate in this species. NUPT's in *H. hammondi*, unlike the other coccidian species show two distinct age populations, one quite recent (Fig. S2B) but this may be an artifact related to contaminating apicoplast sequences in this highly fragmented assembly. Taken together, these results reveal that the species most closely related to *T. gondii*, also have a very high NUOT density but that as you move to more distant coccidian species, NUOTs decrease and in apicomplexans outside of the coccidia, NOUTs are mostly non-existent.

Since NUMTs display an overall rapid turnover and decay based on comparisons between *T. gondii* and *N. caninum* and within *T. gondii* strains, we hypothesized that evolutionary conserved NUMTs might be under functional constraint, indicating cooption for cellular function. To identify such coopted candidates, we compared the NUMTs between *T. gondii* and *N. caninum*, two species that diverged ~28 MYA (45), and identified five orthologous NUMTs (Table S9). The orthologous NUMT residing in the upstream region of the U6 snRNA-associated Sm family protein encoding gene (TGME49_286560) was the most conserved NUMT associated with an annotated gene and thus we selected this NUMT, named NUMT[Sm]. NUMT[Sm] is 28.8% divergent from mtDNA SB U which encodes portions of LSU and SSU rRNA fragments. Next we selected a younger *T. gondii* specific NUMT, called NUMT[Myo] located in the upstream region of a putative myosin heavy chain gene (TGME49_254850) of all examined *T. gondii* strains. NUMT[Myo] is 16.7% divergent from mtDNA SB F which encodes RNA 10 (Fig. S3)(38). The upstream regions of both genes show evidence of active transcription in *T. gondii* based on H3K9ac acetylation and H3K4 tri-methylation marks (ChIP data from ToxoDB.org)(46)(Fig. S3). The upstream regions of both *T. gondii* genes were cloned and used for functional characterization. To test the promoter activity of these upstream regions, we measured their ability to drive luciferase gene expression in transient *in vitro* reporter assays conducted in *T. gondii* cultured parasites (see Methods). For each promoter we generated luciferase reporter constructs where the upstream region was wild-type (WT) or without the identified NUMT (Δ NUMT) (Fig. 4A, B). Transient transfection assays were performed to compare the luciferase expression levels between the WT and Δ NUMT promoter constructs in *T. gondii*. Deletion of the 70 bp NUMT[Sm] significantly decreased promoter activity whereas deletion of the 66 bp NUMT[Myo] dramatically increased its activity (Fig. 4C, D). These results suggest that NUMTs can carry, or can evolve, into cis-regulatory elements capable of modulating gene expression. Taken together, these data demonstrate that NUMTs undergo rapid turnover and that evolutionarily conserved insertions may have been retained due to functional constraints for regulatory function.

NUOT acquisition involves non-homologous end joining repair

It has been demonstrated that organellar DNA fragments can become integrated into the nuclear genome during the process of double-strand break (DSB) repair via a mechanism called non-homologous end joining (NHEJ) (20, 21). NHEJ can repair DSBs without a template or sequence homology but it often results in small deletions and in some cases insertion of short DNA sequences (47). This mechanism of integration of DNA fragments, usually NUOTs but also transposable elements was first demonstrated in *Saccharomyces cerevisiae* and subsequently

proven in mammalian cells and plants (20, 21, 48, 49). In *T. gondii*, it has been observed that DNA is readily integrated into the nuclear genome at random sites via NHEJ (50). Indeed, it was necessary to knock-out the gene encoding a critical NHEJ protein, Ku80, in order to increase the efficiency of homologous recombination in this species (50). Interestingly, apicomplexans that lack the NHEJ machinery also contain very few NUOTs (Fig. 5A, Table 1), suggesting that NHEJ is a prominent mechanism underlying NUOT integration in coccidian species.

To experimentally test if NHEJ drives NUOT integration we used the CRISPR-Cas9 system to generate targeted DSBs in the genome of a *T. gondii* RH wild-type (WT) strain as well as a *ku80* knock-out mutant strain, Δ*ku80* (51, 52). We chose the *uprt* locus as the target for the DSB since disruption of this gene leads to 5-fluorodeoxyribose (FUDR) resistance thus facilitating efficient selection for improper repair of a CRISPR-Cas9 guide-directed cut (51). Resistant parasites were recovered, DNA was extracted, the *uprt* locus was amplified and the amplicons were deep-sequenced using Illumina paired-end sequencing. The paired-end reads were filtered for quality and merged when possible, *i.e.* the insert was less than ~250 bp (Table S10). Processed reads were then analyzed to compare the number of insertions between WT and Δ*ku80* parasites and specifically for insertions of extra-chromosomal DNA, namely, mtDNA, ptDNA and CRISPR-Cas9 plasmid DNA. It should be noted that nuclear DNA was also observed to be inserted at the site of DSB but was not further investigated in this study.

Overall, we observed that the number of unique reads obtained from WT parasites was seven times higher than that obtained from the Δ*ku80* strain (Table S10). Less than 1% of the reads from the Δ *ku80* parasites contained inserts of extrachromosomal DNA, whereas such insertions were observed in ~58% of the amplicons from WT *T. gondii* (Fig. 5B). Analysis of the source of DNA inserted at the DSB site revealed the transfected CRISPR-Cas9 plasmid as by far the most common source of inserted extra-chromosomal DNA both in terms of number of reads with plasmid DNA inserts as well as number of base pairs inserted (Fig. 5B, C). We expected the majority of the inserts to be from plasmid DNA since it is an abundant source of DNA readily available for repair following transfection. However, both mtDNA and ptDNA sequences were recovered as well, but surprisingly ptDNA was slightly more common than mtDNA. This finding is in contrast to our observation of higher levels of NUMTs versus NUPTs in the nuclear chromosomes of tissue coccidia (Table 1). Importantly. DNA capture was drastically reduced in the *ku80* mutant strain with no organellar DNA recovered at the site of DSB repair in this mutant strain (Fig 5B,C).

We next compared the length of the inserts. Since only 150 bp from each end of the amplicon was sequenced, this analysis is limited to detecting a maximum insert length of ~250 bp. We observed that a large proportion of inserts from mtDNA were shorter (~110 bp) than those from ptDNA and plasmid DNA with the latter being the longest (Fig. 5 D, E). Inserts as long as 250 bp are detected from all three DNA sources. We investigated the kind of modifications observed at the DSB. As typically seen in NHEJ mediated repair we detected insertion or deletion of a few base pairs at the repair site. We also observed simple insertions of one continuous stretch of DNA and a number of complex insertions where non-contiguous stretches of DNA from the same or multiple DNA sources were observed (Fig. 5F, Table S11). The insertions derived from all parts of the mt, pt or plasmid DNA and are of varying lengths (Table S11). Taken together these experiments and data indicate that DSB repair via NHEJ is likely the mechanism underlying NUOT integration in the *T. gondii* genome and that the properties of the insertions observed in an artificially induced DSB system are similar to those observed in naturally acquired NUOTs with the exception that we did not observe mtDNA being captured more frequently than ptDNA.

Discussion

Much of the literature on NUOTs is focused on how they may confound genome analyses (53-56) or affect pseudo heteroplasmy (57). However, there is also evidence that NUOTs can also be an important source of genetic variation. For example, a recent study of NUMTs in 66,038 human genome sequences revealed a complex and dynamic landscape with *de novo* NUMTs appearing in the germline once in every 10^4 births and somatic insertions once in every 10^3 cancer cells highlighting the ongoing formation and mutagenic potential of NUMTs in our species (17).

Our study shows that in the *Toxoplasmatinae*, which lack mobile elements, NUOTs are remarkably common and have been an important force in genome variation and evolution. To our knowledge, the *T. gondii* nuclear genome harbors the highest density of DNA of organellar origin ever reported at 1.61%. This percentage is more than 5-fold higher than the previous record holders *Arabidopsis thaliana* (0.25%) and *Ustilago maydis* (0.28%) (43).

Similar to the pattern observed in humans, the majority of the NUOTs in *T. gondii* arose from independent insertions rather than segmental duplications of previous insertions (58). However, unlike humans where NUMT insertions are concentrated in pericentromeric and subtelomeric regions (59). NUOTs in *T. gondii* and *N. caninum* occur evenly throughout the chromosome. A previous study reporting the presence of NUMTs in *T. gondii* found inverted repeats flanking the insertions (42). We have determined that the inverted repeat sequences are derived from mtDNA sequence block J, which is part of the *coxIII* gene. In our systematic catalog of 9379 NUMTs in the *T. gondii* genome, we rarely see inverted repeats flanking the NUMTs and element J does not make up a higher proportion of the observed NUMT sequence, suggesting it is unlikely the inverted repeats are involved in propagating the NUMTs as had been suggested. The mtDNA of *T. gondii* is full of inverted repeats, nearly all sequence blocks are redundantly present and often in alternative orientations within the same molecule (38). It should be noted that the high level of NUMTs in the nuclear genome, especially those with high similarity to the mtDNA was a significant confounding factor for the elucidation of the bizarre, highly fragmented and redundant mtDNAs of *T. gondii*, *H. hammondi* and *N. caninum* as they made it impossible to use nucleic acid hybridization, PCR and genome assembly software (38).

