Reevaluation of the *Toxoplasma gondii* and *Neospora caninum* genomes reveals misassembly, karyotype differences and chromosomal rearrangements

Luisa Berna¹, Pablo Marquez¹, Andrés Cabrera¹, Gonzalo Greif¹, María E. Francia²,³,∗, Carlos Robello¹,4,∗

**AUTHORS AFFILIATIONS**

¹ Laboratory of Host Pathogen Interactions-Molecular Biology Unit. Institut Pasteur de Montevideo. Montevideo, Uruguay

² Laboratory of Apicomplexan Biology. Institut Pasteur de Montevideo. Montevideo, Uruguay.

³ Departamento de Parasitología y Micología. Facultad de Medicina-Universidad de la República. Montevideo, Uruguay.

⁴ Departamento de Bioquímica. Facultad de Medicina-Universidad de la República. Montevideo, Uruguay.

∗Corresponding authors

**ABSTRACT**

*Neospora caninum* primarily infects cattle causing abortions with an estimated impact of a billion dollars on worldwide economy, annually. However, the study of its biology has been unheed by the established paradigm that it is virtually identical to its close relative, the widely studied human pathogen, *Toxoplasma gondii*. By revisiting the genome sequence, assembly and annotation using third generation sequencing technologies, here we show that the *N. caninum* genome was originally incorrectly assembled under the presumption of synteny with *T. gondii*. We show that major chromosomal rearrangements have occurred between these species. Importantly, we show that chromosomes originally annotated as ChrVIIb and VIII are indeed fused, reducing the karyotype of both *N. caninum* and *T. gondii* to 13 chromosomes. We reannotate the *N. caninum* genome, revealing over 500 new genes. We sequence and annotate the non-photosynthetic plastid and mitochondrial genomes, and show that while apicoplast genomes are virtually identical, high levels of gene fragmentation and reshuffling exist between species and strains. Our results correct assembly artifacts that are currently widely distributed in the genome database of *N. caninum* and *T. gondii*, but more importantly, highlight the mitochondria as a previously oversighted source of variability and pave the way for a change in the paradigm of synteny, encouraging rethinking the genome as basis of the comparative unique biology of these pathogens.

**INTRODUCTION**

The Apicomplexa comprise a large phylum of parasitic alveolates of medical and veterinary importance, causing deadly diseases such as malaria, cryptosporidiosis, neosporosis and toxoplasmosis, among others. With the exception of a few commonalities such as their obligatory intracellular lifestyle, the presence of specialized secretory organelles and of secondary endosymbionts, the apicomplexans differ greatly in
morphology, host range specificity, pathogenicity, reproductive strategy and transmission. Understanding the molecular basis of these differences has been the focus of much research. Comparative genomic analyses revealed that, albeit all small, apicomplexans genomes vary greatly in size, ranging from 9 to 130 Mb\textsuperscript{1,2}. Having diverged from a common ancestor 350-824 million years ago\textsuperscript{3}, shy of 900 genes are conserved amongst them, whereby major genomic rearrangements can be observed\textsuperscript{1}.

High synteny, defined as conserved content and order of a given genomic locus, is rarely observed\textsuperscript{1}. A seemingly stark exception to this are the genomes of *Toxoplasma gondii* and *Neospora caninum*. Morphologically, these parasites are virtually indistinguishable, so much so, that *N. caninum* was only recognized as a separate species in 1988\textsuperscript{4,5}. Moreover, both species exhibit similar tropism within their hosts, where they can infect virtually any nucleated cell. They both exhibit a fast replicating form (tachyzoite) causing acute disease, that transitions into a slow dividing form (bradyzoite) which persists in immune-privileged sites, such as the brain, establishing chronic infection. In line with this, initial comparative analysis concluded that these species have largely conserved genomic content, and are largely syntenic\textsuperscript{6}. Despite their commonalities, however, the biology of these pathogens also differs significantly. *T. gondii* infects a wide range of intermediate hosts, including humans, causing deadly disease in immunocompromised individuals or by congenital transmission. In contrast, *N. caninum* infects primarily cattle, causing abortions with an estimated impact of a billion dollars on worldwide economy, annually\textsuperscript{7}. Feline species act as definitive hosts of *T. gondii* whilst sexual replication of *N. caninum* occurs only in canids\textsuperscript{8-10}. These biological differences have been largely ascribed to absence, point mutations and pseudogenization of *T. gondii* virulence factors in *N. caninum*, and the comparative amplification of surface protein-coding gene families in *N. caninum*\textsuperscript{6,11}.

Advancements in genome sequencing technologies have accompanied the fast-paced genomics era. Particularly, third generation sequencing technologies, such as Pacific Bioscience (PacBio) and Oxford Nanopore sequencing outperform prior technologies by providing very long reads that can span regions containing repetitive sequences. This has led to improvements in the assembly of previously unattainable genomes, such as those presenting high proportions of repetitive sequences, allowing whole new genomes to be assembled with high accuracy. Here, we set out to sequence and *de novo* assemble two *N. caninum* strain’s genomes and the *T. gondii* genome, using PacBio and Oxford Nanopore. Our *de novo* assembly of the three genomes identifies largely incorrectly assembled reference genomes. Strikingly, our work reveals that previously annotated chromosomes VIIb and VIII are in fact a single chromosome in both species, reducing the total number of chromosomes in these coccidia to 13. Greater sequencing coverage and corrected assembly uncovers major chromosomal rearrangements, hundreds of previously unidentified *N. caninum* genes, and more accurate account for gene copy number. In addition, we fully annotate the previously uncharacterized apicoplast genome of *N. caninum*, and note large levels of mitochondrial heteroplasmy exist in both species. More importantly, our results ultimately challenge the paradigm of gene content synteny between *T. gondii* and *N. caninum*, prompting us to further explore these species unique genome structures as potentially overlooked sources of important biological information.
MATERIAL AND METHODS

Cell culture
*N. caninum Liverpool* was acquired from ATCC (50845). *N. caninum Uru1* was isolated from a local congenitally infected calf as described in \(^{12}\). *T. gondii RH ΔKu80\(^{13}\) strain was kindly provided by Boris Striepen. All strains were maintained and grown as described in Cabrera, et. al. 2019\(^{12}\).

