The genomic basis of host and vector specificity in non-pathogenic trypanosomatids

Guy R Oldrieve* (0000-0003-1428-0608), Beatrice Malacart (0000-0002-7705-1506), Javier López-Vidal (0000-0002-0144-2316), Keith R Matthews (0000-0003-0309-9184)

Institute for Immunology and Infection Research, School of Biological Sciences, University of Edinburgh, Edinburgh EH9 3FL, UK

***Corresponding author:** guy.oldrieve@ed.ac.uk

Keywords

Trypanosoma theileri, *Trypanosoma melophagium*, non-pathogenic, host and vector specificity

Summary statement

Comparing closely related non-pathogenic trypanosomes, we highlight differential investment in cell surface protein encoding genes and predict this differential investment is associated with the life histories of their hosts and vectors.

Abstract

Trypanosoma theileri, a non-pathogenic parasite of bovines, has a predicted surface protein architecture that likely aids survival in its mammalian host. Their surface proteins are encoded by genes which account for ~10% of their genome. A non-pathogenic parasite of sheep, *Trypanosoma melophagium*, is transmitted by the sheep ked and is closely related to *T. theileri*. To explore host and vector specificity between these species, we sequenced the *T. melophagium* genome and transcriptome and an annotated draft genome was assembled. *T. melophagium* was

compared to 43 kinetoplastid genomes, including *T. theileri. T. melophagium* and *T. theileri* have an AT biased genome, the greatest bias of publicly available trypanosomatids. This trend may result from selection acting to decrease the genomic nucleotide cost. The *T. melophagium* genome is 6.3Mb smaller than *T. theileri* and large families of proteins, characteristic of the predicted surface of *T. theileri,* were found to be absent or greatly reduced in *T. melophagium.* Instead, *T. melophagium* has modestly expanded protein families associated with the avoidance of complement-mediated lysis. We propose that the contrasting genomic features of these species is linked to their mode of transmission from their insect vector to their mammalian host.

Introduction

Trypanosomatidae are a family of single celled eukaryotes, characterised by a specialised mitochondrial genome, the kinetoplast. Trypanosomatidae are monoxenous (single host) or dixenous (two host) species. Dixenous trypanosomatids are obligate parasites of a broad diversity of animals and plants whilst monoxenous species are largely restricted to insects (Podlipaev, 2001). However, the taxonomy of trypanosomatids cannot be distilled into these two broad categories, as many monoxenous species opportunistically infect vertebrates (Kaufer et al., 2017) and some dixenous species have subsequently reverted to a monoxenous lifecycle (Schnaufer et al., 2002).

Expansion into vertebrate hosts gave rise to clades of trypanosomatids which represent medical and veterinary threats. Notably, *Trypanosoma cruzi* and *Leishmania* spp. cause important diseases in humans, Chagas disease and Leishmaniasis, respectively. Also, *Trypanosoma brucei* has been subjected to intense molecular and cytological study as two of its subspecies, *T. b. gambiense* and *T. b. rhodesiense*, are causative agents of Human African Trypanosomiases (HAT) (Cayla et al., 2019). In addition to the medical impact of human African trypanosomes, other African trypanosome species such as *T. b. brucei*, *Trypanosoma vivax* and *Trypanosoma congolense* cause morbidity and mortality in livestock, constraining agricultural development (Mehlitz and Molyneux, 2019).

The lineage including trypanosomatids diverged from other eukaryotes over a billion years ago (Burki, 2014; Cavalier-Smith, 2010) and possess unique adaptations at the genome level (Maslov et al., 2019). As an example, the *T. brucei* genome comprises 11 megabase chromosomes (Melville et al., 1998) along with ~5 intermediate chromosomes and ~100 mini-chromosomes (Daniels et al., 2010; Wickstead et al., 2004). Nuclear DNA is highly compact, and genes are organised into co-transcribed units, the primary transcripts of which are trans-spliced and polyadenylated to resolve mature mRNA (Clayton, 2019; Parsons et al., 1984). Their distinctive mitochondrial genome, the kinetoplast, consists of mini (Kleisen, 1975; Steinert, 1960) and maxicircles (Borst and Fase-Fowler, 1979; Simpson, 1979) that rely upon RNA editing to generate functional mRNA (Maslov et al., 2019). These features are not always retained across trypanosomatid species.

The divergence in trypanosomatid morphology, transmission, and lifecycle development has facilitated an adaption to a diverse range of hosts and vectors. The different surface protein adaptions of *T. brucei*, *T. cruzi* and *Leishmania* spp. exemplify the variant strategies adopted by these parasites to evade the immune systems of their hosts and vectors. For example, *T. brucei* is extracellular and proliferates in the blood and tissue of mammals. Its cell surface is encoded by a very extensive repertoire of variant surface glycoproteins (VSGs) (Berriman et al., 2005; Wickstead et al., 2004). VSG variation is essential to sustain long-term infections, during which antigenically distinct VSG protein types dominate at each peak, facilitating host immune evasion (Borst, 2002; Pays et al., 2004). *T. cruzi* is an intracellular parasite of wild and domestic mammals. The nuclear genome contains an expanded family of mucin genes that represent up to 6% of the genome (Buscaglia et al., 2006) and, along with trans-sialidases (Nardy et al., 2016), these genes enable sustained infections. *Leishmania* spp. are intracellular trypanosomatids whose cell surface is covered by a thick layer of glycoconjugates, including families of GP63 major surface proteases (MSPs), also known as leishmanolysin (Yao, 2010).

Studies have largely focused on pathogenic dixenous trypanosomatids, which can hold broad host niches, capable of infecting multiple mammalian species (Funk et al., 2013). In contrast, non-pathogenic dixenous trypanosomes, such as

Trypanosome theileri and *Trypanosoma melophagium*, can be highly specific to their host and vector. These species represent an attractive model to study the basis of host and vector specificity. For the remainder of this manuscript, we refer to species which cause no overt pathogenicity in immunocompetent hosts as non-pathogenic. We recognise that “non-pathogenic” species may still cause slight detriment to their hosts, which is often difficult to detect, especially when prevalence is high in the host population.

Trypanosoma theileri is a non-pathogenic bovine parasite which has a reported prevalence of 80% in cattle in the US and Europe when screened via culture-based methods (Farrar and Klei, 1990; Matthews et al., 1979; Mott et al., 2011; Schlafer, 1979). *T. theileri* is transmitted by tabanid flies (Bose and Heister, 1993). It can cause lifelong infections but remains at an extremely low parasitaemia in immunocompetent animals, indicating the presence of an effective host immune evasion mechanism or strict self-imposed population control, which prevents overt disease (Doherty et al., 1993; Seifi, 1995). Experimentally, *T. theileri* can sustain an infection for at least 12 weeks (Mott et al., 2011) which, combined with their non-pathogenic nature, has stimulated development of *T. theileri* as a potential vaccine delivery vehicle (Mott et al., 2011). The genome of *T. theileri* encodes five predicted surface protein families. These are four unique protein families, *T. theileri* putative surface proteins (TTPSPs) and one MSP family (Kelly et al., 2017). Together these genes represent 9% of the genome of *T. theileri*, comparable to the representation of the VSG gene family in *T. brucei brucei* (TREU927/4) (Berriman et al., 2005). Lastly, the trans-sialidases that characterise *T. cruzi* were also found to be highly expressed in *T. theileri*. These findings led to the suggestion of a novel immune evasion mechanism in *T. theileri,* contrasting with the well-known system in African trypanosomes and distinct from that in *T. cruzi* or *Leishmania* (Kelly et al., 2017). *T. theileri* is at the base of a clade which comprises *Trypanosoma rangeli*, *Trypanosoma cruzi* and *Trypanosoma grayi* (Kelly et al., 2014; Kelly et al., 2017), distinct from African trypanosomes, such as *T. brucei*. *T. grayi* is transmitted via the tsetse fly between African crocodilians via ingestion of tsetse faeces containing infective metacyclic forms (Hoare, 1929; Hoare, 1931).

Trypanosoma melophagium is a non-pathogenic trypanosome of the subgenus *megatrypanum*, transmitted between sheep via the sheep ked. This flightless insect vector has been eradicated from much of its original geographic distribution due to widespread pesticide use. However, where the sheep ked persists, it often carries *T. melophagium* (Gibson et al., 2010; Martinkovic et al., 2012). In a study of organic sheep farms, *T. melophagium* was found to be present in 86% of keds, however, blood smears from sheep on the same farms did not detect trypanosomes (Martinkovic et al., 2012). Other surveys via blood culture found 7.8% of sheep to be infected with *T. melophagium* (Serpil, 2008). It has historically been argued that *T. melophagium* is a monoxenous parasite of the sheep ked and that the mammalian host is obsolete for transmission (Flu, 1908; Porter; Swingle, 1911). However, extensive studies demonstrated a mammalian host is required (Hoare, 1923). Experimental infections of sheep with *T. melophagium* suggested the longest infection lasts three months and there is no lasting immunity as sheep can be reinfected with *T. melophagium* after several months of isolation (Gibson et al., 2010; Hoare, 1972)

Molecular markers place *T. melophagium* as a close relative to *T. theileri* (Martinkovic et al., 2012). SSU rRNA shares ~98% identity between *T. theileri* and *T. melophagium* isolates (Gibson et al., 2010). Presumably the divergence of *T. theileri* and *T. melophagium* is associated with their discrete host niches (Gibson et al., 2010; Martinkovic et al., 2012). Nonetheless, *T. theileri* and *T. melophagium* undergo a similar transmission cycle where metacyclic forms are produced in the insect hindgut and the infective forms are then believed to be transmitted to their mammalian host via the mouth, by ingestion of insect faeces or the whole insect body. Trypanosomes then invade their mammalian hosts and proliferate in the blood and, potentially, tissues before being taken up as a bloodmeal by their insect vector (Bose and Heister, 1993; Hoare, 1923).

A notable contrast in the biology of these parasites is the divergence in the life history of their vectors. Sheep keds, which transmit *T. melophagium*, spend their entire life attached to either the skin or wool and hair of sheep. Both male and female

keds feed on their mammalian host (Underwood et al., 2015). Tabanids, which transmit *T. theileri*, breed and lay their eggs in soil, water, or trees. The larvae and pupae stages live on vegetation and soil. Only female adults feed on mammalian blood, which is essential for egg production. Although adult tabanids show considerable adaptation to blood feeding, they also feed on the sugars of plants (Chainey, 1993).

Here we derive the *T. melophagium* genome using a combination of long and short read technologies. The genome, and its protein encoding genes, was compared to *T. theileri,* to provide insight into the biological specificity exhibited by each parasite in the context of their close phylogenetic relationship.

Materials and methods

Trypanosome culture, DNA/RNA extraction and sequencing

The full list of tools used in this study, and the options used to run those tools, can be found in Table S1.

T. melophagium was isolated from sheep blood collected on the island of St Kilda, Scotland (Kindly provided by Professor Josephine Pemberton, University of Edinburgh). Whole blood was collected into heparinized vacutainers and used within 2 days. 1 ml of blood was diluted with 5 volumes of a 50% mix of HMI9 supplemented with 20% fetal bovine serum (FBS) and Madin-Darby bovine kidney (MDBK) conditioned medium. All cultures were kept at 37°C and were examined microscopically every 3 days for 6 weeks. After propagation of *T. melophagium* by culturing of the blood sample, the specimens were transferred and co-cultured with fibroblast-like primary cells as feeder cells, isolated from the same blood sample. Due to the short lifespan of these primary cultured cells, *T. melophagium* was subsequently co-cultivated with MDBK cells and then progressively adapted to axenic conditions with a 50% mix of HMI9 and MDBK conditioned medium.

DNA was extracted from cultured *T. melophagium* using a MagAttract high molecular weight DNA kit, following the manufacturer's instructions (Qiagen) and cleaned via ethanol precipitation. The DNA was sequenced with Oxford Nanopore Technology's

(ONT) MinION (R9.4.1), following the ONT Rapid Sequencing protocol. Base-calling was performed in high accuracy mode using Guppy (available at https://community.nanoporetech.com/) which produced 1.059 gigabases (Gb) of data. PycoQC was used to visualise the data (Leger, 2019). The same DNA was sequenced with BGI's DNBseq (4.201Gb, 150 base pair (bp) reads). RNA was extracted with the RNeasy mini kit (Qiagen) including a DNAse step, following the manufacturer's instructions and sequenced with BGI's DNBseq (5.019Gb, 100bp reads). Raw DNA and RNA DNBseq reads were trimmed with Trimmomatic (Bolger et al., 2014).

T. theileri sequencing data was downloaded from NCBI. 170bp genomic reads were used (SRR13482812).

T. melophagium genome assembly and annotation

Jellyfish and GenomeScope were used to provide a k-mer based estimate of the genome size and heterozygosity using the short DNA reads described above (Marcais and Kingsford, 2011; Vurture et al., 2017).

ONT long reads were assembled using Wtdbg2 (Ruan and Li, 2020). For the polishing steps, BWA-MEM (Li, 2013) was used to align short reads and Minimap2 (Li, 2018) was used to align ONT reads. ONT reads were aligned to the Wtdbg2 draft assembly and three iterations of Racon (Vaser et al., 2017) followed by one round of Medaka (available at https://github.com/nanoporetech/medaka) were performed. DNBseq reads were mapped to the Medaka polished assembly, and two iterations of Racon were performed. Short and long reads were aligned to the Racon polished assembly to complete two final iterations of polishing with Pilon (Walker et al., 2014). At each stage of polishing, the quality of the draft genome was assessed using scaffold_stats.pl (available at https://github.com/sujaikumar/assemblage) and BUSCO (Seppey et al., 2019). BUSCO provides a metric for genome assembly and annotation completeness, based on the presence of near-universal single-copy orthologues in a genome assembly or the corresponding annotated proteins.

Both sets of reads were mapped to the draft assembly. Each contig was subjected to a DIAMOND blastx search against the InterProScan database (Buchfink et al., 2015; Jones et al., 2014b). The resulting alignments and DIAMOND hits were visualised with BlobTools (Laetsch, 2017). All the contigs had a DIAMOND hit to sequences from the *Trypanosoma* genus. Therefore, the assembly was confirmed to be free from contamination. However, two contigs were outliers in comparison to the rest of the assembly at under 100x coverage and consisting of only 3,975 and 1,892 base pairs (Fig. S2A). These contigs were removed from the assembly. BUSCO and BlobTools were re-run on this trimmed assembly to assess the final assembly's completeness and coverage (Fig. S2B).