The fact that there are few to no NUOTs in the apicomplexans, *P. falciparum*, *T. parva* and *B. bovis*, suggests that the high proportion of insertions found in *T. gondii* and other coccidians is not a feature common to all apicomplexans. Molecular and bioinformatic studies performed in yeast, tobacco and human have shown that integration of organellar DNA fragments occur during illegitimate repair of double-strand breaks (DSBR) by nonhomologous end-joining (NHEJ) (23, 47, 60). We and others (61) do not find homologs for the NHEJ proteins, KU70, KU80 and DNA ligase IV in *P. falciparum* and *T. parva* and *B. bovis* only contains DNA ligase IV. Indeed, these organisms are known to predominantly employ homologous recombination for DNA repair (21). The NHEJ proteins are present in *T. gondii* and the other coccidians and we have validated their role in NUOT insertion in *T. gondii* by examining repair of a CRISPR induced DSB in a *ku80* knock out strain. This strain was created because it was experimentally demonstrated that *T. gondii* prefers NHEJ over other methods for DSB repair (50). The phylogenetic profile of the NHEJ pathway genes and the fact that *T. gondii* uses NHEJ preferentially suggests that the proficiency of the NHEJ pathway is sufficient to explain the high NUOT content in *T. gondii*. However, it doesn't explain the variation in insertion levels observed between coccidian species.

We have shown that the NUMT content even within closely related apicomplexan species varies significantly. These data are largely consistent with previous observations where a large variation in NUMT content has been described among insect species like *Drosophila melanogaster*, *Anopheles gambiae*, and *A. mellifera*, and even among mammals like human, mouse and rat (17, 18, 43, 62-64). This variation may be explained by two major forces: 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 (63, 65). The frequency at which mtDNA is transferred to the nucleus can be influenced by many factors, including the total number of mitochondria within a given cell, and the level of vulnerability to stressful agents that may damage the mitochondria thereby releasing mtDNA. However, given that apicomplexans generally contain only one mitochondrion (and apicoplast) per cell, this cannot sufficiently explain the observable differences in NUMT (and NUPT) accumulation (66). There is one significant difference between *Eimeria* and *Sarcocystis* which have fewer NUOTs and the high NUOT group

of *T. gondii, H. hammondi* and *N. caninum* and that is the fact that the former species contain transposable elements (31, 67) whereas the others do not. Perhaps *Sarcocystis* and *Eimeria* have evolved mechanisms for keeping the insertion of DNA under control and *Toxoplasma* and its closely related species lost this ability when they cleared mobile elements from their genome. Alternatively, *Toxoplasma* and it is close relatives may benefit from this high level of genome variation precisely because they have lost their mobile elements and the sources of genetic variation they provide. *Toxoplasma* in particular, also has the added biology of being able to propagate asexually. This means it can spend long periods of time as a haploid organism without a homologous chromosome to use as a template for repair, which could create additional advantages to having a steady supply of organellar DNA patches. There are many hypotheses that deserve exploration.

A major unanswered question is how organellar DNA enters the nucleus and is available for NHEJ repair. Analysis of existing NUOTs suggests that smaller fragments (100-200 bp) are most frequently observed, but that significantly longer inserts are occasionally seen. NUMT inserts appear to be random genome segments and do not correspond to the unique sequence blocks that are annotated in the unusual *T. gondii* mtDNA. It is possible that oxidative stress in the mitochondria may play a role in the generation of random mtDNA fragments as single and double strand mtDNA breaks are observed in yeast (68-70). But this explanation will not work for ptDNA fragments. The CRISPR DSB repair experiment revealed slight differences in the insert lengths depending on insert DNA source. CRISPR plasmid DNA insertions were longer than ptDNA insertions which were slightly longer than mtDNA inserts. In this experimental setting NUPTs were also observed slightly more frequently than NUMTs a finding that is at odds with the much higher abundance of NUMTs than NUPTs in *T. gondii* and all the other coccidian genomes. The reason(s) why NUMTs greatly exceed NUPTs is unclear but likely reflects cell biological differences in these independently acquired organelles. There are observed differences in the calculated ages of NUOTs in *T. gondii* and its closely related species suggesting that different lineages have experienced different waves of insertion or variation in removal that could not be recapitulated in the experiments used here. We did observe the insertion of multiple non-continuous DNA fragments including DNA from different sources both at the induced CRISPR DSB and naturally occurring in the genome sequence attesting to the efficiency of DSB repair in *T. gondii*. These mosaic inserts, termed NUMINs have been observed previously in plants (71).

Our attempt to identify NUOTs that were not ubiquitous in the 16 *T. gondii* strains examined revealed 50 insertions in or near genes that were enriched in a Gene Ontology analysis for processes related to microtubules and nucleotides. The distribution of NUMTs among the strains revealed both recent insertion in particular *T. gondii* clades and also patterns that are most easily explained as the result of sexual recombination. NUOTs may be an additional source of markers for following particular strains and recombination events. Our search for orthologous NUOTs between *T. gondii* and *N. caninum* resulted in only 5 orthologous NUMTs and no orthologous NUPTs. This finding highlighted the efficiency with which NUOTs are removed from the genome. These species diverged an estimated 28 MY ago (45). Of the 5 detectable NUMTs, one NUMT showed a higher level of conservation when compared to the surrounding sequence, suggesting it may have evolved under functional constraint and shed light on the types of evolutionary innovation NUOTs can provide. Indeed, functional assays suggest it might act as a cis-regulatory element affecting promoter activity. Surprisingly, the *T. gondii*-specific NUMT inserted near the putative myosin heavy chain ATPase gene dramatically decreases the gene promoter activity. These findings are intriguing because they suggest that both ancient and recent NUMT insertions can influence gene regulation. Because the putative myosin heavy chain ATPase gene implicated here belongs to a diverse family of proteins involved in parasite motility, division and penetration of host cells (72), it is possible that the NUMT inserted within its promoter region, which is conserved in all *T. gondii* strains examined, could modulate the pathogenicity and virulence of the parasite. However, this would require experimental validation. The non-orthologous U6 snRNA associated Sm family protein is also intriguing. The NUMT insertion in the promotor region of this gene is associated with an increase in expression. All coccidians have a large number of introns relative to other members of the phylum thus, one could speculate that this might confer an effect, but this would need to be tested. NUMTs are increasingly associated with functional consequences, particularly in human diseases (59) and mitigating deletions in the

evolution of primates (47). For example, mtDNA instability and increased NUMT somatic integration has been associated with colorectal adenocarcinoma (73).

Conclusion

Following identification of both *T. gondii* organellar genome sequences, it became possible to characterize the NUMTs and NUPTs that have hindered some molecular studies of these organellar DNAs. Shockingly, this protist pathogen and its closest relatives *H. hammondi* and *N. caninum* harbor the highest NUOT density ever observed with more than 10,000 organellar DNA insertions accounting for >1 MB of sequence. While most NUOTs exhibit various levels of sequence decay, several inserts are identical in sequence to their organellar source suggesting they are recent and that nuclear insertion and turnover of organellar DNA is an ongoing process in these species. NHEJ was shown to be essential for repair of an induced DSB. A survey of 16 *T. gondii* strains revealed 56 insertions that were not present in all strains and a search for orthologous insertions shared with *N. caninum* yielded 5 NUMTs. Deletion of one orthologous and one non-orthologous NUMT insertion affected gene expression of reporter constructs. To our knowledge this is the first study to experimentally establish that NUMTs can also influence gene expression. NUOTs provide a previously unexplored source of genomic variation in these important pathogens.

Materials and Methods

Source of genome Sequences.
The nuclear genomic sequences of *Toxoplasma gondi* ME49 (NCBI assembly GCA_000006565.1)*, Hammondia hammondi* H.H.34 (GCA_000258005.1) *Neospora caninum* Liverpool (GCA_000208865.1), *Sarcocystis neurona* SN3 (GCA_000727475.1)*, and Eimeria tenella* Houghton (GCA_000499545.1) were downloaded from ToxoDB.org (46). *Plasmodium falciparum* 3D7 (GCA_000002765.1)*, Babesia bovis* (GCA_000165395.1) and *Theileria annulata* (GCA_000003225.1) nuclear as well as organellar genome sequences were obtained from PlasmoDB.org and PiroplasmaDB.org respectively.