High Molecular Weight DNA extraction and sequencing
For “Oxford Nanopore” and Illumina sequencing, DNA was extracted from *N. caninum* Liverpool, *N. caninum Uru1* and *T. gondii RH ΔKu80\(^{13}\) strain, using a DNA purification kit from Zymo Research (#D4074). Minion Oxford Nanopore sequencing was done *in house* as described in \(^{14}\). Sequencing libraries were prepared using a ligation sequencing kit and a Native barcoding expansion kit (EXP-NBD103/SQK-LSK108, Nanopore, England) according to \(^{15}\), starting from 1 µg of total genomic DNA. 12 hours sequencing was performed in an R9.4 Flow Cell (FLO-MIN106, Oxford Nanopore). Base calling and sequence retrieval was done using Guppy basecaller version 3.0.3. For PacBio sequencing, DNA was extracted from *N. caninum Liverpool*, *N. caninum Uru1* and *T. gondii RH ΔKu80\(^{13}\) strain by overnight incubation in lysis buffer (Tris-EDTA buffer supplemented with 1 µg/mL of RNAse A, 10% SDS, 20% Proteinase K) at 55°C, phenol/chloroform extraction, ethanol precipitation and resuspension in ultrapure water. PacBio Sequencing was performed at the Integrative Genomics Core (Beckman Research Institute, Monrovia, CA.), using 5 SMRT cells per sample for *N. caninum* and *T. gondii* genomes. Illumina sequencing of *N. caninum* DNA was also done at the Integrative Genomics Core to 65x coverage.

Genome assembly and annotation
PacBio reads were assembled using HGAP Assembly software\(^{16}\). Oxford Nanopore reads were assembled using CANU\(^{17,18}\). Assemblies were merged using Quickmerge software\(^{19}\). A single conflict region in the assembly was solved by PCR using primers: M1_F: GAGGCGCTTACAATCAACCC, H37_F_M1_R: GAGACAGGACGGACTGAAGA, H37_R: CTGCTCTGTCTGAACAGGTT, M37_F: GCGAACAGCAGAAGTCTT, M37_R: TCGTGTCTGGCAGCTCCTCT. Short insertions and deletions, a common artifact produced by long read technologies, were corrected using illumina reads in Pilon\(^{20}\). Reads used include our own, from DNA purified as described above, as well as raw reads obtained from Sequence Read Archive (SRA) NCBI repositories (Accession IDs: PRJNA531306, ERR012899 and ERR012900). Gene annotation was performed using the automated annotation tool COMPANION\(^{21}\) using Augustus Threshold 0.2, Taxon ID 5811, Align reference proteins to target sequence as parameters (others as default), and supporting RNA seq data to produce a transcript assembly. This was first aligned with Cufflinks\(^{22}\) and assembled with TopHat\(^{23}\) using SRAs with accession IDs ERR690607, ERR690608, SRR4013168, SRR4013169, SRR4013170, SRR4013171, SRR4013172 and SRR4013173 as input sequences. Apicoplast and mitochondrial genomes were identified by manual GC filtering and confirmed by Blast. Apicoplast genome was annotated using Mfannot\(^{24}\). Mitochondrial genome annotation was done in MITOS (http://mitos2.bioinf.uni-leipzig.de/index.py) using Opisthokont as reference. PacBio and MinION data has been deposited in NCBI repository (BioProject ID PRJNA597814).
Comparative Genomic Analysis
Assembled genomes were compared using NUCmer\textsuperscript{25} to create the alignments between the assemblies being compared and Assemblytics\textsuperscript{26} for visualization. Plots comparing the synteny between assemblies were obtained with the visualization tool Circos\textsuperscript{27}, using the output from Blast to create links between chromosomes. Repetitive regions were analyzed using YASS\textsuperscript{28}. Individual chromosome comparisons were performed with Artemis\textsuperscript{29} and ACT. IGV\textsuperscript{30} was used for visual inspection of aligned reads (wgs and rna-seq) on the assemblies. Specific scripts generated in this study were written in R environment\textsuperscript{31} and Bash to parse results and automate pipelines. Telomeres were identified by searching on chromosome ends for the typical TTTAGGG and AAACCCT septameric repeats. Centromeric regions in *Neospora* were determined by BLAST against previously identified centromere sequences in *Toxoplasma gondii* by ChIP-chip\textsuperscript{32}.
RESULTS

*Neospora caninum* and *Toxoplasma gondii* long-read assembly genomes

To assemble the *Neospora caninum* genome, we sequenced DNA from the reference strain *N. caninum* Liverpool (NcLiv), and a recent isolate from an experimental farm in Uruguay, named *N. caninum* Uru1 (NcUru1)\(^\text{12}\). Sequencing was done by Pacific Biosciences technology, Oxford Nanopore and Illumina. Reads derived from each sequencing were assembled independently, and then combined into a single assembly per strain (*SFig 1A*). For NcLiv, PacBio and Nanopore assemblies matched completely, with the exception of a single conflict region, which was resolved by PCR (*SFig 1B-C*). The assembled genomes were corrected with publicly available Illumina reads or those obtained in house. For NcLiv, the sequencing resulted in over 100x depth, vastly improving the currently available genome coverage (*Table 1*). Genomes of *N. caninum* strains were assembled separately.

Assemblies for both *Neospora* strains were practically indistinguishable, indicating high genome similarity. NcLiv genome consisted of 13 large contigs and 31 short (less than 122kb) unplaced fragments. With an N50 of 6.4Mb, 75% of the 61.6 Mb nuclear genome is in 8 of the largest chromosomes (*Table 1*). The 13 largest contigs correspond to complete chromosomes; both 5' and 3' telomeres were mapped for 10, while we found one telomere at one end and sub-telomeric regions at the other end for 3 (*Supp.Table 1*). Putative centromeric sequences were identified *in silico* either by blasting the *T. gondii* centromeres (Chr Ia, II, IV and V) or by identifying large regions devoid of gene-coding sequences flanked by syntenic genes flanking the centromeric regions of *T. gondii* (Chr Ib, IX and XII) (*Fig 1B-E and Supp. Table 2*). Our assembly revealed that, in contrast to what has been reported, the *N. caninum* genome is not organized in 14 chromosomes. Rather, previous ChrVIIb and ChrVIII are in fact a single chromosome (*Fig 1B*).