Repeat sequences in the genome were identified and soft-masked using RepeatModeler2 and RepeatMasker (Flynn et al., 2020). BRAKER2 was used to annotate the soft-masked genome in ETP mode. The OrthoDB v10 protozoa database was utilised for protein hints (Kriventseva et al., 2019) with the addition of the *T. theileri* proteome and *T. melophagium* RNAseq evidence (Barnett et al., 2011; Bruna et al., 2021; Bruna et al., 2020; Buchfink et al., 2015; Gotoh, 2008; Hoff et al., 2016; Hoff et al., 2019; Iwata and Gotoh, 2012; Li et al., 2009; Lomsadze et al., 2014; Stanke et al., 2008; Stanke et al., 2006). BRAKER2 produced the protein and transcript files utilised in the following analysis. The *T. melophagium* proteome was functionally annotated using InterProScan (Jones et al., 2014b) using the Pfam and SignalP databases (Mistry et al., 2021; Nielsen, 2017). Genome conservation of collinearity was compared using D-Genies (Cabanettes and Klopp, 2018).

Publicly available genomes, transcriptomes and proteomes were accessed from TriTrypDB (Aslett et al., 2010) along with the *Phytomonas* EM1 assembly which was accessed from NCBI (Leinonen et al., 2011). The quality of the proteomes were assessed using BUSCO; only isolates which had > 85% complete proteomes were included, which left 43 isolates along with *T. melophagium*. A list of all the isolates used in this study can be found in File S1. The assembly statistics of the 44 genomes were assessed using scaffold_stats.pl (available at https://github.com/sujaikumar/assemblage).

The genomes from each of these trypanosomatid isolates were screened for transfer RNA genes using tRNAscan-SE (Chan et al., 2021). The outputs of these results were used to infer the strength of selection acting on translational efficiency and nucleotide cost for each isolate, along with the background mutation bias, using CodonMuSe (Seward and Kelly, 2016; Seward and Kelly, 2018). Each isolate was assessed using 1) every CDS, 2) single copy universal orthologs (identified in the orthologous protein clustering steps below) (n=992) and 3) single copy universal orthologs which are essential for every stage of the *T. brucei* life cycle (n=158). The last list of genes were identified by screening the universal single copy orthologs for genes which had a significant reduction in transcript levels in every library of an RNAi phenotype screen (>1.5 log fold decrease) (Alsford et al., 2011).

Each trypanosomatid genome was also screened for retrotransposons and long terminal repeat (LTR) retrotransposons using TransposonPSI (available at http://transposonpsi.sourceforge.net) and LTR-harvest (Ellinghaus et al., 2008), respectively.

Orthology inference

Orthologous proteins from 44 trypanosomatid proteomes were identified with OrthoFinder (Emms and Kelly, 2019) and protein clusters were summarised with KinFin (Laetsch and Blaxter, 2017), using InterProScan annotations based on the Pfam and signalP databases (Jones et al., 2014b; Mistry et al., 2021; Nielsen, 2017). A minimal cut-off threshold was not applied to the orthogroup annotation. The orthogroup annotation summary can be found in File S2. A species tree was produced by STAG and STRIDE, as part of the OrthoFinder analysis (Emms and Kelly, 2017; Emms and Kelly, 2018) which was visualised with iTOL (Letunic and Bork, 2007). STRIDE identified *Bodo saltans* as the best root for the consensus species tree.

To confirm the absence of VSGs in *T. melophagium*, all CDS sequences labelled as 'VSG' in the *T. brucei* TREU927/4 reference genome were downloaded from TriTrypDB. A blastn search was performed using these VSG sequences as the query and the *T. melophagium* genome as the database using a loose cut-off (e-value = 1e-5). A similar search was performed to confirm the reduced TTPSP counts in *T.*

melophagium. For this, the *T. theileri* transcripts were used to query a database made from the *T. melophagium* transcripts (e-value = 1e-25).

The cell surface orthogroups in this study were annotated with the orthogroups from Kelly *et al*. 2017 (Kelly et al., 2017) based on the orthogroup membership of genes in the two analyses. Unless stated otherwise, all of the figures in this study were plotted in R (Team, 2019) using ggplot2 (Wickham, 2016) and ggrepel (Slowikowski, 2018).

Results

T. melophagium genome assembly

An initial assessment of the *T. melophagium* genome, via k-mer counting, predicted that it was smaller than that of *T. theileri* (22.3 Mb and 27.6 Mb, respectively), this variation being observed in its repeat and unique sequence (Table 1). Notably, both genomes are predicted to have extremely low heterozygosity (0.30 and 0.41 for *T. melophagium* and *T. theileri,* respectively) in comparison to other *Trypanosoma* isolates (Oldrieve et al., 2021). Large gene families, such as the TTPSPs, only account for ~10% of the *T. theileri* genome and so are predicted to have a minor effect in the heterozygosity calculation. The k-mer counting prediction was similar in size to the final assembly (Table 1). The *T. melophagium* assembly consisted of 64 contigs in comparison to 253 for *T. theileri*. BUSCO assessments predict that both assemblies are 100% complete, although *T. theileri* is slightly fragmented (Table 1).

The *T. melophagium* and *T. theileri* genomes were aligned to highlight the conservation of collinearity. The conservation is more similar to the conservation between the species *T. brucei* TREU927/4 *and T. congolense* IL3000 2019 than between isolates of the same species, *T. brucei* TREU927/4 and *T. brucei* Lister 427 2018 (Figs. 1, S3). The *T. melophagium* genome was annotated with 10,057 protein encoding genes, in comparison to 11,312 in *T. theileri* (Fig. 2C, Table 1) (Kelly et al., 2017).

Trypanosomatid genomes, proteomes and transcriptomes were downloaded from TriTrypDB (Aslett et al., 2010). Proteome completeness was assessed with BUSCO. Those above 85% complete were retained, leaving 44 isolates from 9 genus (File S1). *T. theileri* and *T. melophagium* have some of the smallest genomes in this study and have the lowest GC content (40.1% and 41.2%, respectively), contrasting with the trypanosomatid mean of 48.4% (Fig. 2A).

Environmental temperature (Lao and Forsdyke, 2000; Paz et al., 2004), generation time (Shah and Gilchrist, 2011), neutral drift (Eyrewalker, 1991; Rao et al., 2011), tRNAs (Plotkin et al., 2004), translational accuracy/efficiency (Akashi, 1994; Hu et al., 2013; Shah and Gilchrist, 2011; Sorensen et al., 1989) and gene splicing and protein folding (Novoa and de Pouplana, 2012) have all been hypothesised to cause codon bias. However, codon bias can be influenced by nutrient availability (Seward and Kelly, 2016). Species with low nitrogen availability, such as the plant trypanosomatid, *Phytomonas*, have an AT rich genome, potentially to mitigate the lack of nitrogen in their plant hosts (Seward and Kelly, 2016). Selection acting on genome nucleotide cost (Sc) is in competition with selection acting to alter the translational efficiency of the genome (St) (Seward and Kelly, 2016; Seward and Kelly, 2018). Alternative hypotheses for the cause of codon bias can be excluded by analysing closely related species (Seward and Kelly, 2016). Based on universal single copy orthologues, selection pressure acting on the translational efficiency is minimal for both *T. theileri* and *T. melophagium* (Fig. 2B). In contrast, they show the greatest predicted selection pressure acting to reduce the nucleotide cost of any trypanosomatid genome, including the closely related *T. grayi*. The AT biased content can be interpreted as a remodelling of the *T. theileri* and *T. melophagium* genomes to reduce the cost of the genome (Fig. 2A,B). This pattern was consistent when every CDS was compared (Fig. S4A,B) and in universal single copy orthologs which are essential for every life cycle stage of *T. brucei* (Fig. S4C,D). *T. rangeli* displays the highest level of selection pressure acting to increase translational efficiency (Fig. 2B), potentially linked to its reduced genome length (Fig. 2C).

Orthologous protein clustering and phylogenetic inference

Orthologous clustering identifies genes that descended from a gene in the last common ancestor of the 44 trypanosomatid proteomes used in this study (Fig. 2, File S1). From the 44 proteomes, 18,274 orthogroups were identified (96.5% of the proteins used in this study were included in one of these orthogroups), 992 orthogroups were single copy and contained all isolates.

A species tree was generated as part of the orthologous protein clustering. Using genetic markers, previous studies have noted the similarity between *T. melophagium* and *T. theileri* (Gibson et al., 2010; Martinkovic et al., 2012). Based on 2,312 gene trees, *T. theileri* is the closest isolate to *T. melophagium* and groups with the stercorarian trypanosomes, which include *T. cruzi*, *T. grayi* and *T. rangeli,* rather than with salivarian trypanosomes such as *T. brucei* (Fig. 3). *T. grayi* is the closest isolate to the *T. melophagium* and *T. theileri* clade and is closer in size to the genome length of *T. melophagium* than to *T. theileri* (Fig. 2C). *T. melophagium* and *T. theileri* are closely related species with relatively short branch lengths (0.093 and 0.088 substitutions per site, respectively). In comparison, *T. congolense* is more divergent from *T. brucei* (0.249) whilst the isolates within *T. brucei* (*T. brucei brucei* Lister 427 2018:0.0009, *T. brucei evansi* STIB805: 0.003, *T. brucei brucei* TREU927/4: 0.003 and *T. brucei gambiense* DAL972:0.004) show less divergence than seen between *T. theileri* and *T. melophagium*.

T. theileri has a greater number of species-specific orthogroups than *T. melophagium* and 12.9% of its genes are assigned to one of these orthogroups, while *T. melophagium* has only 2.7% of its genes in a species specific orthogroup (Table 2). Terminal branch length is correlated with specific orthogroup counts, which could account for the discrepancy. However, *T. melophagium* and *T. theileri* are each other's most recent common ancestor and have been evolving at a roughly similar rate since this time, with similar terminal branch lengths (Fig. 3). To visualise these differences, the number of genes in each orthogroup was compared between *T. melophagium* and *T. theileri*. Orthogroups associated with host interaction protein families were highlighted based on their identification as a putative cell surface protein by Kelly *et al*. (2017). Many of the *T. theileri* species specific orthogroups expansions belong to a cell surface family (Fig. 4A).

Interaction with the mammalian host and predicted cell surface proteins

Firstly, and as expected, *T. melophagium* was found to lack any genes in orthogroups which contained VSGs, characteristic of African trypanosomes (File S4). To validate this, a blast search was performed using a relaxed cut off (1e-5) using the *T. melophagium* genome as the database and *T. brucei* TREU927/4 VSGs as the query. No hits were identified.

To enable comparison to the *T. theileri* genome analysis, the genes and orthogroups from this study were annotated with the host interaction genes from Kelly *et al*. (2017). Across the entire genome, *T. theileri* is predicted to contain 1,265 more genes than *T. melophagium* (Table 1). Examination of orthogroups which were associated with a *T. theileri* putative surface protein (TTPSP) revealed a large expansion in *T. theileri* (1,251 genes) compared to *T. melophagium* (10 genes) which could equate to much of the disparity in genome size (Fig. 4). To confirm the difference in TTPSPs, the *T. theileri* transcripts were subjected to a blastn search against a database consisting of the *T. melophagium* transcripts (1e-25 cut-off). The *T. melophagium* transcripts were derived from the genome annotation analysis. Only 9 *T. theileri* TTPSPs aligned to *T. melophagium*.

TTPSPs were split between four orthogroups in the original *T. theileri* genome analysis but were split between 66 orthogroups in this analysis (Table 3) (Kelly et al., 2017). TTPSPs share conserved C terminal GPI addition and N terminal signal sequences and contain regions of high divergence in the remainder of the sequence. TTPSPs are highly expressed as a family at the population level and are largely contained within tandem arrays, highlighting a similarity to the VSGs of *T. brucei* (Kelly et al., 2017). Excluding *T. melophagium*, TTPSPs are absent from all other kinetoplastid species analysed (File S4) revealing their specific innovation in these related trypanosomatids.

Following the TTPSPs, the largest expanded gene family in *T. theileri* are the MSPs. This protein family belongs to the peptidase M8 family of metalloproteinases. MSPs are likely to have contrasting roles in different life stages but their best understood function is to bind and cleave members of the complement system and for evasion of other cellular and antimicrobial immune defences (Yao, 2010). This role presumably

allows for evasion of complement-mediated proteolysis and therefore assists survival in the insect and mammalian hosts (Yao, 2010). Orthogroups which annotated as MSPs were expanded in *T. melophagium*, such as OG0008865. *T. melophagium* also has a species specific orthogroup (OG0009903) consisting of 11 proteins and annotated as MSP. Combined, these results suggest that the surface protein environment of *T. melophagium* and *T. theileri* are distinct and that genes encoding these proteins account for most of the discrepancy in the genome sizes.

Clade and species specific orthogroups

T. grayi, *T. melophagium* and *T. theileri* contain 42 clade-specific orthogroups including the MSP orthogroups OG0000590 and OG0008095. Both MSP orthogroups were expanded in *T. theileri* in comparison to *T. grayi* and *T. melophagium*. *T. melophagium* and *T. theileri* share 81 orthogroups specific to the two species. Twelve of the orthogroups are putative cell surface protein families, including five representatives from TTPSP and seven MSP orthogroups. Of the nine-remaining annotated orthogroups, OG0008096 is the largest and contains trans-sialidsases. The orthogroup is expanded in *T. theileri* (n=17) compared to *T. melophagium* (n=5).

T. theileri has 121 species specific orthogroups. Sixty-four of these were putative host interaction genes (Kelly et al., 2017). Two annotated orthogroups remained, which included a leucine rich repeat family associated with protein binding (OG0011868) and a calpain cysteine protease family (OG0018204). *T. melophagium* has 76 species specific orthogroups, 68 of these were unannotated. MSP families were annotated in three of the remaining orthogroups (OG0009903, OG0013746 and OG0013747). Other annotated species-specific orthogroups consisted of an actin family (OG0018163) along with several families associated with protein binding, WD domain G-beta repeat (OG0018131) and a leucine rich repeat family (OG0018098).