Identification of NUMTs and NUPTs.
RepeatMasker version 4.0.5 (http://www.repeatmasker.org) rather than the frequently used BLASTN was used to identify organellar DNA sequences in the nuclear genomes as it provides better accuracy, handles many-to-one hits, avoids overestimates and provides a detailed annotation of the identified sequences, which facilitates subsequent characterization. Mitochondrial and apicoplast genome sequences were used as the “repeat library” and “cross_match” was used as the search engine, with filtering out low complexity sequences and other parameters as default, to mask the corresponding nuclear genomes to identify NUMTs and NUPTs respectively. For *T. gondii* and *N. caninum* the species-specific 21 sequence blocks as well as cytochrome coding sequences, *coxI*, *coxIII* and *cob* (38) were used to mask the corresponding assembled nuclear chromosomal sequences. Since the 21 sequence blocks occur in a number of random arrangements and a NUMT can arise from a non-contiguous mitochondrial sequence, the cytochromes sequences were included to minimize the overestimation of the number of NUMTs. The apicoplast sequence *T. gondii* (GenBank: U87145.2) was used to detect NUPTs in *T. gondii* ME49 and *N. caninum* Liverpool. Since the mtDNA sequence could not be distinguished from NUMTs in *H. hammondi* (ToxoDB.org) (46), *T. gondii* mtDNA sequences were used to identify NUMTs in *H. hammondi*. Additionally, the publicly available apicoplast sequence for *H. hammondi* is incomplete in comparison to the well annotated plastid sequence of *T. gondii.* Therefore, *T. gondii's* apicoplast sequence was used to screen for NUPTs in *H. hammondi*. The partial cytochrome sequences annotated in *S. neurona* and *T. gondii* mtDNA sequences were used to identify NUMTs in *S. neurona* and the parasites' own apicoplast DNA sequence was used to identify NUPTs. NUMTs and NUPTs in *E. tenella* were detected using its own organellar genome sequences (GenBank Accessions: AB564272.1 and AY217738.1. Similarly, *Plasmodium, Babesia* and *Theileria* species-specific organellar sequences obtained from VEuPathDB.org were used to screen their respective nuclear

chromosome sequences for NUOTs and genes involved in the NHEJ DSB repair pathway. NUOTs shorter than 28 bp and overlapping insertions were only counted once.

Identification of NUMT location

Coordinates of coding regions, introns, exons, intergenic and 1 kb upstream and downstream flanking regions were obtained from the annotation data available on ToxoDB.org release 13 (46). Based on coordinates, a custom script was used to identify the genomic location for each NUMT. NUMTs located in non-coding exons were classified as being present in the UTR. NUMTs spanning more than one genomic feature were classified separately. Insertion sites were inspected to see if a target site preference could be observed.

Identification of segmental duplications, shared and uniquely-present NUMTs

Segmental duplications were identified by extracting NUMT sequences along with 150 bp of the flanking regions. The extracted sequences were used as BLASTN queries against themselves. If a NUMT along with at least 100bp of flanking sequence was present at least twice it was counted as a segmental duplication (29).

Following the identification NUMTs in *T. gondii* ME49 reference nuclear genome, NUMTs were similarly identified in nuclear genomes of two other fully assembled *T. gondii* strains, GT1 and VEG using the same mtDNA library repeat described above. Subsequently, the *de novo* assembled genomes of 13 *T. gondii* strains namely, ARI, CAST, TgCtBr5, TgCtBr9, TgCtPRC2, COUG, TgCtCo5, FOU, GAB2-2007-GAL-DOM2, MAS, p89, RUB and VAND, were downloaded from ToxoDB.org (46) and NUMTs were similarly identified. It should be noted that the genomes of these 13 strains have not be assembled into chromosomes and occur as scaffolds and contigs. Therefore, NUMTs that are present at the ends of these sequences could not be included in the analyses due to the lack of flanking regions. Additionally, any scaffold or contig that contained only mitochondrial DNA sequence was not included in further analysis. In order to identify NUMTs that were shared or unique among the strains, each NUMT sequence along with 200 bp of flanking sequences from one strain was searched against the NUMTs with 200 bp flanking sequences of all other strains using BLASTN. For each pairwise comparison, the NUMT was considered to be shared only when >150 bp of the upstream and downstream flanking sequence was present. If both flanking sequences were not present, the match was not considered. NUMTs that were present in only one strain were called 'strain-specific *T. gondii* NUMTs' and NUMTs that present in more than one strain were called 'differentially present *T. gondii* NUMTs'. NUMTs that were present in all 16 strains were called 'shared *T. gondii* NUMTs'. Strain-specific NUMTs identified from the comparison of just the three fully-assembled strains ME49, GT1 and VEG were manually verified. Finally, a similar approach was utilized to identify NUMTs that were shared or uniquely present among *T. gondii* ME49 and *N. caninum*. All BLASTN search results were filtered for an E-value = 0 and query coverage of >90%.

Estimation of NUOT age

The relative date of insertion of a NUMT can be inferred from the NUMTs age. RepeatMasker reports a percent divergence between the NUMT and the corresponding mtDNA sequence. The percent divergences were converted into genetic distances using the Jukes-Cantor Model formula $K = -\frac{3}{4}\ln\left(1-\frac{4}{3}p\right)$, where K is the genetic distance and p is the percent divergence between the mtDNA and NUMT. Percent divergence was calculated by dividing the number of substitutions between the two sequences by the sequence length. Using the genetic distance the age of each NUMT was calculated using the formula, $T = K/2\mu$, where T is the age of the NUMT, K is the genetic distance (Dataset S1), μ is the intron mutation rate. The intron mutation rate was used to calculate the age of the NUMTs since NUMTs are not expected to be under selective pressure and should have a neutral mutation rate similar to the introns. The intron mutation rate was previously calculated as 2.12×10^{-8} (10). The same rate was for used for all three species.

Expression constructs and Transient Transfection.

PCR primers (Table S5) containing the attB1 and attB2 sites were used to amplify the appropriate promoter and 5′-UTR regions from ME49 genomic DNA for the two genes tested. A two-step

overlap-extension PCR technique (74) was employed to delete the NUMT sequence from each promoter. The Gateway™ cloning system (Invitrogen) was used to clone the WT and deletion promoters. These two kinds of promoters were first cloned into pDONR221 via the BP reaction. Following sequencing verification, these promoter fragments were moved into a firefly luciferase-expressing vector (destination vectors) via the LR reaction. As constitutive *T. gondii* promoter (α-tubulin)-driven renilla luciferase-expressing construct was co-transfected along with the experimental construct. Nucleotide positions in these deletion studies are referenced with respect to the start of translation (+1). Parasite culture and transient transfections were performed as described (75). In short, *T. gondii* RH tachyzoites were transiently transfected via electroporation. 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 on the extract using the Promega Dual Luciferase kit. The different substrate requirements for Firefly and Renilla luciferase allowed us to assay reporter expression for each construct sequentially within the same extract. Reporter activity from the WT or mutagenized promoter was measured relative to the internal control, eliminating errors due to variation in parasite populations and individual transfections. Each electroporation experiment was performed 6 times and luciferase assays were performed in duplicate 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.

CRISPR-Cas9 transfection assays, amplicon sequencing and analysis of inserts

T. gondii strains RH and an RH mutant TATiΔ*ku80*, referred here as WT and Δ*ku80* (76) respectively were cultured on human immortalized fibroblast (hTERT) cells grown in Dulbecco's Modified Eagle's Medium (DMEM) supplemented with 10% heat-inactivated Cosmic Calf Serum, 0.5% 10mg/ml penicillin-streptomycin and 0.05% 10 mg/ml gentamycin. *T. gondii* RH WT and Δ*ku80* parasites were filtered, washed and resuspended in cytomix (supplemented with ATP and glutamine) at a concentration of 3.3×10^7 parasites/ml. 10^7 parasites were electroporated with 8 µg of the CRISPR/Cas9 plasmid with a single guide RNA targeting the UPRT locus utilizing a BTX electroporator (CRISPR/Cas9 plasmid and guide RNA kindly provided by Michael Cipriano and Boris Striepen). hTERT cells were infected with the transfected parasites (WT or Δ*ku80* in triplicate) and grown for 48 hours without drug selection. After 48 hours, lysed parasites were passed to infect fresh hTERT cells and placed under 10 µM of 5-fluorodeoxyuridine (5-FUDR) drug selection to select for parasites with a mutation in the UPRT gene. Parasites were harvested following four rounds of drug selection and DNA was extracted using Qiagen Blood and Tissue kit following the manufacturer's protocol. PCR was performed using primers (Forward primer: 5′-TATGGTAATTGTGAAGTGACAACCCCTCTGGA-3′; Reverse primer: 5′-AGTCAGTCAGCCCAAGCCACTTTCCATCGACT-3′) targeting the UPRT locus to generate 66 bp amplicons from DNA extracted from all six transfections. Amplicons were then indexed and paired-end sequencing (PE 150 bp) was performed using the Illumina MiSeq sequencer.
Sequenced paired-end reads were filtered and trimmed using trimmomatic (minimum length:36, sliding window:4:20) (77), The processed paired-end reads were then merged using FLASH (default parameters) (77, 78). Following quality control and merging of paired-end reads unique reads were identified using the usearch fastx_uniques script (79). The unique reads were then investigated for the presence of *T. gondii* mtDNA (21 sequence blocks), ptDNA, and CRISPR/Cas9 plasmid DNA sequences using RepeatMasker and each genome sequence as a repeat library. RepeatMasker results were then analyzed to identify the characteristics of the inserts reported in Fig. 5.

Acknowledgments

We would like to thank Dr. Michael Cipriano (University of Georgia) and Dr. Boris Striepen (University of Pennsylvania) for generously providing the RH mutant TATiΔ*ku80* and the CRISPR/Cas9 plasmid. This work was funded by NIH R01AI068908 to JCK.

Data sharing

Amplicon sequences are available via NCBI BioProject PRJNA961209 and BioSamples: SAMN34355938 and SAMN34355939. The coordinates of each *T. gondii* ME49 NUMT and NUPT and their corresponding age calculations are provided as a supplementary dataset.