**Table 1. Metrics of the de novo genome assembly of *N. caninum* Liverpool**

| Genome properties               |       |
|--------------------------------|-------|
| Chromosomes                     | 13    |
| Unplaced contigs                | 31    |
| Total length                    | 61.5Mb|
| GC (%)                         | 54.8  |
| N50                            | 6.4Mb |
| N75                            | 3.6Mb |
| L50                            | 4     |
| L75                            | 8     |
| # N's per 100 kbp              | 0     |

**Protein-coding genes**

| Number of gene models                      | 7540  |
Pseudogene 258
Mean exons by gene 5.8
Percentage coding 59.2%
Mean CDS length (pb) 4872
Mean exonic length (pb) 423
Coding GC (%) 60.1%

Intergenic regions
Mean Length (pb) 3091

RNA genes
- tRNA 157
- rRNA genes 22
- snRNA 12

Apicoplast Genome (NcLiv)
- Contigs 2
- Total length 52.3 Kb
- GC (%) 23.7%

Mitochondrial Genome (NcLiv)
- Contigs 9
- Largest contig 32.4 Kb
- Total contigs length 121 Kb
- GC (%) 36.8%

Given that the previous *N. caninum* genome assembly had been constructed using as the reference the *Toxoplasma gondii* genome, by assuming that these species conserve synteny and gene order, we next examined how our new *N. caninum* genome assembly compared to that of the *T. gondii* reference genome (*Fig 1A*). To our surprise, not only did the newly assembled *N. caninum* genome differ from that of *T. gondii* in the total number of chromosomes, but a number of large chromosomal rearrangements could be observed (*Fig 1B*). Strikingly, only half of the chromosomes' structure, corresponding to *N. caninum*’s chromosomes Ib, II, III, IX, X and XII, are maintained between these species.

Next, we wondered whether the observable differences between the species were an artifact of the *T. gondii* genome assembly based on shorter reads. To this end, we sequenced the genome of a Type I strain (RH) of *T. gondii* both by PacBio and Oxford Nanopore, and *de novo* assembled the genome. Our long read sequencing resulted in a 108x coverage, and its assembly closely matched that reported by using Sanger and 454 sequencing (*Fig 1C*) confirming that the differences observed with *N. caninum* are not
artifactual. However, we were able to once again observe the fusion of chromosomes VIIb and VIII in *Toxoplasma*, thereby indicating that the karyotype for both these apicomplexans is 13, and not 14 as previously reported (Fig 1D).

Figure 1. Comparative analysis of genome assemblies of *Neospora caninum* and *Toxoplasma gondii* using third generation sequencing data reveals misassembly and karyotype differences. A. Comparative analysis of the *Toxoplasma gondii* type II (TgME49) genome assembly and the *Neospora caninum* Liverpool strain genome assembly, obtained based on Sanger technology sequencing data. B. Comparative alignment of the *Neospora caninum* Liverpool genome assemblies using Sanger and third generation (long read) technology. C. Comparative alignment of the *Toxoplasma gondii* type II (TgME49) genome assemblies based on Sanger technology sequencing data or third generation (long read) technology of *Toxoplasma gondii* type I (TgRH). D. Comparative alignment of the *Toxoplasma gondii* type I (TgRH) and the *Neospora caninum* Liverpool genome assemblies based on third generation (long read) sequencing technology. E. Chromosomal layout of *Neospora caninum*. Karyotype, chromosome length, telomeres, putative centromeres and large repeats, are shown.

We next explored whether the breaks in synteny between these closely related species correlated with any distinguishable genomic features. We noticed that the level of
sequence identity between the species varies along chromosomes. Coding regions seem to be highly conserved (82.4%). Nonetheless, we noticed that there is a detectable shift in codon usage between the two species, whereby *N. caninum* tends to use GC richer codons than *T. gondii*. The difference in codon usage conveys a detectable difference in composition whereby the mean GC content for *N. caninum* is 58.3% while it is 56.4% for *T. gondii*, being this a statistically significant difference (*p*-value $1.4\times 10^{-9}$). The same trend can be observed for GC3 (63.7% and 61.0%, respectively) and less markedly for GC1 (SFig 2). On the other hand, non-coding regions, including intronic and intergenic regions are on average 80.7% identical. Non-coding regions throughout the genome of *N. caninum* are interspaced with low complexity repetitive sequences rich in A/G or A/T. In addition, we identified three conserved sequence motifs present at a number of regions where synteny is interrupted. Particularly, five chromosomes (IV-VIII) exhibit comparative chromosomal rearrangements at regions that coincide with repetitive units of such domains (Fig2A and B, and Supp. Table 3).

![Figure 2](image)

**Figure 2.** Regions of synteny breaks between *N. caninum* and *T. gondii* are populated by three conserved domains. **A.** Sequence identity of domains identified at regions where chromosomal rearrangements have occurred. **B.** Graphical representation of chromosome VII of *N. caninum Liverpool*. Comparative alignment to the *Toxoplasma gondii*
chromosomes. Percentages of sequence identity are shown. Regions examined for the presence of motifs are indicated (a, b, and c, light green). The position of the putative centromere is indicated in orange. Note that large repetitive regions were not identified in this chromosome. 5’ (light purple) and 3’ (dark purple) telomeres are indicated. Identity and number of domains found per region, in chromosome VII, are indicated.

**Gene Annotation**

Annotation of the newly assembled NcLiv genome, using homology-based searches and RNA-seq datasets, resulted in the annotation of 7540 genes and 254 pseudogenes. Of these, 7502 are distributed in 13 chromosomes while 38 are in unplaced sequences. The products of 4553 genes have a putative assigned function, while 2987 code for hypothetical proteins. Comparatively, 554 new genes were identified in our annotation; 9 are coded for in unassembled sequences, while 545 are distributed in the chromosomes. Albeit not originally annotated, 516 of these can be found in the original Sanger sequence reads of the NcLiv genome, with over 95% identity in over 95% of the gene’s length. Of these, 494 mapped to the originally assembled chromosomes, while 22 belong to unplaced sequences.