Cell surface modifying enzymes

Trans-sialidases are differentially expanded in *T. theileri* and *T. melophagium*, with 38 and 8 genes, respectively. Of the four orthogroups containing trans-sialidase, two do not contain proteins from *T. melophagium* (OG0011818 and OG0015016). These two orthogroups contain proteins from *T. cruzi* and so are likely to have been lost by

T. melophagium and *T. grayi.* In *T. cruzi*, trans-sialidases are involved in host immune evasion (Nardy et al., 2016).

Invertases are typical of the cell surface of *Leishmania* and are thought to transform sucrose into hexose in the gut of the vector. The orthogroups (OG0000150 and OG0000409) that include 20 *T. theileri* invertase genes only have 3 members in *T. melophagium* and 3 members in *T. grayi*. This expansion might indicate an adaptation of *T. theileri* to its vector, the tabanid fly, which can feed on sugary flower nectar (Kniepert, 1980). In contrast, the *T. melophagium* vector, the sheep ked, exclusively feeds on mammalian blood. The orthogroups which contain the *T. theileri* invertase genes (OG0000150 and OG0000409) contain only one gene from the plant parasite *Phytomonas* (Jaskowska et al., 2015; Sanchez-Moreno et al., 1992), suggesting a different mechanism for sucrose metabolism in these parasites.

Other putative cell surface modifying molecules, such as UDP-galactose/UDP-N-acetylglucosamine transferases (OG0000001) were expanded in *T. theileri* (n=60) compared to *T. melophagium* (n=11).

Glycolysis

Kelly *et al*., (2017) compared *T. brucei* and *T. theileri* transcriptomes which revealed differences in the abundance of glycosomal enzyme mRNAs. Particularly, pyruvate orthophosphate dikinase, phosphoenolpyruvate carboxykinase and malate dehydrogenase were found to be >10 fold more abundant in *T. theileri* than in *T. brucei* (Kelly et al., 2017). We confirm that enzymes associated with the glycolytic pathway are present in *T. melophagium* (File S4) and found that *T. melophagium* has expanded the orthogroups associated with three glycolytic enzymes. These included pyruvate orthophosphate dikinase (OG0000570), phosphoenolpyruvate carboxykinase (OG0000120) and malate dehydrogenase (OG0000332), whilst *T. melophagium* has a reduced numbers of genes in the fumarate reductase orthogroup (OG0000078). Orthogroups associated with peroxisome targeting were found to be in equal numbers (OG0004108 PEX5 and OG0003998 – PEX7). All other glycolytic enzymes are present in equal numbers in the two species. Therefore, it is likely that the glycolysis pathway is conserved in *T. melophagium.*

Life cycle

Extensive studies of *T. brucei* have identified genes which are associated with key stages of the *T. brucei* life cycle. These studies tracked genes associated with stumpy formation in the blood stream form (Cayla et al., 2020; Ling et al., 2011; Liu et al., 2020; Mony et al., 2014) and regulators of metacyclogenesis (Toh et al., 2021). These genes were combined with a list of validated development associated genes such as the RNA binding proteins RBP6, RBP7, RBP10 and ZFP2 and ZFP3 along with developmental regulators NRK A, NRK B, RDK1, RDK2, MAPK2 and phosphatases such as PTP1 and PIP39 (Domingo-Sananes et al., 2015; Gale et al., 1994; Gale and Parsons, 1993; Jones et al., 2014a; Muller et al., 2002; Szoor et al., 2010; Szoor et al., 2006; Walrad et al., 2009). Most of the orthogroups containing these genes were represented with a similar number of genes in each species, indicating the presence of an environmental sensing ability and developmental competence (Fig. 5). However, there were notable differences. There is an expansion in the orthogroups containing KRIPP14, which is a mitochondrial SSU component (Mony et al., 2014), in *T. theileri*. *T. melophagium* has expanded its orthogroups containing the kinases NRK (Domingo-Sananes et al., 2015; Gale et al., 1994), NEK and (Gale and Parsons, 1993) ADKF (Mony et al., 2014) along with a dual specificity phosphatase (DsPho) and protein phosphatases 1 (PP1) (Mony et al., 2014; Mony and Matthews, 2015). Both *T. theileri* and *T. melophagium* are missing metacaspase (MCA1) which is associated with the later stages of progression towards metacyclic forms in *T. brucei* (Toh et al., 2021) and Hyp12 which upregulates bound mRNAs during development based on tethering assays in *T. brucei* (Erben et al., 2014; Lueong et al., 2016; Mony et al., 2014; Mony and Matthews, 2015). Puf11, an effector molecule required for kinetoplast repositioning in epimastigotes (Toh et al., 2021), is also absent in *T. melophagium*.

Life cycle regulatory genes and genes controlling meiosis SPO11, MND1, HOP1 and DMC1, along with the cell fusion protein HAP2/GCS1 (Peacock et al., 2021), were found in *T. melophagium* and *T. theileri* (Fig. 5), suggesting maintenance of a sexual stage.

RNA interference and transposable elements

All 5 core genes that represent the trypanosome RNAi machinery (AGO1, DCL1, DCL2, RIF4 and RIF5) were present in *T. melophagium*, with an extra gene in the orthogroup containing DLC2 (File S4). Therefore, a functional gene silencing pathway is likely to be present in *T. melophagium,* matching the prediction in *T. theileri*.

Retrotransposon counts highlighted an expansion in *T. cruzi* and *T. vivax* isolates, along with *T. brucei* Lister 427 2018 (Fig. S5A, File S1). *T. melophagium* and *T. theileri* have not expanded their retrotransposon repertoire. A similar pattern was observed for long terminal repeat (LTR) retrotransposon counts, which show a positive correlation with genome size (Fig. S5B).

Discussion

T. melophagium and *T. theileri* are closely related trypanosomes that have distinct hosts and vectors (Gibson et al., 2010; Martinkovic et al., 2012). Here, the genome of *T. melophagium* was sequenced and a draft assembly was produced and annotated. The annotated proteome was incorporated into a comparison with *T. theileri,* and other publicly available trypanosomatid proteomes, to determine their phylogenetic relationship and to explore the genomic basis of the host and vector specificity of these non-pathogenic trypanosomatids.

Although the two genomes compared in this analysis are predicted to be complete (Table 1), their assembly and annotation were performed five years apart using different assembly pipelines and sequencing technologies. Comparing assemblies obtained using the same sequencing technologies and assembly pipelines would allow for greater confidence in the observed variations in their genome content as updated methods and long read sequencing continue to improve the quality of genome assemblies and, therefore, completeness. This was not possible since many of the assembly tools used are specific for the sequencing technologies whilst there has been substantial development in assembly methods between the two studies. However, based on the convergence of the k-mer counting based prediction with the assembly sizes, along with 100% complete BUSCO scores, we were reassured by

the quality of the draft assemblies and the subsequent comparisons of their genome content.

Using k-mer counting based predictions, *T. melophagium* was anticipated to have a smaller genome than *T. theileri*. This held true when the data were assembled (Table 1). The *T. melophagium* genome is more similar in size to *T. grayi*, the closest relative to *T. melophagium* and *T. theileri,* than to *T. theileri* (Figs. 2, 3). *T. theileri* has likely expanded its genome size since speciation occurred. A peculiarity of the *T. theileri* and *T. melophagium* isolates analysed in this study is their highly reduced heterozygosity, in contrast to African trypanosomes (Oldrieve et al., 2021). Whilst only one genome is available for both species, should these isolates represent the species as a whole, the reduced heterozygosity could be linked to a founder effect (Pool and Nielsen, 2007). As *T. theileri* and *T. melophagium* have utilised specific host and vector niches, the small population that initially expanded into the niches possibly underwent a significant population bottleneck, especially as host domestication caused eradication of wild progenitors and wild relatives, which could have facilitated a reduction in heterozygosity, induced by genetic drift. Alternatively, the absence of a sexual cycle could contribute to the reduced heterozygosity. Although *T. melophagium* and *T. theileri* contain genes required for meiosis, this does not confirm the species undergo sexual reproduction.

Selection appears to be acting to reduce the genome wide nucleotide biosynthesis cost in both *T. theileri* and *T. melophagium* (Fig. 2B) which has remodelled their genomes toward an AT bias, contrasting all other trypanosomatid genomes analysed in this study (Fig. 2 A,C). The predicted selection pressure acting to reduce nucleotide cost is at the expense of translational efficiency (Fig. 2C) and is greater than for the free-living *Bodo saltans*, or monoxenous insect parasites such as the early branching *Paratrypanosoma confusum* and *Phytomonas* EM1. *Phytomonas* has limited access to nitrogen as it infects nitrogen deficient plants and has been highlighted as an example where diet can cause selection to reduce the species genome nucleotide cost, through a reduction in GC content (Seward and Kelly, 2016). We propose that the reduction in the selection cost of *T. theileri* and *T. melophagium* may be related to their non-pathogenic nature. By remodelling their genome to an AT bias, they may have reduced their cost to their host, facilitating

reduced pathogenicity. Closely related species of bacteria exist on a spectrum from pathogen to symbiont, highlighting how a selective advantage can arise from a parasite reducing the cost to its host (Toft and Andersson, 2010). It is possible that *T. melophagium* and *T. theileri* are part of a similar spectrum amongst trypanosomatids. We acknowledge that alternative hypotheses exist for the reduced nucleotide cost associated with the AT rich genome, such as host tissue niche adaptation in their mammalian host or selection primarily operating within the arthropod vector rather than mammalian host.

This clade specific genome remodelling provides an example of the similarity between *T. theileri* and *T. melophagium*. However, the species have contrasting hosts, vectors, and genome sizes. Genome annotation and orthology inference identified candidates for their discrepancy in genome size. When species specific orthogroups were compared, the greatest contrast was between orthogroups associated with the putative cell surface, with the largest expansions detected in *T. theileri* being of TTPSP and MSP surface protein families (Fig. 4). Although both species undergo a cyclical transmission cycle, which includes mammalian and insect stages, we hypothesise the respective prevalence in their mammalian hosts, and the contrasting life history of their respective vectors could explain the genome expansion in *T. theileri*. *T. theileri*, spread by tabanids, are found in over 80% of livestock (Farrar and Klei, 1990; Matthews et al., 1979; Mott et al., 2011; Schlafer, 1979). In comparison, *T. melophagium* exhibits lower detected prevalence, being rarely identified in its mammalian host via blood smears (Martinkovic et al., 2012) or after blood culture (Serpil, 2008). Moreover, sheep keds, which transmit *T. melophagium*, are intimately associated with their mammalian host, spending their entire life either on the sheep's skin or wool. Here, males and females feed solely on mammalian blood providing many opportunities for transmission of *T. melophagium* from the sheep to the sheep ked (Hoare, 1923). Therefore, there is potentially less advantage for *T. melophagium* to invest in mammalian immune evasion mechanisms required to extend the length of its infection in sheep, since it has many transmission opportunities. The limited investment in *T. melophagium* is emphasised by their relatively unsophisticated putative TTPSP-related repertoire alongside modestly expanded species-specific MSP families (Fig. 4). Instead, *T. melophagium* could rely on its ancestral ability to sustain infections in invertebrate hosts which, although able

to be primed to defend against a specific pathogen, rely upon an innate immune response (Cooper and Eleftherianos, 2017).

In contrast, *T. theileri* has a transient host-vector interaction. Tabanid flies of either gender survive on plant sugars whilst adult females occasionally feed on mammalian blood (Chainey, 1993). Therefore, potentially *T. theileri* requires extended survival in its mammalian host to sustain transmission between cattle, compared to the intimate long-term association of sheep keds with *T. melophagium*. The investment from *T. theileri* in an expanded surface protein repertoire is likely to support adaptive immune evasion and prolonged survival in the mammalian stage of its life cycle. Alternatively, or additionally, differences between the bovine and ovine immune responses could contribute (Wang et al., 2013). It should be noted that both *T. melophagium* and *T. theileri* prevalence was surveyed via blood smear or blood culture. Although this is a standard approach, studies have highlighted the prevalence of *T. brucei* in adipose tissue (Trindade et al., 2016) and we cannot exclude one species preferentially infecting these tissues, rather than the bloodstream.

Many of the gene families identified in *T. theileri* (Kelly et al., 2017) were divided into multiple orthogroups in this study. The discrepancy is likely to be explained by evolution of the methods utilised by OrthoFinder. At the time of publication of the *T. theileri* study, OrthoFinder v.1 was available, while our analysis used version v.2.5. For instance, OrthoFinder v2.5 uses updated sequence alignment tools, such as DIAMOND ultra-sensitive. For this reason, we can speculate that the clustering in this study is more refined, such that the TTPSP families should be divided into smaller protein families. However, large paralogous orthogroups remain the toughest challenge for orthogroup clustering software and so the relationships between this set of proteins will likely continue to evolve alongside the software (Emms and Kelly, 2020).

Genes involved in the trypanosome life cycle, cellular quiescence and meiosis were all detected in *T. melophagium*, suggesting a competent developmental cycle along with the machinery for sexual recombination. There is an expansion of the *T. theileri* invertase orthogroup which was not present in *T. melophagium*. This is potentially

associated with the utilisation of sucrose in the tabanid fly's diet. Although the glycolysis pathway is present in *T. melophagium*, there was an expansion in the pyruvate orthophosphate dikinase, phosphoenolpyruvate carboxykinase and malate dehydrogenase orthogroups. These genes are associated with the branch of the glycolysis pathway that converts pyruvate to succinate to facilitate the recovery of NAD^+ (Kelly et al., 2017). This branch of the glycolytic pathway was upregulated in *T. theileri* in contrast to *T. brucei* (Kelly et al., 2017). Interestingly, the core RNAi genes were detected in *T. melophagium*, consistent with *T. theileri* but distinct from *T. cruzi* which is also a stercorarian trypanosome, but which lacks the requisite molecular machinery (Ullu et al., 2004).

In summary, we have found that *T. theileri* and *T. melophagium* are closely related species which display substantial remodelling of their genomes to facilitate a reduction in their nucleotide costs, which might reduce the costs they impose on their hosts. *T. theileri* displays a considerable genome expansion which is associated with a large repertoire of unique proteins that characterise its cell surface and host interaction gene repertoire. These genes could facilitate a lifelong infection in its mammalian host. In contrast, the comparatively unsophisticated immune evasion repertoire displayed by *T. melophagium* suggests more limited adaptation to its mammalian host.

Acknowledgments

We thank Steve Kelly (University of Oxford) for his comments and insights on the manuscript.

Competing interests

The authors declare no conflicts of interest.