References

1. A. M. Tenter, A. R. Heckeroth, L. M. Weiss, *Toxoplasma gondii*: from animals to humans. *Int J Parasitol* **30**, 1217-1258 (2000).
2. J. P. Dubey, *Toxoplasmosis of Animals and Humans* (CRC Press, ed. Third Edition, 2021).
3. F. Robert-Gangneux, M. L. Darde, Epidemiology of and diagnostic strategies for toxoplasmosis. *Clin Microbiol Rev* **25**, 264-296 (2012).
4. F. Belanger, F. Derouin, L. Grangeot-Keros, L. Meyer, Incidence and risk factors of toxoplasmosis in a cohort of human immunodeficiency virus-infected patients: 1988-1995. HEMOCO and SEROCO Study Groups. *Clinical infectious diseases : an official publication of the Infectious Diseases Society of America* **28**, 575-581 (1999).
5. R. E. Gilbert *et al.*, 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 (2001).
6. C. Su *et al.*, Globally diverse *Toxoplasma gondii* isolates comprise six major clades originating from a small number of distinct ancestral lineages. *Proceedings of the National Academy of Sciences of the United States of America* **109**, 5844-5849 (2012).
7. H. Lorenzi *et al.*, Local admixture of amplified and diversified secreted pathogenesis determinants shapes mosaic Toxoplasma gondii genomes. *Nat Commun* **7**, 10147 (2016).
8. E. K. Shwab *et al.*, Geographical patterns of *Toxoplasma gondii* genetic diversity revealed by multilocus PCR-RFLP genotyping. *Parasitology* **141**, 453-461 (2014).
9. E. Bertranpetit *et al.*, Phylogeography of *Toxoplasma gondii* points to a South American origin. *Infect Genet Evol* **48**, 150-155 (2017).
10. C. Su *et al.*, *Recent exp*ansion of *Toxoplasma* through enhanced oral transmission. *Science* **299**, 414-416 (2003).
11. T. Kleine, U. G. Maier, D. Leister, DNA Transfer from Organelles to the Nucleus: The Idiosyncratic Genetics of Endosymbiosis. *Annual Review of Plant Biology* **60**, 115-138 (2009).
12. D. Leister, Origin, evolution and genetic effects of nuclear insertions of organelle DNA. *Trends in Genetics* **21**, 655-663 (2005).
13. M. W. Gray, G. Burger, B. F. Lang, Mitochondrial evolution. *Science* **283**, 1476-1481 (1999).
14. J. L. Boore, Animal mitochondrial genomes. *Nucleic acids research* **27**, 1767-1780 (1999).
15. W. Martin *et al.*, Evolutionary analysis of Arabidopsis, cyanobacterial, and chloroplast genomes reveals plastid phylogeny and thousands of cyanobacterial genes in the nucleus. *Proceedings of the National Academy of Sciences of the United States of America* **99**, 12246-12251 (2002).
16. F. M. Calabrese *et al.*, NumtS colonization in mammalian genomes. *Sci Rep* **7**, 16357 (2017).
17. W. Wei *et al.*, Nuclear-embedded mitochondrial DNA sequences in 66,083 human genomes. *Nature* **611**, 105-114 (2022).
18. G. Dayama, W. Zhou, J. Prado-Martinez, T. Marques-Bonet, R. E. Mills, Characterization of nuclear mitochondrial insertions in the whole genomes of primates. *NAR Genom Bioinform* **2**, lqaa089 (2020).
19. M. Lang *et al.*, Polymorphic NumtS trace human population relationships. *Hum Genet* **131**, 757-771 (2012).
20. X. Yu, A. Gabriel, Patching broken chromosomes with extranuclear cellular DNA. *Mol Cell* **4**, 873-881 (1999).

21. M. Ricchetti, C. Fairhead, B. Dujon, Mitochondrial DNA repairs double-strand breaks in yeast chromosomes. *Nature* **402**, 96-100 (1999).
22. G. Barja, Updating the mitochondrial free radical theory of aging: an integrated view, key aspects, and confounding concepts. *Antioxid Redox Signal* **19**, 1420-1445 (2013).
23. K. Henze, W. Martin, How do mitochondrial genes get into the nucleus? *Trends Genet* **17**, 383-387 (2001).
24. J. Wasmuth, J. Daub, J. M. Peregrín-Alvarez, C. A. M. Finney, J. Parkinson, The origins of apicomplexan sequence innovation. *Genome Research* **19**, 1202-1213 (2009).
25. J. D. DeBarry, J. C. Kissinger, Jumbled genomes: missing Apicomplexan synteny. *Molecular biology and evolution* **28**, 2855-2871 (2011).
26. C. Feschotte, E. J. Pritham, DNA transposons and the evolution of eukaryotic genomes. *Annu Rev Genet* **41**, 331-368 (2007).
27. J. N. Wells, C. Feschotte, A Field Guide to Eukaryotic Transposable Elements. *Annu Rev Genet* **54**, 539-561 (2020).
28. D. Almojil *et al.*, The Structural, Functional and Evolutionary Impact of Transposable Elements in Eukaryotes. *Genes (Basel)* **12** (2021).
29. A. B. Barrie (2010) Role of Repetitive DNA in Apicomplexan Genome Evolution. in *Biology* (University of Texas at Arlington), p 79.
30. T. J. Templeton *et al.*, A genome-sequence survey for *Ascogregarina taiwanensis* supports evolutionary affiliation but metabolic diversity between a Gregarine and *Cryptosporidium*. *Molecular biology and evolution* **27**, 235-248 (2010).
31. T. Blazejewski *et al.*, Systems-based analysis of the *Sarcocystis neurona* genome identifies pathways that contribute to a heteroxenous life cycle. *mBio* **6** (2015).
32. A. Seidi *et al.*, Elucidating the mitochondrial proteome of *Toxoplasma gondii* reveals the presence of a divergent cytochrome c oxidase. *Elife* **7** (2018).
33. D. S. Roos *et al.*, Origin, targeting, and function of the apicomplexan plastid. *Curr Opin Microbiol* **2**, 426-432 (1999).
34. N. Egea, N. Lang-Unnasch, Phylogeny of the large extrachromosomal DNA of organisms in the phylum Apicomplexa. *J Eukaryot Microbiol* **42**, 679-684 (1995).
35. R. J. Wilson, M. J. Gardner, J. E. Feagin, D. H. Williamson, Have malaria parasites three genomes? *Parasitol Today* **7**, 134-136 (1991).
36. G. I. McFadden, M. E. Reith, J. Munholland, N. Lang-Unnasch, Plastid in human parasites. *Nature* **381**, 482 (1996).
37. R. F. Waller, G. I. McFadden, The apicoplast: a review of the derived plastid of apicomplexan parasites. *Curr Issues Mol Biol* **7**, 57-79 (2005).
38. S. Namasivayam *et al.*, A novel fragmented mitochondrial genome in the protist pathogen *Toxoplasma gondii* and related tissue coccidia. *Genome Res* **31**, 852-865 (2021).
39. L. Berna, N. Rego, M. E. Francia, The Elusive Mitochondrial Genomes of Apicomplexa: Where Are We Now? *Front Microbiol* **12**, 751775 (2021).
40. C. J. Jackson, S. G. Gornik, R. F. Waller, The mitochondrial genome and transcriptome of the basal dinoflagellate *Hematodinium* sp.: character evolution within the highly derived mitochondrial genomes of dinoflagellates. *Genome Biol Evol* **4**, 59-72 (2012).
41. R. F. Waller, C. J. Jackson, Dinoflagellate mitochondrial genomes: stretching the rules of molecular biology. *Bioessays* **31**, 237-245 (2009).
42. P. N. Ossorio, L. D. Sibley, J. C. Boothroyd, Mitochondrial-like DNA sequences flanked by direct and inverted repeats in the nuclear genome of *Toxoplasma gondii*. *J Mol Biol* **222**, 525-536 (1991).
43. E. Hazkani-Covo, R. M. Zeller, W. Martin, Molecular Poltergeists: Mitochondrial DNA Copies (numts) in Sequenced Nuclear Genomes. *PLoS genetics* **6**, e1000834 (2010).
44. B. A. Fox, D. J. Bzik, De novo pyrimidine biosynthesis is required for virulence of *Toxoplasma gondii. Nature* **415**, 926-929 (2002).
45. A. J. Reid *et al.*, Comparative genomics of the apicomplexan parasites *Toxoplasma gondii* and *Neospora caninum*: Coccidia differing in host range and transmission strategy. *PLoS pathogens* **8**, e1002567 (2012).
46. J. C. Kissinger, B. Gajria, L. Li, I. T. Paulsen, D. S. Roos, ToxoDB: accessing the *Toxoplasma gondii* genome. *Nucleic acids research* **31**, 234-236 (2003).