No common theme was observed among the newly annotated genes. Of these, the most frequently found are SRS domain-containing proteins, followed by WD-domain, and zinc-finger containing proteins. This is consistent with previous findings showing that the SAG-related family of surface proteins is amplified in *N. caninum*. Previously, 227 members of this family had been identified. Here, we annotate a total of 231 SAG-related proteins. Finally, 38 novel genes were uniquely identified in the newly sequenced genome; 36 are distributed in 11 of the assembled chromosomes, while 2 remain in unplaced sequences (**Fig3 A**). Of these, 6 have putative assigned functions or recognizable domains; a Syntaxin 6, N-terminal/SNARE domain containing protein, an SPX domain/VTC domain containing protein, a Kelch motif/Galactose oxidase, central domain/BTB/POZ domain containing protein, a glutathione S-transferase, N-terminal domain containing protein, and a Cyclophilin type peptidyl-prolyl cis-trans isomerase (**Fig3B** and **Supp. Table 4**). The products of the remaining 32 genes are regarded as hypothetical. We note that many of these genes resulted either from the annotation of sequencing gaps (**Fig 3A**) or sequencing that extended through repetitive regions where previous sequencing had failed (See **Fig3C** for a representative example).

Next, we explored whether the new annotation revealed presence of uncharacterized virulence factors. We surveyed the genome for homologs of major virulence factors characterized in *T. gondii*, the majority being kinases involved in protecting the parasitophorous vacuole from host-cell intrinsic defenses; ROP5, ROP16, ROP17, ROP18 and ROP38, dense granule secreted effectors such as GRA16, GRA24, GRA25 and GRA44, as well as, IST. No major differences were found between the annotations regarding virulence factors with the exception of ROP38, whereby 4 gene copies had been reported, we could resolve 9 ROP38 copies arranged in tandem. Our annotation is consistent with previous reports describing the absence of GRA24 and IST homologs in *N. caninum*.
Figure 3. Gene annotation reveals previously unknown genes in the genome of *N. caninum*. **A.** Graphical representation of the position of novel genes in the newly assembled chromosomes. Red lines mark the position of novel genes along chromosomes. Alignment to previously assembled NcLiv genome is partially shown for reference. Three insets are shown to highlight the annotation of three new genes in three newly assembled genomic regions. **B.** Putative function of six out of the 38 newly identified genes. The remaining 32 genes are annotated as hypothetical. **C.** Representative example of the improvement in annotation in regions that had been previously collapsed due to the presence of tandem repeats. Several new genes were annotated, all of whose annotation is supported by RNA-seq data.

**Neospora caninum** apicoplast genome

The apicoplast is a relic non-photosynthetic chloroplast-like organelle of bacterial origin present in most apicomplexan parasites, whereby essential lipid synthesis occurs. This organelle is of great importance as it has been validated as the target of anti-parasitic drugs such as clindamycin. Most apicoplast proteins are encoded for in the nucleus, and later imported. However, the apicoplast harbors its own genome, which has been traditionally regarded as coding for proteins needed for its maintenance.

Surprisingly, we identified two contigs of 44 and 26 kb, with markedly lower GC content than that of the nuclear genome (21.5% vs. 59.2%), and a 10 kb inverted repeat sequence (Fig 4A-H). Annotation of these revealed 60 open reading frames most of which correspond RNA polymerase subunits and ribosomal proteins. In addition, the coding...
regions for SulfB and Clp protease, and a hypothetical protein were identified (Fig 4C, E and Supp. Table 5). These have been shown to be resident proteins of the apicoplast, thereby indicating these two contigs correspond to the apicoplast genome of *N. caninum*.

The apicoplast genome has been reported to be circular. However, we do not find compelling evidence to favor circularization of the apicoplast genome over a two linear chromosome model, in spite of using long reads, regarded as optimal to resolve these kinds of conundrums (Fig 4I and SFig 3). The assembled sequences bear significant homology at their ends, allowing their collapse into a single molecule of 56kb. However, mapped reads to this pseudomolecule do not support its circularization (Fig 4I and SFig 3C). This could be explained by lower sequencing coverage of the apicoplast genome due to its markedly lower GC content, or alternatively, it could suggest that the genome is indeed linear.

**Figure 4.** *Neospora caninum* Apicoplast genome structure and annotation. A. Read Coverage count along the length of the 44kb apicoplast genome contig. B. Percentage of GC
along the length of the 44kb apicoplast genome contig. Average GC is shown above the dotted line. C. Annotation along the length of the 44kb apicoplast genome contig (see supplementary Table 5 for details) D. The presence of an inverted repeat sequence at the 5’ end of the contig is graphically represented in a YASS plot. E. Annotation along the length of the 26kb apicoplast genome contig (see supplementary Table 5 for details) F. Percentage of GC along the length of the 26kb apicoplast genome contig. Average GC is shown above the dotted line. G. Read Coverage count along the length of the 26kb apicoplast genome contig H. The presence of an inverted repeat sequence at the 3’ end of the contig is graphically represented in a YASS plot. I. Graphical representation of the apicoplast genome structure if both contigs are fused together Read Coverage count supporting the fusion, along the length of the resulting 56kb apicoplast genome contig, are shown. Note that both annotation and the position of the inverted repeat sequences are shown. The YASS plot graphically shows the presence of both the 5’ and 3’ inverted repeats. Note, however, that the presence of this highly similar repetitive elements at both ends is not sufficient to support circularization of the genome.

Mitochondrial genomes

A number of large contigs (up to 32.4kb) amounting to a total of 121kb were identified upon sequencing NcLiv (Table 1, Fig5 and Fig4). These contain sequences with homology to classically known mitochondrial genes such as cox1 and cob1, and ribosomal RNAs (Fig4 and Supp Table 6). In addition, these contigs display a GC content averaging 36.8%, markedly lower than that of the nuclear and apicoplast genomes (Table 1). Interestingly, we could not assemble these contigs together, and neither one of them circularizes on its own. All of these contigs contain fragments of cox1 and cob1 open reading frames, however, the vast majority of them are interrupted by internal STOP codons. Three of the contigs feature replication origins of the heavy strand (OH) sequences, and only one of them features a potentially functional cox1 gene with no internal STOPs. One other contig lacks an origin of replication but features a potentially functional cob1 copy (Supp Table 6).

The mitochondrial genome of NcUru1 is distributed in 20 contigs, which vary in size ranging from 1437bp to 86kb with a medium of 12kb in size, and average a GC contents of 36.95%. The largest contig, of 86 kb, features three origin of replication (OH sequences), 36 fragments with homology to cytochrome b, 56 fragments with homology to cytochrome c oxidase subunits I and III, and several ribosomal RNA encoding genes. All genes encoding for cob, cox1 and cox3 contain internal STOP sequences which would result in truncated proteins. Likewise, both genomes encode for the intronic endonuclease LAGLI.