Funding

This research was funded in part by the Wellcome Trust (103740/Z/14/Z and 108905/B/15/Z), the 'Supporting Evidence Based Interventions', Bill and Melinda

Gates foundation, and a Royal Society GCRF Challenge grant CH160034, both awarded to KM. For the purpose of open access, the author has applied a CC BY public copyright licence to any Author Accepted Manuscript version arising from this submission.

KM was supported by a Wellcome Trust Investigator award grant (103740/Z/14/Z), GO by a Wellcome Trust PhD studentship (108905/B/15/Z), and JLP by a Royal Society GCRF Challenge grant CH160034.

Data availability

T. melophagium DNA and RNA sequencing data, along with the draft genome assembly and its annotation, can be found under the NCBI BioProject PRJNA786535. The *T. melophagium* genome analysed in this study refers to version GCA_022059095.1.

Author contributions

Guy Oldrieve (GO) 0000-0003-1428-0608
Beatrice Malacart (BM)
Javier López Vidal (JLV)
Keith Matthews (KM) 0000-0003-0309-9184

Contributor Role	Author
Conceptualisation	GO, KM
Methodology	GO, BM
Software	GO
Validation	GO
Formal Analysis	GO, BM
Investigation	GO, BM
Resources	GO, KM, JLV
Data Curation	GO
Writing – Original Draft	GO, BM

Preparation	
Writing – Review and Editing	GO, KM, BM, JLV
Visualisation	GO, KM, BM
Supervision	KM
Project Administration	GO, KM
Funding	KM

References

Akashi, H. (1994). Synonymous Codon Usage in Drosophila-Melanogaster - Natural-Selection and Translational Accuracy. *Genetics* **136**, 927-935.

Alsford, S., Turner, D. J., Obado, S. O., Sanchez-Flores, A., Glover, L., Berriman, M., Hertz-Fowler, C. and Horn, D. (2011). High-throughput phenotyping using parallel sequencing of RNA interference targets in the African trypanosome. *Genome Research* **21**, 915-924.

Aslett, M., Aurrecoechea, C., Berriman, M., Brestelli, J., Brunk, B. P., Carrington, M., Depledge, D. P., Fischer, S., Gajria, B., Gao, X., et al. (2010). TriTrypDB: a functional genomic resource for the Trypanosomatidae. *Nucleic Acids Research* **38**, D457-D462.

Barnett, D. W., Garrison, E. K., Quinlan, A. R., Stromberg, M. P. and Marth, G. T. (2011). BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics* **27**, 1691-1692.

Berriman, M., Ghedin, E., Hertz-Fowler, C., Blandin, G., Renauld, H., Bartholomeu, D. C., Lennard, N. J., Caler, E., Hamlin, N. E., Haas, B., et al. (2005). The genome of the African trypanosome Trypanosoma brucei. *Science* **309**, 416-422.

Bolger, A. M., Lohse, M. and Usadel, B. (2014). Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114-2120.

Borst, P. (2002). Antigenic variation and allelic exclusion. *Cell* **109**, 5-8.

Borst, P. and Fase-Fowler, F. (1979). The maxi-circle of Trypanosoma brucei kinetoplast DNA. *Biochim Biophys Acta* **565**, 1-12.

Bose, R. and Heister, N. C. (1993). Development of Trypanosoma (M.) theileri in tabanids. *J Eukaryot Microbiol* **40**, 788-792.

Bruna, T., Hoff, K. J., Lomsadze, A., Stanke, M. and Borodovsky, M. (2021). BRAKER2: automatic eukaryotic genome annotation with GeneMark-EP+ and AUGUSTUS supported by a protein database. *NAR Genom Bioinform* **3**, lqaa108.

Bruna, T., Lomsadze, A. and Borodovsky, M. (2020). GeneMark-EP+: eukaryotic gene prediction with self-training in the space of genes and proteins. *NAR Genom Bioinform* **2**, lqaa026.

Buchfink, B., Xie, C. and Huson, D. H. (2015). Fast and sensitive protein alignment using DIAMOND. *Nature Methods* **12**, 59-60.

Burki, F. (2014). The eukaryotic tree of life from a global phylogenomic perspective. *Cold Spring Harb Perspect Biol* **6**, a016147.

Buscaglia, C. A., Campo, V. A., Frasch, A. C. and Di Noia, J. M. (2006). Trypanosoma cruzi surface mucins: host-dependent coat diversity. *Nat Rev Microbiol* **4**, 229-236.

Cabanettes, F. and Klopp, C. (2018). D-GENIES: dot plot large genomes in an interactive, efficient and simple way. *PeerJ* **6**, e4958.

Cavalier-Smith, T. (2010). Kingdoms Protozoa and Chromista and the eozoan root of the eukaryotic tree. *Biol Lett* **6**, 342-345.

Cayla, M., McDonald, L., MacGregor, P. and Matthews, K. (2020). An atypical DYRK kinase connects quorum-sensing with posttranscriptional gene regulation in Trypanosoma brucei. *Elife* **9**.

Cayla, M., Rojas, F., Silvester, E., Venter, F. and Matthews, K. R. (2019). African trypanosomes. *Parasit Vectors* **12**, 190.

Chainey, J. E. (1993). Horse-flies, deer-flies and clegs (Tabanidae). In *Medical insects and arachnids* (ed. R. P. Lane & R. W. Crosskey), pp. 310-332. Dordrecht: Springer Netherlands.

Chan, P. P., Lin, B. Y., Mak, A. J. and Lowe, T. M. (2021). tRNAscan-SE 2.0: improved detection and functional classification of transfer RNA genes. *Nucleic Acids Research* **49**, 9077-9096.

Clayton, C. (2019). Regulation of gene expression in trypanosomatids: living with polycistronic transcription. *Open Biol* **9**, 190072.

Cooper, D. and Eleftherianos, I. (2017). Memory and Specificity in the Insect Immune System: Current Perspectives and Future Challenges. *Front Immunol* **8**, 539.

Daniels, J. P., Gull, K. and Wickstead, B. (2010). Cell biology of the trypanosome genome. *Microbiol Mol Biol Rev* **74**, 552-569.

Doherty, M. L., Windle, H., Voorheis, H. P., Larkin, H., Casey, M., Clery, D. and Murray, M. (1993). Clinical disease associated with Trypanosoma theileri infection in a calf in Ireland. *Vet Rec* **132**, 653-656.

Domingo-Sananes, M. R., Szoor, B., Ferguson, M. A., Urbaniak, M. D. and Matthews, K. R. (2015). Molecular control of irreversible bistability during trypanosome developmental commitment. *J Cell Biol* **211**, 455-468.

Ellinghaus, D., Kurtz, S. and Willhoeft, U. (2008). LTRharvest, an efficient and flexible software for de novo detection of LTR retrotransposons. *Bmc Bioinformatics* **9**.

Emms, D. M. and Kelly, S. (2017). STRIDE: Species Tree Root Inference from Gene Duplication Events. *Mol Biol Evol* **34**, 3267-3278.

---- (2018). STAG : Species Tree Inference from All Genes. *BioRxiv*.

---- (2019). OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* **20**, 238.

---- (2020). Benchmarking Orthogroup Inference Accuracy: Revisiting Orthobench. *Genome Biology and Evolution* **12**, 2258-2266.

Erben, E. D., Fadda, A., Lueong, S., Hoheisel, J. D. and Clayton, C. (2014). A genome-wide tethering screen reveals novel potential post-transcriptional regulators in Trypanosoma brucei. *PLoS Pathog* **10**, e1004178.

Eyrewalker, A. C. (1991). An Analysis of Codon Usage in Mammals - Selection or Mutation Bias. *J Mol Evol* **33**, 442-449.

Farrar, R. G. and Klei, T. R. (1990). Prevalence of Trypanosoma theileri in Louisiana cattle. *J Parasitol* **76**, 734-736.

Flu, P. C. (1908). Über die Flagellaten im Darm von Melophagus ovinus. In *Arch. f. Protist.*, pp. 147.

Flynn, J. M., Hubley, R., Goubert, C., Rosen, J., Clark, A. G., Feschotte, C. and Smit, A. F. (2020). RepeatModeler2 for automated genomic discovery of transposable element families. *Proc Natl Acad Sci U S A* **117**, 9451-9457.

Funk, S., Nishiura, H., Heesterbeek, H., Edmunds, W. J. and Checchi, F. (2013). Identifying Transmission Cycles at the Human-Animal Interface: The Role of Animal Reservoirs in Maintaining Gambiense Human African Trypanosomiasis. *Plos Computational Biology* **9**, e1002855.

Gale, M., Jr., Carter, V. and Parsons, M. (1994). Translational control mediates the developmental regulation of the Trypanosoma brucei Nrk protein kinase. *J Biol Chem* **269**, 31659-31665.

Gale, M., Jr. and Parsons, M. (1993). A Trypanosoma brucei gene family encoding protein kinases with catalytic domains structurally related to Nek1 and NIMA. *Mol Biochem Parasitol* **59**, 111-121.

Gibson, W., Pilkington, J. G. and Pemberton, J. M. (2010). Trypanosoma melophagium from the sheep ked Melophagus ovinus on the island of St Kilda. *Parasitology* **137**, 1799-1804.

Gotoh, O. (2008). A space-efficient and accurate method for mapping and aligning cDNA sequences onto genomic sequence. *Nucleic Acids Res* **36**, 2630-2638.

Hoare, C. A. (1923). An Experimental Study of the Sheep-Trypanosome (T. melophagium Flu, 1908), and its Transmission by the Sheep-Ked (Melophagus ovinus L.). *Parasitology* **15**, 365-424.

---- (1929). Studies on Trypanosoma grayi. II. Experimental Transmission to the Crocodile. *Transactions of the Royal Society of Tropical Medicine and Hygiene* **23**.

---- (1931). Studies on Trypanosoma grayi. III. Life-cycle in the tsetse-fly and in the crocodile. *Parasitology* **23**, 449-484.

---- (1972). *The trypanosomes of mammals. A zoological monograph*. Oxford, England: Blackwell.

Hoff, K. J., Lange, S., Lomsadze, A., Borodovsky, M. and Stanke, M. (2016). BRAKER1: Unsupervised RNA-Seq-Based Genome Annotation with GeneMark-ET and AUGUSTUS. *Bioinformatics* **32**, 767-769.

Hoff, K. J., Lomsadze, A., Borodovsky, M. and Stanke, M. (2019). Whole-Genome Annotation with BRAKER. *Methods Mol Biol* **1962**, 65-95.

Hu, H., Gao, J., He, J., Yu, B., Zheng, P., Huang, Z. Q., Mao, X. B., Yu, J., Han, G. Q. and Chen, D. W. (2013). Codon Optimization Significantly Improves the Expression Level of a Keratinase Gene in Pichia pastoris. *Plos One* **8**.

Iwata, H. and Gotoh, O. (2012). Benchmarking spliced alignment programs including Spaln2, an extended version of Spaln that incorporates additional species-specific features. *Nucleic Acids Research* **40**, e161.

Jaskowska, E., Butler, C., Preston, G. and Kelly, S. (2015). Phytomonas: trypanosomatids adapted to plant environments. *PLoS Pathog* **11**, e1004484.

Jones, N. G., Thomas, E. B., Brown, E., Dickens, N. J., Hammarton, T. C. and Mottram, J. C. (2014a). Regulators of Trypanosoma brucei cell cycle progression and differentiation identified using a kinome-wide RNAi screen. *PLoS Pathog* **10**, e1003886.

Jones, P., Binns, D., Chang, H. Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014b). InterProScan 5: genome-scale protein function classification. *Bioinformatics* **30**, 1236-1240.

Kaufer, A., Ellis, J., Stark, D. and Barratt, J. (2017). The evolution of trypanosomatid taxonomy. *Parasit Vectors* **10**, 287.

Kelly, S., Ivens, A., Manna, P. T., Gibson, W. and Field, M. C. (2014). A draft genome for the African crocodilian trypanosome Trypanosoma grayi. *Sci Data* **1**, 140024.

Kelly, S., Ivens, A., Mott, G. A., O'Neill, E., Emms, D., Macleod, O., Voorheis, P., Tyler, K., Clark, M., Matthews, J., et al. (2017). An Alternative Strategy for Trypanosome Survival in the Mammalian Bloodstream Revealed through Genome and Transcriptome Analysis of the Ubiquitous Bovine Parasite Trypanosoma (Megatrypanum) theileri. *Genome Biology and Evolution* **9**, 2093-2109.

Kleisen, C. M. a. B. P. (1975). Sequence heterogeneity of the mini-circles of kinetoplast DNA of Crithidia luciliae and evidence for the presence of a component more complex than mini-circle DNA in the kinetoplast network. *Biochimica et Biophysica Acta (BBA) - Nucleic Acids and Protein Synthesis* **407**, 473-478.

Kniepert, F. W. (1980). Blood-feeding and nectar-feeding in adult tabanidae (Diptera). *Oecologia* **46**, 125-129.

Kriventseva, E. V., Kuznetsov, D., Tegenfeldt, F., Manni, M., Dias, R., Simao, F. A. and Zdobnov, E. M. (2019). OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. *Nucleic Acids Res* **47**, D807-D811.

Laetsch, D. R. and Blaxter, M. L. (2017). KinFin: Software for Taxon-Aware Analysis of Clustered Protein Sequences. *G3 (Bethesda)* **7**, 3349-3357.

Laetsch, D. R. a. B. M. L. (2017). BlobTools : Interrogation of genome assemblies. *F1000Research* **6**, 1287.

Lao, P. J. and Forsdyke, D. R. (2000). Thermophilic bacteria strictly obey Szybalski's transcription direction rule and politely purine-load RNAs with both adenine and guanine. *Genome Research* **10**, 228-236.

Leger, A. a. L. T. (2019). pycoQC , interactive quality control for Oxford Nanopore Sequencing. *The Journal of Open Source Software* **4**, 1236.

Leinonen, R., Sugawara, H., Shumway, M. and International Nucleotide Sequence Database, C. (2011). The sequence read archive. *Nucleic Acids Res* **39**, D19-21.

Letunic, I. and Bork, P. (2007). Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *Bioinformatics* **23**, 127-128.

Li, H. (2013). Aligning sequence reads, clone sequences and assembly contigs with BWA - MEM. *arXiv:1303.3997*.

Li, H. (2018). Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics* **34**, 3094-3100.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. and Genome Project Data Processing, S. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078-2079.

Ling, A. S., Trotter, J. R. and Hendriks, E. F. (2011). A zinc finger protein, TbZC3H20, stabilizes two developmentally regulated mRNAs in trypanosomes. *J Biol Chem* **286**, 20152-20162.