47. E. Hazkani-Covo, S. Covo, Numt-mediated double-strand break repair mitigates deletions during primate genome evolution. *PLoS genetics* **4**, e1000237 (2008).
48. Y. Lin, A. S. Waldman, Promiscuous patching of broken chromosomes in mammalian cells with extrachromosomal DNA. *Nucleic acids research* **29**, 3975-3981 (2001).
49. Y. Lin, A. S. Waldman, Capture of DNA sequences at double-strand breaks in mammalian chromosomes. *Genetics* **158**, 1665-1674 (2001).
50. B. A. Fox, J. G. Ristuccia, J. P. Gigley, D. J. Bzik, Efficient gene replacements in *Toxoplasma gondii* strains deficient for nonhomologous end joining. *Eukaryot Cell* **8**, 520-529 (2009).
51. B. Shen, K. M. Brown, T. D. Lee, L. D. Sibley, Efficient gene disruption in diverse strains of *Toxoplasma gondii* using CRISPR/CAS9. *mBio* **5**, e01114-01114 (2014).
52. S. M. Sidik, C. G. Hackett, F. Tran, N. J. Westwood, S. Lourido, Efficient genome engineering of *Toxoplasma gondii* using CRISPR/Cas9. *PloS one* **9**, e100450 (2014).
53. A. de Flamingh *et al.*, Numt Parser: automated identification and removal of nuclear mitochondrial pseudogenes (numts) for accurate mitochondrial genome reconstruction in *Panthera*. *J Hered* 10.1093/jhered/esac065 (2022).
54. J. C. Cihlar *et al.*, Distinguishing mitochondrial DNA and NUMT sequences amplified with the precision ID mtDNA whole genome panel. *Mitochondrion* **55**, 122-133 (2020).
55. A. E. Woerner, J. C. Cihlar, U. Smart, B. Budowle, Numt identification and removal with RtN! *Bioinformatics* **36**, 5115-5116 (2020).
56. J. D. Ring, K. Sturk-Andreaggi, M. Alyse Peck, C. Marshall, Bioinformatic removal of NUMT-associated variants in mitotiling next-generation sequencing data from whole blood samples. *Electrophoresis* **39**, 2785-2797 (2018).
57. A. T. Pagnamenta, W. Wei, S. Rahman, P. F. Chinnery, Biparental inheritance of mitochondrial DNA revisited. *Nat Rev Genet* **22**, 477-478 (2021).
58. D. Bensasson, M. W. Feldman, D. A. Petrov, Rates of DNA duplication and mitochondrial DNA insertion in the human genome. *Journal of molecular evolution* **57**, 343-354 (2003).
59. M. J. Puertas, M. Gonzalez-Sanchez, Insertions of mitochondrial DNA into the nucleus-effects and role in cell evolution. *Genome* **63**, 365-374 (2020).
60. J. L. Blanchard, G. W. Schmidt, Mitochondrial DNA migration events in yeast and humans: integration by a common end-joining mechanism and alternative perspectives on nucleotide substitution patterns. *Molecular biology and evolution* **13**, 893 (1996).
61. A. Nenarokova *et al.*, Causes and Effects of Loss of Classical Nonhomologous End Joining Pathway in Parasitic Eukaryotes. *mBio* **10** (2019).
62. P. Pamilo, L. Viljakainen, A. Vihavainen, Exceptionally high density of NUMTs in the honeybee genome. *Molecular biology and evolution* **24**, 1340-1346 (2007).
63. E. Richly, D. Leister, NUMTs in sequenced eukaryotic genomes. *Molecular biology and evolution* **21**, 1081-1084 (2004).
64. D. Bensasson, D. Zhang, D. L. Hartl, G. M. Hewitt, Mitochondrial pseudogenes: evolution's misplaced witnesses. *Trends Ecol Evol* **16**, 314-321 (2001).
65. K. L. Adams, J. D. Palmer, Evolution of mitochondrial gene content: gene loss and transfer to the nucleus. *Mol Phylogenet Evol* **29**, 380-395 (2003).
66. F. Seeber, J. Limenitakis, D. Soldati-Favre, Apicomplexan mitochondrial metabolism: a story of gains, losses and retentions. *Trends Parasitol* **24**, 468-478 (2008).
67. M. Rodriguez, W. Makalowski, Mobilome of Apicomplexa Parasites. *Genes (Basel)* **13** (2022).
68. P. M. Girard, S. Boiteux, Repair of oxidized DNA bases in the yeast *Saccharomyces cerevisiae*. *Biochimie* **79**, 559-566 (1997).
69. N. A. Doudican, B. Song, G. S. Shadel, P. W. Doetsch, Oxidative DNA damage causes mitochondrial genomic instability in *Saccharomyces cerevisiae*. *Mol Cell Biol* **25**, 5196-5204 (2005).
70. P. E. Thorsness, E. R. Weber, Escape and migration of nucleic acids between chloroplasts, mitochondria, and the nucleus. *Int Rev Cytol* **165**, 207-234 (1996).
71. S. Portugez, W. F. Martin, E. Hazkani-Covo, Mosaic mitochondrial-plastid insertions into the nuclear genome show evidence of both non-homologous end joining and homologous recombination. *BMC Evol Biol* **18**, 162 (2018).

72. M. Meissner, D. Schluter, D. Soldati, Role of *Toxoplasma gondii* myosin A in powering parasite gliding and host cell invasion. *Science* **298**, 837-840 (2002).
73. V. Srinivasainagendra *et al.*, Migration of mitochondrial DNA in the nuclear genome of colorectal adenocarcinoma. *Genome Med* **9**, 31 (2017).
74. J. Sambrook, E. R. Fritsch, T. Maniatis, *Molecular Cloning: A laboratory Manual* (Cold Spring Harbor Laboratory Press, Cold spring Harbor, NY, ed. 2nd, 1989).
75. N. Mullapudi, S. J. Joseph, J. C. Kissinger, Identification and functional characterization of cis-regulatory elements in the apicomplexan parasite *Toxoplasma gondii*. *Genome Biol* **10**, R34 (2009).
76. L. Sheiner *et al.*, A systematic screen to discover and analyze apicoplast proteins identifies a conserved and essential protein import factor. *PLoS pathogens* **7**, e1002392 (2011).
77. A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120 (2014).
78. T. Magoc, S. L. Salzberg, FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics* **27**, 2957-2963 (2011).
79. R. C. Edgar, B. J. Haas, J. C. Clemente, C. Quince, R. Knight, UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27**, 2194-2200 (2011).

Figure Legends

Figure 1. Characteristics of NUMTs and NUPTs in *T. gondii* ME49 *(A)* Cladogram of apicomplexan parasite relationships *(B)* Circos plot (Circos version 0.51) represents the distribution of NUMTs and NUPTs in the nuclear chromosomes of *T. gondii* ME49. The outer circle of yellow bands indicates the chromosomes as labeled and each tick on the band equals 100 kb. The ticks interior to the chromosomes are different features as indicated in the key. *(C)* The length and the percent identity of a NUMT and NUPT to mt and ptDNA respectively was calculated from RepeatMasker results and the distribution is plotted. *(D)* The age of each NUMT and NUPT was calculated based on the percent divergence from the mt and ptDNA of the corresponding species as described in the Methods and the distribution of the age is plotted against the number of base pairs.

Figure 2. Strain-specific presence and absence of NUMTs in *T. gondii* *(A, B)* An example of a possible strain-specific insertion in *T. gondii* (TG) GT1 *(A)* and deletion in ME49 *(B)* is shown. A multiple sequence alignment of the loci of interest was performed for the three strains, GT1, VEG and ME49, and the co-ordinates of the loci are indicated with reference with GT1 chromosome VI. The NUMT and flanking nuclear sequence are indicated in red and black respectively. *(C)* NUMTs in 16 *T. gondii* strains were compared to identify a strain-specific presence and absence. Strains are grouped and colored by clades as previously identified (7). The 14 chromosomes divided by black lines are represented on the x-axis. NUMTs that show differential presence/absence in the 16 *T. gondii* strains are numbered and ordered based on their location on the ME49 chromosomes. The coordinates of each NUMT plus 200 bp flanking region on the ME49 chromosomes are shown at the bottom. The 200 bp regions were included to provide a location for NUMTs missing in ME49. The percent divergence of the NUMT from the mtDNA sequence is indicated above the coordinates. In the matrix, "+" and "-" indicate presence and absence of a NUMT in a particular strain. Each cell in the matrix is colored based on local clade ancestry of that location identified from analyses conducted using blocks of 1,000 SNPs (7) although, it does not imply that location originated from that clade. Yellow cells indicate clade E ancestry and was not included the NUMT analysis as an assembled genome sequence for a strain from this clade was not available

Figure 3. Characteristics of NUMTs and NUPTs in *N. caninum LIV* *(A)* Circos plot represents the distribution of NUMTs and NUPTs in the nuclear chromosomes of *N. caninum LIV*. The blue bands indicate the chromosomes as labeled and each tick on the band equals 100 kb. The ticks interior to the chromosomes are indicate the location of NUMTs and NUPTs as referenced in the key. *(B)* The length and the percent identity of a NUMT and NUPT to *N. caninum* mt and ptDNA respectively was calculated from RepeatMasker results and the distribution is plotted. *(C)* The age of each NUMT and NUPT was calculated based on the percent divergence from the mt and ptDNA as described in the Methods and the distribution of the age is plotted against the number of base pairs.