Interestingly, while apicoplast genomes are 95% identical between the two N. caninum strains sequenced, the contigs corresponding to the mitochondrial genomes of NcUru1 and NcLiv did not coincide with each other (Fig5). They both showcase a similar structure consisting of multiple linear contigs, featuring pseudogenized copies of cox1 and cob1, whose order seems to reshuffle in every contig (Fig5). The gene order, distribution and length is unique to each strain.

Twenty nine contigs ranging from 1.1 to 39 kb in size, with a median of 8.0kb were identified as the mitochondrial genome of T. gondii. The contigs’ GC content averages 36.6%. A similar structure of shuffled pseudogenes can be observed, whereby multiple copies of different size coding fragments of cob, cox1 and cox3 populate the mitochondrial
chromosomes, together with rRNAs (Fig 5). Origins of replication of the heavy strand (OH sequences) can be identified in six contigs. Multiple coding sequences for the intronic endonucleases GIY and LAGLI can be identified in six contigs. Strikingly, both between *N. caninum* strains, or between *N. caninum* and *T. gondii*, no single contig bearing the same combination or order of gene coding fragments could be identified (Fig 5).
DISCUSSION

Our ability to sequence genomes and annotate genes have greatly enhanced our understanding of the molecular basis of life, health and disease. Comparative genomics has allowed us to establish evolutionary relationships among living organisms, and aided in the development of specific molecular diagnosis and the rational prediction of selective drug targets. Widely used sequencing technologies, such as Sanger, 454 and Illumina, have played a pivotal part in these advancements. However, the limitations of these technologies, namely their trouble reading through repetitive regions and their short read outputs, have led to assembly artifacts that are currently widely distributed in genome and proteome databases. A number of protozoan parasite genomes have been recently revisited using third generation sequencing technologies. One noteworthy example is the genome of Trypanosoma cruzi, the causative agent of Chagas disease. This genome was massively improved for several strains. Its assembly went from being highly fragmented, in over 4000 short contigs, to less than 2000 contigs, doubling the overall genome size. Such improvements have allowed accurate account for gene copy number and the identification of an underlying genome structure. Such advances are essential in order to profiting from the advent of highly specific “guided” genome editing technologies, such as CRISPR/Cas9 technology, understand virulence linked to gene copy number, and inter and intra-species differences.

Here, we disentangled the genomes of two closely related species; Neospora caninum and Toxoplasma gondii by de novo assembling them. The N. caninum genome had been previously assembled under the presumption that they were largely syntenic, using short reads. We uncovered that even though half of the genome is indeed structured quite similarly between N. caninum and T. gondii, seven out of thirteen chromosomes differ significantly from each other. Importantly, our unbiased assembly of both genomes revealed that, strikingly, the karyotype of these apicomplexans is of 13 chromosomes. Our results suggest that chromosomes previously mapped as ChrVIIb and VIII are, in fact, a single chromosome. A number of previous findings support this fusion. Several linkage analysis assays to uncover virulence factors demonstrated that these chromosomes always segregate together. Genome interactions mapped by Hi-c analysis revealed chrVIIb and chrVIII had a higher number of interactions with each other than any other combination of chromosomes did. The latter study also showed that the number of contacts between the right telomere of chrVIIb and the left telomere of chrVIII were the highest of all, suggesting
that these could be physically linked\textsuperscript{48}. In addition, identification of the centromere of ChrVIb was unattainable by ChIP-Chip or ChIP-seq\textsuperscript{32,49}.

Interestingly, we detect a large inversion in ChrXII of \textit{N. caninum} with respect to the \textit{T. gondii} chromosome. This inversion is suggested in the 3D analysis of genome structure done by Hi-C using the \textit{Toxoplasma gondii} Type II strain ME49\textsuperscript{48}. However, we do not detect such inversion in our \textit{T. gondii} RH genome assembly suggesting that the ChrXII inversion could be strain-specific.

Genome structure differences and rearrangements are widely observed amongst apicomplexans\textsuperscript{1}. However, the driving forces of these differences are ill understood. It is well established that genomic divergence amongst Trypanosomatids, for example, can be partially ascribed to the presence of transposable elements within their genomes.

Apicomplexans’ genomes, however, are devoid of such sequences.\textsuperscript{1} We identified low complexity, repetitive regions but their appearance did not correlate with recombination prone chromosomes, but were rather evenly distributed throughout the genome. We did, however, identify a number of repetitive motifs frequently located in the vicinity of regions where synteny is lost. Experimental validation of these motifs as drivers or “soft spots” for recombination would be needed to mechanistically link them to chromosomal rearrangements.

Unlike the situation for \textit{T. gondii} whereby multiple strains have been fully sequenced, whole genomes of \textit{N. caninum} were so far limited to the reference strain \textit{N. caninum Liverpool}. Detailed population genomics studies based on whole-genome sequences from multiple strains worldwide are lacking, and so is our current understanding of population genetic structure of \textit{Neospora}. Very recently, however, a study analyzing 19 linked and unlinked genetic markers of 50 isolates collected worldwide resolved a single genotype of \textit{N. caninum}\textsuperscript{50}. This is consistent with our results whereby our whole genome assembly of NcUru1 is practically indistinguishable from that of NcLiv, despite deriving from completely different geographical locations (Europe vs. South America). Nonetheless, it is well established that great genetic variability exists amongst \textit{N. caninum} strains in the form of SNPs at particular loci and that these genetic differences underlie phenotypic variability\textsuperscript{12,51,52}. In this context, it is noteworthy that linkage assays, whereby these minimal differences can be correlated to virulence phenotypes, rely on correct genome assemblies.