Liu, B., Kamanyi Marucha, K. and Clayton, C. (2020). The zinc finger proteins ZC3H20 and ZC3H21 stabilise mRNAs encoding membrane proteins and mitochondrial proteins in insect-form Trypanosoma brucei. *Mol Microbiol* **113**, 430-451.

Lomsadze, A., Burns, P. D. and Borodovsky, M. (2014). Integration of mapped RNA-Seq reads into automatic training of eukaryotic gene finding algorithm. *Nucleic Acids Research* **42**, e119.

Lueong, S., Merce, C., Fischer, B., Hoheisel, J. D. and Erben, E. D. (2016). Gene expression regulatory networks in Trypanosoma brucei: insights into the role of the mRNA-binding proteome. *Mol Microbiol* **100**, 457-471.

Marcais, G. and Kingsford, C. (2011). A fast, lock-free approach for efficient parallel counting of occurrences of k-mers. *Bioinformatics* **27**, 764-770.

Martinkovic, F., Matanovic, K., Rodrigues, A. C., Garcia, H. A. and Teixeira, M. M. (2012). Trypanosoma (Megatrypanum) melophagium in the sheep ked Melophagus ovinus from organic farms in Croatia: phylogenetic inferences support restriction to sheep and sheep keds and close relationship with trypanosomes from other ruminant species. *J Eukaryot Microbiol* **59**, 134-144.

Maslov, D. A., Opperdoes, F. R., Kostygov, A. Y., Hashimi, H., Lukes, J. and Yurchenko, V. (2019). Recent advances in trypanosomatid research: genome organization, expression, metabolism, taxonomy and evolution. *Parasitology* **146**, 1-27.

Matthews, D. M., Kingston, N., Maki, L. and Nelms, G. (1979). Trypanosoma theileri Laveran, 1902, in Wyoming cattle. *Am J Vet Res* **40**, 623-629.

Mehlitz, D. and Molyneux, D. H. (2019). The elimination of Trypanosoma brucei gambiense? Challenges of reservoir hosts and transmission cycles: Expect the unexpected. *Parasite Epidemiol Control* **6**, e00113.

Melville, S. E., Leech, V., Gerrard, C. S., Tait, A. and Blackwell, J. M. (1998). The molecular karyotype of the megabase chromosomes of Trypanosoma brucei and the assignment of chromosome markers. *Mol Biochem Parasitol* **94**, 155-173.

Mistry, J., Chuguransky, S., Williams, L., Qureshi, M., Salazar, G. A., Sonnhammer, E. L. L., Tosatto, S. C. E., Paladin, L., Raj, S., Richardson, L. J., et al. (2021). Pfam: The protein families database in 2021. *Nucleic Acids Res* **49**, D412-D419.

Mony, B. M., MacGregor, P., Ivens, A., Rojas, F., Cowton, A., Young, J., Horn, D. and Matthews, K. (2014). Genome-wide dissection of the quorum sensing signalling pathway in Trypanosoma brucei. *Nature* **505**, 681-685.

Mony, B. M. and Matthews, K. R. (2015). Assembling the components of the quorum sensing pathway in African trypanosomes. *Mol Microbiol* **96**, 220-232.

Mott, G. A., Wilson, R., Fernando, A., Robinson, A., MacGregor, P., Kennedy, D., Schaap, D., Matthews, J. B. and Matthews, K. R. (2011). Targeting Cattle-Borne Zoonoses and Cattle Pathogens Using a Novel Trypanosomatid-Based Delivery System. *Plos Pathogens* **7**, e1002340.

Muller, I. B., Domenicali-Pfister, D., Roditi, I. and Vassella, E. (2002). Stage-specific requirement of a mitogen-activated protein kinase by Trypanosoma brucei. *Mol Biol Cell* **13**, 3787-3799.

Nardy, A. F., Freire-de-Lima, C. G., Perez, A. R. and Morrot, A. (2016). Role of Trypanosoma cruzi Trans-sialidase on the Escape from Host Immune Surveillance. *Front Microbiol* **7**, 348.

Nielsen, H. (2017). Predicting Secretory Proteins with SignalP. *Methods Mol Biol* **1611**, 59-73.

Novoa, E. M. and de Pouplana, L. R. (2012). Speeding with control: codon usage, tRNAs, and ribosomes. *Trends Genet* **28**, 574-581.

Oldrieve, G., Verney, M., Jaron, K. S., Hebert, L. and Matthews, K. R. (2021). Monomorphic Trypanozoon: towards reconciling phylogeny and pathologies. *Microb Genom* **7**.

Parsons, M., Nelson, R. G., Watkins, K. P. and Agabian, N. (1984). Trypanosome mRNAs share a common 5' spliced leader sequence. *Cell* **38**, 309-316.

Pays, E., Vanhamme, L. and Perez-Morga, D. (2004). Antigenic variation in Trypanosoma brucei: facts, challenges and mysteries. *Curr Opin Microbiol* **7**, 369-374.

Paz, A., Mester, D., Baca, I., Nevo, E. and Korol, A. (2004). Adaptive role of increased frequency of polypurine tracts in mRNA sequences of thermophilic prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 2951-2956.

Peacock, L., Kay, C., Farren, C., Bailey, M., Carrington, M. and Gibson, W. (2021). Sequential production of gametes during meiosis in trypanosomes. *Commun Biol* **4**.

Plotkin, J. B., Robins, H. and Levine, A. J. (2004). Tissue-specific codon usage and the expression of human genes. *Proceedings of the National Academy of Sciences of the United States of America* **101**, 12588-12591.

Podlipaev, S. (2001). The more insect trypanosomatids under study-the more diverse Trypanosomatidae appears. *Int J Parasitol* **31**, 648-652.

Pool, J. E. and Nielsen, R. (2007). Population size changes reshape genomic patterns of diversity. *Evolution* **61**, 3001-3006.

Porter, A. The Structure and Life-history of Crithidia melophagia (Flu), an Endoparasite of the Sheep Ked, Melophagus ovinus.

Rao, Y. S., Wu, G. Z., Wang, Z. F., Chai, X. W., Nie, Q. H. and Zhang, X. Q. (2011). Mutation Bias is the Driving Force of Codon Usage in the Gallus gallus genome. *DNA Res* **18**, 499-512.

Ruan, J. and Li, H. (2020). Fast and accurate long-read assembly with wtdbg2. *Nat Methods* **17**, 155-158.

Sanchez-Moreno, M., Lasztity, D., Coppens, I. and Opperdoes, F. R. (1992). Characterization of carbohydrate metabolism and demonstration of glycosomes in a Phytomonas sp. isolated from Euphorbia characias. *Mol Biochem Parasitol* **54**, 185-199.

Schlafer, D. H. (1979). Trypanosoma theileri: a literature review and report of incidence in New York cattle. *Cornell Vet* **69**, 411-425.

Schnaufer, A., Domingo, G. J. and Stuart, K. (2002). Natural and induced dyskinetoplastic trypanosomatids: how to live without mitochondrial DNA. *International Journal for Parasitology* **32**, 1071-1084.

Seifi, H. A. (1995). Clinical Trypanosomosis Due to Trypanosoma-Theileri in a Cow in Iran. *Tropical Animal Health and Production* **27**, 93-94.

Seppey, M., Manni, M. and Zdobnov, E. M. (2019). BUSCO: Assessing Genome Assembly and Annotation Completeness. *Methods Mol Biol* **1962**, 227-245.

Serpil (2008). Trypanosoma melophagium in blood cell culture. *Ankara Üniversitesi Veteriner Fakültesi Dergisi* **50**, 1-1.

Seward, E. A. and Kelly, S. (2016). Dietary nitrogen alters codon bias and genome composition in parasitic microorganisms. *Genome Biology* **17**.

---- (2018). Selection-driven cost-efficiency optimization of transcripts modulates gene evolutionary rate in bacteria. *Genome Biology* **19**.

Shah, P. and Gilchrist, M. A. (2011). Explaining complex codon usage patterns with selection for translational efficiency, mutation bias, and genetic drift. *Proceedings of the National Academy of Sciences of the United States of America* **108**, 10231-10236.

Simpson, L. (1979). Isolation of maxicircle component of kinetoplast DNA from hemoflagellate protozoa. *Proc Natl Acad Sci U S A* **76**, 1585-1588.

Slowikowski, K. (2018). ggrepel: Automatically Position Non-Overlapping Text Labels with'ggplot2'. *R package version 0.8. 0 ed*.

Sorensen, M. A., Kurland, C. G. and Pedersen, S. (1989). Codon Usage Determines Translation Rate in Escherichia-Coli. *J Mol Biol* **207**, 365-377.

Stanke, M., Diekhans, M., Baertsch, R. and Haussler, D. (2008). Using native and syntenically mapped cDNA alignments to improve de novo gene finding. *Bioinformatics* **24**, 637-644.

Stanke, M., Schoffmann, O., Morgenstern, B. and Waack, S. (2006). Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources. *Bmc Bioinformatics* **7**, 62.

Steinert, M. (1960). Mitochondria Associated with the Kinetonucleus of Trypanosoma Mega. *J Biophys Biochem Cytol* **8**, 542-546.

Swingle, L. D. (1911). The Relation of Crithidia melophagia to the Sheep's Blood, with Remarks upon the Controversy between Dr. Porter and Dr. Woodcock. *Transactions of the American Microscopical Society* **30**, 275.

Szoor, B., Ruberto, I., Burchmore, R. and Matthews, K. R. (2010). A novel phosphatase cascade regulates differentiation in Trypanosoma brucei via a glycosomal signaling pathway. *Genes Dev* **24**, 1306-1316.

Szoor, B., Wilson, J., McElhinney, H., Tabernero, L. and Matthews, K. R. (2006). Protein tyrosine phosphatase TbPTP1: A molecular switch controlling life cycle differentiation in trypanosomes. *J Cell Biol* **175**, 293-303.

Team, R. C. (2019). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing.

Toft, C. and Andersson, S. G. E. (2010). Evolutionary microbial genomics: insights into bacterial host adaptation. *Nat Rev Genet* **11**, 465-475.

Toh, J. Y., Nkouawa, A., Sanchez, S. R., Shi, H., Kolev, N. G. and Tschudi, C. (2021). Identification of positive and negative regulators in the stepwise developmental progression towards infectivity in Trypanosoma brucei. *Sci Rep* **11**, 5755.

Trindade, S., Rijo-Ferreira, F., Carvalho, T., Pinto-Neves, D., Guegan, F., Aresta-Branco, F., Bento, F., Young, S. A., Pinto, A., Van den Abbeele, J., et al. (2016). Trypanosoma brucei Parasites Occupy and Functionally Adapt to the Adipose Tissue in Mice. *Cell Host Microbe* **19**, 837-848.

Ullu, E., Tschudi, C. and Chakraborty, T. (2004). RNA interference in protozoan parasites. *Cellular Microbiology* **6**, 509-519.

Underwood, W. J., Blauwiekel, R., Delano, M. L., Gillesby, R., Mischler, S. A. and Schoell, A. (2015). Biology and diseases of ruminants (sheep, goats, and cattle). In *Laboratory Animal Medicine*, pp. 623-694: Elsevier.

Vaser, R., Sovic, I., Nagarajan, N. and Sikic, M. (2017). Fast and accurate de novo genome assembly from long uncorrected reads. *Genome Res* **27**, 737-746.

Vurture, G. W., Sedlazeck, F. J., Nattestad, M., Underwood, C. J., Fang, H., Gurtowski, J. and Schatz, M. C. (2017). GenomeScope: fast reference-free genome profiling from short reads. *Bioinformatics* **33**, 2202-2204.

Walker, B. J., Abeel, T., Shea, T., Priest, M., Abouelliel, A., Sakthikumar, S., Cuomo, C. A., Zeng, Q., Wortman, J., Young, S. K., et al. (2014). Pilon: an integrated tool for comprehensive microbial variant detection and genome assembly improvement. *PLoS One* **9**, e112963.

Walrad, P., Paterou, A., Acosta-Serrano, A. and Matthews, K. R. (2009). Differential trypanosome surface coat regulation by a CCCH protein that co-associates with procyclin mRNA cis-elements. *PLoS Pathog* **5**, e1000317.

Wang, F., Ekiert, D. C., Ahmad, I., Yu, W., Zhang, Y., Bazirgan, O., Torkamani, A., Raudsepp, T., Mwangi, W., Criscitiello, M. F., et al. (2013). Reshaping antibody diversity. *Cell* **153**, 1379-1393.

Wickham, H. (2016). ggplot2: elegant graphics for data analysis.

Wickstead, B., Ersfeld, K. and Gull, K. (2004). The small chromosomes of Trypanosoma brucei involved in antigenic variation are constructed around repetitive palindromes. *Genome Res* **14**, 1014-1024.

Yao, C. Q. (2010). Major Surface Protease of Trypanosomatids: One Size Fits All? *Infect Immun* **78**, 22-31.