Figure 4. Cis-regulatory activity of two selected NUMTs *(A, B)* The structure of WT and NUMT deletion promoter constructs for U6 snRNA-associated Sm family protein gene TGME49_286560 *(A)* and myosin heavy chain gene TGME49_254850 *(B)*. Nucleotide positions are referenced with respect to the start of translation (+1) of host gene and the red region indicates the position of the NUMT (Fig. S3). *(C, D)* Reporter assay results for the promoter of Sm-like protein *(C)* and mysoin heavy chain *(D)* genes. The graphs depict luciferase activity as ratios of Firefly:Renilla activity in relative luciferase units (RLU) from the different constructs containing either WT or mutagenized promoter. All luciferase readings are relative to an internal control (α-tubulin-Renilla). Error bars represent standard error calculated across the means of six independent electroporations. P-values were calculated using Student's *t*-test.

Figure 5. Capture of DNA insertions at the site of CRISPR/Cas9-directed double-strand breaks on the nuclear chromosome of *T. gondii.* Utilizing the CRISPR-Cas9 system double-strands were induced at the *uprt* locus in wild-type *T. gondii* parasites and in parasites with the

NHEJ pathway gene, *ku80*, knocked out. The locus was then amplified and deep-sequenced to evaluate the extrachromosomal DNA insertions (mtDNA, ptDNA, CRISPR-Cas9 plasmid DNA) at the site of the double strand break. *(A)* Apicomlexan cladogram with presence (+) or absence (-) of key NHEJ pathway genes and NUOT content in each species indicated. The coccidian branch is shown in bold. *(B, C)* The percent of amplicons with insertions from each DNA type *(B)* and the percentage of base pairs from each type present in the amplicons *(C)* are plotted for WT and Δ*ku80* parasites. The amplicons analyzed comprise all read pairs that could be merged as well as unmerged reads. Each group consists of three replicates and error bars represent standard error of means. Statistical significance was calculated using Student's *t*-test (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). *(D, E)* Distribution of the insert length for each DNA type is plotted as a proportion *(D)* and as box plots *(E)*. The data were generated from merged read pairs for one WT replicate. Color key is as shown in *(B).* Statistical significance was calculated using Student's *t*-test (**** $p < 0.0001$). *(F)* Schematic representation of the modifications observed in the amplicon reads for WT parasites. The numbers under 'WT amplicon' indicate the length of the amplicon with the inverted triangle denoting the approximate site of the double strand break. The type of DNA inserted is colored as indicated in the key in *(A).* Not all combination of insertion patterns observed in the reads are represented. Examples of reads with the kind of modifications shown in this schematic are provided in Table S11. Amplicons with WT sequence (yellow) on both ends represent reads where the pairs could be merged. Numerous unmerged read with similar insertion patterns as shown in this figure exist but are not drawn here.

Table 1. NUMTs and NUPTs detected in apicomplexan other eukaryotic genome sequences

Apicomplexan Species	NUMTs			NUPTs		
	Density (%)	Number of insertions	Number of base pairs	Density (%)	Number of insertions	Number of base pairs
T. gondii ME49[1]	1.430	9,379	895,312	0.180	1,108	112,443
H. hammondi H.H.34[1]	1.480	10,643	999,296	0.190	1,206	117,917
N. caninum LIV[1]	0.660	4,534	380,139	0.060	373	34,236
S. neurona SN*3*	0.030	454	38,353	0.020	54	27,624
E. tenella Houghton	0.040	227	22,024	0.088	338	45,513
B. bovis	0.000	0	0	0.020	11	1,697
T. parva	0.000	0	0	0.010	7	1,036
P. falciparum	0.002	5	515	0.020	30	5,408
Other Eukaryotes (43)						
A. thaliana	0.2564	820	305,602	0.03	301	50
O. sativa	0.1768	5,357	823,923	0.15	2090	782
H. sapiens	0.0087	871	263,478			
S. cerevisiae	0.0081	18	983			
D. melanogaster	0.0057	50	10,331			
M. musculus	0.0015	137	37,670			
C. elegans	0.0001	1	126			
M. domestica	0.0698	1859	2,093,630			
A. mellifera	0.0861	1,790	172,131			

[1]Non-contiguous DNA from organellar genomes can be inserted into the same locus during the same insertion event. Therefore one caveat of the approach used is overestimation of the number of insertion events. However, the identification of the amount of DNA inserted is not affected.

Table 2. Categorization of NUOTs by *T. gondii* ME49 genome feature

NUMT	Number of insertions	Number of bp	% of NUOT in each genomic feature	% of each feature in total genome
CDS	30	2,151	0.24	31.69
Intron	4,784	458,133	50.63	31.76
UTR	1,053	97,854	10.81	14.86
Within 1 kb upstream of gene	1,151	109,124	12.06	21.31
Within 1 kb downstream of gene	1,155	105,692	11.68	21.31
Further Intergenic	1,596	343,207	12.65	21.64
Overall*	9,379	895,312	1.43	100
NUPT				
CDS	4	430	0.38	31.69
Intron	724	74,995	66.70	31.76
UTR	95	8,824	7.85	14.86
Within 1 kb upstream of gene	80	7,168	6.37	21.31
Within 1 kb downstream of gene	103	9,875	8.78	21.31
Intergenic	86	10,448	9.29	21.64
Overall*	1,108	112,443	0.18	100.00

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

Figure 1

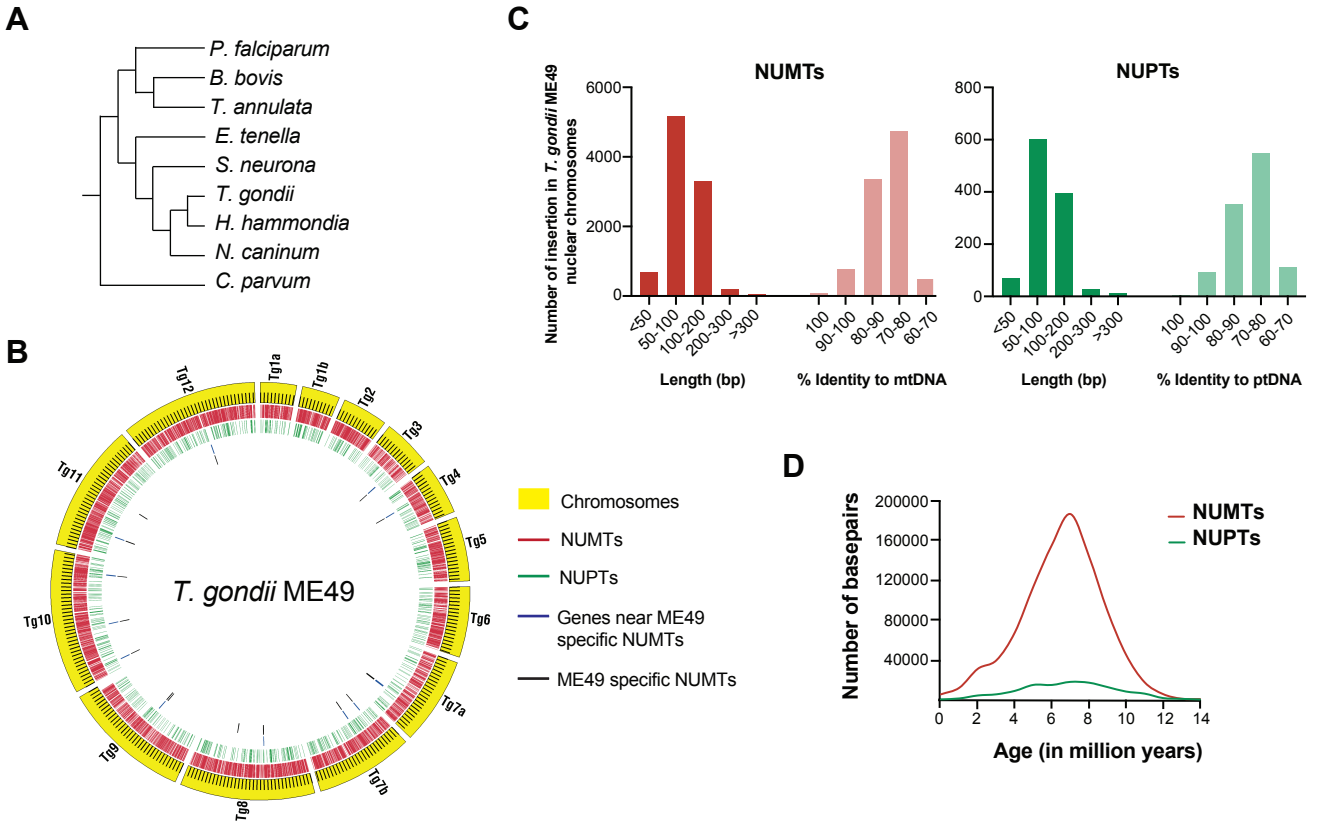

A
P. falciparum
B. bovis
T. annulata
E. tenella
S. neurona
T. gondii
H. hammondia
N. caninum
C. parvum
B
T. gondii ME49
Tg1a
Tg1b
Tg2
Tg3
Tg4
Tg5
Tg6
Tg7a
Tg7b
Tg8
Tg9
Tg10
Tg11
Tg12
Chromosomes
NUMTs
NUPTs
Genes near ME49 specific NUMTs
ME49 specific NUMTs
C
NUMTs
NUPTs
Number of insertion in T. gondii ME49 nuclear chromosomes
Length (bp)
% Identity to mtDNA
% Identity to ptDNA
D
Number of basepairs
Age (in million years)
NUMTs
NUPTs