The apicoplast is a validated drug target to fight apicomplexan caused diseases such as toxoplasmosis and malaria\textsuperscript{34,53}. Despite its importance, however, very few complete apicomplexan apicoplast genome sequences have been reported. The \textit{N. caninum} plastid genome sequence had not been identified prior to this study. However, plastid genome physical characterization suggested a size of approximately 35 kb; whereby formation of oligomeric molecules, migrating as linear molecules in approximate multiples of the unit length, was detected\textsuperscript{54}. Despite the ample coverage, and average read lengths that comfortably exceeding the putative apicoplast genome size, we were not able to find evidence that preferentially supported a circularize plastid genome of 35kb over two linear molecules adding up to over a 60kb, fostering an inverted repeat on their ends. It is feasible, however, that the comparatively lower GC content of the apicoplast genome hinders its effective sequencing yielding lower coverage. Nonetheless, our sequencing data cannot distinguish between the presence of a circular molecule harboring an inverted repeat or of two linear molecules, of high identity to each other. Further experiments would be required to elucidate this conundrum.
On the other hand, the apicomplexan mitochondrial genome had been described before as consisting of repeated elements of 6-7kb in length\textsuperscript{55}. Mitochondrial genomes commonly encode a number of proteins required for its maintenance; part of the translation apparatus, tRNAs, large and small ribosomal RNAs \textit{rns} and \textit{rnl}; membrane associated proteins which catalyze oxidative phosphorylation: cytochrome b (\textit{cob}), subunits of cytochrome c oxidase (\textit{cox}), ATP synthase subunits 6, 8 and 9, subunits of the NADH dehydrogenase complex; and additional ORFs of unknown functions.

Here, we found that the mitochondrial genomes of \textit{N. caninum} code for fragments of \textit{cob}, \textit{cox1}, \textit{cox3}, \textit{rns}, \textit{rnl}, \textit{Lagli} and \textit{Giy}.

Interestingly, while the nuclear and apicoplast genomes are virtually identical between the two \textit{N. caninum} strains sequenced, mitochondria genomes are strikingly different. Noteworthy, different combinations of ORFs are observed between species (\textit{N. caninum} and \textit{T. gondii}), and between strains of \textit{N. caninum}. Mitochondria also differed greatly in size. Varying size mtDNA fragments had been inadvertently observed before for \textit{Eimeria tenella}, whereby southern blot of \textit{Cox3} yielded a smear pattern of 4 to 20 kb\textsuperscript{56}. In addition, despite our ample sequencing coverage, we did not find any circular molecule, suggesting that the mtDNA is composed of linear fragments. This has been reported for \textit{Babesia}, \textit{Theileria} and \textit{Plasmodium} mtDNA\textsuperscript{57,58}. Gene structure organization differences are observable amongst these closely related hematozoa apicomplexans.

Strikingly, no mitochondrial contig in \textit{N. caninum} nor \textit{T. gondii} contained a fully functional copy of \textit{cox1} or \textit{cob}. Likewise, \textit{Plasmodium} mitochondrial ribosomal RNAs display high degree of fragmentation. It is unclear how these small RNA fragments come together to form a ribosome\textsuperscript{59}. A similar gene structure of fragmented gene pieces (divided in modules) has been described in the free living kinetoplastids. The mitochondrial gene modules of the unicellular flagellate \textit{ Diplonemid} are separately transcribed, followed by the joining of partial transcripts to contiguous RNAs\textsuperscript{60}. Mitochondrial gene editing, whereby transcripts are altered post-transcriptionally to render a functional product, has so far not described for apicomplexans. However, this process is quite common and mechanistically diverse in this organelle in a wide range of species ranging from protozoa to plants. On the other hand, functional copies of mitochondrial proteins could be encoded for in the nuclear genome and imported into the organelle. Protein import is largely used by the apicoplast, whereby proteins required for its metabolic functions are all imported from the nucleus\textsuperscript{38–41,61}. Only those proteins required for the maintenance of its genome are fully transcribed and translated within the organelle\textsuperscript{42}. Mitochondrial tRNAs, and ATP synthase subunits are known to be imported in \textit{T. gondii}\textsuperscript{55,62,63}. Noteworthy, a functional \textit{cox1} copy has been annotated in the \textit{T. gondii} nuclear genome (TgME49 _209260), and a homolog is present in the reference NcLiv genome (NCLIV_003650). In addition, nuclear-encoded divergent \textit{cox} related proteins have been identified in the mitochondrial proteome of \textit{T. gondii}\textsuperscript{64}. Nonetheless, our results pose questions regarding the mechanisms of mitochondrial protein synthesis and transport which merit consideration.

The molecular mechanisms underpinning such high variability of genetic content among mitochondrial genomes are unknown. However, we identified the presence of LAGLI and GIY among the mitochondria encoded genes. Both these proteins are endonucleases encoded for in invasive introns shown to be mobile elements. Intraspecific variability of fungal mitochondrial genomes has been mechanistically linked to the movement of these endonucleases\textsuperscript{65}. Likewise, a GIY type endonuclease present in the second intron of the mitochondrial \textit{cytochrome b} gene in the fungus \textit{Podospora curvicolla} was shown to
autonomously transfer from an ORF-containing intron to an ORF-less allele. Our results pave the way for exploring the exact contribution of these endonucleases to genome variability. It remains to be determined, as well, whether the mitochondrial genome variability observed corresponds to a clonal collection of multiple fragment, homogenously present within a population, or whether the contigs assembled represent a cohort of heteroplasmic mtDNA differentially distributed at the population level.

Overall, our results highlight distinct nuclear genomic structures, and a variable mitochondrial genome, as previously unexplored sources of genetic variability among apicomplexans. Variability is not only observed amongst closely related species, but also between strains. Efforts to explore the mechanistic contributions of this variability could shed light onto the molecular underpinnings of virulence related traits ranging from fitness to differences in drug susceptibility.

ACKNOWLEDGEMENTS
We thank Dr. Fernando Alvarez and Dr. Jon Boyle for critical reading of this manuscript. This work was funded by a grant from the National Agency for Research and Innovation (ANII) and the National Institute for Agricultural Research (INIA) number FSSA_X_2014_1_106026 awarded to CR, and an installment grant from the Banco de Seguros del Estado (BSE) awarded to MEF.