Figures and Tables

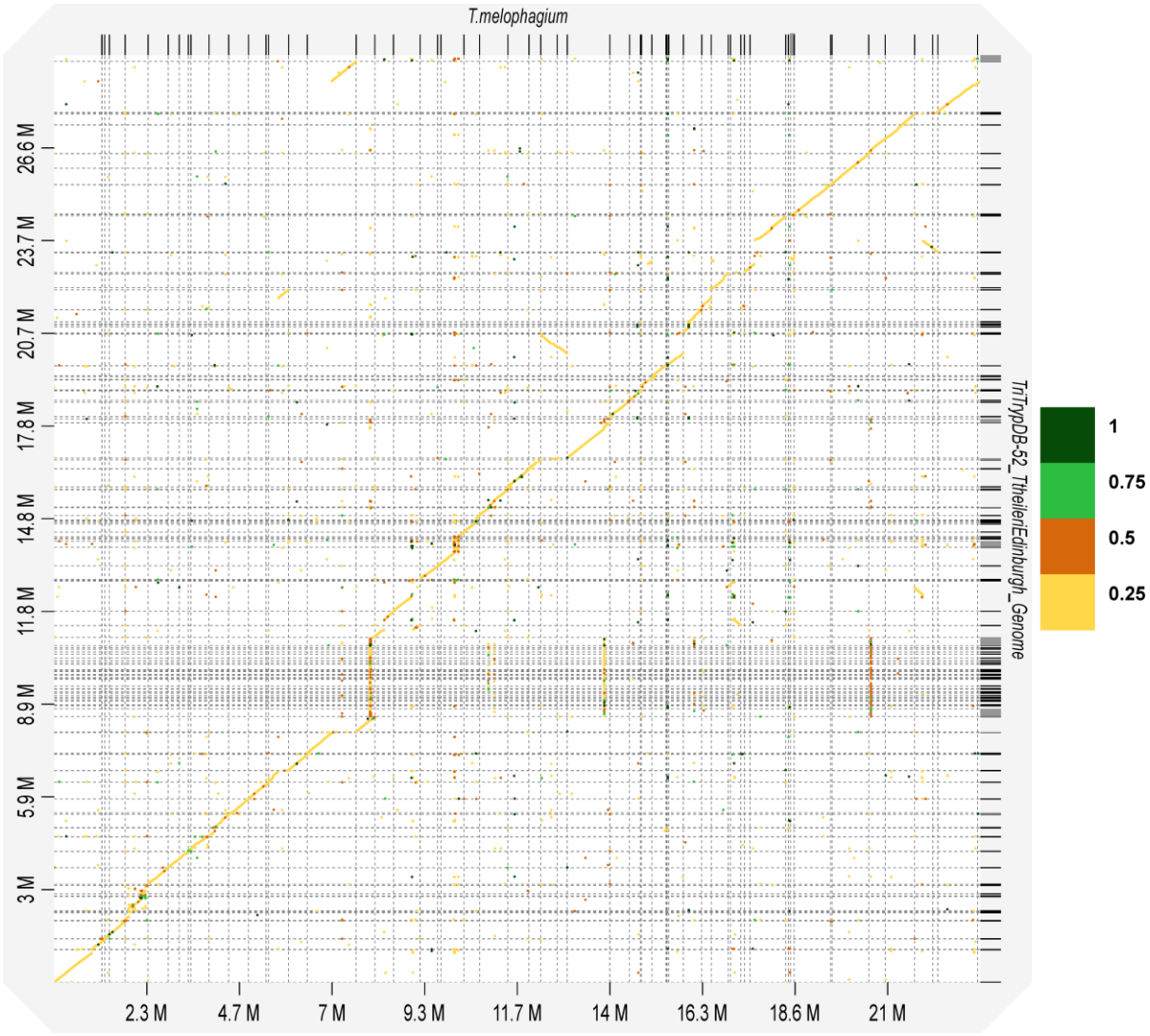

Figure 1: Synteny of the *T. melophagium* and *T. theileri* genome sequences highlights conservation and identity between the two species. The legend refers to the percentage identity between the sequences.

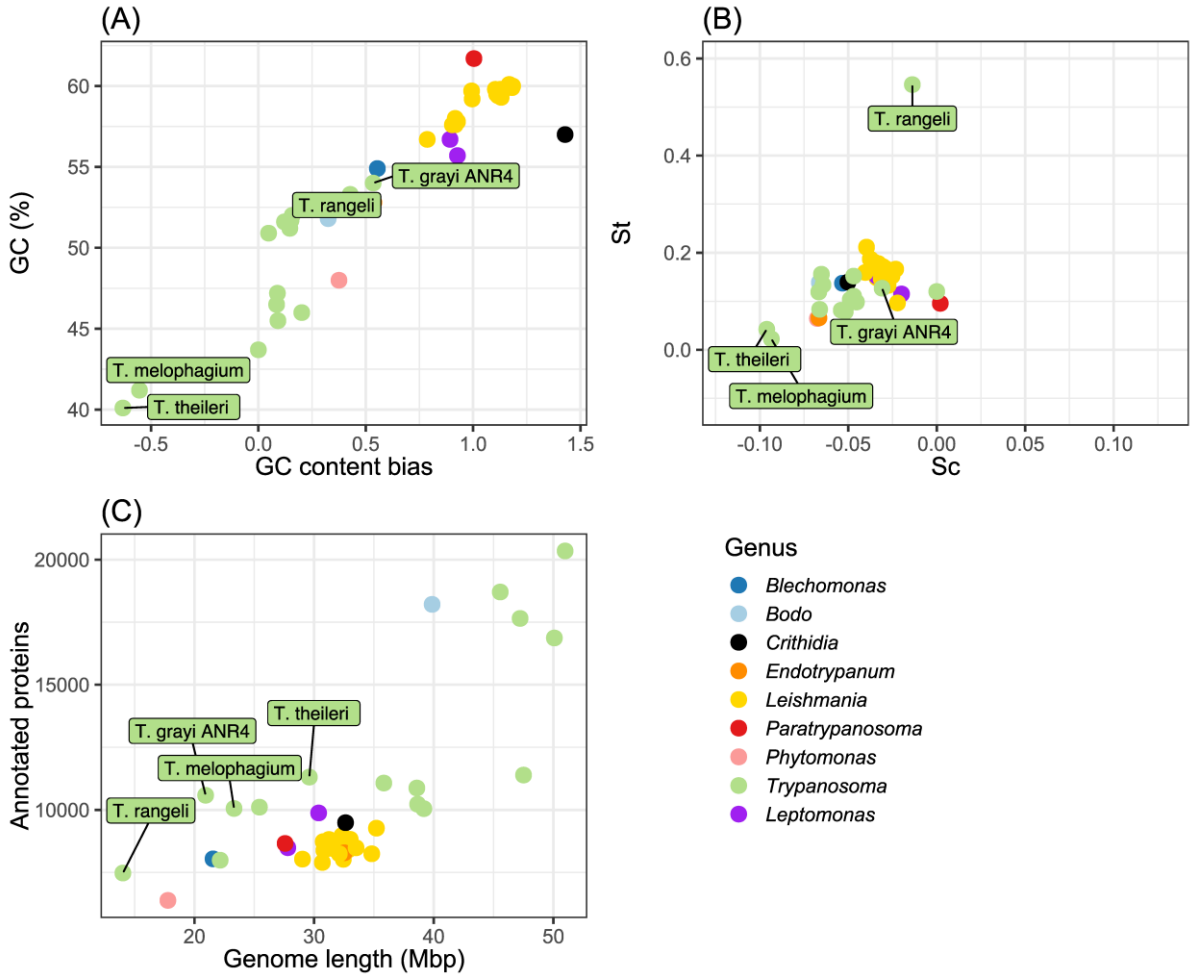

Figure 2: (A) GC content across the whole genome and GC content bias in the CDS of trypanosomatid universal single copy orthologues (n=992). GC content (GC) > 0 = GC content bias. GC < 0 = AT content bias. (B) Selection acting on translational efficiency (St) and selection acting on nucleotide cost (Sc) in trypanosomatid universal single copy orthologues. Sc > 0 = Selection acting to increase codon nucleotide cost. Sc < 0 = Selection is acting to decrease codon nucleotide cost. St > 0 = Selection is acting to increase codon translational efficiency. St < 0 = Selection acting to decrease codon translational efficiency. (C) Counts of annotated protein sequences of publicly available trypanosomatids compared by genome size.

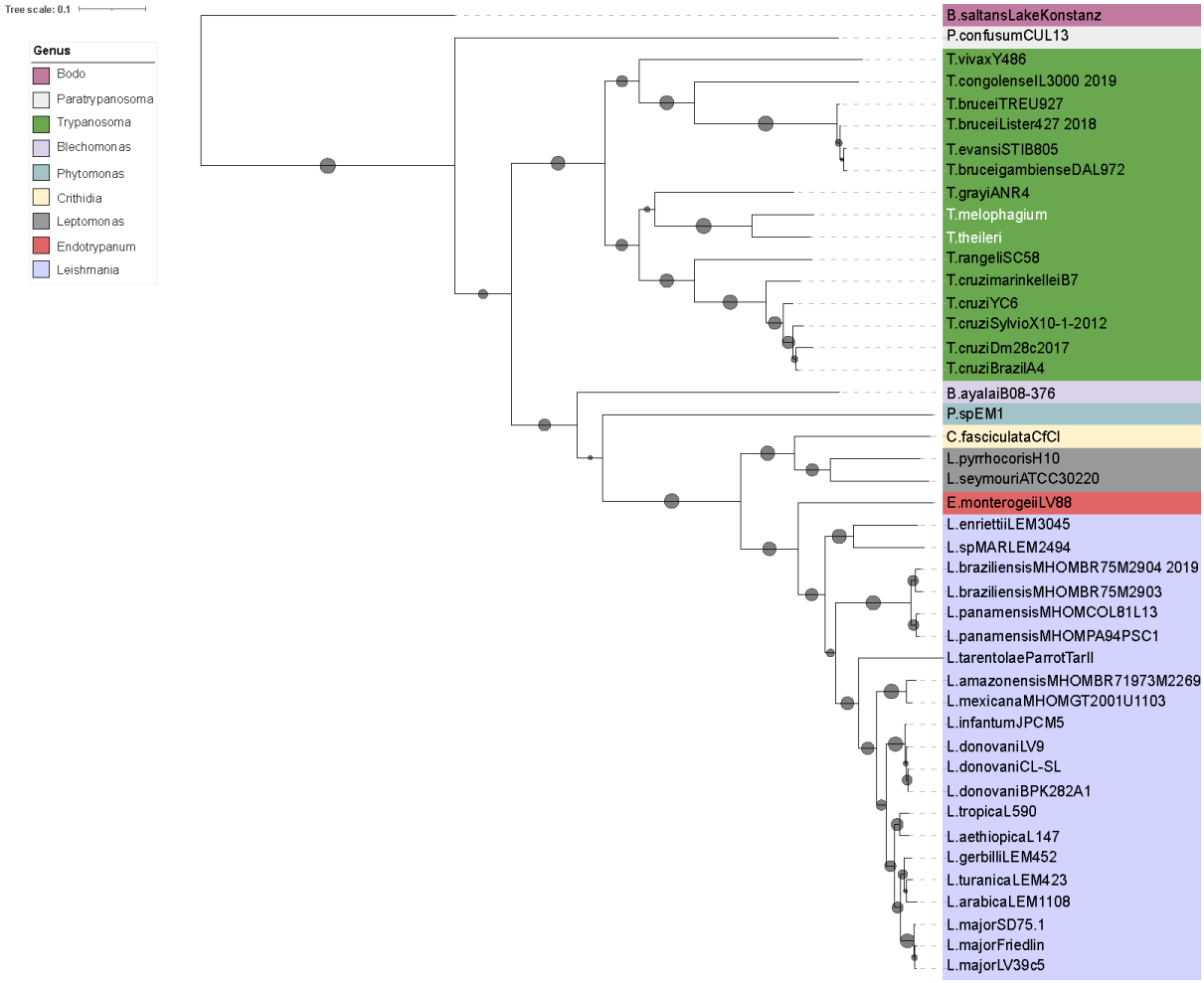

Figure 3: Species consensus tree based on 2,312 species trees created by STAG and STRIDE, OrthoFinder. The support values are represented by circles. Support values correlate to the proportion of times that the bipartition is seen in each of the individual trees used to create the consensus tree. The scale represents substitutions per site.

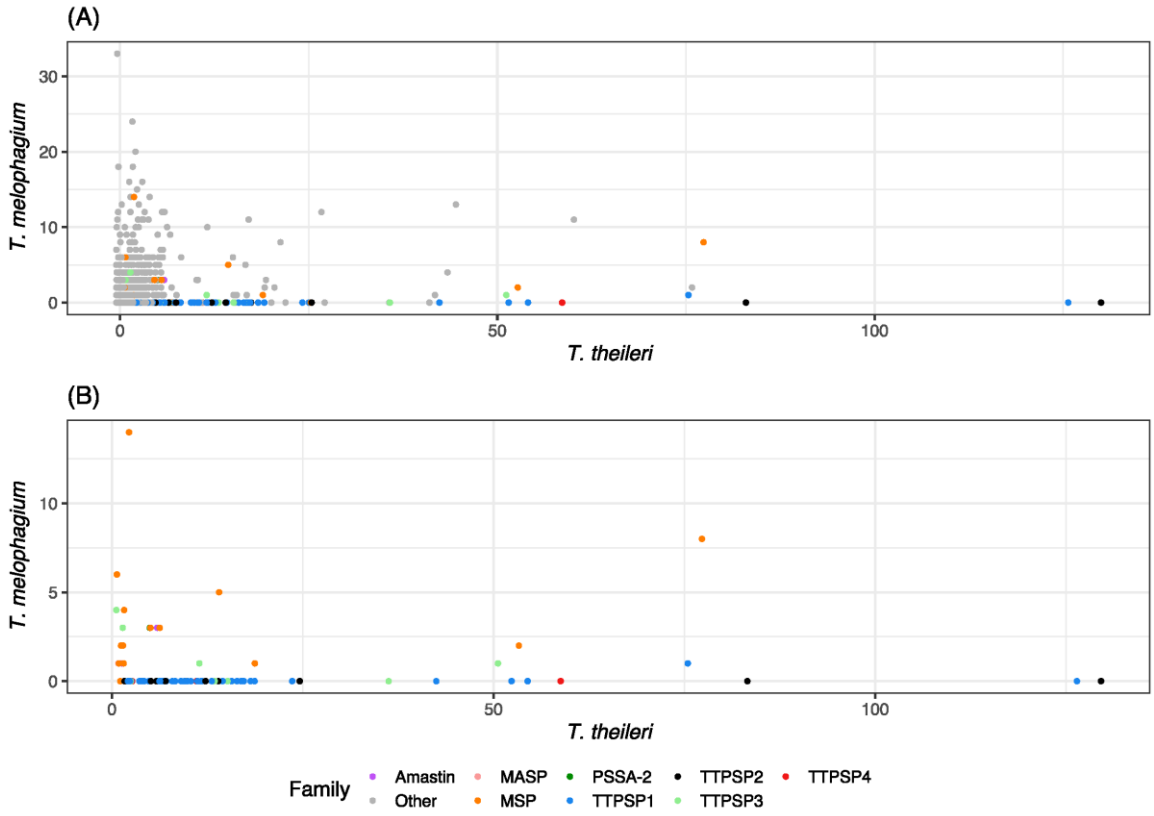

Figure 4: (A) All orthogroups and (B) orthogroups associated with host interaction size comparison between *T. melophagium* and *T. theileri*. Each dot represents the numbers of genes found in each orthogroup for both species. The orthogroups have been annotated with their designation as either a putative cell surface protein family or 'other' (Kelly et al., 2017).