Figure 2

A

```
TGGT1_chrVI   TGAAGCGAAGAAGAGAAATCGAAAAAAAAACGCAAGAGAGTGTAGACATTTAGCATCTGT  699307
TGME49_chrVI  TGAAGCGAAGAAGAGAAATCGAAAAAAAA-CGCGAGAGAGTG------------------
TGVEG_chrVI   TGAAGCGAAGAAGAGAAATCGAAAAAAAA-CGCGAGAGAGTG------------------

TGGT1_chrVI   CGTTAACATATGAGGATAAAAGGCAACTTTAAGCGCGGTATCAATACCTGCAGGATTGCT  699367
TGME49_chrVI  ------------------------------------------------------------
TGVEG_chrVI   ------------------------------------------------------------

TGGT1_chrVI   AGAACCATTTAAATGTAAATAGAGAGAGTGTGCACGCCCGCTGTTGCTGCGTCTTTCTTC  699427
TGME49_chrVI  ----------------------AGAGAGTGTGCACGCCCGCTGTTGCTGCGTCTTTCTTC
TGVEG_chrVI   ----------------------AGAGAGTGTGCACGCCCGCTGTTGCTGCGTCTTTCTTC
```

B

```
TGGT1_chrIV   TGTGCGACAGGTCCGTCAGCACATGCCTCGTCAAAAGAAAGGGGATTTCAGTATCCTA  1045199
TGME49_chrIV  TGTGCGACAGGTCCGTCAGCACATGCCTCGTCAAAAGAAAGGGGATTTCAGTATC---
TGVEG_chrIV   TGTGCGACAGGTCCGTCAGCACATGCCTCGTCAAAAGAAAGGGGATTTCAGTATCCTA

TGGT1_chrIV   CGGCACTCAGATTCTTCCACATCGGATTTGTTCTCGCGCAATACCTTGACTACTGTTA  1045259
TGME49_chrIV  ----------------------------------TCGCGCAGTACCTTGACTACTGTTA
TGVEG_chrIV   CGGCACTCAGATTCTTCCACATCGGATTTGTTCTCGCGCAATACCTTGACTACTGTTA
```

C

Clade	Chromosome	Ia	Ib		II	III	IV		V			VI			VIIa				VIIb					VIII			IX			X								XI										XII						
	NUMT #	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56
A	CAST	+	-	+	-	+	-	-	-	-	+	+	+	-	+	+	+	+	-	+	-	-	+	+	+	-	-	+	+	+	-	+	+	+	+	-	+	-	-	+	+	-	-	-	-	+	-	+	-	+	+	-	+	-	-	+
	TgCtCo5	+	-	-	-	+	-	-	-	-	+	+	+	-	+	-	+	+	-	+	-	-	+	+	-	+	-	+	+	+	-	+	+	+	+	+	+	-	-	+	+	-	-	-	+	-	-	-	-	+	+	-	+	-	+	-
	GT1	+	-	+	+	+	-	-	-	-	+	+	+	-	+	+	-	+	-	+	-	-	+	+	-	+	-	+	+	+	-	+	+	+	+	-	+	-	+	+	+	+	-	-	+	-	-	-	+	-	+	-	+	-	-	+
	TgCATBr9	+	-	-	+	-	-	+	-	-	+	+	+	-	-	+	-	+	-	+	-	-	+	+	-	+	-	+	-	-	-	+	+	+	+	-	+	-	-	+	-	+	+	-	-	+	+	+	+	+	+	+	+	-	-	+
	GAB2	+	-	-	-	+	-	-	-	-	+	+	+	-	+	+	-	+	-	+	-	-	+	+	-	+	-	+	+	+	-	-	-	+	+	-	+	+	-	+	-	-	-	-	-	+	+	-	-	+	+	+	-	-	+	-
	FOU	+	-	-	+	+	-	-	-	-	+	+	+	-	+	+	-	+	-	+	-	-	+	+	-	+	-	+	+	+	-	+	+	+	+	-	+	-	-	+	-	+	-	-	-	+	+	-	+	+	+	+	+	-	+	-
B	TgCATBr5	+	-	-	+	+	-	+	-	-	-	+	+	-	-	+	+	+	+	+	-	-	+	+	-	-	-	+	-	-	-	+	+	-	+	-	+	-	+	+	+	+	+	+	+	-	+	-	+	-	+	+	+	-	-	+
	MAS	+	+	+	-	+	+	+	-	-	-	-	-	-	+	+	+	+	+	+	-	-	+	+	+	+	+	+	-	-	+	+	+	+	+	-	-	-	-	+	+	+	+	+	+	-	+	-	+	+	+	-	+	-	+	+
C	p89	+	+	-	-	+	-	-	-	-	+	+	+	+	+	-	+	+	-	+	-	-	+	+	+	-	-	+	+	+	-	+	+	-	+	-	-	-	+	-	+	+	-	+	-	+	-	+	-	+	+	+	-	+	+	-
	VEG	+	-	-	-	+	-	-	-	-	+	-	+	-	+	-	+	-	-	+	-	-	-	-	+	-	-	+	+	+	-	+	+	-	+	-	-	-	-	+	+	-	-	+	-	-	+	-	+	+	+	+	-	+	-	+
D	ME49	+	-	-	-	+	-	-	-	+	-	-	+	-	+	+	+	+	-	-	+	-	-	-	+	-	+	+	+	+	+	-	-	+	+	-	+	+	-	+	+	+	-	-	+	-	-	+	-	+	-	-	-	-	+	+	-	+
	ARI	+	+	-	-	+	-	-	-	+	-	-	+	-	+	+	+	+	-	-	+	-	-	-	+	-	+	+	+	+	+	-	+	-	+	-	+	-	-	+	+	-	-	-	+	-	-	-	-	+	+	-	-	+	+	-	+
	TgCat PRC2	+	+	-	-	+	-	-	+	-	-	-	+	-	+	+	+	+	-	+	+	+	+	+	+	-	+	+	+	+	+	+	+	+	+	-	+	-	-	+	+	-	-	-	+	-	+	-	-	+	+	-	-	-	+	+	+
	COUG	+	+	+	-	+	-	-	-	+	-	-	+	-	+	+	+	+	-	-	+	+	+	-	+	-	+	+	+	+	+	+	+	+	+	-	+	-	-	+	+	-	-	-	+	-	+	-	-	+	+	-	+	+	-	+
F	VAND	-	-	-	+	+	-	-	-	-	+	+	+	+	+	+	+	-	-	+	-	-	+	+	-	-	-	+	+	+	-	+	+	-	-	+	+	-	+	+	+	+	-	-	+	-	+	-	-	-	+	-	+	-	+	+	+
	RUB	-	+	-	+	+	-	-	-	-	+	+	+	-	+	+	-	+	-	+	-	-	+	+	+	+	-	+	+	+	-	+	+	+	+	-	+	-	-	+	+	+	-	-	+	-	+	-	-	-	+	-	+	-	-	-	+
% Divergence		14	20	8	2	2	0	1.9	6	16	0	1	23	0	26	28	15	31	0	15	27	2	23	21	22	23	14	29	29	21	24	18	22	18	14	0	23	0	25	19	26	2	0	5	24	0	24	12	25	3	3	23	1	25	0	20	0	14
Chromosome Location (TGME49)		1018474-1018944	213457-213857	1596582-1596989	261484-261902	2457884-2458465	1660519-1660918	2006407-2006806	1768829-1769234	2529301-2529857	2798255-2798857	706677-707077	2259261-2259751	2415880-2416280	605108-605597	2259268-2259788	2673706-2674160	3250336-3250897	668720-669119	1025866-1026297	1710159-1710710	1904265-1904665	3218491-3219027	4646913-4647351	2497086-2497686	3466455-3466854	4435378-4435833	3008823-3009336	3038491-3039017	4251589-4252066	4753118-475768	2699533-2699941	2977559-2977959	3011324-3011796	5158557-5159008	6191403-6191803	6399526-6400012	6727704-6728163	6863803-6864202	1532798-1533403	1603990-1604629	2153288-2153772	2168335-2168735	2292428-2293114	2922696-2929196	3138783-3139183	3380178-3380628	3550551-3551008	5020102-5020569	5133457-5133946	328834-329234	905189-905588	2430105-2430505	3864644-3865116	4009932-4010391	4599466-4599866	4641543-4642129