AUTHOR DECLARATION
All authors have read and approved the current version of this manuscript and have agreed to its submission
REFERENCES

1. Debarry, J. D. & Kissinger, J. C. Jumbled genomes: Missing apicomplexan synteny. *Mol. Biol. Evol.* (2011). doi:10.1093/molbev/msr103
2. Blazejewski, T. et al. Systems-Based Analysis of the Sarcocystis neurona Genome Identifies Pathways That Contribute to a Heteroxenous Life Cycle. *MBio* 6, 1–16 (2015).
3. Escalante, A. A. & Ayala, F. J. Evolutionary origin of Plasmodium and other Apicomplexa based on rRNA genes. *Proc. Natl. Acad. Sci.* (1995). doi:10.1073/pnas.92.13.5793
4. Dubey, J. P. Review of Neospora caninum and neosporosis in animals. *The Korean journal of parasitology* 41, 1–16 (2003).
5. Dubey, J. P. et al. Redescription of Neospora caninum and its differentiation from related coccidia. *Int. J. Parasitol.* (2002). doi:10.1016/S0020-7519(02)00094-2
6. Reid, A. J. et al. Comparative genomics of the apicomplexan parasites Toxoplasma gondii and neospora caninum: Coccidia differing in host range and transmission strategy. *PloS Pathog.* (2012). doi:10.1371/journal.ppat.1002567
7. Reichel, M. P., Alejandra Ayanegui-Alcérreca, M., Gondim, L. F. P. & Ellis, J. T. What is the global economic impact of Neospora caninum in cattle - The billion dollar question. *Int. J. Parasitol.* (2013). doi:10.1016/j.ijpara.2012.10.022
8. McAllister, M. M. et al. Dogs are definitive hosts of Neospora caninum. *Int. J. Parasitol.* (1998). doi:10.1016/S0020-7519(98)00138-6
9. King, J. S. et al. Australian dingoes are definitive hosts of Neospora caninum. *Int. J. Parasitol.* (2010). doi:10.1016/j.ijpara.2010.01.008
10. Gondim, L. F. P., McAllister, M. M., Pitt, W. C. & Zemlicka, D. E. Coyotes (Canis latrans) are definitive hosts of Neospora caninum. *Int. J. Parasitol.* (2004). doi:10.1016/j.ijpara.2004.01.001
11. Khan, A., Taylor, S., Ajioka, J. W., Rosenthal, B. M. & Sibley, L. D. Selection at a single locus leads to widespread expansion of toxoplasma gondii lineages that are virulent in mice. *PloS Genet.* (2009). doi:10.1371/journal.pgen.1000404
12. Cabrera, A. et al. Isolation and molecular characterization of four novel Neospora caninum strains. *Parasitol. Res.* (2019). doi:10.1007/s00436-019-06474-9
13. Huynh, M.-H. & Carruthers, V. B. Tagging of Endogenous Genes in a Toxoplasma gondii Strain Lacking Ku80. *Eukaryot. Cell* (2009). doi:10.1128/ec.00358-08
14. Díaz-Viraqué, F. et al. Nanopore Sequencing Significantly Improves Genome Assembly of the Protozoan Parasite Trypanosoma cruzi. *Genome Biol. Evol.* (2019). doi:10.1093/gbe/evz129
15. Quick, J. et al. Multiplex PCR method for MinION and Illumina sequencing of Zika and other virus genomes directly from clinical samples. *Nat. Protoc.* (2017). doi:10.1038/nprot.2017.066
16. Chin, C.-S. et al. Nonhybrid, finished microbial genome assemblies from long-read SMRT sequencing data. *Nat. Methods* (2013). doi:10.1038/nmeth.2474
17. Koren, S. et al. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *bioRxiv* (2016). doi:10.1101/071282
18. Koren, S. et al. Canu: Scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. *Genome Res.* (2017). doi:10.1101/gr.215087.116
19. Chakraborty, M., Baldwin-Brown, J. G., Long, A. D. & Emerson, J. J. Contiguous and
accurate de novo assembly of metazoan genomes with modest long read coverage. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw654

20. Walker, B. J. *et al.* Pilon: An integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* (2014). doi:10.1371/journal.pone.0112963

21. Steinbiss, S. *et al.* Companion: a web server for annotation and analysis of parasite genomes. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw292

22. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* (2010). doi:10.1038/nbt.1621

23. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: Discovering splice junctions with RNA-Seq. *Bioinformatics* (2009). doi:10.1093/bioinformatics/btp120

24. Beck, N. & Lang, B. MFannot, organelle genome annotation webserver. *Canada Univ. Montréal QC* (2010).

25. Delcher, A. L. *et al.* Alignment of whole genomes. *Nucleic Acids Res.* (1999). doi:10.1093/nar/27.11.2369

26. Nattestad, M. & Schatz, M. C. Assemblytics: A web analytics tool for the detection of variants from an assembly. *Bioinformatics* (2016). doi:10.1093/bioinformatics/btw369

27. Krzywinski, M. *et al.* Circos: An information aesthetic for comparative genomics. *Genome Res.* (2009). doi:10.1101/gr.092759.109

28. Noé, L. & Kucherov, G. YASS: Enhancing the sensitivity of DNA similarity search. *Nucleic Acids Res.* (2005). doi:10.1093/nar/gki478

29. Carver, T., Harris, S. R., Berriman, M., Parkhill, J. & McQuillan, J. A. Artemis: An integrated platform for visualization and analysis of high-throughput sequence-based experimental data. *Bioinformatics* (2012). doi:10.1093/bioinformatics/btr703

30. Robinson, J. T., Thorvaldsdóttir, H., Wenger, A. M., Zehir, A. & Mesirov, J. P. Variant review with the integrative genomics viewer. *Cancer Research* (2017). doi:10.1158/0008-5472.CAN-17-0337

31. RStudio Team. RStudio: Integrated Development for R. (2015).

32. Brooks, C. F. *et al.* Toxoplasma gondii sequesters centromeres to a specific nuclear region throughout the cell cycle. *Proc. Natl. Acad. Sci. U. S. A.* 108, (2011).

33. Waller, R. F. & McFadden, G. I. The apicoplast: A review of the derived plastid of apicomplexan parasites. *Current Issues in Molecular Biology* (2005). doi:10.21775/cimb.007.057

34. Fichera, M. E. & Roos, D. S. A plastid organelle as a drug target in apicomplexan parasites. *Nature* (1997). doi:10.1038/37132

35. Fellows, J. D., Cipriano, M. J., Agrawal, S. & Striepena, B. A plastid protein that evolved from ubiquitin and is required for apicoplast protein import in Toxoplasma gondii. *MBio* (2017). doi:10.1128/mBio.00950-17

36. Lim, L., Kalanon, M. & McFadden, G. I. New proteins in the apicoplast membranes: time to rethink apicoplast protein targeting. *Trends in Parasitology* (2009). doi:10.1016/j.pt.2009.02.001

37. Sheiner, L. *et al.* A systematic screen to discover and analyze apicoplast proteins identifies a conserved and essential protein import factor. *PLoS Pathog.* (2011). doi:10.1371/journal.ppat.1002392

38. Van Dooren, G. G., Tomova, C., Agrawal, S., Humbel, B. M. & Striepen, B. Toxoplasma
gondii Tic20 is essential for apicoplast protein import. *Proc. Natl. Acad. Sci. U. S. A.* (2008). doi:10.1073/pnas.0803862105

39. Glaser, S. *et al.* Tic22 is an essential chaperone required for protein import into the apicoplast. *J. Biol. Chem.* (2012). doi:10.1074/jbc.M112.405100

40. Waller, R. F. Protein trafficking to the plastid of Plasmodium falciparum is via the secretory pathway. *EMBO J.* (2000). doi:10.1093/emboj/19.8.1794

41. DeRocher, A., Hagen, C. B., Froehlich, J. E., Feagin, J. E. & Parsons, M. Analysis of targeting sequences demonstrates that trafficking to the Toxoplasma gondii plastid branches off the secretory system. *J. Cell Sci.* (2000).