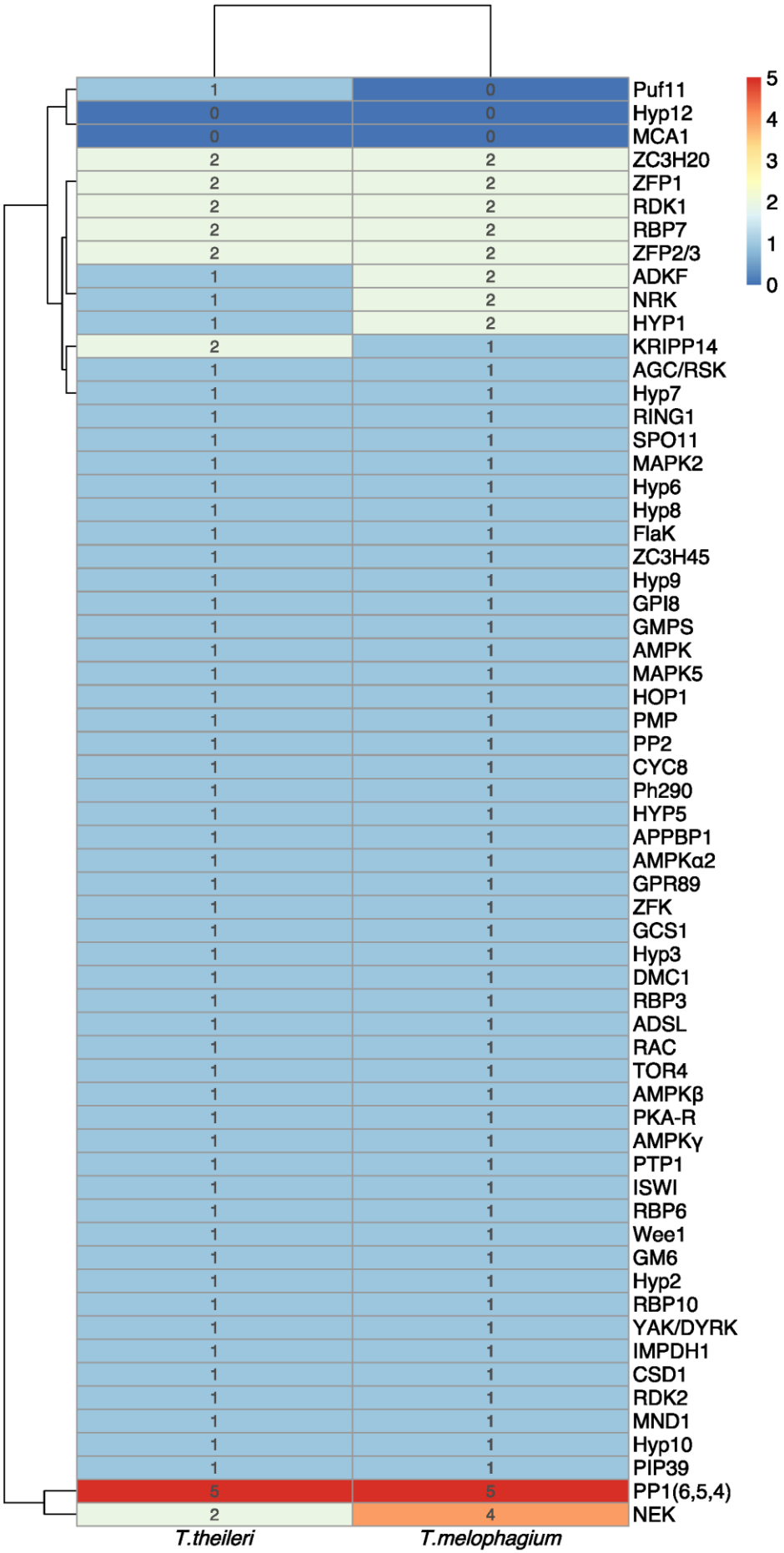

Figure 5: Genes associated with development, and related proteins, at various stages throughout the *T. brucei* life cycle. The number of genes in orthogroups associated with developmental regulation have been quantified in *T. theileri* and *T. melophagium*.

Table 1: Genome assessment. The assemblies used in this assessment represent the final draft of the *T. melophagium* assembly produced during this study and the *T. theileri* assembly available from TriTrypDB (Aslett et al., 2010; Kelly et al., 2017). The k-mer spectra plots associated with the first section of the table are found in Fig. S1.

	T. melophagium	***T. theileri*** (32)
k-mer based genome survey		
Heterozygosity (%)	0.30	0.41
Genome length (Mbp) / Repeat / Unique	22.28 / 4.16 / 18.13	27.62 / 7.97 / 19.65
Genome assembly statistics (for scaffolds longer than 200 bp)		
Number of contigs	64	253
Length (Mbp)	23.3	29.6
Minimium / Maximium / Mean (bp)	5,186 / 1,230,212 / 364,120	791 / 1,635,300 / 117,005
N50	505,851	517,122
GC (%)	41.2	40.1
Genome BUSCO assessment		
Complete/ Single copy/ Duplicated/ Fragmented	100 / 100 / 0 / 0	99.2 / 99.2 / 0 / 0.8
Annotation		
Number of genes	10,057	11,312
Annotation BUSCO assessment		
Complete/ Single copy/ Duplicated/ Fragmented	100 / 100 / 0 / 0	99.2 / 99.2 / 0 / 0.8

Table 2: Orthologous protein clustering statistics of *T. melophagium* and *T. theileri*. The full species-specific clustering summary can be found in File S3.

	T. melophagium	***T. theileri***
Number of genes	10,057	11,312
Genes in orthogroups (%)	97.3	97.4
Orthogroups containing species (%)	43.9	44.2
Number of species-specific orthogroups	76	121
Genes in species-specific orthogroups (%)	2.7	12.9
Comparatively expanded orthogroups	787	563

Table 3: Cell surface orthogroup counts from Kelly *et al*. (2017) (32) and this study along with counts of genes present in each category.

Cell surface conservation	**Annotation**	**Orthogroup count**		**Gene count**	
		Kelly et al., 2017	This study	*T. theileri*	*T. melophagium*
Conserved	Amastin	1	2	7	4
Conserved	MASP	1	1	1	1
Conserved	MSP	1	18	229	52
Conserved	PSSA-2	1	1	5	3
Unique	TTPSP1	1	42	720	1
Unique	TTPSP2	1	10	301	0
Unique	TTPSP3	1	11	157	9
Unique	TTPSP4	1	3	73	0

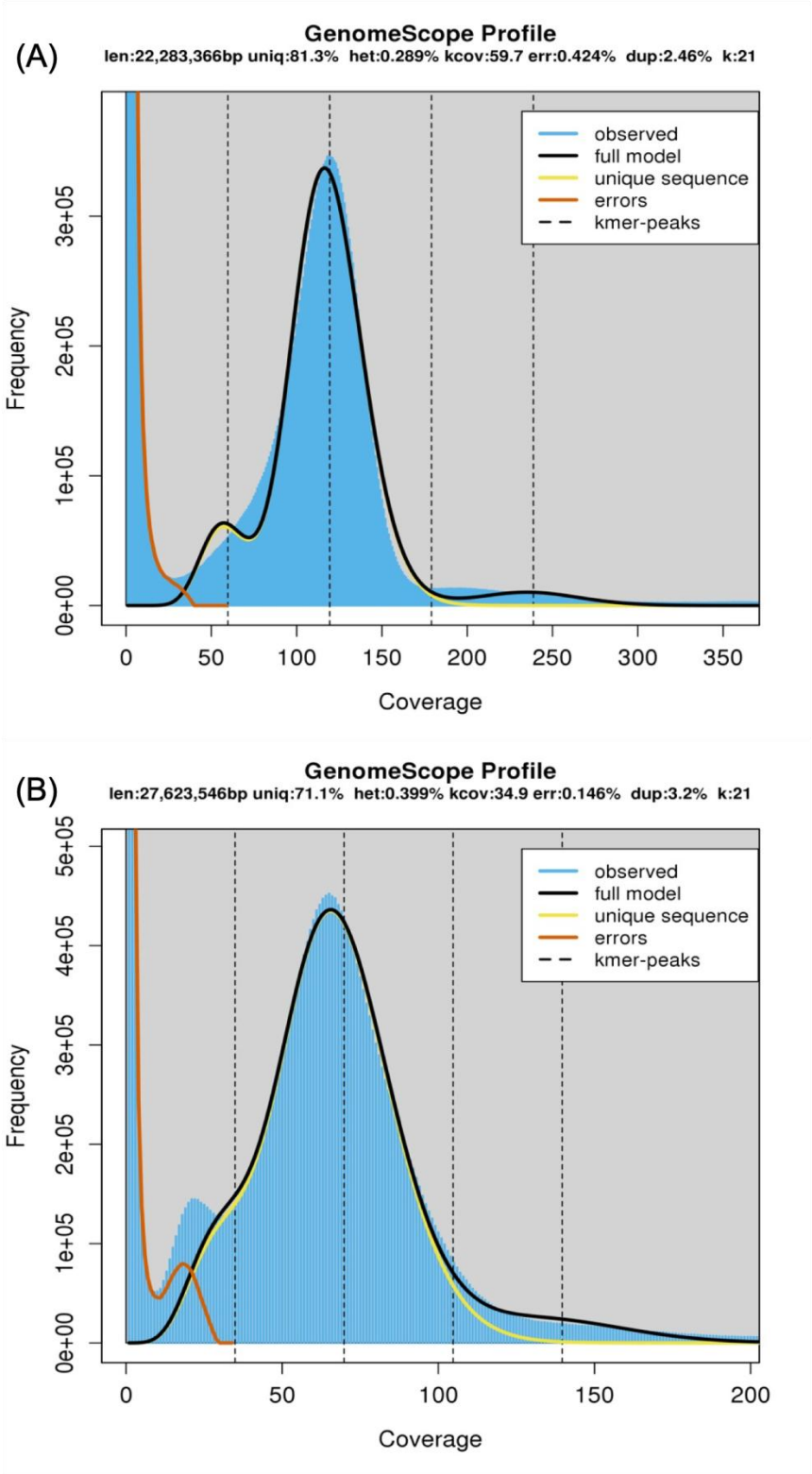

Fig. S1. k-mer spectra of (A) *T. melophagium* and (B) *T. theileri* using a k-mer size of 21. The statistics presented in these images represent the lowest estimate size, the highest estimate was used in Table 1 as these had a greater model fit.

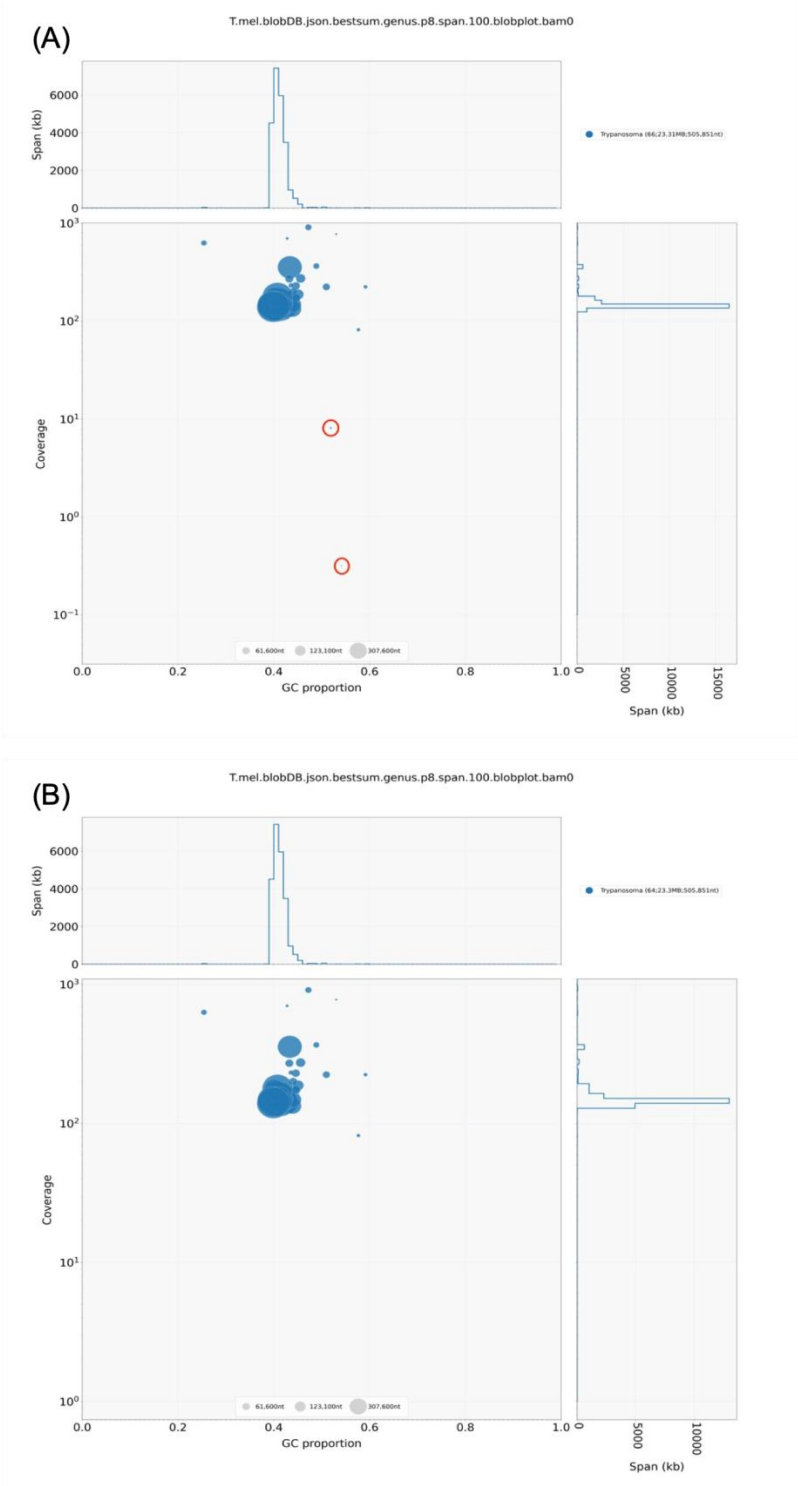

Fig. S2. Coverage and blast annotation for (A) the polished assembly and (B) the manually trimmed assembly.