Figure 3

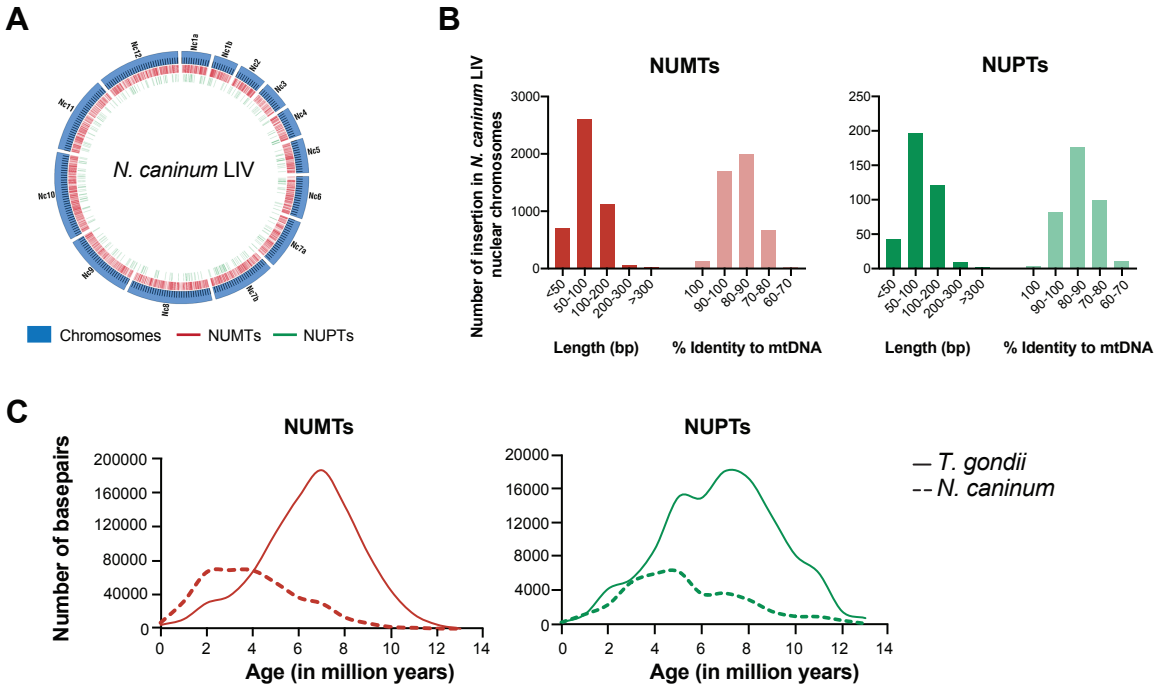

A
N. caninum LIV
Chromosomes
NUMTs
NUPTs
B
NUMTs
NUPTs
Number of insertion in N. caninum LIV nuclear chromosomes
Length (bp)
% Identity to mtDNA
C
NUMTs
NUPTs
Number of basepairs
Age (in million years)
T. gondii
N. caninum

Figure 4

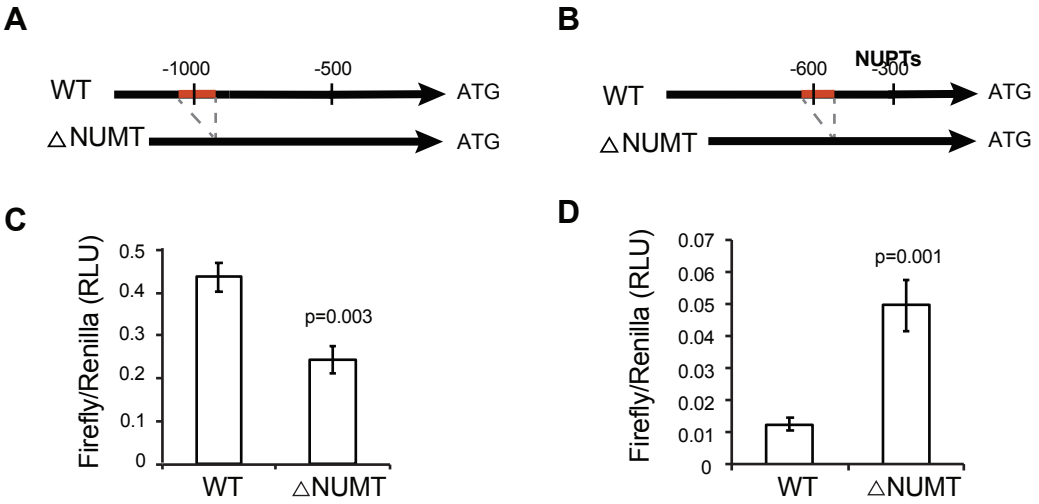
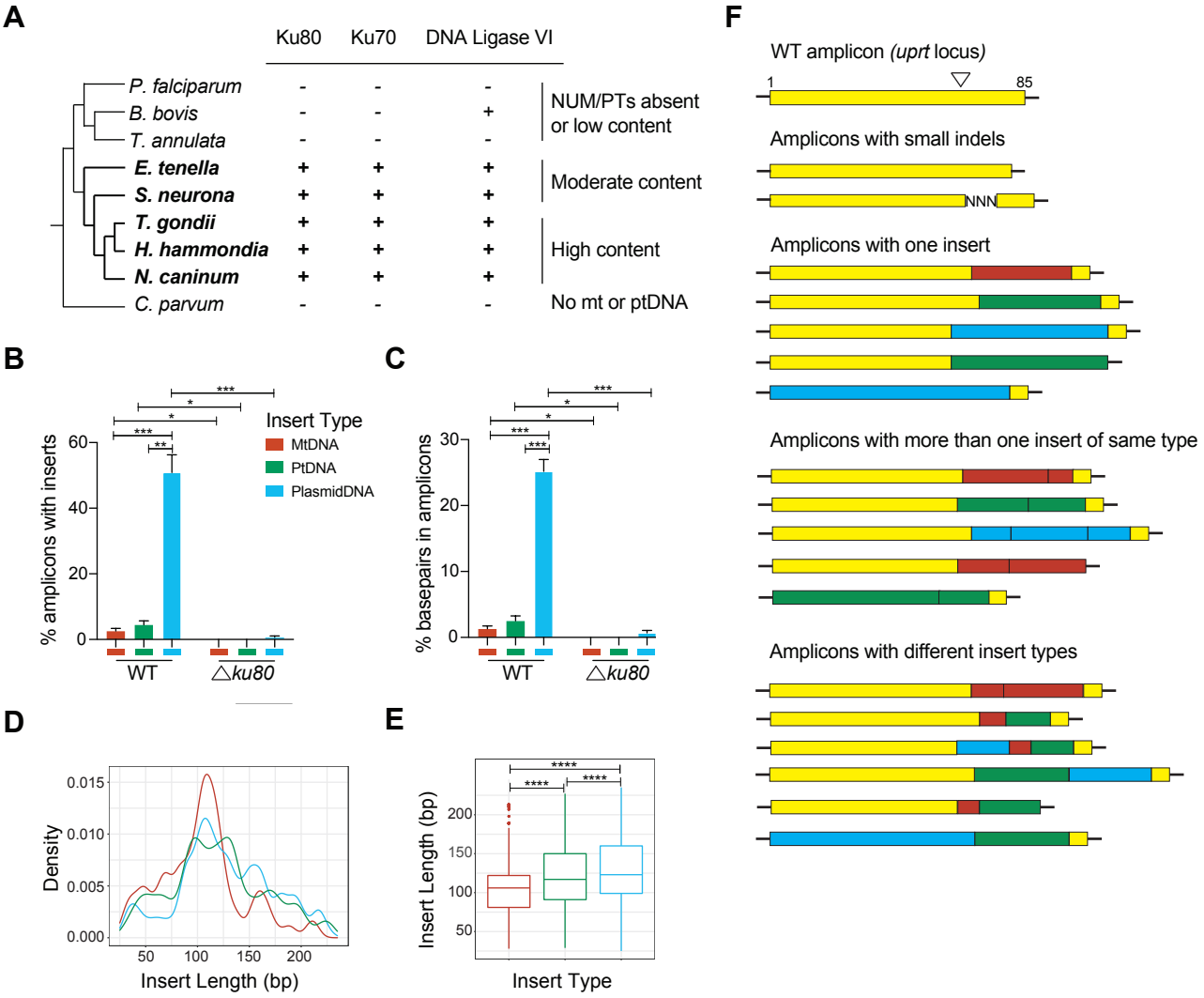

A
-1000
-500
WT
ATG
△NUMT
ATG
B
-600
NUPTs
-300
WT
ATG
△NUMT
ATG
C
Firefly/Renilla (RLU)
0.5
0.4
0.3
0.2
0.1
0
p=0.003
WT
△NUMT
D
Firefly/Renilla (RLU)
0.07
0.06
0.05
0.04
0.03
0.02
0.01
0
p=0.001
WT
△NUMT

Figure 5


A
Ku80
Ku70
DNA Ligase VI
P. falciparum - - -
B. bovis - - +
T. annulata - - -
E. tenella + + +
S. neurona + + +
T. gondii + + +
H. hammondia + + +
N. caninum + + +
C. parvum - - -
NUM/PTs absent or low content
Moderate content
High content
No mt or ptDNA
B
% amplicons with inserts
Insert Type
MtDNA
PtDNA
PlasmidDNA
WT
△ku80
C
% basepairs in amplicons
D
Density
Insert Length (bp)
E
Insert Length (bp)
Insert Type
F
WT amplicon (uprt locus)
1
85
Amplicons with small indels
NNN
Amplicons with one insert
Amplicons with more than one insert of same type
Amplicons with different insert types