42. Reiff, S. B., Vaishnava, S. & Striepen, B. The HU protein is important for apicoplast genome maintenance and inheritance in Toxoplasma gondii. *Eukaryot. Cell* (2012). doi:10.1128/EC.00029-12

43. Tørresen, O. K. *et al.* Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. *Nucleic Acids Res.* (2019). doi:10.1093/nar/gkz841

44. Berná, L. *et al.* Expanding an expanded genome: long-read sequencing of Trypanosoma cruzi. *Microb. genomics* (2018). doi:10.1099/mgen.0.000177

45. Berná, L. *et al.* Biology of the Trypanosoma cruzi Genome. in *Biology of Trypanosoma cruzi [Working Title]* (2019). doi:10.5772/intechopen.86144

46. Su, C., Howe, D. K., Dubey, J. P., Ajioka, J. W. & Sibley, L. D. Identification of quantitative trait loci controlling acute virulence in Toxoplasma gondii. *Proc. Natl. Acad. Sci. U. S. A.* (2002). doi:10.1073/pnas.172117099

47. Khan, A. *et al.* Composite genome map and recombination parameters derived from three archetypal lineages of Toxoplasma gondii. *Nucleic Acids Res.* (2005). doi:10.1093/nar/gki604

48. Bunnik, E. M. *et al.* Comparative 3D genome organization in apicomplexan parasites. *Proc. Natl. Acad. Sci.* 116, 3183–3192 (2019).

49. Gissot, M. *et al.* Toxoplasma gondii chromodomain protein 1 binds to heterochromatin and colocalises with centromeres and telomeres at the nuclear periphery. *PLoS One* (2012). doi:10.1371/journal.pone.0032671

50. Khan, A. *et al.* Global selective sweep of a highly inbred genome of the cattle parasite Neospora caninum. *Proc. Natl. Acad. Sci. U. S. A.* (2019). doi:10.1073/pnas.1913531116

51. Al-Qassab, S., Reichel, M. P., Ivens, A. & Ellis, J. T. Genetic diversity amongst isolates of Neospora caninum, and the development of a multiplex assay for the detection of distinct strains. *Mol. Cell. Probes* (2009). doi:10.1016/j.mcp.2009.01.006

52. Regidor-Cerrillo, J. *et al.* Genetic Diversity and Geographic Population Structure of Bovine Neospora caninum Determined by Microsatellite Genotyping Analysis. *PLoS One* (2013). doi:10.1371/journal.pone.0072678

53. Lell, B. & Kremsner, P. G. Clindamycin as an antimalarial drug: Review of clinical trials. *Antimicrobial Agents and Chemotherapy* (2002). doi:10.1128/AAC.46.8.2315-2320.2002

54. Gleeson, M. T. & Johnson, A. M. Physical characterisation of the plastid DNA in Neospora caninum. *Int. J. Parasitol.* 29, 1563–1573 (1999).

55. Esseliva, A. C., Naguleswaran, A., Hemphill, A. & Schneider, A. Mitochondrial tRNA import in Toxoplasma gondii. *J. Biol. Chem.* (2004). doi:10.1074/jbc.M404519200

56. Hikosaka, K. *et al.* Concatenated mitochondrial DNA of the coccidial parasite Eimeria
57. Hikosaka, K. et al. Divergence of the mitochondrial genome structure in the apicomplexan parasites, babesia and theileria. *Mol. Biol. Evol.* (2010). doi:10.1093/molbev/msp320

58. Preiser, P. R. et al. Recombination associated with replication of malarial mitochondrial DNA. *EMBO J.* (1996). doi:10.1002/j.1460-2075.1996.tb00401.x

59. Hillebrand, A. et al. Identification of clustered organellar short (cos) RNAs and of a conserved family of organellar RNA-binding proteins, the heptatricopeptide repeat proteins, in the malaria parasite. *Nucleic Acids Res.* (2018). doi:10.1093/nar/gky710

60. Moreira, S., Valach, M., Aoulad-Aissa, M., Otto, C. & Burger, G. Novel modes of RNA editing in mitochondria. *Nucleic Acids Res.* (2016). doi:10.1093/nar/gkw188

61. Agrawal, S. et al. An Apicoplast Localized Ubiquitylation System Is Required for the Import of Nuclear-encoded Plastid Proteins. *PLoS Pathog.* (2013). doi:10.1371/journal.ppat.1003426

62. Huet, D., Rajendran, E., van Dooren, G. G. & Lourido, S. Identification of cryptic subunits from an apicomplexan ATP synthase. *Elife* (2018). doi:10.7554/eLife.38097

63. Salunke, R., Mourier, T., Banerjee, M., Pain, A. & Shanmugam, D. Highly diverged novel subunit composition of apicomplexan F-type ATP synthase identified from Toxoplasma gondii. *PLoS Biol.* (2018). doi:10.1371/journal.pbio.2006128

64. Seidi, A. et al. Elucidating the mitochondrial proteome of toxoplasma gondii reveals the presence of a divergent cytochrome c oxidase. *Elife* (2018). doi:10.7554/eLife.38131

65. Juhász, Á. et al. Comparative analysis of the complete mitochondrial genomes of Aspergillus niger mtDNA type 1a and Aspergillus tubingensis mtDNA type 2b. *FEMS Microbiol. Lett.* (2008). doi:10.1111/j.1574-6968.2008.01077.x

66. Saguez, C. Intronic GIY-YIG endonuclease gene in the mitochondrial genome of Podospora curvicolla: evidence for mobility. *Nucleic Acids Res.* (2000). doi:10.1093/nar/28.6.1299