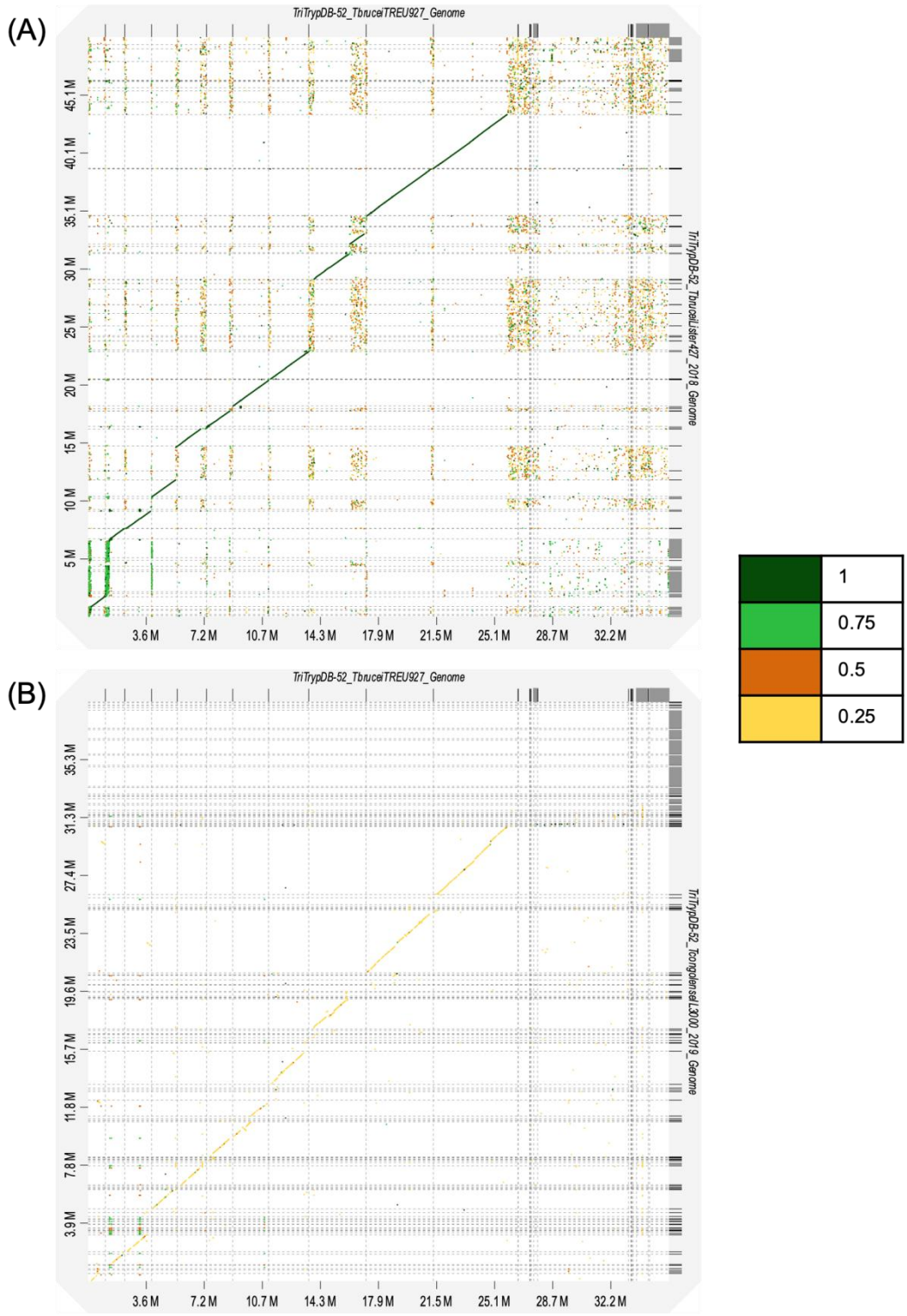

Fig. S3. Synteny plot of (A) *T. brucei* TREU927/4 and *T. brucei* Lister 427 (2018) and (B) *T. brucei* TREU927/4 *and T. congolense* IL3000 2019 genome sequences. The legend refers to the identity between the sequences.

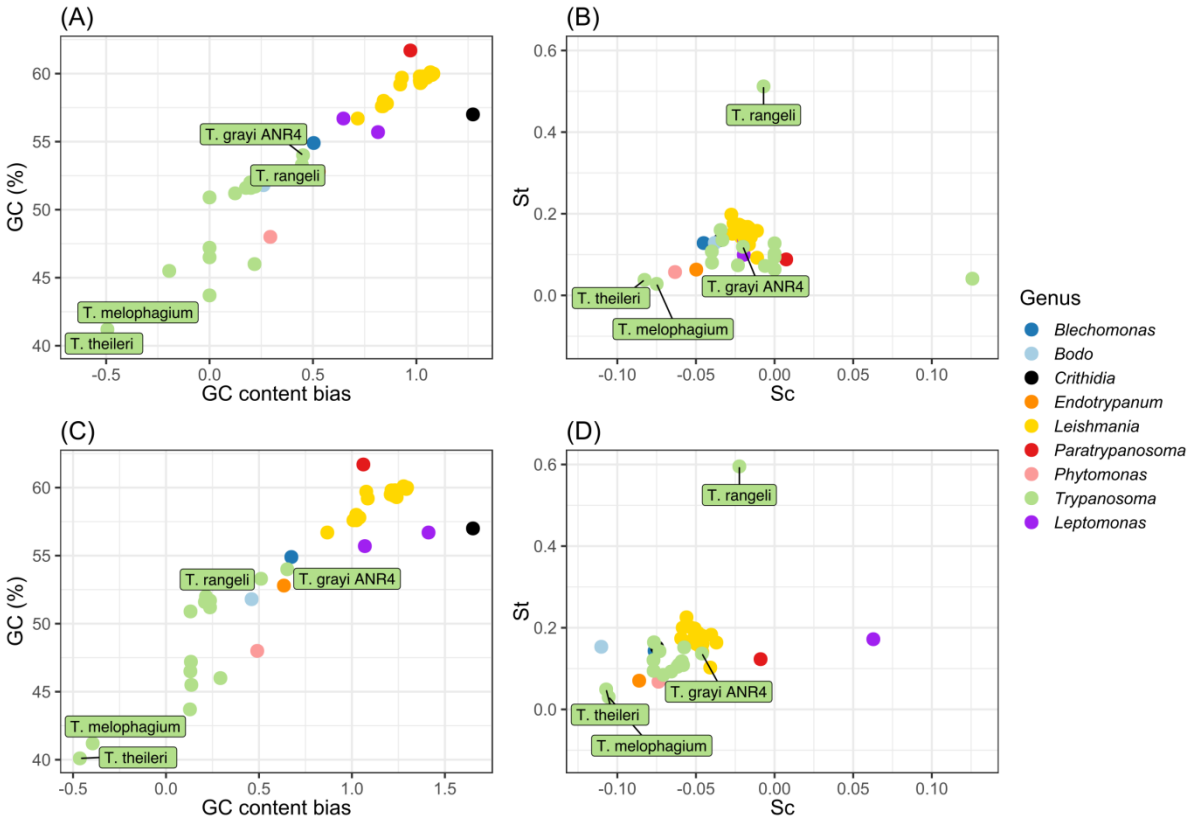

Fig. S4. (A) GC content across the whole genome and GC content bias in a every CDS. (B) Selection acting on translational efficiency (St) and selection acting on nucleotide cost (Sc). (C) GC content across the whole genome and GC content bias in CDS of a trypanosomatid universal single copy orthologue which is essential in every life cycle stage in *T. brucei* (n=158). (D) Selection acting on translational efficiency (St) and selection acting on nucleotide cost (Sc) in trypanosomatid universal single copy orthologues which is essential in every life cycle stage in *T. brucei* (n=158).

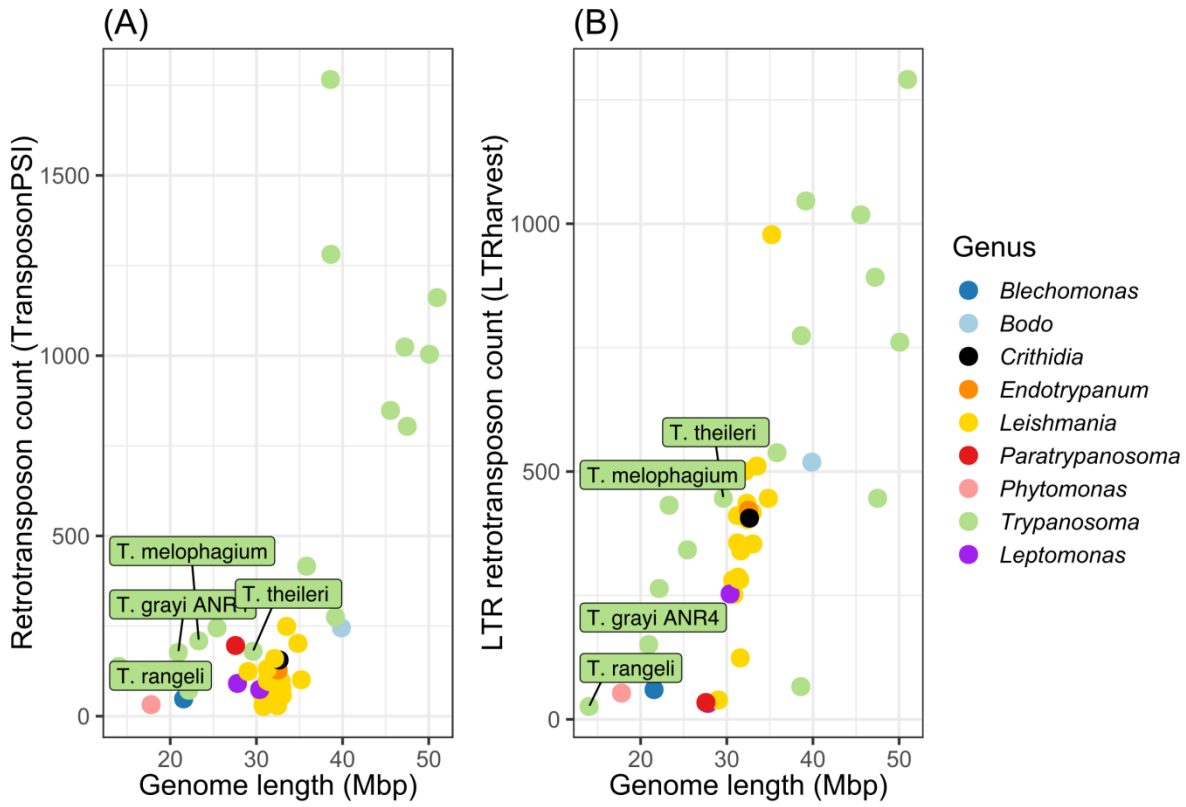

Fig. S5. (A) Retrotransposon (B) long terminal repeat (LTR) retrotransposon counts compared to genome length.

Table S1. Software and any options used.

Tool	Version	Flag	Available at
Guppy	4.0.15	--flowcell FLO-MIN106 --kit SQK-RAD004	https://community.nanoporetech.com/
PycoQC	2.5.0.3		https://github.com/a-slide/pycoQC
Trimmomatic	0.39	-SLIDINGWINDOW 4:20 -MINLEN 50	https://github.com/timflutre/trimmomatic
GenomeScope	1	-Kmer 21 -Max kmer cov. 1000	http://qb.cshl.edu/genomescope/
Jellyfish	2.2.10	-C -m 21 -s 1000000000	https://github.com/gmarcais/Jellyfish
Wtdbg2	3	-x ont -g 30m -A -t 10	https://github.com/ruanjue/wtdbg2
Minimap2	2.17-r941	-ax map-ont	https://github.com/lh3/minimap2
Racon	v1.4.13	-m 8 -x -6 -g -8 -w 500	https://github.com/isovic/racon
Medaka	V5	-m r941_min_high_g360	https://github.com/nanoporetech/medaka
BWA-MEM	0.7.17-r1188		https://github.com/lh3/bwa
Pilon	1.24	--genome --diploid –bam --bam	https://github.com/broadinstitute/pilon
Blast	2.6.0+	blastn -task megablast -outfmt '6 qseqid	ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/

		staxids bitscore std' -max_target_seqs 1 -max_hsps 1 -evalue 1e-25 or 1e-5	
DIAMOND	2.0.5.143	Blastx --sensitive --max-target-seqs 1 --evalue 1e-25	http://www.diamondsearch.org
BlobTools	1.1.1	-r genus	https://github.com/DRL/blobtools
Scaffold_stats.pl	N/A		https://github.com/sujaikumar/assemblage
BUSCO	V5	-l euglenozoa_odb10	https://busco.ezlab.org/
RepeatModeler	N/A	-database - LTRStruct -pa 8	http://www.repeatmasker.org/RepeatModeler/
RepeatMasker	N/A	-lib consensi.fa.classified -pa 8 -xsmall -nolow –gff	http://www.repeatmasker.org/RMDownload.html
BRAKER2	2.1.6	--prot_seq --bam --softmasking --etpmode --gff3	https://github.com/Gaius-Augustus/BRAKER
D-Genies	1.2.0	-Minimap2	http://dgenies.toulouse.inra.fr/run
tRNAscan-SE	2.0.9	-E	http://lowelab.ucsc.edu/tRNAscan-SE/
CodonMuse	0.1.0	-f -tscan	https://github.com/easeward/CodonMuSe

TransposonPSI	1.0.0	-nuc	http://transposonpsi.sourceforge.net
gt suffixerator	1.6.2	-tis -suf -lcp -des -ssp -sds -dna	http://genometools.org/tools/gt_ltrharvest.html
gt LTRharvest	1.6.2		http://genometools.org/tools/gt_ltrharvest.html
OrthoFinder	2.5.2	-S diamond_ultra_sens	https://github.com/davidemms/OrthoFinder
InterProScan	5.52-86.0	--dp --goterms -appl SignalP-EUK-4.1, Pfam -f TSV	https://www.ebi.ac.uk/interpro/
KinFin	1.0.3		https://github.com/DRL/kinfin
iTOL	6		https://itol.embl.de
R	3.6.1		https://www.r-project.org/
ggplot2	3.3.3		https://cran.r-project.org/web/packages/ggplot2/index.html
ggrepel	0.9.1		https://cran.r-project.org/web/packages/ggrepel/index.html

Table S2.

Click here to download Table S2

Table S3.

Click here to download Table S3

Table S4.

Click here to download Table S4

Table S5.

Click here to download Table S5